



CD-ROM includes
statistical software

Primer of Biostatistics

SEVENTH EDITION

Stanton A. Glantz

Primer of Biostatistics

Notice

Medicine is an ever-changing science. As new research and clinical experience broaden our knowledge, changes in treatment and drug therapy are required. The author and the publisher of this work have checked with sources believed to be reliable in their efforts to provide information that is complete and generally in accord with the standards accepted at the time of publication. However, in view of the possibility of human error or changes in medical sciences, neither the author nor the publisher nor any other party who has been involved in the preparation or publication of this work warrants that the information contained herein is in every respect accurate or complete, and they disclaim all responsibility for any errors or omissions or for the results obtained from use of the information contained in this work. Readers are encouraged to confirm the information contained herein with other sources. For example, and in particular, readers are advised to check the product information sheet included in the package of each drug they plan to administer to be certain that the information contained in this work is accurate and that changes have not been made in the recommended dose or in the contraindications for administration. This recommendation is of particular importance in connection with new or infrequently used drugs.

Primer of Biostatistics

Seventh Edition

Stanton A. Glantz, PhD

Professor of Medicine

American Legacy Foundation Distinguished Professor in Tobacco Control

Director, Center for Tobacco Control Research and Education

Member, Cardiovascular Research Institute

Member, Philip R. Lee Institute for Health Policy Studies

Member, Helen Diller Family Comprehensive Cancer Center

University of California, San Francisco

San Francisco, California



New York Chicago San Francisco Lisbon London Madrid Mexico City
Milan New Delhi San Juan Seoul Singapore Sydney Toronto

Copyright © 2012 by The McGraw-Hill Companies, Inc. All rights reserved. Except as permitted under the United States Copyright Act of 1976, no part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.

ISBN: 978-0-07-179440-4

MHID: 0-07-179440-9

The material in this eBook also appears in the print version of this title: ISBN: 978-0-07-178150-3,
MHID: 0-07-178150-1.

All trademarks are trademarks of their respective owners. Rather than put a trademark symbol after every occurrence of a trademarked name, we use names in an editorial fashion only, and to the benefit of the trademark owner, with no intention of infringement of the trademark. Where such designations appear in this book, they have been printed with initial caps.

McGraw-Hill eBooks are available at special quantity discounts to use as premiums and sales promotions, or for use in corporate training programs. To contact a representative please e-mail us at bulksales@mcgraw-hill.com.

Previous editions copyright © 2005, 2002, 1997, 1992, 1987, 1981 by The McGraw-Hill Companies, Inc.

TERMS OF USE

This is a copyrighted work and The McGraw-Hill Companies, Inc. (“McGraw-Hill”) and its licensors reserve all rights in and to the work. Use of this work is subject to these terms. Except as permitted under the Copyright Act of 1976 and the right to store and retrieve one copy of the work, you may not decompile, disassemble, reverse engineer, reproduce, modify, create derivative works based upon, transmit, distribute, disseminate, sell, publish or sublicense the work or any part of it without McGraw-Hill’s prior consent. You may use the work for your own noncommercial and personal use; any other use of the work is strictly prohibited. Your right to use the work may be terminated if you fail to comply with these terms.

THE WORK IS PROVIDED “AS IS.” MCGRAW-HILL AND ITS LICENSORS MAKE NO GUARANTEES OR WARRANTIES AS TO THE ACCURACY, ADEQUACY OR COMPLETENESS OF OR RESULTS TO BE OBTAINED FROM USING THE WORK, INCLUDING ANY INFORMATION THAT CAN BE ACCESSED THROUGH THE WORK VIA HYPERLINK OR OTHERWISE, AND EXPRESSLY DISCLAIM ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. McGraw-Hill and its licensors do not warrant or guarantee that the functions contained in the work will meet your requirements or that its operation will be uninterrupted or error free. Neither McGraw-Hill nor its licensors shall be liable to you or anyone else for any inaccuracy, error or omission, regardless of cause, in the work or for any damages resulting therefrom. McGraw-Hill has no responsibility for the content of any information accessed through the work. Under no circumstances shall McGraw-Hill and/or its licensors be liable for any indirect, incidental, special, punitive, consequential or similar damages that result from the use of or inability to use the work, even if any of them has been advised of the possibility of such damages. This limitation of liability shall apply to any claim or cause whatsoever whether such claim or cause arises in contract, tort or otherwise.

To Marsha Kramar Glantz

This page intentionally left blank

What I've proposed is that we have a panel of medical experts that are making determinations about what protocols are appropriate for what diseases. There's going to be some disagreement, but if there's broad agreement that, in this situation the blue pill works better than the red pill, and it turns out the blue pills are half as expensive as the red pill, then we want to make sure that doctors and patients have that information available to them.

President Barack Obama, 2009*

*Interview with ABC News' Dr. Timothy Johnson, July 15, 2009.

This page intentionally left blank

Contents

Preface.....	xiii	4. The Special Case of Two Groups: The <i>t</i> test.....	49
1. Biostatistics and Clinical Practice	1	The General Approach	49
What Do Statistical Procedures Tell You?	3	The Standard Deviation of a Difference or a Sum	51
Why Not Just Depend on the Journals?	4	Use of <i>t</i> to test Hypotheses About Two Groups.....	52
Why Has the Problem Persisted?.....	5	What if the Two Samples Are Not the Same Size?	56
2. How to Summarize Data	7	Cell Phones Revisited.....	56
Three Kinds of Data	7	The <i>t</i> test is an Analysis of Variance.....	59
The Mean	9	Common Errors in the Use of the <i>t</i> test and How to Compensate for Them.....	60
Measures of Variability	9	How to Use <i>t</i> tests to Isolate Differences Between Groups in Analysis of Variance.....	62
The Normal Distribution	10	The Bonferroni <i>t</i> test	62
Getting the Data	10	More on Cell Phones and Rabbit Sperm.....	63
Random Sampling.....	10	A Better Approach to Multiple Comparisons: The Holm <i>t</i> test.....	64
Bias	12	The Holm-Sidak <i>t</i> test.....	65
Experiments and Observational Studies.....	13	Multiple Comparisons Against a Single Control	67
Randomized Clinical Trials.....	14	The Meaning of <i>P</i>	67
How to Estimate the Mean and Standard Deviation from a Sample.....	15	Statistical Versus Real (Clinical) Thinking.....	69
How Good Are These Estimates?.....	15	Why $P < .05$?	70
Percentiles	19	Problems.....	71
Pain Associated with Diabetic Neuropathy.....	23	5. How to Analyze Rates and Proportions	73
Summary.....	25	Back to Mars	73
Problems.....	25	Estimating Proportions from Samples	75
3. How to Test for Differences between Groups	27	Hypothesis Tests for Proportions.....	78
The General Approach	27	The Yates Correction for Continuity.....	79
Two Different Estimates of the Population Variance.....	31	Effect of Counseling on End-of-Life Planning in Homeless People	79
What is a “Big” <i>F</i> ?.....	34	Another Approach to Testing Nominal Data: Analysis of Contingency Tables.....	80
Cell Phones and Sperm	38	The Chi-Square Test Statistic	81
An Early Study.....	38	The Yates Correction for Continuity.....	84
A Better Control Group.....	40		
An Experimental Study.....	41		
Unequal Sample Size	41		
Two Way Analysis of Variance	44		
Problems.....	44		

Chi-Square Applications to Experiments with More than Two Treatments or Outcomes	86
Multiple Comparisons.....	87
The Fisher Exact Test.....	87
Measures of Association Between	
Two Nominal Variables.....	91
Prospective Studies and Relative Risk	91
Absolute Risk Increase (or Reduction) and Number Needed to Treat	92
Case-Control Studies and the Odds Ratio.....	93
Passive Smoking and Breast Cancer.....	94
Problems.....	95
6. What Does “Not Significant” Really Mean?	101
An Effective Diuretic	101
Two Types of Errors	104
What Determines a Test’s Power?.....	105
The Size of the Type I Error, α	105
The Size of the Treatment Effect.....	109
The Population Variability.....	110
Bigger Samples Mean More Powerful Tests	112
What Determines Power? A Summary	112
Muscle Strength in People with Chronic Obstructive Pulmonary Disease.....	113
Power and Sample Size for	
Analysis of Variance	116
Power and Sperm Motility	116
Power and Sample Size for Comparing Two	
Proportions.....	118
Power and Polyethylene Bags	119
Sample Size for Comparing Two Proportions	120
Power and Sample Size for	
Relative Risk and Odds Ratio	121
Power and Sample Size for	
Contingency Tables	121
Power and Polyethylene Bags (Again)	121
Practical Problems in Using Power	122
What Difference Does it Make?.....	122
Problems.....	124
7. Confidence Intervals.....	125
The Size of the Treatment Effect Measured as the Difference of Two Means	125
The Effective Diuretic.....	127
More Experiments	127
What Does “Confidence” Mean?	129
Confidence Intervals Can Be Used to Test Hypotheses	130
Confidence Interval for the Population Mean....	132
The Size of the Treatment Effect Measured as the Difference of Two Rates or Proportions	132
Difference in Survival for Two Methods for Keeping Extremely Low Birth Weight Infants Warm	133
How Negative Is a “Negative” Clinical Trial?	133
Meta-Analysis	133
Confidence Interval for Rates and Proportions	136
Quality of Evidence Used as a Basis for Interventions to Improve Hospital Antibiotic Prescribing.....	136
Exact Confidence Intervals for Rates and Proportions	137
Confidence Intervals for Relative Risk and Odds Ratio.....	138
Effect of Counseling on Filing Advance Directives for End of Life Care Among Homeless People	139
Passive Smoking and Breast Cancer.....	140
Confidence Interval for the Entire Population	140
Problems.....	142
8. How to Test for Trends	143
More About the Martians	143
The Population Parameters	144
How to Estimate the Trend from a Sample.....	147
The Best Straight Line through the Data.....	148
Variability about the Regression Line	150
Standard Errors of the Regression Coefficients	151
How Convincing Is the Trend?.....	154
Confidence Interval for the Line of Means.....	155
Confidence Interval for an Observation	156
Cell Phone Radiation, Reactive Oxygen Species, and DNA Damage in Human Sperm.....	156
How to Compare Two Regression Lines.....	160
Overall Test for Coincidence of Two Regression Lines.....	161
Relationship between Weakness and Muscle Wasting in Rheumatoid Arthritis.....	161
Correlation and Correlation Coefficients	164
The Pearson Product-Moment Correlation Coefficient.....	164
The Relationship Between Regression and Correlation.....	165

How to Test Hypotheses about Correlation Coefficients	168
Journal Size and Selectivity.....	168
The Spearman Rank Correlation Coefficient	169
Cell Phone Radiation and Mitochondrial Reactive Oxygen Species in Sperm.....	172
Power and Sample Size in Regression and Correlation	173
Comparing Two Different Measurements of the Same Thing: The Bland- Altman Method	174
Assessing Mitral Regurgitation with Echocardiography	175
Multiple Regression.....	177
Summary.....	178
Problems.....	178
9. Experiments When Each Subject Receives More Than One Treatment.....	185
Experiments When Subjects Are Observed Before and After a Single Treatment: the Paired t test	185
Cigarette Smoking and Platelet Function	187
Another Approach to Analysis of Variance.....	189
Some New Notation	190
Accounting for All the Variability in the Observations.....	193
Experiments When Subjects Are Observed After Many Treatments: Repeated Measures Analysis of Variance.....	194
Anti-asthmatic Drugs and Endotoxins.....	197
How to Isolate Differences in Repeated Measures Analysis of Variance.....	200
Power in Repeated Measures Analysis of Variance	200
Experiments When Outcomes Are Measured on a Nominal Scale: McNemar's Test.....	200
p7 Antigen Expression in Human Breast Cancer.....	200
Problems.....	202
10. Alternatives to Analysis of Variance and the t test Based on Ranks	205
How to Choose Between Parametric and Nonparametric Methods	205
Two Different Samples: The Mann-Whitney Rank-Sum test.....	207
Use of a Cannabis-Based Medicine in Painful Diabetic Neuropathy.....	211
Each Subject Observed Before and After One Treatment: The Wilcoxon Signed-Rank Test	213
Cigarette Smoking and Platelet Function	217
Experiments with Three or More Groups When Each Group Contains Different Individuals: The Kruskal-Wallis test	218
Prenatal Marijuana Exposure and Child Behavior.....	219
Nonparametric Multiple Comparisons	219
Experiments in Which Each Subject Receives More than One Treatment: The Friedman Test	222
Anti-asthmatic Drugs and Endotoxin.....	224
Multiple Comparisons After the Friedman Test	225
Summary.....	225
Problems.....	227
11. How to Analyze Survival Data	229
Censoring on Pluto.....	230
Estimating the Survival Curve.....	230
Median Survival Time	234
Standard Errors and Confidence Limits for the Survival Curve	234
Comparing Two Survival Curves	235
Bone Marrow Transplantation to Treat Adult Leukemia.....	237
The Yates Correction for the Log Rank Test ...	242
Gehan's Test.....	242
Power and Sample Size	242
Power.....	243
Sample Size	243
Summary.....	243
Problems.....	244
12. What Do the Data Really Show?	247
Cell Phones: Putting All the Pieces Together	247
When to Use Which Test	248
Issues in Study Design	250
Randomize and Control.....	250
Internal Mammary Artery Ligation to Treat Angina Pectoris	251
The Portacaval Shunt to Treat Cirrhosis of the Liver	251
Is Randomization of People Ethical?	253
Is a Randomized Controlled Trial Always Necessary?	253
Does Randomization Ensure Correct Conclusions?	254

Problems with the Population	256	Appendix B. Statistical Tables and	
How You Can Improve Things	257	Power Charts	263
Appendix A. Computational Forms	259	Statistical Tables	263
To Interpolate Between Two Values in		Critical Values of F Corresponding	
a Statistical Table	259	to $P < .05$ and $P < .01$	265
Variance	259	Critical Values of t (Two-Tailed)	268
One-Way Analysis of Variance	259	Holm-Sidak Critical P Values for	
Given Sample Means and		Individual Comparisons to Maintain	
Standard Deviations.....	259	a 5% Family Error Rate ($\alpha_T = .05$)	270
Given Raw Data	259	Critical Values for the χ^2 Distribution	271
Unpaired t Test	260	Critical Values of t (One-Tailed).....	272
Given Sample Means and		Critical Values for Spearman Rank	
Standard Deviations.....	260	Correlation Coefficient.....	274
Given Raw Data	260	Critical Values (Two-Tailed) of the Mann-	
2×2 Contingency Tables (Including Yates		Whitney Rank-Sum T	275
Correction for Continuity)	260	Critical Values (Two-Tailed) of	
Chi Square	260	Wilcoxon W	276
McNemar's Test	260	Critical Values for Friedman χ_r^2	276
Fisher Exact Test.....	260	Power Charts for Analysis of Variance	277
Linear Regression and Correlation.....	260	Appendix C. Answers to Exercises	287
Repeated Measures Analysis of Variance	261	Index	297
Kruskal-Wallis Test	261		
Friedman Test	261		

Preface

I have always thought of myself as something of an outsider and troublemaker, so it is with some humility that I prepare the seventh edition of this book, 30 years after the first edition appeared. Then, as now, the book had an unusual perspective: that many papers in the medical literature contained avoidable errors. At the time, the publisher, McGraw-Hill, expressed concern that this “confrontational approach” would put off readers and hurt sales. They also worried that the book was not organized like a traditional statistics text.

Time has shown that the biomedical community was ready for such an approach and the book has achieved remarkable success.

The nature of the problems with the medical literature, however, has evolved over time and this new edition reflects that evolution. Many journals now have formal statistical reviewers so the kinds of simple errors that used to dominate have been replaced with more subtle problems of biased samples and underpowered studies (although there are still more than enough inappropriate t tests to go around). Over time, this book has evolved to include more topics, such as power and sample size, more on multiple comparison procedures, relative risks and odds ratios, and survival analysis.

In this edition I actually pruned back the discussion of multiple comparison testing to focus on Bonferonni, Holm, and Holm-Sidak corrected tests for both parametric and nonparametric methods.

At the same time, this is the most extensive revision done for a new edition since the book was first published. The book is now published in a larger, more open text format with more worked out examples. There are new brief introductions to higher order analysis of variance, multiple regression and logistic regression,* as well as expanded discussions of problems with study designs and more information on how to combine information from many different studies. The examples and problems have

been extensively reworked, with almost all coming from studies published in the twenty-first century.

This book has its origins in 1973, when I was a post-doctoral fellow. Many friends and colleagues came to me for advice and explanations about biostatistics. Since most of them had even less knowledge of statistics than I did, I tried to learn what I needed to help them. The need to develop quick and intuitive, yet correct, explanations of the various tests and procedures slowly evolved into a set of stock explanations and a two-hour slide show on common statistical errors in the biomedical literature and how to cope with them. The success of this slide show led many people to suggest that I expand it into an introductory book on biostatistics, which led to the first edition of *Primer of Biostatistics* in 1981.

As a result, this book is oriented as much to the individual reader—whether he or she is a student, postdoctoral research fellow, professor, or practitioner—as to the student attending formal lectures.

This book can be used as a text at many levels. It has been the required text for the biostatistics portion of the epidemiology and biostatistics course required of medical students, covering the material in the first eight chapters in eight one-hour lectures. The book has also been used for a more abbreviated set of lectures on biostatistics (covering the first three chapters) given to our dental students. In addition, it has served me (and others) well in a one-quarter four-unit course in which we cover the entire book in depth. This course meets for four lecture hours and has a one-hour problem session. It is attended by a wide variety of students, from undergraduates through

*These issues are treated in detail in a second book on the subject of multiple regression and analysis of variance, written with the same approach in *Primer of Biostatistics*. It is Glantz SA, Slinker BK. *Primer of Applied Regression and Analysis of Variance*, 2nd ed. New York: McGraw-Hill; 2001.

graduate students and postdoctoral fellows, as well as faculty members.

Because this book includes the technical material covered in any introductory statistics course, it is suitable as either the primary or the supplementary text for a general undergraduate introductory statistics course (which is essentially the level at which this material is taught in medical schools), especially for a teacher seeking a way to make statistics relevant to students majoring in the life sciences.

This book differs from other introductory texts on biostatistics in several ways, and it is these differences which seem to account for the book's enduring popularity.

First, because inappropriate use of the t test to analyze multigroup studies continues to be a common error, probably because the t test is usually the first procedure presented in a statistics book that will yield the highly prized P value. Analysis of variance, if presented at all, is deferred to the end of the book to be ignored or rushed through at the end of the term. Since so much is published that probably should be analyzed with analysis of variance, and since analysis of variance is really the paradigm of all parametric statistical tests, I present it first, then discuss the t test as a special case.

Second, in keeping with the problems that I see in the literature, there is a discussion of multiple comparison testing.

Third, the book is organized around hypothesis testing and estimation of the size of treatment effects, as opposed to the more traditional (and logical from a theory of statistics perspective) organization that goes from one-sample to two-sample to general k -sample estimation and hypotheses testing procedures. This approach goes directly to the kinds of problems one most commonly encounters when reading about or doing biomedical research.

The examples are based mostly on interesting studies from the literature and are reasonably true to the original data. I have, however, taken some liberty in recreating the

raw data to simplify the statistical problems (for example, making the sample sizes equal) so that I could focus on the important intuitive ideas behind the statistical procedures rather than getting involved in the algebra and arithmetic. There are still some topics common in introductory texts that I leave out or treat implicitly. There is not an explicit discussion of probability calculus and expected values and I still blur the distinction between P and α .

As with any book, there are many people who deserve thanks. Julien Hoffman gave me the first really clear and practically oriented course in biostatistics, which allowed me to stay one step ahead of the people who came to me for expert help. Over the years, Virginia Ernster, Susan Sacks, Philip Wilkinson, Marion Nestle, Mary Giammona, Bryan Slinker, Jim Lightwood, Kristina Thayer, Joaquin Barnoya, Jennifer Ibrahim, and Sara Shain helped me find good examples to use in the text and as problems. Bart Harvey and Evelyn Schlenker were particularly gracious in offering suggestions and detailed feedback on the new material in this edition. I thank them all. Finally, I thank the many others who have used the book, both as students and as teachers of biostatistics, who took the time to write me questions, comments, and suggestions on how to improve it. I have done my best to heed their advice in preparing this seventh edition.

Many of the pictures in this book are direct descendants of my original slides. In fact, as you read this book, you would do best to think of it as a slide show that has been set to print. Most people who attend my slide show leave more critical of what they read in the biomedical literature and people who have read earlier editions said that the book had a similar effect on them. Nothing could be more flattering or satisfying to me. I hope that this book will continue to make more people more critical and help improve the quality of the biomedical literature and, ultimately, the care of people.

Stanton A. Glantz

Biostatistics and Clinical Practice

Until the second quarter of the 20th century, medical treatment had little positive effect on when, or even whether, sick people recovered. With the discovery of ways to reverse the biochemical deficiencies that caused some diseases and the development of antibacterial drugs, it became possible to cure sick people. These early successes and the therapeutic optimism they engendered stimulated the biomedical research community to develop a host of more powerful agents to treat heart disease, cancer, neurological disorders, and other ailments. These increasing opportunities for productive intervention as well as a fundamental restructuring of the market away from non-profit health care providers to for-profit entities and the expansion of the pharmaceutical, medical device, and insurance industries that saw opportunities to make money providing medical services, together with increasing expectations by the public, have led to spending an accelerating amount of money on medical services, reaching \$2.6 trillion and nearly one-fifth of the United States' entire gross domestic product in 2011 (Fig. 1-1).

This situation has led to continuous calls for reform from a wide spectrum of stakeholders, from business leaders who saw their costs skyrocketing, to labor leaders who saw health insurance costs putting downward pressure on wages, to advocates for the growing number of uninsured people who were simply priced out of the system, to political decision makers who saw out-of-control costs of providing medical care through government programs such as Medicare and Medicaid, jeopardizing other important government services.

Because of the fact that medical care touches everyone's life in one way or another and because of the high stakes—financial and otherwise—for the individuals and organizations that provide these services, reforming the health care system has been a controversial and politically charged issue.

After over a year of increasingly partisan debate, in March 2010 the Democrats in Congress passed the Patient Protection and Affordable Care Act without a single Republican vote. On March 23, 2010, President Barack Obama signed the bill into law.

While this law has many provisions, including requiring people to have or purchase health insurance and imposing many regulations on the health insurance industry, it also recognizes that the current medical system is unsustainable financially and includes several provisions designed to get the costs of the medical system under control. (Indeed, one of the main facts driving the debate was the observation, from an ongoing research project at Dartmouth University, the Dartmouth Atlas of Health Care,* that 30% of the nation's medical spending would be unnecessary if all regions of the United States the provided services at the level observed in low-spending regions that achieved that same equal quality.) The law

*The research behind this statement, together with many other findings about geographical variations in medical services and health outcomes is available at www.dartmouthatlas.org.

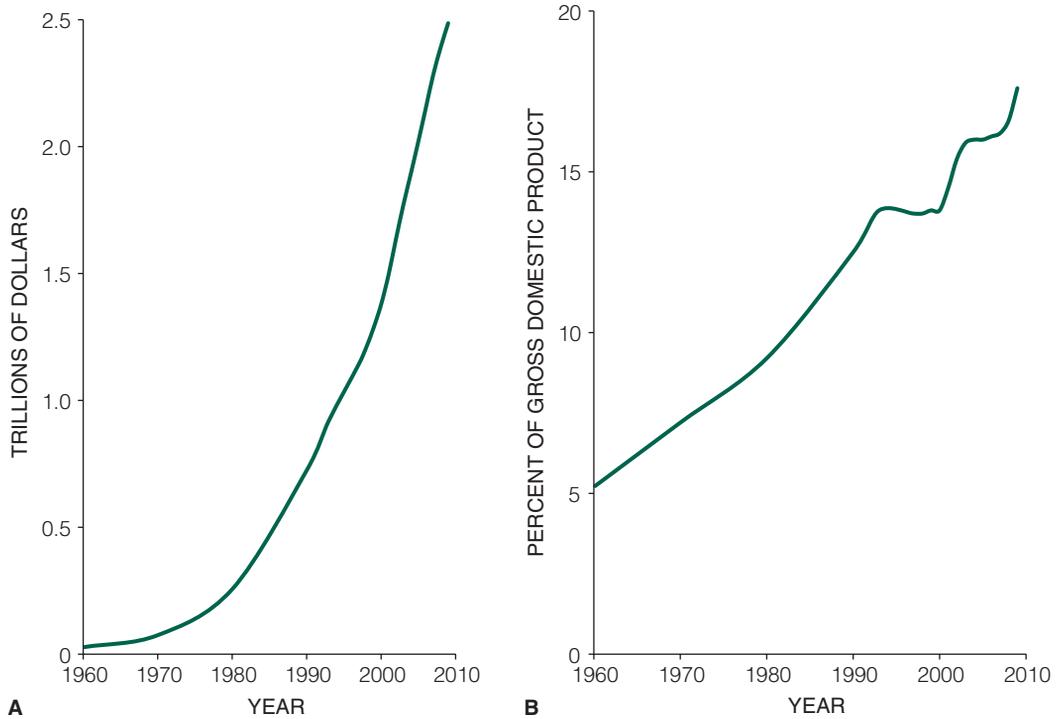


FIGURE 1-1. (A) Total annual expenditures for medical services in the United States between 1960 and 2010. **(B)** Expenditures for medical services as a percentage of the gross domestic product. (Source: Statistical Abstract of the United States, 2011. Washington, DC: US Department of Commerce, pp. 99.)

established a Patient-Centered Outcomes Research Institute to conduct *comparative effectiveness research* on the “relative health outcomes, clinical effectiveness, and appropriateness” of different medical treatments. The law also created task forces on Preventive Services and Community Preventive Services to develop, update, and disseminate evidenced-based recommendations on the use of clinical and community prevention services.

These issues are, at their heart, statistical issues. Because of factors such as natural biological variability between individual patients and the placebo effect,* one usually cannot conclude that some therapy was beneficial on the basis of simple experience. Biostatistics provides the tools for turning clinical and laboratory experience

into quantitative statements about whether and by how much a treatment or procedure affects a group of patients.

Hence, evidence collected and analyzed using biostatistical methods can potentially affect not only how clinicians choose to practice their profession but what choices are open to them. Intelligent participation in these decisions requires an understanding of biostatistical methods and models that will permit one to assess the quality of the evidence and the analysis of that evidence used to support one position or another.

Clinicians have not, by and large, participated in debates on these quantitative questions, probably because the issues appear too technical and seem to have little impact on their day-to-day activities. Clinicians need to be able to make more informed judgments about claims of medical efficacy so that they can participate more intelligently in the debate on how to allocate health care resources. These judgments will be based, in large part, on statistical reasoning.

*The placebo effect is a positive response to therapy *per se* as opposed to the therapy’s specific effects. For example, about one-third of people given placebos in place of painkillers report experiencing relief. We will discuss the placebo effect in detail later in this book.

■ WHAT DO STATISTICAL PROCEDURES TELL YOU?

Suppose researchers believe that administering some drug increases urine production in proportion to the dose and to study it they give different doses of the drug to five different people, plotting their urine production against the dose of drug. The resulting data, shown in Figure 1-2A, reveal a strong relationship between the drug dose and daily urine production in the five people who were studied. This result would probably lead the investigators to publish a paper stating that the drug was an effective diuretic.

The only statement that can be made with absolute certainty is that as the drug dose increased, so did urine production *in the five people in the study*. The real question of interest, however, is: How is the drug likely to affect *all people who receive it*? The assertion that the drug is effective requires a leap of faith from the limited experience, shown in Figure 1-2A, to all people.

Now, pretend that we knew how every person who would ever receive the drug would respond. Figure 1-2B shows this information. There is no systematic relationship between the drug dose and urine production! The drug is not an effective diuretic.

How could we have been led so far astray? The dark points in Figure 1-2B represent the specific individuals who happened to be studied to obtain the results shown in Figure 1-2A. While they are all members of the population of people we are interested in studying, the five specific individuals we happened to study, taken as a group, were not really representative of how the entire population of people responds to the drug.

Looking at Figure 1-2B should convince you that obtaining such an unrepresentative sample of people, though possible, is not very probable. One set of statistical procedures, called *tests of hypotheses*, permit you to estimate the likelihood of concluding that two things are related as Figure 1-2A suggests when the relationship is really due to bad luck in selecting people for study, and not a true effect

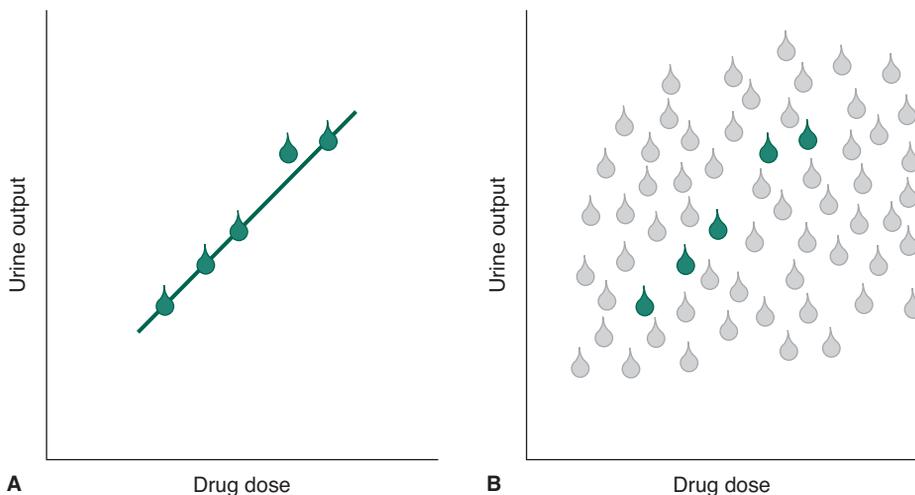


FIGURE 1-2. (A) Results of an experiment in which researchers administered five different doses of a drug to five different people and measured their daily urine production. Output increased as the dose of drug increased in these five people, suggesting that the drug is an effective diuretic in all people similar to those tested. **(B)** If the researchers had been able to administer the drug to all people and measure their daily urine output, it would have been clear that there is no relationship between the dose of drug and urine output. The five specific individuals who happened to be selected for the study in panel A are shown as shaded points. It is possible, but not likely, to obtain such an unrepresentative sample that leads one to believe that there is a relationship between the two variables when there is none. A set of statistical procedures called tests of hypotheses permits one to estimate the chance of getting such an unrepresentative sample.

of the drug. In this example, one can estimate that such a sample of people will turn up in a study of the drug only about 5 times in 1000 when the drug actually has no effect.

Of course it is important to realize that although statistics is a branch of mathematics, there can be honest differences of opinion about the best way to analyze a problem. This fact arises because all statistical methods are based on relatively simple mathematical models of reality, so the results of the statistical tests are accurate only to the extent that the reality and the mathematical model underlying the statistical test are in reasonable agreement.

■ WHY NOT JUST DEPEND ON THE JOURNALS?

Aside from direct personal experience, most health care professionals rely on medical journals to keep them informed about the current concepts on how to diagnose and treat their patients. Since few members of the clinical or biomedical research community are conversant in the use and interpretation of biostatistics, most readers assume that when an article appears in a journal, the reviewers and editors have scrutinized every aspect of the manuscript, including the use of statistics. Unfortunately, this is often not so.

Beginning in the 1950s, several critical reviews* of the use of statistics in the general medical literature consistently found that about half the articles used incorrect statistical methods. This situation led many of the larger journals to incorporate formal statistical reviews (by a statistician) into the peer review process. Reviews of the efficacy of providing secondary statistical reviews of tentatively accepted papers have revealed that about half (or more) of the papers, tentatively accepted for publication, have statistical problems.† For the most part, these errors are resolved before publication, together with substantive issues raised by the other (content) reviewers, and the rate

of statistical problems in the final published papers is much lower.

By 1995, most (82%) of the large-circulation general medical journals had incorporated a formal statistical review into the peer review process. There was a 52% chance that a paper published in one of these journals would receive a statistical review before it was published.‡ This situation was not nearly as common among the smaller specialty and subspecialty journals. Only 31% of these journals had a statistical reviewer available and only 27% of published papers had been reviewed by a statistician.

As the demands for evidence of efficacy have increased, so has the appreciation of the problem of biased studies in which the outcome is influenced by the selection of people included in the study or the precise therapies that are being compared. Sponsorship of the research by companies with a financial interest in the outcome of the study can influence the conclusions of the resulting papers. These problems are more subtle than just applying the wrong statistical test. Indeed, reviews of specialty journals continue to show a high frequency of statistical problems in published papers.§

*Ross OB Jr. Use of controls in medical research. *JAMA*. 1951;145:72–75; Badgley RF. An assessment of research methods reported in 103 scientific articles from two Canadian medical journals. *Can MAJ*. 1961;85:256–260; Schor S, Karten I. Statistical evaluation of medical journal manuscripts. *JAMA*. 1966;195:1123–1128; Gore S, Jones IG, Rytter EC. Misuses of statistical methods: critical assessment of articles in *B.M.J.* from January to March, 1976. *Br Med J*. 1977;1(6053):85–87.

†For a discussion of the experiences of two journals, see Gardner MJ, Bond J. An exploratory study of statistical assessment of papers published in the *British Medical Journal*. *JAMA*. 1990;263:1355–1357; Glantz SA. It is all in the numbers. *J Am Coll Cardiol*. 1993;21:835–837.

*Goodman SN, Altman DG, George SL. Statistical reviewing policies of medical journals: caveat lector? *J Gen Intern Med*. 1998;13:753–756.

‡More recent reviews, while dealing with a more limited selection of journals, have shown that this problem still persists. See Rushton L. Reporting of occupational and environmental research: use and misuse of statistical and epidemiological methods. *Occup Environ Med*. 2000;57:1–9; Dimick JB, Diener-West M, Lipsett PA. Negative results of randomized clinical trials published in the surgical literature. *Arch Surg*. 2001;136:796–800; Dijkers M, Kropp GC, Esper RM, Yavuzer G, Cullen N, Bakdaliieh Y. Quality of intervention research reporting in medical rehabilitation journals. *Am J Phys Med Rehab*. 2002;81:21–33; Welch GE II, Gabbe SG. Statistics usage in the *American Journal of Obstetrics and Gynecology*: has anything changed? *Am J Obstet Gynecol*. 2002;186:584–586; Maggard MA, O’Connell JB, Liu JH, Etzioni DA, Ko CY. Sample size calculations in surgery: are they done correctly. *Surgery*. 2003;134:275–279; Bedard PL, Kryznanowska MK, Pintille M, Tannock IF. Statistical power of negative randomized controlled trials presented at American Society for Clinical Oncology annual meetings. *J Clin Oncol*. 2007;25:3482–3487; Tsang R, Colley L, Lynd LD. Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials. *J Clin Epidemiol*. 2009;62:609–616; Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*. 2010;303:2058–2064.

When confronted with this observation—or the confusion that arises when two seemingly comparable articles arrive at different conclusions—people often conclude that statistical analyses are maneuverable to one's needs, or are meaningless, or are too difficult to understand.

Unfortunately, except when a statistical procedure merely confirms an obvious effect (or the paper includes the raw data), a reader cannot tell whether the data, in fact, support the author's conclusions or not. Ironically, the errors rarely involve sophisticated issues that provoke debate among professional statisticians but are simple mistakes, such as neglecting to include a control group, not allocating treatments to subjects at random, or misusing elementary tests of hypotheses. These errors generally bias the study on behalf of the treatments.

The existence of errors in experimental design or biased samples in observational studies and misuse of elementary statistical techniques in a substantial fraction of published papers is especially important in clinical studies. These errors may lead investigators to report a treatment or diagnostic test to be of statistically demonstrated value when, in fact, the available data fail to support this conclusion. Health care professionals who believe that a treatment has been proved effective on the basis of publication in a reputable journal may use it for their patients. Because all medical procedures involve some risk, discomfort, or cost, people treated on the basis of erroneous research reports gain no benefit and may be harmed. On the other hand, errors could produce unnecessary delay in the use of helpful treatments. Scientific studies which document the effectiveness of medical procedures will become even more important as efforts grow to control medical costs without sacrificing quality. Such studies must be designed and interpreted correctly.

In addition to indirect costs, there are significant direct costs associated with these errors: money is spent, animals may be sacrificed, and human study participants are inconvenienced and may even be put at risk to collect data that are not interpreted correctly.

■ WHY HAS THE PROBLEM PERSISTED?

Because so many people are making these errors, there is little peer pressure on academic investigators to use statistical techniques carefully. In fact, one rarely hears a word of criticism. Quite the contrary, some investigators fear that their colleagues—and, especially, reviewers—will view a correct analysis as unnecessarily theoretical and complicated.

Most editors still assume that the reviewers will examine the statistical methodology in a paper with the same level of care that they examine the clinical protocol or experimental preparation. If this assumption were correct, one would expect all papers to describe, in detail as explicit as the description of the protocol or preparation, how the authors have analyzed their data. Yet, often the statistical procedures used to test hypotheses in medical journals are not even identified. It is hard to believe that the reviewers examined the methods of data analysis with the same diligence with which they evaluated the experiment used to collect the data.

To read the medical literature intelligently, you will have to be able to understand and evaluate the use of the statistical methods used to analyze the experimental results as well as the laboratory methods used to collect the data. Fortunately, the basic ideas needed to be an intelligent reader—and, indeed, to be an intelligent investigator—are quite simple. The next chapter begins our discussion of these ideas and methods.

This page intentionally left blank

How to Summarize Data

An investigator collecting data generally has two goals: to obtain descriptive information about the population from which the sample was drawn and to test hypotheses about that population. We focus here on the first goal: to summarize data collected on a single variable in a way that best describes the larger, unobserved population.

When the value of the variable associated with any given individual is more likely to fall near the mean (average) value for all individuals in the population under study than far from it and equally likely to be above the mean and below it, the *mean* and *standard deviation* for the sample observations describe the location and amount of variability among members of the population. When the value of the variable is more likely than not to fall below (or above) the mean, one should report the *median* and values of at least two other *percentiles*.

To understand these rules, assume that we observe *all* members of the population, not only a limited (ideally representative) sample as in an experiment.

For example, suppose we wish to study the height of Martians and, to avoid any guesswork, we visit Mars and measure the entire population—all 200 of them. Figure 2-1 shows the resulting data with each Martian's height rounded to the nearest centimeter and represented by a circle. There is a *distribution* of heights of the Martian population. Most Martians are between about 35 and 45 cm tall, and only a few (10 out of 200) are 30 cm or shorter, or 50 cm or taller.

Having successfully completed this project and demonstrated the methodology, we submit a proposal to measure the height of Venusians. Our record of good work assures funding, and we proceed to make the measurements. Following the same conservative approach, we measure the heights of *all* 150 Venusians. Figure 2-2 shows the measured heights for the entire population of Venus, using the same presentation as Figure 2-1. As on Mars, there is a distribution of heights among members of the population, and all Venusians are around 15 cm tall, almost all of them being taller than 10 cm and shorter than 20 cm.

Comparing Figures 2-1 and 2-2 demonstrates that Venusians are shorter than Martians and that the variability of heights within the Venusian population is smaller. Whereas almost all (194 of 200) the Martians' heights fall in a range 20 cm wide (30 to 50 cm), the analogous range for Venusians (144 of 150) is only 10 cm (10 to 20 cm). Despite these differences, there are important similarities between these two populations. In both, any given member is more likely to be near the middle of the population than far from it and equally likely to be shorter or taller than average. In fact, despite the differences in population size, average height, and variability, the *shapes* of the distributions of heights of the inhabitants of both the planets are almost identical. A most striking result!

■ THREE KINDS OF DATA

The heights of Martians and Venusians are known as *interval* data because heights are measured on a scale with constant intervals, in this case, centimeters. For

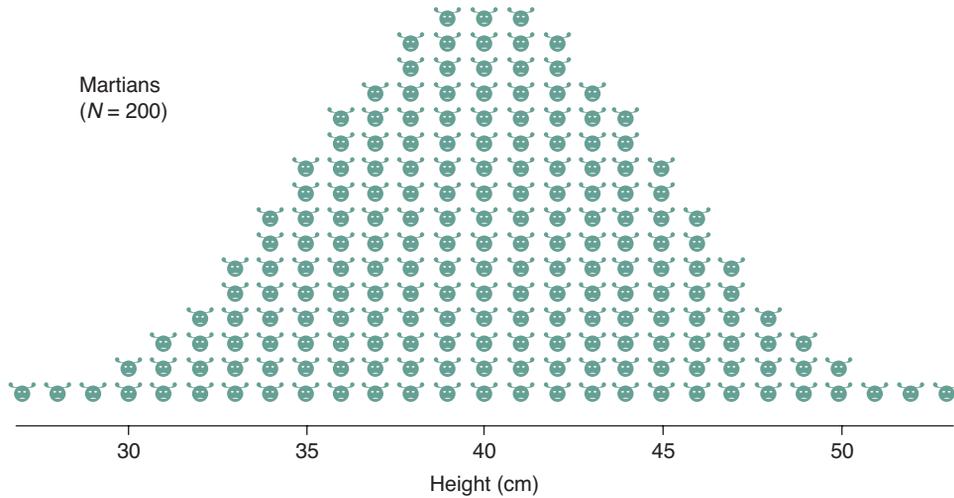


FIGURE 2-1. Distribution of heights of 200 Martians, with each Martian's height represented by a single point. Notice that any individual Martian is more likely to have a height near the mean height of the population (40 cm) than far from it and is equally likely to be shorter or taller than average.

interval data, the absolute difference between two values can always be determined by subtraction.* The difference in heights of Martians who are 35 and 36 cm tall is the same as the difference in height of Martians who are 48 and 49 cm tall. Other variables measured on interval scales include temperature (because a 1°C difference always means the same thing), blood pressure (because a 1 mmHg difference in pressure always means the same thing), height, or weight.

There are other data, such as gender, state of birth, or whether or not a person has a certain disease, that are not measured on an interval scale. These variables are examples of *nominal* or *categorical data*, in which individuals are classified into two or more *mutually exclusive* and *exhaustive categories*. For example, people could be categorized as male or female, dead or alive, or as being born in one of the 50 states, District of Columbia, or outside the United States. In every case, it is possible to categorize each individual into one and only one category. In

addition, there is no arithmetic relationship or even ordering between the categories.†

Ordinal data fall between interval and nominal data. Like nominal data, ordinal data fall into categories, but there is an inherent ordering (or ranking) of the categories. Level of health (excellent, very good, good, fair, or poor) is a common example of a variable measured on an ordinal scale. The different values have a natural order, but the differences or “distances” between adjoining values on an ordinal scale are not necessarily the same and may not even be comparable. For example, excellent health is better than very good health, but this difference is not necessarily the same as the difference between fair and poor health. Indeed, these differences may not even be strictly comparable.

For the remainder of this chapter, we will concentrate on how to describe interval data, particularly how to describe the location and shape of the distributions.‡ Because of the similar shapes of the distributions of heights of Martians and Venusians, we will reduce all the information in Figures 2-1 and 2-2 to a few numbers, called *parameters*, of

*Relative differences can only be computed when there is a *true zero point*. For example, height has a true zero point, so a Martian that is 45 cm tall is 1.5 times as tall as a Martian that is 30 cm tall. In contrast, temperature measured in degrees Celsius or Fahrenheit does not have a true zero point, so it would be inaccurate to say that 100°C is twice as hot as 50°C . However, the Kelvin temperature scale does have a true zero point. Interval data that has a true zero point is called *ratio data*. The methods we will be developing only require interval data.

†Variables measured on a nominal scale in which there are only two categories are also known as *dichotomous variables*.

‡We will present the corresponding approaches for nominal (in Chapters 5 and 11) and ordinal data (in Chapter 10). The basic principles are the same for all three kinds of data.

Venusians
($N = 150$)

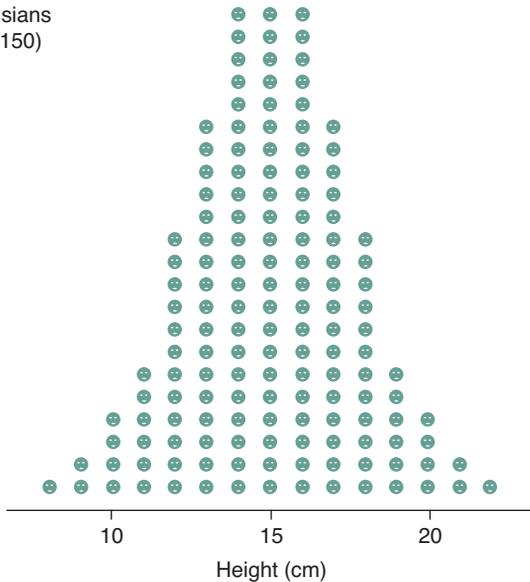


FIGURE 2-2. Distribution of heights of 150 Venusians. Notice that although the average height and dispersion of heights about the mean differ from those of Martians (Fig. 2-1), they both have a similar bell-shaped appearance.

the distributions. Indeed, since the shapes of the two distributions are so similar, we only need to describe how they differ; we do this by computing the mean height and the *variability* of heights about the mean.

■ THE MEAN

To indicate the location along the height scale, define the *population mean* to be the average height of all members of the population. Population means are often denoted by μ , the Greek letter mu. When the population is made up of discrete members,

$$\text{Population mean} = \frac{\text{Sum of values, e.g., heights, for each member of population}}{\text{Number of population members}}$$

The equivalent mathematical statement is

$$\mu = \frac{\sum X}{N}$$

in which Σ , Greek capital letter sigma, indicates the sum of the values of the variable X for all N members of the population. Applying this definition to the data in Figures 2-1 and 2-2 yields the result that the mean height of Martians

is 40 cm and the mean height of Venusians is 15 cm. These numbers summarize the qualitative conclusion that the distribution of heights of Martians is higher than the distribution of heights of Venusians.

■ MEASURES OF VARIABILITY

Next, we need a measure of dispersion about the mean. A value an equal distance above or below the mean should contribute the same amount to our index of variability, even though in one case the deviation from the mean is positive and in the other it is negative. Squaring a number makes it positive, so let us describe the variability of a population about the mean by computing the *average squared deviation from the mean*. The average squared deviation from the mean is larger when there is more variability among members of the population (compare the Martians and Venusians). It is called the *population variance* and is denoted by σ^2 , the square of the lower case Greek sigma. Its precise definition for populations made up of discrete individuals is

$$\text{Population variance} = \frac{\text{Sum of (value associated with member of population - population mean)}^2}{\text{Number of population members}}$$

The equivalent mathematical statement is

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Note that the units of variance are the square of the units of the variable of interest. In particular, the variance of Martian heights is 25 cm² and the variance of Venusian heights is 6.3 cm². These numbers summarize the qualitative conclusion that there is more variability in heights of Martians than in heights of Venusians.

Since variances are often hard to visualize, it is more common to present the square root of the variance, which we might call the *square root of the average squared deviation from the mean*. Since that is quite a mouthful, this quantity has been named the *standard deviation*, σ . Therefore, by definition,

$$\text{Population standard deviation} = \sqrt{\text{Population variance}}$$

$$= \sqrt{\frac{\text{Sum of (value associated with member of population - population mean)}^2}{\text{Number of population members}}}$$

or mathematically,

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(X - \mu)^2}{N}}$$

where the symbols are defined as before. Note that the standard deviation has the same units as the original observations. For example, the standard deviation of Martian heights is 5 cm, and the standard deviation of Venusian heights is 2.5 cm.

■ THE NORMAL DISTRIBUTION

Table 2-1 summarizes what we found out about Martians and Venusians. The three numbers in the table tell a great deal: the population size, the mean height, and how much the heights vary about the mean. The distributions of heights on both the planets have a similar shape, so that *roughly 68% of the heights fall within 1 standard deviation from the mean and roughly 95% within 2 standard deviations from the mean*. This pattern occurs so often that mathematicians have studied it and found that if the observed measurement is the sum of many independent small random factors, the resulting measurements will take on values that are distributed, like the heights we observed on both Mars and Venus. This distribution is called the *normal (or Gaussian) distribution*.

Its height at any given value of X is

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{X - \mu}{\sigma}\right)^2\right]$$

Note that the distribution is completely defined by the population mean μ and population standard deviation σ . Therefore, the information given in Table 2-1 is not just a good abstract of the data, it is *all* the information one needs to describe the population fully *if the distribution of values follows a normal distribution*.

■ GETTING THE DATA

So far, everything we have done has been exact because we followed the conservative course of examining every single member of the population. Usually it is physically or fiscally impossible to do this, and we are limited to examining a *sample* of n individuals drawn from the population in the hope that it is representative of the complete population. Without knowledge of the entire population, we can no longer know the population mean, μ , and population standard deviation, σ . Nevertheless, we can estimate them from the sample. To do so, however, the sample has to be “representative” of the population from which it is drawn.

Random Sampling

All statistical methods are built on the assumption that the individuals included in your sample represent a *random sample* from the underlying (and unobserved) population. In a random sample *every member of the population has an equal probability (chance) of being selected for the sample*. For the results of any of the methods developed in this book to be reliable, this assumption has to be met.

The most direct way to create a simple random sample would be to obtain a list of every member of the population of interest, number them from 1 to N (where N is the number of population members), then use a computerized *random number generator* to select the n individuals for the sample. Table 2-2 shows 100 random numbers between 1 and 150 created with a random number generator. Every number has the same chance of appearing and there is no relationship between adjacent numbers.

We could use this table to select a random sample of Venusians from the population shown in Figure 2-2. To do this, we number the Venusians from 1 to 150, beginning with number 1 for the far left individual in Figure 2-2, numbers 2 and 3 for the next two individuals in the second column in Figure 2-2, numbers 4, 5, 6, and 7 for the individuals in the next column, until we reach the individual

■ TABLE 2-1. Population Parameters for Heights of Martians and Venusians

	Size of Population	Population Mean (cm)	Population Standard Deviation (cm)
Martians	200	40	5.0
Venusians	150	15	2.5

■ **TABLE 2-2. One Hundred Random Numbers between 1 and 150**

2	135	4	138	57
101	26	116	131	77
49	99	146	137	129
54	83	4	121	129
30	102	7	128	15
137	85	71	114	7
40	67	109	34	123
6	23	120	6	72
112	7	131	58	38
74	30	126	47	79
108	82	96	57	123
55	32	16	114	41
7	81	81	37	21
4	52	131	62	7
7	38	55	102	5
37	61	142	42	8
116	5	41	111	109
76	83	51	37	40
100	82	49	11	93
83	146	42	50	35

at the far right of the distribution, who is assigned the number 150. To obtain a simple random sample of six Venusians from this population, we take the first six numbers in the table—2, 101, 49, 54, 30, and 137—and select the corresponding individuals. Figure 2-3 shows the result of this process. (When a number repeats, as with the two 7s in the first column of Table 2-2, simply skip the repeats because the corresponding individual has already been selected.)

We could create a second random sample by simply continuing in the table beginning with the seventh entry, 40, or starting in another column. The important point is not to reuse any sequence of random numbers already used to select a sample. (As a practical matter, one would probably use a computerized random number generator, which automatically makes each sequence of random numbers independent of the other sequences it generates.) In this way, we ensure that every member of the population is equally likely to be selected for observation in the sample.

The list of population members from which we drew the random sample is known as a *sampling frame*. Sometimes it is possible to obtain such a list (for example, a list of all people hospitalized in a given hospital on a given day), but

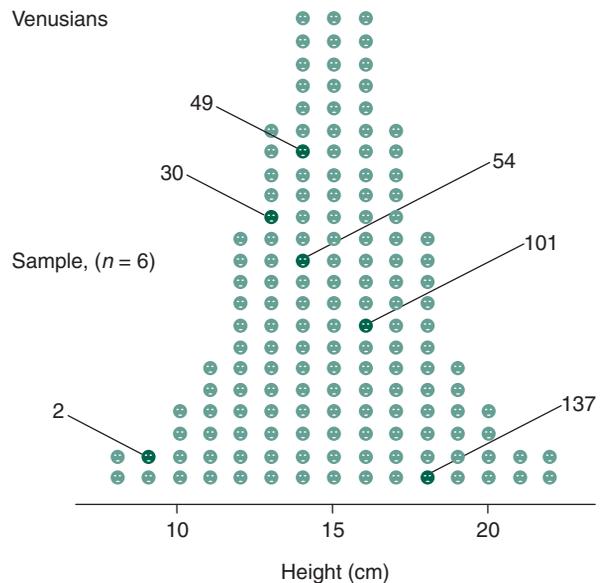


FIGURE 2-3. To select $n = 6$ Venusians at random, we number the entire population of $N = 150$ Venusians from 1 to 150, beginning with the first individual on the far left of the population as number 1. We then select six random numbers from Table 2-2 and select the corresponding individuals for the sample to be observed.

often no such list exists. When there is no list, investigators use other techniques for creating a random sample, such as dialing telephone numbers at random for public opinion polling or selecting geographic locations at random from maps. The issue of how the sampling frame is constructed can be very important in terms of how well and to whom the results of a given study generalize to individuals beyond the specific individuals in the sample.*

The procedure we just discussed is known as a *simple random sample*. In more complex designs, particularly in large surveys or clinical trials, investigators sometimes use *stratified random samples* in which they first divide the population into different subgroups (perhaps based on gender, race, or geographic location), then construct simple random samples within each subgroup (strata). This procedure is used when there are widely varying numbers of people in the different subpopulations so that obtaining adequate sample sizes in the smaller subgroups would require collecting more data than necessary in the larger subpopulations if the sampling was done with a simple random sample. Stratification reduces data collection costs by reducing the total sample size necessary to obtain the desired precision in the results, but makes the data analysis more complicated. The basic need to create a random sample in which each member of each subpopulation (strata) has the same chance of being selected is the same as in a simple random sample.

Bias

The primary reason for random sampling—whether a simple random sample or a more complex stratified sample—is to avoid *bias* in selecting the individuals to be included in the sample. A bias is a systematic difference between the characteristics of the members of the sample and the population from which it is drawn.

Biases can be introduced purposefully or by accident. For example, suppose you are interested in describing the age distribution of the population. The easiest way to obtain a sample would be to simply select the people whose age is to be measured from the people in your biostatistics class. The problem with this *convenience sample* is that you will be leaving out everyone not old enough to be learning biostatistics or those who have outgrown the desire to do so. The results obtained from this convenience sample

would probably underestimate both the mean age of people in the entire population as well as the amount of variation in the population. Biases can also be introduced by selectively placing people in one comparison group or another. For example, if one is conducting an experiment to compare a new drug with conventional therapy, it would be possible to bias the results by putting the sicker people in the conventional therapy group with the expectation that they would do worse than people who were not as sick and were receiving the new drug. Random sampling protects against both these kinds of biases.

Biases can also be introduced when there is a systematic error in the measuring device, such as when the zero on a bathroom scale is set too high or too low, so that all measurements are above or below the real weight.[†]

Another source of bias can come from the people making or reporting the measurements if they have hopes or beliefs that the treatment being tested is or is not superior to the control group or conventional therapy being studied. It is common, particularly in clinical research, for there to be some room for judgment in making and reporting measurements. If the investigator wants the study to come out one way or another, there is always the possibility for reading the measurements systematically low in one group and systematically high in the other.

The best way to avoid this measurement bias is to have the person making the measurements *blinded* to which treatment led to the data being measured. For example, suppose that one is doing a comparison of the efficacy of two different stents (small tubes inserted into arteries) to keep coronary arteries (arteries in the heart) open. To blind the measurements, the person reading the data on artery size would not know whether the data came from a person in the control group (who did not receive a stent), or which of the different stents was used in a given person.

Another kind of bias is due to the *placebo effect*, the tendency of people to report a change in condition simply because they received a treatment, even if the treatment had no biologic effect. For example, about one-third of people given an inert injection that they thought was an anesthetic reported a lessening of dental pain. To control for this effect in clinical experiments, it is common to

*We will return to this issue in Chapter 12, with specific emphasis on doing clinical research on people being served at academic medical centers.

[†]For purposes of this text, we assume that the measurements themselves are unbiased. Random errors associated with the measurement process are absorbed into the other random elements associated with the sampling process.

give one group a placebo so that they think that they are receiving a treatment. Examples of placebos include an injection of saline, a sugar pill, or surgically opening and closing without performing any specific procedure on the target organ. Leaving out a placebo control can seriously bias the results of an experiment in favor of the treatment. Ideally, the experimental subject would not know if they were receiving a placebo or an active treatment. When the subject does not know whether they received a placebo or not, the subject is *blinded*.

When neither the investigator nor the subject knows who received which treatment, the study is *double blinded*. For example, in double-blind drug studies, people are assigned treatments at random and neither the subject nor the person delivering the drug and measuring the outcome knows whether the subject received an active drug or a placebo. The drugs are delivered with only a number code identifying them. The code is broken only after all the data have been collected.

Experiments and Observational Studies

There are two ways to obtain data: *experiments* and *observational studies*. Experiments permit drawing stronger conclusions than observational studies, but often it is only possible to do observational studies.

In an *experiment*, the investigator selects individuals from the population of interest (using an appropriate sampling frame), then assigns the selected individuals to different *treatment groups*, applies the treatments, and measures the variables of interest. Drug trials where people are randomly assigned to receive conventional therapy or a drug that is thought to improve their condition are common biomedical experiments. Since the only systematic difference between the different treatment groups is the treatment itself, one can be reasonably confident that the treatment *caused* the observed differences.

Selecting people and randomly assigning them to different experimental conditions is not always possible or ethical. In an *observational study* the investigators obtain data by simply observing events without controlling them. Such studies are prone to two potentially serious problems. First, the groups may vary in ways the investigators do not notice or choose to ignore and these differences, rather than the treatment itself, may account for the differences the investigators find. Second, such studies can be subject to bias in patient recall, investigator assessment, or selection of the treatment group or the control group.

Observational studies do, however, have advantages. First, they are relatively inexpensive because they are often based on reviews of existing information or information that is already being collected for other purposes (like medical records) and because they generally do not require direct intervention by the investigator. Second, ethical considerations or prevailing medical practice can make it impossible to carry out active manipulation of the variable under study.

Because of the potential difficulties in all observational studies, it is critical that the investigators explicitly specify the criteria they used for classifying each subject in the control or case group. Such specifications help minimize biases when the study is done as well as help you, as the consumer of the resulting information, judge whether the classification rules made sense.

For example, epidemiologists have compared the rates of lung cancer and heart disease in nonsmokers whose spouses or coworkers smoke with the rates observed in nonsmokers living in smokefree environments. These studies have shown higher rates of lung cancer and heart disease in the people exposed to secondhand smoke, leading to the conclusion that secondhand smoke increases the risk of disease (Fig. 2-4A).

When doing an observational study, however, one always has to worry that the association observed in the data is not due to a cause-and-effect link between the two variables (in this case, secondhand smoke causing lung cancer), but rather the presence of some unobserved *confounding variable* that was related causally to the other two variables and so makes it appear that the two observed variables were causally linked when they were not (Fig. 2-4B). For example, a tobacco industry consultant has claimed that nonsmokers married to smokers are more likely to own pet birds and that the birds spread diseases that increase the risk of lung cancer.*

The only way to completely exclude the possibility of confounding variables would be to conduct a randomized trial in which nonsmokers were randomly selected from the population, randomly allocated to marry other nonsmokers or smokers, then monitored for many years to see who developed heart disease or lung cancer. (Presumably the ownership of pet birds would be randomly distributed between the people assigned to marry nonsmokers and assigned to marry smokers.) Such an experiment could never be done.

*Gardiner A, Lee P. Pet birds and lung cancer. *BMJ*. 1993;306(6869):60.

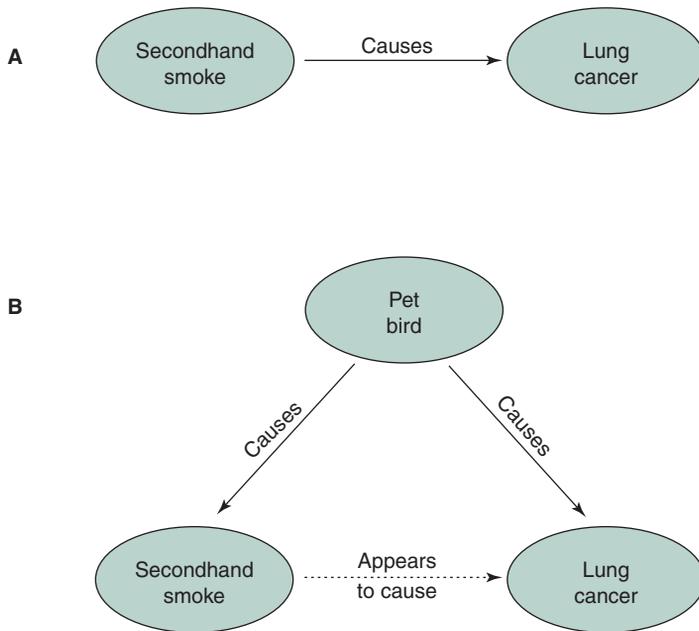


FIGURE 2-4. Panel **A** shows the situation that would exist if breathing secondhand smoke caused lung cancer. Panel **B** shows the situation that would exist if, as suggested by a tobacco industry consultant, people exposed to secondhand smoke were more likely to own pet birds and the birds carried diseases that caused lung cancer, while there was no connection between breathing secondhand smoke and lung cancer. Since owning a pet bird would be linked both to exposure to secondhand smoke and lung cancer this (unobserved) *confounding variable* could make it appear that secondhand smoke caused lung cancer when, in fact, there was no link.

It is, however, still possible to conclude that there are causal links between exposure to some agent (such as secondhand smoke) and an outcome (such as lung cancer) from observational studies. Doing so requires studies that account for known confounding variables either through an experimental design that separates people based on the effect of the confounding variable (by stratifying the confounding variable) or by controlling for their effects using more advanced statistical procedures,* and considering other related experimental evidence that helps explain the biologic mechanisms that cause the disease. These considerations have led reputable scientists and health authorities to conclude that secondhand smoke causes both lung cancer and heart disease.

The statistical techniques for analyzing data collected from experiments and observational studies are the same. The differences lie in how you interpret the results, particularly how confident you can be in using the word “cause.”

Randomized Clinical Trials

One procedure, called a *randomized clinical trial*, is the method of choice for evaluating therapies because it

avoids the selection biases that can creep into observational studies. The randomized clinical trial is an example of what statisticians call an *experimental study* because the investigator actively manipulates the treatment under study, making it possible to draw much stronger conclusions than are possible from observational studies about whether or not a treatment produced an effect. Experimental studies are the rule in the physical sciences and animal studies in the life sciences but are less common in studies involving human subjects.

Randomization reduces biases that can appear in observational studies and, since all clinical trials are *prospective*, no one knows how things will turn out at the beginning. This fact also reduces the opportunity for bias. Perhaps for these reasons, randomized clinical trials often show therapies to be of little or no value, even when observational studies have suggested that they were efficacious.†

Why, then, are not all therapies subjected to randomized clinical trials? Once something has become part of generally accepted medical practice—even if it did so

*For a discussion of the statistical approaches to control for confounding variables, see Glantz SA, Slinker BK. Regression with a qualitative dependent variable. In: *Primer of Applied Regression and Analysis of Variance*, 2nd ed. New York: McGraw-Hill; 2001:chap 12.

†For a readable and classic discussion of the place of randomized clinical trials in providing useful clinical knowledge, together with a sobering discussion of how little of commonly accepted medical practice has ever been actually shown to do any good, see Cochran K. *Effectiveness and Efficiency: Random Reflections on Health Services*. London: Nuffield Provincial Hospitals Trust; 1972.

without any objective demonstration of its value—it is extremely difficult to convince patients and their physicians to participate in a study that requires withholding it from some of the patients. Second, randomized clinical trials are always prospective; a person recruited into the study must be monitored for some time, often many years. People move, lose interest, or die for reasons unrelated to the study. Simply keeping track of people in a randomized clinical trial is often a major task.

To collect enough patients to have a meaningful sample, it is often necessary to have many groups at different institutions participating. While it is great fun for the people running the study, it is often just one more task for the people at the collaborating institutions. All these factors often combine to make randomized clinical trials expensive and difficult to execute. Nevertheless, when done well, they provide the most definitive answers to questions regarding the relative efficacy of different treatments.

■ HOW TO ESTIMATE THE MEAN AND STANDARD DEVIATION FROM A SAMPLE

Having obtained a random sample from a population of interest, we are ready to use information from that sample to estimate the characteristics of the underlying population. The estimate of the population mean is called the *sample mean* and is defined analogously to the population mean:

$$\text{Sample mean} = \frac{\text{Sum of values, e.g., heights, of each observation in sample}}{\text{Number of observations in sample}}$$

The equivalent mathematical statement is

$$\bar{X} = \frac{\sum X}{n}$$

in which the bar over the X denotes that it is the mean of the n observations of X .

The estimate of the population standard deviation is called the *sample standard deviation* s and is defined as

$$\text{Sample standard deviation} = \sqrt{\frac{\text{Sum of (value of observation in the sample - sample mean)}^2}{\text{Number of observations in sample} - 1}}$$

or, mathematically,*

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$$

(The standard deviation is also often denoted as SD.)

The definition of the sample standard deviation, s , differs from the definition of the population standard deviation σ in two ways: (1) the population mean μ has been replaced by our estimate of it, the sample mean \bar{X} , and (2) we compute the “average” squared deviation of a sample by dividing by $n - 1$ rather than n . The precise reason for dividing by $n - 1$ rather than n requires substantial mathematical arguments, but we can present the following intuitive justification: The sample will never show as much variability as the entire population and dividing by $n - 1$ instead of n compensates for the resultant tendency of the sample standard deviation to underestimate the population standard deviation.

In conclusion, if you are willing to assume that the sample was drawn from a normal distribution, summarize data with the sample mean and sample standard deviation, the best estimates of the population mean and population standard deviation, because these two parameters completely define the normal distribution. When there is evidence that the population under study does not follow a normal distribution, summarize data with the median and upper and lower percentiles discussed later in this chapter.

■ HOW GOOD ARE THESE ESTIMATES?

The mean and standard deviation computed from a random sample are estimates of the mean and standard deviation of the entire population from which the sample was drawn. There is nothing special about the specific random sample used to compute these statistics, and different random samples will yield slightly different estimates of the true population mean and standard deviation. To quantify how accurate these estimates are likely to be, we can compute their *standard errors*. It is possible to compute a standard error for any statistic, but here we shall focus on the *standard error of the mean*. This statistic quantifies the

*All equations in the text will be presented in the form most conducive to understanding statistical concepts. Often there is another, mathematically equivalent, form of the equation which is more suitable for computation. These forms are tabulated in Appendix A.

certainty with which the mean computed from a random sample estimates the true mean of the population from which the sample was drawn.

What is the standard error of the mean?

Figure 2-5A shows the same population of Martian heights we considered before. Since we have complete knowledge of every Martian’s height, we will use this example to explore how accurately statistics computed from a random sample describe the entire population. Suppose that we draw a random sample of 10 Martians from the entire population of 200, then compute the sample mean and sample standard deviation. The 10 Martians in the sample are indicated by solid points in Figure 2-5A.

Figure 2-5B shows the results of this random sample as it might be reported in a journal article, together with the sample mean ($\bar{X} = 41.5$ cm) and sample standard deviation ($s = 3.8$ cm). The values are close, but not equal, to the population mean ($\mu = 40$ cm) and standard deviation ($\sigma = 5$ cm).

There is nothing special about this sample—after all, it was drawn at random—so let us consider a second random sample of 10 Martians from the same population of 200. Figure 2-5C shows the results of this sample, with the corresponding Martians that comprise the sample identified in Figure 2-5A. While the mean and standard deviation, 36 and 5 cm, of this second random sample are also

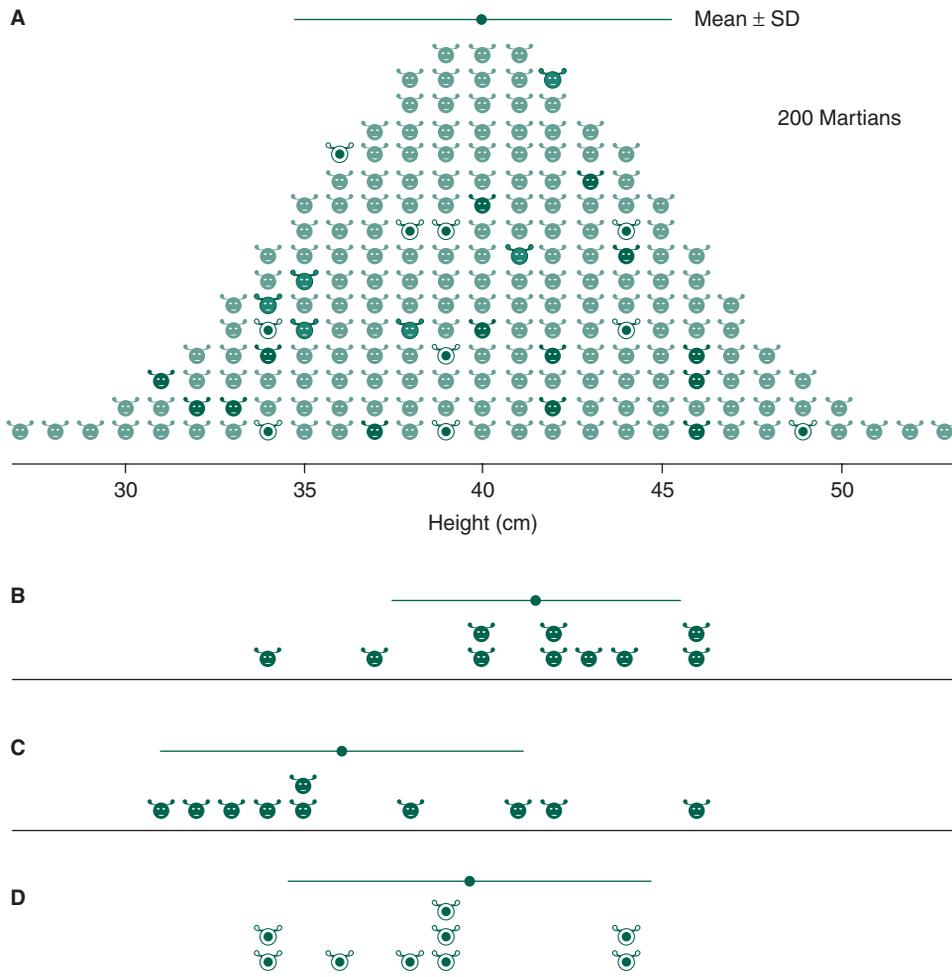


FIGURE 2-5. If one draws three different samples of 10 members each from a single population, one will obtain three different estimates of the mean and standard deviation.

similar to the mean and standard deviation of the whole population, they are not the same. Likewise, they are also similar, but not identical, to those from the first sample.

Figure 2-5D shows a third random sample of 10 Martians, identified in Figure 2-5A with circles containing dots. This sample leads to estimates of 40 and 5 cm for the mean and standard deviation.

Now, we make an important change in emphasis. Instead of concentrating on the population of all 200 Martians, let us examine the *means of all possible random samples of 10 Martians*. We have already found three possible values for this mean, 41.5, 36, and 40 cm, and there are many more possibilities. Figure 2-6 shows these three means, using the same symbols as Figure 2-5. To better understand the amount of variability in the means of samples of 10 Martians, let us draw another 22 random samples of 10 Martians each and compute the mean of each sample. These additional means are plotted in Figure 2-6 as open points.

Now that we have drawn 25 random samples of 10 Martians each, have we exhausted the entire population of 200 Martians? No. There are more than 10^{16} different ways to select 10 Martians at random from the population of 200 Martians.

Look at Figure 2-6. The collection of the means of 25 random samples, each of 10 Martians, has a roughly bell-shaped distribution, which is similar to the normal distribution. When the variable of interest is the sum of

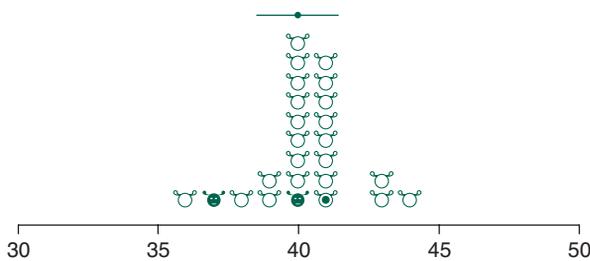


FIGURE 2-6. If one draws more and more samples—each with 10 members—from a single population, one eventually obtains the population of all possible sample means. This figure illustrates the means of 25 samples of 10 Martians each drawn from the population of 200 Martians shown in Figures 2-1 and 2-5A. The means of the three specific samples shown in Figure 2-5 are shown using corresponding symbols. This new population of all possible sample means will be normally distributed regardless of the nature of the original population; its mean will equal the mean of the original population; its standard deviation is called the standard error of the mean.

many other independent random variables, its distribution will tend to be normal, regardless of the distributions of the variables used to form the sum. Since the sample mean is just such a sum, its distribution will tend to be normal, with the approximation improving as the sample size increases. (If the sample were drawn from a normally distributed population, the distribution of the sample means would have a normal distribution regardless of the sample size.) Therefore, it makes sense to describe the data in Figure 2-6 by computing their mean and standard deviation. Since the mean value of the 25 points in Figure 2-6 is the mean of the means of 25 samples, we will denote it $\bar{\bar{X}}$. The standard deviation is the *standard deviation of the means* of 25 independent random samples of 10 Martians each, and so we will denote it $\sigma_{\bar{X}}$. Using the formulas for mean and standard deviation presented earlier, we compute $\bar{\bar{X}} = 40$ cm and $\sigma_{\bar{X}} = 1.6$ cm.

The mean of the sample means $\bar{\bar{X}}$ is (within measurement and rounding error) equal to the mean height μ of the entire population of 200 Martians from which we drew the random samples. This is quite a remarkable result, since $\bar{\bar{X}}$ is *not* the mean of a sample drawn directly from the original population of 200 Martians; $\bar{\bar{X}}$ is the mean of 25 random samples of size 10 drawn from the *population consisting of all 10^{16} possible values of the mean of random samples of size 10 drawn from the original population of 200 Martians*.

Is $\sigma_{\bar{X}}$ equal to the standard deviation σ of the population of 200 Martians? No. In fact, it is quite a bit smaller; the standard deviation of the collection of sample means $\sigma_{\bar{X}}$ is 1.6 cm while the standard deviation for the whole population is 5 cm. Just as the standard deviation of the original sample of 10 Martians s is an estimate of the variability of Martians' heights, $\sigma_{\bar{X}}$ is an estimate of the *variability of possible values of means of samples of 10 Martians*. Since when one computes the mean, extreme values tend to balance each other, there will be less variability in the values of the sample means than in the original population. $\sigma_{\bar{X}}$ is a measure of the precision with which a sample mean \bar{X} estimates the population mean μ . We might name $\sigma_{\bar{X}}$ “standard deviation of means of random samples of size 10 drawn from the original population.” To be brief, statisticians have coined a shorter name, the *standard error of the mean* (SEM).

Since the precision with which we can estimate the mean increases as the sample size increases, the standard error of the mean decreases as the sample size increases. Conversely, the more variability in the original population,

the more variability will appear in possible mean values of samples; therefore, the standard error of the mean increases as the population standard deviation increases. The true standard error of the mean of samples of size n drawn from a population with standard deviation σ is*

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The best estimate of $\sigma_{\bar{x}}$ from a single sample is

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Since the possible values of the sample mean tend to follow a normal distribution, the true (and unobserved) mean of the original population will lie within 2 standard errors of the sample mean about 95% of the time.

As already noted, mathematicians have shown that the distribution of mean values will always approximately follow a normal distribution *regardless* of how the population from which the original samples were drawn is distributed. We have developed what statisticians call the *Central Limit Theorem*. It says:

- The distribution of sample means will be approximately normal regardless of the distribution of values in the original population from which the samples were drawn.
- The mean value of the collection of all possible sample means will equal the mean of the original population.
- The standard deviation of the collection of all possible means of samples of a given size, called the standard error of the mean, depends on both the standard deviation of the original population and the size of the sample.

Figure 2-7 illustrates the relationship between the sample mean, the sample standard deviation, and the standard error of the mean and how they vary with sample size as we measure more and more Martians.[†] As we add more Martians to our sample, the sample mean \bar{X} and standard deviation s estimate the popula-

tion mean μ and standard deviation σ with increasing precision. This increase in the precision with which the sample mean estimates the population mean is reflected by the smaller standard error of the mean with larger sample sizes. Therefore, the standard error of the mean tells not about variability in the original population, as the standard deviation does, but about the certainty with which a sample mean estimates the true population mean.

The *standard deviation and standard error of the mean* measure two very different things and are often confused. Most medical investigators summarize their data with the standard error of the mean because it is always smaller than the standard deviation. It makes their data look better. However, unlike the standard deviation, which quantifies the *variability in the population*, the standard error of the mean quantifies *uncertainty in the estimate of the mean*. Since readers are generally interested in knowing about the population, data should generally not be summarized with the standard error of the mean.

To understand the difference between the standard deviation and standard error of the mean and why one ought to summarize data using the standard deviation, suppose that in a sample of 20 patients an investigator reports that the mean cardiac output was 5.0 L/min with a standard deviation of 1 L/min. Since about 95% of all population members fall within about 2 standard deviations of the mean, this report would tell you that, assuming that the population of interest followed a normal distribution, it would be unusual to observe a cardiac output below about 3 or above about 7 L/min. Thus, you have a quick summary of the population described in the paper and a range against which to compare specific patients you examine. Unfortunately, it is unlikely that these numbers would be reported, the investigator being more likely to say that the cardiac output was 5.0 ± 0.22 (SEM) L/min. If you confuse the standard error of the mean with the standard deviation, you would believe that the range of most of the population was narrow indeed—4.56 to 5.44 L/min. These values describe the range which, with about 95% confidence, contains the mean cardiac output of the entire population from which the sample of 20 patients was drawn. (Chapter 7 discusses these ideas in detail.) In practice, one generally wants to compare a specific patient's cardiac output not only with the population mean but with the spread in the population taken as a whole.

*This equation is derived in Chapter 4.

[†]Figure 2-7 was obtained by selecting two Martians from Figure 2-1 at random, then computing \bar{X} , s , and $\sigma_{\bar{x}}$. Then one more Martian was selected and the computations done again. Then, a fourth, a fifth, and so on, always adding to the sample already drawn. Had we selected different random samples or the same samples in a different order, Figure 2-7 would have been different.

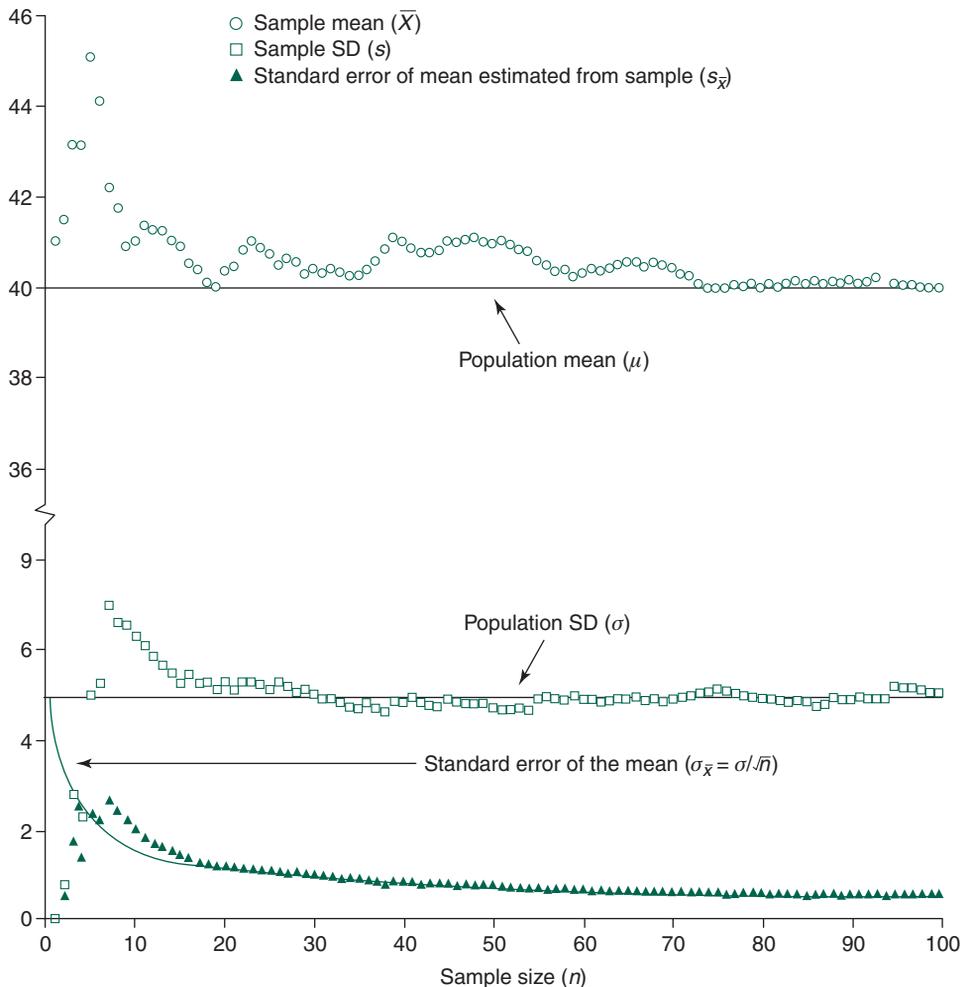


FIGURE 2-7. As the size of a random sample of Martians drawn from the population depicted in Figure 2-1 grows, the precision with which the sample mean and sample standard deviation, \bar{X} and s , estimate the true population mean and standard deviation, μ and σ , increases. This increasing precision appears in two ways: (1) the difference between the statistics computed from the sample (the points) moves closer to the true population values (the lines), and (2) the size of the standard error of the mean decreases.

■ PERCENTILES

Armed with our understanding of how to describe normally distributed populations using the mean and standard deviation, we extend our research efforts and measure the heights of all Jupiter's inhabitants but also to compute the mean and standard deviation of the heights of all Jovians. The resulting data show the mean height to be 37.6 cm and the standard deviation of heights to be

4.5 cm. By comparison with Table 2-1, Jovians appear quite similar in height to Martians, since these two parameters completely specify a normal distribution.

The actual distribution of heights on Jupiter, however, tells a different story. Figure 2-8A shows that, unlike those living on the other two planets, a given Jovian is not equally likely to have a height above average as below average; the distribution of heights of all population members is no longer symmetric but *skewed*. The few individuals

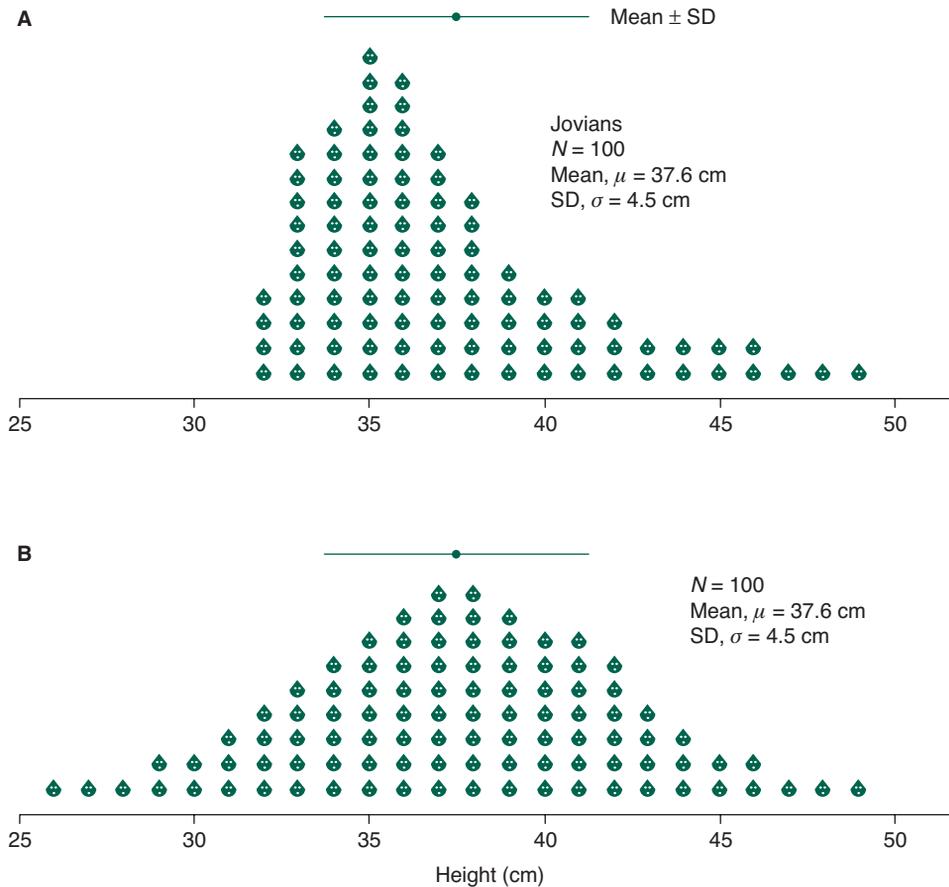


FIGURE 2-8. When the population values are not distributed symmetrically about the mean, reporting the mean and standard deviation can give the reader an inaccurate impression of the distribution of values in the population. Panel **A** shows the true distribution of the heights of the 100 Jovians (note that it is skewed toward taller heights). Panel **B** shows a normally distributed population with 100 members and the same mean and standard deviation as in panel **A**. Despite the fact that the means and standard deviations are the same, the distributions of heights in the two populations are quite different.

who are much taller than the rest increase the mean and standard deviation in a way that led us to think that most of the heights were higher than they actually are and that the variability of heights was greater than it actually is. Specifically, Figure 2-8B shows a population of 100 individuals whose heights are distributed according to a normal or Gaussian distribution with the same mean and standard deviation as the 100 Jovians in Figure 2-8A. It is quite different. So, although we can compute the mean and standard deviation of heights of Jupiter's—or, for that matter, any—population, these two numbers do not summarize the distribution of heights nearly as accurately

as they did when the heights in the population followed a normal distribution.

An alternative approach that better describes such data is to report the *median*. The median is the value that half the members of the population fall below. Figure 2-9A shows that half the Jovians are shorter than 36 cm; 36 cm is the median. Since 50% of the population values fall below the median, it is also called the *50th percentile*.

Calculation of the *median* and other *percentiles* is simple. First, list the n observations in order. The median, the value that defines the lower half of the observations, is simply the $.5(n + 1)$ observation. When there are an odd

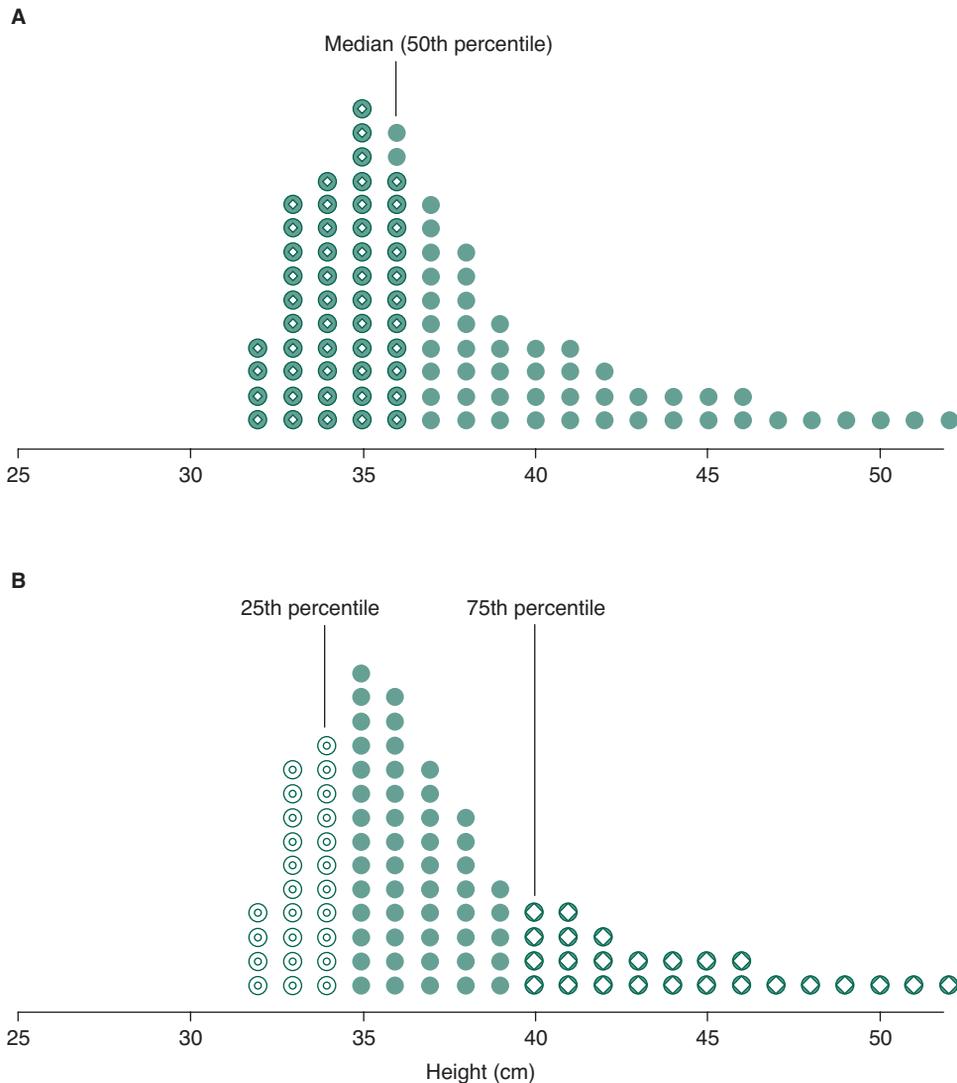


FIGURE 2-9. One way to describe a skewed distribution is with percentiles. The median is the point that divides the population in half. Panel **A** shows that 36 cm is the median height on Jupiter. Panel **B** shows the 25th and 75th percentiles, the points locating the lowest and highest quarter of the heights, respectively. The fact that the 25th percentile is closer to median than the 75th percentile indicates that the distribution is skewed toward higher values.

number of observations, the median falls on one of the observations. For example, if there are 27 observations, the $.5(27 + 1) = 14$ th observation (listed from smallest to largest) is the median. When there is an even number of observations, the median falls between two observations. For example, if there are 40 observations, the median would be the $.5(40 + 1) = 20.5$ th observation. Since there

is no 20.5th observation, we take the average of 20th and 21st observation.

Other percentile points are defined analogously. For example, the 25th percentile point, the point that defines the lowest quarter of the observations, is just the $.25(n + 1)$ observation. Again, if the value falls between two observations, take the mean of the two surrounding observations.

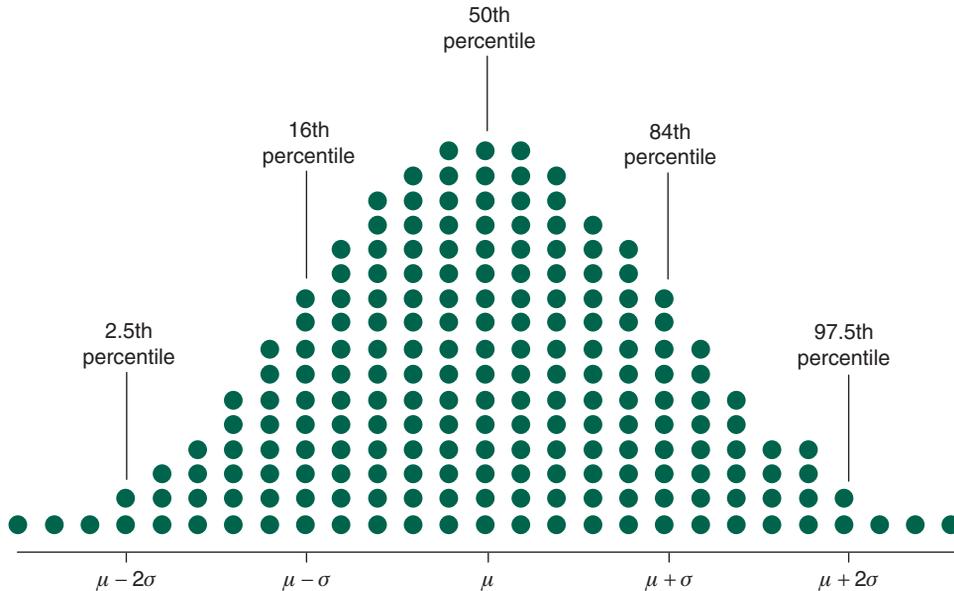


FIGURE 2-10. Percentile points of the normal distribution.

In general, the p th percentile point is the $(p/100)(n + 1)$ observation.*

To give some indication of the dispersion of heights in the population, report the value that separates the lowest (shortest) 25% of the population from the rest and the value that separates the shortest 75% of the population from the rest. These two points are called the *25th* and *75th percentile* points, respectively, and the interval they define is called the *interquartile range*. For the Jovians, Figure 2-9B shows that these percentiles are 34 and 40 cm. While these three numbers (the 25th, 50th, and 75th percentile

points, 34, 36, and 40 cm) do not precisely describe the distribution of heights, they do indicate what the range of heights is and that there are a few very tall Jovians but not many very short ones.

Although these percentiles are often used, one could equally well report the 5th and 95th percentile points, or, for that matter, report the 5th, 25th, 50th, 75th, and 95th percentile points.

Computing the percentile points of a population is a good way to see how close to a normal distribution it is. Recall that we said that in a population that exhibits a normal distribution of values, about 95% of the population members fall within 2 standard deviations of the mean and about 68% fall within 1 standard deviation of the mean. Figure 2-10 shows that, for a normal distribution, the values of the associated percentile points are:

2.5th percentile	Mean - 2 standard deviation
16th percentile	Mean - 1 standard deviation
25th percentile	Mean - 0.67 standard deviation
50th percentile (median)	Mean
75th percentile	Mean + 0.67 standard deviation
84th percentile	Mean + 1 standard deviation
97.5th percentile	Mean + 2 standard deviation

If the values associated with the percentiles are not too different from what one would expect on the basis of the

*An alternative definition for the percentile value when the percentile point falls between two observations is to interpolate between the observation above and below the percentile point, rather than just averaging the observations. For example, in a problem in which there are 14 data points, the 75th percentile would be the $(p/100)(n + 1) = (75/100)(14 + 1) = 11.25$ observation. Using the approach in the text, we would just average the 11th and 12th observation. Using the alternative definition we would use the value 0.25 of the way between the 11th and 12th observations. If the 11th observation is 34 and the 12th observation is 40, using the definition of percentile in the text, we would estimate the 75th percentile as $(34 + 40)/2 = 37$. Interpolating between the two observations, we would compute the 75th percentile as $34 + 0.25(40 - 34) = 35.5$. (Appendix A describes how to interpolate in general.) Most computer programs use the interpolation approach. As a practical matter, when sample sizes are large, there is little or no difference between the two different ways of computing percentiles.

mean and standard deviation, the normal distribution is a good approximation to the true population and then the mean and standard deviation describe the population adequately.

Why care whether or not the normal distribution is a good approximation? Because many of the statistical procedures used to test hypotheses—including the ones we will develop in Chapters 3, 4, and 9—require that the population follow a normal distribution at least approximately for the tests to be reliable. (Chapters 10 and 11 present alternative tests that do not require this assumption.)

Pain Associated with Diabetic Neuropathy

Peripheral neuropathy is a complication of diabetes mellitus in which peripheral nerves are damaged, leading to many symptoms, including spasms, tingling, numbness and pain. Because conventional treatments are often ineffective or have serious side effects, Dinesh Selvarajah and colleagues* conducted a randomized placebo-controlled double blind clinical trial of a cannabis-based medicinal extract in people with intractable pain.

They recruited people for the study who had not had their pain controlled using other drugs and randomly assigned them to receive the cannabis extract or a placebo for 12 weeks. The use of a placebo was particularly important because of the *placebo effect*, when people report feeling better because they are being treated, even if the treatment had no biological effect on the underlying disease process. The experiment was also *double blind*, with neither the experimental subjects nor the investigators knowing who was receiving the drug or placebo. Double blinding was particularly important because the outcome was a subjective measure of pain that could be biased not only by the placebo effect, but a desire on the part of the experimental subjects to please the investigators by reporting less pain. It was also important that the investigators were blinded to the treatment group to avoid biasing clinical assessments or subtly encouraging the experimental subjects to bias their reported subjective pain scores.

*Selvarajah D, Gandhi R, Emery CJ, Tesfaye S. Randomized placebo-controlled double-blind clinical trial of cannabis-based medicinal product (Sativex) in painful diabetic neuropathy. *Diabetes Care* 2010;33:128–130.

■ TABLE 2-3. Measured Pain in 29 People with Diabetic Neuropathy ($n = 29$)

13	4	19
8	16	37
46	23	13
61	33	8
28	18	28
7	51	25
93	26	4
10	19	12
7	20	12
100	54	

The investigators used standard questionnaires that measured superficial, deep and muscular pain, and then averaged the three scores to get a total pain score. Higher scores indicate greater pain. The data for the placebo appear in Table 2-3.

Figure 2-11 shows a plot of these data in a way that shows how they are distributed along the pain scale. Such a plot is called a *histogram*.[†] Simply looking at this histogram suggests that the data are not drawn from a normally distributed population because the observations do not seem to be symmetrically distributed about the mean following the bell-shaped curve that describes the normal distribution.

As Box 2-1 shows, the mean pain score is 27.4 with a standard deviation of 24.5. If these data had been drawn from a normal distribution, about 95% of population members would have been within about 2 standard deviations of the mean, from about $27.4 - 2 \times 24.5 = -21.6$ to about $27.4 + 2 \times 24.5 = 76.4$. The pain score ranges from 0 to 100, while the upper end of this range is plausible, the lower end is not: the pain score cannot be negative, so the population is highly unlikely to be normally distributed. (Such a comparison can be used as an informal test for normality when the measurement cannot be negative.)

[†]In general histograms can display the data over a range of values in each bin. The histogram in Figure 2-11 that has bins 1 unit wide (i.e., that shows the number of observations at each observed value) is also called a *dot plot*.

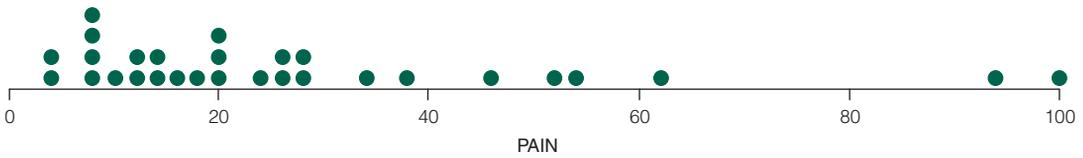


FIGURE 2-11. Level of pain reported among people with diabetic neuropathy after 12 weeks of taking a placebo.

BOX 2-1 • Descriptive Statistics for the Data on Diabetic Neuropathy in Table 2-3

Sorted Data from Table 2-3

Data	Observation Number	Data	Observation Number
4	1	20	16
4	2	23	17
7	3	25	18
7	4	26	19
8	5	28	20
8	6	28	21
10	7	33	22
12	8	37	23
12	9	46	24
13	10	51	25
13	11	54	26
16	12	61	27
18	13	93	28
19	14	100	29
19	15		

To estimate the mean, we simply add up all the observations and divide by the number of observations. From the data in Table 2-3,

$$\bar{X} = \frac{\sum X}{n} = \frac{13 + 8 + 46 + \cdots + 12 + 12}{29} = 27.4$$

Therefore the estimate of the standard deviation from the sample is

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} = \sqrt{\frac{(13 - 27.4)^2 + (13 - 27.4)^2 + (46 - 27.4)^2 + \cdots + (12 - 27.4)^2 + (12 - 27.4)^2}{29 - 1}} = 24.5$$

To compute the median and percentile points, we first sort the observations in Table 2-3 in ascending order, as shown in the table in this box. The median, 50th percentile point, of the $n = 29$ observations is the $(p/100)(n + 1) = (50/100)(29 + 1) = 14$ th data point, a value of 19. The 25th percentile is the $(25/100)(29 + 1) = 7.5$ th point. Taking the mean of the 7th and 8th observation, we find that the 25th percentile point is $(12 + 12)/2 = 11$. Likewise, the 75th percentile point is the $(75/100)(29 + 1) = 22.5$ th observation. Taking the mean of the 22nd and 23rd observations, we find that the 75th percentile point is $(33 + 37)/2 = 35$.

Because these data do not seem to follow a normal distribution, the best way to describe them is with the median and top and bottom quartiles. Box 2-1 shows that the median of these data is 19 and the 25th and 75th percentile points are 11 and 35. The fact that the 25th percentile point is much closer to the median than the 75th percentile point is a reflection of the fact that the distribution is not symmetrical, which is further evidence that the underlying population is not normally distributed.

■ SUMMARY

When a population follows a normal distribution, we can describe its location and variability completely with two parameters—the mean and standard deviation. When the population does not follow a normal distribution at least roughly, it is more appropriate to describe it with the median and other percentiles. Since one can rarely observe all members of a population, we will estimate these parameters from a sample drawn at random from the population. The standard error quantifies the precision of these estimates. For example, the standard error of the mean quantifies the precision with which the sample mean estimates the population mean.

In addition to being useful for describing a population or sample, these numbers can be used to estimate how compatible measurements are with clinical or scientific assertions that an intervention affected some variable. We now turn our attention to this problem.

■ PROBLEMS

2-1 The pain scores for the people treated with the cannabis medicinal in the study discussed earlier in this chapter are 90, 10, 45, 70, 13, 27, 11, 70, 14, 15, 13, 75, 50, 30, 80, 40, 29, 13, 9, 7, 20, 85, 55, and 94. Find the mean, median, standard deviation, and 25th and 75th percentiles. Do these data seem to be drawn from a normally distributed population? Why or why not?

2-2 Viral load of HIV-1 is a known risk factor for heterosexual transmission of HIV; people with higher viral loads of HIV-1 are significantly more likely to transmit the virus to their uninfected partners. Thomas Quinn and associates.*

studied this question by measuring the amount of HIV-1 RNA detected in blood serum. The following data represent HIV-1 RNA levels in the group whose partners seroconverted, which means that an initially uninfected partner became HIV positive during the course of the study; 79,725, 12,862, 18,022, 76,712, 25,6440, 14,013, 46,083, 6808, 85,781, 1251, 6081, 50,397, 11,020, 13,633, 1064, 496, 433, 25,308, 6616, 11,210, 13,900 RNA copies/mL. Find the mean, median, standard deviation, and 25th and 75th percentiles of these concentrations. Do these data seem to be drawn from a normally distributed population? Why or why not?

2-3 When data are not normally distributed, researchers can sometimes *transform* their data to obtain values that more closely approximate a normal distribution. One approach to this is to take the logarithm of the observations. The following numbers represent the same data described in Prob. 2-1 following log (base 10) transformation: 4.90, 4.11, 4.26, 4.88, 5.41, 4.15, 4.66, 3.83, 4.93, 3.10, 3.78, 4.70, 4.04, 4.13, 3.03, 5.70, 4.40, 3.82, 4.05, 4.14. Find the mean, median, standard deviation, and 25th and 75th percentiles of these concentrations. Do these data seem to be drawn from a normally distributed population? Why or why not?

2-4 Polychlorinated biphenyls (PCBs) are a class of environmental chemicals associated with a variety of adverse health effects, including intellectual impairment in children exposed *in utero* while their mothers were pregnant. PCBs are also one of the most abundant contaminants found in human fat. Tu Binh Minh and colleagues[†] analyzed PCB concentrations in the fat of a group of Japanese adults. They detected 1800, 1800, 2600, 1300, 520, 3200, 1700, 2500, 560, 930, 2300, 2300, 1700, 720 ng/g lipid weight of PCBs in the people they studied. Find the mean, median standard deviation, and 25th and 75th percentiles of these concentrations. Do these data seem to be drawn from a normally distributed population? Why or why not?

2-5 Sketch the distribution of all possible values of the number on the upright face of a die. What is the mean of this population of possible values?

*Quinn TC, Wawer MJ, Sewankambo N, Serwadda D, Li C, Wabwire-Mangen F, Meehan MO, Lutalo T, Gray RH. Viral load and heterosexual transmission of human immunodeficiency virus type 1. *N Engl J Med*. 2000;342:921–929.

[†]Minh TB, Watanabe M, Tanabe S, Yamada T, Hata J, Watanabe S. Occurrence of tris (4-chlorophenyl)methane, tris (4-chlorophenyl)methanol, and some other persistent organochlorines in Japanese human adipose tissue. *Environ Health Perspect*. 2000;108:599–603.

2-6 Roll a *pair* of dice and note the numbers on each of the upright faces. These two numbers can be considered a sample of size 2 drawn from the population described in Prob. 2-4. This sample can be averaged. What does this

average estimate? Repeat this procedure 20 times and plot the averages observed after each roll. What is this distribution? Compute its mean and standard deviation. What do they represent?

How to Test for Differences between Groups

Statistical methods are used to summarize data and test hypotheses with those data. Chapter 2 discussed how to use the mean, standard deviation, median, and percentiles to summarize data and how to use the standard error of the mean to estimate the precision with which a sample mean estimates the population mean. Now we turn our attention to how to use data to test scientific hypotheses. The statistical techniques used to perform such tests are called *tests of significance*; they yield the highly prized *P value*. We now develop procedures to test the hypothesis that, on the average, different treatments all affect some variable identically. Specifically, we will develop a procedure to test the hypothesis that diet has no effect on the mean cardiac output of people living in a small town. Statisticians call this hypothesis of no effect the *null hypothesis*.

The resulting test can be generalized to analyze data obtained in experiments involving any number of treatments. In addition, it is the archetype for a whole class of related procedures known as *analysis of variance*.

■ THE GENERAL APPROACH

To begin our experiment, we randomly select four groups of seven people each from a small town with 200 healthy adult inhabitants. All participants give informed consent. People in the control group continue eating normally; people in the second group eat only spaghetti; people in the third group eat only steak; and people in the fourth group eat only fruit and nuts. After 1 month, each person

has a cardiac catheter inserted and his or her cardiac output is measured.

As with most tests of significance, we begin with the hypothesis that all treatments (diets) have the same effect (on cardiac output). Since the study includes a control group (as experiments generally should), this hypothesis is equivalent to the hypothesis that diet has no effect on cardiac output. Figure 3-1 shows the distribution of cardiac outputs for the entire population, with each individual's cardiac output represented by a circle. The specific individuals who were randomly selected for each diet are indicated by shaded circles, with different shading for different diets. Figure 3-1 shows that the null hypothesis is, in fact, true. Unfortunately, as investigators we cannot observe the entire population and are left with the problem of deciding whether or not to reject the null hypothesis from the limited data shown in Figure 3-2. There are obviously differences between the samples; the question is: *Are these differences due to the fact that the different groups of people ate differently or are these differences simply a reflection of the random variation in cardiac output between individuals?*

To use the data in Figure 3-2 to address this question, we proceed under the assumption that the null hypothesis that diet has no effect on cardiac output is correct. Since we assume that it does not matter which diet any particular individual ate, we *assume* that the four experimental groups of seven people each are four random samples of size 7 drawn from a single population of 200 individuals.

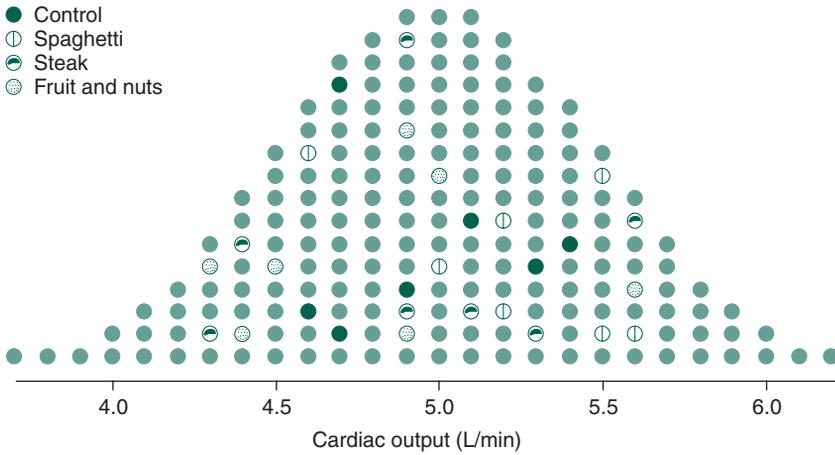


FIGURE 3-1. The values of cardiac output associated with all 200 members of the population of a small town. Since diet does not affect cardiac output, the four groups of seven people each selected at random to participate in our experiment (control, spaghetti, steak, and nuts) simply represent four random samples drawn from a single population.

Since the samples are drawn at random from a population with some variance, we expect the samples to have different means and standard deviations, but *if our null hypothesis that the diet has no effect on cardiac output is true*, the observed differences are simply due to random sampling.

Forget about statistics for a moment. What is it about different samples that leads you to believe that they are representative samples drawn from different populations? Figures 3-2, 3-3, and 3-4 show three different possible sets of samples of some variable of interest. Simply looking at

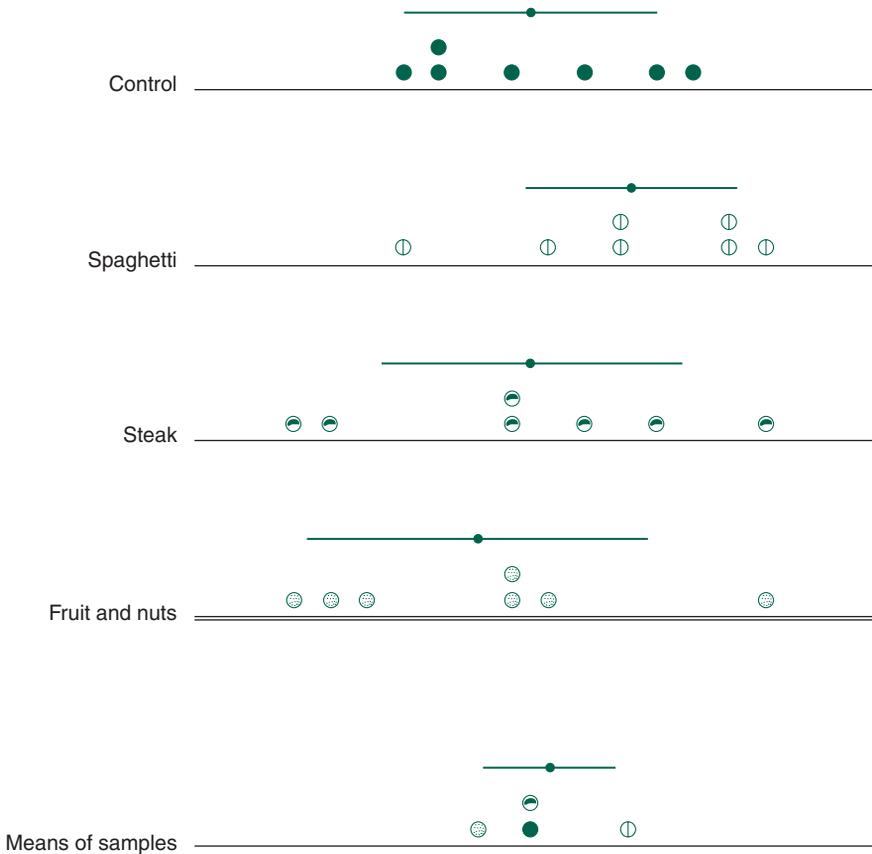


FIGURE 3-2. An investigator cannot observe the entire population but only the four samples selected at random for treatment. This figure shows the same four groups of individuals as in Figure 3-1 with their means and standard deviations as they would appear to the investigator. The question facing the investigator is: Are the observed differences due to the different diets or simply random variation? The figure also shows the collection of sample means together with their standard deviation, which is an estimate of the standard error of the mean.

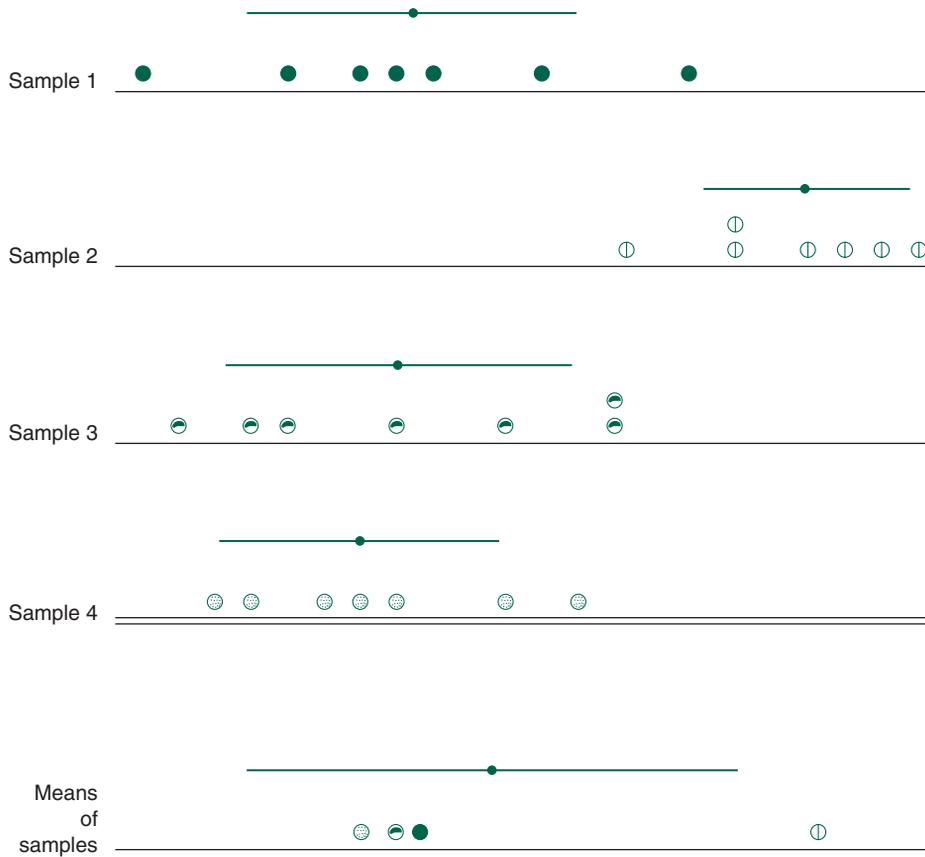


FIGURE 3-3. The four samples shown are identical to those in Figure 3-2 except that the variability in the mean values has been increased substantially. The samples now appear to differ from each other because the variability between the sample means is larger than one would expect from the variability within each sample. Compare the relative variability in mean values with the variability within the sample groups with that seen in Figure 3-2.

these pictures makes most people think that the four samples in Figure 3-2 were all drawn from a single population, while the samples in Figures 3-3 and 3-4 were not. Why? The variability within each sample, quantified with the standard deviation, is approximately the same. In Figure 3-2, the variability in the mean values of the samples is consistent with the variability one observes within the individual samples. In contrast, in Figures 3-3 and 3-4, the variability among sample means is much larger than one would expect from the variability within each sample. Notice that we reach this conclusion whether all (Fig. 3-3) or only one (Fig. 3-4) of the sample means appear to differ from the others.

Now let us formalize this analysis of variability to analyze our diet experiment. The standard deviation or its

square, the variance, is a good measure of variability. We will use the variance to construct a procedure to test the hypothesis that diet does not affect cardiac output.

Chapter 2 showed that two population parameters—the mean and standard deviation (or, equivalently, the variance)—completely describe a normally distributed population. Therefore, we will use our raw data to compute these parameters and then base our analysis on their values rather than on the raw data directly. Since the procedures, we will now develop are based on these parameters they are called *parametric statistical methods*. Because these methods assume that the population from which the samples were drawn can be completely described by these parameters, they are valid only when the real population approximately follows the normal

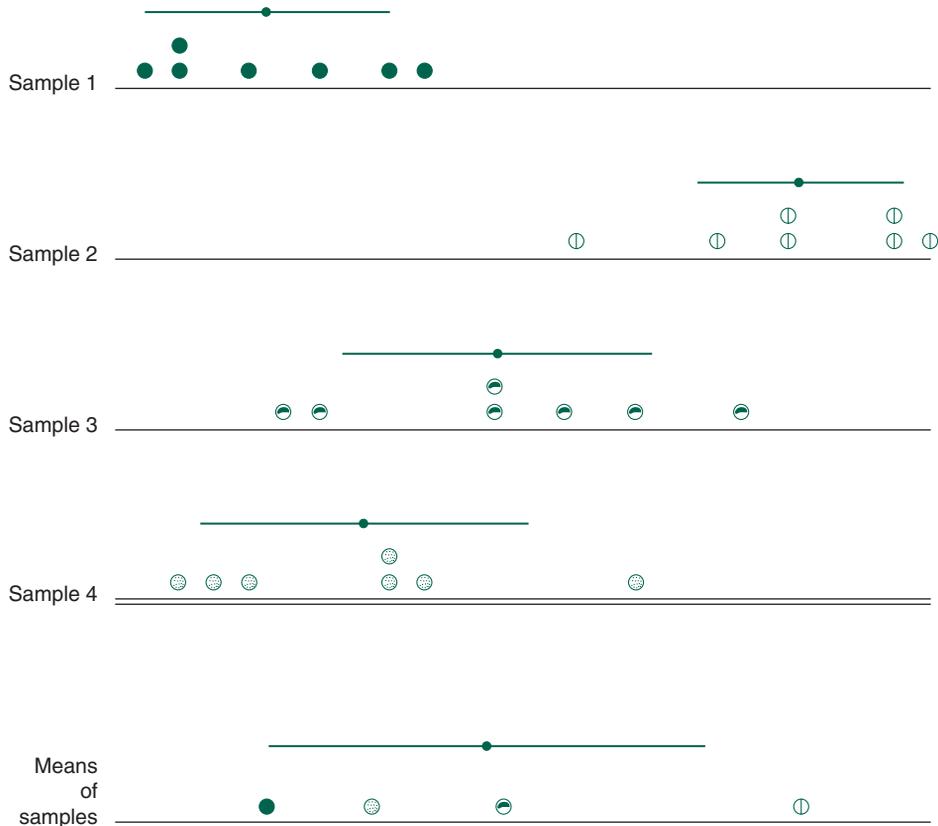


FIGURE 3-4. When the mean of even one of the samples (sample 2) differs substantially from the other samples, the variability computed from within the means is substantially larger than one would expect from examining the variability within the groups.

distribution. Other procedures, called *nonparametric statistical methods*, are based on frequencies, ranks, or percentiles do not require this assumption.* Parametric methods generally provide more information about the treatment being studied and are more likely to detect a real treatment effect when the underlying population is normally distributed.

We will estimate the parameter population variance in two different ways: (1) The standard deviation or variance computed from each sample is an estimate of the standard deviation or variance of the entire population. Since each of these estimates of the population variance

is computed from within each sample group, the estimates will not be affected by any differences in the mean values of different groups. (2) We will use the values of the means of each sample to determine a second estimate of the population variance. In this case, the differences between the means will obviously affect the resulting estimate of the population variance. If all the samples were, in fact, drawn from the same population (i.e., the diet had no effect), these two different ways to estimate the population variance should yield approximately the same number. When they do, we will conclude that the samples were likely to have been drawn from a single population; otherwise, we will reject this hypothesis and conclude that at least one of the samples was drawn from a different population. In our experiment, rejecting the original hypothesis would lead to the conclusion that diet *does* alter cardiac output.

*In fact, these methods make no assumption about the specific shape of the distribution of the underlying population; they are also called *distribution-free* methods. We will study these procedures in Chapters 5, 8, 10, and 11.

■ TWO DIFFERENT ESTIMATES OF THE POPULATION VARIANCE

How shall we estimate the population variance from the four sample variances? When the hypothesis that the diet does not affect cardiac output is true, the variances of each sample of seven people, regardless of what they ate, are equally good estimates of the population variance, so we simply average our four estimates of *variance within the treatment groups*:

Average variance in cardiac output within treatment groups = $1/4$ (variance in cardiac output of controls + variance in cardiac output of spaghetti eaters + variance in cardiac output of steak eaters + variance in cardiac output of fruit and nut eaters)

The mathematical equivalent is

$$s_{\text{wit}}^2 = \frac{1}{4} \left(s_{\text{con}}^2 + s_{\text{spa}}^2 + s_{\text{st}}^2 + s_{\text{f}}^2 \right)$$

where s^2 represents variance. The variance of each sample is computed with respect to the mean of that sample. Therefore, the population variance estimated from within the groups, *the within-groups variance* s_{wit}^2 , will be the same whether or not diet altered cardiac output.

Next, we estimate the population variance from the means of the samples. Since we have hypothesized that all four samples were drawn from a single population, the standard deviation of the four sample means will approximate the standard error of the mean. Recall that the standard error of the mean $\sigma_{\bar{x}}$ is related to the sample size n

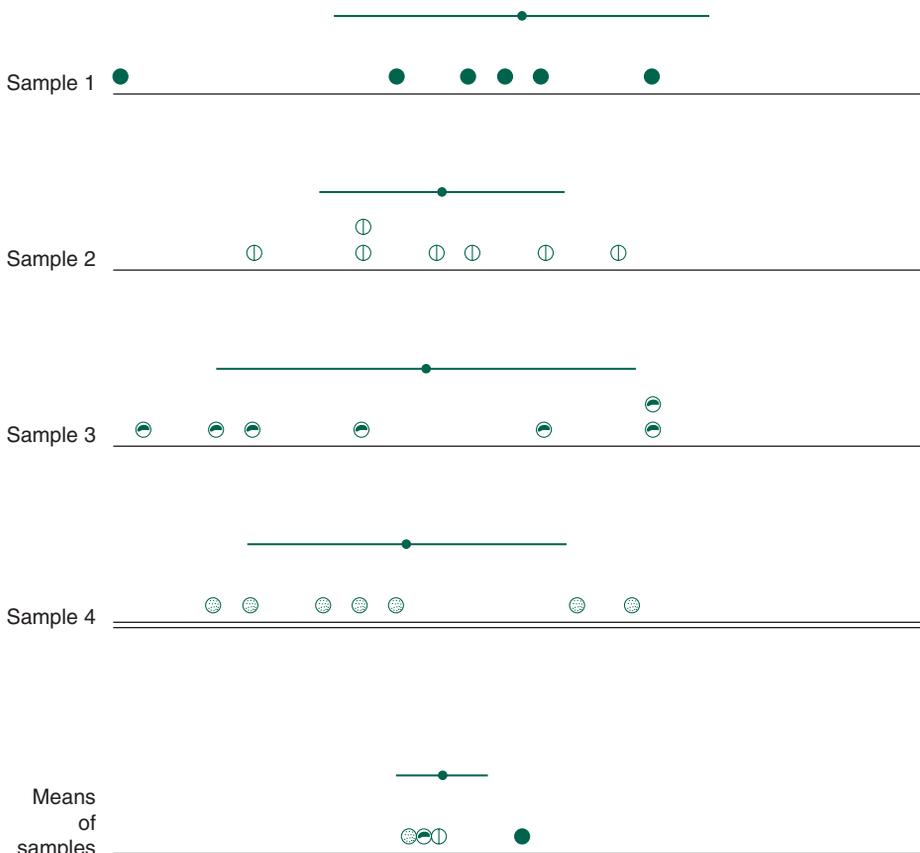


FIGURE 3-5. Four samples of seven members each drawn from the population shown in Figure 3-1. Note that the variability in sample means is consistent with the variability within each of the samples, $F = 0.5$.

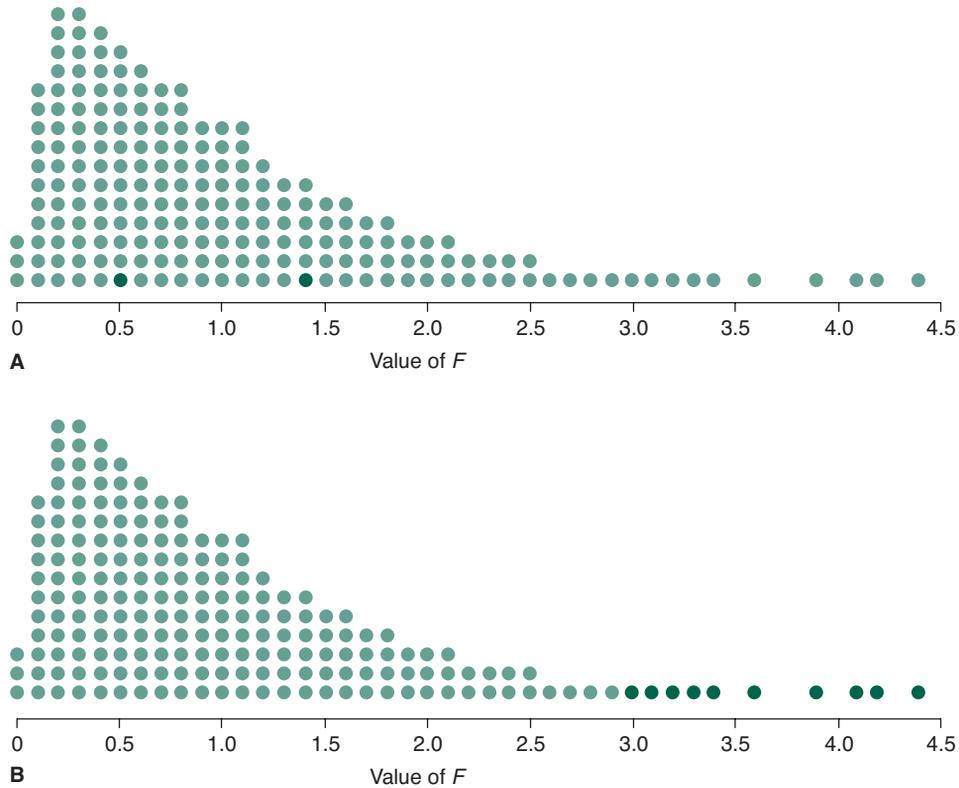


FIGURE 3-6. (A) Values of F computed from 200 experiments involving four samples, each of size 7, drawn from the population in Figure 3-1. **(B)** We expect F to exceed 3.0 only 5% of the time when all samples were, in fact, drawn from a single population. (continued)

(in this case 7) and the population standard deviation σ according to

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Therefore, the true population variance σ^2 is related to the sample size and standard error of the mean according to

$$\sigma^2 = n\sigma_{\bar{x}}^2$$

We use this relationship to estimate the population variance from the variability between the sample means using

$$s_{\text{bet}}^2 = ns_{\bar{x}}^2$$

where s_{bet}^2 is the estimate of the population variance computed from between the sample means and $s_{\bar{x}}$ is the standard deviation of the means of the four sample groups, the standard error of the mean. This estimate of the

population variance, computed from between the group means is often called the *between-groups variance*.

If the null hypothesis that all four samples were drawn from the same population is true (i.e., that diet does not affect cardiac output), the within-groups variance and between-groups variance are both estimates of the same population variance and so should be about equal. Therefore, we will compute the following ratio, called the F -test statistic:

$$F = \frac{\text{Population variance estimated from sample means}}{\text{Population variance estimated as average of sample variances}}$$

$$F = \frac{s_{\text{bet}}^2}{s_{\text{wit}}^2}$$

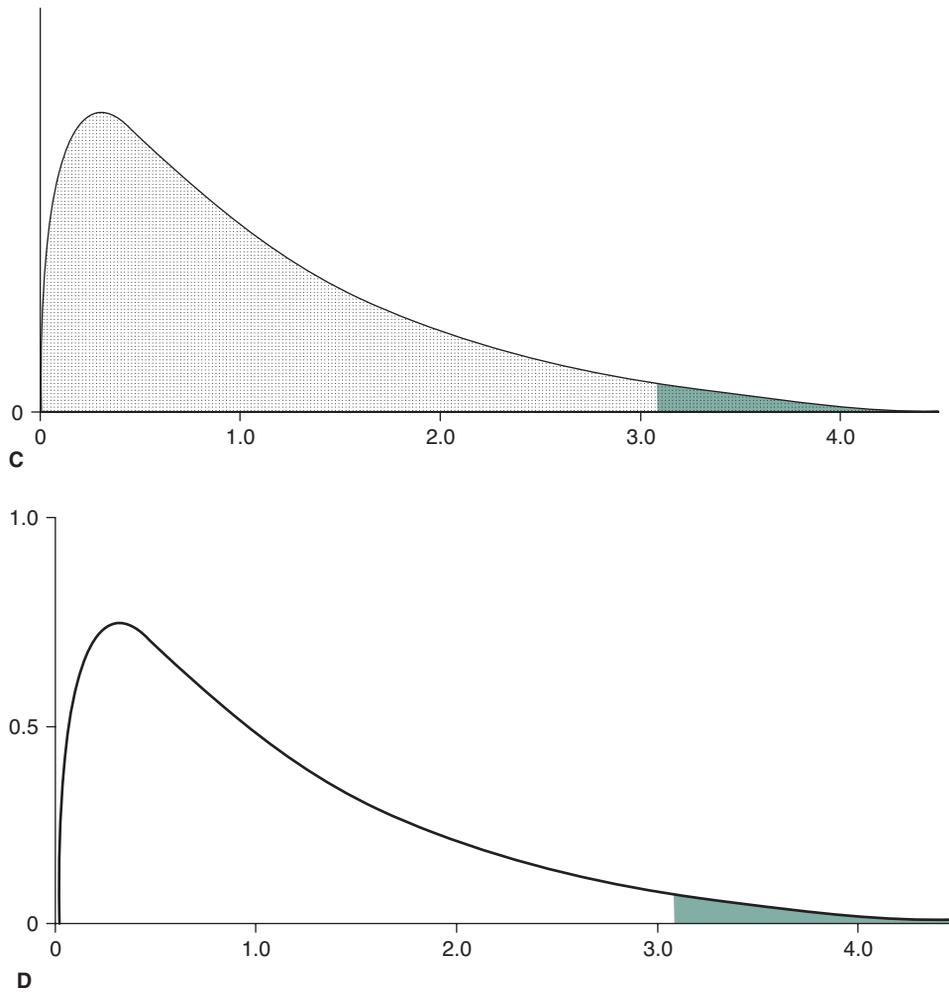


FIGURE 3-6. (Continued) **(C)** Results of computing the F ratio for all possible samples drawn from the original population. The 5% of most extreme F values are shown darker than the rest. **(D)** The F distribution one would obtain when sampling an infinite population. In this case, the cutoff value for considering F to be “big” is that value of F that subtends the upper 5% of the total area under the curve.

Since both the numerator and the denominator are estimates of the same population variance σ^2 , F should be about equal to 1, we conclude that the data in Figure 3-2 are not inconsistent with the hypothesis that diet does not affect cardiac output and we continue to accept that hypothesis.

Now we have a rule for deciding when to reject the null hypothesis that all the samples were drawn from the same population:

If F is a big number, the variability between the sample means is larger than expected from the variability within the samples, so reject the null hypothesis that all the samples were drawn from the same population.

This quantitative statement formalizes the qualitative logic we used when discussing Figures 3-2 to 3-4. The F associated with Figure 3-3 is 68.0, and that associated with Figure 3-4 is 24.5.

■ WHAT IS A “BIG” F ?

The exact value of F one computes depends on which individuals were drawn for the random samples. For example, Figure 3-5 shows yet another set of four samples of seven people drawn from the population of 200 people in Figure 3-1. In this example $F = 0.5$. Suppose we repeated our experiment 200 times on the same population. Each time we would draw four different samples of people and—even if the diet had no effect on cardiac output—get slightly different values for F due to random variation. Figure 3-6A shows the result of this procedure, with the resulting F s rounded to one decimal place and represented with a circle; the two dark circles represent the values of F computed from the data in Figures 3-2 and 3-5. The exact shape of the distribution of values of F depend on how many samples were drawn, the size of each sample, and the distribution of the population from which the samples were drawn.

As expected, most of the computed F s are around 1 (i.e., between 0 and 2), but a few are much larger. Thus, even though most experiments will produce relatively small values of F , it is possible that, by sheer bad luck, one could select random samples that are not good representatives of the whole population. The result is an occasional relatively large value for F even though the treatment had no effect. Figure 3-6B shows, however, that such values are unlikely. Only 5% of the 200 experiments (i.e., 10 experiments) produced F values equal to or greater than 3.0. We now have a tentative estimate of what to call a “big” value for F . Since F exceeded 3.0 only 10 out of 200 times *when all the samples were drawn from the same population*, we might decide that F is big when it exceeds 3.0 and reject the null hypothesis that all the samples were drawn from the same population (i.e., that the treatment had no effect). In deciding to reject the hypothesis of no effect when F is big, we accept the risk of erroneously rejecting this hypothesis 5% of the time because F will be 3.0 or greater about 5% of the time, even when the treatment does not alter mean response.

When we obtain such a “big” F , we reject the original null hypothesis that all the means are the same and report $P < .05$. $P < .05$ means that there is less than a 5% chance of getting a value of F as big or bigger than the computed value if the original hypothesis were true (i.e., diet did not affect cardiac output).

The critical value of F should be selected not on the basis of just 200 experiments but all 10^{42} possible experi-

ments. Suppose we did all 10^{42} experiments and computed the corresponding F values, then plotted the results such as we did for Figure 3-6B. Figure 3-6C shows the results with grains of sand to represent each observed F value. The darker sand indicates the biggest 5% of the F values. Notice how similar it is to Figure 3-6B. This similarity should not surprise you, since the results in Figure 3-6B are just a random sample of the population in Figure 3-6C. Finally, recall that everything so far has been based on an original population containing only 200 members. In reality, populations are usually much larger, so that there can be many more than 10^{42} possible values of F . Often, there are essentially an infinite number of possible experiments. In terms of Figure 3-6C, it is as if all the grains of sand melted together to yield the continuous line in Figure 3-6D.

Therefore, *areas under the curve* are analogous to the fractions of total number of circles or grains of sand in Figures 3-6B and 3-6C. Since the shaded region in Figure 3-6D represents 5% of the total area under the curve, it can be used to compute that the cutoff point for a “big” F with the number of samples and sample size in this study is 3.01. This and other cutoff values that correspond to $P < .05$ and $P < .01$ are listed in Table 3-1.

To construct these tables, mathematicians have assumed four things about the underlying population that must be at least approximately satisfied for the tables to be applicable to real data:

- *Each sample must be independent of the other samples.*
- *Each sample must be randomly selected from the population being studied.*
- *The populations from which the samples were drawn must be normally distributed.**
- *The variances of each population must be equal, even when the means are different, i.e., when the treatment has an effect.†*

When the data suggest that these assumptions do not apply, one ought not to use the procedure we just developed, the analysis of variance. Since there is one factor (the diet) that distinguishes the different experimental

*This is another reason parametric statistical methods require data from normally distributed populations.

†You can formally compare two variances with an F test; the numerator and denominator degrees of freedom are one less than the number of observations in the variance in the numerator and denominator that are being compared.

■ TABLE 3-1. Critical Values of F Corresponding to P < .05 (Lightface) and P < .01 (Boldface)

v _d	v _n																							
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	249	250	251	252	253	253	254	254	254
	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6082	6106	6142	6169	6208	6234	6261	6286	6302	6323	6334	6352	6361	6366
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44	19.45	19.46	19.47	19.47	19.48	19.48	19.49	19.49	19.50
	98.49	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.41	99.42	99.43	99.44	99.45	99.46	99.47	99.48	99.48	99.49	99.49	99.49	99.50	99.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.56	8.54	8.54	8.53
	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05	26.92	26.83	26.69	26.60	26.50	26.41	26.35	26.27	26.23	26.18	26.14	26.12
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.66	5.65	5.64	5.63
	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.54	14.45	14.37	14.24	14.15	14.02	13.93	13.83	13.74	13.69	13.61	13.57	13.52	13.48	13.46
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36
	16.26	13.27	12.06	11.39	10.97	10.67	10.45	10.29	10.15	10.05	9.96	9.89	9.77	9.68	9.55	9.47	9.38	9.29	9.24	9.17	9.13	9.07	9.04	9.02
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67
	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.60	7.52	7.39	7.31	7.23	7.14	7.09	7.02	6.99	6.94	6.90	6.88
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24	3.23
	12.25	9.55	8.45	7.85	7.46	7.19	7.00	6.84	6.71	6.62	6.54	6.47	6.35	6.27	6.15	6.07	5.98	5.90	5.85	5.78	5.75	5.70	5.67	5.65
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.93
	11.26	8.65	7.59	7.01	6.63	6.37	6.19	6.03	5.91	5.82	5.74	5.67	5.56	5.48	5.36	5.28	5.20	5.11	5.06	5.00	4.96	4.91	4.88	4.86
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72	2.71
	10.56	8.02	6.99	6.42	6.06	5.80	5.62	5.47	5.35	5.26	5.18	5.11	5.00	4.92	4.80	4.73	4.64	4.56	4.51	4.45	4.41	4.36	4.33	4.31
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55	2.54
	10.04	7.56	6.55	5.99	5.64	5.39	5.21	5.06	4.95	4.85	4.78	4.71	4.60	4.52	4.41	4.33	4.25	4.17	4.12	4.05	4.01	3.96	3.93	3.91
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.45	2.42	2.41	2.40
	9.65	7.20	6.22	5.67	5.32	5.07	4.88	4.74	4.63	4.54	4.46	4.40	4.29	4.21	4.10	4.02	3.94	3.86	3.80	3.74	3.70	3.66	3.62	3.60
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.35	2.32	2.31	2.30
	9.33	6.93	5.95	5.41	5.06	4.82	4.65	4.50	4.39	4.30	4.22	4.16	4.05	3.98	3.86	3.78	3.70	3.61	3.56	3.49	3.46	3.41	3.38	3.36
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.55	2.51	2.46	2.42	2.38	2.34	2.32	2.28	2.26	2.24	2.22	2.21
	9.07	6.70	5.74	5.20	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.85	3.78	3.67	3.59	3.51	3.42	3.37	3.30	3.27	3.21	3.18	3.16
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13
	8.86	6.51	5.56	5.03	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.70	3.62	3.51	3.43	3.34	3.26	3.21	3.14	3.11	3.06	3.02	3.00
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07
	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.56	3.48	3.36	3.29	3.20	3.12	3.07	3.00	2.97	2.92	2.89	2.87
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01
	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.61	3.55	3.45	3.37	3.25	3.18	3.10	3.01	2.96	2.98	2.86	2.80	2.77	2.75
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96
	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.45	3.35	3.27	3.16	3.08	3.00	2.92	2.86	2.79	2.76	2.70	2.67	2.65
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19	2.15	2.11	2.07	2.04	2.00	1.98	1.95	1.93	1.92
	8.28	6.01	5.09	4.58	4.25	4.01	3.85	3.71	3.60	3.51	3.44	3.37	3.27	3.19	3.07	3.00	2.91	2.83	2.78	2.71	2.68	2.62	2.59	2.57
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31	2.26	2.21	2.15	2.11	2.07	2.02	2.00	1.96	1.94	1.91	1.90	1.88
	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.19	3.12	3.00	2.92	2.84	2.76	2.70	2.63	2.60	2.54	2.51	2.49

v_n = degrees of freedom for numerator; v_d = degrees of freedom for denominator.

(continued)

■ TABLE 3-1. Critical Values of F Corresponding to $P < .05$ (Lightface) and $P < .01$ (Boldface) (Continued)

v_d	v_n																								
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞	
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.90	1.87	1.85	1.84	1.84
	8.10	5.85	4.94	4.43	4.10	3.87	3.71	3.56	3.45	3.37	3.30	3.23	3.13	3.05	2.94	2.86	2.77	2.69	2.63	2.56	2.53	2.47	2.44	2.42	2.42
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20	2.15	2.09	2.05	2.00	1.96	1.93	1.89	1.87	1.84	1.82	1.81	1.81
	8.02	5.78	4.87	4.37	4.04	3.81	3.65	3.51	3.40	3.31	3.24	3.17	3.07	2.99	2.88	2.80	2.72	2.63	2.58	2.51	2.47	2.42	2.38	2.36	2.36
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23	2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.80	1.78	1.78
	7.94	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.02	2.94	2.83	2.75	2.67	2.58	2.53	2.46	2.42	2.37	2.33	2.31	2.31
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20	2.14	2.10	2.04	2.00	1.96	1.91	1.88	1.84	1.82	1.79	1.77	1.76	1.76
	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	2.97	2.89	2.78	2.70	2.62	2.53	2.48	2.41	2.37	2.32	2.28	2.26	2.26
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13	2.09	2.02	1.98	1.94	1.89	1.86	1.82	1.80	1.76	1.74	1.73	1.73
	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.25	3.17	3.09	3.03	2.93	2.85	2.74	2.66	2.58	2.49	2.44	2.36	2.33	2.27	2.23	2.23	2.31
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16	2.11	2.06	2.00	1.96	1.92	1.87	1.84	1.80	1.77	1.74	1.72	1.71	1.71
	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.21	3.13	3.05	2.99	2.89	2.81	2.70	2.62	2.54	2.45	2.40	2.32	2.29	2.23	2.19	2.17	2.17
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.10	2.05	1.99	1.95	1.90	1.85	1.82	1.78	1.76	1.72	1.70	1.69	1.69
	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.17	3.09	3.02	2.96	2.86	2.77	2.66	2.58	2.50	2.41	2.36	2.28	2.25	2.19	2.15	2.13	2.13
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13	2.08	2.03	1.97	1.93	1.88	1.84	1.80	1.76	1.74	1.71	1.68	1.67	1.67
	7.68	5.49	4.60	4.11	3.79	3.56	3.39	3.26	3.14	3.06	2.98	2.93	2.83	2.74	2.63	2.55	2.47	2.38	2.33	2.25	2.21	2.16	2.12	2.10	2.10
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12	2.06	2.02	1.96	1.91	1.87	1.81	1.78	1.75	1.72	1.69	1.67	1.65	1.65
	7.64	5.45	4.57	4.07	3.76	3.53	3.36	3.23	3.11	3.03	2.95	2.90	2.80	2.71	2.60	2.52	2.44	2.35	2.30	2.22	2.18	2.13	2.09	2.06	2.06
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05	2.00	1.94	1.90	1.85	1.80	1.77	1.73	1.71	1.68	1.65	1.64	1.64
	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.08	3.00	2.92	2.87	2.77	2.68	2.57	2.49	2.41	2.32	2.27	2.19	2.15	2.10	2.06	2.03	2.03
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09	2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.64	1.62	1.62
	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.06	2.98	2.90	2.84	2.74	2.66	2.55	2.47	2.38	2.29	2.24	2.16	2.13	2.07	2.03	2.01	2.01
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07	2.02	1.97	1.91	1.86	1.82	1.76	1.74	1.69	1.67	1.64	1.61	1.59	1.59
	7.50	5.34	4.46	3.97	3.66	3.42	3.25	3.12	3.01	2.94	2.86	2.80	2.70	2.62	2.51	2.42	2.34	2.25	2.20	2.12	2.08	2.02	1.98	1.96	1.96
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05	2.00	1.95	1.89	1.84	1.80	1.74	1.71	1.67	1.64	1.61	1.59	1.57	1.57
	7.44	5.29	4.42	3.93	3.61	3.38	3.21	3.08	2.97	2.89	2.82	2.76	2.66	2.58	2.47	2.38	2.30	2.21	2.15	2.08	2.04	1.98	1.94	1.91	1.91
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03	1.98	1.93	1.87	1.82	1.78	1.72	1.69	1.65	1.62	1.59	1.56	1.55	1.55
	7.39	5.25	4.38	3.89	3.58	3.35	3.18	3.04	2.94	2.86	2.78	2.72	2.62	2.54	2.43	2.35	2.26	2.17	2.12	2.04	2.00	1.94	1.90	1.87	1.87

38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.96	1.92	1.85	1.80	1.76	1.71	1.67	1.63	1.60	1.57	1.54	1.53
	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.91	2.82	2.75	2.69	2.59	2.51	2.40	2.32	2.22	2.14	2.08	2.00	1.97	1.90	1.86	1.84
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00	1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51
	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.88	2.80	2.73	2.66	2.56	2.49	2.37	2.29	2.20	2.11	2.05	1.97	1.94	1.88	1.84	1.81
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99	1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.60	1.57	1.54	1.51	1.49
	7.27	5.15	4.29	3.80	3.49	3.26	3.10	2.96	2.86	2.77	2.70	2.64	2.54	2.46	2.35	2.26	2.17	2.08	2.02	1.94	1.91	1.85	1.80	1.78
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.50	1.48
	7.24	5.12	4.26	3.78	3.46	3.24	3.07	2.94	2.84	2.75	2.68	2.62	2.52	2.44	2.32	2.24	2.15	2.06	2.00	1.92	1.88	1.82	1.78	1.75
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97	1.91	1.87	1.80	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46
	7.21	5.10	4.24	3.76	3.44	3.22	3.05	2.92	2.82	2.73	2.66	2.60	2.50	2.42	2.30	2.22	2.13	2.04	1.98	1.90	1.86	1.80	1.76	1.72
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96	1.90	1.86	1.79	1.74	1.70	1.64	1.61	1.56	1.53	1.50	1.47	1.45
	7.19	5.08	4.22	3.74	3.42	3.20	3.04	2.90	2.80	2.71	2.64	2.58	2.48	2.40	2.28	2.20	2.11	2.02	1.96	1.88	1.84	1.78	1.73	1.70
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95	1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44
	7.17	5.06	4.20	3.72	3.41	3.18	3.02	2.88	2.78	2.70	2.62	2.56	2.46	2.39	2.26	2.18	2.10	2.00	1.94	1.86	1.82	1.76	1.71	1.68
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.86	1.81	1.75	1.70	1.65	1.59	1.56	1.50	1.48	1.44	1.41	1.39
	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.40	2.32	2.20	2.12	2.03	1.93	1.87	1.79	1.74	1.68	1.63	1.60
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89	1.84	1.79	1.72	1.67	1.62	1.56	1.53	1.47	1.45	1.40	1.37	1.35
	7.01	4.92	4.08	3.60	3.29	3.07	2.91	2.77	2.67	2.59	2.51	2.45	2.35	2.28	2.15	2.07	1.98	1.88	1.82	1.74	1.69	1.62	1.56	1.53
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88	1.82	1.77	1.70	1.65	1.60	1.54	1.51	1.45	1.42	1.38	1.35	1.32
	6.96	4.88	4.04	3.56	3.25	3.04	2.87	2.74	2.64	2.55	2.48	2.41	2.32	2.24	2.11	2.03	1.94	1.84	1.78	1.70	1.65	1.57	1.52	1.49
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85	1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.30	1.28
	6.90	4.82	3.98	3.51	3.20	2.99	2.82	2.69	2.59	2.51	2.43	2.36	2.26	2.19	2.06	1.98	1.89	1.79	1.73	1.64	1.59	1.51	1.46	1.43
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.87	1.84	1.78	1.73	1.66	1.61	1.56	1.50	1.46	1.39	1.37	1.32	1.28	1.25
	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.40	2.34	2.23	2.15	2.03	1.95	1.86	1.76	1.70	1.61	1.56	1.48	1.42	1.38
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75	1.69	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.00
	6.63	4.60	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.24	2.18	2.07	1.99	1.87	1.79	1.69	1.59	1.52	1.41	1.36	1.25	1.15	1.00

v_n = degrees of freedom for numerator; v_d = degrees of freedom for denominator.

Reproduced from Snedecor GW, Cochran WG. *Statistical Methods*, 8th ed. Copyright © 1989. Reproduced with the permission of John Wiley & Sons, Inc.

groups, this is known as a *single factor or one way analysis of variance*. Other forms of analysis of variance (not discussed here) can be used to analyze experiments in which there is more than one experimental factor.

Since the distribution of possible F values depends on the size of each sample and number of samples under consideration, so does the exact value of F which corresponds to the 5% cutoff point. For example, in our diet study, the number of samples was 4 and the size of each sample was 7. This dependence enters into the mathematical formulas used to determine the value at which F gets “big” as two parameters known as *degree-of-freedom* parameters, often denoted as ν (Greek “nu”). For this analysis, the between-groups degrees of freedom (also called the numerator degrees of freedom because the between-groups variance is in the numerator of F) is defined to be the number of samples m minus 1, or $\nu_n = m - 1$. The within-groups (or denominator) degrees of freedom is defined to be the number of samples times 1 less than the size of each sample, $\nu_d = m(n - 1)$. For our diet example, the numerator degrees of freedom are $4 - 1 = 3$, and the denominator degrees of freedom are $4(7 - 1) = 24$. Degrees of freedom often confuse and mystify people who are trying to work with statistics. They simply represent the way *number of samples* and *sample size* enter the mathematical formulas used to construct all statistical tables.

■ CELL PHONES AND SPERM

We now have the tools needed to form conclusions using statistical reasoning. We will examine examples, all based on results published in the medical literature. I have exercised some literary license with these examples for two reasons: (1) Medical and scientific authors usually summarize their raw data with descriptive statistics (like those developed in Chapter 2) rather than including the raw data. As a result, the “data from the literature” shown in this chapter—and the rest of the book—are usually my guess at what the raw data probably looked like based on the descriptive statistics in the original article.* (2) The analysis of variance as we developed it requires that each sample contain the same number of members. This is often not the case in reality, so I adjusted the sample sizes in the original studies to meet this restriction. We later generalize our statistical methods

to handle experiments with different numbers of individuals in each sample or treatment group.

An Early Study

Cell phones have become ubiquitous all over the world, exposing people to radiofrequency radiation. The phones are almost always held close to the body, exposing potentially sensitive tissues to relatively high levels of this radiation. Based on an earlier small study suggesting declining levels of rapidly moving spermatozoa in a small number of cell phone users, Imre Fejes and colleagues[†] obtained semen samples from two groups of young men 30.8 ± 4.4 (standard deviation, range 17 to 41) years old who were patients at an infertility clinic: A low use group who used cell phones less than 15 minutes/day and a high use group who used their phones for over 60 minutes/day. (They collected their data between November 2002 and March 2004, when cell phone use was probably lower than in subsequent years.)

Because this is an *observational* study, Fejes and colleagues tried to minimize the effects of confounding variables by excluded men with conditions that could affect sperm function, including smoking (but only more than 10 cigarettes/day), regular alcohol consumption, drug abuse, illness, reproductive or testicular abnormalities, abnormal hormone levels, or genital tract infection.

Figure 3-7 shows the percentage of rapidly moving sperm for each individual. Figure 3-7 shows that for the 61 men in the low use group the mean percentage of rapidly mobile sperm was 49% and for the 61 men in the high use group it was 41%. The standard deviations were 21% and 22%, respectively.

How consistent are these data with the null hypothesis that rapid sperm mobility does not differ in men who use their cell phones less than 15 minutes/day compared to men who use them more than 60 minutes/day? In other words, how likely are the differences in the two samples of men depicted in Figure 3-7 to be due to random sampling rather than the difference in cell phone usage?

To answer this question, we perform an analysis of variance.

We begin by estimating the within-groups variance by averaging the variances of the two samples of men. Since this estimate of the underlying population variance is

*Since authors often failed to include a complete set of descriptive statistics, I had to simulate them from the results of their hypothesis tests.

[†]Fejes I, Závacki Z, Szöllosi J, Koloszar S, Daru J, Kovács L, Pál A. Is there a relationship between cell phone use and semen quality? *Arch Androl.* 2005;51:385–393.

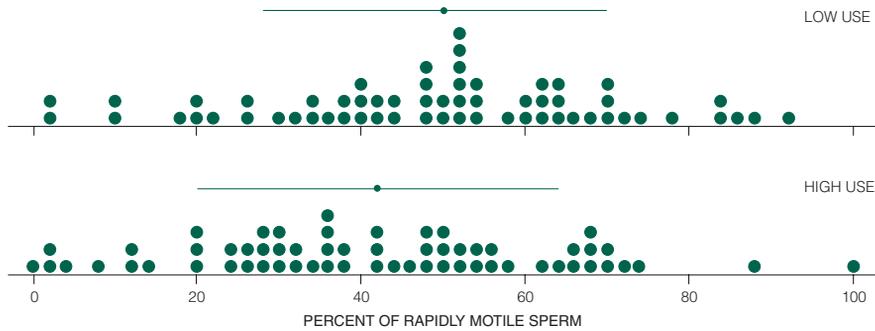


FIGURE 3-7. Results of a study comparing fraction of sperm with rapid motility associated with low and high intensity cell phone use. The fraction of rapid motility for each man’s sperm is indicated by a circle at the appropriate fraction of rapidly motile sperm. (Such a plot is called a *histogram* of the data. The mean fraction of rapidly motile sperm in men with lower cell phone use (49%) is higher than for the men with high cell phone use (41%). The statistical question is whether this difference is due to random sampling or due to an actual effect of cell phone use. The horizontal lines show one standard deviation on either side of the means (21% and 22%, respectively).

computed from the variances of the separate samples, does not depend on whether the means are different or not:

$$s_{\text{wit}}^2 = \frac{1}{2}(s_{\text{low}}^2 + s_{\text{high}}^2)$$

$$s_{\text{wit}}^2 = \frac{1}{2}(21^2 + 22^2) = 462.5\%^2$$

We then go on to compute the between-groups variance assuming that the null hypothesis is correct and the differences between the observed means is due to random sampling variation, not any systematic effects of the level of cell phone usage. The first step is to estimate the standard error of the mean by computing the standard deviation of the two sample means. The mean of the two sample means is

$$\bar{X} = \frac{1}{2}(\bar{X}_{\text{low}} + \bar{X}_{\text{high}})$$

$$\bar{X} = \frac{1}{2}(49 + 41) = 45\%$$

Therefore, the standard deviation of the sample means is

$$s_{\bar{X}} = \sqrt{\frac{(\bar{X}_{\text{low}} - \bar{X})^2 + (\bar{X}_{\text{high}} - \bar{X})^2}{m - 1}}$$

$$s_{\bar{X}} = \sqrt{\frac{(49 - 45)^2 + (41 - 45)^2}{2 - 1}} = 5.66\%$$

Since the sample size n is 61, the estimation of the population variance from between the groups is

$$s_{\text{bet}}^2 = n s_{\bar{X}}^2 = 61(5.66^2) = 1952\%^2$$

Finally, the ratio of these two different estimates of the underlying population variance (assuming that the null hypothesis is correct) is

$$F = \frac{s_{\text{bet}}^2}{s_{\text{wit}}^2} = \frac{1952}{462.5} = 4.22$$

The degrees of freedom for the numerator are the number of sample groups minus 1, so $v_n = 2 - 1 = 1$, and the degrees of freedom for the denominator are the number of groups times one less than the sample size of each group, so $v_d = 2(61 - 1) = 120$. Look in the column headed 1 and the row headed 120 in Table 3-1. The resulting entry indicates that there is less than a 5% chance of F exceeding 3.92 by chance if, in fact, the null hypothesis that cell phone use did not affect mean sperm mobility was true. We therefore concluded that the value of F associated with our observations is “big” and reject the null hypothesis that there is no difference in sperm mobility in the two groups of men ($P < .05$) shown in Figure 3-7.

By rejecting the null hypothesis of no difference, we conclude that there are different levels of rapid sperm motility associated with higher levels of cell phone use, with the heavier users having fewer sperm with rapid motility.

■ TABLE 3-2. Sperm Motility (%)

Observed Cell Phone Use	Number of Subjects (<i>n</i>)	Mean	Standard Deviation
Control (no cell phone use)	40	68	6
Low use (<2 h/d)	40	65	8
Medium use (2 to 4 h/d)	40	54	11
High use (>4 h/d)	40	45	16

A Better Control Group

One problem with Fejes and colleagues' study is that it did not include a completely unexposed (clean) *control group* of men who did not use cell phones at all. Ashok Agarwal and colleagues* avoided this problem when they did a similar observational study of men aged 32 ± 6 (standard deviation) years old who were attending their infertility clinic. They also had stricter exclusion criteria than the earlier study. They excluded anyone with a history of smoking or other tobacco use, alcohol use, diabetes, high blood pressure, or other diseases. Unlike the study just discussed, they measured the fraction of sperm that exhibited any motility (as opposed to rapid motility). Table 3-2 shows the data. As before, the question is whether cell phone use is associated with changes in sperm motility.

To answer this question, we perform an analysis of variance to test the null hypothesis that the level of cell phone use is not associated with differences in sperm motility among the four groups.

As before, we begin by estimating the within-groups variance by averaging the variances of the four samples of men:

$$s_{\text{wit}}^2 = \frac{1}{4}(s_{\text{control}}^2 + s_{\text{low}}^2 + s_{\text{medium}}^2 + s_{\text{high}}^2)$$

$$s_{\text{wit}}^2 = \frac{1}{4}(6^2 + 8^2 + 11^2 + 16^2) = 119.3\%$$

To compute the between-groups variance estimate on the assumption that the null hypothesis is true, so that all the observed means are simply estimating the same

underlying (constant) population mean in sperm motility. The mean of the four sample means is

$$\bar{X} = \frac{1}{4}(\bar{X}_{\text{control}} + \bar{X}_{\text{low}} + \bar{X}_{\text{medium}} + \bar{X}_{\text{high}})$$

$$\bar{X} = \frac{1}{4}(68 + 65 + 54 + 45) = 58\%$$

The standard deviation of the $m = 4$ sample means is

$$s_{\bar{X}} = \sqrt{\frac{(\bar{X}_{\text{control}} - \bar{X})^2 + (\bar{X}_{\text{low}} - \bar{X})^2 + (\bar{X}_{\text{medium}} - \bar{X})^2 + (\bar{X}_{\text{high}} - \bar{X})^2}{m - 1}}$$

$$s_{\bar{X}} = \sqrt{\frac{(68 - 58)^2 + (65 - 58)^2 + (54 - 58)^2 + (45 - 58)^2}{4 - 1}} = 10.44\%$$

Since the sample size n is 40, the estimation of the population variance from between the groups is

$$s_{\text{bet}}^2 = ns_{\bar{X}}^2 = 40(10.44^2) = 4360\%$$

To test whether these two estimates of the underlying population variance are consistent with each other under the assumption that the null hypothesis is correct (i.e., that sperm motility is not detectably different between the different cell phone sample groups), we compute

$$F = \frac{s_{\text{bet}}^2}{s_{\text{wit}}^2} = \frac{4360}{119.3} = 36.53$$

The numerator degrees of freedom are $v_n = m - 1 = 4 - 1 = 3$ and denominator degrees of freedom are $v_d = m(n - 1) = 4(40 - 1) = 156$. Table 3-1 does not have an entry for $v_d = 156$, but the critical value for $P < .01$ will be between 3.95, the value corresponding to 120 denominator degrees of freedom and 3.78, the value for an infinite number of degrees of freedom. The value of F associated with the data

*Agarwal A, Deepinder F, Sharma RK, Ranga G, Li J. Effect of cell phone usage on semen analysis in men attending infertility clinic: an observational study. *Fertil Steril*. 2008;89:124–128.

exceeds this range, so, as before, we reject the null hypothesis that sperm motility is not related to cell phone use.

An important question, however, remains: Which of the four sample groups differed from the others? Is any cell phone use associated with a reduction in sperm motility or is there evidence of the threshold for an effect? Does the effect increase with how much a man used the cell phone? We will have to defer answering these questions until we develop additional statistical tools, the *t* test and the associated *multiple comparison procedures*, in Chapter 4.

An Experimental Study

As discussed in Chapter 2, the strength of the conclusions about *cause and effect* are always limited in observational studies because one can never totally exclude the possibility that there is some unobserved confounding variable that is influencing the results which makes it appear that there is a relationship between the conditions being studied and the outcome variable when no such relationship exists. One can draw much stronger conclusions in an *experiment* in which the investigator randomly assigns experimental subjects to the different treatment conditions, which he or she controls. In such a case, the only systematic difference between the different experimental groups is the presence or absence of the condition being studied.

Motivated in part by the two observational studies just discussed, Nader Salama and colleagues* conducted an experiment in which they exposed adult male rabbits to cell phone radiation for 8 hours a day for 12 weeks. The rabbits were exposed to the cell phones by being housed during the 8 hour exposure time in a specially designed cage which kept the rabbits' testes positioned over the cell phone during the whole time. (They were housed in larger cages the rest of the time.) Because being in such a constrained environment might prove stressful to the rabbit, which could, in turn, affect sperm production and function—a confounding variable—Salama and coworkers had two control groups: a *stress control*, where the rabbit was housed in the same specially designed cage as the cell phone-exposed rabbits, but without the cell phone, and an *ordinary control* in which the rabbit was housed in their usual cage all the time. They studied 24 rabbits, randomizing 8 to each experimental condition.

The data and associated analysis of variance are presented in Box 3-1. Notice that while the average values are

roughly comparable to the values observed in the two human studies (compare the data in Box 3-1 with that in Fig. 3-7 and Table 3-2), the standard deviations are smaller in the experimental study using rabbits. This difference is probably because all the rabbits were the same strain (New Zealand White rabbits) and same age, whereas the human observational studies involved men with a range of ages and other differences. Indeed, one benefit of doing such an experimental study is to obtain this standardization and the associated reduction in between-individual random differences. At the same time, the fact that the observational studies used real people makes the results more relevant in the real world. This tradeoff between a tightly controlled subject population and generality in the real world is a common tension in most biomedical and clinical research.

There is a statistically significant difference between the treatment groups ($P < .01$). (Resolving whether all three groups are different from each other or whether there is some subgrouping of responses will have to wait until we develop procedures for multiple comparison testing in Chapter 4.) Because the rabbits were the same, randomly assigned to treatment groups, and except for the presence of the cell phone and cage situation, treated identically, we can confidently conclude that the experimental condition affected the sperm motility.

While one always has to be cognizant of cross-species comparisons, the fact that two independent observational studies of men done under different conditions and an experimental study—albeit using rabbits rather than people—substantially strengthens a conclusion that the cell phone exposure is *causing* the reduction in sperm function. Combining different sources of information with different strengths and weaknesses to identify points of concordance and disagreement is the key to drawing conclusions about causality, particularly when a substantial part of the evidence comes from observational studies.

■ UNEQUAL SAMPLE SIZE

We have developed analysis of variance for the case of equal sample sizes because doing so allowed us to develop and present the formulas to compute *F* in a way that makes it easy to understand the underlying concepts. It is also possible to do an analysis of variance when the sample sizes are not the same, although the formulas and notation are much more opaque in terms of what they mean. Appendix A gives these computational formulas and Box 3-2 illustrates how to use them.

*Salama N, Kishimoto T, Kanayama H. Effects of exposure to a mobile phone on testicular function and structure in adult rabbit. *Int J Androl.* 2010;33:88–94.

BOX 3-1 • Effect of Cell Phone Radiation on Rabbit Sperm Motility*Rabbit Sperm Motility after 12 Weeks (%)*

Experimental Condition	Sample Size (<i>n</i>)	Mean	Standard Deviation
Ordinary control	8	72	3.2
Stress control	8	61	2.2
Cell phone exposure	8	50	2.5

The within-groups variance is computed by averaging the three sample variances:

$$s_{\text{wit}}^2 = \frac{1}{3}(s_{\text{ordinary}}^2 + s_{\text{stress}}^2 + s_{\text{phone}}^2)$$

$$s_{\text{wit}}^2 = \frac{1}{3}(3.2^2 + 2.2^2 + 2.5^2) = 7.11\%^2$$

The between-groups variance estimate begins with computing the mean in sperm motility in the three samples,

$$\bar{X} = \frac{1}{3}(\bar{X}_{\text{ordinary}} + \bar{X}_{\text{stress}} + \bar{X}_{\text{phone}})$$

$$\bar{X} = \frac{1}{3}(72 + 61 + 50) = 61\%$$

which is then used to compute the standard deviation of the $m = 3$ sample means:

$$s_{\bar{X}} = \sqrt{\frac{(\bar{X}_{\text{ordinary}} - \bar{X})^2 + (\bar{X}_{\text{stress}} - \bar{X})^2 + (\bar{X}_{\text{phone}} - \bar{X})^2}{m - 1}}$$

$$s_{\bar{X}} = \sqrt{\frac{(72 - 61)^2 + (61 - 61)^2 + (50 - 61)^2}{3 - 1}} = 11.0\%$$

Since the sample size of each group, n , is 8, the between-groups variance estimate is

$$s_{\text{bet}}^2 = ns_{\bar{X}}^2 = 8(11.0\%)^2 = 968\%^2$$

So

$$F = \frac{s_{\text{bet}}^2}{s_{\text{wit}}^2} = \frac{968}{7.11} = 136.15$$

We compare this to the critical value of F for $v_n = m - 1 = 3 - 1 = 2$ numerator and $v_d = m(n - 1) = 3(8 - 1) = 21$ denominator degrees of freedom. From Table 3-1, the critical value for $P < .01$ is 5.78; the value associated with our data exceeds this value, so we conclude that there is a statistically significant difference between the three treatment groups.

BOX 3-2 • Effect of Seeing Smoking in Movies on Smokers' Brains

Seeing onscreen smoking in movies is a major stimulus for youth and young adults to start smoking. It also stimulates smoking behavior among people who are already smokers. Smoking is a highly practiced motor skill that often occurs automatically without conscious awareness. There are certain areas of the brain (called the frontoparietal network) that is activated when people observe, plan, or imitate actions. To investigate whether this action observation would be preferentially activated in smokers when watching smoking in a movie, Dylan Wagner and colleagues* did functional MRI (magnetic resonance imaging) on the brains of 17 smokers and 15 nonsmokers and measured the extent to which blood flow increased in brain regions in the frontoparietal network. (The units are arbitrary.)

Here are the data:

	Sample Size (n)	Mean	Standard Deviation
Smokers	17	.65	.20
Nonsmokers	15	.22	.15

To test the null hypothesis that the levels of blood flow are no different among smokers and nonsmokers, we compute an analysis of variance using the formulae in Appendix A.

We first compute the total sample size by adding up the sample sizes:

$$N = \sum n_t = n_{\text{smokers}} + n_{\text{nonsmokers}} = 17 + 15 = 32$$

Next, we estimate the within-groups variance estimate based on a weighted average of the variances within the two sample groups:

$$SS_{\text{wit}} = \sum (n_t - 1)s_t^2 = (17 - 1) \times .20^2 + (15 - 1) \times .15^2 = .955$$

The degrees of freedom associated with the within-groups variance estimate is

$$v_{\text{wit}} = DF_{\text{wit}} = N - k = 32 - 2 = 30$$

So

$$s_{\text{wit}}^2 = \frac{SS_{\text{wit}}}{DF_{\text{wit}}} = \frac{.955}{30} = .0318$$

The formula for the between-groups variance estimate is

$$SS_{\text{bet}} = \sum n_t \bar{X}_t^2 - \frac{(\sum n_t \bar{X}_t)^2}{N}$$

$$SS_{\text{bet}} = (17 \times .65^2 + 15 \times .22^2) - \frac{(17 \times .65 + 15 \times .22)^2}{32} = 1.473$$

The degrees of freedom associated with the between-groups variance estimate is

$$v_{\text{bet}} = DF_{\text{bet}} = k - 1 = 2 - 1 = 1$$

So

$$s_{\text{bet}}^2 = \frac{SS_{\text{bet}}}{DF_{\text{bet}}} = \frac{1.473}{1} = 1.473$$

As in the equal sample size case,

$$F = \frac{s_{\text{bet}}^2}{s_{\text{wit}}^2} = \frac{1.473}{.0318} = 46.32$$

From Table 3-1, this value of F exceeds the critical value of 7.56 that defines the largest 1% of values under the null hypothesis with 1 numerator and 30 denominator degrees of freedom, so we reject the null hypothesis of no difference, and conclude that seeing images of smoking in movies stimulates the brain regions associated with repetitive actions in smokers more than nonsmokers ($P < .01$).

*Wagner DD, Dal Cin S, Sargent JD, Kelley WM, Heatherton TF. Spontaneous action representation in smokers when watching movie characters smoke. *J Neurosci*. 2001;31:894-898.

■ **TABLE 3-3. Experimental Design for One Way Analysis of Variance**

Diet			
Control	Spaghetti	Steak	Fruit and Nuts
<i>data</i>	<i>data</i>	<i>data</i>	<i>data</i>

■ TWO WAY ANALYSIS OF VARIANCE

The analysis of variance we have been discussing in this chapter is more precisely called *one way* or *single factor* analysis of variance, because the different treatment groups are defined by one factor (such as diet or level of cell phone use). It turns out that this one way analysis of variance is just the simplest case of much more general analysis of variance, in which it is possible to consider the effects of two (or more) factors acting simultaneously.

To illustrate the next level of complexity in experimental design, let us return to the diet example from the beginning of this chapter. In the original example, we evaluated the effects of a single *factor*, diet, on cardiac output of people. Table 3-3 shows the layout for the data from this study. A more sophisticated design would be to simultaneously consider the effects of diet and gender on cardiac output, using a *two way* (or *two factor*) design in Table 3-4. Based on these data, we could use a generalization of the analysis of variance presented in this chapter to test three null hypotheses using the resulting data:

1. Diet has no effect on cardiac output, controlling for gender.
2. Gender has no effect on cardiac output, controlling for diet.
3. The effect of diet on cardiac output is the same regardless of gender and vice versa.

The third null hypothesis states that there is no *interaction* between the two *main effects*, diet and gender. A significant interaction would mean that the effects of diet are different for different genders.

While we will not go into the details of how to compute and interpret two way (and higher order) analyses of variance,* the overall principles are the same as those discussed in this chapter.

*See Glantz S, Slinker B. *Primer of Applied Regression and Analysis of Variance*. 2nd ed. New York: McGraw-Hill; 2001 for details on how to analyze two way and higher-order analyses of variance.

We now turn our attention to developing the *t* test and adapting it to do multiple comparisons between pairs of means following a significant analysis of variance.

■ PROBLEMS

3-1 In order to study the cellular changes in people with tendencies to develop diabetes, Kitt Petersen and her colleagues[†] studied the ability of muscle cells in normal children and insulin-resistant children to convert glucose into adenosine triphosphate (ATP), the “energy molecule” muscle cells produce to power contraction. The body produces insulin to permit cells to process glucose, and muscle cells of insulin-resistant people do not respond normally to process glucose. They measured the amount of ATP produced per gram of muscle tissue after giving the study participants a dose of glucose. Persons in the control group produced 7.3 $\mu\text{mol/g}$ of muscle/min of ATP (standard deviation 2.3 $\mu\text{mol/g}$ of muscle/min) and insulin-resistant persons produced 5.0 $\mu\text{mol/g}$ of muscle/min (standard deviation 1.9 $\mu\text{mol/g}$ of muscle/min). There were 15 children in each test group. Is there a difference in the mean rate of ATP production in these two groups of people?

3-2 It was once generally believed that infrequent and short-term exposure to pollutants in tobacco, such as carbon monoxide, nicotine, benzo[*a*]pyrene, and oxides of nitrogen, will not permanently alter lung function in healthy adult nonsmokers. To investigate this hypothesis, James White and Herman Froeb[‡] measured lung function in cigarette smokers and nonsmokers during a “physical fitness profile” at the University of California, San Diego. They measured how rapidly a person could force air from

[†]Petersen K, et al. Impaired mitochondrial activity in the insulin-resistant offspring of patients with type 2 diabetes. *N Engl J Med*. 2004; 350:664–671.

[‡]White J, Froeb H. Small-airways dysfunction in nonsmokers chronically exposed to tobacco smoke. *N Engl J Med*. 1980;302:720–723.

■ **TABLE 3-4. Experimental Design for Two Way Analysis of Variance**

Gender	Diet			
	Control	Spaghetti	Steak	Fruit and Nuts
Male	<i>data</i>	<i>data</i>	<i>data</i>	<i>data</i>
Female	<i>data</i>	<i>data</i>	<i>data</i>	<i>data</i>

■ **TABLE 3-5. Mean Forced Midexpiratory Flow (L/s)**

Group	Sample Size (<i>n</i>)	Mean	Standard Deviation
Nonsmokers			
Worked in smoke-free environment	200	3.17	0.74
Worked in smoky environment	200	2.72	0.71
Light smokers	200	2.63	0.73
Moderate smokers	200	2.29	0.70
Heavy smokers	200	2.12	0.72

the lungs (mean forced midexpiratory flow). Reduced forced midexpiratory flow is associated with small-airways disease of the lungs. Table 3.5 shows the data for the women that White and Froeb tested. Is there evidence that the presence of small-airways disease, as measured by this test, is different among the different experimental groups?

3-3 The stair climb power test is a functional test used among older people to measure leg muscle power. To assess whether this test could be used to assess leg muscle power in people with chronic obstructive pulmonary disease (COPD) Marc Roig and colleagues* measured the power delivered by people with mild-to-severe COPD with age and sex matched controls with no disease. Subjects were told to climb 10 stairs as quickly as they could and the power computed as the vertical velocity (the gain in height of the 10 stairs divided by the length of time it took the subject to climb the stairs) times the subject's weight. The 21 people in the control group developed a mean of 378 watts (standard deviation 121 watts) and the 21 people with COPD developed 266 watts (standard deviation 81 watts). Test the hypothesis that there is no difference in the amount of power these two groups of people developed.

*Roig M, et al. Associations of the Stair Climb Power Test with muscle strength and functional performance in people with chronic obstructive pulmonary disease: a cross-sectional study. *Phys Ther.* 2010;90:1774–1782.

3-4 In the study of cell phone use and sperm function, the investigators also measured sperm viability for the different categories of cell phone users. Is there a difference in viability among these groups? (See Table 3-6.)

3-5 Men and women differ in risk of spinal fracture. Men are at increased risk for all types of bone fractures until approximately 45 years of age, an effect probably due to the higher overall trauma rate in men during this time. However, after age 45, women are at increased risk for spinal fracture, most likely due to age-related increases in osteoporosis, a disease characterized by decreased bone density. S. Kudlacek and colleagues† wanted to investigate the relationship between gender and bone density in a group of older adults who have had a vertebral bone fracture. Their data are presented in Table 3-7. Are there differences in vertebral bone density between similarly aged men and women who have had a vertebral bone fracture?

3-6 Burnout is a term that loosely describes a condition of fatigue, frustration, and anger manifested as a lack of

†Kudlacek S, et al. Gender differences in fracture risk and bone mineral density. *Maturitas*, 2000;36:173–180.

■ TABLE 3-6. Sperm Viability (%)

Observed Cell Phone Use	Sample Size (n)	Mean	Standard Deviation
Control (no cell phone use)	40	72	7
Low use (<2 h/d)	40	68	9
Medium use (2 to 4 h/d)	40	58	11
High use (>4 h/d)	40	47	17

enthusiasm for and feeling of entrapment in one's job. This situation can arise when treating people who have serious diseases. In recent years, AIDS has joined the list of diseases that may have a negative impact on professionals serving people suffering from this disease. To investigate whether there were differences in burnout associated with caring for people who have AIDS compared with other people who have serious diseases, J. López-Castillo and coworkers* administered the Maslach Burnout Inventory questionnaire to health professionals working in four clinical units: infectious disease, hemophilia, oncology, and internal medicine in Spain (see Table 3-8). (Ninety percent of the people in the infectious disease and 60% of the people in the hemophilia unit were HIV positive.) Are there differences in burnout scores between health professionals working in these different units?

3-7 High doses of estrogen interfere with male fertility in many animals, including mice. However, there may be significant differences in the response to estrogen in different mouse strains. To compare estrogen responsiveness in different strains of mice, Spearow and colleagues† implanted

*López-Castillo J, et al. Emotional distress and occupational burnout in health care professionals serving HIV-infected patients: a comparison with oncology and internal medicine services. *Psychother Psychosom.* 1999;68:348–356.

†Spearow JL, et al. Genetic variation in susceptibility to endocrine disruption by estrogen in mice. *Science.* 1999;285:1259–1261.

capsules containing 1 µg of estrogen into four different strains of juvenile male mice. After 20 days, they measured their testicular weight, shown in Table 3-9. Is there sufficient evidence to conclude that any of these strains differ in response to estrogen? (The formulas for analysis of variance with unequal sample sizes are in Appendix A.)

3-8 Several studies suggest that schizophrenic patients have lower IQ scores measured before the onset of schizophrenia (premorbid IQ) than would be expected based on family and environmental variables. These deficits can be detected during childhood and increase with age. Catherine Gilvarry and colleagues‡ investigated whether this was also the case with patients diagnosed with affective psychosis, which encompasses schizoaffective disorder, mania, and major depression. In addition, they also wanted to assess whether any IQ deficits could be detected in first-degree relatives (parents, siblings, and children) of patients with affective psychosis. They administered the National Adult Reading Test (NART), which is an indicator of premorbid IQ, to a set of patients with affective psychosis, their first-degree relatives, and a group of normal subjects without any psychiatric history. Gilvarry and colleagues also

‡Gilvarry C, et al. Premorbid IQ in patients with functional psychosis and their first-degree relatives. *Schizophr Res.* 2000;41:417–429.

■ TABLE 3-7. Vertebral Bone Density (mg/cm³)

Group	Sample Size (n)	Mean	SEM
Women with bone fractures	50	70.3	2.55
Men with bone fractures	50	76.2	3.11

■ **TABLE 3-8. The Maslach Burnout Inventory Questionnaire**

	Infectious Disease	Hemophilia	Oncology	Internal Medicine
Mean	46.1	35.0	44.4	47.9
Standard deviation	16.1	11.1	15.6	18.2
Sample size (<i>n</i>)	25	25	25	25

considered whether there was an obstetric complication (OC) during the birth of the psychotic patient, which is another risk factor for impaired intellectual development. Is there any evidence that NART scores differ among these groups of people (see Table 3-10)? (The formulas for analysis of variance with unequal sample sizes are in Appendix A.)

■ **TABLE 3-9. Testes Weight (mg)**

Mouse Strain	Sample Size (<i>n</i>)	Mean	SEM
CD-1	13	142	6
S15/JIs	16	82	3
C17/JIs	17	60	5
B6	15	38	3

■ **TABLE 3-10. National Adult Reading Test Score**

Group	Sample Size (<i>n</i>)	Mean	Standard Deviation
Controls	50	112.7	7.8
Psychotic patients (no obstetric complications)	28	111.6	10.3
Relatives of psychotic patients (no obstetric complications)	25	114.3	12.1
Psychotic patients with obstetric complications	13	110.4	10.1
Relatives of psychotic patients with obstetric complications	19	116.4	8.8

This page intentionally left blank

The Special Case of Two Groups: The t Test

As we have just seen in Chapter 3, many investigations require comparing only two groups. In addition, as the last example in Chapter 3 illustrated, when there are more than two groups, analysis of variance only allows you to conclude that the data are not consistent with the hypothesis that all the samples were drawn from a single population. It does not help you decide *which* one or ones are most likely to differ from the others. To answer these questions, we now develop a procedure that is specifically designed to test for differences in two groups: the t test or *Student's t test*. While we will develop the t test from scratch, we will eventually show that it is just a different way of doing an analysis of variance. In particular, we will see that $F = t^2$ when there are two groups.

The t test is the most common statistical procedure in the medical literature; you can expect it to appear in more than half the papers you read in the general medical literature. In addition to being used to compare two group means, it is widely applied incorrectly to compare multiple groups, by doing all the pairwise comparisons, for example, by comparing more than one intervention with a control condition or the state of a patient at different times following an intervention. As we will see, this incorrect use increases the chances of rejecting the null hypothesis of no effect above the nominal level, say 5%, used to select the cutoff value for a “big” value of the test statistic t . In practical terms, this boils down to increasing the chances of reporting that some therapy

had an effect when the evidence does not support this conclusion.

■ THE GENERAL APPROACH

Suppose we wish to test a new drug that may be an effective diuretic. We assemble a group of 10 people and divide them at random into two groups, a control group that receives a placebo and a treatment group that receives the drug; then we measure their urine production for 24 hours. Figure 4-1A shows the resulting data. The average urine production of the group receiving the diuretic is 240 mL higher than that of the group receiving the placebo. Simply looking at Figure 4-1A, however, does not provide very convincing evidence that this difference is due to anything more than random sampling.

Nevertheless, we pursue the problem and give the placebo or drug to another 30 people to obtain the results shown in Figure 4-1B. The mean responses of the two groups of people as well as the standard deviations are almost identical to those observed in the smaller samples shown in Figure 4-1A. Even so, most observers are more confident in claiming that the diuretic increased average urine output from the data in Figure 4-1B than the data in Figure 4-1A, even though the samples in each case are good representatives of the underlying population. Why?

As the sample size increases, most observers become more confident in their estimates of the population means so they can begin to discern a difference between

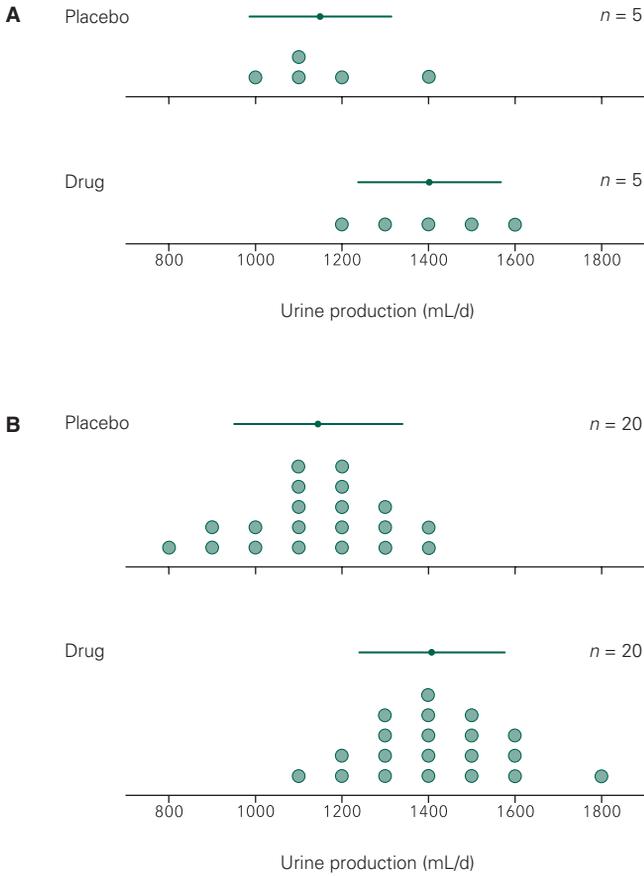


FIGURE 4-1. (A) Results of a study in which five people were treated with a placebo and five people were treated with a drug thought to increase daily urine production. On the average, the five people who received the drug produced more urine than the placebo group. Are these data convincing evidence that the drug is an effective diuretic? (B) Results of a similar study with 20 people in each treatment group. The means and standard deviations associated with the two groups are similar to the results in panel A. Are these data convincing evidence that the drug is an effective diuretic? If you changed your mind, why did you do it?

the people taking the placebo or the drug. Recall that the standard error of the mean quantifies the uncertainty of the estimate of the true population mean based on a sample. Furthermore, as the sample size increases, the standard error of the mean decreases according to

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where n is the sample size and σ is the standard deviation of the population from which the sample was drawn. As the sample size increases the uncertainty in the estimate of the difference of the means between the people who received placebo and the patients who received the drug decreases relative to the difference of the means. As a result, we become more confident that the drug actually has an effect. More precisely, we become less confident in the hypothesis that the drug had no effect, in which case

the two samples of patients could be considered two samples drawn from a single population.

To formalize this logic, we will examine the ratio

$$t = \frac{\text{Difference in sample means}}{\text{Standard error of difference of sample means}}$$

When this ratio is small we will conclude that the data are compatible with the hypothesis that both samples were drawn from a single population. When this ratio is large we will conclude that it is unlikely that the samples were drawn from a single population and assert that the treatment (e.g., the diuretic) produced an effect.

This logic, while differing in emphasis from that used to develop the analysis of variance, is essentially the same. In both cases, we are comparing the relative magnitude of the differences in the sample means with the amount of

variability that would be expected from looking within the samples.

To compute the t ratio we need to know two things: the difference of the sample means and the standard error of this difference. Computing the difference of the sample means is easy; we simply subtract. Computing an estimate for the standard error of this difference is a bit more involved. We begin with a slightly more general problem, that of finding the standard deviation of the difference of two numbers drawn at random from the same population.

■ THE STANDARD DEVIATION OF A DIFFERENCE OR A SUM

Figure 4-2A shows a population with 200 members. The mean is 0, and the standard deviation is 1. Now, suppose we draw two samples at random and compute their difference. Figure 4-2B shows this result for the two members indicated by solid circles in Figure 4-2A. Drawing five more pairs of samples (indicated by different symbols in Fig. 4-2A) and computing their differences yields the corresponding shaded points in Figure 4-2B. Note that there seems to be more variability in the differences of the samples than in the samples themselves. Figure 4-2C shows the results of Figure 4-2B, together with the results of

drawing another 50 pairs of numbers at random and computing their differences. The standard deviation of the population of differences is about 40% larger than the standard deviation of the population from which the samples were drawn.

In fact, it is possible to demonstrate mathematically that *the variance of the difference (or sum) of two variables selected at random equals the sum of the variances of the two populations from which the samples were drawn*. In other words, if X is drawn from a population with standard deviation σ_X and Y is drawn from a population with standard deviation σ_Y , the distribution of all possible values of $X - Y$ (or $X + Y$) will have variance

$$\sigma_{X-Y}^2 = \sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$$

This result should seem reasonable to you because when you select pairs of values that are on opposite (the same) sides of the population mean and compute their difference (sum), the result will be even farther from the mean. Returning to the example in Figure 4-2, we can observe that both the first and second numbers were drawn from the same population whose variance was 1 and so the variance of the difference should be

$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 = 1 + 1 = 2$$

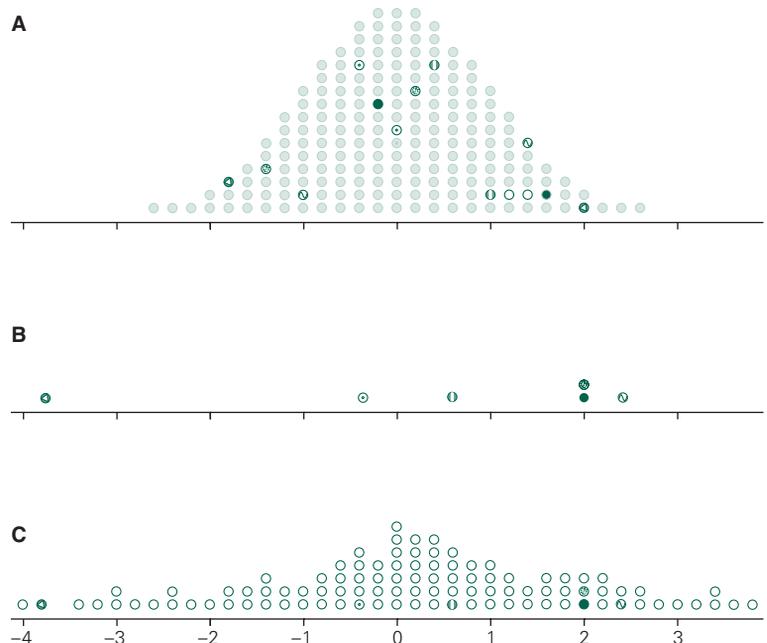


FIGURE 4-2. If one selects pairs of members of the population in panel **A** at random and computes the difference the population of differences, shown in panel **B**, has a wider variance than the original population. Panel **C** shows another 100 values for differences of pairs of members selected at random from the population in **A** to make this point again.

Since the standard deviation is the square root of the variance, the standard deviation of the population of differences will be $\sqrt{2}$ times the standard deviation of the original population, or about 40% bigger, confirming our earlier subjective impression.*

When we wish to estimate the variance in the difference or sum of members of two populations based on the observations, we simply replace the population variances σ^2 in the equation above with the estimates of the variances computed from our samples:

$$s_{X-Y}^2 = s_X^2 + s_Y^2$$

The standard error of the mean is just the standard deviation of the population of all possible sample means of samples of size n , and so we can find the standard error of the difference of two means using the equation above. Specifically,

$$s_{\bar{X}-\bar{Y}}^2 = s_{\bar{X}}^2 + s_{\bar{Y}}^2$$

in which case

$$s_{\bar{X}-\bar{Y}} = \sqrt{s_{\bar{X}}^2 + s_{\bar{Y}}^2}$$

Now we are ready to construct the t ratio from the definition in the last section.

*The fact that the sum of randomly selected variables has a variance equal to the sum of the variances of the individual numbers explains why the standard error of the mean equals the standard deviation divided by \sqrt{n} . Suppose we draw n numbers at random from a population with standard deviation s . The mean of these numbers will be

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + X_3 + \cdots + X_n)$$

so

$$n\bar{X} = X_1 + X_2 + X_3 + \cdots + X_n$$

Since the variance associated with each of the X is a σ^2 , the variance of $n\bar{X}$ will be

$$\sigma_{n\bar{X}}^2 = \sigma^2 + \sigma^2 + \sigma^2 + \cdots + \sigma^2 = n\sigma^2$$

and the standard deviation will be

$$\sigma_{n\bar{X}} = \sqrt{n}\sigma$$

But we want the standard deviation of \bar{X} , which is $n\bar{X}/n$ therefore

$$\sigma_{\bar{X}} = \sqrt{n}\sigma/n = \sigma/\sqrt{n}$$

which is the formula for the standard error of the mean. Note that we made no assumptions about the population from which the sample was drawn. (In particular, we did *not* assume that it had a normal distribution.)

■ USE OF t TO TEST HYPOTHESES ABOUT TWO GROUPS

Recall that we decided to examine the ratio

$$t = \frac{\text{Difference in sample means}}{\text{Standard error of difference of sample means}}$$

We can now use the result of the last section to translate this definition into the equation

$$\begin{aligned} t &= \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} \\ &= \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_{\bar{X}_1}^2 + s_{\bar{X}_2}^2}} \end{aligned}$$

Alternatively, we can write t in terms of the sample standard deviations rather than the standard errors of the mean:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(s_1^2/n) + (s_2^2/n)}}$$

in which n is the size of each sample.

If the hypothesis that the two samples were drawn from the same population is true, the variances s_1^2 and s_2^2 computed from the two samples are both estimates of the same population variance σ^2 . Therefore, we replace the two different estimates of the population variance in the equation above with a single estimate, s^2 , that is obtained by averaging these two separate estimates:

$$s^2 = \frac{1}{2}(s_1^2 + s_2^2)$$

This is called the *pooled-variance estimate* since it is obtained by pooling the two estimates of the population variance to obtain a single estimate. The t test statistic based on the pooled-variance estimate is

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(s^2/n) + (s^2/n)}}$$

The specific value of t one obtains from any two samples depends not only on whether or not there actually is a difference in the means of the populations from which the samples were drawn but also on which specific individuals happened to be selected for the samples. Thus, as for F , there will be a range of possible values that t can have, even when both samples are drawn from the same population. Since the means computed from the two

samples will generally be close to the mean of the population from which they were drawn, the value of t will tend to be small when the two samples are drawn from the same population. Therefore, we will use the same procedure to test hypotheses with t as we did with F in Chapter 3. Specifically, we will compute t from the data then reject the assertion that the two samples were drawn from the same population if the resulting value of t is “big.”

Let us return to the problem of assessing the value of the diuretic we were discussing earlier. Suppose the entire population of interest contains 200 people. In addition, we will assume that the diuretic had no effect, so that the two groups of people being studied can be considered to represent two samples drawn from a single population. Figure 4-3A shows this population, together with two samples of 10 people each selected at random for study.

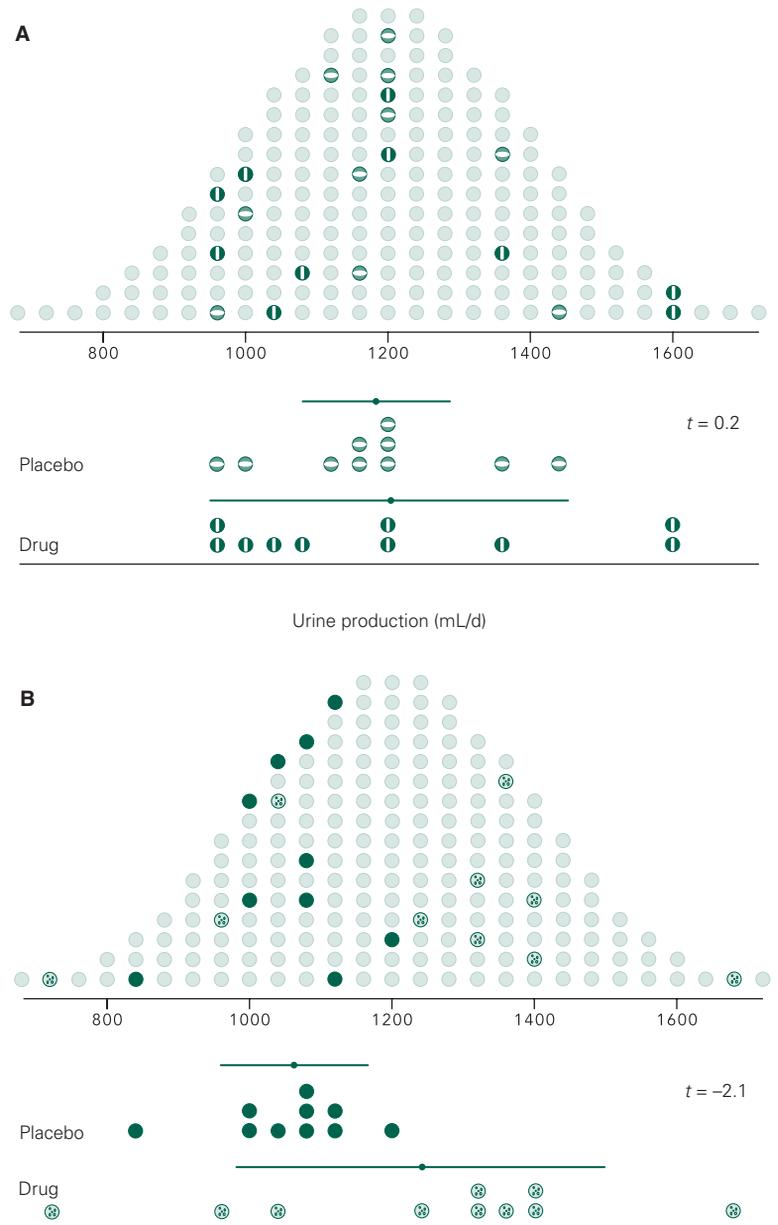


FIGURE 4-3. A population of 200 individuals and two groups selected at random for study of a drug designed to increase urine production but which is totally ineffective. The people shown as dark circles received the placebo and those with the lighter circles received the drug. An investigator would not see the entire population but just the information as reflected in the lower part of panel **A**; nevertheless, the two samples show very little difference and it is unlikely that one would have concluded that the drug had an effect on urine production. Of course, there is nothing special about the two random samples shown in panel **A**, and an investigator could just as well have selected the two groups of people in panel **B** for study. There is more difference between these two groups than the two shown in panel **A** and there is a chance that the investigator would think that this difference is due to the drug's effect on urine production (rather than simple random sampling). (continued)

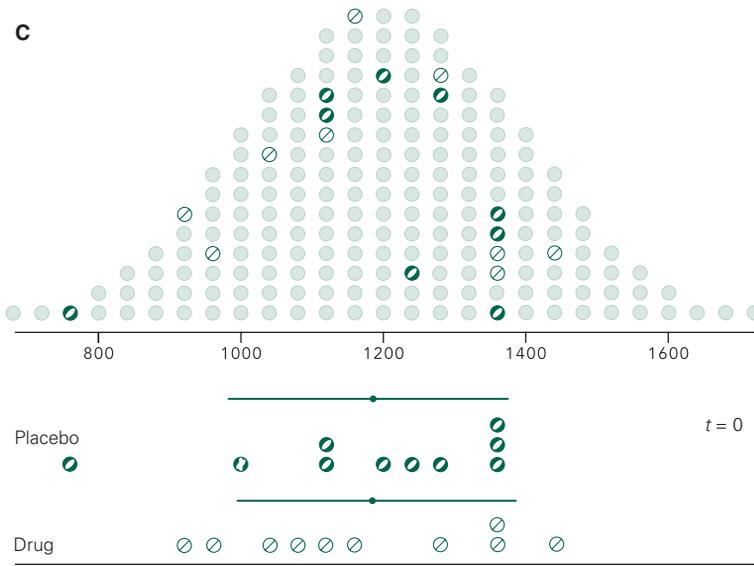


FIGURE 4-3. (Continued) Panel **C** shows yet another pair of random samples the investigator might have drawn for the study.

The people who received the placebo are shown as dark circles, and the people who received the diuretic are shown as lighter circles. The lower part of Figure 4-3A shows the data as they would appear to the investigator together with the mean and standard deviations computed from each of the two samples. Looking at these data certainly does not suggest that the diuretic had any effect. The value of t associated with these samples is -0.2 .

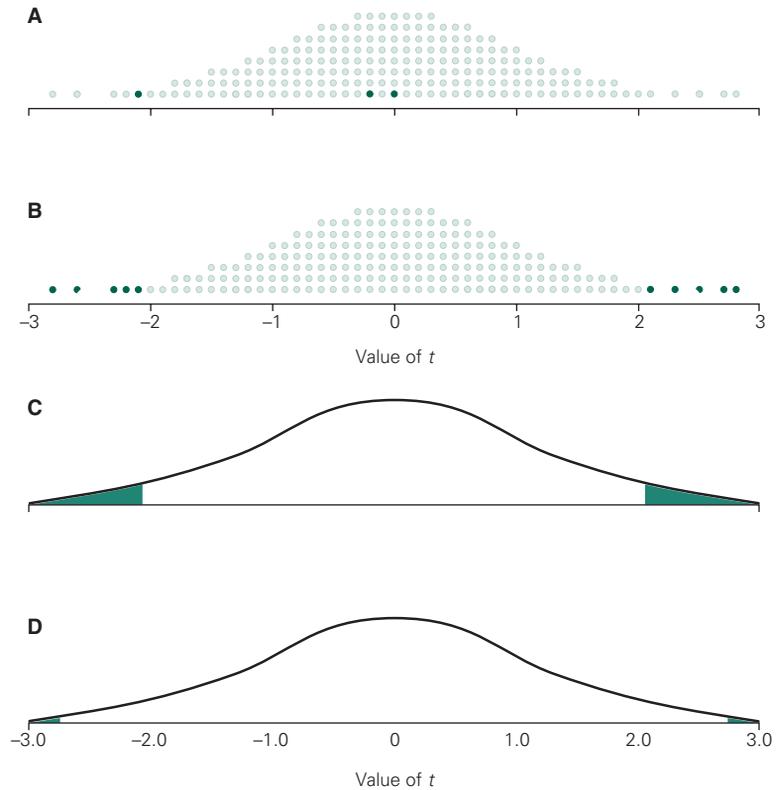
Of course, there is nothing special about these two samples and we could just as well have selected two different groups of people to study. Figure 4-3B shows another collection of people that could have been selected at random to receive the placebo (dark circles) or diuretic (light circles). Not surprisingly, these two samples differ from each other as well as the samples selected in Figure 4-3A. Given only the data in the lower part of Figure 4-3B we might think that the diuretic increases urine production. The t value associated with these data is -2.1 . Figure 4-3C shows yet another pair of samples. They differ from each other and the other samples considered in Figure 4-3A and 4-3B. The samples in Figure 4-3C yield a value of 0 for t .

We could continue this process for quite a long time since there are more than 10^{27} different pairs of samples of 10 people each that we could draw from the population of 200 individuals shown in Figure 4-3A. We can compute a value of t for each of these 10^{27} different pairs of samples. Figure 4-4 shows the values of t associated with 200 different

pairs of random samples of 10 people each drawn from the original population, including the three specific pairs of samples shown in Figure 4-3. The distribution of possible t values is symmetrical about $t = 0$ because it does not matter which of the two samples we subtract from the other. As predicted, most of the resulting values of t are close to zero; t rarely is below about -2 or above $+2$.

Figure 4-4 allows us to determine what a “big” t is. Figure 4-4B shows that t will be less than -2.1 or greater than $+2.1$ 10 out of 200, or 5% of the time. In other words, there is only a 5% chance of getting a value of t more extreme than -2.1 or $+2.1$ when the two samples are drawn from the same population. Just as with the F distribution, the number of possible t values rapidly increases beyond 10^{27} as the population size grows, and the distribution of possible t values approaches a smooth curve. Figure 4-4C shows the result of this limiting process. We define the cutoff values for t that are large enough to be called “big” on the basis of the total area in the two tails. Figure 4-4C shows that only 5% of the possible values of t will lie beyond -2.1 or $+2.1$ when the two samples are drawn from a single population. When the data are associated with a value of t beyond this range, it is customary to conclude that the data are inconsistent with the null hypothesis of no difference between the two samples and report that there was a difference in treatment.

FIGURE 4-4. The results of 200 studies like that described in Figure 4-3; the three specific studies from Figure 4-3 are indicated in panel **A**. Note that most values of the t statistic cluster around 0, but it is possible for some values of t to be quite large, exceeding 1.5 or 2. Panel **B** shows that there are only 10 chances in 200 of t exceeding 2.1 in magnitude if the two samples were drawn from the same population. If one continues examining all possible samples drawn from the population and our pairs of samples drawn from the same population, one obtains a distribution of all possible t values which becomes the smooth curve in panel **C**. In this case, one defines the critical value of t by saying that it is unlikely that this value of t statistic was observed under the hypothesis that the drug had no effect by taking the 5% most extreme error areas under the tails of distribution and selecting the t value corresponding to the beginning of this region. Panel **D** shows that if one required a more stringent criterion for rejecting the hypothesis for no difference by requiring that t be in the most extreme 1% of all possible values, the cutoff value of t is 2.878.



The extreme values of t that lead us to reject the hypothesis of no difference lie in both tails of the distribution. Therefore, the approach we are taking is sometimes called a *two-tailed t test*. Occasionally, people use a one-tailed t test, and there are indeed cases where this is appropriate. One should be suspicious of such one-tailed tests, however, because the cutoff value for calling t “big” for a given value of P is smaller. In reality, people are almost always looking for a *difference* between the control and treatment groups so a two-tailed test is appropriate. This book always assumes a two-tailed test.

Note that the data in Figure 4-3B are associated with a t value of -2.1 , which we have decided to consider “big.” If all we had were the data shown in Figure 4-4B, we would conclude that the observations were inconsistent with the hypothesis that the diuretic had no effect and report that it *increased* urine production, and even though we did the statistical analysis correctly, *our conclusion about the drug would be wrong*.

Reporting $P < .05$ means that if the treatment had no effect, there is less than a 5% chance of getting a value of t from the data as far or farther from 0 as the critical value for t to be called “big.” It does not mean it is impossible to get such a large value of t when the treatment has no effect. We could, of course, be more conservative and say that we will reject the hypothesis of no difference between the populations from which the samples were drawn if t is in the most extreme 1% of possible values. Figure 4-4D shows that this would require t to be beyond -2.88 or $+2.88$ in this case, so we would not erroneously conclude that the drug had an effect on urine output in any of the specific examples shown in Figure 4-3. In the long run, however, we will make such errors about 1% of the time. The price of this conservatism is decreasing the chances of concluding that there is a difference when one really exists. Chapter 6 discusses this trade-off in more detail.

The critical values of t , like F , have been tabulated and depend not only on the level of confidence with which one rejects the hypothesis of no difference—the P

value—but also on the sample size. As with the F distribution, this dependence on sample size enters the table as the *degrees of freedom*, v , which is equal to $2(n - 1)$ for this t test, where n is the size of each sample. As the sample size increases the value of t needed to reject the hypothesis of no difference decreases. In other words, as sample size increases it becomes possible to detect smaller differences with a given level of confidence. Reflecting on Figure 4-1 should convince you that this is reasonable.

■ WHAT IF THE TWO SAMPLES ARE NOT THE SAME SIZE?

It is easy to generalize the t test to handle problems in which there are different numbers of members in the two samples being studied. Recall that t is defined by

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_{\bar{X}_1}^2 + s_{\bar{X}_2}^2}}$$

in which $s_{\bar{X}_1}$ and $s_{\bar{X}_2}$ are the standard errors of the means of the two samples. If the first sample is of size n_1 and the second sample contains n_2 members,

$$s_{\bar{X}_1}^2 = \frac{s_1^2}{n_1} \quad \text{and} \quad s_{\bar{X}_2}^2 = \frac{s_2^2}{n_2}$$

in which s_1 and s_2 are the standard deviations of the two samples. Use these definitions to rewrite the definition of t in terms of the sample standard deviations

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}$$

When the two samples are different sizes, the pooled estimate of the variance is given by

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

so that

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(s^2/n_1) + (s^2/n_2)}}$$

This is the definition of t for comparing two samples of unequal size. There are $v = n_1 + n_2 - 2$ degrees of freedom.

Notice that this result reduces to our earlier results when the two sample sizes are equal, that is, when $n_1 = n_2 = n$.

■ CELL PHONES REVISITED

The study by Fejes and colleagues about the relationship of cell phone use and rapid sperm motility we discussed in Chapter 3 had two observational groups, 61 men who used cell phones less than 15 minutes/day and 61 men who used cell phones more than 60 minutes/day, so we can analyze their data using a t test, as well as an analysis of variance. From Figure 3-7, the mean percentage of rapidly mobile sperm was 49% for the low use group and 41% for the high use group. The standard deviations were 21% and 22%, respectively. Because the sample sizes are equal,*

$$s^2 = \frac{1}{2}(s_{\text{low}}^2 + s_{\text{high}}^2)$$

$$s^2 = \frac{1}{2}(21^2 + 22^2) = 462.5\%{}^2$$

and

$$t = \frac{\bar{X}_{\text{low}} - \bar{X}_{\text{high}}}{\sqrt{\frac{s^2}{n_{\text{low}}} + \frac{s^2}{n_{\text{high}}}}}$$

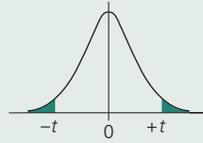
$$t = \frac{49 - 41}{\sqrt{\frac{462.5}{61} + \frac{462.5}{61}}} = 2.054$$

with $v = 2(n - 1) = 2(61 - 1) = 120$ degrees of freedom. Table 4-1 shows that the magnitude of t should only exceed 1.980 only 5% of the time by chance when the null hypothesis is true, in this case, that cell phone exposure does not affect rapid sperm motility ($P < .05$). Since the magnitude of t associated with the data exceeds 1.980, we reject the null hypothesis and conclude that cell phone use is associated with rapid sperm motility.

*We would have obtained precisely the same value had we used the general formula for the pooled variance:

$$s^2 = \frac{(n_{\text{low}} - 1)s_{\text{low}}^2 + (n_{\text{high}} - 1)s_{\text{high}}^2}{n_{\text{low}} + n_{\text{high}} - 2}$$

$$s^2 = \frac{(61 - 1)21^2 + (61 - 1)22^2}{61 + 61 - 2} = \frac{60 \cdot 21^2 + 60 \cdot 22^2}{2 \cdot 60} = 462.5\%{}^2$$

■ TABLE 4-1. Critical Values of t (Two-Tailed)

ν	Probability of Greater Value (P)								
	0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
1	1.000	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	1.415	1.895	2.365	2.998	3.449	4.029	4.785	5.408
8	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
31	0.682	1.309	1.696	2.040	2.453	2.744	3.022	3.375	3.633
32	0.682	1.309	1.694	2.037	2.449	2.738	3.015	3.365	3.622
33	0.682	1.308	1.692	2.035	2.445	2.733	3.008	3.356	3.611
34	0.682	1.307	1.691	2.032	2.441	2.728	3.002	3.348	3.601
35	0.682	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591
36	0.681	1.306	1.688	2.028	2.434	2.719	2.990	3.333	3.582
37	0.681	1.305	1.687	2.026	2.431	2.715	2.985	3.326	3.574
38	0.681	1.304	1.686	2.024	2.429	2.712	2.980	3.319	3.566

(continued)

■ TABLE 4-1. Critical Values of t (Two-Tailed) (Continued)

ν	Probability of Greater Value (P)								
	0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
39	0.681	1.304	1.685	2.023	2.426	2.708	2.976	3.313	3.558
40	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
42	0.680	1.302	1.682	2.018	2.418	2.698	2.963	3.296	3.538
44	0.680	1.301	1.680	2.015	2.414	2.692	2.956	3.286	3.526
46	0.680	1.300	1.679	2.013	2.410	2.687	2.949	3.277	3.515
48	0.680	1.299	1.677	2.011	2.407	2.682	2.943	3.269	3.505
50	0.679	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
52	0.679	1.298	1.675	2.007	2.400	2.674	2.932	3.255	3.488
54	0.679	1.297	1.674	2.005	2.397	2.670	2.927	3.248	3.480
56	0.679	1.297	1.673	2.003	2.395	2.667	2.923	3.242	3.473
58	0.679	1.296	1.672	2.002	2.392	2.663	2.918	3.237	3.466
60	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
62	0.678	1.295	1.670	1.999	2.388	2.657	2.911	3.227	3.454
64	0.678	1.295	1.669	1.998	2.386	2.655	2.908	3.223	3.449
66	0.678	1.295	1.668	1.997	2.384	2.652	2.904	3.218	3.444
68	0.678	1.294	1.668	1.995	2.382	2.650	2.902	3.214	3.439
70	0.678	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435
72	0.678	1.293	1.666	1.993	2.379	2.646	2.896	3.207	3.431
74	0.678	1.293	1.666	1.993	2.378	2.644	2.894	3.204	3.427
76	0.678	1.293	1.665	1.992	2.376	2.642	2.891	3.201	3.423
78	0.678	1.292	1.665	1.991	2.375	2.640	2.889	3.198	3.420
80	0.678	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
90	0.677	1.291	1.662	1.987	2.368	2.632	2.878	3.183	3.402
100	0.677	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
120	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
140	0.676	1.288	1.656	1.977	2.353	2.611	2.852	3.149	3.361
160	0.676	1.287	1.654	1.975	2.350	2.607	2.846	3.142	3.352
180	0.676	1.286	1.653	1.973	2.347	2.603	2.842	3.136	3.345
200	0.676	1.286	1.653	1.972	2.345	2.601	2.839	3.131	3.340
∞	0.6745	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.0902	3.2905
Normal	0.6745	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.0902	3.2905

Adapted from Zar JH. *Biostatistical Analysis*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall; 1984, 484–485:table B.3, by permission of Pearson Education, Inc., Upper Saddle River, NJ.

This is the same conclusion and same P value we obtained when analyzing the data using analysis of variance.

Fejes and colleagues also measured total sperm motility, this time in different numbers of men in the two samples of cell phone users. Box 4-1 shows the data and calculation of the associated t test. The t value for these data fall between the critical values of 0.667 and 1.289 that

define the 50% and 20% extremes of the t distribution, which is nowhere near the value of 1.980 that defines the most extreme 5%, the cutoff used to define traditional statistical significance. Thus, we do not have strong enough evidence to reject the null hypothesis that there is no relationship between cell phone exposure and total sperm motility.

BOX 4-1 • Effect of Low Versus High Cell Phone Use on Overall Sperm Motility**Total Sperm Motility (%)**

Observed Cell Phone Use	Sample Size (n)	Mean	Standard Deviation
Low use (<15 min/d)	120	60	19
High use (>60 min/d)	62	57	17

Because the sample sizes are different, we compute the pooled variance estimate with

$$s^2 = \frac{(n_{\text{low}} - 1)s_{\text{low}}^2 + (n_{\text{high}} - 1)s_{\text{high}}^2}{n_{\text{low}} + n_{\text{high}} - 2}$$

$$s^2 = \frac{(120 - 1)19^2 + (62 - 1)17^2}{120 + 62 - 2} = 336.6\%^2$$

and so

$$t = \frac{\bar{X}_{\text{low}} - \bar{X}_{\text{high}}}{\sqrt{\frac{s^2}{n_{\text{low}}} + \frac{s^2}{n_{\text{high}}}}}$$

$$t = \frac{60 - 57}{\sqrt{\frac{336.6}{120} + \frac{336.6}{62}}} = 1.045$$

with $\nu = n_{\text{low}} + n_{\text{high}} - 2 = 180$ degrees of freedom. This value of t does not even approach 1.973, the critical value for the most extreme 5% of the t distribution used to define conventional statistical significance, so we do not reject the null hypothesis of no effect of cell phone use on overall sperm motility.

Does this result *prove* that there really is not an effect? No. It just means that we do not have enough evidence to reject the null hypothesis of no effect. (We will return to the question of how confident we can be in drawing negative conclusions when results do not reach statistical significance in Chapter 6.)

■ THE t TEST IS AN ANALYSIS OF VARIANCE*

The t test we just developed and analysis of variance we developed in Chapter 3 are really two different ways of doing the same thing. Since few people recognize this, we

will prove that when comparing the means of two groups, $F = t^2$. In other words, the t test is simply a special case of analysis of variance applied to two groups.

We begin with two samples, each of size n , with means and standard deviations \bar{X}_1 and \bar{X}_2 and s_1 and s_2 , respectively.

To form the F ratio used in analysis of variance, we first estimate the population variance as the average of the variances computed for each group

$$s_{\text{wit}}^2 = \frac{1}{2}(s_1^2 + s_2^2)$$

Next, we estimate the population variance from the sample means by computing the standard deviation of the sample means with

$$s_{\bar{X}} = \sqrt{\frac{(\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2}{2 - 1}}$$

*This section represents the only mathematical proof in this book and as such is a bit more technical than everything else. The reader can skip this section with no loss of continuity.

Therefore

$$s_{\bar{X}}^2 = (\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2$$

in which \bar{X} is the mean of the two sample means

$$\bar{X} = \frac{1}{2}(\bar{X}_1 + \bar{X}_2)$$

Eliminate \bar{X} from the equation for $s_{\bar{X}}^2$ to obtain

$$\begin{aligned} s_{\bar{X}}^2 &= [\bar{X}_1 - \frac{1}{2}(\bar{X}_1 + \bar{X}_2)]^2 + [\bar{X}_2 - \frac{1}{2}(\bar{X}_1 + \bar{X}_2)]^2 \\ &= (\frac{1}{2}\bar{X}_1 - \frac{1}{2}\bar{X}_2)^2 + (\frac{1}{2}\bar{X}_2 - \frac{1}{2}\bar{X}_1)^2 \end{aligned}$$

Since the square of a number is always positive, $(a - b)^2 = (b - a)^2$ and the equation above becomes

$$\begin{aligned} s_{\bar{X}}^2 &= (\frac{1}{2}\bar{X}_1 - \frac{1}{2}\bar{X}_2)^2 + (\frac{1}{2}\bar{X}_2 - \frac{1}{2}\bar{X}_1)^2 \\ &= 2[\frac{1}{2}(\bar{X}_1 - \bar{X}_2)]^2 = \frac{1}{2}(\bar{X}_1 - \bar{X}_2)^2 \end{aligned}$$

Therefore, the estimate of the population variance from between the groups is

$$s_{\text{bet}}^2 = ns_{\bar{X}}^2 = (n/2)(\bar{X}_1 - \bar{X}_2)^2$$

Finally, F is the ratio of these two estimates of the population variance

$$\begin{aligned} F &= \frac{s_{\text{bet}}^2}{s_{\text{wit}}^2} = \frac{(n/2)(\bar{X}_1 - \bar{X}_2)^2}{\frac{1}{2}(s_1^2 + s_2^2)} = \frac{(\bar{X}_1 - \bar{X}_2)^2}{(s_1^2/n) + (s_2^2/n)} \\ &= \left[\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(s_1^2/n) + (s_2^2/n)}} \right]^2 \end{aligned}$$

The quantity in the brackets is t , hence

$$F = t^2$$

The degrees of freedom for the numerator of F equals the number of groups minus 1, that is, $2 - 1 = 1$ for all comparisons of two groups. The degrees of freedom for the denominator equals the number of groups times 1 less than the sample size of each group, $2(n - 1)$, which is the same as the degrees of freedom associated with the t test.

In sum, the t test and analysis of variance are just two different ways of looking at the same test for two groups. Of course, if there are more than two groups, one cannot use the t test form of analysis of variance but must use the more general form we developed in Chapter 3.

As noted earlier, we drew the same conclusion about the effects of cell phone use on rapid sperm motility when analyzing the results using analysis of variance in Chapter 3 and using a t test in this chapter. As expected, the degrees of freedom for the t test, ν , is 120, the same as the denominator degrees of freedom for the analysis of variance, ν_d and the square of the t value we obtained, 2.054^2 , equals the value of F we obtained from the analysis of variance, 4.22.

■ COMMON ERRORS IN THE USE OF THE t TEST AND HOW TO COMPENSATE FOR THEM

The t test is used to compute the probability of being wrong, the P value, when asserting that the mean values of two treatment groups are different, when, in fact, they were drawn from the same population. It is also used widely but erroneously to test for differences between more than two groups by comparing all possible pairs of means with t tests.

For example, suppose an investigator measured blood sugar under control conditions, in the presence of drug A, and in the presence of drug B. It is common to perform three t tests on these data: one to compare control versus drug A, one to compare control versus drug B, and one to compare drug A versus drug B. This practice is incorrect because the true probability of erroneously concluding that the drug affected blood sugar is actually higher than the nominal level, say 5%, used when looking up the “big” cutoff value of the t statistic in a table.

To understand why, reconsider the experiment described in the last paragraph. Suppose that if the value of the t statistic computed in one of the three comparisons just described is in the most extreme 5% of the values that would occur if the drugs really had no effect, we will reject that assumption and assert that the drugs changed blood sugar. We will be satisfied if $P < .05$; in other words, in the long run we are willing to accept the fact that 1 statement in 20 will be wrong. Therefore, when we test control versus drug A, we can expect erroneously to assert a difference 5% of the time. Similarly, when testing control versus drug B, we expect erroneously to assert a difference 5% of the time, and when testing drug A versus drug B, we expect erroneously to assert a difference 5% of the time. Therefore, when considering the three

tests together, we expect to conclude that at least one pair of groups differs about $5\% + 5\% + 5\% = 15\%$ of the time, even if in reality the drugs did not affect blood sugar. (As we will see later, P actually equals 14%.) If there are not too many comparisons, simply adding the P values obtained in multiple tests produces a realistic and conservative estimate of the true P value for the set of comparisons.

In the example above, there were three t tests, so the effective P value was about $3(.05) = .15$ or 15%. When comparing four groups, there are six possible t tests (1 versus 2, 1 versus 3, 1 versus 4, 2 versus 3, 2 versus 4, 3 versus 4), so if the author concludes that there is a difference and reports $P < .05$, the effective P value is about $6(.05) = .30$; there is about a 30% chance of at least one incorrect statement if the author concludes that the treatments had an effect!

In Chapter 2, we discussed random samples of Martians to illustrate the fact that different samples from the same population yield different estimates of the population mean and standard deviation. Figure 2-5 showed

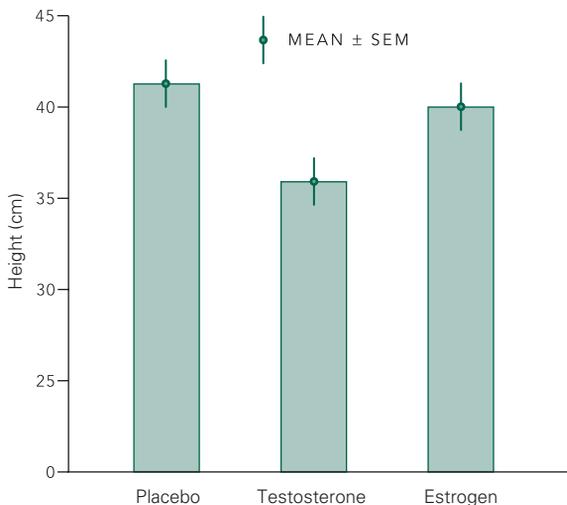


FIGURE 4-5. Results of a study of human hormones on Martians as it would be commonly presented in the medical literature. Each large bar has a height equal to the mean of the group; the small vertical bars indicate 1 standard error of the mean on either side of the mean (not 1 standard deviation).

three such samples of the heights of Martians, all drawn from a single population. Suppose we chose to study how these Martians respond to human hormones. We draw three samples at random, give one group a placebo, one group testosterone, and one group estrogen. Suppose that these hormones have no effect on the Martians' heights. Thus, the three groups shown in Figure 2-5 represent three samples drawn at random from the same population.

Figure 4-5 shows how these data would probably appear in a typical medical journal. The large vertical bars denote the value of the mean responses, and the small vertical bars denote 1 standard error of the mean above or below the sample means. (Showing 1 standard deviation would be the appropriate way to describe variability in the samples.) Many authors would analyze these data by performing three t tests: placebo against testosterone, placebo against estrogen, and testosterone against estrogen. These three tests yield t values of 2.39, 0.93, and 1.34, respectively. Since each test is based on 2 samples of 10 Martians each, there are $2(10 - 1) = 18$ degrees of freedom. From Table 4-1, the critical value of t with a 5% chance of erroneously concluding that a difference exists is 2.101. Thus, the author would conclude that testosterone produced shorter Martians than placebo, whereas estrogen did not differ significantly from placebo, and that the two hormones did not produce significantly different results.

Think about this result for a moment. What is wrong with it? If testosterone produced results not detectably different from those of estrogen and estrogen produced results not detectably different from those of placebo, how can testosterone have produced results different from placebo? Far from alerting medical researchers that there is something wrong with their analysis, this illogical result usually leads to a very creatively written "Discussion" section in their paper.

An analysis of variance of these data yields $F = 2.74$ [with numerator degrees of freedom $= m - 1 = 3 - 1 = 2$ and denominator degrees of freedom $m(n - 1) = 3(10 - 1) = 27$], which is below the critical value of 3.35 we have decided is required to assert that the data are incompatible with the hypothesis that all three treatments acted as placebos.

Of course, performing an analysis of variance does not ensure that we will not reach a conclusion that is actually wrong, but it will make it less likely.

We end our discussion of common errors in the use of the t test with three rules of thumb:

- The t test can be used to test the hypothesis that two group means are not different.
- When the experimental design involves multiple groups, analysis of variance should be used.
- When t tests are used to test for differences between multiple groups, it is not appropriate simply to use multiple t tests to do pairwise comparisons of the groups.

■ HOW TO USE t TESTS TO ISOLATE DIFFERENCES BETWEEN GROUPS IN ANALYSIS OF VARIANCE

The last section demonstrated that when presented with data from experiments with more than two groups of subjects, one must do an analysis of variance to determine how inconsistent the observations are with the hypothesis that all the treatments had the same effect. Doing pairwise comparisons with t tests increases the chances of erroneously reporting an effect above the nominal value, say 5%, used to determine the value of a “big” t . The analysis of variance, however, only tests the global hypothesis that *all* the samples were drawn from a single population. In particular, it does not provide any information on which sample or samples differed from the others.

There are a variety of methods, called *multiple-comparison procedures*, that can be used to provide information on this point. All are essentially based on the t test but include appropriate corrections for the fact that we are comparing more than one pair of means. We will develop several approaches, beginning with the *Bonferroni corrected t test*, or, more simply, the *Bonferroni t test*. The general approach we take is first to perform an analysis of variance to test the overall null hypothesis of no differences, then use a multiple-comparison procedure to isolate the treatment or treatments producing the different results.*

*Some statisticians believe that this approach is too conservative and that one should skip the analysis of variance and proceed directly to the multiple comparisons of interest.

The Bonferroni t Test

In the previous section, we saw that if one analyzes a set of data with three t tests, each using the 5% critical value for concluding that there is a difference, there is about a $3(5) = 15\%$ chance of finding it. This result is a special case of a formula called the *Bonferroni inequality*, which states that if k statistical tests are performed with the cutoff value for the test statistics, for example, t or F , at the α level, the likelihood of observing a value of the test statistic exceeding the cutoff value at least once when the treatments did not produce an effect is no greater than k times α . Mathematically, the Bonferroni inequality states

$$\alpha_T < k\alpha$$

where α_T is the true probability of erroneously concluding a difference exists at least once. α_T is the error rate we want to control. From the equation above,

$$\frac{\alpha_T}{k} < \alpha$$

Thus, if we do *each* of the t tests using the critical value of t corresponding to α_T/k , the error rate for *all* the comparisons taken as a group will be at most α_T . For example, if we wish to do three comparisons with t tests while keeping the probability of making at least one false-positive error to less than 5%, require that the P value associated with each value of t be smaller than $.05/3 = 1.67\%$ for each of the individual comparisons. This procedure is called the *Bonferroni corrected t test* or, more simply, the *Bonferroni t test*, because it is based on the Bonferroni inequality.

This procedure works reasonably well when there are only a few groups to compare, but as the number of comparisons k increases above 3 or 4, the value of t required to conclude that a difference exists becomes much larger than it really needs to be and the method becomes overly conservative. One way to make the Bonferroni t test less conservative is to use the estimate of the population variance computed from within the groups in the analysis of variance. Specifically, recall that we defined t as

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(s^2/n_1) + (s^2/n_2)}}$$

where s^2 is an estimate of the population variance. We will replace this estimate with the population variance estimated

from within the groups as part of the analysis of variance, s_{wit}^2 , to obtain

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(s_{\text{wit}}^2/n_1) + (s_{\text{wit}}^2/n_2)}}$$

The degrees of freedom for this test are the same as the denominator degrees of freedom for the analysis of variance and will be higher than for a simple t test based on the two samples being compared. Since the critical value of t decreases as the degrees of freedom increase, it will be possible to detect a difference with a given confidence with smaller absolute differences in the means.

More on Cell Phones and Rabbit Sperm

In Chapter 3 we analyzed the data in Box 3-1 and concluded that they were inconsistent with the null hypothesis that the three sample groups of rabbits—ordinary controls in regular rabbit cages, stress controls in more restrictive cages, and cell phone exposed rabbits in restricted cages where their testes were exposed to cell phone radiation for 8 hours a day—were drawn from populations with the same mean sperm motility. At the time, however, we were unable to isolate where the difference came from. Now we can use the Bonferroni t test to compare the three groups pairwise.

From Box 3-1, our best estimate of the within-groups variance s_{wit}^2 is 7.11% . There are $m = 3$ samples, each consisting of $n = 8$ rabbits, so there are $m(n - 1) = 3(8 - 1) = 21$ degrees of freedom associated with the estimate of the within-groups variance. (By comparison, if we just used the pooled variance from the two samples in each pairwise comparison, there would only be $2[n - 1] = 2[8 - 1] = 14$ degrees of freedom.)

We do the three pairwise comparisons by computing the corresponding three values of t using the within-groups

variance from the analysis of variance. To compare the ordinary control with the cell phone exposure,

$$t_{\text{ord vs phone}} = \frac{\bar{X}_{\text{ordinary}} - \bar{X}_{\text{phone}}}{\sqrt{\frac{s_{\text{wit}}^2}{n_{\text{ordinary}}} + \frac{s_{\text{wit}}^2}{n_{\text{phone}}}}} = \frac{72 - 50}{\sqrt{\frac{7.11}{8} + \frac{7.11}{8}}} = 1.501$$

To compare the stress control with the cell phone exposure,

$$t_{\text{stress vs phone}} = \frac{\bar{X}_{\text{stress}} - \bar{X}_{\text{phone}}}{\sqrt{\frac{s_{\text{wit}}^2}{n_{\text{stress}}} + \frac{s_{\text{wit}}^2}{n_{\text{phone}}}}} = \frac{61 - 50}{\sqrt{\frac{7.11}{8} + \frac{7.11}{8}}} = 8.251$$

To compare the ordinary control with the stress control,

$$t_{\text{ord vs stress}} = \frac{\bar{X}_{\text{ordinary}} - \bar{X}_{\text{stress}}}{\sqrt{\frac{s_{\text{wit}}^2}{n_{\text{ordinary}}} + \frac{s_{\text{wit}}^2}{n_{\text{stress}}}}} = \frac{72 - 61}{\sqrt{\frac{7.11}{8} + \frac{7.11}{8}}} = 8.251$$

There are three comparisons, so to have an overall family error rate of less than 5% we require that the P value associated with each of these three comparisons to be smaller than $.05/3 = .0167$. (Table 4-2 summarizes the three pairwise comparisons.) All three t values exceed 4.140, the critical value for $P < .001$ with 14 degrees of freedom (in Table 3-1), which is much smaller than the required .0167, so we conclude that all three groups are different from each other.

In other words, rabbits in the stress cage have significantly lower sperm motility than rabbits in the ordinary cage and rabbits exposed to cell phone radiation have significantly lower sperm motility than rabbits in the stress cage (as well as the rabbits in the ordinary cage). Thus, we conclude that, while the limited cage space led to lower sperm motility, the cell phone radiation further lowered

■ **TABLE 4-2. Pairwise Comparisons of Sperm Motility in Rabbit Cell Phone Experiment Using Bonferroni t Tests (Family Error Rate, $\alpha_T = 0.05$)**

Comparison	t	P	$P_{\text{crit}} \alpha_T/k$	$P < P_{\text{crit}}?$
Ordinary control vs. cell phone	16.501	<.001	.0167	Yes
Stress control vs. cell phone	8.251	<.001	.0167	Yes
Ordinary control vs. stress control	8.251	<.001	.0167	Yes

$v = 21$ degrees of freedom; $k = 3$ comparisons.

sperm motility. Since these data came from an experimental rather than an observational study, we can conclude that the cell phone radiation *caused* the reduction in sperm motility in these rabbits (as did the stress of being in a cramped cage).

Note that the three t pairwise comparisons are listed in declining order based on the value of t associated with the comparison, from largest to smallest difference. While doing the tests in the order from largest to smallest difference is not required when doing Bonferroni t tests, it is standard practice. The more powerful Holm-Sidak corrected t test that we discuss next requires that the tests be done in order from largest to smallest difference, as measured by the t values associated with the individual comparisons.

A Better Approach to Multiple Comparisons: The Holm t Test

There have been several refinements of the Bonferroni t test designed to maintain the computational simplicity while avoiding the excessive caution that the Bonferroni correction brings. We begin with the *Holm corrected t test* or, more simply, the *Holm t test*.^{*} The Holm correction is nearly as easy to compute as the Bonferroni correction, but yields a more powerful test.[†] The Holm t test is a so-called sequentially rejective, or step-down, procedure because it applies an accept/reject criterion to a set of ordered null hypotheses, starting with the smallest P value, and proceeding until it fails to reject a null hypothesis.

To perform the Holm t test, we compute the family of pairwise comparisons of interest (with t tests using the pooled variance estimate from the analysis of variance as we did with the Bonferroni t test) and determine the P value for each test in the family. We then compare these

P values to critical values that have been adjusted to control the overall family error rate when doing the multiple comparisons.

In contrast to the Bonferroni correction, however, we take into account how many tests we have already done and become less conservative with each subsequent comparison. We begin with a correction just as conservative as the Bonferroni correction, then take advantage of the conservatism of the earlier tests and become less cautious with each subsequent comparison.

Suppose we wish to make k pairwise comparisons.[‡] Order these k uncorrected P values from smallest to largest, with the smallest uncorrected P value considered first in the sequential step-down test procedure. (Because all the P values are based on the same number of degrees of freedom, this ordering is the same as ordering the comparisons based on the magnitude of t from largest to smallest, without regard for the signs associated with the individual t tests.) P_1 is the smallest P value in the sequence (corresponding to the most extreme pairwise comparison) and P_k is the largest. For the j th hypothesis test in this ordered sequence, Holm's test applies the Bonferroni criterion in a step-down manner that depends on k and j , beginning with $j = 1$, and proceeding until we fail to reject the null hypothesis or run out of comparisons to do. Specifically, the uncorrected P value for the j th test is compared to $\alpha_j = \alpha_T / (k - j + 1)$. For the first comparison, $j = 1$, and the uncorrected P value needs to be smaller than $\alpha_1 = \alpha_T / (k - 1 + 1) = \alpha_T / k$, the same as the Bonferroni correction. If this smallest observed P value is less than α_1 , we reject that null hypothesis and then compare the next smallest uncorrected P value with $\alpha_2 = \alpha_T / (k - 2 + 1) = \alpha_T / (k - 1)$, which is a larger cutoff than we would obtain just using the Bonferroni correction. Because this critical value is larger, the test is less conservative and more powerful.

In the example of the relationship between cell phone exposure and sperm motility we have been discussing, there are $k = 3$ pairwise comparisons of interest, so to maintain an overall family error rate, α_T , of 5%, the P value associated with the first ($j = 1$) of these ordered hypotheses (comparisons) will have to be smaller than $.05 / (3 - 1 + 1) = .05 / 3 = .0167$, which is identical to the Bonferroni correction we applied previously to each of the members of this family of three tests. The comparison

^{*}Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat.* 1979;6:65–70.

[†]Other multiple comparisons include the Tukey t test, Student-Neuman-Keuls test, and Dunnett test. The Holm test is superior to these older tests. For more details, see Ludbrook J. Multiple comparison procedures updated. *Clin Exp Pharmacol Physiol.* 1998;25:1032–1037; Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: the Bonferroni vs. Holm methods. *Am J Public Health.* 1996;86:726–728; Levin B. Annotation: on the Holm, Simes, and Hochberg multiple test procedures. *Am J Public Health.* 1996;86:628–629; Brown BW, Russel K. Methods for correcting for multiple testing: operating characteristics. *Stat Med.* 1997;16:2511–2528; Morikawa T, Terao A, Iwasaki M. Power evaluation of various modified Bonferroni procedures by a Monte Carlo study. *J Biopharm Stat.* 1996;6:343–359.

[‡]Like the Bonferroni correction, the Holm correction can be applied to any family of hypothesis tests, not just multiple pairwise comparisons.

■ **TABLE 4-3. Pairwise Comparisons of Sperm Motility in Rabbit Cell Phone Experiment Using Holm t Tests (Family Error Rate, $\alpha_T = .05$)**

Comparison	t	P	j	$P_{\text{crit}} = \alpha_T / (k - j + 1)^*$	$P < P_{\text{crit}}?$
Ordinary control vs. cell phone	16.501	<.001	1	.0167	Yes
Stress control vs. cell phone	8.251	<.001	2	.0250	Yes
Ordinary control vs. stress control	8.251	<.001	3	.0500	Yes

$v = 21$ degrees of freedom; $k = 3$ comparisons.
 *The Holm-Sidak calculation of $P_{\text{crit}} = 1 - (1 - \alpha_T)^{1/(k-j+1)}$ gives .0170, .0253, and .0500.

with the largest magnitude of t , Ordinary Control versus Cell Phone, has a t equal to 13.503. For 21 degrees of freedom, this value of t is associated with $P < .001$, so we reject the null hypothesis that these two samples were drawn from populations with the same mean sperm motility (Table 4-3).

Because the null hypothesis was rejected at this first step, we proceed to the next step, $j = 2$, using the rejection criterion that the P value associated with the second t is smaller than $.05/(3 - 2 - 1) = .0250$. The t value for this second test, 8.250, is associated with $P < .001$, which is smaller than .0250, so we reject the null hypothesis that the Stress Control and Cell Phone samples were drawn from populations with the same mean sperm motility and proceed to the third comparison.

For the third, and final, comparison, $j = 3$, we use the rejection criterion that the P value associated with the second t is smaller than $.05/(3 - 3 - 1) = .05$, which means that we do no adjustment at all for the final comparison. The t value for this third test, 6.752, is associated with $P < .001$, which is smaller than .050, so we reject the null hypothesis that the Ordinary Control and Stress Control samples were drawn from populations with the same mean sperm motility and are finished.

As when we used the Bonferroni t test, all three experimental groups differed from each other. If, however, any of the comparisons had been associated with P values larger than the appropriate P_{crit} , we would have stopped the test and declared that all subsequent comparisons nonsignificant.

Despite reaching the same conclusion that we did when using the single-step Bonferroni t test, you can see by comparing the P_{crit} values in Tables 4-2 and 4-3 that the progressively less stringent requirement for rejecting the null hypothesis with the Holm test it becomes easier to

reject the null hypothesis for all but the first comparison compared with the Bonferroni procedure.

The Holm-Sidak t Test

As noted earlier, the Bonferroni inequality, which forms the basis for the Bonferroni t test and, indirectly, the Holm test, gives a reasonable approximation for the total risk of a false-positive in a family of k comparisons when the number of comparisons is not too large, around 3 or 4. The actual probability of at least one false-positive conclusion (when the null hypothesis of no difference is true) is given by the formula

$$\alpha_T = 1 - (1 - \alpha)^k$$

When there are $k = 3$ comparisons, each done at the $\alpha = 0.05$ level, the Bonferroni inequality says that the total risk of at least one false-positive is less than $k\alpha = 3 \times 0.05 = .150$. This probability is reasonably close to the actual risk of at least one false-positive statement given by the equation above, $1 - (1 - 0.05)^3 = .143$. As the number of comparisons increases, the Bonferroni inequality more and more overestimates the true false-positive risk. For example, if there are $k = 6$ comparisons, $k\alpha = 6 \times 0.05 = .300$ compared with the actual probability of at least one false-positive of .265, nearly 10% lower. If there are 12 comparisons, the Bonferroni inequality says that the risk of at least one false-positive is below $12 \times 0.05 = .600$, 25% above the true risk of .460. Table 4-4 gives the Holm-Sidak critical P values for various numbers of comparisons.

The *Holm-Sidak corrected t test*, or *Holm-Sidak t test*, is a further refinement of the Holm corrected t test that is based on the exact formula for α_T rather than the Bonferroni inequality. The Holm-Sidak corrected t test works just like the Holm corrected t test, except that the criteria for rejecting the j th hypothesis test in an ordered

■ TABLE 4-4. Holm-Sidak Critical *P* Values for Individual Comparisons to Maintain a 5% Family Error Rate ($\alpha_T = .05$)

Comparison Number (<i>j</i>)	Total Number of Comparisons (<i>k</i>)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.0500	.0253	.0170	.0127	.0102	.0085	.0073	.0064	.0057	.0051	.0047	.0043	.0039	.0037	.0034
2		.0500	.0253	.0170	.0127	.0102	.0085	.0073	.0064	.0057	.0051	.0047	.0043	.0039	.0037
3			.0500	.0253	.0170	.0127	.0102	.0085	.0073	.0064	.0057	.0051	.0047	.0043	.0039
4				.0500	.0253	.0170	.0127	.0102	.0085	.0073	.0064	.0057	.0051	.0047	.0043
5					.0500	.0253	.0170	.0127	.0102	.0085	.0073	.0064	.0057	.0051	.0047
6						.0500	.0253	.0170	.0127	.0102	.0085	.0073	.0064	.0057	.0051
7							.0500	.0253	.0170	.0127	.0102	.0085	.0073	.0064	.0057
8								.0500	.0253	.0170	.0127	.0102	.0085	.0073	.0064
9									.0500	.0253	.0170	.0127	.0102	.0085	.0073
10										.0500	.0253	.0170	.0127	.0102	.0085
11											.0500	.0253	.0170	.0127	.0102
12												.0500	.0253	.0170	.0127
13													.0500	.0253	.0170
14														.0500	.0253
15															.0500

$P_{\text{crit}} = 1 - (1 - \alpha_T)^{1/(k-j+1)}$.

sequence of k tests is an uncorrected P value below $1 - (1 - \alpha_T)^{1/(k-j+1)}$ rather than the $\alpha_T/(k - j + 1)$ used in the Holm test. This further refinement makes the Holm-Sidak test slightly more powerful than the Holm test.

The differences between the Holm and Holm-Sidak corrections are small. For example, if there are $k = 20$ comparisons, the differences between the resulting threshold values for P are in the fourth decimal place. To illustrate this difference, the Holm-Sidak value of P_{crit} for the first comparison ($j = 1$) of the $k = 3$ pairwise comparisons in the cell phone example in Box 4-2 to control the family error rate α_T at .05 is $1 - (1 - \alpha_T)^{1/(k-j+1)} = 1 - (1 - .05)^{1/(3-1+1)} = 1 - .95^{1/3} = 1 - .95^{.3333} = 1 - .983 = .0170$, slightly larger than the .0167 for the Holm critical value.

Box 4-2 shows all pairwise comparisons for sperm motility in the four groups of cell phone users that we analyzed in Chapter 3 using all three multiple comparison methods discussed in this chapter. In this case, we conclude that sperm motility in the Control and Low Use (<2 hour/day) are not significantly different, but that this subset of two groups differs significantly from the Medium Use (2 to 4 hour/day) use group, which also differs from the High Use (>4 hour/day) group. The fact that many, if not most of the men in the High Use group in the study by Fejes and colleagues (defined as using cell phones more than 1 hour/day) were probably in the Low Use group in the study by Agarwal and colleagues. (<2 hour/day) may explain why Fejes and colleagues (Box 4-1) did not find a significant difference in overall sperm motility—as opposed to rapid sperm motility—associated with cell phone use. In this case, all three multiple comparison procedures yield the same conclusions.

Note, however, that the critical values of P are larger for the Holm method than the Bonferroni method, and the Holm-Sidak method are larger than for the Holm method, demonstrating the progressively less conservative standard for the three methods. Because of the improved power while controlling the overall false-positive error rate for the family of comparisons at the desired level, we recommend the Holm-Sidak t test over the Bonferroni t test for multiple comparisons following a positive result in analysis of variance.

■ MULTIPLE COMPARISONS AGAINST A SINGLE CONTROL

In addition to all pairwise comparisons, the need sometimes arises to compare the values of multiple treatment groups to a single control group. One alternative would be

to use Bonferroni, Holm, or Holm-Sidak t tests to do all pairwise comparisons, then only consider the ones that involve the control group. The problem with this approach is that it requires many more comparisons than are actually necessary, with the result that each individual comparison is done much more conservatively than is necessary based on the actual number of comparisons of interest. We can use these methods just as before for multiple comparisons against a single control group by reducing the number of comparisons, k , accordingly. As with all pairwise multiple comparisons, use these tests *after* finding significant differences among all the groups with an analysis of variance.

For example, if we only wanted to compare the low, medium, and high cell phone users against the Control nonusers in the example we just discussed, we would only have to account for $k = 3$ comparisons (Table 4-5), as opposed to the 6 we had to allow for when doing all pairwise comparisons (Box 4-1). We compute the values of t and the associated P values, just as before. The critical values of P_{crit} needed to reject the null hypothesis of no difference, however, are computed based on the smaller number of comparisons, k , so are larger than when doing all pairwise comparisons with the same family error rate (compare the values of P_{crit} for all three multiple comparison tests in Box 4-1 and Table 4-5). All three multiple comparison procedures show that the High and Medium, but not the Low users have significantly different levels of sperm motility than the Control nonusers. *No statement can be made about the comparison of low, medium, and high users against each other.*

■ THE MEANING OF P

Understanding what P means requires understanding the logic of statistical hypothesis testing. For example, suppose an investigator wants to test whether or not a drug alters body temperature. The obvious experiment is to select two similar groups of people, administer a placebo to one and the drug to the other, measure body temperature in both groups, then compute the mean and standard deviation of the temperatures measured in each group. The mean responses of the two groups will probably be different, regardless of whether the drug has an effect or not for the same reason that different random samples drawn from the same population yield different estimates for the mean. Therefore, the question becomes: Is the observed difference in mean temperature of the two

BOX 4-2 • All Pairwise Multiple Comparisons for Effects of Cell Phone Use on Human Sperm Motility

To test all pairwise comparisons for the data in Table 3-2, we first compute the t test statistics for all the six comparisons using the within-groups variance estimate, $s_{\text{wit}}^2 = 119.3\%$ and associated degrees of freedom, $v_d = 156$, from the analysis of variance we completed in Chapter 3. Comparing the low user (<2 h/d) with control (no cell phone use) yields

$$t_{\text{low vs. control}} = \frac{\bar{X}_{\text{low}} - \bar{X}_{\text{control}}}{\sqrt{\frac{s_{\text{wit}}^2}{n_{\text{low}}} + \frac{s_{\text{wit}}^2}{n_{\text{control}}}}} = \frac{65 - 68}{\sqrt{\frac{119.3}{40} + \frac{119.3}{40}}} = -1.228$$

Likewise, comparing the high user (>4 h/d) and medium user (2 to 4 h/d) yields,

$$t_{\text{high vs. medium}} = \frac{\bar{X}_{\text{high}} - \bar{X}_{\text{medium}}}{\sqrt{\frac{s_{\text{wit}}^2}{n_{\text{high}}} + \frac{s_{\text{wit}}^2}{n_{\text{medium}}}}} = \frac{45 - 55}{\sqrt{\frac{119.3}{40} + \frac{119.3}{40}}} = -3.685$$

We compute the t values for the four other comparisons, similarly, then list all the comparisons in declining order of magnitude of the associated t values (without regard to sign), as listed in the second column of the table below.

Pairwise Comparisons for Human Cell Phone Use ($\alpha_T = .05$)

Comparison	t	P	j	Bonferroni		Holm		Holm-Sidak	
				$P_{\text{crit}} \alpha_T/k$	$P < P_{\text{crit}}?$	$P_{\text{crit}} \alpha_T/(k - j + 1)$	$P < P_{\text{crit}}?$	$1 - (1 - \alpha_T)^{1/(k-j+1)}$	$P < P_{\text{crit}}?$
High vs. Control	9.417	<.001	1	.0083	Yes	.0083	Yes	.0085	Yes
Medium vs. Low	8.189	<.001	2	.0083	Yes	.0100	Yes	.0102	Yes
Medium vs. Control	5.732	<.001	3	.0083	Yes	.0125	Yes	.0127	Yes
Medium vs. Low	4.504	<.001	4	.0083	Yes	.0167	Yes	.0170	Yes
High vs. Medium	3.685	<.001	5	.0083	Yes	.0250	Yes	.0253	Yes
Low vs. Control	1.228	>.10	6	.0083	No	.0050	No	.0500	No

$v_d = 156$.

Next, look up the P values for each t in Table 4-1 using the denominator degrees of freedom from the analysis of variance (third column in the table above) and compare these P values with the critical value, P_{crit} , for the multiple comparison procedure.

For the Bonferroni corrected t test, this critical value is just the family error rate, $\alpha_T = .05$, divided by the total number of comparisons, $k = 6$, for all the comparisons (the fifth column in the table).

For the Holm corrected t test, we begin with a critical value of P as small as for the Bonferroni t test, but make each critical value less conservative (larger) as the differences in the sample means (quantified using the corresponding t) get smaller, until for the last comparison, we are using an unadjusted value of P , equal to the family error rate, in this case .05. For example, for the second comparison, $j = 2$, and $P_{\text{crit}} = \alpha_T/(k - j + 1) = .05/(6 - 2 + 1) = .05/5 = .0100$.

For the Holm-Sidak corrected t test, as with the Holm corrected t test, the first comparison is done with the full Bonferroni correction, and we make the critical values less conservative with each subsequent comparison, we make the critical value of P larger with each subsequent comparison, but use the formula that better models the actual accumulation of false positive risks. In this case, for the second comparison, when $j = 2$, $P_{\text{crit}} = 1 - (1 - \alpha_T)^{1/(k-j+1)} = 1 - (1 - .05)^{1/(6-2+1)} = 1 - .95^{1/5} = 1 - .95^{.2} = 1 - .9898 = .0102$.

To determine whether each pair of means differs significantly from each other, compare the P value associated with the t test with the corresponding P_{crit} adjusted for the multiple comparison test. In this case, the P values are smaller than P_{crit} all the pairs of means except the Control and Low Use groups for all three multiple comparison procedures. Thus, we conclude that sperm motility is different between all the sample groups, except for the Control and Low Use groups, which are not detectably different.

■ **TABLE 4-5. Multiple Comparisons Against a Single Control Group for Human Sperm Motility and Cell Phone Use ($\alpha_T = .05$)**

Comparison	<i>t</i>	<i>P</i>	<i>j</i>	Bonferroni		Holm		Holm-Sidak	
				$P_{crit} = \alpha_T/k$	$P < P_{crit}?$	$P_{crit} = \alpha_T/(k - j + 1)$	$P < P_{crit}?$	$1 - (1 - \alpha_T)^{1/(k-j+1)}$	$P < P_{crit}?$
High vs. Control	9.417	<.001	1	.0167	Yes	.0167	Yes	.0167	Yes
Medium vs. Control	5.732	<.001	2	.0167	Yes	.0250	Yes	.0253	Yes
Low vs. Control	1.288	>.10	3	.0167	No	.0050	No	.0500	No

$v_d = 156.$

groups likely to be due to random variation associated with the allocation of individuals to the two experimental groups or due to the drug?

To answer this question, statisticians first quantify the observed difference between the two samples with a single number, called a *test statistic*, such as *F* or *t*. These statistics, like most test statistics, have the property that the greater the difference between the samples, the greater their value. If the drug has no effect, the test statistic will be a small number. But what is “small”?

To find the boundary between “small” and “big” values of the test statistic, statisticians assume that the drug does *not* affect temperature (the *null hypothesis*). If this assumption is correct the two groups of people are simply random samples from a single population, all of whom received a placebo (because the drug is, in effect, a placebo). Now, in theory, the statistician repeats the experiment using all possible samples of people and computes the test statistic for each hypothetical experiment. Just as random variation produced different values for means of different samples, this procedure will yield a range of values for the test statistic. Most of these values will be relatively small, but sheer bad luck requires that there be a few samples that are not representative of the entire population. These samples will yield relatively large values of the test statistic *even if the drug had no effect*. This exercise produces only a few of the possible values of the test statistic, say 5% of them, above some cutoff point. The test statistic is “big” if it is larger than this cutoff point.

Having determined this cutoff point, we execute an experiment on a drug with unknown properties and compute the test statistic. It is “big.” Therefore, we conclude that *there is less than a 5% chance of observing data which*

led to the computed value of the test statistic on the assumption that the drug has had no effect was true. Traditionally, if the chances of observing the computed test statistic when the intervention has no effect are below 5%, one rejects the working assumption that the drug has no effect and asserts that the drug *does* have an effect. There is, of course, a chance that this assertion is wrong: about 5%. This 5% is known as the *P value* or *significance level*.

Precisely,

The P value is the probability of obtaining a value of the test statistic as large as or larger than the one computed from the data when in reality there is no difference between the different treatments.

As a result of this logic, if we are willing to assert a difference when $P < .05$, we are tacitly agreeing to accept the fact that, over the long run, we expect 1 out of every 20 assertions of a difference to be wrong.

Statistical versus Real (Clinical) Thinking

As we have said several times, statistical hypothesis testing as presented in this book and generally practiced is an argument by contradiction. One begins with the null hypothesis of no difference and estimates the probability of obtaining the observed data assuming that the null hypothesis is true. If that probability is sufficiently low, we reject the null hypothesis. Even though this formalism is widely used, the simple fact is that investigators rarely begin a study actually *expecting* the null hypothesis to be true. Quite the contrary, generally one expects that some alternative hypothesis—that the treatment or observational factor being studied—*does* have an effect.

Indeed, in terms of practical thinking, if the results of the study reject the null hypothesis of no effect, it actually reinforces the “real” hypothesis that there was an effect, which is what motivated the study in the first place. If, on the other hand, you fail to reject the null hypothesis of no effect, that fact is evidence that the “real” hypothesis is not correct. This use of information in an incremental way, which involves beginning with some prior expectation of what the underlying relationship between the treatment (or observational factor) and the outcome is, then modifying that belief on the basis of the experimental data is how scientific and clinical decision making is actually done.

There is a branch of statistical reasoning called *Bayesian decision making*, based on simple probability calculations known as *Bayes’ rule*,* that allows you to use the results of an experiment to modify, in a quantitative way, your prior expectations of the relationship you are studying.

Bayes’ rule allows you to begin with a *prior* distribution of possible outcomes (each with a probability attached to it, much like the *F* and *t* sampling distributions we have already discussed) then mathematically modify that distribution based on the information obtained in your study to obtain your *posterior* distribution of probabilities associated with different possible outcomes. Indeed, at a qualitative level, that is the process that people use to integrate new information in making decisions—be they scientific, clinical, or personal.

Many statisticians,[†] especially those concerned with clinical decision making, have argued that the simple null hypothesis approach to statistical decision making both oversimplifies the process of using data to make clinical

*Bayes’ Rule states:

$$\left(\begin{array}{c} \text{Posterior odds} \\ \text{of null hypothesis} \end{array} \right) = \left(\begin{array}{c} \text{Prior odds} \\ \text{of null hypothesis} \end{array} \right) \times \frac{\text{Pr}(\text{data, given the null hypothesis})}{\text{Pr}(\text{data, given the alternative hypothesis})}$$

where *Pr* means the probability of the stated situation. For a detailed discussion of the application of this formulation of Bayes’ Rule to biomedical data, see Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. *Ann Intern Med.* 1999;130:1005–1013.

[†]For a discussion of the Bayesian approach, with a comparison to the frequentist approach used in this book and several clinical examples, see Browner WS, Newman TB. Are all significant *P* values created equal? The analogy between diagnostic tests and clinical research. *JAMA.* 1987;257:2459–2463; Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. *Ann Intern Med.* 1999;130:1005–1013; Diamond GA, Kaul S. Bayesian approaches to the analysis and interpretation of clinical megatrends. *J Am Coll Cardiol.* 2004;43:1929–1939.

and scientific decisions and leads to being overly reluctant to conclude that the treatment actually had an effect.

There are two reasons for this view. First, traditional statistical hypothesis testing based on the null hypothesis of no effect is equivalent to saying that at the outset of the study you do not believe that there is any evidence to support the possibility that the treatment actually had an effect, which is, as discussed above, rarely the case. Second, each hypothesis is tested without taking in to account anything else you know about the likely effects of the intervention. These two factors combine to lead you to implicitly underestimate the prior probability that the treatment has an effect, which makes it harder to conclude that there is an effect than the data may warrant.

They are correct. Why, then, do people persist in using the classic approach to statistical decision making described in this book?

The primary reason is the difficulty in obtaining good estimates of the prior probabilities of the possible outcomes before the experiment was conducted. Indeed, despite repeated entreaties to use Bayesian decision making by its enthusiasts, they can point to few examples where it has been used in routine clinical or scientific research because of the difficulties in obtaining meaningful prior probability distributions.

Nevertheless, it is worth keeping in mind this process and recognizing that the results of classic statistical hypothesis testing—embodied as the *P* value—need to be integrated into the larger collection of knowledge that creators and consumers of scientific and clinical results possess in order to further refine their understanding of the problems at hand. From this perspective, the *P* value is not the arbiter of truth but rather an assistant in making evolving judgments as to what the truth is.

Why *P* < .05?

The convention of considering a difference “statistically significant” when *P* < .05 is widely accepted. In fact, it came from an arbitrary decision by one person, Ronald A. Fisher, who invented much of modern parametric statistics (including the *F* statistic, which is named for him). In 1926, Fisher published a paper[†] describing how to assess

[†]Fisher RA. The arrangement of field experiments. *J Min Agr.* 1926;33:503–513. For a discussion of this paper in its historical context, including evidence that the logic of hypothesis testing dates back to Blaise Pascal and Pierre Fermat, in 1654, see Cowles M, Davis C. On the origins of the .05 level of statistical significance. *Am Psychol.* 1982;37:533–558.

whether adding manure to a field would increase crop yields, which introduced the idea of statistical significance and established the 5% standard. He said:

To an acre of ground the manure is applied; a second acre, sown with similar seed and treated in all other ways like the first, receives none of the manure. When the produce is weighed, it is found that the acre which received the manure has yielded a crop larger indeed by, say, 10 percent. The manure has scored a success, but the confidence with which such a result should be received by the purchasing public depends wholly on the manner in which the experiment was carried out.

First, if the experimenter could say that in twenty years of experience with uniform treatment the difference in favour of the acre treated with manure had never before touched 10%, the evidence would have reached a point which may be called the verge of significance; for it is convenient to draw the line at about the level at which we can say: "Either there is something in the treatment, or a coincidence has occurred such as does not occur more than one in twenty trials." This level, which we may call the 5% point, would be indicated, though very roughly, by the greatest chance deviation observed in twenty successive trials. To locate the 5% point with any accuracy we should need about 500 years' experience, for we could then, supposing no progressive changes in fertility were in progress, count out the 25 largest deviations and draw the line between the 25th and 26th largest deviation. If the difference between the two acres in our experimental year exceeded this value, we should have reasonable grounds for calling this value significant.

If one in 20 does not seem high enough odds, we may, if we prefer it, draw the line at 1 in 50 (the 2% point) or 1 in 100 (the 1% point.) *Personally, the writer prefers to set a low standard of significance at the 5% point, and ignore entirely all results which fails to reach this level.*

Although $P < .05$ is widely accepted, and you will certainly not generate controversy if you use it, a more sensible approach is to consider the P value in making decisions about how to interpret your results without slavishly considering 5% a rigid criterion for "truth."

It is commonly believed that the P value is the probability of making a mistake. There are obviously two ways

an investigator can reach a mistaken conclusion based on the data, reporting that the treatment had an effect when in reality it did not or reporting that the treatment did not have an effect when in reality it did. As noted above, the P value only quantifies the probability of making the first kind of error (called a *Type I* or α error), that of erroneously concluding that the treatment had an effect when in reality it did not. It gives no information about the probability of making the second kind of error (called a *Type II* or β error), that of concluding that the treatment had no effect when in reality it did. Chapter 6 discusses how to estimate the probability of making Type II errors.

■ PROBLEMS

4-1 In the randomized controlled trial of the use of a cannabis-based medicinal to treat pain associated with diabetic neuropathy discussed in Chapter 3, the 29 people randomized to the control group had a mean age of 54.4 years old and the 24 people randomized to the treatment group had a mean age of 58.2 years old, with standard deviations of 11.6 and 8.8 years. Was there a detectable difference in the ages of these two groups?

4-2 Hypothermia is problem for extremely low birth weight infants. One idea to help these infants maintain body temperature is to wrap them in polyethylene bags in the delivery room and while they are being transferred to the neonatal intensive care unit. Patrick Carroll and colleagues* reviewed medical records and located 70 infants who were kept warm with polyethylene bags and 70 infants who were kept warm with traditional methods. The skin temperature for the infants who were kept warm with the polyethylene bags was 36°C and for the infants kept warm using traditional techniques was 35°C. The standard deviations for both groups were 1°C. Is there a difference in skin temperature between these two treatment groups?

4-3 In addition to the stair climbing test discussed in Chapter 3, Mark Roig and colleagues also conducted 6 minute walk tests in which they measured how far people could walk (in meters) in 6 minutes to compare the ability of normal people and people with chronic obstructive pulmonary disease (COPD) to exercise. Based on the data in Table 4-6 is there a detectable difference in performance?

*Carroll P, Nankervis CA, Giannone PJ, Cordero L. Use of polyethylene bags in extremely low birth weight infant resuscitation for the prevention of hypothermia. *J Reprod Med.* 2010;55: 9–13.

■ TABLE 4-6. Distance Walked in 6 Minutes (meters)

Control	COPD
619	283
512	402
523	407
586	402
436	340
515	445
562	548
544	344
531	358
534	419
572	393
541	469
551	393
492	420
698	463
700	438
571	428
502	364
557	336
482	256
627	368

4-4 To assess whether providing in-person counseling would increase the use of advance directives in homeless people, John Song and colleagues* recruited 262 volunteers at emergency night shelters and other programs serving homeless people and randomly allocated them to receive in-person counseling or just be provided written materials. The mean age for the 145 people randomized to receive in-person counseling was 43.1 years and for the 117 people randomized to receive written materials was 43.3 years. The standard errors of the mean for the two groups were .87 and .96, respectively. Is there a difference in the ages of the two study groups?

*Song J. Effect of an end-of life planning intervention on the completion of advance directives in homeless persons. *Ann Intern Med.* 2010;153:76–84.

4-5 Rework Problems 3-1, 3-3, and 3-5 using the t test. What is the relationship between the value of t computed here and the value of F computed for these data in Chapter 3?

4-6 Problem 3-2 presented the data that White and Froeb collected on the lung function of nonsmokers working in smoke-free environments, nonsmokers working in smoky environments, and smokers of various intensity. Analysis of variance revealed that these data were inconsistent with the hypothesis that the lung function was the same in all these groups. Isolate the various subgroups with similar lung function. What does this result mean in terms of the original question they posed: Does chronic exposure to other people's smoke affect the health of healthy adult nonsmokers?

4-7 Directly test the limited hypothesis that exposure to other people's smoke affects the health of healthy nonsmokers by comparing each group of involuntary smokers and active smokers with the nonsmokers working in a clean environment as the control group. Use the data from Problem 3-2.

4-8 Problem 3-4 led to the conclusion that there were differences in sperm viability among men with different levels of cell phone use. What are the detectable subgroups in this response? Use a Holm-Sidak t test.

4-9 What conclusions would you draw if you were only interested in whether sperm viability among men with different levels of cell phone use were significantly different from men who did not use cell phones at all?

4-10 In Problem 3-6 you determined there was a difference in burnout among staffs in different patient care units. Isolate these differences and discuss them.

4-11 In a test of significance, the P value of the test statistic is .063. Are the data statistically significant at

- both $\alpha = .05$ and $\alpha = .01$ levels?
- $\alpha = .05$ level but not at $\alpha = .01$ level?
- $\alpha = .01$ level but not at $\alpha = .05$ level?
- neither $\alpha = .05$ nor $\alpha = .01$ levels?

How to Analyze Rates and Proportions

The statistical procedures developed in Chapters 2 to 4 are appropriate for analyzing the results of experiments in which the variable of interest is measured on an interval scale, such as blood pressure, urine production, or length of hospital stay. Much of the information physicians, nurses, other health professionals, and medical scientists use cannot be measured on interval scales. For example, an individual may be male or female, dead or alive, or Caucasian, African American, Hispanic, or Asian. These variables are measured on *nominal scales*, in which there is no arithmetic relationship between the different classifications. We now develop the statistical tools necessary to describe and analyze such information.

It is easy to describe things measured on a nominal scale: simply count the number of patients or experimental subjects with each condition and (perhaps) compute the corresponding percentages.

For example, John Song and colleagues* wanted to study whether or not providing homeless people with personal counseling on end-of-life care and advanced directives would lead more of them to complete such directives. (This question had been studied among insured general adult populations, but not among the homeless, who have more health problems and less access to stable health care relationships.) To investigate this question, they recruited

people at emergency night shelters, 24-hour shelters, a day program and treatment programs. They conducted an experiment in which volunteers were randomly assigned to either receive written material on advance directives or invited to attend a 1-hour in-person counseling session on advance directives. The outcome of the study was whether the people returned a completed advance directive within 3 months. Among the 262 people who participated in the study 37.9% of the people who received the in-person counseling returned the advanced directives within 3 months, compared with 12.8% of the people who were just given written instructions. Is this difference likely to be a real effect of the counseling or simply a reflection of random sampling variation?

To answer this and other questions about nominal data, we must first invent a way to estimate the precision with which percentages based on limited samples approximate the true rates that would be observed if we could examine the entire population, in this case, *all* homeless people. We will use these estimates to construct statistical procedures to test hypotheses.

■ BACK TO MARS

Before we can quantify the certainty of our descriptions of a population on the basis of a limited sample, we need to know how to describe the population itself. Since we have already visited Mars and met all 200 Martians (in Chapter 2), we will continue to use them to develop ways to describe populations. In addition to measuring the Martians' heights, we noted that 50 of them were left-footed

*Song J, Ratner ER, Wall HM, Bartels DM, Ulvestad N, Petroskas D, West M, Weber-Main AM, Grengs L, Gelberg L. Effect of an end-of-life planning intervention on the completion of advance directives in homeless persons. *Ann Intern Med.* 2010;153:76–84.

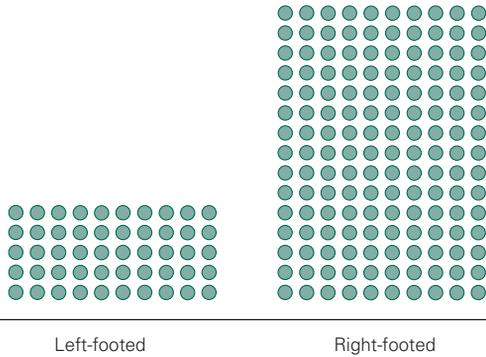


FIGURE 5-1. Of the 200 Martians 50 are left-footed, and the remaining 150 are right-footed. Therefore, if we select one Martian at random from this population, there is a $p_{\text{left}} = 50/200 = 0.25 = 25\%$ chance it will be left-footed.

and the remaining 150 were right-footed. Figure 5-1 shows the entire population of Mars divided according to footedness. The first way in which we can describe this population is by giving the *proportion* p of Martians who are in each class. In this case, $p_{\text{left}} = 50/200 = 0.25$ and $p_{\text{right}} = 150/200 = 0.75$. Since there are only two possible classes, notice that $p_{\text{right}} = 1 - p_{\text{left}}$. Thus, whenever there are only two possible classes and they are mutually exclusive, we can completely describe the division in the population with the single parameter p , the proportion of members with one of the attributes. The proportion of the population with the other attribute is *always* $1 - p$.

Note that p also is the *probability* of drawing a left-footed Martian if one selects one member of the population at random.

Thus p plays a role exactly analogous to that played by the population mean μ in Chapter 2. To see why, suppose we associate the value $X = 1$ with each left-footed Martian and a value of $X = 0$ with each right-footed Martian. The mean value of X for the population is

$$\begin{aligned} \mu &= \frac{\sum X}{N} = \frac{1+1+\cdots+1+0+0+\cdots+0}{200} \\ &= \frac{50(1)+150(0)}{200} = \frac{50}{200} = 0.25 \end{aligned}$$

which is p_{left} .

This idea can be generalized quite easily using a few equations. Suppose M members of a population of N individuals have some attribute and the remaining $N - M$ members of the population do not. Associate a value of $X = 1$ with the population members having the attribute

and a value of $X = 0$ with the others. The mean of the resulting collection of numbers is

$$\mu = \frac{\sum X}{N} = \frac{M(1) + (N - M)(0)}{N} = \frac{M}{N} = p$$

the proportion of the population having the attribute.

Since we can compute a mean in this manner, why not compute a standard deviation in order to describe variability in the population? Even though there are only two possibilities, $X = 1$, and $X = 0$, the amount of variability will differ, depending on the value of p . Figure 5-2 shows three more populations of 200 individuals each. In Figure 5-2A only 10 of the individuals are left-footed; it exhibits less variability than the population shown in Figure 5-1. Figure 5-2B shows the extreme case in which half the members of the population fall into each of the two classes; the variability is greatest. Figure 5-2C shows the other extreme; all the members fall into one of the two classes, and there is no variability at all.

To quantify this subjective impression, we compute the standard deviation of the 1s and 0s associated with each member of the population when we computed the mean. By definition, the population standard deviation is

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$$

$X = 1$ for M members of the population and 0 for the remaining $N - M$ members, and $\mu = p$; therefore

$$\begin{aligned} \sigma &= \sqrt{\frac{(1-p)^2 + (1-p)^2 + \cdots + (1-p)^2}{N} + \frac{(0-p)^2 + (0-p)^2 + \cdots + (0-p)^2}{N}} \\ &= \sqrt{\frac{M(1-p)^2 + (N-M)p^2}{N}} = \sqrt{\frac{M}{N}(1-p)^2 + \left(1 - \frac{M}{N}\right)p^2} \end{aligned}$$

But since $M/N = p$ is the proportion of population members with the attribute,

$$\sigma = \sqrt{p(1-p)^2 + (1-p)p^2} = \sqrt{[p(1-p) + p^2](1-p)}$$

which simplifies to

$$\sigma = \sqrt{p(1-p)}$$

This equation for the population standard deviation produces quantitative results that agree with the qualitative impressions we developed from Figures 5-1 and 5-2.

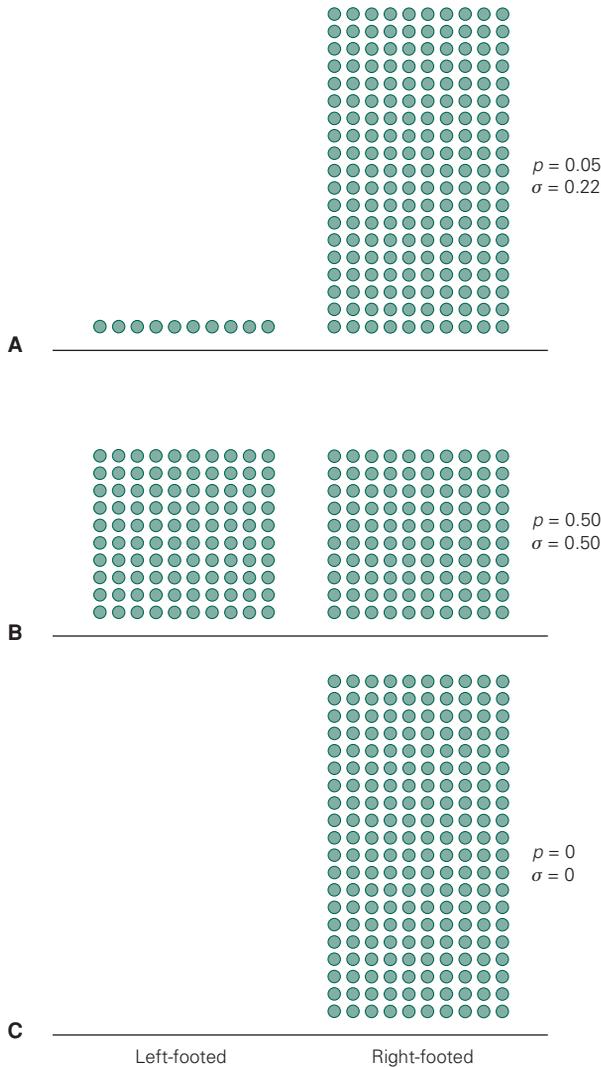


FIGURE 5-2. This figure illustrates three different populations, each containing 200 members but with different proportions of left-footed members. The standard deviation, $\sigma = \sqrt{p(1 - p)}$ quantifies the variability in the population. **(A)** When most of the members fall in one class, σ is a small value, 0.2, indicating relatively little variability. **(B)** In contrast, if half the members fall into each class, σ reaches its maximum value of .5, indicating the maximum possible variability. **(C)** At the other extreme, if all members fall into the same class, there is no variability at all and $\sigma = 0$.

As Figure 5-3 shows, $\sigma = 0$ when $p = 0$ or $p = 1$, that is, when all members of the population either do or do not have the attribute, and σ is maximized when $p = .5$, that is, when any given member of the population is as likely to have the attribute as not.

Since σ depends only on p , it really does not contain any additional information (in contrast to the mean and standard deviation of a normally distributed variable, where μ and σ provide two independent pieces of information). It will be most useful in computing a standard error associated with estimates of p based on samples drawn at random from populations such as those shown in Figures 5-1 or 5-2.

ESTIMATING PROPORTIONS FROM SAMPLES

Of course, if we could observe all members of a population, there would not be any statistical question. In fact, all we ever see is a limited, hopefully representative, sample drawn from that population. How accurately does the proportion of members of a sample with an attribute reflect the proportion of individuals in the population with that attribute? To answer this question, we do a sampling experiment, just as we did in Chapter 2 when we asked how well the sample mean estimated the population mean.

Suppose we select 10 Martians at random from the entire population of 200 Martians. Figure 5-4A shows which Martians were drawn; Figure 5-4B shows all the information the investigators who drew the sample would have. Half the Martians in the sample are left-footed and half are right-footed. Given only this information, one would probably report that the proportion of left-footed Martians is 0.5%, or 50%.

Of course, there is nothing special about this sample, and one of the four other random samples shown in Figure 5-5 could just as well have been drawn, in which case the investigator would have reported that the proportion of left-footed Martians was 30%, 30%, 10%, or 20%, depending on which random sample happened to be drawn. In each case, we have computed an estimate of the population proportion p based on a sample. Denote this estimate \hat{p} . Like the sample mean, the possible values of \hat{p} depend on both the nature of the underlying population and the specific sample that is drawn. Figure 5-6 shows the five values of \hat{p} computed from the specific samples in Figures 5-4 and 5-5 together with the results of drawing another 20 random samples of 10 Martians each. Now we change our focus from the population of Martians to the population of all values of \hat{p} computed from random samples of 10 Martians each. There are more than 10^{16} such samples with their corresponding estimates \hat{p} of the value of p for the population of Martians.

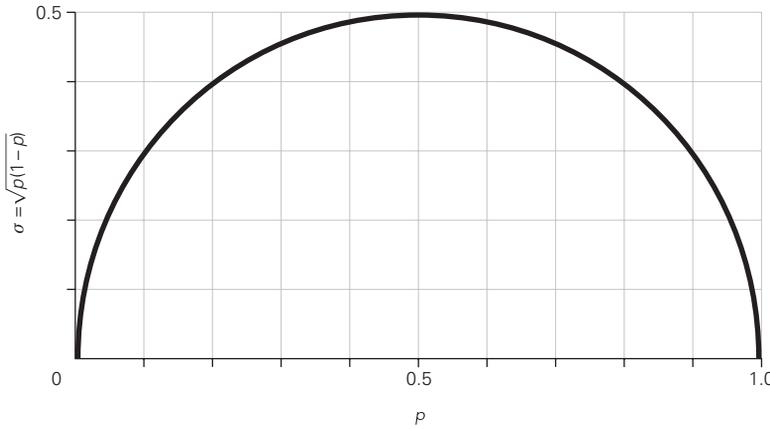


FIGURE 5-3. The relationship between the standard deviation of a population divided into two categories varies with p , the proportion of members in one of the categories. There is no variation if all members are in one category or the other (so $\sigma = 0$ when $p = 0$ or 1) and maximum variability when a given member is equally likely to fall in one class or the other ($\sigma = 0.5$ when $p = 0.5$).

The mean estimate of \hat{p} for the 25 samples of 10 Martians each shown in Figure 5-6 is 30%, which is remarkably close to the true proportion of left-footed Martians in the population (25% or 0.25). There is some variation in the estimates. To quantify the variability in the possible values of \hat{p} , we compute the *standard deviation* of values of \hat{p} computed from random samples of 10 Martians each. In this case, it is about 14% or 0.14. This number describes

the variability in the population of all possible values of the proportion of left-footed Martians computed from random samples of 10 Martians each.

Does this sound familiar? It should. It is just like the standard error of the mean. Therefore, we define the *standard error of the proportion* to be the standard deviation of the population of all possible values of the proportion computed from samples of a given size. Just as with the standard error of the mean

$$\sigma_{\hat{p}} = \frac{\sigma}{\sqrt{n}}$$

in which $\sigma_{\hat{p}}$ is the standard error of the proportion, σ is the standard deviation of the population from which the sample was drawn, and n is the sample size. Since

$$\sigma = \sqrt{p(1-p)}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

We estimate the standard error from a sample by replacing the true value of p in this equation with our estimate \hat{p} obtained from the random sample. Thus,

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The standard error is a very useful way to describe the uncertainty in the estimate of the proportion of a population with a given attribute because the central-limit theorem (Chapter 2) also leads to the conclusion that the distribution of \hat{p} is approximately normal, with mean p

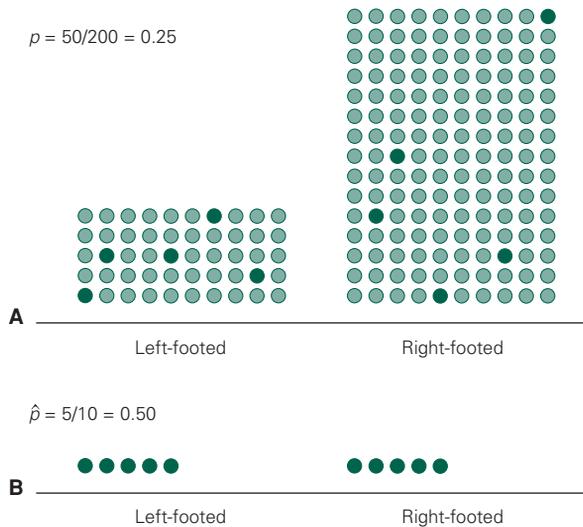


FIGURE 5-4. Panel **A** shows one random sample of 10 Martians selected from the population in Figure 5-1; panel **B** shows what the investigator would see. Since this sample included five left-footed Martians and five right-footed Martians, the investigator would estimate the proportion of left-footed Martians to be $\hat{p}_{\text{left}} = 5/10 = .5$, where the circumflex denotes an estimate.



FIGURE 5-5. Four more random samples of 10 Martians each, together with the sample as it would appear to the investigator. Depending which sample happened to be drawn, the investigator would estimate the proportion of left-footed Martians to be 30%, 30%, 10%, or 20%.

and standard deviation $\sigma_{\hat{p}}$ for large enough sample sizes. On the other hand, this approximation fails for values of p near 0 or 1 or when the sample size n is small. When can you use the normal distribution? Statisticians

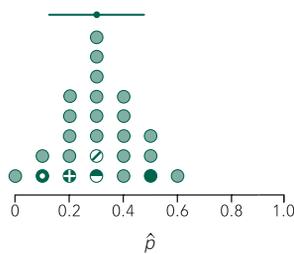


FIGURE 5-6. There will be a distribution of estimates of the proportion of left-footed Martians \hat{p}_{eff} depending on which random sample the investigator happens to draw. This figure shows the five specific random samples drawn in Figures 5-4 and 5-5 together with 20 more random samples of 10 Martians each. The mean of the 25 estimates of p and the standard deviation of these estimates are also shown. The standard deviation of this distribution is the standard error of the estimate of the proportion $\sigma_{\hat{p}}$; it quantifies the precision with which \hat{p} estimates p .

have shown that it is adequate when $n\hat{p}$ and $n(1 - \hat{p})$ both exceed about 5.* Recall that about 95% of all members of a normally distributed population fall within 2 standard deviations of the mean. When the distribution of \hat{p} approximates the normal distribution, we can assert, with about 95% confidence, that the true proportion of population members with the attribute of interest p lies within $2s_{\hat{p}}$ of \hat{p} .

These results provide a framework within which to consider the question we posed earlier in this chapter regarding whether in-person counseling led to higher levels of completing end-of-life advance directives among homeless people. Of the 145 people who received in-person counseling 37.9% completed the advance directives

*When the sample size is too small to use the normal approximation, you need to solve the problem exactly using the binomial distribution (or use a table of exact values). For a discussion of the binomial distribution, see Zar JH. Dichotomous variables. *Biostatistical Analysis*, 5th ed. Upper Saddle River, NJ: Prentice Hall; 2010:chap 24.

and 12.8% of the 117 people who just received written instructions did so. The standard errors of these proportions are

$$s_{\hat{p}_{\text{counsel}}} = \sqrt{\frac{.379(1-.379)}{145}} = .040 = 4.0\%$$

for the people who received counseling and

$$s_{\hat{p}_{\text{paper}}} = \sqrt{\frac{.128(1-.128)}{117}} = .031 = 3.1\%$$

for written instructions. Given that there was a 25.1% difference in the rate that people returned the advance directive, it seems likely that the counseling had an effect beyond just random sampling variation.

Before moving on, we should pause to list explicitly the assumptions that underlie this approach. We have been analyzing what statisticians call *independent Bernoulli trials*, in which

- Each individual trial has two mutually exclusive outcomes.
- The probability p of a given outcome remains constant.
- All the trials are independent.

In terms of a population, we can phrase these assumptions as follows:

- Each member of the population belongs to one of two classes.
- Each member of the sample is selected independently of all other members.

■ HYPOTHESIS TESTS FOR PROPORTIONS

In Chapter 4, the sample mean and standard error of the mean provided the basis for constructing the t test to quantify how compatible observations were with the null hypothesis. We defined the t statistic as

$$t = \frac{\text{Difference of sample means}}{\text{Standard error of difference of sample means}}$$

The role of \hat{p} is analogous to that of the sample mean in Chapters 2 and 4, and we have also derived an expression for the standard error of \hat{p} . We now use the observed proportion of individuals with a given attribute and its standard error to construct a test statistic analogous to t to test the hypothesis that the two samples were drawn from

populations containing the same proportion of individuals with a given attribute.

The test statistic analogous to t is

$$z = \frac{\text{Difference of sample proportions}}{\text{Standard error of difference of sample proportions}}$$

Let \hat{p}_1 and \hat{p}_2 be the observed proportions of individuals with the attribute of interest in the two samples. The standard error is the standard deviation of the population of all possible values of \hat{p} associated with samples of a given size, and since variances of differences add, the standard error of the difference in proportions is

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{s_{\hat{p}_1}^2 + s_{\hat{p}_2}^2}$$

Therefore

$$z = \frac{\hat{p}_1 - \hat{p}_2}{s_{\hat{p}_1 - \hat{p}_2}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{s_{\hat{p}_1}^2 + s_{\hat{p}_2}^2}}$$

If n_1 and n_2 are the sizes of the two samples,

$$s_{\hat{p}_1} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} \quad \text{and} \quad s_{\hat{p}_2} = \sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

then

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{[\hat{p}_1(1-\hat{p}_1)/n_1] + [\hat{p}_2(1-\hat{p}_2)/n_2]}}$$

is our test statistic.

z replaces t because this ratio is approximately normally distributed for large enough sample sizes,* and it is customary to denote a normally distributed variable with the letter z .

Just as it was possible to improve the sensitivity of the t test by pooling the observations in the two sample groups to estimate the population variance, it is possible to increase the sensitivity of the z test for proportions by pooling the information from the two samples to obtain a single estimate of the population standard deviation s . Specifically, if the null hypothesis that the two samples

*The criterion for a large sample is the same as in the last section, namely that $n\hat{p}$ and $n(1-\hat{p})$ both exceed about 5 for both samples. When this is not the case, one should use the *Fisher exact test* discussed later in this chapter.

were drawn from the same population is true, $\hat{p}_1 = m_1/n_1$ and $\hat{p}_2 = m_2/n_2$, in which m_1 and m_2 are the number of individuals in each sample with the attribute of interest, are both estimates of the same population proportion p . In this case, we could consider all the individuals drawn as a single sample of size $n_1 + n_2$ containing a total of $m_1 + m_2$ individuals with the attribute and use this single pooled sample to estimate \hat{p} :

$$\hat{p} = \frac{m_1 + m_2}{n_1 + n_2} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

in which case

$$s = \sqrt{\hat{p}(1 - \hat{p})}$$

and we can estimate

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}} = \sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

Therefore, our test statistic, based on a pooled estimate of the uncertainty in the population proportion, is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})(1/n_1 + 1/n_2)}}$$

Like the t statistic, z will have a range of possible values depending on which random samples happen to be drawn to compute \hat{p}_1 and \hat{p}_2 , even if both samples were drawn from the same population. If z is sufficiently “big” we will conclude that the data are inconsistent with this null hypothesis and assert that there is a difference in the proportions. This argument is exactly analogous to that used to define the critical values of the t for rejecting the hypothesis of no difference. The only change is that in this case we use the standard normal distribution (Fig. 2-5) to define the cutoff values. In fact, the standard normal distribution and the t distribution with an infinite number of degrees of freedom are identical, so we can get the critical values for 5 or 1% confidence levels from the last line in Table 4-1. This table shows that there is less than a 5% chance of z being beyond -1.96 or $+1.96$ and less than a 1% chance of z being beyond -2.58 or $+2.58$ when, in fact, the two samples were drawn from the same population.

The Yates Correction for Continuity

The standard normal distribution only approximates the actual distribution of the z test statistic in a way that yields P values that are always smaller than they should be. Thus,

the results are biased toward concluding that the treatment had an effect when the evidence does not support such a conclusion. The mathematical reason for this problem has to do with the fact that the z test statistic can only take on discrete values, whereas the theoretical standard normal distribution is continuous. To obtain values of the z test statistic which are more compatible with the theoretical standard normal distribution statisticians have introduced the *Yates correction* (or *continuity correction*), in which the expression for z is modified to become

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - \frac{1}{2}(1/n_1 + 1/n_2)}{\sqrt{\hat{p}(1 - \hat{p})(1/n_1 + 1/n_2)}}$$

This adjustment slightly reduces the value of z associated with the data and compensates for the mathematical problem just described.

Effect of Counseling on End-of-Life Planning in Homeless People

We can now formally test the null hypothesis that counseling and just giving homeless people written instructions on end-of-life care leads to the same rate of completing advance directives. (Note that we can say “leads” or “causes” rather than just “is associated with” because this is a randomized experiment, not an observational study.) Since 55 (37.9% of 145) people who received in-person counseling completed the advance directives and 15 (12.8% of 117) people who just received written instructions did so,

$$\hat{p} = \frac{55 + 15}{145 + 117} = .267$$

Since $n\hat{p}$ for the two samples, $.267 \cdot 145 = 38.7$ and $.267 \cdot 117 = 31.2$ both exceed 5, we can use the test described in the last section.* Our z test statistic is therefore

$$\begin{aligned} z &= \frac{\hat{p}_{\text{counsel}} - \hat{p}_{\text{paper}}}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_{\text{counsel}}} + \frac{1}{n_{\text{paper}}} \right)}} \\ &= \frac{.379 - .128}{\sqrt{.267(1 - .267) \left(\frac{1}{145} + \frac{1}{117} \right)}} = 4.565 \end{aligned}$$

* $n(1 - \hat{p})$ also exceeds 5 for both samples because $\hat{p} < .5$, so $n\hat{p} < n(1 - \hat{p})$.

Including the Yates correction, it is

$$z = \frac{|\hat{p}_{\text{counsel}} - \hat{p}_{\text{paper}}| - \frac{1}{2} \left(\frac{1}{n_{\text{counsel}}} + \frac{1}{n_{\text{paper}}} \right)}{\sqrt{\hat{p}(1-\hat{p}) \left(\frac{1}{n_{\text{counsel}}} + \frac{1}{n_{\text{paper}}} \right)}}$$

$$= \frac{|.379 - .128| - \frac{1}{2} \left(\frac{1}{145} + \frac{1}{117} \right)}{\sqrt{.267(1-.267) \left(\frac{1}{145} + \frac{1}{117} \right)}} = 4.443$$

Note that the Yates correction reduced the value of the z test statistic. (Since the sample sizes are reasonably large, the effect was small.) The value of the z test static, 4.443, exceeds 3.2905, the value that defines the most extreme 1% of the normal distribution (from Table 4-1), so we reject the null hypothesis of no difference and conclude that the in-person counseling significantly increased the rate at which homeless people returned the advance directives.

■ ANOTHER APPROACH TO TESTING NOMINAL DATA: ANALYSIS OF CONTINGENCY TABLES

The methods we just developed based on the z statistic are perfectly adequate for testing hypotheses when there are only two possible attributes or outcomes of interest. The z statistic plays a role analogous to the t test for data measured on an interval scale. There are many situations, however, where there are more than two samples to be compared or more than two possible outcomes. To do this, we need to develop a testing procedure, analogous to analysis of variance, which is more flexible than the z test just described. While the following approach may seem quite different from the one we just used to

design the z test for proportions, it is essentially the same.

To keep things simple, we begin with the problem we just solved, assessing the effectiveness of in-person counseling of homeless people to prepare advance directives. In the last section we based the analysis on the *proportion* of people in each of the two treatment groups (in-person counseling or written materials). Now we change our emphasis slightly and base the analysis on the *number* of people in each group who did and did not file advance directives. Since the procedure we will develop does not require assuming anything about the nature of the parameters of the population from which the samples were drawn, it is called a *nonparametric* method.

Table 5-1 presents the data from this experiment in terms of the number of people in each treatment group who did and did not file advance directives. This table is called a 2×2 *contingency table*. Most of the people in the study fall along the diagonal in this table, suggesting an association between the experimental intervention and whether or not the person filed and advance directive. Table 5-2 shows what the experimental results might have looked like *if the experimental intervention had no effect on the results*, if the null hypothesis of no effect was true. It also shows the total number of people who received each intervention as well as the total who did and did not file advance directives. (The sums of the rows and columns are the same as in Table 5-1.) In Table 5-2, fewer people in both intervention groups filed advance directives than did not; the differences in the absolute numbers occur because more people were randomized into the counseling group than the written instructions group. In contrast to Table 5-1, there does not seem to be a relationship between the intervention and whether people filed advance directives.

To understand why most people have this subjective impression, let us examine where the numbers in Table 5-2 came from. Of all the 262 people in the study, 70, or

■ TABLE 5-1. Advance Directives Filed in People Who Received In-Person Counseling or Written Instructions

Intervention	Number of People		
	Filed Advance Directive	Did Not File Advance Directive	Total in Intervention Group
In-person counseling	55	90	145
Written instructions	15	102	117
Total	70	192	262

■ **TABLE 5-2. Expected Advance Directives Filed if Intervention Had No Effect**

Intervention	Number of People		
	Filed Advance Directive	Did Not File Advance Directive	Total in Intervention Group
In-person counseling	38.74	106.26	145
Written instructions	31.26	85.74	117
Total	70.00	192.00	262

$70/262 = 26.7\%$, filed advance directives and 192, or $192/262 = 73.3\%$, did not. Now, let us assume that the null hypothesis is true and that the intervention had no effect on the likelihood that a person would file an advance directive. In this case, we would expect 26.7% of the 145 people who received in-person counseling to file advance directives (38.74 people) and 26.7% of the 117 people who just received written materials (31.26) to file advance directives. We would expect the remaining 73.3% of people in each group to not have filed advance directives.* (We compute the expected frequencies to two decimal places to ensure accurate results in the computation of the χ^2 test statistic below.) Thus, Table 5-2 shows how we would *expect* the data to look if 145 people received in-person counseling and 117 received written materials and 70 of them were destined to file advance directives *regardless of which intervention they received*. Compare Tables 5-1 and 5-2. Do they seem similar? Not really; the actual pattern of observations seems quite different from what we expected if the intervention had no effect.

The next step in designing a statistical procedure to test the hypothesis that the pattern of observations is due to random sampling rather than the intervention is to reduce this subjective impression to a single number, a test statistic, such as F , t , or z , so that we can reject the null hypothesis of no effect when this statistic is “big.”

Before constructing this test statistic, however, let us consider another example. Hypothermia is a problem for extremely low birth weight infants. To investigate whether wrapping these infants in polyethylene bags in the delivery room and while they are being transferred to the neonatal

intensive care unit affected survival, Patrick Carroll and colleagues[†] reviewed medical records and located 70 infants who were kept warm with polyethylene bags and 70 infants who were kept warm with traditional methods. In an effort to avoid problems created by confounding variables in this observational study, they matched the infants according to birth weight, gestational age, and gender. They found that the infants wrapped in the polyethylene bags had statistically significantly higher skin temperatures, by an average of 1°C (see Prob. 4-2). The more important question was whether or not there was a mortality benefit.

Table 5-3 shows the results of this study, presented in the same format as Table 5-1. Table 5-4 shows the expected pattern of observations if the null hypothesis that the warming treatment had no effect on mortality was true. Out of the 140 infants, 124, or $124/140 = 88.6\%$, lived. If the warming treatment had no effect on survival, we would expect 88.6% of the 70 infants in each treatment group, 62 to live and the remaining 8 in each group to die. Comparing the observed mortality pattern in Table 5-3 with the expected pattern if the null hypothesis of no effect was true shows little difference, suggesting that there is no association between the kind of warming treatment and mortality.

The Chi-Square Test Statistic

Now we are ready to design our test statistic. It should describe, with a single number, how much the observed frequencies in each cell in the table differ from the frequencies we would expect if there is no relationship between the

*We could also have computed the estimated numbers by multiplying the number of people who did or did not file advance directives times the fraction of all the 262 people in the study, 145, or $145/262 = 55.3\%$, received in-person counseling and 117, or $117/262 = 44.7\%$, received written materials. The result would be the same.

[†]Carroll PD, Nanketvis CA, Giannone PJ, Cordero L. Use of polyethylene bags in extremely low birth weight infant resuscitation for the prevention of hyperthermia. *J Reprod Med.* 2010;55:9–13.

■ TABLE 5-3. Mortality Associated With Extreme Low Birth Weight

Warming Treatment	Number of Infants		
	Lived	Died	Total in Treatment Group
Polyethylene bag	63	7	70
Traditional	61	9	70
Total	124	16	140

treatments and the outcomes that define the rows and columns of the table. In addition, it should allow for the fact that if we expect a large number of people to fall in a given cell, a difference of one person between the expected and observed frequencies is less important than in cases where we expect only a few people to fall in the cell.

We define the test statistic χ^2 (the square of the Greek letter chi) as

$$\chi^2 = \text{Sum of } \frac{(\text{Observed} - \text{expected number of individuals in cell})^2}{\text{Expected number of individuals in cell}}$$

The sum is calculated by adding the results for all cells in the contingency table. The equivalent mathematical statement is

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

in which O is the observed number of individuals (frequency) in a given cell, E is the expected number of individuals (frequency) in that cell, and the sum is over all the cells in the contingency table. Note that if the observed frequencies are similar to the expected frequencies, χ^2 will be a small number and if the observed and expected frequencies differ, χ^2 will be a big number.

We can now use the information in Tables 5-1 and 5-2 to compute the χ^2 statistic associated with the data on counseling and filing advanced directives. Table 5-1 gives the observed frequencies, and Table 5-2 gives the expected frequencies. Thus,

$$\begin{aligned} \chi^2 &= \sum \frac{(O - E)^2}{E} = \frac{(55 - 38.74)^2}{38.74} + \frac{(90 - 106.26)^2}{102.6} \\ &\quad + \frac{(15 - 31.26)^2}{31.26} + \frac{(102 - 85.74)^2}{85.74} = 20.854 \end{aligned}$$

To begin getting a feeling of whether or not 20.854 is “big,” let us compute χ^2 for the data on warming technique for extreme low birth weight infants and mortality using the observed and expected counts in Tables 5-3 and 5-4:

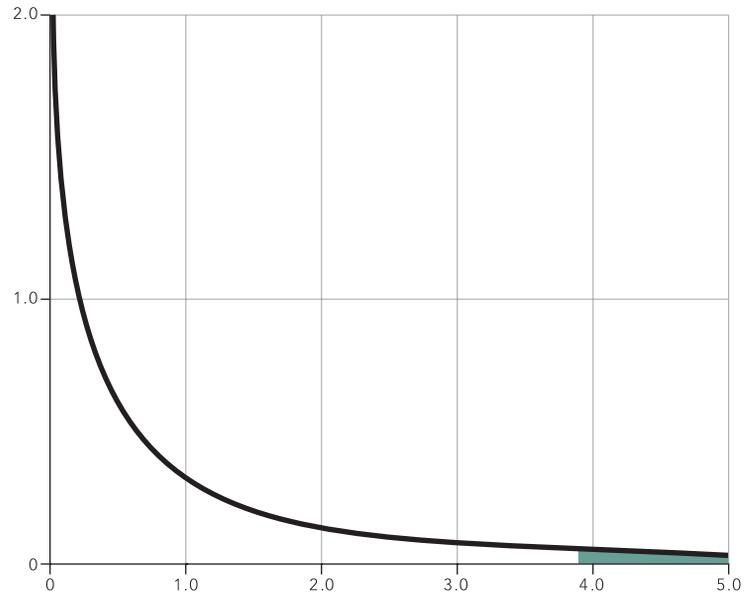
$$\begin{aligned} \chi^2 &= \sum \frac{(O - E)^2}{E} = \frac{(63 - 62)^2}{62} + \frac{(7 - 8)^2}{8} \\ &\quad + \frac{(61 - 62)^2}{62} + \frac{(9 - 8)^2}{8} = .282 \end{aligned}$$

which is pretty small, in agreement with our intuitive impression that the observed and expected frequencies are quite similar. (Of course, it is also in agreement with our earlier analysis of the same data using the z

■ TABLE 5-4. Expected Mortality if Treatment Had No Effect

Warming Treatment	Number of Infants		
	Lived	Died	Total in Treatment Group
Polyethylene bag	62	8	70
Traditional	62	8	70
Total	124	16	140

FIGURE 5-7. The χ^2 distribution with 1 degree of freedom. The shaded area denotes the biggest 5% of possible values of the χ^2 test statistic when there is no relationship between the treatments and observations.



statistic in the last section.) In fact, it is possible to show that $\chi^2 = z^2$ when there are only two samples and two possible outcomes.

Like all test statistics, χ^2 can take on a range of values even when there is no relationship between the treatments and outcomes because of the effects of random sampling. Figure 5-7 shows the distribution of possible values for χ^2 computed from data in 2×2 contingency tables such as those in Tables 5-1 or 5-3. It shows that when the hypothesis of no relationship between the rows and columns of the table is true, χ^2 would be expected to exceed 3.841 only 5% of the time.

Because the observed value of χ^2 for the counseling study in Table 5-1, 20.854, exceeds this critical value of 3.841, we conclude that the data in Table 5-1 are unlikely to occur if the null hypothesis that the counseling has no effect on filing advance directives was true. We report that counseling leads to higher rates of homeless people filing advance directives ($P < .05$).

Like all the other procedures we have been using to test hypotheses, however, when we reject the null hypothesis of no association at the 5% level, we are implicitly willing to accept the fact that, in the long run, about 1 reported effect in 20 will be due to random variation rather than a real treatment effect.

In contrast, the data in Table 5-3 seem very compatible with the null hypothesis that the warming technique did

not have any effect on mortality in extremely low birth weight infants.

Of course, neither of these studies *proves* that the in-person counseling session did or did not have an effect on homeless people filing advanced directives or use of polyethylene bags had an effect on extreme low birth weight infant mortality. What they show is that in the first example the pattern of the observations is unlikely to arise if the counseling session did not have an effect, whereas in the second example the pattern of observations is likely to arise if the polyethylene bag produced the same mortality rate as conventional warming techniques.

As with all theoretical distributions of test statistics used for testing hypotheses, there are assumptions built into the use of χ^2 . For the resulting theoretical distribution to be reasonably accurate, *the expected number of individuals in all the cells must be at least 5.** (This is essentially the same as the restriction on the z test in the last section.)

Like most test statistics, the distribution of χ^2 depends on the number of treatments being compared. It also depends on the number of possible outcomes. This dependency is quantified in a *degrees of freedom* parameter

*When the data do not meet this requirement, one should use the Fisher exact test, discussed later in this chapter.

v equal to the number of rows in the table minus 1 times the number of columns in the table minus 1

$$v = (r - 1)(c - 1)$$

where r is the number of rows and c is the number of columns in the table. For the 2×2 tables we have been dealing with so far, $v = (2 - 1)(2 - 1) = 1$.

Table 5-5 presents a table of critical values for the χ^2 test statistic. The critical value that defines the .1% largest values of χ^2 under the assumption that the null hypothesis is true with $v = 1$ degree of freedom is 10.828. The value associated with our data on the effects of in-person counseling on filing advance directives is 20.854, which exceeds this value. Consequently, we can reject the null hypothesis of no effect and conclude that in-person counseling increases the likelihood that a homeless person will file an advance directive ($P < .001$).

Likewise, the value of χ^2 for the study of the value of polyethylene wraps for extremely low birth weight infants was only .282, which is smaller than the value of .455 which defines the upper half of the χ^2 distribution with 1 degree of freedom, so we do not come close to rejecting the null hypothesis that the polyethylene wrap is no better than traditional warming methods in terms of infant mortality.

This study illustrates the importance of looking at *outcomes* in clinical trials. The human body has tremendous capacity to adapt not only to disease but also to medical manipulation. Therefore, simply showing that some intervention (such as a difference in warming technique) changed a patient's physiological state (by producing different body temperature) does not mean that in the long run it will make any difference in the clinical *outcome*. Focusing on these intermediate variables, often called *process variables*, rather than the more important outcome variables may lead you to think something made a clinical difference when it did not. For example, in this study there was the expected change in the process variable, skin temperature, but not the outcome variable, mortality. If we had stopped with the process variables we might have concluded that the polyethylene wrap was superior to traditional warming methods, even though the choice of warming method does not appear to have affected the most important variable, whether or not the infant survived.

Keep this distinction in mind when reading medical journals and listening to proponents argue for their tests, procedures, and therapies. It is much easier to show that something affects process variables than the more important outcome variables. In addition to being easier to produce a

demonstrable change in process variables than outcome variables, process variables are generally easier to measure. Observing outcomes may require following the patients for some time and often present difficult subjective problems of measurement, especially when one tries to measure "quality of life" variables. Nevertheless, when assessing whether or not some new procedure deserves to be adopted in an era of limited medical resources, you should seek evidence that something affects the patient's outcome. The patient and the patient's family care about outcome, not process.

The Yates Correction for Continuity

As with the z test statistic discussed earlier in this chapter, when analyzing 2×2 contingency tables ($v = 1$), the value of χ^2 computed using the formula above and the theoretical χ^2 distribution leads to P values that are smaller than they ought to be. Thus, the results are biased toward concluding that the treatment had an effect when the evidence does not support such a conclusion. The mathematical reason for this problem has to do with the fact that the theoretical χ^2 distribution is continuous whereas the set of all possible values that the χ^2 test statistics can take on is not. To obtain values of the test statistic that are more compatible with the critical values computed from the theoretical χ^2 distribution when $v = 1$, apply the *Yates correction* (or *continuity correction*) to compute a corrected χ^2 test statistic according to

$$\chi^2 = \sum \frac{(|O - E| - \frac{1}{2})^2}{E}$$

This correction slightly reduces the value of χ^2 associated with the contingency table and compensates for the mathematical problem just described. The Yates correction is used only when $v = 1$, that is, for 2×2 tables.

To illustrate the use and effect of the continuity correction, let us recompute the value of χ^2 associated with the data on counseling and filing of advance directives in Table 5-1. From the observed and expected frequencies in Tables 5-1 and 5-2, respectively,

$$\begin{aligned} \chi^2 &= \sum \frac{(|O - E| - \frac{1}{2})^2}{E} = \frac{(|55 - 38.74| - \frac{1}{2})^2}{38.74} \\ &+ \frac{(|90 - 106.26| - \frac{1}{2})^2}{102.6} + \frac{(|15 - 31.26| - \frac{1}{2})^2}{31.26} \\ &+ \frac{(|102 - 85.74| - \frac{1}{2})^2}{85.74} = 19.591 \end{aligned}$$

■ TABLE 5-5. Critical Values for the χ^2 Distribution

v	Probability of Greater Value P							
	.50	.25	.10	.05	.025	.01	.005	.001
1	.455	1.323	2.706	3.841	5.024	6.635	7.879	10.828
2	1.386	2.773	4.605	5.991	7.378	9.210	10.597	13.816
3	2.366	4.108	6.251	7.815	9.348	11.345	12.838	16.266
4	3.357	5.385	7.779	9.488	11.143	13.277	14.860	18.467
5	4.351	6.626	9.236	11.070	12.833	15.086	16.750	20.515
6	5.348	7.841	10.645	12.592	14.449	16.812	18.548	22.458
7	6.346	9.037	12.017	14.067	16.013	18.475	20.278	24.322
8	7.344	10.219	13.362	15.507	17.535	20.090	21.955	26.124
9	8.343	11.389	14.684	16.919	19.023	21.666	23.589	27.877
10	9.342	12.549	15.987	18.307	20.483	23.209	25.188	29.588
11	10.341	13.701	17.275	19.675	21.920	24.725	26.757	31.264
12	11.340	14.845	18.549	21.026	23.337	26.217	28.300	32.909
13	12.340	15.984	19.812	22.362	24.736	27.688	29.819	34.528
14	13.339	17.117	21.064	23.685	26.119	29.141	31.319	36.123
15	14.339	18.245	22.307	24.996	27.488	30.578	32.801	37.697
16	15.338	19.369	23.542	26.296	28.845	32.000	34.267	39.252
17	16.338	20.489	24.769	27.587	30.191	33.409	35.718	40.790
18	17.338	21.605	25.989	28.869	31.526	34.805	37.156	42.312
19	18.338	22.718	27.204	30.144	32.852	36.191	38.582	43.820
20	19.337	23.828	28.412	31.410	34.170	37.566	39.997	45.315
21	20.337	24.935	29.615	32.671	35.479	38.932	41.401	46.797
22	21.337	26.039	30.813	33.924	36.781	40.289	42.796	48.268
23	22.337	27.141	32.007	35.172	38.076	41.638	44.181	49.728
24	23.337	28.241	33.196	36.415	39.364	42.980	45.559	51.179
25	24.337	29.339	34.382	37.652	40.646	44.314	46.928	52.620
26	25.336	30.435	35.563	38.885	41.923	45.642	48.290	54.052
27	26.336	31.528	36.741	40.113	43.195	46.963	49.645	55.476
28	27.336	32.020	37.916	41.337	44.461	48.278	50.993	56.892
29	28.336	33.711	39.087	42.557	45.722	49.588	52.336	58.301
30	29.336	34.800	40.256	43.773	46.979	50.892	53.672	59.703
31	30.336	35.887	41.422	44.985	48.232	52.191	55.003	61.098
32	31.336	36.973	42.585	46.194	49.480	53.486	56.328	62.487
33	32.336	38.058	43.745	47.400	50.725	54.776	57.648	63.870
34	33.336	39.141	44.903	48.602	51.966	56.061	58.964	65.247
35	34.336	40.223	46.059	49.802	53.203	57.342	60.275	66.619
36	35.336	41.304	47.212	50.998	54.437	58.619	61.581	67.985
37	36.336	42.383	48.363	52.192	55.668	59.893	62.883	69.346
38	37.335	43.462	49.513	53.384	56.896	61.162	64.181	70.703
39	38.335	44.539	50.660	54.572	58.120	62.428	65.476	72.055
40	39.335	45.616	51.805	55.758	59.342	63.691	66.766	73.402
41	40.335	46.692	52.949	56.942	60.561	64.950	68.053	74.745
42	41.335	47.766	54.090	58.124	61.777	66.206	69.336	76.084
43	42.335	48.840	55.230	59.304	62.990	67.459	70.616	77.419
44	43.335	49.913	56.369	60.481	64.201	68.710	71.893	78.750
45	44.335	50.985	57.505	61.656	65.410	69.957	73.166	80.077
46	45.335	52.056	58.641	62.830	66.617	71.201	74.437	81.400
47	46.335	53.127	59.774	64.001	67.821	72.443	75.704	82.720
48	47.335	54.196	60.907	65.171	69.023	73.683	76.969	84.037
49	48.335	55.265	62.038	66.339	70.222	74.919	78.231	85.351
50	49.335	56.334	63.167	67.505	71.420	76.154	79.490	86.661

Adapted from Zar JH. *Biostatistical Analysis*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall; 1984, 479–482:table B.1, by permission of Pearson Education, Inc., Upper Saddle River, NJ.

■ **TABLE 5-6. Medical Students Who Signed Organ Donation Cards**

Race/Ethnicity	Yes	No	Total
White	290	57	347
Asian	40	17	57
Black	14	21	35
Total	344	95	439

Note that this value of χ^2 is smaller than the uncorrected value, although the effect is so large that even the smaller value of χ^2 still exceeds the critical value that defines the largest .1% of the χ^2 distribution with 1 degree of freedom, 10.282, so we still reject the null hypothesis of no effect ($P < .001$), as before. This situation is not always the case; when the effect or sample sizes are smaller than in this example including the Yates correction can affect the P value and even whether or not the result reaches conventional statistical significance.

You should always include the Yates continuity correction when analyzing 2×2 contingency tables.

■ CHI-SQUARE APPLICATIONS TO EXPERIMENTS WITH MORE THAN TWO TREATMENTS OR OUTCOMES

It is easy to generalize what we have just done to analyze the results of experiments with more than two treatments or outcomes. The z test we developed earlier in this chapter will not work for such studies.

There is a chronic shortage of donated organs. To develop better educational programs on organ donation for medical students Teresa Edwards and colleagues* surveyed 439 students in three Ohio medical schools to investigate whether there were any racial or ethnic differences in attitudes toward organ donation. Table 5-6 shows the results for whether the students had already signed an organ donation card. Are these data consistent with the null hypothesis that race/ethnicity is not related to having signed an organ donation card?

*Edwards TM, Essma C, Thornton JD. Assessing racial and ethnic differences in medical student knowledge, attitudes and behaviors regarding organ donation. *J Natl Med Assoc.* 2007;99:131–137.

■ **TABLE 5-7. Expected Number of Students Who Signed Organ Donation Cards if Race/Ethnicity Did Not Matter***

Race/Ethnicity	Yes	No	Total
White	271.90	75.10	347
Asian	44.67	12.33	57
Black	27.43	7.57	35
Total	344.00	95.00	439

We compute the expected numbers in each cell assuming that the null hypothesis is true just as we did before. Three hundred forty-four of the 439 students, $344/439 = 78.36\%$, signed donor cards. Thus, if race/ethnicity does not affect the likelihood that a student signed a donor card, then we would expect 78.36% of the 347 white students (271.90 students), 78.36% of the 57 Asian students (44.67 students) and 78.36% of the 35 Black students (27.43 students) in the sample to have signed donor cards, with the remaining students in each group not signing donor cards (Table 5-7).

We now compute the χ^2 test statistic as before. (Because this is not a 2×2 table, we do not need to include the Yates correction.)

$$\begin{aligned}\chi^2 &= \sum \frac{(O-E)^2}{E} = \frac{(290-271.90)^2}{271.90} + \frac{(57-75.10)^2}{75.10} \\ &+ \frac{(40-44.67)^2}{44.67} + \frac{(17-12.33)^2}{12.33} + \frac{(14-27.43)^2}{27.43} \\ &+ \frac{(21-7.57)^2}{7.57} = 38.186\end{aligned}$$

The contingency table in Table 5-6 has three rows and two columns, so the χ^2 test statistic has

$$v = (r - 1)(c - 1) = (3 - 1)(2 - 1) = 2$$

degrees of freedom associated with it. Table 5-5 shows that χ^2 will exceed 38.186 less than .1% of the time when the difference between the observed and expected frequencies is due to random variation rather than an effect of the sample group (in this case, race/ethnicity). Thus, we conclude that there is a significant difference in the likelihood that a student will have signed a donor card, depending on his or her race/ethnicity.

Note, however, that we do not know where this difference comes from. Answering that question will require doing multiple comparisons.

Let us now sum up how to use the χ^2 statistic:

- *Tabulate the data in a contingency table.*
- *Sum the number of individuals in each row and each column and figure the percentage of all individuals who fall in each row and column, independent of the column or row in which they fall.*
- *Use these percentages to compute the number of people that would be expected in each cell of the table if the treatment had no effect.*
- *Summarize the differences between these expected frequencies and the observed frequencies by computing χ^2 . If the data form a 2×2 table, include the Yates correction.*
- *Compute the number of degrees of freedom associated with the contingency table and use Table 5-5 to see whether the observed value of χ^2 exceeds what would be expected from random variation.*

Recall that when the data fell into a 2×2 contingency table, all the expected frequencies had to exceed about 5 for the χ^2 test to be accurate. In larger tables, most statisticians recommend that the expected number of individuals in each cell never be less than 1 and that no more than 20% of cells be less than 5. When this is not the case, the χ^2 test can be quite inaccurate. The problem can be remedied by collecting more data to increase the cell numbers or by reducing the number of categories to increase the numbers in each cell of the table.

Multiple Comparisons

Because the contingency table has two columns, we can *subdivide* it into three 2×2 contingency tables to do all pairwise comparisons, just as we did following rejecting the null hypothesis in analysis of variance. As then, we can use Bonferroni, Holm or Holm-Sidak corrections to determine if individual comparisons are significant. The reason that we can use these corrections is because they all adjust the critical value of P required to reject individual pairwise (or comparisons against a single control group) based on considerations of how the risks of erroneously rejecting the null hypothesis accumulate as we do multiple comparisons. The values of P_{crit} depend on the overall family error rate one seeks to control (α_T) and the number of comparisons (k), but not the details of how one

obtained the individual P values that are compared to P_{crit} . In fact, the Bonferroni, Holm and Holm-Sidak procedures can be applied to control the family error rate for *any* collection of hypothesis tests that you wish to consider a family of comparisons.

To apply this general principle to the problem of identifying what difference or differences between the racial/ethnic groups in Table 5-6 led us to reject the null hypothesis of no difference in the likelihood of a student having a donor card, we first test for differences in each of the three 2×2 tables we can construct from Table 5-6. Box 5-1 shows these three tables and the associated χ^2 test statistics. (Note that we have to include the Yates correction because these are 2×2 tables.)

Once we have the values of χ^2 associated with each of these three pairwise comparisons, we can determine whether or not they are big enough to reject the null hypothesis of no difference for the individual comparisons, while controlling the overall family error rate at $\alpha_T = 5\%$. As when we used the Holm-Sidak correction with t tests following a significant analysis of variance, we order the comparisons according to descending values of the χ^2 associated with each comparison (Table 5-8). In each case, the P value exceeds the Holm-Sidak P_{crit} , so we conclude that each racial/ethnic group has significantly different rates of signing organ donor cards from the other two, with 84% of whites, 70% of Asians and 40% of Blacks signing the donor cards.

There is no generally accepted procedure for subdividing contingency tables that are 3×3 or larger.

■ THE FISHER EXACT TEST

The χ^2 test can be used to analyze 2×2 contingency tables when each cell has an expected frequency of at least 5. In small studies, when the expected frequency is smaller than 5, the *Fisher exact test* is the appropriate procedure. This test turns the liability of small sample sizes into a benefit. When the sample sizes are small, it is possible to simply *list* all the possible arrangements of the observations, and then compute the exact probabilities associated with each possible arrangement of the data. The total (two-tailed) probability of obtaining the observed data or more extreme patterns in the data is the P value associated with the hypothesis that the rows and columns in the data are independent.

The Fisher exact test begins with the fact that the probability of observing any given pattern in the 2×2

BOX 5-1 • All Pairwise Multiple Comparisons for Effects of Race/Ethnicity on Having Signed an Advance Directive for Organ Donation

There are three 2×2 contingency tables; note that the marginal sums are only based on the two groups of students represented in each table. There is $v = 1$ degree of freedom associated with each table, which can be used to look up the appropriate P value for each comparison in Table 5-8.

White vs. Asian Students

Race/Ethnicity	Yes	No	Total
White	290	57	347
Asian	40	17	57
Total	330	74	404

Overall $330/404 = 81.68\%$ of students signed advance directives, so under the null hypothesis we expect 81.68% of the 347 white students (283.44) and 81.68% of the 57 Asian students (63.56) to have advanced directives, so

$$\chi_{W \text{ vs } A}^2 = \sum \frac{(|O - E| - \frac{1}{2})^2}{E} = \frac{(|290 - 283.44| - \frac{1}{2})^2}{283.44} + \frac{(|57 - 63.56| - \frac{1}{2})^2}{63.56} + \frac{(|40 - 46.56| - \frac{1}{2})^2}{46.56} + \frac{(|17 - 10.44| - \frac{1}{2})^2}{10.44} = 5.012; P < .05$$

White vs. Black Students

Race/Ethnicity	Yes	No	Total
White	290	57	347
Black	14	21	35
Total	304	78	382

Overall, $304/382 = 79.58\%$ of students had advance directives, so

$$\chi_{W \text{ vs } B}^2 = \sum \frac{(|O - E| - \frac{1}{2})^2}{E} = \frac{(|290 - 276.15| - \frac{1}{2})^2}{276.15} + \frac{(|57 - 70.85| - \frac{1}{2})^2}{70.85} + \frac{(|14 - 27.85| - \frac{1}{2})^2}{27.85} + \frac{(|21 - 7.15| - \frac{1}{2})^2}{7.15} = 34.515; P < .001$$

Asian vs. Black

Race/Ethnicity	Yes	No	Total
Asian	40	17	57
Black	14	21	35
Total	54	38	92

Likewise,

$$\chi_{A \text{ vs } B}^2 = \sum \frac{(|O - E| - \frac{1}{2})^2}{E} = \frac{(|40 - 33.46| - \frac{1}{2})^2}{33.46} + \frac{(|17 - 23.54| - \frac{1}{2})^2}{63.56} + \frac{(|14 - 20.55| - \frac{1}{2})^2}{20.55} + \frac{(|21 - 14.45| - \frac{1}{2})^2}{14.45} = 6.947; P < .01$$

TABLE 5-8. Pairwise Comparisons of Sperm Motility in Rabbit Cell Phone Experiment using Holm-Sidak Adjustment (Family Error Rate, $\alpha_T = 0.05$)

Comparison	χ^2	P	j	$P_{crit} = \alpha_T / (k - j + 1)$	$P < P_{crit}?$
White vs. Black	34.515	<.001	1	.0170	Yes
Asian vs. Black	6.947	<.010	2	.0253	Yes
White vs. Asian	5.012	<.050	3	.0500	Yes

v = 1 degree of freedom; *k* = 3 comparisons.

contingency table with the observed row and column totals in Table 5-9 is

$$p = \frac{R_1!R_2!C_1!C_2!}{N! O_{11}!O_{12}!O_{21}!O_{22}!}$$

where O_{11} , O_{12} , O_{21} and O_{22} are the observed frequencies in the four cells of the contingency table, C_1 and C_2 are the sums of the two columns, R_1 and R_2 are the sums of the two rows, N is the total number of observations, and the exclamation mark “!” indicates the factorial operator.*

Unlike the χ^2 test statistic, there are one- and two-tailed versions of the Fisher exact test. Unfortunately, most descriptions of the Fisher exact test simply describe the one-tailed version and many computer programs compute the one-tailed version without clearly identifying it as such. Because many researchers do not recognize this issue, results (i.e., P values) may be reported for a single tail without the researchers realizing it.

To determine whether or not investigators recognized whether they were using one- or two-tailed Fisher exact tests, W. Paul McKinney and colleagues† examined the use of the Fisher exact test in papers published in the medical literature to see whether or not the authors noted the type of Fisher exact test that was used. Table 5-10 shows the data for the two journals: *New England Journal of Medicine* and *The Lancet*. Because the numbers are small, χ^2 is not an appropriate test statistic. From the equation above, the

probability of obtaining the pattern of observations in Table 5-10 for the given row and column totals is

$$p = \frac{9!14!11!12!}{23! 1!8!10!4!} = .00666$$

Thus, it is very unlikely that *this particular* table would be observed. To obtain the probability of observing a pattern in the data this extreme *or more extreme* in the direction of the table, reduce the smallest observation by 1, and recompute the other cells in the table to maintain the row and column totals constant.

TABLE 5-9. Notation for the Fisher Exact Test

	Row Totals		
	O_{11}	O_{12}	R_1
	O_{21}	O_{22}	R_2
Column Total	C_1	C_2	N

TABLE 5-10. Reporting of Use of Fisher Exact Test in the *New England Journal of Medicine* and *The Lancet*

Group	Test Identified?		
	Yes	No	Total
<i>New England Journal of Medicine</i>	1	8	9
<i>The Lancet</i>	10	4	14
Total	11	12	23

*The definition $n!$ is $n! = (n)(n-1)(n-2) \times \dots \times (2)(1)$; e.g., $5! = 5 \times 4 \times 3 \times 2 \times 1$.

†McKinney WP, Young MJ, Harta A, Lee MB. The inexact use of Fisher’s exact test in six major medical journals. *JAMA*. 1989;261:3430–3433.

■ **TABLE 5-11. More Extreme Pattern of Observations in Table 5-11, Using Smallest Observed Frequency (in This Case, 1)**

Group	Test Identified?		Total
	Yes	No	
<i>New England Journal of Medicine</i>	0	9	9
<i>The Lancet</i>	<u>11</u>	<u>3</u>	<u>14</u>
Totals	11	12	23

In this case, there is one more extreme table, given in Table 5-11. This table has a probability of occurring of

$$p = \frac{9!14!11!12!}{23!} = .00027$$

(Note that the numerator only depends on the row and column totals associated with the table, which does not change, and so only needs to be computed once.) Thus, the one-tailed Fisher exact test yields a P value of $P = .00666 + .00027 = .00693$. This probability represents the probability of obtaining a pattern of observations as extreme or more extreme in one direction as the actual observations in Table 5-10.

To find the other tail, we list all the remaining possible patterns in the data that would give the same row and column totals. These possibilities, together with the associated probabilities, appear in Table 5-12. These tables are obtained by taking each of the remaining three elements in Table 5-10 one at a time and progressively making it smaller by one, then eliminating the duplicate tables. Two of these tables have probabilities at or below the probability of obtaining the original observations, .00666: the ones with probabilities of .00242 and .00007. These two tables constitute the “other” tail of the Fisher exact test. There is a total

■ **TABLE 5-12. Other Patterns of Observations in Table 5-11 with the Same Row and Column Totals**

		Total			Total	
	2	7	9	6	3	
	<u>9</u>	<u>5</u>	<u>14</u>	<u>5</u>	<u>9</u>	
Total	11	12	23	Total	11 12 23	
	$P = .05330$			$P = .12438$		
		Total			Total	
	3	6	9	7	2	
	<u>8</u>	<u>6</u>	<u>14</u>	<u>4</u>	<u>10</u>	
Total	11	12	23	Total	11 12 23	
	$P = .18657$			$P = .02665$		
		Total			Total	
	4	5	9	8	1	
	<u>7</u>	<u>7</u>	<u>14</u>	<u>3</u>	<u>11</u>	
Total	11	12	23	Total	11 12 23	
	$P = .31983$			$P = .00242$		
		Total			Total	
	5	4	9	9	0	
	<u>6</u>	<u>8</u>	<u>14</u>	<u>2</u>	<u>12</u>	
Total	11	12	23	Total	11 12 23	
	$P = .27985$			$P = .00007$		

probability of being in this table of $.00242 + .00007 = .00249$.^{*} Thus, the total probability of obtaining a pattern of observations as extreme or more extreme than that observed is $P = .00693 + .00249 = .00942$, and we conclude there is a significant difference in the correct presentation of the Fisher exact test in the *New England Journal of Medicine* and *The Lancet* ($P = .009$). Indeed, it is important when reading papers that use the Fisher exact test to make sure the authors know what they are doing and report the results appropriately.

Let us now sum up how to do the Fisher exact test:

- Compute the probability associated with the observed data.
- Identify the cell in the contingency table with the smallest frequency.
- Reduce the smallest element in the table by 1 and then compute the elements for the other three cells so that the row and column sums remain constant.
- Compute the probability associated with the new table.
- Repeat this process until the smallest element becomes its lowest possible value, which is often but not always zero.
- List the remaining tables by repeating this process for the other three elements.[†] List each pattern of observations only once.
- Sequentially compute the probabilities associated with the tables from most extreme to least extreme until reaching a table that has a probability greater than the observed result.
- Add all the probabilities together that are equal to or smaller than the probability associated with the observed data.

This probability is the *two-tail* probability of observing a pattern in the data as extreme or more extreme than observed. Many computer programs show P values for the Fisher exact test, without clearly indicating whether they are one- or two-tail values. Make sure that you know which value is being reported before you use it in your work; the two-tailed P value is generally the one you want.

^{*}Note that the two tails have different probabilities; this is generally the case. The one exception is when either the two rows or two columns have the same sums, in which case the two-tail probability is simply twice the one-tail probability. Some books say that the two-tail value of P is always simply twice the one-tail value. This is not correct unless the row or column sums are equal.

[†]Many of these computations can be avoided, see Appendix A.

■ MEASURES OF ASSOCIATION BETWEEN TWO NOMINAL VARIABLES

In addition to testing whether there are significant differences between two rates or proportions, people often want a measure of the strength of association between some event and different treatments or conditions, particularly in *clinical trials* and *epidemiological studies*. In a *prospective clinical trial*, such as the study of the effect of in-person counseling on filing of advanced directives by homeless people discussed earlier in this chapter (Table 5-1), investigators randomly assign people to treatment (in-person counseling) or control (written materials only), then followed them to see whether they filed an advance directive or not. In that example, 38% (55 out of 145) of the people receiving in-person counseling filed advance directives and 13% (15 out of 117) of the people receiving written materials filed advanced directives. These proportions are estimates of the probability of filing an advanced directive associated with each of these treatments; these results indicate that the probability of filing an advance directive was nearly tripled by in-person counseling. We will now examine different ways to quantify this effect: *relative risk* and *odds ratio*.

Prospective Studies and Relative Risk

We quantify the size of the association between treatment and outcome with the *relative risk*, RR, which is defined as

$$RR = \frac{\text{Probability of event in treatment group}}{\text{Probability of event in control in group}}$$

For the advanced directive study, in which 37.9% of people who received in-person counseling completed the advance directives and 12.8% of people who just received written instructions did so,

$$RR = \frac{\hat{p}_{\text{counsel}}}{\hat{p}_{\text{written}}} = \frac{.379}{.128} = 2.92$$

The fact that the relative risk exceeds 1 indicates that in-person counseling increases the likelihood (“risk”) that a homeless person will file an advance directive. In clinical trials evaluating treatments against placebo (or standard treatment, when it would be unethical to administer a placebo) and the outcome is a negative event (such as death or disease recurrence), a relative risk of less than 1 indicates that the treatment leads to better outcomes.

■ **TABLE 5-13. Arrangement of Data to Compute Relative Risk**

Sample Group	Number of People		
	Disease	No Disease	Total
Treated (or exposed to risk factor)	n_{TD}	n_{TN}	n_T
Control (or not exposed to risk factor)	n_{CD}	n_{CN}	n_C
Total	n_D	n_N	

In an *epidemiological study*, the probability of an event among people *exposed* to some potential toxin or risk factor is compared to people who are *not exposed*. The calculations are the same as for clinical trials.*

Relative risks greater than 1 indicate that exposure to the toxin *increases* the risk of disease. For example, breathing secondhand smoke is associated with a relative risk of heart disease in nonsmokers of 1.3,[†] indicating that nonsmokers married to smokers are 1.3 times more likely to die from heart disease as nonsmokers married to nonsmokers (and so not breathing secondhand smoke at home).

Table 5-13 shows the general layout for a calculation of relative risk; it is simply a 2×2 contingency table. The probability of an event in the treatment group (also called the *experimental event rate*) is n_{TD}/n_T and the probability of an event in the control group (also called the *control event rate*) is n_{CD}/n_C . Therefore, the formula for relative risk is

$$RR = \frac{n_{TD}/n_T}{n_{CD}/n_C}$$

*In clinical trials and epidemiological studies one often wants to adjust for other so-called *confounding variables* that could be affecting the probability of an event. It is possible to account for such variables using multivariate techniques using *logistic regression* or *Cox proportional hazards regression*. For a discussion of these issues, see Glantz SA, Slinker BK. Regression with a qualitative dependent variable. *Primer of Applied Regression and Analysis of Variance*, 2nd ed. New York: McGraw Hill; 2001:chap 12.

[†]Barnoya J, Glantz SA. Cardiovascular effects of secondhand smoke: nearly as large as smoking. *Circulation*. 2005;24:111:2684–2698.

This formula is simply a restatement of the definition of relative risk presented above.

Using the results of the advance directives trial in Table 5-1, we compute

$$RR = \frac{55/(55+90)}{15/(15+102)} = \frac{.38}{.13} = 2.92$$

The most common null hypothesis that people wish to test related to relative risks is that the relative risk equals 1 (i.e., that the treatment or risk factor does not affect event rate). Although it is possible to test this hypothesis using the standard error of the relative risk, most people simply apply a χ^2 test to the contingency table used to compute the relative risk.[‡]

To compute a relative risk, the data must be collected as part of a *prospective study* in which people are randomized to treatment or control or subjects in an epidemiological study[§] are followed forward in time after they are exposed (or not exposed) to the toxin or risk factor of interest. It is necessary to conduct the study prospectively to estimate the absolute event rates in people in the treatment (or exposed) and control groups.

Absolute Risk Increase (or Reduction) and Number Needed to Treat

Another way to quantify this difference is to present the *absolute risk increase*, which is simply the difference of the probability of an event (in this case, filing an advance directive) with and without the treatment (in-person counseling), $.38 - .13 = .25$. The in-person counseling increases the probability that a homeless person will file an advanced directive by .25. This information can also be used to compute the *number needed to treat*, which is the number of people that would have to be treated to have one additional event. The number needed to treat is simply 1 divided by the absolute risk increase, in this case $1/.25 = 4$. Thus, one would expect to have one additional advance directive filed for each 4 homeless people that receives in-person counseling. If studying a clinical intervention that *reduces* the risk of an adverse event, we

[‡]Traditionally, direct hypothesis testing of relative risks is done by examining confidence intervals, see Chapter 7.

[§]Prospective epidemiological studies are also called *cohort studies*.

compute the *absolute risk reduction* and the number needed to treat (1 divided by the absolute risk reduction) is the number of people that need to be treated to avoid one adverse event.

Number needed to treat is often used as a measure of the cost-effectiveness of a treatment. With a large clinical trial it is often possible to detect small benefits of a therapy. When the benefits of the control therapy are small but positive with the new therapy being studied, it is possible that the relative risk (in this case, the benefit) for the new therapy will be larger even though the absolute risk improvement is small and, so, the number needed to treat to obtain one additional successful outcome is very large. If the therapy is very expensive or has serious side effects, it may not be sensible to use the therapy even though it produces a statistically significant improvement in outcomes.

Case-Control Studies and the Odds Ratio

Prospective studies are often difficult and expensive to do, particularly if it takes several years for events to occur after treatment or exposure. It is, however, possible to conduct a similar analysis *retrospectively* based on so-called *case-control studies*.

Unlike prospective studies, case-control studies are done after the fact. In a case-control study, people who experienced the outcome of interest are identified and the number exposed to the risk factor of interest are counted. These people are the *cases*. You then identify people who did not experience the outcome of interest, but are similar to the cases in all other relevant ways and count the number that were exposed to the risk factor. These people are the *controls*. (Often investigators include more than one control per case in order to increase the sample size.) Table 5-14 shows the layout for data from a case-control study.

This information can be used to compute a statistic similar to the relative risk known as the *odds ratio*. The odds ratio, OR, is defined as

$$\text{OR} = \frac{\text{Odds of exposure in cases}}{\text{Odds of exposure in controls}}$$

The proportion of cases (people with the disease) exposed to the risk factor is n_{ED}/n_D and the proportion of cases not exposed to the risk factor is n_{UD}/n_D . (Note that

■ TABLE 5-14. Arrangement of Data to Compute Odds Ratio

Sample Group	Number of People	
	Disease “Cases”	No Disease “Controls”
Exposed to risk factor (or treatment)	n_{ED}	n_{EN}
Unexposed to risk factor (or treatment)	n_{UD}	n_{UN}
Total	n_D	n_N

each of the denominators is appropriate for the numerator; this situation would not exist if one was using case-control data to compute a relative risk.) The odds of exposure in the cases is the ratio of these two percentages:

$$\text{Odds of exposure in cases} = \frac{n_{ED}/n_D}{n_{UD}/n_D} = \frac{n_{ED}}{n_{UD}}$$

Likewise, the odds of exposure in the controls is

$$\text{Odds of exposure in controls} = \frac{n_{EN}/n_N}{n_{UN}/n_N} = \frac{n_{EN}}{n_{UN}}$$

Finally, the odds ratio is

$$\text{OR} = \frac{n_{ED}/n_{UD}}{n_{EN}/n_{UN}} = \frac{n_{ED}n_{UN}}{n_{EN}n_{UD}}$$

Because the number of controls (n_{EN} and n_{UN} in Table 5-14) depends on how the investigator designs the study, you cannot use data from a case-control study to compute a relative risk. In a case-control study the investigator decides how many subjects with and without the disease will be studied. This is the opposite of the situation in prospective studies (clinical trials and epidemiological cohort studies), when the investigator decides how many subjects with and without the risk factor will be included in the study. The odds ratio may be used in both case-control and prospective studies, but *must* be used in case-control studies.

While the odds ratio is distinct from the relative risk, the odds ratio is a reasonable estimate of the relative risk when the number of people with the disease is small compared to the number of people without the disease.*

As with the relative risk, the most common null hypothesis that people wish to test related to relative risks is that the odds ratio equals 1 (i.e., that the treatment or risk factor does not affect the event rate). While it is possible to test this hypothesis using the standard error of the odds ratio, most people simply apply a χ^2 test to the contingency table used to compute the odds ratio.†

Passive Smoking and Breast Cancer

Breast cancer is the second leading cause of cancer death among women (behind lung cancer). Smoking could cause breast cancer because of the cancer-causing chemicals in the smoke that enter the body and some of these chemicals appear in breast milk, indicating that they reach the breast. To examine whether exposure to secondhand tobacco smoke increased the risk of breast cancer in lifelong nonsmokers, Kenneth Johnson and colleagues‡ conducted a case-control study using cancer registries in Canada to identify premenopausal women with histologically confirmed invasive primary breast cancer. They contacted the women and interviewed them about their smoking habits and exposure to secondhand smoke at home and at work. They obtained a group of controls who did not have breast cancer, matched by age group, from a mailing to women using

■ **TABLE 5-15. Passive Smoking and Breast Cancer**

Sample Group	Number of People	
	Cases (Breast Cancer)	Controls
Exposed to second-hand smoke	50	43
Not exposed to second-hand smoke	14	35
Total	64	78

lists obtained from the provincial health insurance authorities. Table 5-15 shows the resulting data.

The fraction of women with breast cancer (cases) who were exposed to secondhand smoke is $50/(50 + 14) = 0.781$ and the fraction of women with breast cancer not exposed to secondhand smoke is $14/(50 + 14) = 0.218$, so the odds of the women with breast cancer having been exposed to secondhand smoke is $0.781/0.218 = 3.58$. Similarly, the fraction of controls exposed to secondhand smoke is $43/(43 + 35) = 0.551$ and the fraction not exposed to secondhand smoke is $35/(43 + 35) = 0.449$, so the odds of the women without breast cancer having been exposed to secondhand smoke is $0.551/0.449 = 1.23$. Finally, the odds ratio of breast cancer associated with secondhand smoke exposure is

$$\text{OR} = \frac{\text{Odds of secondhand smoke exposure in women with breast cancer}}{\text{Odds of secondhand smoke exposure in controls}} = \frac{3.58}{1.23} = 2.91$$

Alternatively, we could use the direct formula for odds ratio and compute

$$\text{OR} = \frac{n_{ED}n_{UN}}{n_{UD}n_{EN}} = \frac{50 \cdot 35}{14 \cdot 43} = 2.91$$

Based on this study, we conclude that exposure to secondhand smoke increases the odds of having breast cancer by 2.91 times among this population. A χ^2 analysis of the data in Table 5-15 shows that this difference is statistically significant ($P = .007$).

*In this case, the number of people who have the disease, n_{TD} and n_{CD} , is much smaller than the number of people without the disease, n_{TN} and n_{CN} , so $n_T = n_{TD} + n_{TN} \approx n_{TN}$ and $n_C = n_{CD} + n_{CN} \approx n_{CN}$. As a result,

$$\text{RR} = \frac{\frac{n_{TD}}{n_T}}{\frac{n_{CD}}{n_C}} \approx \frac{\frac{n_{TD}}{n_{TN}}}{\frac{n_{CD}}{n_{CN}}} = \frac{n_{TD}n_{CN}}{n_{CD}n_{TN}} = \text{OR}$$

Because $n_{TD} = n_{ED}$, $n_{TN} = n_{EN}$, $n_{CD} = n_{UD}$, and $n_{CN} = n_{UN}$. For a more detailed and practical discussion of how the odds ratio and relative risk relate to each other, see Guyat GG, Rennie D, Meade MO, Cook DJ. Understanding the results: more about odds ratios. In: *Users' Guide to the Medical Literature*, 2nd ed. New York: McGraw-Hill; 2008:chap 10.2.

†Direct hypothesis testing regarding odds ratios is usually done with confidence intervals; see Chapter 7.

‡Johnson KC, Hu J, Mao Y, Canadian Cancer Registries Epidemiology Research Group. Passive and active smoking and breast cancer risk in Canada, 1994–1997. *Cancer Causes Control*. 2000;11:211–221.

We now have the tools to analyze data measured on a nominal scale. So far we have been focusing on how to demonstrate a difference and quantify the certainty with which we can assert this difference or effect with the P value. Now we turn to the other side of the coin: What does it mean if the test statistic is *not* big enough to reject the hypothesis of no difference?

■ PROBLEMS

5-1 Obtaining a blood sample of arterial blood permits measuring blood pH, oxygenation, and CO_2 elimination in order to see how well the lungs are functioning at oxygenating blood. The blood sample is often drawn from an artery in the wrist, which can be a painful procedure. Shawn Aaron and colleagues* compared the effectiveness of a topical anesthetic gel applied to the skin over the puncture point with a placebo cream. They observed adverse effects (redness, swelling, itching, or bruising) within 24 hours of administering the gel. Three of 36 people receiving the anesthetic gel and 8 of 40 receiving the placebo gel suffered an adverse reaction. Is there evidence of a difference in the rate of adverse effects between the anesthetic gel and the placebo gel?

5-2 Adolescent suicide is commonly associated with alcohol misuse. In a retrospective study involving Finnish adolescents who committed suicide, Sami Pirkola and colleagues†

compared situational factors and family background between victims who abused alcohol and those who did not. Alcohol use was determined by family interview several months following the suicide. Adolescents with alcohol problems, ranging from mild to severe, were classified together in a group called SDAM (Subthreshold or Diagnosable Alcohol Misuse) and compared to victims with no such reported alcohol problems. Some of Pirkola's findings are shown in Table 5-16. Use these data to identify the characteristics of SDAM suicides. Are these factors specific enough to be of predictive value in a specific adolescent? Why or why not?

5-3 The 106 suicides analyzed in Prob. 5-2 were selected from 116 suicides that occurred between April 1987 and March 1988. Eight of the 10 suicides not included in the study were due to lack of family interviews. Discuss the potential problems, if any, associated with these exclusions.

5-4 Major depression can be treated with medication, psychotherapy or a combination of the two. M. Keller and colleagues‡ compared the efficacy of these approaches in outpatients diagnosed with a chronic major depressive disorder. Depression was diagnosed using the 24-item Hamilton Rating Scale for Depression, where a higher score indicates more severe depression. All subjects began the study with a score of at least 20. The investigators

*Aaron, et al. Topical tetracaine prior to arterial puncture: a randomized, placebo-controlled clinical trial. *Respir Med.* 2003;97:1195–1199.

†Pirkola, et al. Alcohol-related problems among adolescent suicides in Finland. *Alcohol Alcohol.* 1999;34:320–328.

‡Keller M, et al. A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression. *N Engl J Med.* 2000;342:1462–1470.

■ TABLE 5-16. Characteristics of Swedish Adolescents Who Committed Suicide

Factor	SDAM Group (n = 44)	Not in SDAM Group (n = 62)
Violent death (shooting, hanging, jumping, traffic)	32	51
Suicide under influence of alcohol	36	25
Blood alcohol concentration (BAC) \geq 150 mg/dL	17	3
Suicide during weekend	28	26
Parental divorce	20	15
Parental violence	14	5
Parental alcohol abuse	17	12
Paternal alcohol abuse	15	9

randomly assigned patients who met study criteria to the three groups—medication (nefazodone), psychotherapy, or both—for 12 weeks then measured remission, defined as having a follow-up score of 8 or less after 10 weeks of treatment. The responses of the people they studied are shown in Table 5-17. Is there any evidence that the different treatments produced different responses? If so, which one seems to work best?

■ TABLE 5-17. Responses to Treatment of Depression

Treatment	Remission	No Remission
Nefazodone	36	131
Psychotherapy	41	132
Nefazodone and psychotherapy	75	104

5-5 In debates over whether or not to pass legislation making all restaurants and bars smoke free opponents of the laws routinely claim that such laws harm the hospitality industry economically and produce economic studies supporting this claim. To assess the association of funding for economic studies supporting this claim, Michelle Scollo and colleagues* tabulated the conclusions of the studies according to the funding source (Table 5-18). Does the data support the claim that tobacco industry funded studies are more likely to conclude that these laws would have negative economic effects? What is the odds ratio for a study concluding a negative economic effect having been supported by the tobacco industry or one of its allies?

5-6 Meta-analysis is an important way to summarize the biomedical literature because they pull together information from many different studies to provide a quantitative estimate of the effect of a treatment or exposure to a toxin. As a result they are often widely cited and influential. To determine whether there are biases in meta-analyses supported by a single pharmaceutical company, Veronica Yank and colleagues† examined the results and conclu-

sions of meta-analyses of the efficacy of anti-hypertensive drugs and the source of funding for the analyses. Table 5-19 presents their data.

■ TABLE 5-18. Relationship between Tobacco Industry Funding and Concluding that Smokefree Laws Hurt the Hospitality Industry

Funded by Tobacco Industry or an Industry Ally?	Study Conclusion	
	Negative Economic Effect	No Effect or Positive Effect
Yes	29	2
No	2	60

5-7 Authorship in biomedical publications establishes accountability, responsibility, and credit. The International Committee of Medical Journal Editors established authorship criteria in 1985, which boil down to playing an active role in the research and writing of the paper and being in a position to take responsibility for a paper’s scientific content.‡ Misappropriation of authorship undermines the integrity of the authorship system. There are two ways that authorship is misappropriated: honorary authorship, when someone (typically a department or division chair or the person who obtained funding for the project) who did not actually participate in preparing the paper, is listed as an author, and ghost authorship, when someone who played an important role in writing the paper is not listed as an author. To investigate the prevalence of honorary and ghost authorship in medical journals, Annette Flanagin and colleagues§ sent questionnaires to a random sample of corresponding authors for papers published in three highly circulated general medical journals (*Annals of Internal Medicine*, *Journal of the American Medical Association*, and *New England Journal of Medicine*) and three specialty journals (*American Journal of Cardiology*, *American Journal of Medicine*, and *American Journal of Obstetrics and Gynecology*). Their results are shown in Table 5-20. Are there differences

*Scollo M, et al. Review of the quality of studies on the economic effects of smoke-free policies on the hospitality industry. *Tobacco Control*. 2003;12:13–20.

†Yank V, et al. Financial ties and concordance between results and conclusions in meta-analyses: retrospective cohort study. *Br Med J*. 2007;335:1202–1205.

‡The full guidelines, which are accepted by most medical journals, are available at: International Committee of Medical Journal Editors. Guidelines on authorship. *BMJ*. 1985;291:722.

§Flanagin A, et al. Prevalence of articles with honorary authors and ghost authors in peer-reviewed medical journals. *JAMA*. 1998;280:222–224.

■ **TABLE 5-19. Relationship between Drug Company Funding and Conclusions of Meta-analyses on Effects of Their Drugs**

Funding Source	Number of Studies	Study Outcome	
		Number (%) with Favorable Results	Number (%) with Favorable Conclusions
One drug company	49	27 (55%)	45 (92%)
All other*	75	49 (65%)	55 (73%)

*Includes studies supported by several drug companies, nonprofits, and papers where the source of funding was not stated.

■ **TABLE 5-20. Authorship Patterns in Several Well-regarded Journals**

Journal	Total Number of Articles	Articles with Honorary Authors	Articles with Ghost Authors
<i>American Journal of Cardiology</i>	137	22	13
<i>American Journal of Medicine</i>	113	26	15
<i>American Journal of Obstetrics and Gynecology</i>	125	14	13
<i>Annals of Internal Medicine</i>	104	26	16
<i>Journal of the American Medical Association</i>	194	44	14
<i>New England Journal of Medicine</i>	136	24	22

in the patterns of honorary authorship and ghost authorship among the different journals? Are there differences in patterns of honorary and ghost authorship between the specialty journals and large circulation generalist journals?

5-8 Dioxin is one of the most toxic synthetic environmental contaminants. An explosion at a herbicide plant in Sevaso, Italy in 1976 released large amounts of this long-lasting contaminant into the environment. Because exposure to dioxin during development is known to be dangerous, researchers have been carefully following the health status of exposed people and their children in Sevaso and surrounding areas. Peter Mocarelli and colleagues* measured the serum concentration of dioxin in potentially exposed parents and analyzed the number of male and female babies born after 1976. They found that when both parents were exposed to greater than 15 parts per trillion (ppt) of dioxin the proportion of girl babies born was significantly increased compared to couples not exposed to this amount of dioxin. Mocarelli and col-

*Mocarelli P, et al. Paternal concentrations of dioxin and sex ratio of offspring. *Lancet*. 2000;355:1858–1863.

leagues also investigated whether there were differences in the proportion of female babies born if only one parent was exposed to greater than 15 ppt of dioxin and whether the sex of the parent (mother or father) made a difference (Table 5-21). Are there differences in the proportion of female babies born when only one parent is exposed to greater than 15 ppt of dioxin?

5-9 Bipolar disorder is a disabling mental illness that is characterized by episodes of elevated or irritable mood and depression. Lithium carbonate has been the standard therapy for treating bipolar disorder. In more recent years lithium has been replaced with sodium valproate because of a wider range of useful doses and fewer side effects. The BALANCE investigators* conducted a randomized open label (in which the subjects and investigators knew who was taking which drug) clinical trial comparing valproate in combination with lithium with valproate alone. The end point was having an emergent mood episode during a 24-month period. Fifty-nine (54%) of the 110 people in the lithium group and 76 (69%) of the 110 in the valproate group had events during follow-up. What is the relative risk for an event for people being treated with valproate compared to lithium (the tradi-

tional treatment)? Is this difference statistically significant? What is the number needed to treat?

TABLE 5-21. Prenatal Dioxin Exposure and Baby Gender

Parental Exposure to Dioxin	Female Babies	Male Babies
	Father exposed; mother unexposed	105
Father unexposed; mother exposed	100	120

5-10 The chance of contracting disease X is 10%, regardless of whether or not a given individual has disease A or disease B. Assume that you can diagnose all three diseases with perfect accuracy and that in the entire population 1000 people have disease A and 1000 have disease B. People with X, A, and B have different chances of being hospitalized. Specifically, 50% of the people with A, 20% of the people with B, and 40% of the people with X are hospitalized. Then

- Out of the 1000 people with A, 10% (100 people) also have X; 50% (500 people) are hospitalized because they have A. Of the remaining 900 (who also have X), 40% (360 people) are hospitalized because of X. Therefore, 460 people will be hospitalized with both A and X.
- Out of the 900 people with A but not X, 50% are hospitalized for disease A (450 people).
- Out of the 1000 with B, 10% (100 people) also have X; 20% (200 people) are hospitalized because of B, and of the 800 people who are not hospitalized because of B, 40% (320 patients) are hospitalized because they have X. Thus, 520 people with B and X are in the hospital.
- Of the 900 with B but not X, 20% (180 people) are hospitalized because they have disease B.

Thus, Table 5-22 summarizes how a hospital-based investigator will encounter these patients in the hospital. Is there a statistically significant difference in the chances that an individual has X depending on whether or not he has A or B in the sample of patients the hospital-based investigator will encounter? Would the investigator reach the same conclusion if she could observe the entire population? If not, explain why.[†]

*The BALANCE Investigators. Lithium plus valproate combination therapy versus monotherapy for relapse prevention in bipolar i disorder (BALANCE): a randomized open-label trial. *Lancet*. 2010;375:385–394.

TABLE 5-22. Relationship Between Three Diseases in Hospitalized Patients

	Disease X	No Disease X
Disease A	70	450
Disease B	52	180

5-11 Cigarette smoking is associated with increased incidence of many types of cancers. Jian-Min Yuan and colleagues[‡] wanted to investigate whether cigarette smoking was also associated with increased risk of renal cell cancer. They recruited patients with renal cell cancer from the Los Angeles County Cancer Surveillance Program to serve as cases in a retrospective case-control study. Control subjects without renal cell cancer were matched on sex, age (within 5 years), race, and neighborhood of residence to each case subject. After recruiting a total of 2314 subjects for the study, Yuan and colleagues visited subjects in their homes and interviewed them about their smoking habits, both past and present (Table 5-23). What effect does smoking cigarettes have on the risk of developing renal cell cancer?

TABLE 5-23. Smoking and Renal Cell Cancer

	Number of People	
	Renal Cell Cancer	No Cancer
Ever smoked cigarettes	800	713
Never smoked cigarettes	357	444

5-12 Yuan and colleagues also collected information from subjects who had quit smoking. Based on the data in Table 5-24, is there any evidence that stopping smoking reduces risk of developing renal cell cancer compared to current smokers?

[†]This example is from Mainland D, The risk of fallacious conclusions from autopsy data on the incidence of diseases with applications to heart disease. *Am Heart J*.1953;45:644–654.

[‡]Yuan J-M, et al. Tobacco use in relation to renal cell carcinoma. *Cancer Epidemiol Biomarkers Prev*. 1998;7:429–433.

■ **TABLE 5-24. Former vs. Current Smokers and Renal Cell Cancer**

	Number of People	
	Renal Cell Cancer	No Cancer
More than 20 years since quitting	169	177
Current smokers	337	262

5-13 Many postmenopausal women are faced with the decision of whether they want to take hormone replacement therapy or not. Benefits of hormone replacement include decreased risk of cardiovascular disease and osteoporosis. However, hormone replacement therapy has also been associated with increased risk of breast cancer and endometrial cancer. Francine Grodstein and colleagues* investigated the relationship between hormone replacement therapy and overall mortality in a large group of postmenopausal women. The women used in this study were selected from a sample of registered nurses participating in the Nurses' Health Study. This prospective study has been tracking the health status of a large group of registered nurses since 1976, updating information every 2 years. Women became eligible for Grodstein's study when they became menopausal and were included as long as they did not report a history of cardiovascular disease or cancer on the original 1976 questionnaire. Based on the data in Table 5-25, is there any evidence that the risk of death differs in women who were identified as currently using hormone replacement therapy?

*Grodstein F, et al. Postmenopausal hormone therapy and mortality. *N Engl J Med.* 1997;336:1769–1775.

■ **TABLE 5-25. Current Use of Hormone Replacement Therapy and Survival**

	Number of People	
	Deceased	Alive
Currently using hormone replacement therapy	574	8483
Never used hormone replacement therapy	2051	17,520

5-14 Based on the data in Table 5-26, is there an increase in risk of death in women who reported past hormone replacement therapy use compared to women who never used it?

■ **TABLE 5-26. Past Use of Human Therapy and Survival**

	Number of People	
	Deceased	Alive
Past use of hormone replacement therapy	1012	8621
Never used hormone replacement therapy	2051	17,520

This page intentionally left blank

What Does “Not Significant” Really Mean?

Thus far, we have used statistical methods to reach conclusions by seeing how compatible the observations were with the null hypothesis that the treatment had no effect. When the data were unlikely to occur if this null hypothesis was true, we rejected it and concluded that the treatment had an effect. We used a test statistic (F , t , z , or χ^2) to quantify the difference between the actual observations and those we would expect if the null hypothesis of no effect were true. We concluded that the treatment had an effect if the value of this test statistic was bigger than 95% of the values that would occur if the treatment had no effect. When this is so, it is common for medical investigators to report a *statistically significant* effect. On the other hand, when the test statistic is not big enough to reject the hypothesis of no treatment effect, investigators often report *no statistically significant difference* and then discuss their results as if they had proven that the treatment had no effect. *All they really did was fail to demonstrate that it did have an effect.* The distinction between positively demonstrating that a treatment had no effect and failing to demonstrate that it did have an effect is subtle but very important, especially in the light of the small numbers of subjects included in most clinical studies.*

*This problem is particularly encountered in small clinical studies in which there are no “failures” in the treatment group. This situation often leads to overly optimistic assessments of therapeutic efficacy. See Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. *JAMA*. 1983;249:1743–1745.

As already mentioned in our discussion of the t test, the ability to detect a treatment effect with a given level of confidence depends on the size of the treatment effect, the variability within the population, and the size of the samples used in the study. Just as bigger samples make it more likely that you will be able to detect an effect, smaller sample sizes make it harder. In practical terms, this fact means that studies of therapies that involve only a few subjects and fail to reject the null hypothesis of no treatment effect may arrive at this result because the statistical procedures lacked the *power* to detect the effect because of a too small sample size, even though the treatment did have an effect. Conversely, considerations of the power of a test permit you to compute the sample size needed to detect a treatment effect of given size that you believe is present.

■ AN EFFECTIVE DIURETIC

Now, we make a radical departure from everything that has preceded: we assume that the treatment *does* have an effect.

Figure 6-1 shows the same population of people we studied in Figure 4-3 except that this time the drug given to increase daily urine production works. It increases the average urine production for members of this population from 1200 to 1400 mL/day. Figure 6-1A shows the distribution of daily urine production for all 200 members of the population in the control (placebo) group, and Figure 6-1B shows the distribution of urine

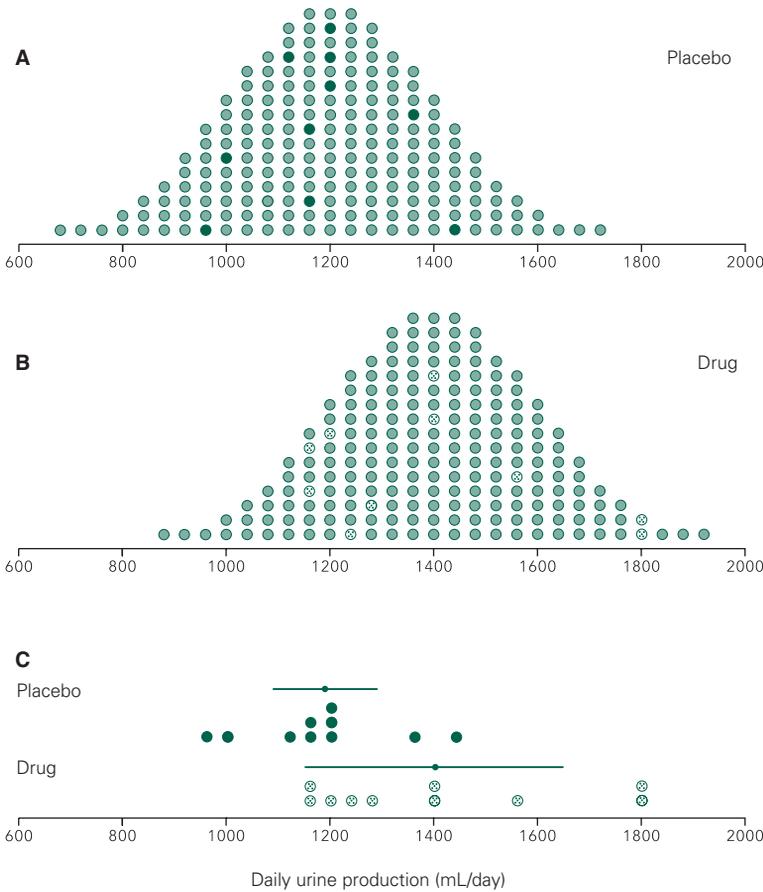


FIGURE 6-1. Daily urine production in a population of 200 people while they are taking a placebo and while they are taking an effective diuretic that increases urine production by 200 mL/day on the average. Panels **A** and **B** show the specific individuals selected at random for study. Panel **C** shows the results as they would appear to the investigator. $t = 2.447$ for these observations. Since the critical value of t for $P < .05$ with $2(10 - 1) = 18$ degrees of freedom is 2.101, the investigator would probably report that the diuretic was effective.

production for all 200 members of the population in the diuretic group.

More precisely, the population of people taking the placebo consist of a normally distributed population with mean $\mu_{\text{pla}} = 1200$ mL/day and the population of people taking the drug consist of a normally distributed population with a mean of $\mu_{\text{dr}} = 1400$ mL/day. Both populations have the same standard deviation, $\sigma = 200$ mL/day.

Of course, an investigator cannot observe all members of the population, so he or she selects two groups of 10 people at random, gives one group the diuretic and the other a placebo, and measures their daily urine production. Figure 6-1C shows what the investigator would see. The people receiving a placebo produce an average of 1180 mL/day, and those receiving the drug produce an average of 1400 mL/day. The standard deviations of these two samples are 144 and 245 mL/day,

respectively. The pooled estimate of the population variance is

$$s^2 = \frac{1}{2}(s_{\text{dr}}^2 + s_{\text{pla}}^2) = \frac{1}{2}(245^2 + 144^2) = 40,381 = 201^2$$

The value of t associated with these observations is

$$t = \frac{\bar{X}_{\text{dr}} - \bar{X}_{\text{pla}}}{\sqrt{(s^2/n_{\text{dr}}) + (s^2/n_{\text{pla}})}} = \frac{1400 - 1180}{\sqrt{(210^2/10) + (210^2/10)}} = 2.447$$

which exceeds 2.101, the value that defines the most extreme 5% of possible values of the t test statistic when the two samples are drawn from the same population. (There are $\nu = n_{\text{dr}} + n_{\text{pla}} - 2 = 10 + 10 - 2 = 18$ degrees of freedom.) The investigator would conclude that the observations are not consistent with the assumption that two samples came from the same population and report

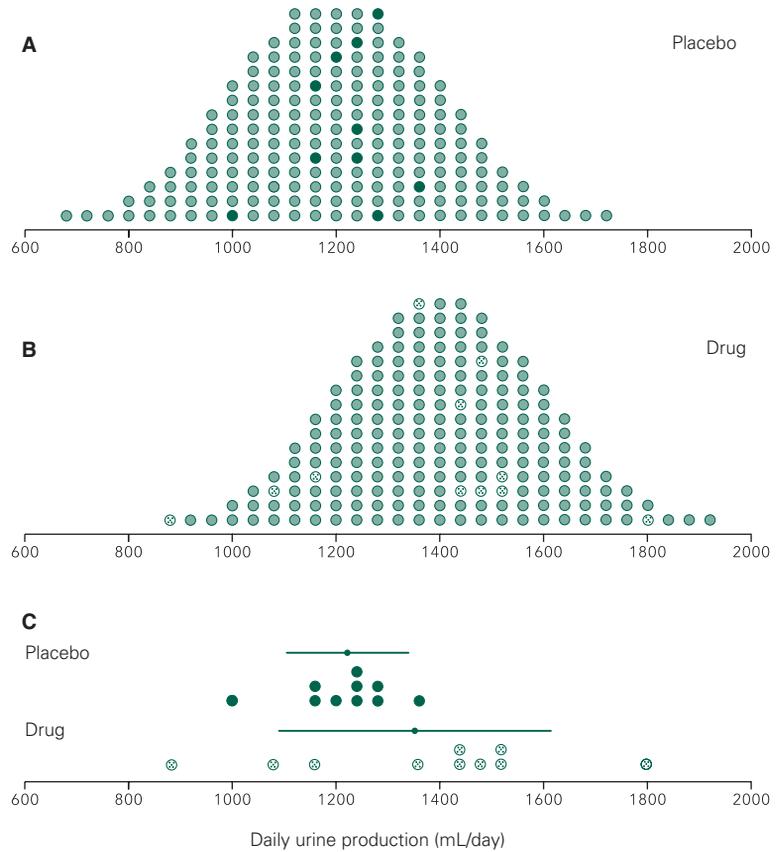


FIGURE 6-2. There is nothing special about the two random samples shown in Figure 6-1. This illustration shows another random sample of two groups of 10 people each selected at random to test the diuretic (**A** and **B**) and the results as they would appear to the investigator (**C**). The value of t associated with these observations is only 1.71, not great enough to reject the hypothesis of no drug effect with $P < 0.05$, that is, $\alpha = 0.05$. If the investigator reported the drug had no effect, he or she would be wrong.

that the drug increased urine production. And he or she would be right.

Of course, there is nothing special about the two random samples of people selected for the experiment. Figure 6-2 shows two more groups of people selected at random to test the drug, together with the results as they would appear to the investigator. In this case, the mean urine production is 1216 mL/day for the people given the placebo and 1368 mL/day for the people taking the drug. The standard deviations of urine production in the two samples are 97 and 263 mL/day, respectively, so the pooled estimate of the variance is $1/2(97^2 + 263^2) = 198^2$. The value of t associated with these observations is

$$t = \frac{1368 - 1216}{\sqrt{(198^2/10) + (198^2/10)}} = 1.71$$

which is less than 2.101. Had the investigator selected these two groups of people for testing, he or she would not

have obtained a value of t large enough to reject the hypothesis that the drug had no effect and probably reported “no significant difference.” If the investigator went on to conclude that the drug had no effect, he or she would be wrong.

Notice that this is a different type of error from that discussed in Chapters 3 to 5. In the earlier chapters, we were concerned with *rejecting* the hypothesis of no effect when it was true. Now we are concerned with *not rejecting it when it is not true*. This situation is called a *Type II error* or β error.

What are the chances of making this second kind of error?

Just as we could repeat this experiment more than 10^{27} times when the drug had no effect to obtain the distribution of possible values of t (compare with the discussion of Fig. 4-4), we can do the same thing when the drug does have an effect. Figure 6-3 shows the results of 200 such experiments; 111 out of the resulting values of t fall at or

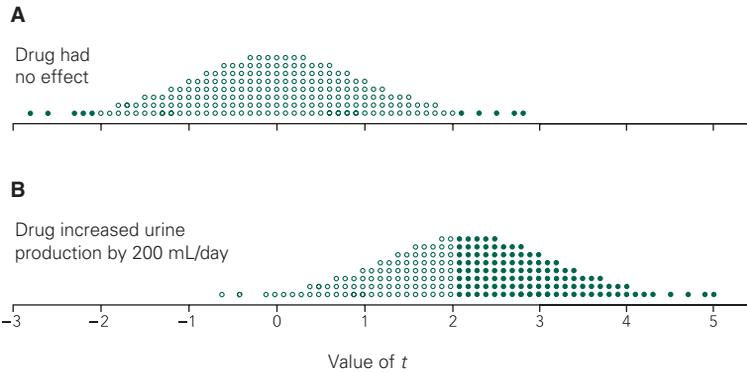


FIGURE 6-3. (A) The distribution of values of the t test statistic computed from 200 experiments that consisted of drawing two samples of size 10 each from a single population; this is the distribution we would expect if the diuretic had no effect on urine production is centered on zero. (compare with Fig. 4-4A.) (B) The distribution of t values from 200 experiments in which the drug increased average urine production by 200 mL/day. $t = 2.1$ defines the most extreme 5% of the possible values of t when the drug has no effect; 111 of the 200 values of t we would expect to observe from our data fall above this point when the drug increases urine production by 200 mL/day. Therefore, there is a 56% chance that we will conclude that the drug actually increases urine production from our experiment.

above 2.101, the value we used to define a “big” t . Put another way, if we wish to keep the P value at or below 5%, there is a $111/200 = 56\%$ chance of concluding that the diuretic increases urine output when average urine output actually increases by 200 mL/day. We say the *power* of the test is .56. *The power quantifies the chance of detecting a real difference of a given size.*

Alternatively, we could concentrate on the 89 of the 200 experiments that produced t values below 2.101, in which case we would fail to reject the hypothesis that the treatment had no effect and be wrong. Thus, there is an $89/200 = 44\% = .44$ chance of continuing to accept the hypothesis of no effect when the drug really increased urine production by 200 mL/day on the average.

■ TWO TYPES OF ERRORS

Now we have isolated the two different ways the random-sampling process can lead to erroneous conclusions. These two types of errors are analogous to the false-positive and false-negative results one obtains from diagnostic tests. Before this chapter we concentrated on controlling the likelihood of making a false-positive error, that is, concluding that a treatment has an effect when it really does

not. In keeping with tradition, we have generally sought to keep the chances of making such an error below 5%; of course, we could arbitrarily select any cutoff value we wanted at which to declare the test statistic “big.” Statisticians denote the maximum acceptable risk of this error by α , the Greek letter alpha. If we reject the hypothesis of no effect whenever $P < .05$, $\alpha = 0.05$ or 5%. If we actually obtain data that lead us to reject the null hypothesis of no effect when the null hypothesis of no effect is true, statisticians say that we have made a *Type I error*. All this logic is relatively straightforward because we have specified how much we believe the treatment affects the variable of interest, that is, not at all.

What about the other side of the coin, the chance of making a false-negative conclusion and not reporting an effect when one exists? Statisticians denote the chance of erroneously accepting the hypothesis of no effect by β , the Greek letter beta. The chance of detecting a true positive, that is, reporting a statistically significant difference when the treatment really produces an effect, is $1 - \beta$. The *power* of the test that we discussed earlier is equal to $1 - \beta$. For example, if a test has power equal to .56, there is a 56% chance of actually reporting a statistically significant effect when one is really present. Table 6-1 summarizes these definitions.

■ **TABLE 6-1. Types of Erroneous Conclusions in Statistical Hypothesis Testing**

Conclude From Observations	Actual Situation	
	Treatment Has an Effect	Treatment Has No Effect
Treatment has an effect	True positive Correct conclusion $1 - \beta$	False positive Type I error α
Treatment has no effect	False negative Type II error β	True negative Correct conclusion $1 - \alpha$

■ WHAT DETERMINES A TEST'S POWER?

So far we have developed procedures for estimating and controlling the Type I, or α , error. Now we turn our attention to keeping the Type II, or β , error as small as possible. In other words, we want the power to be as high as possible. In theory, this problem is not very different from the one we already solved with one important exception. Since the treatment has an effect, *the size of this effect influences how easy it is to detect*. Large effects are easier to detect than small ones. To estimate the power of a test, you need to specify how small an effect is worth detecting.

Just as with false positives and false negatives in diagnostic testing, the Type I and Type II errors are intertwined. As you require stronger evidence before reporting that a treatment has an effect, that is, make α smaller, you also increase the chance of missing a true effect, that is, make β bigger or power smaller. The only way to reduce both α and β simultaneously is to increase the sample size, because with a larger sample you can be more confident in your decision, whatever it is.

In other words, the power of a given statistical test depends on three interacting factors:

- *The risk of error you will tolerate when rejecting the hypothesis of no treatment effect.*
- *The size of the difference you wish to detect relative to the amount of variability in the populations.*
- *The sample size.*

To keep things simple, we will examine each of these factors separately.

The Size of the Type I Error α

Figure 6-3 showed the complementary nature of the maximum size of the Type I error α and the power of the test. The acceptable risk of erroneously rejecting the hypothesis

of no effect, α , determines the critical value of the test statistic above which you will report that the treatment had an effect, $P < \alpha$. (We have usually taken $\alpha = 0.05$.) This critical value is defined from the distribution of the test statistic for all possible experiments with a specific sample size *given that the treatment had no effect*. The power is the proportion of possible values of the test statistic that fall above this cutoff value *given that the treatment had a specified effect* (here a 200 mL/day increase in urine production). Changing α , or the P value required to reject the hypothesis of no difference, moves this cutoff point, affecting the power of the test.

Figure 6-4 illustrates this point further. Figure 6-4A essentially reproduces Figure 6-3 except that it depicts the distribution of t values for all 10^{27} possible experiments involving two groups of 10 people as a continuous distribution. The top part, copied from Figure 4-4D, shows the distribution of possible t values (with $\nu = 10 + 10 - 2 = 18$ degrees of freedom) that would occur if the drug did not affect urine production. Suppose we require $P < .05$ before we are willing to assert that the observations were unlikely to have arisen from random sampling rather than the effect of the drug. According to the table of critical values of the t distribution (see Table 4-1), for $\nu = 18$ degrees of freedom, 2.101 is the (two-tail) critical value that defines the most extreme 5% of possible values of the t test statistic if the null hypothesis of no effect of the diuretic on urine production is true. In other words, when we make $\alpha = 0.05$, in which case -2.101 and $+2.101$ delimit the most extreme 5% of all possible t values we would expect to observe if the diuretic did not affect urine production.

We know, however, that the drug actually increased average urine production by $\mu_{\text{dr}} - \mu_{\text{pla}} = 200$ mL/day. Therefore, the actual distribution of possible values of t associated with our experiment will not be given by the distribution at the top of Figure 6-4 (which assumes that

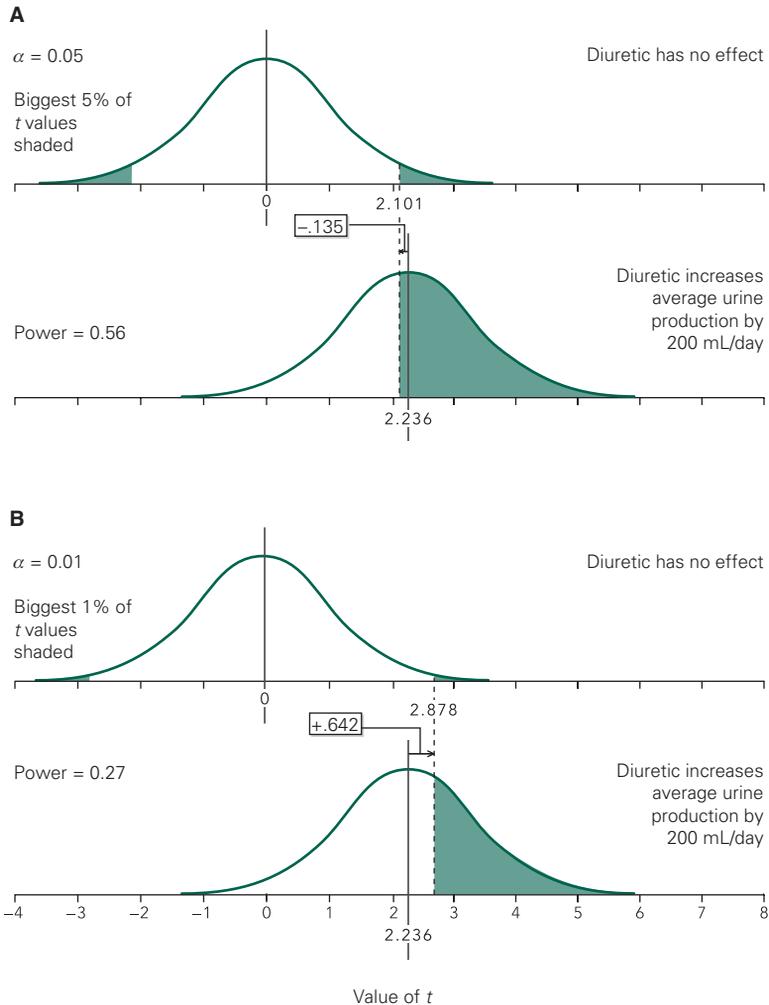


FIGURE 6-4. (A) The top panel shows the distribution of the t test statistic that would occur if the null hypothesis was true and the diuretic did not affect urine production. The distribution is centered on 0 (because the diuretic has no effect on urine production) and, from Table 4-1, $t = +2.101$ (and -2.101) define the (two-tail) 5% most extreme values of the t test statistic that would be expected to occur by chance if the drug had no effect. The second panel shows the actual distribution of the t test statistic that occurs when the diuretic increases urine output by 200 mL/day; the distribution of t values is shifted to the right, so the distribution is now centered on 2.236. The critical value of 2.101 is $-.135$ below 2.236, the center of this shifted distribution. From Table 6-2, .56 of the possible t values fall in the one-tail above $-.135$, so we conclude that the power of a t test to detect a 200 mL/day increase in urine production is 56%. (The power also includes the portion of the t distribution in the lower tail below -2.101 , but because this area is so small we will ignore it.) **(B)** If we require more evidence before rejecting the null hypothesis of no difference by reducing α to 0.01, the critical value of t that must be exceeded to reject the null hypothesis increases to 2.878 (and -2.878). Since the effect of the diuretic is unchanged, the actual distribution of t remains centered on 2.236; the critical value of 2.878 is $.642$ above 2.236, the center of the actual t distribution. From Table 6-2, .27 of the possible t values fall in the tail above $.642$, so the power of the test drops to 27%.

the null hypothesis that $\mu_{dr} - \mu_{pla} = 0$ is true and so is centered on 0).

To determine where the actual distribution of values of the t test statistic will be centered, recall from Chapter 4 that the t test statistic to compare two means, is

$$t = \frac{\bar{X}_{dr} - \bar{X}_{pla}}{\sqrt{(s^2/n_{dr}) + (s^2/n_{pla})}}$$

$\bar{X}_{dr} - \bar{X}_{pla}$ computed from the observations is an estimate of the actual difference in mean urine production between the populations of people taking the drug and taking the placebo, $\mu_{dr} - \mu_{pla} = 200$ mL/day. The observed standard deviation, s , is an estimate of the standard deviation of the underlying populations, σ , which, from Figure 6-1, is 200 mL/day. Therefore, we would expect the actual distribution of the t test statistic to be centered on

$$t' = \frac{\mu_{dr} - \mu_{pla}}{\sqrt{(\sigma^2/n_{dr}) + (\sigma^2/n_{pla})}}$$

n_{dr} and n_{pla} both are 10, so the actual distribution of the t test statistic will be centered on

$$t' = \frac{200}{\sqrt{(200^2/10) + (200^2/10)}} = 2.236$$

The lower distribution in Figure 6-4A shows this actual distribution of possible t values associated with our experiment: the t distribution is moved to the right to be centered on 2.236 (rather than 0, as it was under the null hypothesis). Fifty-six percent of these possible values of t , that is, 56% of the area under the curve, fall above the 2.101 cutoff, so we say the power of the test is .56.

In other words, if the drug increases average urine production by 200 mL/day in this population and we do an experiment using two samples of 10 people each to test the drug, there is a 56% chance that we will conclude that the drug is effective ($P < .05$). To understand how we obtain this estimate of the power, we need to consult another table of critical values of the t distribution, one that gives the *one-tail* probability of being in the upper tail of the distribution as a function of the value of t (Table 6-2). The information in this table is essentially the same as in Table 4-1, with the difference that it presents critical values for one tail only, so the P values associated with each value of t in this table are half the corresponding

values in Table 4-2. For example, the critical value of $t = +2.101$, the two-tail critical value associated with $P = .05$ for $\nu = 18$ degrees of freedom in Table 4-2, corresponds to a one (upper) tail probability of .025 in Table 6-2. This situation arises because in a two-tail test of the null hypothesis of *no difference*, half the risk of a false-positive conclusion resides in the upper tail of the distribution of possible values of t and the other half resides in the lower end of the distribution, below -2.101 in this case. Note, from Table 6-2, that the probability of being in the lower tail of the distribution of possible values of t (with $\nu = 18$) at or below -2.101 is .025. The .025 probability of being *at or below* -2.101 plus the .025 probability of being *at or above* $+2.101$ add up to the .05 two-tailed probability we found in Table 4-1.

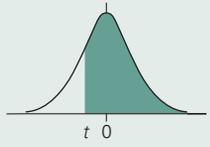
As noted above, the actual distribution of values of the t test statistic given that there is actually a 200 mL/day increase in urine production with the diuretic is centered on 2.236 rather than 0, as it would be if the null hypothesis was true. The critical value of 2.101 that leads us to reject the null hypothesis (from the top distribution in Fig. 6-4A) is below the center of the actual distribution of the t test statistic by $2.101 - 2.236 = -.135$. We can use Table 6-2 to determine the probability of being in the upper tail of this t distribution* (with $\nu = 18$ degrees of freedom) is .56 (between .60, which corresponds to $-.257$ and .50, which corresponds to .000), yielding the power of 56%.

Conversely, we can say that β , the probability that we will make a false negative, or Type II, error and accept the null hypothesis of no effect when it is not true is $1 - .56 = .44 = 44\%$. Alternatively, we can use Table 6-2 to note that the probability of being in the lower tail of the t distribution (at or below $-.135$) is .44.

Now look at Figure 6-4B. The two distributions of t values are identical to those in Figure 6-4A. (After all, the drug's true effect is still the same.) This time, however, we will insist on stronger evidence before concluding that the drug actually increases urine production. We will require that the test statistic fall in the most extreme 1% of possible values before concluding that the data are inconsistent with the null hypothesis that the drug has no effect. Thus, $\alpha = 0.01$ and t must be below -2.878 or above $+2.878$ to fall

*Technically, we should also consider the portion of the actual t distribution in the lower tail of Figure 6-4A below -2.101 , but this portion is extremely small so we will ignore it.

■ TABLE 6-2. Critical Values of t (One-Tailed)



Probability of Larger Value (Upper Tail)										
	.995	.99	.98	.975	.95	.90	.85	.80	.70	.60
Probability of Smaller Value (Lower Tail)										
ν	.005	.01	.02	.025	.05	.10	.15	.20	.30	.40
2	-9.925	-6.965	-4.849	-4.303	-2.920	-1.886	-1.386	-1.061	-0.617	-0.289
4	-4.604	-3.747	-2.999	-2.776	-2.132	-1.533	-1.190	-0.941	-0.569	-0.271
6	-3.707	-3.143	-2.612	-2.447	-1.943	-1.440	-1.134	-0.906	-0.553	-0.265
8	-3.355	-2.896	-2.449	-2.306	-1.860	-1.397	-1.108	-0.889	-0.546	-0.262
10	-3.169	-2.764	-2.359	-2.228	-1.812	-1.372	-1.093	-0.879	-0.542	-0.260
12	-3.055	-2.681	-2.303	-2.179	-1.782	-1.356	-1.083	-0.873	-0.539	-0.259
14	-2.977	-2.624	-2.264	-2.145	-1.761	-1.345	-1.076	-0.868	-0.537	-0.258
16	-2.921	-2.583	-2.235	-2.120	-1.746	-1.337	-1.071	-0.865	-0.535	-0.258
18	-2.878	-2.552	-2.214	-2.101	-1.734	-1.330	-1.067	-0.862	-0.534	-0.257
20	-2.845	-2.528	-2.197	-2.086	-1.725	-1.325	-1.064	-0.860	-0.533	-0.257
25	-2.787	-2.485	-2.167	-2.060	-1.708	-1.316	-1.058	-0.856	-0.531	-0.256
30	-2.750	-2.457	-2.147	-2.042	-1.697	-1.310	-1.055	-0.854	-0.530	-0.256
35	-2.724	-2.438	-2.133	-2.030	-1.690	-1.306	-1.052	-0.852	-0.529	-0.255
40	-2.704	-2.423	-2.123	-2.021	-1.684	-1.303	-1.050	-0.851	-0.529	-0.255
60	-2.660	-2.390	-2.099	-2.000	-1.671	-1.296	-1.045	-0.848	-0.527	-0.254
120	-2.617	-2.358	-2.076	-1.980	-1.658	-1.289	-1.041	-0.845	-0.526	-0.254
∞	-2.576	-2.326	-2.054	-1.960	-1.645	-1.282	-1.036	-0.842	-0.524	-0.253
Normal	-2.576	-2.326	-2.054	-1.960	-1.645	-1.282	-1.036	-0.842	-0.524	-0.253

in the most extreme 1% of values. The top part of Figure 6-1B shows this cutoff point. The actual distribution of the t test statistic is still centered on 2.236, so the 2.878 critical value is now above the center of this distribution by $2.878 - 2.236 = .642$. From Table 6-2, we find that only .27 or 27% of the actual distribution of t falls above 2.878 in Figure 6-4B, so the power of the test has fallen to .27. In other words, there is less than an even chance that we will report that the drug is effective even though it actually is.

By requiring stronger evidence that there be a treatment effect before reporting it we have decreased the chances of erroneously reporting an effect (a Type I error), but we have increased the chances of failing to detect a difference when one actually exists (a Type II error) because we decreased the power of the test. This trade-off always exists.

The Size of the Treatment Effect

We just demonstrated that the power of a test decreases as we reduce the acceptable risk of making a Type I error, α . The entire discussion was based on the fact that the drug increased average urine production by 200 mL/day, from 1200 to 1400 mL/day. Had this change been different, the actual distribution of t values connected with the experiment also would have been different. In other words, the power of a test depends on the size of the difference to be detected.

Let us consider three specific examples. Figure 6-5A shows the t distribution (the distribution of possible values of the t statistic) for a sample size of 10 if the diuretic had no effect and the two treatment groups could be considered two random samples drawn from the same population. The most extreme 5% of the values are shaded, just as in Figure

■ TABLE 6-2. Critical Values of t (One-Tailed) (Continued)

Probability of Larger Value (Upper Tail)										
.50	.40	.30	.20	.15	.10	.05	.025	.02	.01	.005
Probability of Smaller Value (Lower Tail)										
.50	.60	.70	.80	.85	.90	.95	.975	.98	.99	.995
0	0.289	0.617	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925
0	0.271	0.569	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604
0	0.265	0.553	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707
0	0.262	0.546	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355
0	0.260	0.542	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169
0	0.259	0.539	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055
0	0.258	0.537	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977
0	0.258	0.535	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921
0	0.257	0.534	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878
0	0.257	0.533	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845
0	0.256	0.531	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787
0	0.256	0.530	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750
0	0.255	0.529	0.852	1.052	1.306	1.690	2.030	2.133	2.438	2.724
0	0.255	0.529	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704
0	0.254	0.527	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660
0	0.254	0.526	0.845	1.041	1.289	1.658	1.980	2.076	2.358	2.617
0	0.253	0.524	0.842	1.036	1.282	1.645	1.960	2.054	2.326	2.576
0	0.253	0.524	0.842	1.036	1.282	1.645	1.960	2.054	2.326	2.576

6-4. Figure 6-5B shows the distribution of t values we would expect if the drug increased urine production an average of 200 mL/day over the placebo; 56% of the possible values are beyond -2.101 or $+2.101$, so the power of the test is .56. (So far we are just recapitulating the results in Fig. 6-4). Now, suppose that the drug only increased urine production by 100 mL/day. In this case, as Figure 6-5C shows, the actual distribution of the t test statistic will no longer be centered on 0, but on

$$t' = \frac{100}{\sqrt{(200^2/10) + (200^2/10)}} = 1.118$$

Thus, we need to determine the fraction of the actual possible values of the t distribution that fall above $2.101 - 1.118 = .983$. The sample size is the same as before

($n = 10$ in each group), so there are still $\nu = 10 + 10 - 2 = 18$ degrees of freedom. From Table 6-2 we find that .17 of the possible values fall above .983, so the power of the test to detect a 100 mL/day change in urine production is only .17 (or 17%). In other words, there is less than a 1 in 5 chance that doing a study of two groups of 10 people would detect a change in urine production of 100 mL/day if we required that $P < .05$ before reporting an effect.

Finally, Figure 6-5D shows the distribution of t values that would occur if the drug increased urine production by an average of 400 mL/day. Because of this larger effect, the actual distribution of the t test statistic will be centered on

$$t' = \frac{400}{\sqrt{(200^2/10) + (200^2/10)}} = 4.472$$

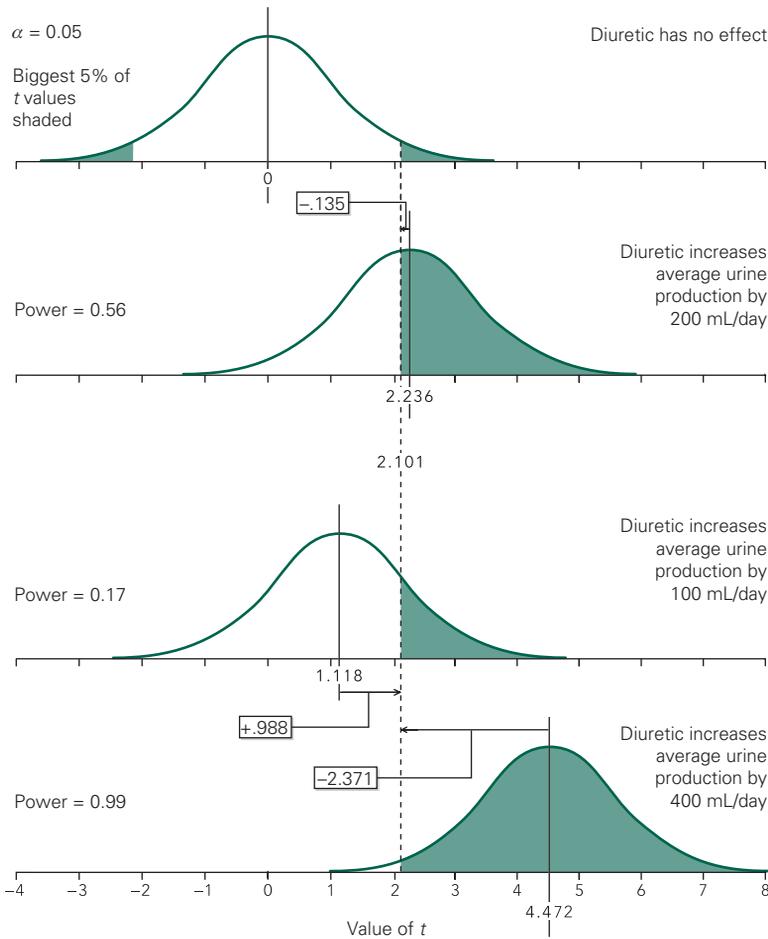


FIGURE 6-5. The larger the size of the treatment effect, the further the actual distribution of the t test statistic will shift away from zero; the more of the actual distribution of t values will exceed the critical value of 2.101 that determines the most extreme (two-tail) 5% of the values of t that will occur if the null hypothesis of no effect is true. As a result, the greater the effect of the diuretic, the greater the power to detect the fact that the diuretic increases urine production.

The power of the test to detect this difference will be the fraction of the t distribution larger than $2.101 - 4.472 = -2.371$. From Table 6-2, with $\nu = 18$ degrees of freedom, .985 of all possible t values fall above 2.371, so the power of the test is 99%. The chance is quite good that our experiment will lead to the conclusion that the diuretic affects urine production (with $P < .05$).

Figure 6-5 illustrates the general rule: *It is easier to detect big differences than small ones.*

We could repeat this process for all possible sizes of the treatment effect, from no effect at all up to very large effects, then plot the power of the test as it varies with the change in urine production actually produced by the drug. Figure 6-6 shows a plot of the results, called a *power function*, of the test. It quantifies how much easier

it is to detect a change (when we require a value of t corresponding to $P < .05$ and two samples of 10 people each) in urine production as the actual drug effect gets larger and larger. This plot shows that if the drug increases urine production by 200 mL/day, there is a 55% chance that we will detect this change with the experiment designed as we have it; if urine production increases by 350 mL/day, the chance of our detecting this effect improves to 95%.

The Population Variability

The power of a test increases as the size of the treatment effect increases, but the variability in the population under study also affects the likelihood with which we can detect a treatment effect of a given size.

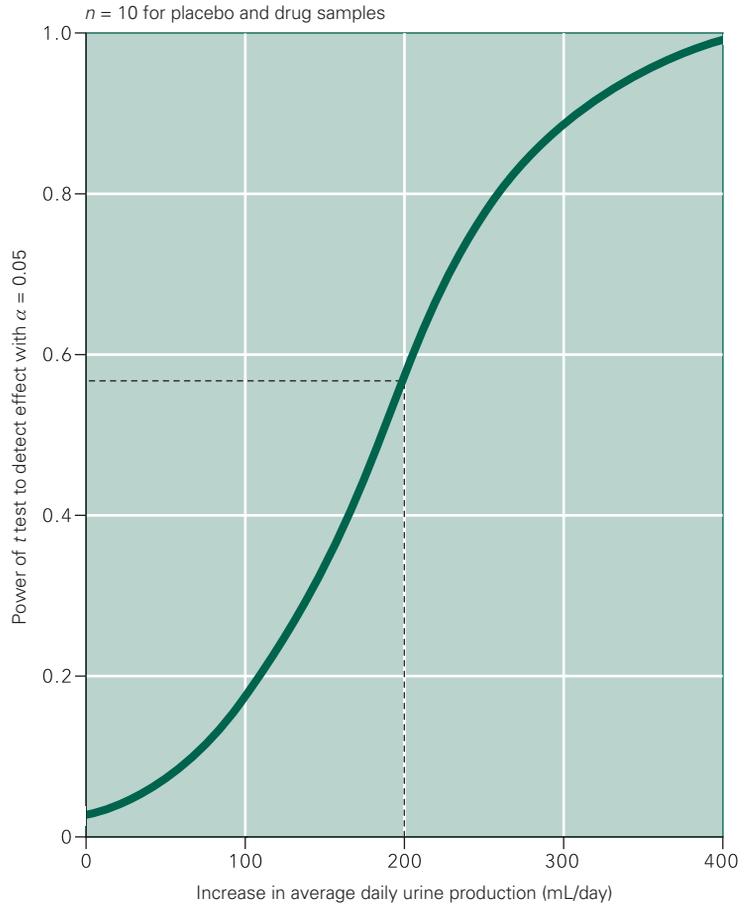


FIGURE 6-6. The power of a t test to detect a change in urine production based on experiments with two groups of people, each containing 10 individuals. The dashed line indicates how to read the graph. A t test has a power of .56 for detecting a 200 mL/day change in urine production.

Recall that the actual distribution of the t test statistic is centered on

$$t' = \frac{\mu_{\text{dr}} - \mu_{\text{pla}}}{\sqrt{(\sigma^2/n_{\text{dr}}) + (\sigma^2/n_{\text{pla}})}}$$

in which $\mu_{\text{dr}} - \mu_{\text{pla}}$ is the actual size of the treatment effect, σ is the standard deviation of the two (different) underlying populations, and n_{dr} and n_{pla} are the sizes of the two samples. In the interest of simplicity, we assume that the two samples are the same size; that is $n_{\text{dr}} = n_{\text{pla}} = n$. Denote the change in the population mean due to the treatment with the Greek letter delta, δ ; then $\mu_{\text{dr}} - \mu_{\text{pla}} = \delta$, and the center of the actual t distribution will be

$$t' = \frac{\delta}{\sqrt{(\sigma^2/n) + (\sigma^2/n)}} = \frac{\delta}{\sigma} \sqrt{\frac{n}{2}}$$

Therefore, t' , how far from 0 the center of the actual distribution of the t test statistic moves, depends on the change in the mean response (δ) normalized by the population standard deviation (σ).

For example, the standard deviation in urine production in the population we are studying is 200 mL/day (from Fig. 6-1). In this context, an increase in urine production of 200 or 400 mL/day can be seen to be 1 or 2 standard deviations, a fairly substantial change. These same absolute changes in urine production would be even more striking if the population standard deviation were only 50 mL/day, in which case a 200 mL/day absolute change would be 4 standard deviations. On the other hand, these changes in urine production would be hard to detect—indeed one wonders if you would want to detect them—if the population standard deviation were 500 mL/day. In this case, 200 mL/day would be only 0.4 standard deviation of the population.

As the variability in the population σ decreases, the power of the test to detect a fixed absolute treatment effect size δ increases and vice versa. In fact, we can combine the influence of these two factors by considering the dimensionless ratio $\phi = \delta/\sigma$, known as the *noncentrality parameter*, rather than each one separately.

Bigger Samples Mean More Powerful Tests

So far we have seen two things: (1) The power of a test to correctly reject the hypothesis that a treatment has no effect decreases as the confidence with which you wish to reject that hypothesis increases; (2) the power increases as the size of the treatment effect, measured with respect to the population standard deviation, increases. In most cases, investigators cannot control either of these factors and for a given sample size are stuck with whatever the power of the test is. However, the situation is not totally beyond their control. They can increase the power of the test without sacrificing the confidence with which they reject the hypothesis of no treatment effect (α) by *increasing the sample size*.

Increasing the sample size generally increases the power for two reasons. First, as the sample size grows the number of degrees of freedom increases, and the value of the test statistic that defines the “biggest” 100α percent of possible values under the assumption of no treatment effect decreases. Second, as the equation for t' above shows, the value of t (and many other test statistics) increases as sample size n increases. As a result, the distribution of t values that occur when the treatment has an effect of a given size δ/σ is located at higher t values as sample size increases.

For example, Figure 6-7A shows the same information as Figure 6-4A, with the sample size equal to 10 in each of the two groups. Figure 6-7B shows the distribution of possible t values if the hypothesis of no effect were true as well as the distribution of t values that would appear if the drug still increased urine production by 200 mL/day but now based on an experiment with 20 people in each group. Even though the size of the treatment effect ($\delta = 200$ mL/day) and the standard deviations of the underlying populations ($\sigma = 200$ mL/day) are the same as before, the actual distribution of the t test statistic moves further to the right to

$$t' = \frac{200}{\sqrt{(200^2/20) + (200^2/20)}} = 3.162$$

because the sample size of each group increased from $n = 10$ to $n = 20$.

In addition, because there are now 20 people in each group, the experiment has $\nu = 2(20 - 1) = 38$ degrees of freedom. From Table 4-1, the critical value of t defining the most extreme (two-tail) 5% of possible t values under the null hypothesis of no effect falls to 2.024. To obtain the power of this test to reject the null hypothesis, we find the proportion of the t distribution at or above $2.024 - 3.162 = -1.138$ with $\nu = 38$ degrees of freedom. From Table 6-2, we find that the power of this study to detect an effect has increased to .86, up substantially from the value of .56 associated with a sample size of 10 in each treatment group.

We could repeat this analysis over and over again to compute the power of this test to detect a 200 mL/day increase in urine production for a variety of sample sizes. Figure 6-8 shows the results of such computations. As the sample size increases, so does the test's power. In fact, estimating the sample size required to detect an effect large enough to be clinically significant is probably the major practical use to which power computations are put. Such computations are especially important in planning randomized clinical trials to estimate how many patients will have to be recruited and how many centers will have to be involved to accumulate enough patients to obtain a large enough sample to complete a meaningful analysis.

What Determines Power? A Summary

Figure 6-9 shows a general power curve for the t test, allowing for a variety of sample sizes and differences of interest. All these curves assume that we will reject the null hypothesis of no treatment effect whenever we compute a value of t from the data that corresponds to $P < .05$ ($\alpha = 0.05$). If we were more or less stringent in our requirement concerning the size of t necessary to report a difference, we would obtain a family of curves different from those in Figure 6-9.

There is one curve for each value of the sample size n in Figure 6-9. This value of n represents the size of *each* of the two sample groups being compared with the t test. Most power charts (and tables) present the results assuming that each of the experimental groups is the same size because, for a given total sample size, power is greatest when there are equal numbers of subjects in each treatment group. Thus, when using power analysis to estimate the sample size for an experiment, the result actually yields the size of each of the sample groups. Power analysis also can be used to estimate the power of a test that yielded a negative finding; in the case of unequal sample sizes, use

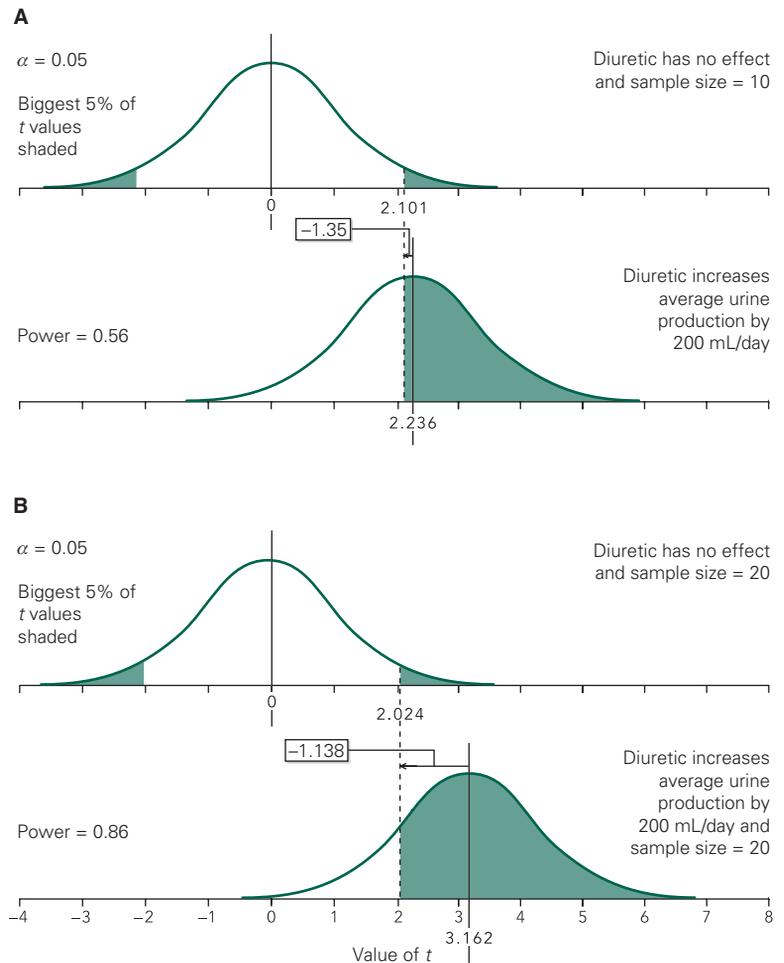


FIGURE 6-7. As the sample size increases from 10 per group (**A**) to 20 per group (**B**), the power of the test increases for two reasons: (1) the critical value of t for a given confidence level in concluding that the treatment had an effect decreases, and (2) the values of the t statistic associated with the experiment increase.

the size of the smaller sample in the power analysis with the charts in this book.* This procedure will give you a conservative (low) estimate for the power of the test.

To illustrate the use of Figure 6-9, again consider the effects of diuretic presented in Figure 6-1. We wish to compute the power of a t test (with a 5% risk of a Type I error, $\alpha = 0.05$) to detect a mean change in urine production of 200 mL/day when the population has a standard deviation of 200 mL/day. Hence

$$\phi = \frac{\delta}{\sigma} = \frac{200 \text{ mL/day}}{200 \text{ mL/day}} = 1$$

*There are computer programs that yield exact power calculations when sample sizes are not equal.

Since the sample size is $n = 10$ (in both the placebo and drug groups), we use the “ $n = 10$ ” line in Figure 6-9 to find that this test will have a power of .56.

All the examples in this chapter so far deal with estimating the power of an experiment that is analyzed with a t test. It is also possible to compute the power for all the other statistical procedures described in this book. Although the details of the computations are different, the same variables are important and play the same general roles in the computation.

Muscle Strength in People with Chronic Obstructive Pulmonary Disease

The stair climb power test is a functional test used among older people to measure leg muscle power.

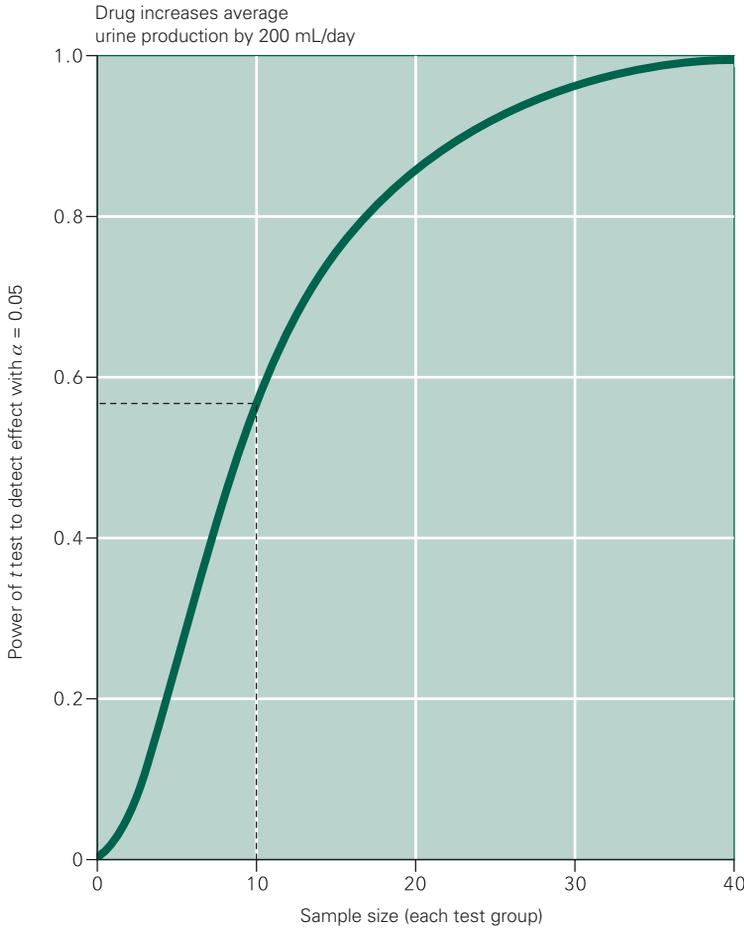


FIGURE 6-8. The effect of sample size on the power of a t test to detect a 200 mL/day increase in urine production with $\alpha = 0.05$ and a population standard deviation in urine production of 200 mL/day. The dashed line illustrates how to read the graph. A sample size of 10 yields a power of .56 for a t test to detect a 200 mL/day change in urine production.

To assess whether this test could be used to assess leg muscle power in people with chronic obstructive pulmonary disease (COPD) Marc Roig and colleagues* measured the power delivered by people with mild-to-severe COPD with age and sex matched controls without disease but lived sedentary lifestyles. Subjects were told to climb 10 stairs as quickly as they could and the power computed as the vertical velocity (the gain in height of the 10 stairs divided by the length of time it took the subject to climb the stairs) times the subject's weight. Based on historical data, Roig and colleagues

expected the normal control people to deliver about 375 W with a standard deviation of about 125 W.

How large a sample size would be necessary to have an 80% power to detect a 100 W change in the power delivered by the people with COPD using conventional statistical significance ($\alpha = .05$)?

The desired effect size, δ , is 100 W and the estimated standard deviation, σ , is 125 W, so the noncentrality parameter is

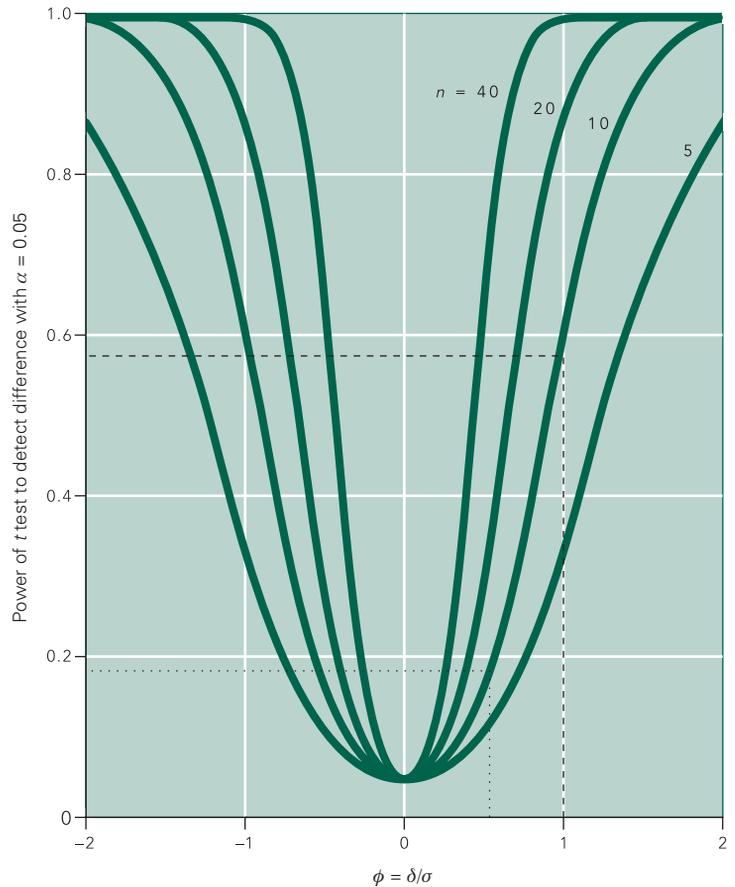
$$\phi = \frac{\delta}{\sigma} = \frac{100}{125} = .80$$

From Figure 6-9, the sample size to obtain a power of 0.8 is $n = 26$ for each sample.

Of course, we could also compute the power of a study with a given sample size to detect a specified effect. Box 6-1 illustrates such a calculation.

*Roig M, Eng JJ, MacIntyre DL, Road JD, Reid WD. Associations of the stair climb power test with muscle strength and functional performance in people with chronic obstructive pulmonary disease: a cross-sectional study. *Phys Ther.* 2010;90:1774–1782.

FIGURE 6-9. The power function for a t test for comparing two experimental groups, each of size n , with $\alpha = 0.05$. δ is the size of the change we wish to detect and σ is the population standard deviation. If we had taken $\alpha = 0.01$ or any other value, we would have obtained a different set of curves. The dashed line indicates how to read the power of a test to detect a $\sigma = 200$ mL/day change in urine production with a $\delta = 200$ mL/day standard deviation in the underlying population with a sample size of $n = 10$ in each test group; the power of this test is .56. The dotted line indicates how to find the power of an experiment designed to study the effects of anesthesia on the cardiovascular system in which $\phi = \delta/\sigma = .55$ with a sample size of 9; the power of this test is only .19.



BOX 6-1 • Power to Detect a Change in Stair Climbing Performance Given the Sample Size

If Roig and colleagues included 20 people in each of the control and COPD groups, what would be the power of the study to detect a 25% change stair climbing power assuming conventional statistical significance ($\alpha = .05$)?

Because normal people have an average stair climbing power of 375 W with a 25% change in pain score would be an effect size, δ , of $.25 \times 375 = 94$ W. The estimated standard deviation, σ , is 125 W, so the noncentrality parameter is

$$\phi = \frac{\delta}{\sigma} = \frac{94}{125} = .75$$

From Figure 6-9, with a sample size of $n = 20$ in each group, the power to detect this effect is .64.

We summarize our discussion of the power of hypothesis-testing procedures with these five statements:

- The power of a test tells the likelihood that the hypothesis of no treatment effect will be rejected when the treatment has an effect.
- The more stringent our requirement for reporting that the treatment produced an effect (i.e., the smaller the chances of erroneously reporting that the treatment was effective), the lower the power of the test.
- The smaller the size of the treatment effect (with respect to the population standard deviation), the harder it is to detect.
- The larger the sample size, the greater the power of the test.
- The exact procedure to compute the power of a test depends on the test itself.

■ POWER AND SAMPLE SIZE FOR ANALYSIS OF VARIANCE*

The issues underlying power and sample size calculations in analysis of variance are no different than for the t test. The only difference is the way in which the size of the minimum detectable treatment effect is quantified and the mathematical relationship relating this magnitude and the risk of erroneously concluding a treatment effect. The measure of the treatment effect to be detected is more complicated than in a t test because it must be expressed as more than a simple difference of two groups (because there are generally more than two groups in an analysis of variance). The size of the treatment effect is again quantified by the *noncentrality parameter*, ϕ , although it is defined differently than for a t test. To estimate the power of an analysis of variance, you specify the number of treatment groups, sample size, risk of a false positive (α) you are willing to accept, and size of the treatment effect you wish to detect (δ), then look the power up in charts for analysis of variance, just as we used Figure 6-9 for t tests.

The first step is to define the size of the treatment effect with the noncentrality parameter. We specify the minimum difference between any two treatment groups we wish to detect, δ , just as when computing the power of the t test. In this case, we define

$$\phi = \frac{\delta}{\sigma} \sqrt{\frac{n}{2k}}$$

where σ is the standard deviation within the underlying population, k is the number of treatment groups, and n is the sample size of each treatment group.[†] (Note the

*In an introductory course, this section can be skipped without interfering with the remaining material in the book.

[†]We present the analysis for equal sample sizes in all treatment groups and the case where all the means but one are equal and the other differs by δ . This arrangement produces the maximum power for a given total sample size. An alternative definition of ϕ involves specifying the means for the different treatment groups that you expect to detect, μ , for each of the k groups. In this case,

$$\phi = \sqrt{\frac{n \sum (\mu_i - \mu)^2}{k \sigma^2}}$$

where

$$\mu = \frac{\sum \mu_i}{k}$$

is the grand population mean. The definition of ϕ in terms of the minimum detectable difference is generally easier to use because it requires fewer assumptions.

similarity with the definition of $\phi = \delta/\sigma$ for the t test.) Once ϕ is determined, obtain the power by looking in a power chart such as Figure 6-10 with the appropriate number of numerator degrees of freedom, $v_n = k - 1$ and denominator degrees of freedom $v_d = k(n - 1)$. (A more complete set of power charts for analysis of variance appears in Appendix B.)

These same charts can be used to estimate the sample size necessary to detect a given effect with a specified power. The situation is a little more complicated than it was in the t test because the sample size, n , appears in the noncentrality parameter, ϕ , and the denominator degrees of freedom, v_d . As a result, you must apply successive guesses to find n . You first guess n , compute the power, then adjust the guess until the computed power is close to the desired value. The example below illustrates this process.

Power and Sperm Motility

Suppose in the experimental study of rabbit sperm motility with three ($k = 3$) experimental conditions – ordinary control, stress control and cell phone exposure – we also wanted to measure the effect of cell phone exposure on sperm count. Normal sperm count in a rabbit is about 350 million sperm/mL with a standard deviation of about 20 million sperm/mL. What would be the power of the study with $n = 8$ rabbits per group we analyzed earlier (Box 3-1) to detect a change of 50 million sperm/mL at conventional statistical significance ($\alpha = .05$)?

Using this information, the noncentrality parameter is

$$\phi = \frac{\delta}{\sigma} \sqrt{\frac{n}{2k}} = \frac{50}{20} \sqrt{\frac{8}{2 \cdot 3}} = 2.88$$

There are $v_n = k - 1 = 3 - 1 = 2$ numerator and $v_d = k(n - 1) = 3(8 - 1) = 21$ denominator degrees of freedom. From the power chart in Figure 6-10, the power to detect a change of 50 million sperm/mL is .99, so we can be very confident of detecting this change.

This is an exceptionally high power. Given the cost of doing the experiments and a desire to minimize the number of animals used in the experiments, suppose that we would be happy with .80 power. We estimate the sample size using the same noncentrality parameter and power chart, but, because the sample size, n , appears both in the noncentrality parameter and the denominator degrees of freedom, v_d , which, in turn, determines which line in Figure 6-10 we use, we need to solve for n iteratively. Box 6-2 shows that this process yields a sample size of 5 per group.

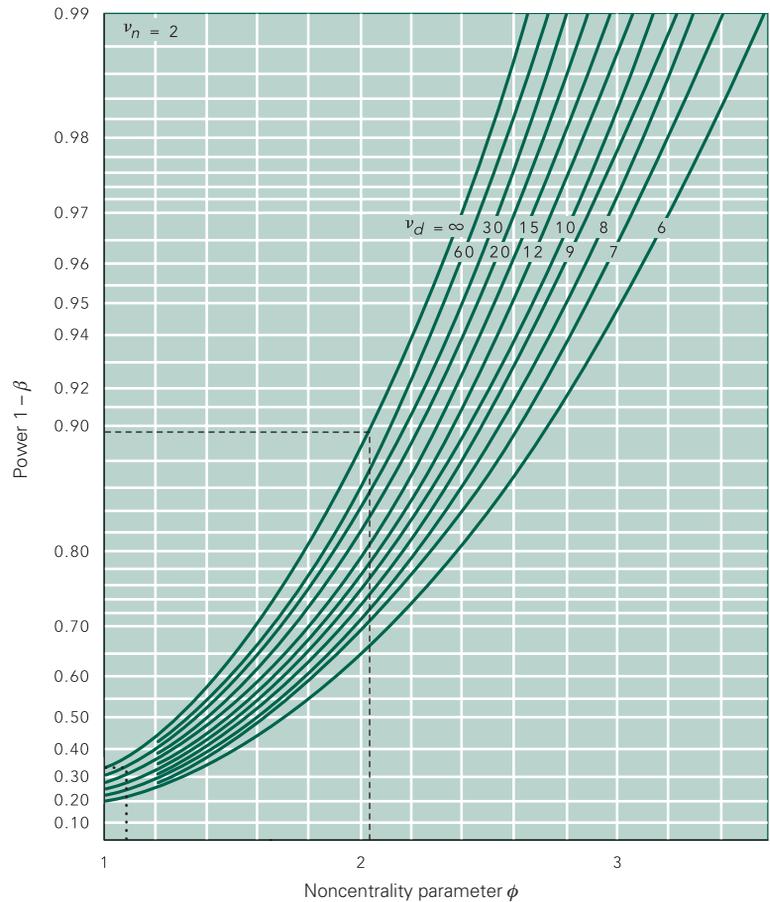


FIGURE 6-10. The power function for analysis of variance for $v_n = 2$ and $\alpha = 0.05$. Appendix B contains a complete set of power charts for a variety of values of v_n and $\alpha = 0.05$ and $.01$. (Source: Adapted from Pearson ES, Hartley HO. Charts for the power function for analysis of variance tests, derived from the non-central f distribution. *Biometrika* 1951;38:112–130.)

BOX 6-2 • Sample Size to Detect a Change of 50 million sperm/mL in Rabbit Study

There are three ($k = 3$) experimental conditions and we want to be able to detect a difference of $\delta = 50$ million sperm/mL with a standard deviation of $\sigma = 20$ million sperm/mL with $\alpha = .05$? We know that $n = 8$ rabbits per group gives more power than we need, so try $n = 4$. In this case the noncentrality parameter would be

$$\phi = \frac{\delta}{\sigma} \sqrt{\frac{n}{2k}} = \frac{50}{20} \sqrt{\frac{4}{2 \cdot 3}} = 2.04$$

There are $v_n = k - 1 = 3 - 1 = 2$ numerator and $v_d = k(n - 1) = 3(4 - 1) = 9$ denominator degrees of freedom. From the power chart in Figure 6-10, the power to detect a change of 50 million sperm/mL is .76, which is a little below our target of .80. Since it is close, try $n = 5$, so

$$\phi = \frac{\delta}{\sigma} \sqrt{\frac{n}{2k}} = \frac{50}{20} \sqrt{\frac{5}{2 \cdot 3}} = 2.28$$

There are still $v_n = k - 1 = 3 - 1 = 2$ numerator degrees of freedom, but now there are $v_d = k(n - 1) = 3(5 - 1) = 12$ denominator degrees of freedom. From Figure 6-10, the power is .89, so we can do this experiment with $n = 5$ rabbits in each group and achieve the desired power.

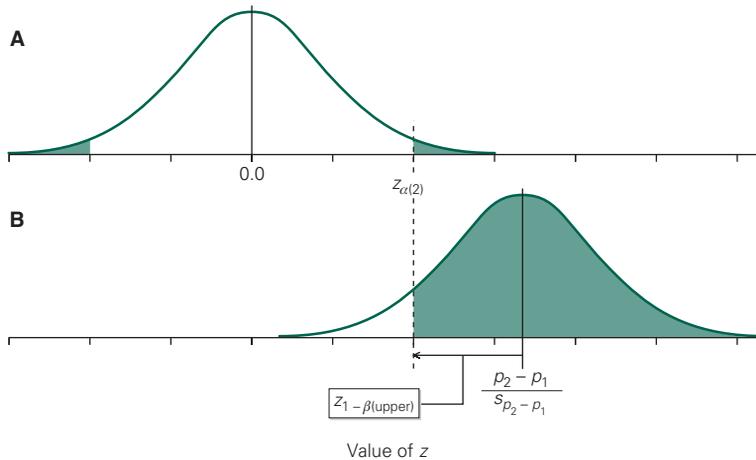


FIGURE 6-11. (A) $z_{\alpha(2)}$ is the two-tail critical value of the z test statistic that defines the α percent most extreme values of the z test statistic that we would expect to observe in an experiment comparing two proportions if the null hypothesis of no differences in the underlying populations was true. (B) If there is a difference in the proportions with the characteristic of interest in the two populations, the distribution of possible values of the z test statistic will no longer be centered on 0, but rather a value that depends on how big the actual differences in proportions between the two populations, $|p_1 - p_2|$, is. The fraction of this actual distribution of the z test statistic that fall above $z_{\alpha(2)}$ approximates the power of the test. (compare with Fig. 6-4.)

■ POWER AND SAMPLE SIZE FOR COMPARING TWO PROPORTIONS*

The development of formulas for power and sample size when comparing two proportions is similar to the procedure that we used for the t test, except that we will be basing the computations on the normal distribution. We wish to find the power of a z test to detect a difference between two proportions, p_1 and p_2 with sample sizes n_1 and n_2 . Recall, from Chapter 5, that the z test statistic used to compare two observed proportions, is

$$z = \frac{\hat{p}_2 - \hat{p}_1}{s_{p_2 - p_1}}$$

Under the null hypothesis of no difference, this test statistic follows the standard normal distribution (with mean 0 and standard deviation 1) given in the

*If time is limited this material can be skipped without loss of continuity.

last row of Table 6-2. We denote the two-tailed critical value of z that we require to reject the null hypothesis of no difference with Type I error α , $z_{\alpha(2)}$. For example, if we follow the convention of accepting a 5% risk of a false positive (i.e., reject the null hypothesis of no difference when $P < .05$), from Table 4-1, $z_{\alpha(2)} = 1.960$ (Fig. 6-11A).

If there is actually a difference in the two proportions, p_1 and p_2 , then the actual distribution of the z test statistic will be centered on

$$z' = \frac{|p_2 - p_1|}{s_{p_2 - p_1}}$$

where

$$s_{p_2 - p_1} = \sqrt{\frac{p_2(1-p_2)}{n_2} + \frac{p_1(1-p_1)}{n_1}}$$

As we did with the t test, we determine the power to detect the difference $p_2 - p_1$ as the proportion of the actual distribution of the z test statistic (Fig. 6-11B) that falls above $z_{\alpha(2)}$. Hence, the power of the test to detect the

specified difference could be estimated as the proportion of the normal distribution above

$$z_{1-\beta(\text{upper})} = z_{\alpha(2)} - z' = z_{\alpha(2)} - \frac{|p_2 - p_1|}{s_{p_2-p_1}}$$

where $z_{1-\beta(\text{upper})}$ is the value of z that defines the upper $(1 - \beta)$ percentage of the normal distribution (from Table 6-2).*

The estimate of power obtained by matching the two distributions in Figure 6-11 based on z values can be improved by adjusting the matching criterion because in real units that standard deviations of the normal distribution under the null hypothesis (analogous to the top panel in Fig. 6-11) and the alternative hypothesis (the bottom panel) are slightly different. We obtain a more accurate estimate of the power by adjusting for this fact, which yields

$$\begin{aligned} z_{1-\beta(\text{upper})} &= \frac{s_{\bar{p}}}{s_{p_2-p_1}} z_{\alpha(2)} - \frac{|p_2 - p_1|}{s_{p_2-p_1}} \\ &= 1.960 - \frac{|.97 - .87|}{.0451} = -.257 \end{aligned}$$

where \bar{p} is the weighted average of the two anticipated probabilities

$$\bar{p} = \frac{n_2 p_2 + n_1 p_1}{n_2 + n_1}$$

and $s_{\bar{p}}$ is the associated standard deviation

$$s_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n_2} + \frac{\bar{p}(1-\bar{p})}{n_1}}$$

Power and Polyethylene Bags

When we evaluated the effect on mortality of keeping extremely low birth weight infants warm by wrapping them in polyethylene bags compared to traditional methods

(Table 5-3) we did not reject the null hypothesis of no effect. To get a sense of how confident we can be in drawing a negative conclusion from these data (and *accept* the null hypothesis of no effect), we will compute the power of this study to detect a 10% difference in survival.

In Chapter 5, we analyzed these data using χ^2 , but because this is a 2×2 contingency table, we could also do the same analysis as a comparison of two proportions. Sixty-one of 70 infants who were warmed using traditional methods, $61/70 = 87\%$ survived. Therefore, we will set the initial proportion, p_1 , to .87 and final proportion, p_2 , to .87 + .10 = .97. The sample size, n , of both groups is 70. We will use conventional statistical significance of $\alpha = .05$.

We begin by using the target proportions to compute

$$\begin{aligned} \bar{p} &= \frac{n_2 p_2 + n_1 p_1}{n_2 + n_1} = \frac{70 \cdot .97 + 70 \cdot .87}{70 + 70} = .92 \\ s_{\bar{p}} &= \sqrt{\frac{\bar{p}(1-\bar{p})}{n_2} + \frac{\bar{p}(1-\bar{p})}{n_1}} \\ &= \sqrt{\frac{.92(1-.92)}{70} + \frac{.92(1-.92)}{70}} = .0459 \end{aligned}$$

and

$$\begin{aligned} s_{p_2-p_1} &= \sqrt{\frac{p_2(1-p_2)}{n_2} + \frac{p_1(1-p_1)}{n_1}} \\ &= \sqrt{\frac{.97(1-.97)}{70} + \frac{.87(1-.87)}{70}} = .0451 \end{aligned}$$

The two-tail 95% critical value of the normal distribution, $z_{\alpha(2)}$, is, from Table 6-2, 1.960 so the power of the test is the fraction of the normal distribution above

$$\begin{aligned} z_{1-\beta(\text{upper})} &= \frac{s_{\bar{p}}}{s_{p_2-p_1}} z_{\alpha(2)} - \frac{p_2 - p_1}{s_{p_2-p_1}} \\ &= \frac{.0459}{.0451} \cdot 1.960 - \frac{.97 - .87}{.0451} = -.223 \end{aligned}$$

From Table 6-2, the power of the test is 59%, which means that we can be reasonably confident in accepting the null hypothesis and concluding that there was no survival benefit from using the polyethylene bags to keep the infants warm.

Box 6-3 shows that this study had a 16% power to detect a 5% survival benefit, so if you did not think changing the technology for keeping extremely low birth weight

*Technically, we should also include the part of the distribution in Figure 6-11A that falls below the lower $z_{\alpha(2)}$ tail of the distribution in Figure 6-11B, but this tail of the distribution rarely contributes anything of consequence. Note that these calculations do not include the Yates correction. It is possible to include the Yates correction by replacing $(p_2 - p_1)$ with $(p_2 - p_1 - 1/2) (1/n_2 + 1/n_1)$. Doing so makes the arithmetic more difficult, but does not represent a theoretical change. Including the Yates correction lowers the power or increases the sample size.

BOX 6-3 • Power to Detect a 5% Survival Benefit of Using Polyethylene Bags to Keep Extremely Low Birth Weight Infants Warm

We use the proportion of infants who survived using traditional methods as the initial proportion so set $p_1 = .87$ and the final proportion 5% higher, so $p_2 = .87 + .05 = .92$. The sample size, n , of both groups is 70.

We begin by using the target proportions to compute

$$\bar{p} = \frac{n_2 p_2 + n_1 p_1}{n_2 + n_1} = \frac{70 \cdot .92 + 70 \cdot .87}{70 + 70} = .895$$

$$s_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n_2} + \frac{\bar{p}(1-\bar{p})}{n_1}} = \sqrt{\frac{.895(1-.895)}{70} + \frac{.895(1-.895)}{70}} = .0518$$

and

$$s_{p_2-p_1} = \sqrt{\frac{p_2(1-p_2)}{n_2} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{.92(1-.92)}{70} + \frac{.87(1-.87)}{70}} = .0516$$

We will use conventional statistical significance of $\alpha = .05$, so we use the two-tail 95% critical value of the normal distribution, $z_{\alpha(2)}$, is, from Table 6-2, 1.960. The power of the test is the fraction of the normal distribution above

$$z_{1-\beta(\text{upper})} = \frac{s_{\bar{p}}}{s_{p_2-p_1}} z_{\alpha(2)} - \frac{p_2 - p_1}{s_{p_2-p_1}} = \frac{.0518}{.0516} \cdot 1.960 - \frac{.92 - .87}{.0516} = .999$$

From Table 6-2, the power of the test is 13%.

infants warm would be worth doing if it had a mortality benefit of 5%, you could not be very confident in reaching a negative conclusion.

Sample Size for Comparing Two Proportions

To obtain the sample size to compare two proportions, simply take $z_{1-\beta(\text{upper})}$ as given and solve the resulting equations for n , the size of each group. Assuming that the two groups are the same size, this process yields

$$n = \frac{A \left[1 + \sqrt{1 + \frac{4\delta}{A}} \right]^2}{4\delta^2}$$

where

$$\bar{p} = \frac{p_2 + p_1}{2}$$

$$\delta = |p_2 - p_1|$$

$$A = \left[z_{\alpha(2)} \sqrt{2\bar{p}(1-\bar{p})} + z_{1-\beta(\text{upper})} \sqrt{p_2(1-p_2) + p_2(1-p_2)} \right]^2$$

We can use these formulas to estimate the sample size necessary to detect a 5% survival improvement in the study of using polyethylene bags to keep extreme low birth weight infants warm with 80% power. $p_1 = .87$ and $p_2 = .92$, so

$$\bar{p} = \frac{.87 + .92}{2} = .895$$

The desired effect size, $\delta = .05$, and to obtain .80 power, $z_{1-\beta(\text{upper})} = -.842$ from Table 6-2, and

$$A = \left[1.960 \sqrt{2 \cdot .895(1-.895)} + (-.842) \sqrt{.87(1-.87) + .92(1-.92)} \right]^2 = .2361$$

Therefore,

$$n = \frac{.2361 \left[1 + \sqrt{1 + \frac{4 \cdot .05}{.2361}} \right]^2}{4(.05^2)} = 131.4$$

Thus, to obtain 80% power to detect a 5% improvement in survival, we would need 132 infants per experimental group.

■ POWER AND SAMPLE SIZE FOR RELATIVE RISK AND ODDS RATIO*

The formulas developed above can be used to estimate power and sample sizes for relative risks and odds ratios. Instead of specifying both proportions, you simply specify one proportion, the desired relative risk or odds ratio, and compute the other proportion. Let p_1 be the probability of disease in the unexposed members of the population and p_2 be the probability of disease in the exposed members of the population.

The relative risk is the ratio of the probability of disease in those exposed to the toxin of interest over those not exposed,

$$RR = \frac{P_{\text{exposed}}}{P_{\text{unexposed}}} = \frac{p_2}{p_1}$$

so use the formulas above with

$$p_2 = RR \cdot p_1$$

Likewise, the odds ratio is

$$OR = \frac{P_{\text{exposed}}/(1 - P_{\text{exposed}})}{P_{\text{unexposed}}/(1 - P_{\text{unexposed}})} = \frac{p_2/(1 - p_2)}{p_1/(1 - p_1)}$$

so

$$p_2 = \frac{OR \cdot p_1}{1 + p_1(OR - 1)}$$

■ POWER AND SAMPLE SIZE FOR CONTINGENCY TABLES†

Figure 6-10 (and the corresponding charts in Appendix B) can also be used to compute the power and sample size for contingency tables. As with other power computations, the first step is to define the pattern you wish to be able to detect. This effect is specified by selecting the proportions of row and column observations that appear in each cell of the contingency table.

*If time is limited this section can be skipped without loss of continuity.

†If time is limited this section can be skipped without loss of continuity.

■ TABLE 6-3. Notation for Computing Power for Contingency Tables

p_{11}	p_{12}	R_1
p_{21}	p_{22}	R_2
$\frac{p_{31}}{C_1}$	$\frac{p_{32}}{C_2}$	$\frac{R_3}{1.00}$

Table 6-3 shows the notation for the computation for a 3×2 contingency table: p_{11} is the proportion of all observations expected in the upper left cell of the table, p_{12} the proportion in the upper right corner, and so on. All the proportions must add up to 1. The r row and c column sums are denoted with R s and C s with subscripts corresponding to the rows and columns. The noncentrality parameter for such a contingency table is defined as

$$\phi = \sqrt{\frac{N}{(r-1)(c-1)+1} \sum \frac{(p_{ij} - R_i C_j)^2}{R_i C_j}}$$

where r is the number of rows, c is the number of columns, and N is the total number of observations. This value of ϕ is used with Figure 6-10 with $v_n = (r-1)(c-1)$ and $v_d = \infty$ degrees of freedom.

To compute the sample size necessary to achieve a given power, simply reverse this process. Determine the necessary value of ϕ to achieve the desired power with $v_n = (r-1)(c-1)$ and $v_d = \infty$ from Figure 6-10 (or the power charts in Appendix B). We obtain the sample size by solving the equation above for N , to obtain

$$N = \frac{\phi^2 [(r-1)(c-1)+1]}{\sum \frac{(p_{ij} - R_i C_j)^2}{R_i C_j}}$$

Power and Polyethylene Bags (Again)

We can now compute the power of the study of different ways to warm extremely low birth weight infants using a contingency table approach. We will again compute the power of the study to detect a 10% improvement in survival, from $p_1 = .87$ to $p_2 = .97$. The total sample size is $N = 140$ (70 in each group). Because the infants are distributed equally in the two treatment groups, the fraction in each group (last column in Table 6-4) is .500. The pattern we

■ TABLE 6-4. Expected Mortality Pattern

Warming Treatment	Fraction of All Infants		
	Lived	Died	Total in Treatment Group
Polyethylene bag	.485	.015	.500
Traditional	.435	.065	.500
Total	.920	.080	1.000

seek to find is one in which 97% of the infants treated with the polyethylene bag lived, $.97 \times .50 = .485$ of all infants and $.03 \times .50 = .015$ of infants would be expected to die. Likewise, for the traditional treatment, 87% of infants would live, amounting to $.50 \times .87 = .435$ of all infants with $.50 \times .13 = .065$ dying. Figure 6-4 shows the pattern we want to detect in the 2×2 contingency table.

There are $r = 2$ rows and $c = 2$ columns, so the value of the noncentrality parameter is

$$\phi = \frac{140}{(2-1)(2-1)+1} \sqrt{\frac{(.485 - .500 \cdot .920)^2}{.500 \cdot .920} + \frac{(.015 - .500 \cdot .080)^2}{.500 \cdot .080} + \frac{(.435 - .500 \cdot .920)^2}{.500 \cdot .920} + \frac{(.065 - .500 \cdot .080)^2}{.500 \cdot .080}} = 2.18$$

Use this value of $\phi = 1.54$ with $v_n = (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$ numerator degree of freedom and $v_d = \infty$ denominator degrees of freedom in the power chart in Appendix B to find that the power is about 60%, as before.

■ PRACTICAL PROBLEMS IN USING POWER

If you know the size of the treatment effect, population standard deviation, α , and sample size, you can use graphs like Figure 6-9 to estimate the power of a t test after the fact. Unfortunately, in practice, one does not know how large an effect a given treatment will have (finding that out is usually the reason for the study in the first place), so you must specify how large a change is *worth detecting* to compute the power of the test.

This requirement to go on record about how small a change is worth detecting may be one reason that very few

people report the power of the tests they use. While such information is not especially important when investigators report that they detected a difference, it can be quite important when they report that they failed to detect one. If the power of the test to detect a clinically significant effect is small, say 25%, this report will mean something quite different than if the test was powerful enough to detect a clinically significant difference 85% of the time.

These difficulties are even more acute when using power computations to decide on the sample size for a study in advance. Completing this computation requires that investigators estimate not only the size of the effect they think is worth detecting and the confidence with which they hope to accept (β) or reject (α) the hypothesis that the treatment is effective but also the standard deviation of the population being studied. Sometimes existing information can be used to estimate these numbers; sometimes investigators do a pilot study to estimate them; sometimes they simply guess.

■ WHAT DIFFERENCE DOES IT MAKE?

In Chapter 4 we discussed the most common error in the use of statistical methods in the medical literature, inappropriate use of the t test. Repeated use of t tests increases the chances of reporting a “statistically significant” difference above the nominal levels one obtains from the t distribution. In the language of this chapter, it increases the Type I error. In practical terms, this increases the chances that an investigator will report some procedure or therapy capable of producing an effect beyond what one would expect from chance variation when the evidence does not actually support this conclusion.

This chapter examined the other side of the coin, the fact that perfectly correctly designed studies employing statistical methods correctly may fail to detect real, perhaps clinically important, differences simply because the

sample sizes are too small to give the procedure enough power to detect the effect. This chapter shows how you can estimate the power of a given test after the results are reported in the literature and also how investigators can estimate the number of subjects they need to study to detect a specified difference with a given level of confidence (say, 95%; that is, $\alpha = 0.05$). Such computations are often quite distressing because they often reveal the need for a large number of experimental subjects, especially compared with the relatively few patients who typically form the basis for clinical studies.* Sometimes the investigators increase the size of the difference they say they wish to detect, decrease the power they find acceptable, or ignore the whole problem in an effort to reduce the necessary sample size. Most medical investigators never confront these problems because they have never heard of power.

In 1978, Jennie Freiman and colleagues[†] examined 71 randomized clinical trials published between 1960 and 1977 in journals, such as *The Lancet*, the *New England Journal of Medicine*, and the *Journal of the American Medical Association*, reporting that the treatment studied did not produce a “statistically significant” ($P < .05$) improvement in clinical outcome. Only 20% of these studies included enough subjects to detect a 25% improvement in clinical outcome with a power of .50 or better. In other words, if the treatment produced a 25% reduction in mortality rate or other clinically important endpoint, there was less than a 50:50 chance that the clinical trial would be able to detect it with $P < .05$. Moreover, Freiman and colleagues found that *only one* of the 71 papers stated that α and β were considered at the start of the study; 18 recognized a trend in the results, whereas 14 commented on the need for a larger sample size.

Fifteen years later, in 1994, D Mohler and colleagues[‡] revisited this question by examining randomized controlled trials in these same journals published in 1975,

1980, 1985, and 1990. While the number of randomized controlled trials published in 1990 was more than twice the number published in 1975, the proportion reporting negative results remained reasonably constant, at about 27% of all the trials. Only 16% and 36% of the negative studies had an adequate power (.80) to detect a 25% or 50% change in outcome, respectively. Only one third of the studies with negative results reported information regarding how the sample sizes were computed. An evaluation of randomized controlled trials published in the surgical literature between 1988 and 1998 found that only 25% of the trials were large enough to detect a 50% difference in therapeutic effect with .80 power, and only 29% of the papers included a formal sample size calculation.[§]

Nine years later, in 2003, Melinda Maggard and colleagues examined papers published between 1999 and 2002 showed that half the studies were powered to detect a 50% difference in therapeutic effect.[¶]

Things are improving, but slowly.

The fact remains, however, that publication of “negative” studies without adequate attention to having a large enough sample size to draw definitive conclusions remains a problem. Thus, in this area, like the rest of statistical applications in the medical literature, it is up to responsible readers to interpret what they read rather than take it at face value.

Other than throwing your hands up when a study with low power fails to detect a statistically significant effect, is there anything an investigator or clinician reading the literature can learn from the results? Yes. Instead of focusing on the accept–reject logic of statistical hypothesis testing,^{||}

*Fletcher RA, Fletcher SW. Clinical research in general medical journals: a 30-year perspective. *N Engl J Med.* 1979;301:180–183 report the median number of subjects included in clinical studies published in the *Journal of the American Medical Association*, *The Lancet*, and the *New England Journal of Medicine* in 1946 to 1976 ranged from 16 to 36 people.

[†]Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized controlled trial. *N Engl J Med.* 1978;299:690–694.

[‡]Mohler D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized clinical trials. *JAMA.* 1994;272:122–124.

[§]Dimick JB, Diener-West M, Lipsett PA. Negative results of randomized clinical trials published in the surgical literature. *Arch Surg.* 2001;136:796–800.

[¶]Maggard MA, O’Connell JB, Liu JH, Etzioni DA, Ko CY. Sample size calculations in surgery: are they done correctly? *Surgery.* 2003;134:275–279.

^{||}There is another approach that can be used in some clinical trials to avoid this accept–reject problem. In a *sequential trial* the data are analyzed after each new individual is added to the study and the decision made to (1) accept the hypothesis of no treatment effect, (2) reject the hypothesis, or (3) study another individual. Sequential tests generally allow one to achieve the same levels of α and β for a given size treatment effect with a smaller sample size than the methods discussed in this book. This smaller sample size is purchased at the cost of increased complexity of the statistical procedures. Sequential analyses are often performed by repeated use of the statistical procedures presented in this book, such as the *t* test. This procedure is incorrect because it produces overoptimistic *P* values, just as the repeated use of *t* tests (without the Bonferroni or Holm-Sidak correction) produces erroneous results when one should do an analysis of variance.

one can try to estimate how strongly the observations suggest an effect by estimating the size of the hypothesized effect together with the uncertainty of this estimate.* We laid the groundwork for this procedure in Chapters 2, 4, and 5 when we discussed the standard error and the t distribution. The next chapter builds on this base to develop the idea of confidence limits.

■ PROBLEMS

6-1 Both diabetes and high cholesterol interact to increase the risk of heart disease. Changing diet affects both blood sugar and cholesterol. To investigate how different diets used to control diabetes affects cardiovascular risk factors, Neal Barnard and colleagues[†] compared the effects of a low-fat vegan diet with the diet recommended by the American Diabetic Association. What is the power of this study to detect a change in mean total cholesterol from 190 to 165 mg/dL with a sample size of 20 people on each diet with 95% confidence? Based on earlier experience, the standard deviation of total cholesterol in the population is about 35 mg/dL.

6-2 How large a sample size would be necessary to increase the power of this study to detect a 25 mg/dL change in total cholesterol?

*One quick way to use a computerized statistical package to estimate if getting more cases would resolve a power problem is to simply copy the data twice and rerun the analysis on the doubled data set. If the results become less ambiguous, it suggests that obtaining more cases (on the assumption that the data will be similar to that which you have already obtained) will yield less ambiguous results. This procedure is, of course, not a substitute for a formal power analysis and would certainly not be reportable in a scientific paper, but it is an easy way to get an idea of whether gathering more data would be worthwhile.

[†]Barnard N, et al. A low-fat vegan diet improves glycemic control and cardiovascular risk factors in a randomized clinical trial in individuals with type 2 diabetes. *Diabetes Care* 2006;29: 1777–1783.

6-3 What is the minimum detectable effect one could obtain with 20 people in each group and 80% power?

6-4 In Problem 3-5 (and again in Prob. 4-5), we decided that there was insufficient evidence to conclude that men and women who have had at least one vertebral fracture differ in vertebral bone density. What is the power of this test to detect average (with $\alpha=0.05$) bone density in men 20% lower than the average bone density for women?

6-5 How large a sample would be necessary to be 90% confident that men have vertebral bone densities that differ by at least 30% of the values for women when you wish to be 95% confident in any conclusion that vertebral bone densities differ between men and women?

6-6 Use the data in Problem 3-2 to find the power of detecting a change in mean forced midexpiratory flow of 0.25 L/s with 95% confidence.

6-7 Use the data in Problem 3-3 to find the power of detecting an increase in a change in stair climbing power of 50 and 100 W with 95% confidence.

6-8 How large must each sample group be to have an 80% power to detect a change of 80 W with 95% confidence?

6-9 What is the power of the experiment in Problem 5-4 to detect a situation in which nefazodone and psychotherapy each causes remission one-third of the time, and nefazodone and psychotherapy combined cause remission one-half of the time? Assume that the same number of people take each treatment as in Problem 5-4. Use $\alpha=0.05$.

6-10 How large would the sample size need to be in Problem 6-9 to reach 80% power?

Confidence Intervals

All the statistical procedures developed so far were designed to help decide whether or not a set of observations is compatible with some hypothesis. These procedures yielded P values to estimate the chance of reporting that a treatment has an effect when it really does not and the power to estimate the chance that the test would detect a treatment effect of some specified size. This decision-making paradigm does not characterize the size of the difference or illuminate results that may not be statistically significant (i.e., not associated with a value of P below .05) but does nevertheless suggest an effect. In addition, since P depends not only on the magnitude of the treatment effect but also the sample size, it is not unusual for experiments with large sample sizes to yield very small values of P (what investigators often call “highly significant” results) when the magnitude of the treatment effect is so small that it is clinically or scientifically unimportant. As Chapter 6 noted, it can be more informative to think not only in terms of the accept–reject approach of statistical hypothesis testing but also to estimate the size of the treatment effect together with some measure of the uncertainty in that estimate.

This approach is not new; we used it in Chapter 2 when we defined the standard error of the mean to quantify the certainty with which we could estimate the population mean from a sample. We observed that since the population of all sample means at least approximately follows a normal distribution, the true (and unobserved) population mean will lie within about 2 standard errors of the mean of the sample mean 95% of the time. We now

develop the tools to make this statement more precise and generalize it to apply to other estimation problems, such as the size of the effect a treatment produces. The resulting estimates, called *confidence intervals*, can also be used to test hypotheses.* This approach yields exactly the same conclusions as the procedures we discussed earlier because it simply represents a different perspective on how to use concepts like the standard error, t , and normal distributions. Confidence intervals are also used to estimate the range of values that include a specified proportion of all members of a population, such as the “normal range” of values for a laboratory test.

■ THE SIZE OF THE TREATMENT EFFECT MEASURED AS THE DIFFERENCE OF TWO MEANS

In Chapter 4, we defined the t statistic to be

$$t = \frac{\text{Difference of sample means}}{\text{Standard error of difference of sample means}}$$

then computed its value for the data observed in an experiment. Next, we compared the result with the value t_α that defined the most extreme 100 α percent of the possible values to t that would occur (in both tails) if the two

*Some statisticians believe that confidence intervals provide a better way to think about the results of experiments than traditional hypothesis testing.

samples were drawn from a single population. If the observed value of t exceeded t_α (given in Table 4-1), we reported a “statistically significant” difference, with $P < \alpha$. As Figure 4-4 showed, the distribution of possible values of t has a mean of zero and is symmetric about zero.

On the other hand, if the two samples are drawn from populations with *different* means, the distribution of values of t associated with all possible experiments involving two samples of a given size is *not* centered on zero; it does not follow the t distribution. As Figures 6-3 and 6-5 showed, the actual distribution of possible values of t has a nonzero mean that depends on the size of the treatment effect. It is possible to revise the definition of t so that it will be distributed according to the t distribution in Figure 4-4 *regardless of whether or not the treatment actually has an effect*. This modified definition of t is

$$t = \frac{\text{Difference of sample means} - \text{true difference in population means}}{\text{Standard error of difference of sample means}}$$

Notice that if the hypothesis of no treatment effect is correct, the difference in population means is zero and this definition of t reduces to the one we used before. The equivalent mathematical statement is

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}}$$

In Chapter 4 we computed t from the observations, then compared it with the critical value for a “big” value of t with $\nu = n_1 + n_2 - 2$ degrees of freedom to obtain a P value. Now, however, we cannot follow this approach since we do not know all the terms on the right side of the equation. Specifically, *we do not know the true difference in mean values of the two populations* from which the samples were drawn, $\mu_1 - \mu_2$. We can, however, use this equation to estimate the size of the treatment effect, $\mu_1 - \mu_2$.

Instead of using the equation to determine t , we will select an appropriate value of t and use the equation to estimate $\mu_1 - \mu_2$. The only problem is that of selecting an appropriate value for t .

By definition, 100α percent of all possible values of t are more negative than $-t_\alpha$ or more positive than $+t_\alpha$. For example, only 5% of all possible t values will fall outside the interval between $-t_{.05}$ and $+t_{.05}$, where $t_{.05}$ is the critical value of t that defines the most extreme 5% of the t distribution (tabulated in Table 4-1). Therefore, $100(1 - \alpha)$

percent of all possible values of t fall between $-t_\alpha$ and $+t_\alpha$. For example, 95% of all possible values of t will fall between $-t_{.05}$ and $+t_{.05}$.

Every different pair of random samples we draw in our experiment will be associated with different values of, $\bar{X}_1 - \bar{X}_2$ and $s_{\bar{X}_1 - \bar{X}_2}$, and $100(1 - \alpha)$ percent of all possible experiments involving samples of a given size will yield values of t that fall between $-t_\alpha$ and $+t_\alpha$. Therefore, for $100(1 - \alpha)$ percent of all possible experiments

$$-t_\alpha < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}} < +t_\alpha$$

Solve this equation for the true difference in sample means

$$(\bar{X}_1 - \bar{X}_2) - t_\alpha s_{\bar{X}_1 - \bar{X}_2} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_\alpha s_{\bar{X}_1 - \bar{X}_2}$$

In other words, the actual difference of the means of the two populations from which the samples were drawn will fall within t_α standard errors of the difference of the sample means of the observed difference in the sample means. (t_α has $\nu = n_1 + n_2 - 2$ degrees of freedom, just as when we used the t distribution in hypothesis testing). This range is called the $100(1 - \alpha)$ percent *confidence interval for the difference of the means*. For example, the 95% confidence interval for the true difference of the population means is

$$(\bar{X}_1 - \bar{X}_2) - t_{.05} s_{\bar{X}_1 - \bar{X}_2} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_{.05} s_{\bar{X}_1 - \bar{X}_2}$$

This equation defines the range that will include the true difference in the means for 95% of all possible experiments that involve drawing samples from the two populations under study.

Since this procedure to compute the confidence interval for the difference of two means uses the t distribution, it is subject to the same limitations as the t test. In particular, the samples must be drawn from populations that follow a normal distribution at least approximately.*

*It is also possible to define confidence intervals for differences in means when there are multiple comparisons, by using a Bonferroni or Holm-Sidak correction to determine the appropriate value of t . For a detailed discussion of these computations, see Zar JH. *Biostatistical Analysis*, 4th ed. Upper Saddle River, NJ: Prentice Hall; 1999.

THE EFFECTIVE DIURETIC

Figure 6-1 showed the distributions of daily urine production for a population of 200 individuals when they are taking a placebo or a drug that is an effective diuretic. The mean urine production of the entire population when all members are taking the placebo is $\mu_{\text{pla}} = 1200$ mL/day. The mean urine production for the population when all members are taking the drug is $\mu_{\text{dr}} = 1400$ mL/day. Therefore, the drug increases urine production by an average of $\mu_{\text{dr}} - \mu_{\text{pla}} = 1400 - 1200 = 200$ mL/day. An investigator, however, cannot observe every member of the population and must estimate the size of this effect from samples of people observed when they are taking the placebo or the drug. Figure 6-1 shows one pair of such samples, each of 10 individuals. The people who received the placebo had a mean urine output of 1180 mL/day, and the people receiving the drug had a mean urine output of 1400 mL/day. Thus, these two samples suggest that the drug increased urine production by $\bar{X}_{\text{dr}} - \bar{X}_{\text{pla}} = 1400 - 1180 = 220$ mL/day. The random variation associated with the sampling procedure led to a different estimate of the size of the treatment effect from that really present. Simply presenting this single estimate of 220 mL/day increase in urine output ignores the fact that there is some uncertainty in the estimates of the true mean urine output in the two populations, so there will be some uncertainty in the estimate of the true difference in urine output. We now use the confidence interval to present an alternative description of how large a change in urine output accompanies the drug. This interval describes the average change seen in the people included in the experiment and also reflects the uncertainty introduced by the random sampling process.

To estimate the standard error of the difference of the means $s_{\bar{X}_{\text{dr}} - \bar{X}_{\text{pla}}}$ we first compute a pooled estimate of the population variance. The standard deviations of observed urine production were 245 and 144 mL/day for people taking the drug and the placebo, respectively. Both samples included 10 people; therefore,

$$s^2 = \frac{1}{2}(s_{\text{dr}}^2 + s_{\text{pla}}^2) = \frac{1}{2}(245^2 + 144^2) = 201^2$$

and

$$s_{\bar{X}_{\text{dr}} - \bar{X}_{\text{pla}}} = \sqrt{\frac{s^2}{n_{\text{dr}}} + \frac{s^2}{n_{\text{pla}}}} = \sqrt{\frac{201^2}{10} + \frac{201^2}{10}} = 89.9 \text{ mL/day}$$

To compute the 95% confidence interval, we need the value of $t_{.05}$ from Table 4-1. Since each sample contains $n = 10$ individuals, we use the value of $t_{.05}$ corresponding to $\nu = 10 + 10 - 2 = 18$ degrees of freedom. From Table 4-1, $t_{.05} = 2.101$.

Now we are ready to compute the 95% confidence interval for the mean change in urine production that accompanies use of the drug

$$(\bar{X}_{\text{dr}} - \bar{X}_{\text{pla}}) - t_{.05} s_{\bar{X}_{\text{dr}} - \bar{X}_{\text{pla}}} < \mu_{\text{dr}} - \mu_{\text{pla}} < (\bar{X}_{\text{dr}} - \bar{X}_{\text{pla}}) + t_{.05} s_{\bar{X}_{\text{dr}} - \bar{X}_{\text{pla}}}$$

$$220 - 2.101 \cdot 89.9 < \mu_{\text{dr}} - \mu_{\text{pla}} < 220 + 2.101 \cdot 89.9$$

$$31 \text{ mL/day} < \mu_{\text{dr}} - \mu_{\text{pla}} < 409 \text{ mL/day}$$

Thus, on the basis of this particular experiment, we can be 95% confident that the drug increases average urine production somewhere between 31 and 409 mL/day. The range of values from 31 to 409 is the 95% confidence interval corresponding to this experiment. As Figure 7-1A shows, this interval includes the actual change in mean urine production, $\mu_{\text{dr}} - \mu_{\text{pla}}$, 200 mL/day.

More Experiments

Of course, there is nothing special about the two samples of 10 people each selected in the study we just analyzed. Just as the values of the sample mean and standard deviation vary with the specific random sample of people we happen to draw, so will the confidence interval we compute from the resulting observations. (This should not be surprising, since the confidence interval is computed from the sample means and standard deviations.) The confidence interval we just computed corresponds to the specific random sample of individuals shown in Figure 6-1. Had we selected a *different random sample* of people, say those in Figure 6-2, we would have obtained a *different 95% confidence interval* for the size of the treatment effect.

The individuals selected at random for the experiment in Figure 6-2 show a mean urine production of 1216 mL/day for the people taking the placebo and 1368 mL/day for the people taking the drug. The standard deviations of the two samples are 97 and 263 mL/day, respectively. In these two samples the drug increased average urine production by $\bar{X}_{\text{dr}} - \bar{X}_{\text{pla}} = 1368 - 1216 = 152$ mL/day. The pooled estimate of the population variance is

$$s^2 = \frac{1}{2}(97^2 + 263^2) = 198^2$$

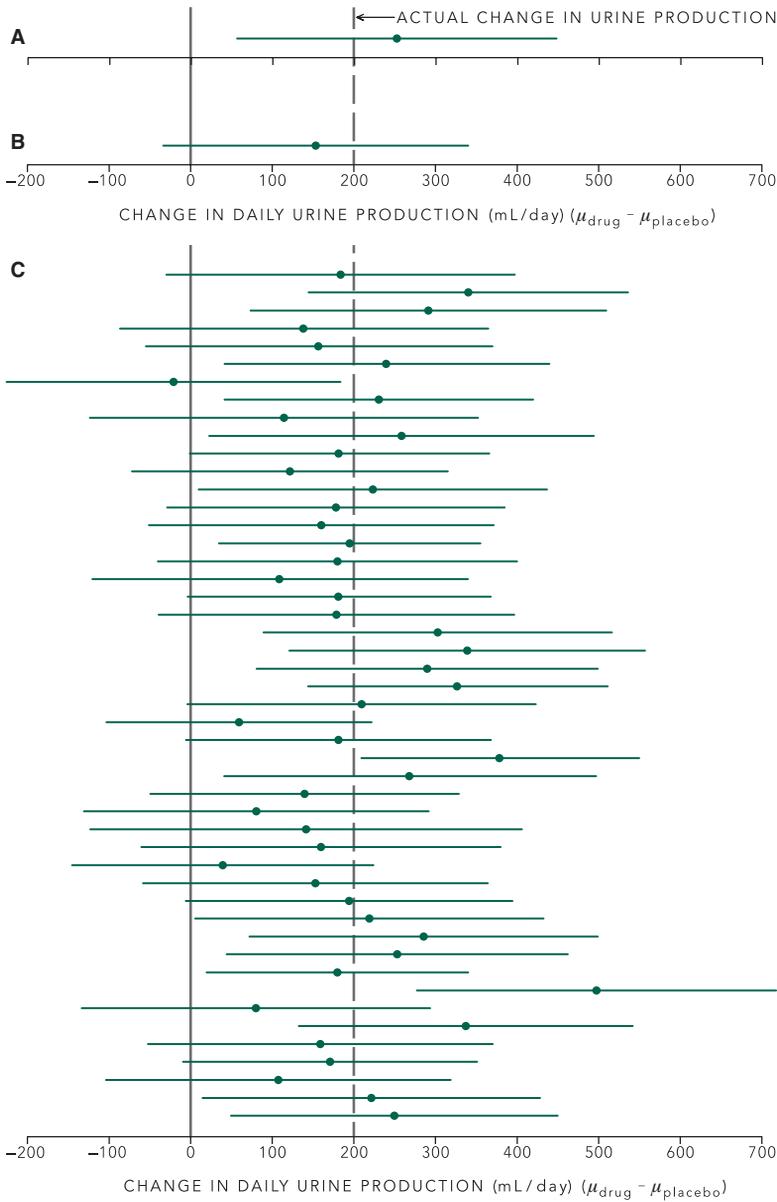


FIGURE 7-1. (A) The 95% confidence interval for the change in urine production produced by the drug using the random samples shown in Figure 6-1. The interval contains the true change in urine production, 200 mL/day (indicated by the dashed line). Since the interval does not include zero (indicated by the solid line), we can conclude that the drug increases urine output ($P < .05$). (B) The 95% confidence interval for change in urine production computed for the random samples shown in Figure 6-2. The interval includes the actual change in urine production (200 mL/day), but it also includes zero, so that it is not possible to reject the hypothesis of no drug effect (at the 5% level). (C) The 95% confidence intervals for 48 more sets of random samples, for example, experiments, drawn from the two populations in Figure 6-1A. All but 3 of the 50 intervals shown in this figure include the actual change in urine production; 5% of all possible 95% confidence intervals will not include the 200 mL/day. Of the 50 confidence intervals, 22 include zero, meaning that the data do not permit rejecting the hypothesis of no difference at the 5% level. In these cases, we would make a Type II error. Since 44% of all possible 95% confidence intervals include zero, the probability of detecting a change in urine production is $1 - \beta = .56$.

in which case,

$$s_{\bar{x}_{dr} - \bar{x}_{pla}} = \sqrt{\frac{198^2}{10} + \frac{198^2}{10}} = 88.5 \text{ mL/day}$$

So the 95% confidence interval for the mean change in urine production associated with the sample shown in Figure 6-2 is

$$152 - 2.101 \cdot 88.5 < \mu_{dr} - \mu_{pla} < 152 + 2.101 \cdot 88.5$$

$$-34 \text{ mL/day} < \mu_{dr} - \mu_{pla} < 340 \text{ mL/day}$$

This interval, while different from the first one we computed, also includes the actual mean increase in urine production, 200 mL/day (Fig. 7-1B). Had we drawn this sample rather than the one in Figure 6-1, we would have been 95% confident that the drug increased average urine production somewhere between -34 and 338 mL/day. (Note that this interval includes negative values, indicating that the data do not permit us to exclude the possibility that the drug decreased as well as increased average urine production. This observation is the basis for using

confidence intervals to test hypotheses later in this chapter.) In sum, the specific 95% confidence interval we obtain depends on the specific random sample we happen to select for observation.

So far, we have seen two such intervals that could arise from random sampling of the populations in Figure 6-1; there are more than 10^{27} possible samples of 10 people each, so there are more than 10^{27} possible 95% confidence intervals. Figure 7-1C shows 48 more of them, computed by selecting two samples of 10 people each from the populations of placebo and drug takers. Of the 50 intervals shown in Figure 7-1, all but 3 (about 5%) include the value of 200 mL/day, the actual change in average urine production associated with the drug.

■ WHAT DOES “CONFIDENCE” MEAN?

We are now ready to attach a precise meaning to the term *95% confident*. The specific 95% confidence interval associated with a given set of data will or will not actually include the true size of the treatment effect, but in the long run 95% of *all possible 95% confidence intervals* will include the true difference of mean values associated with the treatment. As such, it describes not only the size of the effect but quantifies the certainty with which one can estimate the size of the treatment effect.

The size of the interval depends on the level of confidence you want to have that it will actually include the true treatment effect. Since t_α increases as α decreases, requiring a greater and greater fraction of all possible confidence intervals to cover the true effect will make the intervals larger. To see this, let us compute the 90%, 95%, and 99% confidence intervals associated with the data in

Figure 6-1, where the observed mean difference in urine production was 220 mL/day. To do so, we need only substitute the values of $t_{.10}$ and $t_{.01}$ corresponding to $\nu = 18$ from Table 4-1 for t_α in the formula derived above. (We have already solved the problem for $t_{.05}$.)

For the 90% confidence interval, $t_{.10} = 1.734$, so the interval associated with the samples in Figure 6-1 is

$$220 - 1.734 \cdot 89.5 < \mu_{dr} - \mu_{pla} < 220 + 1.734 \cdot 89.5$$

$$65 \text{ mL/day} < \mu_{dr} - \mu_{pla} < 375 \text{ mL/day}$$

which, as Figure 7-2 shows, is narrower than the 95% interval. Does this mean the data now magically yield a more precise estimate of the treatment effect? No. If you are willing to accept the risk that 10% of all possible confidence intervals will not include the true change in mean values, you can get by with a narrower interval.

On the other hand, if you want to specify an interval selected from a population of confidence intervals, 99% of which include the true change in population means, you compute the confidence interval with $t_{.01} = 2.878$. The 99% confidence interval associated with the samples in Figure 6-1 is

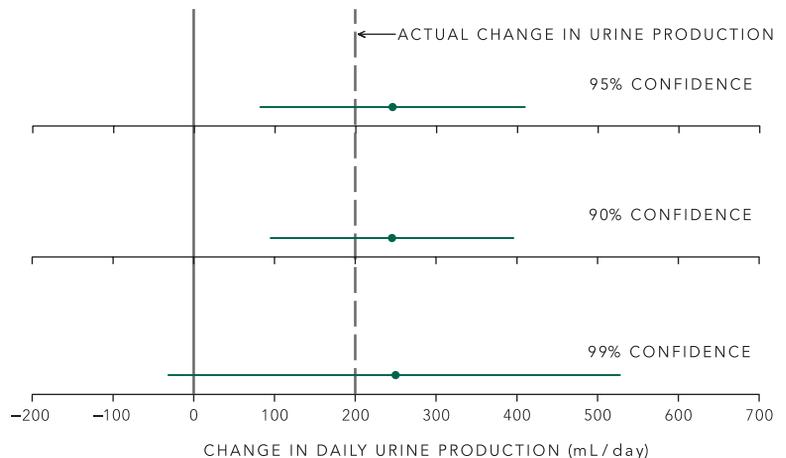
$$220 - 2.878 \cdot 89.5 < \mu_{dr} - \mu_{pla} < 220 + 2.878 \cdot 89.5$$

$$-38 \text{ mL/day} < \mu_{dr} - \mu_{pla} < 478 \text{ mL/day}$$

This interval is wider than the other two in Figure 7-2.

In sum, the confidence interval gives a range that is computed in the hope that it will include the parameter of interest (in this case the difference of two population means). The confidence level associated with the interval (say 95%, 90%, or 99%) gives the percentage of all such possible intervals that will actually include the true value

FIGURE 7-2. Increasing the level of confidence you wish to have that a confidence interval includes the true treatment effect makes the interval wider. All the confidence intervals in this figure were computed from the two random samples shown in Figure 6-1. The 90% confidence interval is narrower than the 95% confidence interval, and the 99% confidence interval is wider. The actual change in urine production, 200 mL/day, is indicated with the dashed line.



of the parameter. A *particular* interval will or will not include the true value of the parameter. Unfortunately, you can never know whether or not that interval does. All you can say is that the chances of selecting an interval that does not include the true value is small (say 5%, 10%, or 1%). The more confidence you wish to have that the interval will cover the true value, the wider the interval.

■ CONFIDENCE INTERVALS CAN BE USED TO TEST HYPOTHESES

As already noted, confidence intervals can provide another way to test statistical hypotheses. This fact should not be surprising because we use all the same ingredients, the difference of the sample means, the standard error of the difference of sample means, and the value of t that corresponds to the biggest α fraction of the possible values defined by the t distribution with v degrees of freedom.

Given a confidence interval one cannot say where within the interval the true difference in population means lies. If the confidence interval contains zero the evidence represented by the experimental observations is not sufficient to rule out the possibility that $\mu_1 - \mu_2 = 0$, that is, that $\mu_1 = \mu_2$, the hypothesis that the t test tests. Hence, we have the following rule:

If the 100(1 - α) percent confidence interval associated with a set of data includes zero, there is not sufficient evidence to reject the hypothesis of no effect with $P < \alpha$. If the confidence interval does not include zero, there is sufficient evidence to reject the hypothesis of no effect with $P < \alpha$.

Apply this rule to the two examples just discussed. The 95% confidence interval in Figure 7-1A does not include zero, so we can report that the drug produced a statistically significant change in urine production ($P < .05$), just as we did using the t test. The 95% confidence interval in Figure 7-1B includes zero, so the random sample (shown in Fig. 6-2) used to compute it does not provide sufficient evidence to reject the hypothesis that the drug has no effect. This, too, is the same conclusion we reached before.

Of the fifty 95% confidence intervals shown in Figure 7-1, twenty-two include zero. Hence $22/50 = 44\%$ of these random samples do not permit reporting a difference with 95% confidence, that is, with $P < .05$. If we looked at all possible 95% confidence intervals computed for these two populations with two samples of 10 people each, we would find that

44% of them include zero, meaning that we would fail to report a true difference, that is, would make a Type II error, 44% of the time. Hence, $\beta = .44$, and the power of the test is .56, which is what we found before (compare with Fig. 6-4).

The confidence interval approach to hypothesis testing offers two potential advantages. In addition to permitting you to reject the hypothesis of no effect when the interval does not include zero, it also gives information about the size of the effect. Thus, if a result reaches statistical significance more because of a large sample size than because of a large treatment effect, the confidence interval will show it. In other words, it will make it easier to recognize effects that can be detected with confidence but are too small to be of clinical or scientific significance.

For example, suppose we wish to study the potential value of a proposed antihypertensive drug. We select two samples of 100 people each and administer a placebo to one group and the drug to the other. The treated group has a mean diastolic pressure of 81 mmHg and a standard deviation of 11 mmHg; the control (placebo) group has a mean blood pressure of 85 mmHg and a standard deviation of 9 mmHg. Are these data consistent with the hypothesis that the diastolic blood pressure among people taking the drug and placebo were actually no different? To answer this question, we use the data to complete a t test. The pooled-variance estimate is

$$s^2 = \frac{1}{2}(11^2 + 9^2) = 10^2 \text{ mmHg}^2$$

so

$$t = \frac{\bar{X}_{\text{dr}} - \bar{X}_{\text{pla}}}{s_{\bar{X}_{\text{dr}} - \bar{X}_{\text{pla}}}} = \frac{81 - 85}{\sqrt{(10^2/100) + (10^2/100)}} = \frac{-4}{1.41} = -2.83$$

This value is more negative than -2.61 , the critical value of t that defines the 1% most extreme of the t distribution with $v = 2(n - 1) = 198$ degrees of freedom (from Table 4-1). Thus, we conclude that the drug lowers diastolic blood pressure ($P < .01$).

But is this result clinically significant? To gain a feeling for this, compute the 95% confidence interval for the mean difference in diastolic blood pressure for people taking the placebo versus the drug. Since $t_{.05}$ for 198 degrees of freedom is (from Table 4-1) 1.973, the confidence interval is

$$\begin{aligned} -4 - 1.972 \cdot 1.41 &< \mu_{\text{dr}} - \mu_{\text{pla}} < -4 + 1.972 \cdot 1.41 \\ -6.8 \text{ mmHg} &< \mu_{\text{dr}} - \mu_{\text{pla}} < -1.2 \text{ mmHg} \end{aligned}$$

BOX 7-1 • The Effect on Temperature of Using Polyethylene Bags to Keep Extreme Low Birth Weight Infants Warm

The skin temperature for the 70 infants wrapped in polyethylene bags was 36°C with a standard deviation of 1°C and 35°C with a standard deviation of 1°C for the 70 infants kept warm using traditional methods. To compute the 95% confidence interval for the difference in temperature, we first compute the observed mean difference in temperature

$$\bar{X}_{\text{bag}} - \bar{X}_{\text{trad}} = 36 - 35 = 1^\circ\text{C}$$

and the standard error of the difference

$$s_{\bar{X}_{\text{bag}} - \bar{X}_{\text{trad}}} = \sqrt{\frac{s^2}{n_{\text{bag}}} + \frac{s^2}{n_{\text{trad}}}} = \sqrt{\frac{1^2}{70} + \frac{1^2}{70}} = .169^\circ\text{C}$$

because

$$s^2 = \frac{(n_{\text{bag}} - 1)s_{\text{bag}}^2 + (n_{\text{trad}} - 1)s_{\text{trad}}^2}{n_{\text{bag}} + n_{\text{trad}} - 2} = \frac{(70 - 1)1^2 + (70 - 1)1^2}{70 + 70 - 2} = 1^\circ\text{C}$$

There are $\nu = n_{\text{bag}} + n_{\text{trad}} - 2 = 70 + 70 - 2 = 138$ degrees of freedom associated with this estimate. From Table 4-1 the critical value of t that defines the 5% most extreme values of the t distribution for 138 degrees of freedom is 1.977, so the 95% confidence interval for the difference in temperature is

$$1 - 1.977 \cdot .169 < \mu_{\text{bag}} - \mu_{\text{trad}} < 1 + 1.977 \cdot .169$$

$$.67^\circ\text{C} < \mu_{\text{bag}} - \mu_{\text{trad}} < 1.33^\circ\text{C}$$

Because the 95% confidence interval does not include 0, we can reject the null hypothesis that the wrapping technique did not affect the infants' temperature ($P < .05$).

From Table 4-1, the critical value of t that defines the 1% most extreme values of the t distribution is 2.611, so the 99% confidence interval for the difference in temperature is

$$1 - 2.611 \cdot .169 < \mu_{\text{bag}} - \mu_{\text{trad}} < 1 + 2.611 \cdot .169$$

$$.54^\circ\text{C} < \mu_{\text{bag}} - \mu_{\text{trad}} < 1.44^\circ\text{C}$$

Because the 99% confidence interval also excludes 0, we can also reject the null hypothesis with $P < .01$. (Compare this result with Prob. 4-2.)

In other words, we can be 95% confident that the drug lowers blood pressure between -6.8 and -1.2 mmHg. This is not a very large effect, especially when compared with standard deviations of the blood pressures observed within each of the samples, which are around 10 mmHg. Thus, while the drug does seem to lower blood pressure on the average, examining the confidence interval permitted us to see that the size of the effect is not very impressive. The small value of P was more a reflection of the sample size than the size of the effect on blood pressure.

The study of the effects of using polyethylene bags to keep extremely low birth weight infants warm discussed in Chapter 5 found no difference in survival between using these bags and traditional methods despite the fact that the bags did statistically significantly increase body temperature by 1°C. Box 7-1 shows that the 95% confidence interval for the actual increase in temperature ranges from $.67^\circ\text{C}$ to 1.33°C . While this difference was *statistically significant*, it does not appear to have been a big enough effect to be *clinically significant*.

This example illustrates the importance of examining not only the P values reported in a study but also the *size* of the treatment effect compared with the variability within each of the treatment groups. Usually this comparison requires converting the reported standard errors of the mean reported in the paper to standard deviations by multiplying them by the square root of the sample size. This simple step often shows clinical studies to be of potential interest in illuminating physiological mechanisms but of little value in diagnosing or managing a specific patient because of person-to-person variability.

■ CONFIDENCE INTERVAL FOR THE POPULATION MEAN

The procedure we developed above can be used to compute a confidence interval for the mean of the population from which a sample was drawn. The resulting confidence interval is the origin of the rule, stated in Chapter 2, that the true (and unobserved) mean of the original population will lie within about 2 standard errors of the mean of the sample mean for 95% of all possible samples.

The confidence intervals we computed up to this point are based on the fact that

$$t = \frac{\text{Difference of sample means} - \text{difference in population means}}{\text{Standard error of difference of sample means}}$$

follows the t distribution. It is also possible to show that

$$t = \frac{\text{Sample mean} - \text{population mean}}{\text{Standard error of mean}}$$

follows the t distribution. The equivalent mathematical statement is

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$$

We can compute the $100(1 - \alpha)$ percent confidence interval for the population mean by obtaining the value of t_{α} corresponding to $\nu = n - 1$ degrees of freedom, in which n is the sample size. Substitute this value for t in the equation and solve for μ (just as we did for $\mu_1 - \mu_2$ earlier).

$$\bar{X} - t_{\alpha} s_{\bar{X}} < \mu < \bar{X} + t_{\alpha} s_{\bar{X}}$$

The interpretation of the confidence interval for the mean is analogous to the interpretation of the confidence interval for the difference of two means: every possible random sample of a given size can be used to compute a, say,

95% confidence interval for the population mean, and this same percentage (95%) of all such intervals will include the true population mean.

It is common to approximate the 95% confidence interval with the sample mean plus or minus twice the standard error of the mean because the values of $t_{.05}$ are approximately 2 for sample sizes above about 20 (see Table 4-1). This approximate rule of thumb does underestimate the size of the confidence interval for the mean, however, especially for the small sample sizes common in biomedical research.

■ THE SIZE OF THE TREATMENT EFFECT MEASURED AS THE DIFFERENCE OF TWO RATES OR PROPORTIONS

It is easy to generalize the procedures we just developed to permit us to compute confidence intervals for rates and proportions. In Chapter 5 we used the statistic

$$z = \frac{\text{Difference of sample proportions}}{\text{Standard error of difference of proportions}}$$

to test the hypothesis that the observed proportions of events in two samples were consistent with the hypothesis that the event occurred at the same rate in the two populations. It is possible to show that even when the two populations have different proportions of members with the attribute, the ratio

$$z = \frac{\text{Difference of sample proportions} - \text{difference in population proportions}}{\text{Standard error of difference of sample proportions}}$$

is distributed approximately according to the normal distribution so long as the sample sizes are large enough.

If p_1 and p_2 are the actual proportions of members of each of the two populations with the attribute, and if the corresponding estimates computed from the samples are \hat{p}_1 and \hat{p}_2 , respectively,

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{s_{\hat{p}_1 - \hat{p}_2}}$$

We can use this equation to define the $100(1 - \alpha)$ percent confidence interval for the difference in proportions by substituting z_{α} for z in this equation and solving just as we did before. z_{α} is the value that defines the most extreme α proportion of the values in the normal

distribution;* $z_{\alpha} = z_{0.05} = 1.960$ is commonly used, since it is used to define the 95% confidence interval. Thus,

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha} s_{\hat{p}_1 - \hat{p}_2} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha} s_{\hat{p}_1 - \hat{p}_2}$$

for 100(1 - α) percent of all possible samples.

Difference in Survival for Two Methods for Keeping Extremely Low Birth Weight Infants Warm

In Chapter 5, we used a contingency table analysis to test the null hypothesis that polyethylene bags and traditional warming methods had the same effect on infant survival. We now use a confidence interval approach to test the same null hypothesis.

The data in Table 5-3 showed that 90% (63 out of 70) of infants kept warm with polyethylene bags survived as did 87% (61 of 70) of infants kept warm using traditional methods. Therefore, the observed difference in survival was $\hat{p}_{\text{bag}} - \hat{p}_{\text{trad}} = .90 - .87 = .03$. The overall proportion of all infants who survived is

$$\hat{p} = \frac{83 + .61}{70 + 70} = .886$$

So the standard error of the difference is

$$\begin{aligned} s_{\hat{p}_{\text{bag}} - \hat{p}_{\text{trad}}} &= \sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_{\text{bag}}} + \frac{1}{n_{\text{trad}}} \right)} \\ &= \sqrt{.886(1 - .886) \left(\frac{1}{70} + \frac{1}{70} \right)} = .054 \end{aligned}$$

Therefore, the 95% confidence interval for the difference in survival rates is

$$\begin{aligned} (\hat{p}_{\text{bag}} - \hat{p}_{\text{trad}}) - z_{.05} s_{\hat{p}_{\text{bag}} - \hat{p}_{\text{trad}}} &< (p_{\text{bag}} - p_{\text{trad}}) \\ &< (\hat{p}_{\text{bag}} - \hat{p}_{\text{trad}}) + z_{.05} s_{\hat{p}_{\text{bag}} - \hat{p}_{\text{trad}}} \\ .03 - 1.960 \cdot .054 &< (p_{\text{bag}} - p_{\text{trad}}) < .03 + 1.960 \cdot .054 \\ -.076 &< (p_{\text{bag}} - p_{\text{trad}}) < .136 \end{aligned}$$

Thus, we can be 95% confident that the true difference in survival lies between a 7.6% better survival rate for

traditional warming methods and a 13.6% better survival rate for the polyethylene bags.[†] Since the 95% confidence interval contains zero, there is not sufficient evidence to reject the null hypothesis that the two warming techniques are associated with the same survival rates. Furthermore, the confidence interval ranges about equally on both sides of zero, so there is not even a suggestion that one method is superior to the other.

■ HOW NEGATIVE IS A “NEGATIVE” CLINICAL TRIAL?

Chapter 6 discussed the study of 71 randomized clinical trials that did not demonstrate a statistically significant improvement in clinical outcome (mortality, complications, or the number of patients who showed no improvement, depending on the study). Most of these trials involved too few patients to have sufficient power to be confident that the failure to detect a treatment effect was not due to an inadequate sample size. To get a feeling for how compatible the data are with the hypothesis of no treatment effect, let us examine the 90% confidence intervals for the proportion of “successful” cases (the definition of success varied with the study) for all 71 trials. Figure 7-3 shows these confidence intervals.

All the confidence intervals include zero, so we cannot rule out the possibility that the treatments had no effect. Note, however, that some of the trials are also compatible with the possibility that the treatments produced sizable improvements in the success rate. Remember that while we can be 90% confident that the true change in proportion of successes lies in the interval, it could be anywhere. Does this prove that some of these treatments improved clinical outcome? No. The important point is that the confidence with which we can assert that there was no treatment effect is often the same as the confidence with which we can assert that the treatment produced a sizable improvement. While the size and location of the confidence interval cannot be used as part of a formal statistical argument to prove that the treatment had an effect, it certainly can help you look for trends in the data.

Meta-Analysis

While the ideal solution to avoiding the problem we have been discussing would be to do large, well-powered studies

*This value can be obtained from a *t* table, for example Table 6-2, by taking the value of *t* corresponding to an infinite number of degrees of freedom.

[†]To include the Yates correction, widen the upper and lower bounds for the confidence interval by $\frac{1}{2}(1/n_{\text{bag}} + 1/n_{\text{trad}})$.

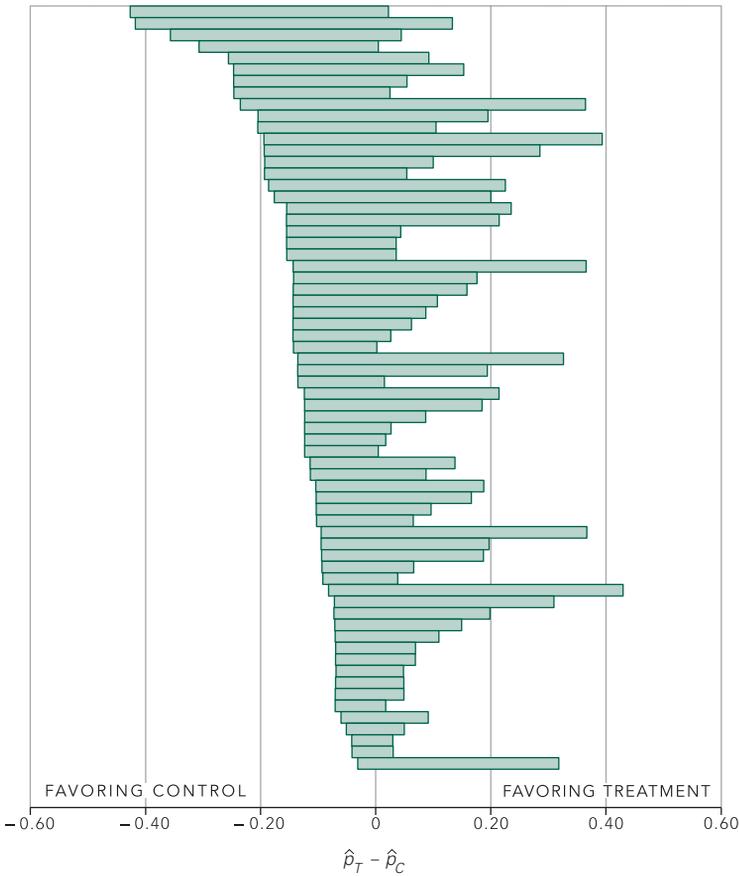


FIGURE 7-3. The 90% confidence intervals for 71 negative clinical trials. Since all the intervals contain zero, there is not sufficient evidence that the success rate is different for the treatment and control groups. Nevertheless, the data are also compatible with the treatment producing a substantial improvement in success rate in many of the trials. While this study was done in 1978, based on clinical trials conducted before then, the problem of drawing negative conclusions based on underpowered clinical trials persists in the 21st century. (From Fig. 2 of Freiman JA, Chalmers TC, Smith H Jr, Keubler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 “negative” trials. *N Engl J Med.* 1978;299:690–694.)

that would yield estimates of the size of the effect that was studied along with a narrow confidence interval, the unfortunate fact remains that doing so is not always possible, either because of practical limitations (such as not being able to recruit enough subjects at the institution doing the study) or financial limitations. Fortunately, there is an approach that permits you to combine the results of several similar studies to obtain a single estimate of the effect that integrates all the available information.

This approach, known as *meta-analysis*, is essentially a procedure for pooling the results of the individual studies as if they were one much larger study.* Because the effec-

*The calculations involved in—and limitations of—meta-analysis are beyond the scope of this book. For a discussion of how to conduct a meta-analysis, see Petitti DB. *Meta-Analysis, Decision Analysis, and Cost-effectiveness Analysis: Methods for Quantitative Synthesis in Medicine*, 2nd ed. New York: Oxford University Press; 2000 or Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-Analysis in Medical Research*. West Sussex, England: John Wiley & Sons; 2000.

tive sample size is increased by combining all the studies, the associated confidence interval is narrower and the power of the combined analysis is increased. These two effects create a situation in which you can be more confident of both positive and negative conclusions than is possible when considering each individual study separately.

Figure 7-4 shows the results of 29 different studies of the relative risk of developing heart disease associated with being regularly exposed to secondhand tobacco smoke (defined as a nonsmoker living or working with a smoker) compared with people not exposed. Each line on the top part of Figure 7-4 represents the results of one of the studies. The points represent the observed risk in each study and the lines span the 95% confidence interval associated with each study. Not surprisingly, there is variability in the estimates of the effect sizes from study to study (because of the random sampling process inherent in making estimates from any sample). Several of the

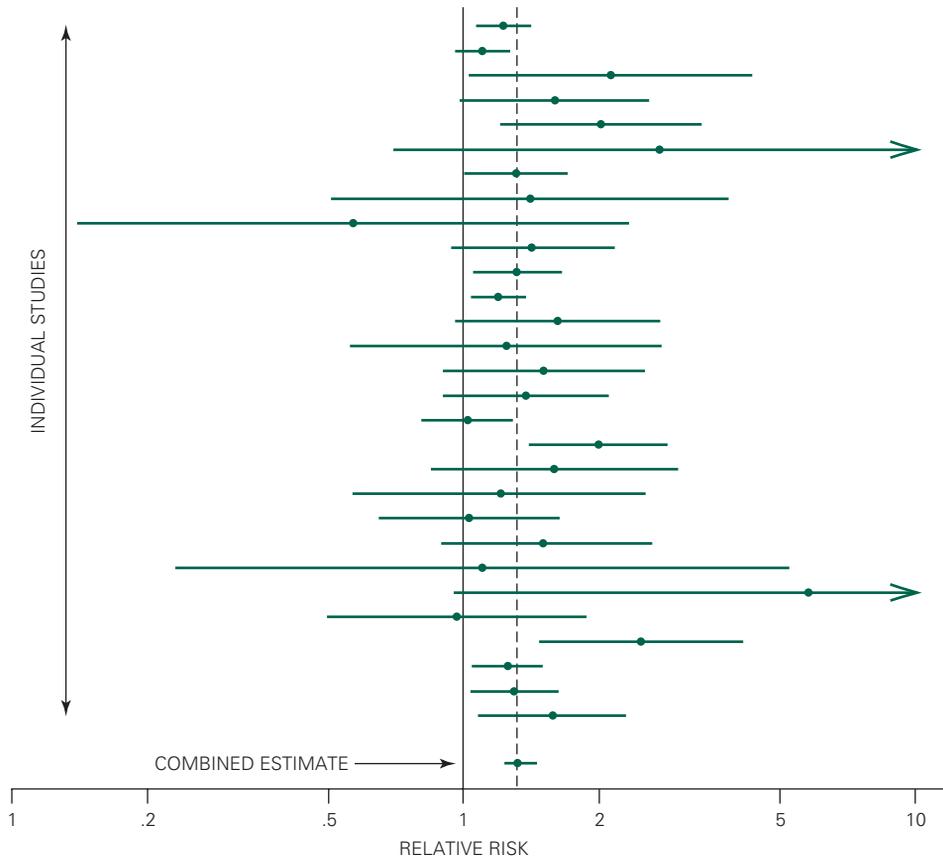


FIGURE 7-4. A meta-analysis of the 29 studies of the relative risk of developing heart disease when exposed to secondhand smoke yields a single combined estimate of the risk with a much narrower 95% confidence interval (bottom of figure) than any of the individual studies (above the combined estimate). This more precise estimate of the risk results from the fact that the combined risk estimate uses all the information from the 29 individual studies and so has a much larger effective sample size than any of the individual studies. (This figure is based on information from Barnoya J, Glantz S. Cardiovascular effects of secondhand smoke: nearly as large as smoking. *Circulation*. 2005;111:2684–2698. For the complete analysis, see the source article.)

confidence intervals exclude a relative risk of 1.0, meaning that those studies found a statistically significant elevation in heart disease risk associated with secondhand smoke exposure. At the same time, several of the studies yielded confidence intervals including 1.0, meaning that you could not conclude that secondhand smoke increased the risk of heart disease based on those individual studies taken alone. Note also that many of the studies had wide confidence intervals associated with them, because of small sample sizes.

The estimate at the bottom of Figure 7-4 shows the results of combining all the individual studies with a meta-

analysis. While only some of the 29 individual studies of the risk of heart disease associated with breathing secondhand smoke were large enough to reach conventional statistical significance (at the .05 level), the combined estimate of a relative risk of 1.31 and the narrow confidence interval (from 1.21 to 1.41) means that we can have a high level of confidence in concluding that there is an increased risk of heart disease in people regularly exposed to secondhand smoke. Because this estimate is based on all the data from all 18 studies, the effective sample size is substantially greater than any of the individual studies, which is the reason that the 95% confidence interval for

the combined estimate of the effect size is so much narrower than for the individual studies.

While not perfect, meta-analysis has become an important tool for combining information from several related studies and dealing with the problem that individual studies lack adequate power to provide high confidence in drawing negative conclusions.

■ CONFIDENCE INTERVAL FOR RATES AND PROPORTIONS

It is possible to use the normal distribution to compute approximate confidence intervals for proportions from observations, so long as the sample size is large enough to make the approximation reasonably accurate.* When it is not possible to use this approximation, we will compute the exact confidence intervals based on the binomial distribution. While we will not go into the computational details of this procedure, we will present the necessary results in graphical form because papers often present results based on small numbers of subjects. Examining the confidence intervals as opposed to only the observed proportion of patients with a given attribute is especially useful in thinking about such studies, because a change of *a single patient* from one group to the other often makes a large difference in the observed proportion of patients with the attribute of interest.

Just as there was an analogous way to use the *t* distribution to relate the difference of means and the confidence interval for a single sample mean, it is possible to show that if the sample size is large enough,

$$z = \frac{\text{Observed proportion} - \text{true proportion}}{\text{Standard error of proportion}}$$

In other words,

$$z = \frac{\hat{p} - p}{s_{\hat{p}}}$$

approximately follows the normal distribution (in Table 6-2). Hence, we can use this equation to define the

100(1 - α) percent confidence interval for the true proportion *p* with

$$\hat{p} - z_{\alpha} s_{\hat{p}} < p < \hat{p} + z_{\alpha} s_{\hat{p}}$$

Quality of Evidence Used as a Basis for Interventions to Improve Hospital Antibiotic Prescribing

Despite many efforts to control antibiotic usage and promote optimal prescribing, practitioners continue to prescribe inappropriately, which contributes not only to increased medical costs but also to the development of antibiotic-resistant bacteria. The British Society for Antimicrobial Chemotherapy and Hospital Infection Society convened a Working Party to address the problem of antibiotic prescribing in hospitals.[†] They did an exhaustive literature search and located 306 papers dealing with recommendations for antibiotic use. They then applied the quality criteria of the Cochrane Collaboration, an international effort that promotes high quality systematic reviews of the literature, and found that 91 of the papers met the minimum criteria for inclusion in a Cochrane review. What is the 95% confidence interval for the fraction of articles that met these quality criteria?

The proportion of acceptable articles is $\hat{p} = 91/306 = .297$ so the standard error of the proportion is

$$s_{\hat{p}} = \sqrt{\frac{.297(1 - .297)}{306}} = .026$$

Therefore, the 95% confidence interval for the proportion of acceptable articles is

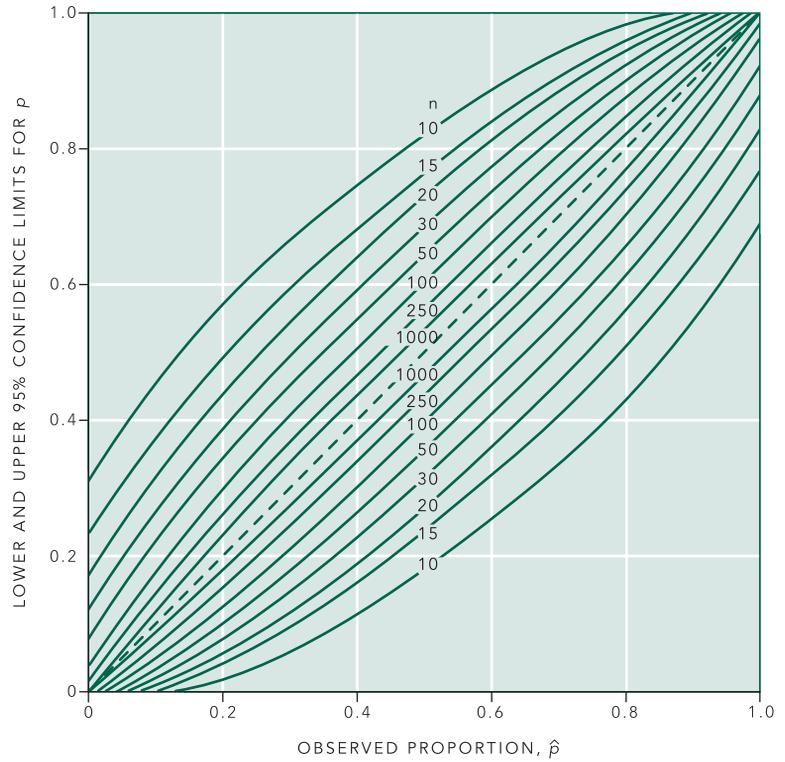
$$.297 - 1.960 \cdot .026 < p < .297 + 1.960 \cdot .026 \\ 246 < p < .348$$

In other words, based on this sample, we can be 95% confident that the true proportion of papers on antibiotic prescribing guidelines that met the Cochrane criteria was between 25% and 35%.

*As discussed in Chapter 5, $n\hat{p}$ and $n(1 - \hat{p})$ must both exceed about 5, where \hat{p} is the proportion of the observed sample having the attribute of interest.

[†]Ramsay C, Brown E, Hartman G, Davey P. Room for improvement: a systematic review of the quality of evaluations to improve hospital antibiotic prescribing. *J Antimicrob Chemother.* 2003;52:764-771.

FIGURE 7-5. Graphical presentation of the exact 95% confidence intervals (based on the binomial distribution) for the population proportion. You read this plot by reading the two limits of the lines defined by the sample size at the point on the horizontal axis at the proportion of the sample with the attribute of interest \hat{p} (Adapted from Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934;26:404–413.)



■ EXACT CONFIDENCE INTERVALS FOR RATES AND PROPORTIONS

When the sample size or observed proportion is too small for the approximate confidence interval based on the normal distribution to be reliable, you have to compute the confidence interval based on the exact theoretical distribution of a proportion, the *binomial distribution*.^{*} Since results based on small sample sizes with low observed rates of events turn up frequently in the medical literature, we present the results of computation of confidence intervals using the binomial distribution.

To illustrate how the procedure we followed above can fall apart when $n\hat{p}$ is below about 5, we consider an example. Suppose a surgeon says that he has done 30 operations without a single complication. His observed complication

rate \hat{p} is $0/30 = 0\%$ for the 30 specific patients he operated on. Impressive as this is, it is unlikely that the surgeon will continue operating forever without a complication, so the fact that $\hat{p} = 0$ probably reflects good luck in the randomly selected patients who happened to be operated on during the period in question. To obtain a better estimate of p , the surgeon's true complication rate, we will compute the 95% confidence interval for p .

Let us try to apply our existing procedure. Since $\hat{p} = 0$,

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0(1-0)}{30}} = 0$$

and the 95% confidence interval is from zero to zero. This result does not make sense. There is no way that a surgeon can *never* have a complication. Obviously, the approximation breaks down.

Figure 7-5 gives a graphical presentation of the 95% confidence intervals for proportions. The upper and lower limits are read off the vertical axis using the pair of curves corresponding to the size of the sample n used to estimate \hat{p} at the point on the horizontal axis corresponding to the

^{*}The reason we could use the normal distribution here and in Chapter 5 is that for large enough sample sizes there is little difference between the binomial and normal distributions. This result is a consequence of the central-limit theorem, discussed in Chapter 2.

observed \hat{p} . For our surgeon, $\hat{p} = 0$ and $n = 30$, so the 95% confidence interval for his true complication rate is from 0 to .12. In other words, we can be 95% confident that his true complication rate, based on the 30 cases we happened to observe, is somewhere between 0% and 12%.

Now, suppose the surgeon had a single complication. Then $\hat{p} = 1/30 = 0.033$ and

$$s_{\hat{p}} = \sqrt{\frac{.033(1-.033)}{30}} = .0326$$

so the 95% confidence interval for the true complication rate, computed using the approximate method, is

$$.033 - 1.960 \cdot .0326 < p < .033 + 1.960 \cdot .0326$$

$$-.031 < p < .097$$

Think about this result for a moment. There is no way a surgeon can have a *negative* complication rate.

Figure 7-5 gives the exact confidence interval, from 0 to .17, or 0% to 17%.^{*} This confidence interval is not too different from that computed when there were no complications, as it should be, since there is little real difference between not having any complications and having only one complication in such a small sample.

Notice how important sample size is, especially for small sample sizes. Had the surgeon been bragging that he had a zero complication rate on the basis of only 10 cases, the 95% confidence interval for his true complication rate would have extended from zero all the way to 31%!

■ CONFIDENCE INTERVALS FOR RELATIVE RISK AND ODDS RATIO[†]

Because the relative risk and odds ratio are ratios, the distributions of the values of these statistics are not normally distributed. It turns out, however, that the logarithm of these ratios is normally distributed. Therefore, we can use approaches similar to those used with

^{*}When there are no “failures” observed, the approximate upper end of the 95% confidence interval for the true failure rate is approximately $3/n$, where n is the sample size. For a more extensive discussion of interpreting results when there are no “failures,” see Hanley JA, Lippman-Hand A. If nothing goes wrong, is everything all right? interpreting zero numerators. *JAMA*. 1983;249:1743–1745.

[†]If time is limited, this section can be skipped without any loss of continuity.

proportions to the logarithms of the relative risk and odds ratio, then invert the results to return to the original scale. By convention, statisticians and epidemiologists use the natural logarithm for these calculations.[‡] Using the notation in Table 5-14, the natural logarithm of the relative risk, $\ln RR$, is normally distributed with standard error

$$s_{\ln RR} = \sqrt{\frac{n_{TD}/n_T + n_{CD}/n_C}{n_{TD}}}$$

Therefore, the $100(1 - \alpha)$ percent confidence interval for the natural logarithm of the true population $\ln RR_{\text{true}}$ is

$$\ln RR - z_{\alpha} s_{\ln RR} < \ln RR_{\text{true}} < \ln RR + z_{\alpha} s_{\ln RR}$$

We convert these estimates back to the original units by applying the exponential function to the terms in this equation to obtain

$$e^{\ln RR - z_{\alpha} s_{\ln RR}} < RR_{\text{true}} < e^{\ln RR + z_{\alpha} s_{\ln RR}}$$

Thus, you could test the null hypothesis that the true $RR = 1$, that the treatment (or risk factor) had no effect, by computing this confidence interval and seeing if it included 1.0.

Likewise, the natural logarithm of the odds ratio, OR, is normally distributed. Using the notation in Table 5-15, the standard error is

$$s_{\ln OR} = \sqrt{\frac{1}{n_{ED}} + \frac{1}{n_{EN}} + \frac{1}{n_{UD}} + \frac{1}{n_{UN}}}$$

and the $100(1 - \alpha)$ percent confidence interval for the true odds ratio is

$$e^{\ln OR - z_{\alpha} s_{\ln OR}} < OR_{\text{true}} < e^{\ln OR + z_{\alpha} s_{\ln OR}}$$

This confidence interval can also be used to test the null hypothesis that the true $OR = 1$, that exposure to the risk factor is not associated with an increase in the odds of having the disease.

[‡]The natural logarithm has the base $e = 2.71828 \dots$ rather than 10, which is the base of the common logarithm. Because e is the base, the natural logarithm and exponential functions are *inverses*, that is, $e^{\ln x} = x$ and $\ln e^x = x$.

Effect of Counseling on Filing Advance Directives for End of Life Care among Homeless People

In Chapter 5 we rejected the null hypothesis that homeless people provided with in-person counseling about end-of-life care filed advance directives at the same rate as people who were simply give written materials (Table 5-1). In that study, 38% (55 out of 145) of the people receiving in-person counseling filed advance directives and 13% (15 out of 117) of the people receiving written materials filed advanced directives. As also presented, the fact that this study was a prospective randomized clinical trial, we could compute a relative risk from these data,

$$RR = \frac{\hat{p}_{\text{counsel}}}{\hat{p}_{\text{written}}} = \frac{.379}{.128} = 2.92$$

Now we will compute the 95% confidence interval for this relative risk to test the null hypothesis that the two ways of educating homeless people lead to the same rates of filing advance directives.

Using the data in Table 5-1 and the notation in Table 5-14, $n_{TD} = 55$, $n_{TN} = 90$, $n_{CD} = 15$, and $n_{CN} = 102$, so $n_T = n_{TD} + n_{TN} = 55 + 90 = 145$ and $n_D = n_{CD} + n_{CN} = 15 + 102 = 117$ and the standard error of $\ln RR$ is

$$s_{\ln RR} = \sqrt{\frac{1 - \frac{55}{145}}{55} + \frac{1 - \frac{15}{117}}{15}} = .263$$

To estimate the 95% confidence interval, we note that $z_{.05} = 1.960$ and compute

$$e^{\ln 2.92 - 1.960 \cdot .263} < RR < e^{\ln 2.92 + 1.960 \cdot .263}$$

$$e^{.556} < RR < e^{1.587}$$

$$1.74 < RR < 4.89$$

Hence, we can be 95% confident that the true relative risk of homeless people filing advance directives if they receive in-person counseling compared to receiving written instructions is between 1.74 and 4.89. Because this range does not include 1 (the same probabilities of filing an advance directive for both treatment groups), we conclude that in-person counseling significantly increases the likelihood that a homeless person will file an advance directive if he or she receives in-person counseling.

Box 7-2 shows that the 95% confidence interval for the relative risk of death for extremely low birth weight

BOX 7-2 • Relative Risk of Death in Extreme Low Birth Weight Infants for Different Warming Techniques

From the data in Table 5-3, 7 of 70 infants kept warm (treated) with polyethylene bags died as did 9 of 70 kept warm using traditional methods. Therefore, $n_{TD} = 7$ out of $n_T = 70$ and $n_{CD} = 9$ out of $n_T = 70$, so

$$RR = \frac{\frac{n_{TD}}{n_T}}{\frac{n_{CD}}{n_C}} = \frac{\frac{7}{70}}{\frac{9}{70}} = \frac{.100}{.129} = .775$$

The fact that the relative risk is below 1 indicates that the death rate is lower in the polyethylene bag-treated group than the traditional treatment group. The standard error of $\ln RR$ is

$$s_{\ln RR} = \sqrt{\frac{1 - \frac{n_{TD}}{n_T}}{n_{TD}} + \frac{1 - \frac{n_{CD}}{n_C}}{n_{CD}}}$$

$$= \sqrt{\frac{1 - \frac{7}{70}}{7} + \frac{1 - \frac{9}{70}}{9}} = .475$$

To estimate the 95% confidence interval, we note that $z_{.05} = 1.960$ and compute

$$e^{\ln .775 - 1.96 \times .475} < RR < e^{\ln .775 + 1.96 \times .475}$$

$$e^{-1.186} < RR < e^{.676}$$

$$.305 < RR < 1.96$$

Because the 95% confidence interval includes 1—equal risks of death for both treatments—we do not reject the null hypothesis of no difference between the two treatments.

infants kept warm with polyethylene bags is $.305 < RR < 1.96$. Since this interval straddles 1, we conclude that the warming procedure does not significantly affect survival, just as we did before. Indeed, testing this null hypothesis using a comparison of proportions, contingency table and relative risk all yield precisely the same results because they are all simply different presentations of the same underlying statistical test.

Passive Smoking and Breast Cancer

We can compute the confidence interval for the odds ratio of a premenopausal woman who is exposed to second-hand smoke developing breast cancer using the data in Table 5-16. To compute the 95% confidence interval for this odds ratio, we note that the observed odds ratio is 2.91 and, from Table 5-16, $n_{ED} = 50$, $n_{EN} = 14$, $n_{UD} = 43$, and $n_{UN} = 35$. Therefore,

$$s_{\ln OR} = \sqrt{\frac{1}{50} + \frac{1}{14} + \frac{1}{43} + \frac{1}{35}} = .378$$

and, so,

$$e^{\ln 2.91 - 1.960 \cdot .378} < OR_{\text{true}} < e^{\ln 2.91 + 1.960 \cdot .378}$$

$$e^{.327} < OR_{\text{true}} < e^{1.809}$$

$$1.39 < OR_{\text{true}} < 6.10$$

Thus, we can be 95% confident that the true odds ratio is somewhere between 1.39 and 6.10. Because the 95% confidence interval for the true odds ratio excludes 1, we conclude that passive smoking significantly increases the odds of breast cancer in premenopausal women.

■ CONFIDENCE INTERVAL FOR THE ENTIRE POPULATION*

So far computed intervals that we can have a high degree of confidence in will include a *population parameter*, such as μ or p . It is often desirable to determine a confidence interval for the *population itself*, most commonly when defining the normal range of some variable. The most common approach is to take the range defined by 2 standard deviations about the sample mean on the grounds that this interval contains 95% of the members of a population that follows the normal distribution (Fig. 2-5). In fact, in carefully worded language Chapter 2 suggested this rule. When the sample used to estimate the mean and standard deviation is large (more than 100 to 200 members), this common rule of thumb is reasonably accurate.

*Confidence intervals for the population are also called *tolerance limits*. The procedures derived in this section are appropriate for analyzing data obtained from a population that is normally distributed. If the population follows other distributions, there are alternate procedures for computing confidence intervals for the population.

Unfortunately, many studies are based on much smaller samples (of the order of 5 to 20 individuals). With such small samples, use of this two standard deviations rule of thumb seriously underestimates the range of values likely to be included in the population from which the samples were drawn.

For example, Figure 2-8 showed the population of the heights of all 200 Martians, together with the results of three random samples of 10 Martians each. Figure 2-8A showed that 95% of all Martians have heights between 31 and 49 cm. The mean and standard deviation of the heights of population of all 200 Martians are 40 and 5 cm, respectively. The three samples illustrated in Figure 2-8 yield estimates of the mean of 41.5, 36, and 40 cm, and of the standard deviation of 3.8, 5, and 5 cm, respectively. Suppose we simply compute the range defined by two *sample* standard deviations above and below the *sample* mean with the expectation that this range will include 95% of the population. Figure 7-6A shows the results of this computation for each of the three samples in Figure 2-8. The light area defines the range of actual heights that covers 95% of the Martians' heights. Two of the three samples yield intervals that do not include 95% of the population.

This problem arises because both the sample mean and standard deviation are only *estimates* of the population mean and standard deviation and so cannot be used interchangeably with the population mean and standard deviation when computing the range of population values. To see why, consider the sample in Figure 2-8B that yielded estimates of the mean and standard deviation of 36 and 5 cm, respectively. By good fortune, the estimate of the standard deviation computed from the sample equaled the population standard deviation. The estimate of the population mean, however, was low. As a result, the interval 2 standard deviations above and below the sample mean did not reach high enough to cover 95% of the entire population values. Because of the potential errors in the estimates of the population mean and standard deviation, we must be conservative and use a range greater than 2 standard deviations around the sample mean to be sure of including, say, 95% of the entire population. However, as the size of the sample used to estimate the mean and standard deviation increases, the certainty with which we can use these estimates to compute the range spanned by the entire population increases, so we do not have to be as conservative (i.e., take fewer multiples of the sample standard deviation) when computing an interval that contains a specified proportion of the population members.

FIGURE 7-6. (A) The range defined by the sample mean ± 2 standard deviations for the three samples of 10 Martians each shown in Figure 2-8. Two of the three resulting ranges do *not* cover the entire range that includes 95% of the population members (indicated by the white area). **(B)** The 95% confidence intervals for the population, computed as the sample mean $\pm K_{.05}$ times the sample standard deviation covers the actual range that includes 95% of the actual population; 95% of all such intervals will cover 95% of the actual population range.

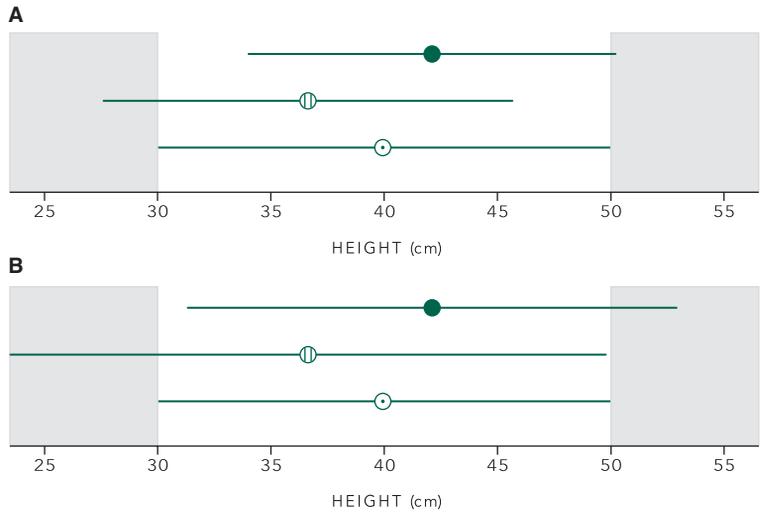
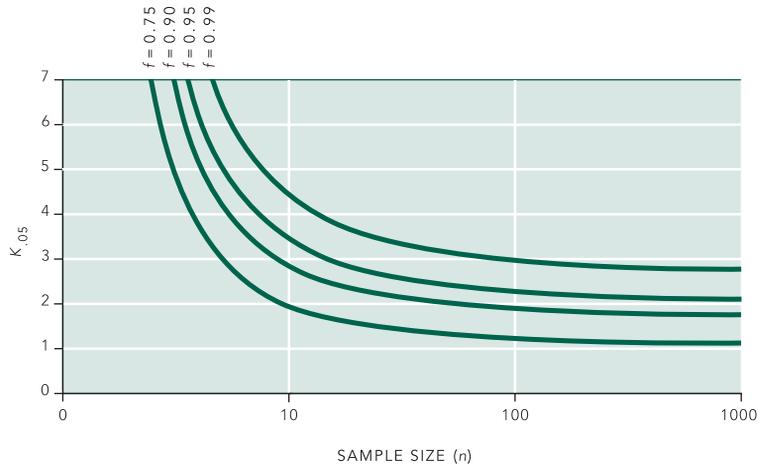


FIGURE 7-7. $K_{.05}$ depends on the size of the sample n used to estimate the mean and standard deviation and the fraction f of the population you want the interval to include.



Specifying the confidence interval for the entire population is more involved than specifying the confidence intervals we have discussed so far because you must specify both the *fraction of the population* if you wish the interval to cover and the *confidence you wish to have that any given interval will cover it*. The size of the interval depends on these two things and the size of the sample used to estimate the mean and standard deviation. The $100(1 - \alpha)$ percent confidence interval for $100f$ percent of the population is

$$\bar{X} - K_{\alpha} s < X < \bar{X} + K_{\alpha} s$$

in which \bar{X} and s are the sample mean and standard deviation and K_{α} is the number of sample standard deviations

about the sample mean needed to cover the desired part of the population. Figure 7-7 shows $K_{.05}$ as a function of sample size for various values of f . It plays a role similar to t_{α} or z_{α} .

K_{α} is larger than t_{α} (which is larger than z_{α}) because it accounts for uncertainty in the estimates of both the mean and standard deviation, rather than the mean alone.*

Notice that K_{α} can be much larger than 2 for sample sizes in the range of 5 to 25, which are common in

*For a derivation of K_{α} that clearly shows how it is related to the confidence limits for the mean and standard deviation, see Lewis AE. Tolerance limits and indices of discrimination. *Biostatistics*. New York: Reinhold; 1966:chap 12.

biomedical research. Thus, simply taking 2 standard deviations about the mean may substantially underestimate the range of the population from which the samples were drawn. Figure 7-6B shows the 95% confidence interval for 95% of the population of Martians' heights based on the three samples of 10 Martians each shown in Figure 2-8. All three of the intervals include 95% of the population.

As Chapter 2 discussed, many people confuse the standard error of the mean with the standard deviation and consider the range defined by “sample mean ± 2 standard errors of the mean” to encompass about 95% of the population. This error leads them to seriously underestimate the possible range of values in the population from which the sample was drawn. We have seen that, for the relatively small sample sizes common in biomedical research, applying the 2 standard deviations rule may underestimate the range of values in the underlying population as well.

■ PROBLEMS

- 7-1** Find the 90% and 95% confidence intervals for the mean levels of polychlorinated biphenyl (PCB) levels in Problem 2-3.
- 7-2** Find the 95% confidence interval for the difference in mean adenosine triphosphate (ATP) production per gram in the two groups of children in Problem 3-1. Based on this confidence interval is the difference significant with $P < .05$?
- 7-3** Find the 95% confidence intervals for the proportions of adverse outcomes as well as the 95% confidence interval for the difference in rates of adverse outcomes in Problem 5-1. Compare this result with the hypothesis test in Problem 5-1.
- 7-4** Find the 95% confidence intervals for the mean difference in the six minute walk distances for the two test groups in Problem 4-3. Compare this result with the hypothesis test in Problem 4-3.
- 7-5** Find the 95% confidence intervals for the percentages of articles with favorable results in the two classes of studies in Problem 5-6.
- 7-6** Use the data in Problem 2-4 to find the 95% confidence interval for 90% and 95% of the population of PCB concentrations in Japanese adults. Plot these intervals together with the observations.
- 7-7** Rework Problem 5-5 using confidence intervals.
- 7-8** Rework Problem 5-11 using confidence intervals.
- 7-9** Rework Problem 5-12 using confidence intervals.
- 7-10** Rework Problem 5-13 using confidence intervals.
- 7-11** Rework Problem 5-14 using confidence intervals.

How to Test for Trends

The first statistical problem we posed in this book, in connection with Figure 1-2A, dealt with a drug that was thought to be a diuretic, but that experiment cannot be analyzed using our existing procedures. In it, we selected different people and gave them different doses of the diuretic, then measured their urine output. The people who received larger doses produced more urine. The statistical question is whether the resulting pattern of points relating urine production to drug dose provided sufficient evidence to conclude that the drug increased urine production in proportion to drug dose. This chapter develops the tools for analyzing such experiments. We will estimate how much one variable increases (or decreases) on the average as another variable changes with a *regression line* and quantifies the *strength* of the association with a *correlation coefficient*.*

■ MORE ABOUT THE MARTIANS

As in all other statistical procedures, we want to use a sample drawn at random from a population to make statements about the population. Chapters 3 and 4 discussed populations whose members are normally distributed with mean μ and standard deviation σ and used estimates of these

parameters to design test statistics (such as F and t) that permitted us to examine whether or not some *discrete* treatment was likely to have affected the mean value of a variable of interest. Now, we add another parametric procedure, *linear regression*, to analyze experiments in which the samples were drawn from populations characterized by a mean response that varies *continuously* with the size of the treatment. To understand the nature of this population and the associated random samples, we return again to Mars, where we can examine the entire population of 200 Martians.

Figure 2-1 showed that the heights of Martians are normally distributed with a mean of 40 cm and a standard deviation of 5 cm. In addition to measuring the heights of each Martian, let us also weigh each one. Figure 8-1 shows a plot in which each point represents the height x and weight y of one Martian. Since we have observed the *entire population*, there is no question that tall Martians tend to be heavier than short Martians.

There are a number of things we can conclude about the heights and weights of Martians as well as the relationship between these two variables. As noted in Chapter 2, the heights are normally distributed with mean $\mu = 40$ cm and standard deviation $\sigma = 5$ cm. The weights are also normally distributed with mean $\mu = 12$ g and standard deviation $\sigma = 2.5$ g. The most striking feature of Figure 8-1, however, is that the *mean weight of Martians at each height* increases as height increases.

For example, the Martians who are 32 cm tall weigh 7.1, 7.9, 8.3, and 8.8 g, so the mean weight of Martians

*Simple linear regression is a special case of the more general method of *multiple regression* in which case there are multiple independent variables. For a discussion of multiple regression and related procedures written in the same style as this book, see Glantz SA, Slinker BK. *Primer of Applied Regression and Analysis of Variance*, 2nd ed. New York: McGraw-Hill; 2001.

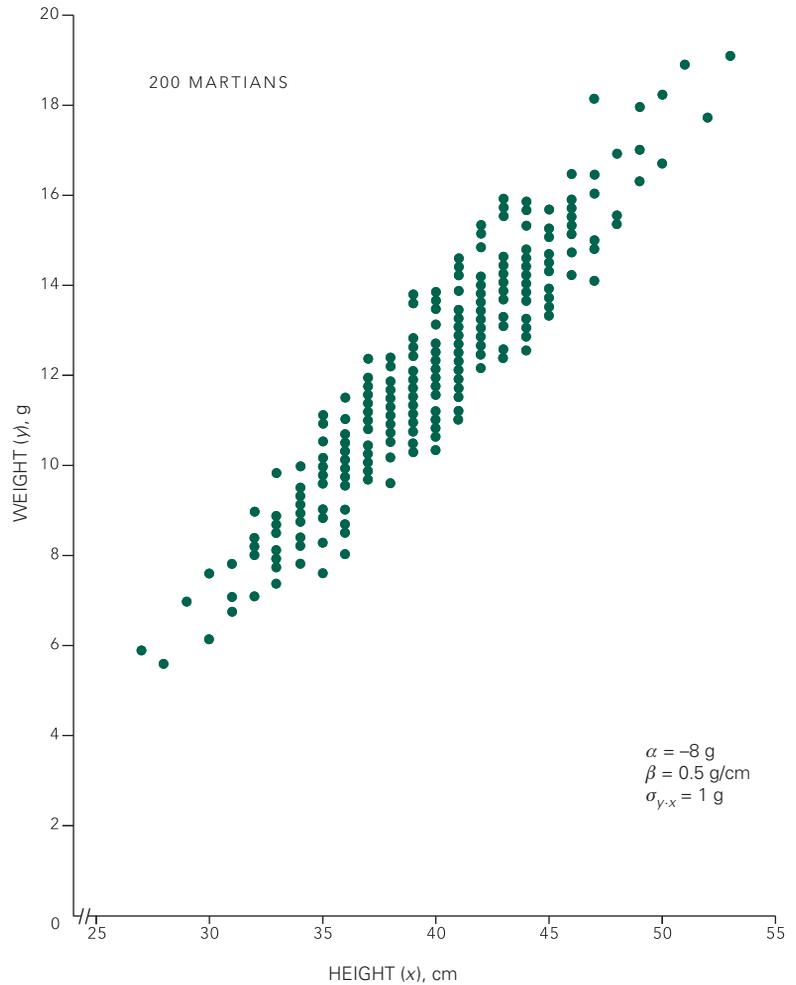


FIGURE 8-1. The relationship between height and weight in the population of 200 Martians, with each Martian represented by a circle. The weights at any given height follow a normal distribution. In addition, the mean weight of Martians at any given height increases linearly with height, and the variability in weight at any given height is the same regardless of height. A population must have these characteristics to be suitable for linear regression or correlation analysis.

who are 32 cm tall is 8 g. The eight Martians who are 46 cm tall weigh 13.7, 14.5, 14.8, 15.0, 15.1, 15.2, 15.3, and 15.8 g, so the mean weight of Martians who are 46 cm tall is 15 g. Figure 8-2 shows that the mean weight of Martians at each height increases *linearly* as height increases.

This line does not make it possible, however, to predict the weight of *an individual* Martian if you know his or her height. Why not? There is variability in weights among Martians at each height. Figure 8-1 reveals that standard deviation of weights of Martians with *any given height* is about 1 g. We need to distinguish this standard deviation from the standard deviation of weights of *all* Martians computed without regard for the fact that mean weight varies with height.

The Population Parameters

Now, let us define some new terms and symbols so that we can generalize from Martians to other populations with similar characteristics. Since we are considering how weight varies with height, call height the *independent variable* x and weight the *dependent variable* y . In some instances including the example at hand, we can only *observe* the independent variable and use it to *predict* the expected mean value of the dependent variable. (There is variability in the dependent variable at each value of the independent variable). In other cases, including controlled experiments, it is possible to *manipulate* the independent variable to control, with some uncertainty, the value of the dependent variable. In the first case, it is only possible to identify an *association* between the two

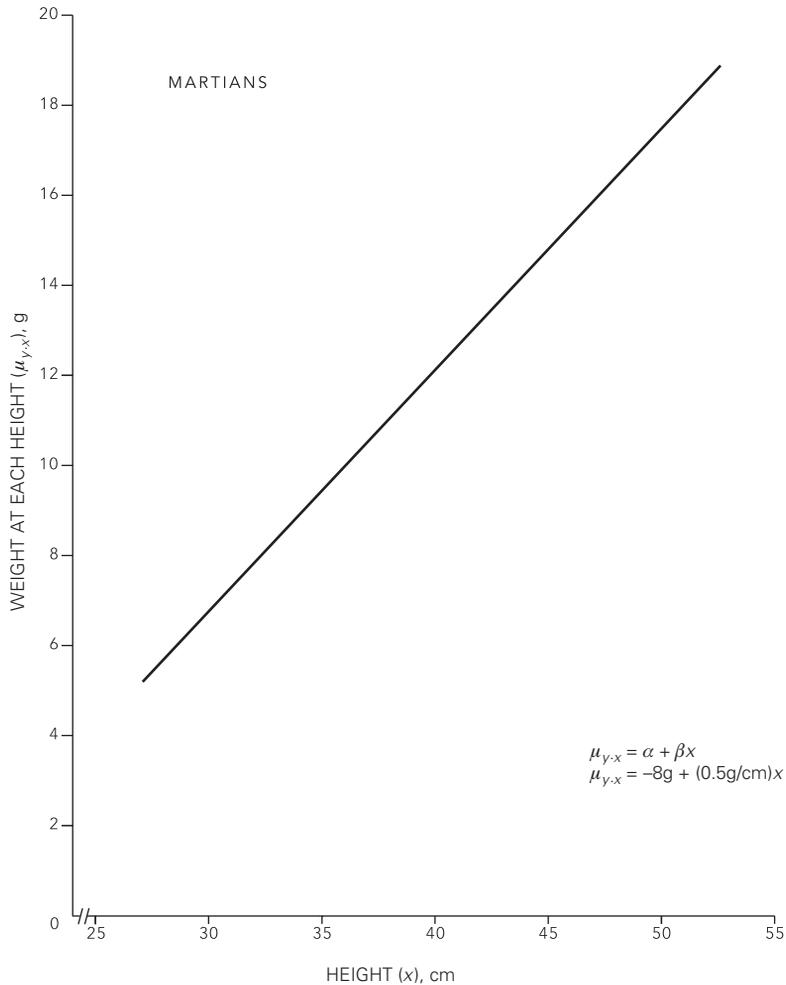


FIGURE 8-2. The line of means for the population of Martians in Figure 8-1.

variables, whereas in the second case it is possible to conclude that there is a *causal* link.*

*In an observational study, statistical analysis alone only permits identification of an association. In order to identify a causal relationship, one generally requires independent evidence to explain the biological (or other) mechanisms that give rise to the observed association. For example, the fact that several epidemiological studies demonstrated an association between passive smoking and heart disease combined with laboratory studies showing short-term effects of secondhand smoke and secondhand smoke constituents on the heart, led to the conclusion that passive smoking *causes* heart disease. For details on how a variety of such evidence is combined to use observational studies as *part* of the case for a causal relationship, see Glantz SA, Parmley WW. Passive smoking and heart disease: epidemiology, physiology, and biochemistry. *Circulation*. 1991;83:1–12 and Barnoya J, Glantz S. Cardiovascular effects of secondhand smoke: nearly as large as smoking. *Circulation* 2005;111:2684–2698.

For any given value of the independent variable x , it is possible to compute the value of the mean of all values of the dependent variable corresponding to that value of x . We denote this mean $\mu_{y,x}$ to indicate that it is the mean of all the values of y in the population at a given value of x . These means fall along a straight line given by

$$\mu_{y,x} = \alpha + \beta x$$

in which α is the intercept and β is the slope[†] of the *line of means*. For example, Figure 8-2 shows that, on the average,

[†]It is, unfortunately, statistical convention to use α and β in this way even though the same two Greek letters also denote the size of the Type I and Type II errors in hypothesis testing. The meaning of α should be clear from the context. β always refers to the slope of the line of means in this chapter.

the average weight of Martians increases by 0.5 g for every 1 cm increase in height, so the slope β of the $\mu_{y|x}$ versus x line is 0.5 g/cm. The intercept α of this line is -8 g. Hence,

$$\mu_{y|x} = -8 \text{ g} + (0.5 \text{ g/cm})x$$

There is variability about the line of means. For any given value of the independent variable x , the values of y for the population are normally distributed with mean $\mu_{y|x}$ and standard deviation $\sigma_{y|x}$. This notation indicates that $\sigma_{y|x}$ is the standard deviation of weights (y) computed after allowing for the fact that mean weight varies with height (x). As noted above, the residual variation about the line of means for our Martians is 1 g; $\sigma_{y|x} = 1$ g. The amount of this variability is an important factor in determining how useful the

line of means is for predicting the value of the dependent variable, for example, weight, when you know the value of the independent variable, for example, height. The methods we develop below require that this standard deviation be *the same* for all values of x . In other words, the variability of the dependent variable about the line of means is the same regardless of the value of the independent variable.

In summary, we will be analyzing the results of experiments in which the observations were drawn from populations with these characteristics:

- The mean of the population of the dependent variable at a given value of the independent variable increases (or decreases) linearly as the independent variable increases.

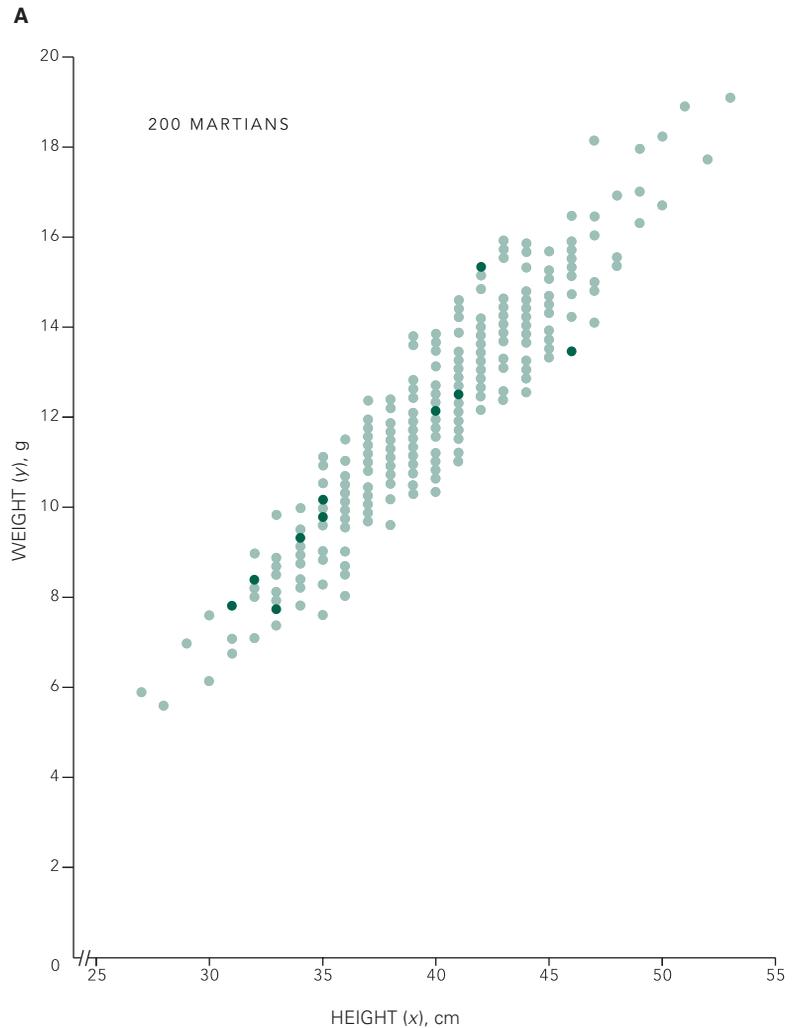


FIGURE 8-3. A random sample of 10 Martians, showing **(A)** the members of the population that were selected together with **(B)** the sample as it appears to the investigator. (continued)

- For any given value of the independent variable, the possible values of the dependent variable are distributed normally.
- The standard deviation of population of the dependent variable about its mean at any given value of the independent variable is the same for all values of the independent variable.

The parameters of this population are α and β , which define the line of means, the dependent-variable population mean at each value of the independent variable, and $\sigma_{y,x}$, which defines the variability about the line of means.

Now let us turn our attention to the problem of estimating these parameters from samples drawn at random from such populations.

■ HOW TO ESTIMATE THE TREND FROM A SAMPLE

Since we observed the entire population of Mars there was no uncertainty how weight varied with height. This situation contrasts with real problems, in which we cannot observe all members of a population and must infer things about it from a limited sample which we hope is representative. To understand the information that such samples contain, let us consider a sample of 10 individuals selected at random from the population of 200 Martians. Figure 8-3A shows the members of the population that happened to be selected; Figure 8-3B shows what an investigator or reader would see. What do the data in Figure 8-3B allow you to say about the underlying

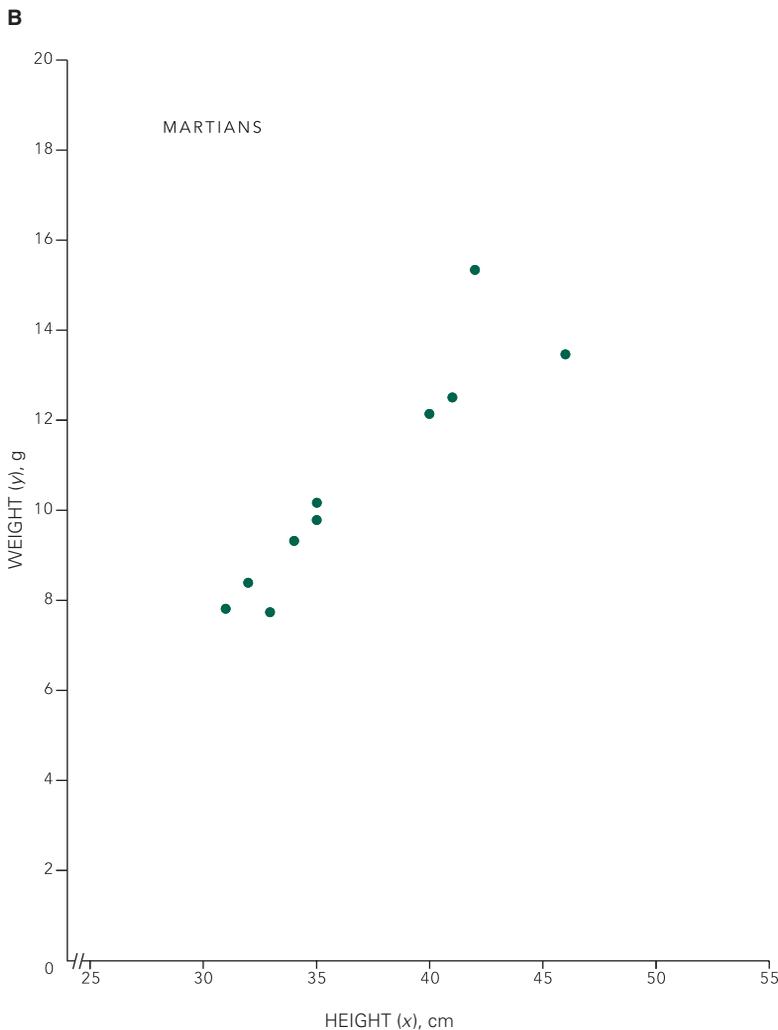


FIGURE 8-3. (Continued)

population? How certain can you be about the resulting statements?

Simply looking at Figure 8-3B reveals that weight increases as height increases among the 10 specific individuals in *this* sample. The real question of interest, however, is: Does weight vary with height in the population the sample came from? After all, there is always a chance that we could draw an unrepresentative sample, just as in Figure 1-2. Before we can test the hypothesis that the apparent trend in the data is due to chance rather than a true trend in the population, we need to estimate the population trend from the sample. This task boils down to estimating the intercept α and slope β of the line of means.

The Best Straight Line through the Data

We will estimate the two population parameters α and β with the intercept and slope, a and b , of a straight line placed through the sample points. Figure 8-4 shows the same sample as Figure 8-3B with four proposed lines, labeled I, II, III and IV. Line I is obviously not appropriate; it does not even pass through the data. Line II passes through the data but has a much steeper slope than the data suggest is really the case. Lines III and IV seem more reasonable; they both pass along the cloud defined by the data points. Which one is best?

To select the best line and so get our estimates a and b of α and β , we need to define precisely what “best” means. To arrive at such a definition, first think about why line II seems

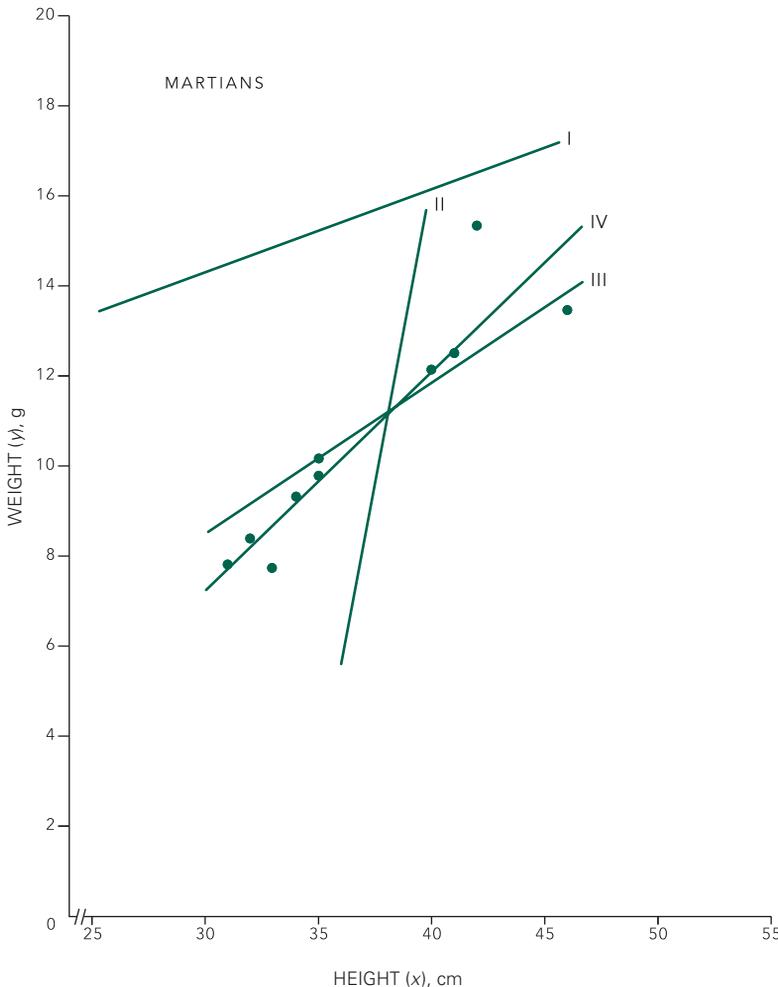


FIGURE 8-4. Four different possible lines to estimate the line of means from the sample in Figure 8-3. Lines I and II are unlikely candidates because they fall so far from most of the observations. Lines III and IV are more promising.

■ TABLE 8-1. Computation of Regression Line in Figure 8-5B

Observed Height X (cm)	Observed Weight Y (g)	X ² (cm ²)	XY (g · cm)
31	7.8	961	241.8
32	8.3	1,024	265.6
33	7.6	1,089	250.8
34	9.1	1,156	309.4
35	9.6	1,225	336.0
35	9.8	1,225	343.0
40	11.8	1,600	472.0
41	12.1	1,681	496.1
42	14.7	1,764	617.4
46	13.0	2,116	598.0
<u>369</u>	<u>103.8</u>	<u>13,841</u>	<u>3,930.1</u>

better than line I and line III seems better than line II. The “better” a straight line is, the closer it comes to all the points taken as a group. In other words, we want to select the line that minimizes the total variability between the data and the line. The farther any one point is from the line, the more the line varies from the data, so let us select the line that leads to the smallest total variability between the observed values and the values predicted from the straight line.

The problem becomes one of defining a measure of variability then selecting values of a and b to minimize this quantity. Recall that we quantified variability in a population with the variance (or standard deviation) by computing the sum of the squared deviations from the mean and then divided by the sample size, n , minus 1. Now we will use the same idea and use *sum of the squared differences between the observed values of the dependent variable and the value on the line at the same value of the independent variable* as our measure of how much any given line varies from the data. We square the deviations so that positive and negative deviations contribute equally. Figure 8-5 shows the deviations associated with lines III and IV in Figure 8-4. The sum of squared deviations is smaller for line IV than line III, so it is the better line. In fact, it is possible to prove mathematically that line IV is the one with the smallest sum of squared deviations between the observations and the line,* making it the “best” line. For this reason, this procedure is often called the *method of least squares* or *least-squares regression*.

*For this proof and a derivation of the formulas for the slope and intercept of this line, see Glantz SA, Slinker BK. *Primer of Applied Regression and Analysis of Variance*, 2 ed. New York: McGraw-Hill; 2001, 19.

The resulting line is called the *regression line* of y on x (in this case, the regression line of weight on height). Its equation is

$$\hat{y} = a + bx$$

\hat{y} denotes the value of y on the regression for a given value of x . This notation distinguishes it from the observed value of the dependent variable Y . The intercept a is given by

$$a = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{n(\sum X^2) - (\sum X)^2}$$

and the slope is given by

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}$$

in which X and Y are the coordinates of the n points in the sample.[†]

Table 8-1 shows these computations for the sample of 10 points in Figure 8-3B. From this table, $n = 10$, $\sum X = 369$ cm, $\sum Y = 103.8$ g, $\sum X^2 = 13,841$ cm², and $\sum XY = 3930.1$ g · cm. Substitute these values into the equations for the intercept and slope of the regression line to find

$$\begin{aligned} a &= \frac{(103.8 \text{ g})(13,841 \text{ cm}^2) - (369 \text{ cm})(3930.1 \text{ g} \cdot \text{cm})}{10(13,841 \text{ cm}^2) - (369 \text{ cm})^2} \\ &= -6.0 \text{ g} \end{aligned}$$

[†]The calculations can be simplified by computing b first, then finding a from $a = \bar{Y} - b\bar{X}$, in which \bar{X} and \bar{Y} are the means of all observations of the independent and dependent variables, respectively.

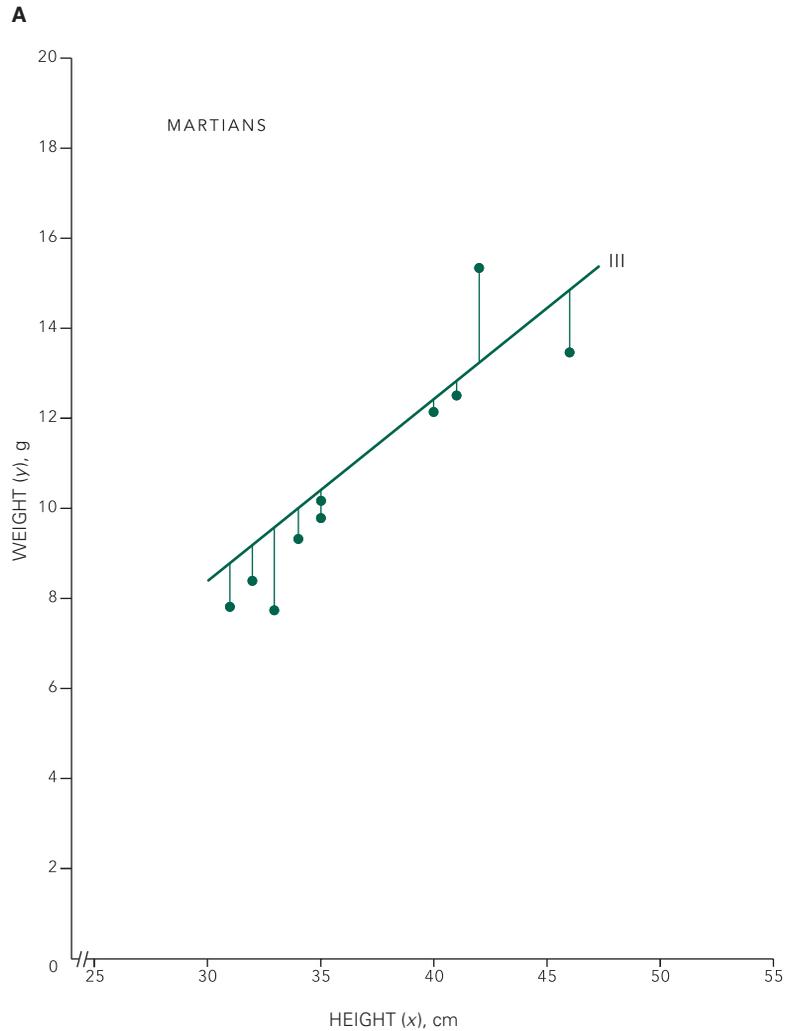


FIGURE 8-5. Lines III (A) and IV (B) in Figure 8-4, together with the deviations between the lines and the observations. Line IV is associated with the smallest sum of squared deviations between the regression line and the observed values of the dependent variable. The vertical lines indicate the deviations. The black line is the line of means for the population of Martians in Figure 8-1. The regression line approximates the line of means but does not precisely coincide with it. Line III is associated with larger deviations than line IV.

and

$$b = \frac{10(3930.1 \text{ g} \cdot \text{cm}) - (369 \text{ cm})(103.8 \text{ g})}{10(13,841 \text{ cm}^2) - (369 \text{ cm})^2} = 0.44 \text{ g/cm}$$

Line IV in Figures 8-4 and 8-5B is this regression line.

$$\hat{y} = -6.0 \text{ g} + (0.44 \text{ g/cm})x$$

These two values are estimates of the population parameters, $\alpha = -8 \text{ g}$ and $\beta = 0.5 \text{ g/cm}$, the intercept and slope of the line of means. The light line in Figure 8-5B shows the line of means.

Variability about the Regression Line

We have the regression line to estimate the line of means, but we still need to estimate the variability of population

members about the line of means, $\sigma_{y \cdot x}$. We estimate this parameter by computing the square root of the “average” squared deviation of the data about the regression line

$$s_{y \cdot x} = \sqrt{\frac{\sum(Y - \hat{y})^2}{n - 2}} = \sqrt{\frac{\sum[Y - (a + bX)]^2}{n - 2}}$$

where $a + bX$ is the value \hat{y} on the regression line corresponding to the observation at X ; Y is the actual observed value of y ; $Y - (a + bX)$ is the amount that the observation deviates about the regression line and Σ denotes the sum, over all the data points, of the squares of these deviations $[Y - (a + bX)]^2$. We divide by $n - 2$ rather than n for reasons analogous to dividing by $n - 1$ when computing the sample standard deviation as an estimate of the population standard deviation. Since the sample will

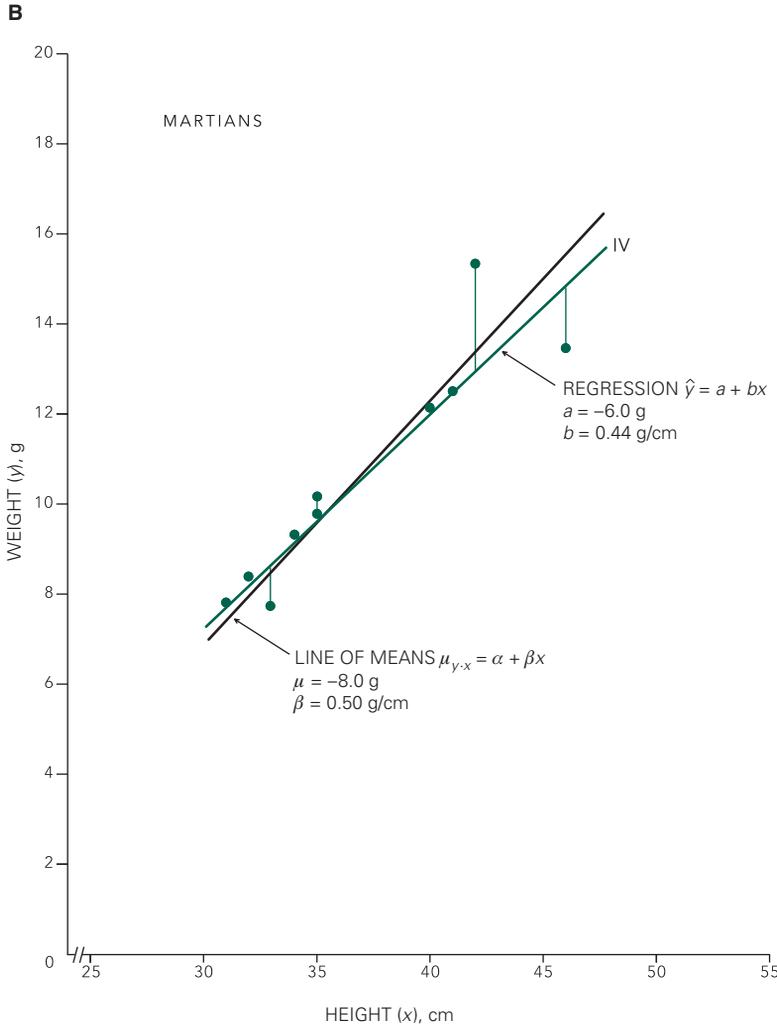


FIGURE 8-5. (Continued)

not show as much variability as the population, we need to decrease the denominator when computing the “average” squared deviation from the line to compensate for this tendency to underestimate the population variability.

$s_{y,x}$ is called the *standard error of the estimate*. It is related to the standard deviations of the dependent and independent variables and the slope of the regression line according to

$$s_{y,x} = \sqrt{\frac{n-1}{n-2}(s_Y^2 - b^2 s_X^2)}$$

where s_Y and s_X are the standard deviations of the dependent and independent variables, respectively.

For the sample shown in Figure 8-3B (and Table 8-1), $s_X = 5.0$ cm and $s_Y = 2.4$ g, so

$$s_{y,x} = \sqrt{\frac{9}{8}[2.4^2 - 0.44^2(5.0^2)]} = 1.02 \text{ g}$$

This number is an estimate of the actual variability about the line of means, $\sigma_{y,x} = 1$ g.

Standard Errors of the Regression Coefficients

Just as the sample mean is only an estimate of the true population mean, the slope and intercept of the regression line are only estimates of the slope and intercept of the line of means in the population. In addition, just as different samples yield different estimates for the population

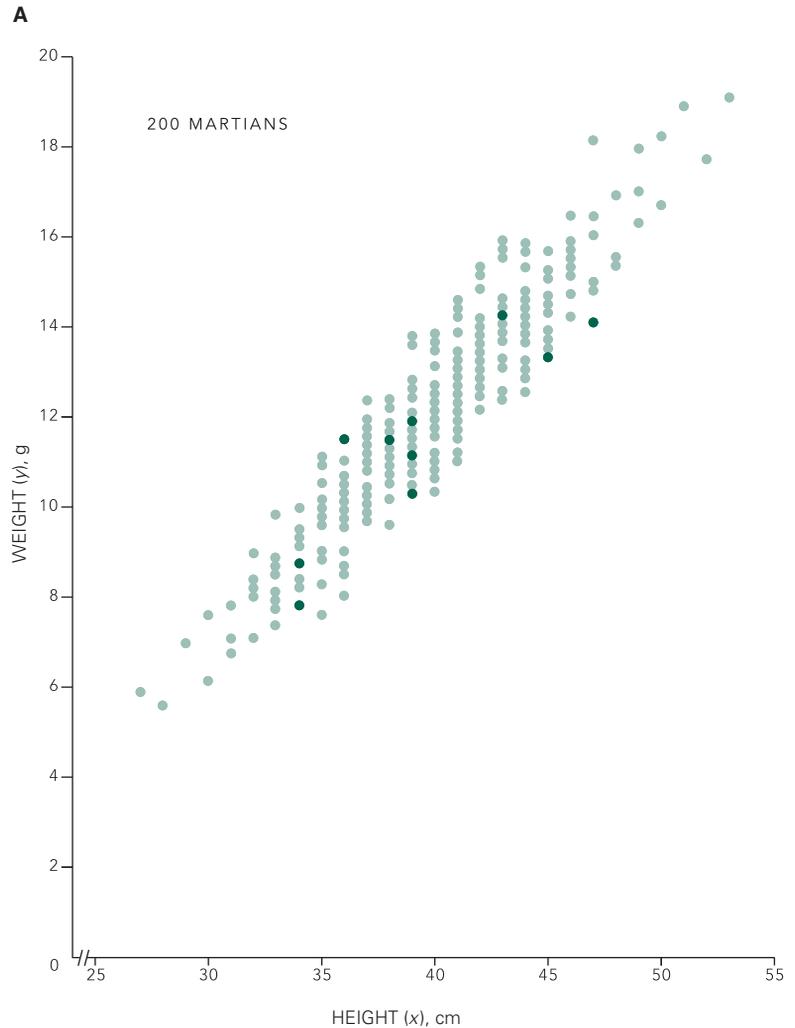


FIGURE 8-6. This figure illustrates a second random sample of 10 Martians drawn from the population in Figure 8-1. This sample is associated with a different regression line than that computed from the first sample, shown in Figure 8-5A.

mean, different samples will yield different regression lines. After all, there is nothing special about the sample in Figure 8-3. Figure 8-6A shows another sample of 10 individuals drawn at random from the population of all Martians. Figure 8-6B shows what you would see. Like the sample in Figure 8-3B, the results of this sample also suggest that taller Martians tend to be heavier, but the relationship looks a little different from that associated with our first sample. This sample yields estimates of $a = -4.0$ g and $b = 0.38$ g/cm as estimates of the intercept and slope of the line of means.

There is a population of possible values of a and b corresponding to all possible samples of a given size drawn from the population in Figure 8-1. These distributions of all possible values of a and b have means α and β , respectively, and standard deviations σ_a and σ_b , called the *standard error of the intercept* and *standard error of the slope*, respectively.

These standard errors can be used just as we used the standard error of the mean and standard error of a proportion. Specifically, we will use them to test hypotheses about, and compute confidence intervals for, the regression coefficients and the regression equation itself.

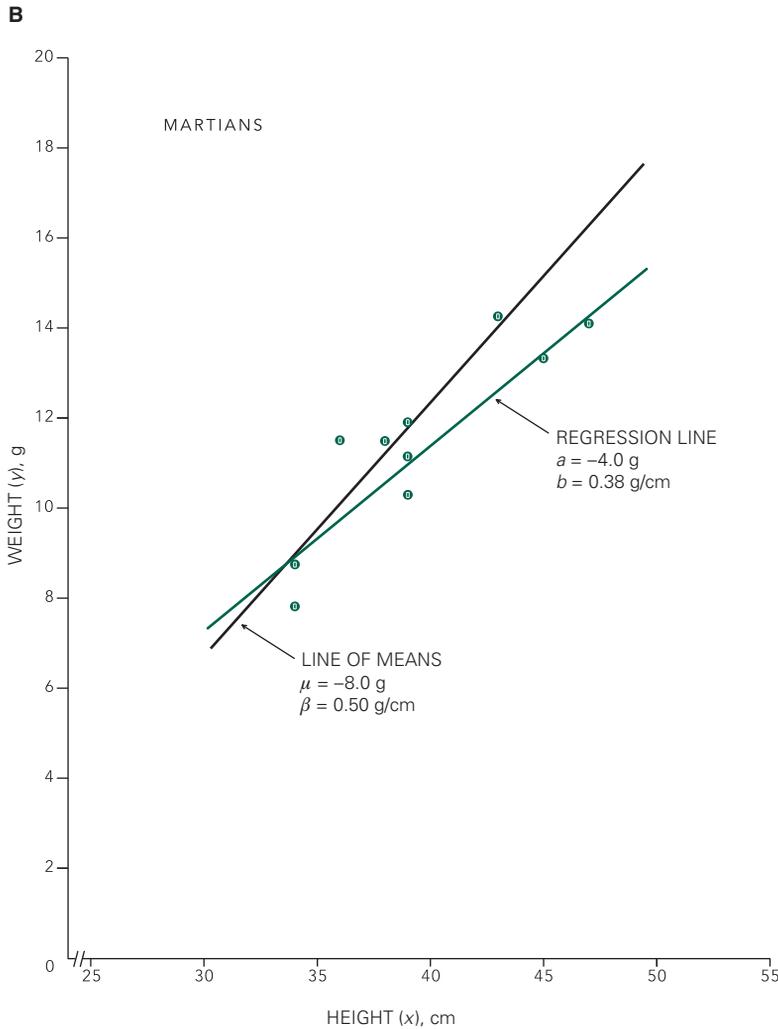


FIGURE 8-6. (Continued)

The standard deviation of the population of all possible values of the regression line intercept, the *standard error of the intercept*, can be estimated from the sample with*

$$s_a = s_{y \cdot x} \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}}$$

*For a derivation of these formulas, see Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. Inferences in regression analysis. *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs*. Boston: WCB McGraw-Hill; 1996:chap 2.

The *standard error of the slope* of the regression line is the standard deviation of the population of all possible slopes. Its estimate is

$$s_b = \frac{1}{\sqrt{n-1}} \frac{s_{y \cdot x}}{s_X}$$

From the data in Figure 8-3B and Table 8-1 it is possible to compute the standard errors for the slope and intercept as

$$s_a = (1.02 \text{ g}) \sqrt{\frac{1}{10} + \frac{(36.9 \text{ cm})^2}{(10-1)(5.0 \text{ cm})^2}} = 2.6 \text{ g}$$

and

$$s_b = \frac{1}{\sqrt{10-1}} \frac{1.02 \text{ g}}{5.0 \text{ cm}} = .064 \text{ g/cm}$$

Like the sample mean, both a and b are computed from sums of the observations. Like the distributions of all possible values of the sample mean, the distributions of all possible values of a and b tend to be normally distributed. (This result is another consequence of the central-limit theorem.) The specific values of a and b associated with the regression line are then randomly selected from normally distributed populations. Therefore, these standard errors can be used to compute confidence intervals and test hypotheses about the intercept and slope of the line of means using the t distribution, just as we did for the sample mean in Chapter 7.

How Convincing Is the Trend?

There are many hypotheses we can test about regression lines, but the most common and important one is that the slope of the line of means is zero. This hypothesis is equivalent to estimating the chance that we would observe a trend as strong or stronger than the data show *when there is actually no relationship* between the dependent and independent variables. The resulting P value quantifies the certainty with which you can reject the hypothesis that there is no *linear* trend relating the two variables.*

Since the population of possible values of the regression slope is approximately normally distributed, we can use the general definition of the t statistic

$$t = \frac{\text{Parameter estimate} - \text{true value of population parameter}}{\text{Standard error of parameter estimate}}$$

to test this hypothesis. The equivalent mathematical statement is

$$t = \frac{b - \beta}{s_b}$$

This equation permits testing the hypothesis that there is no trend in the population from which the sample was

drawn, that is, $\beta = 0$, using either of the approaches to hypothesis testing developed earlier.

To take a classic hypothesis-testing approach (as in Chapter 4), set β to zero in the equation above and compute

$$t = \frac{b}{s_b}$$

then compare the resulting value of t with the critical value t_α defining the 100α percent most extreme values of t that would occur if the hypothesis of no trend in the population was true.

For example, the data in Figure 8-3B (and Table 8-1) yielded $b = 0.44 \text{ g/cm}$ and $s_b = 0.064 \text{ g/cm}$ from a sample of 10 points. Hence, $t = 0.44/0.064 = 6.875$, which exceeds 5.041, the value of t for $P < .001$ with $\nu = 10 - 2 = 8$ degrees of freedom (from Table 4-1). Hence, it is unlikely that this sample was drawn from a population in which there was no relationship between the independent and dependent variables, that is, height and weight. We can use these data to assert that as height increases, weight increases ($P < .001$).

Of course, like all statistical tests of hypotheses, this small P value does not guarantee that there is really a trend in the population, it just means it is unlikely that there is not such a trend. For example, the sample in Figure 1-2A is associated with $P < .0005$. Nevertheless, as Figure 1-2B shows, there is no trend in the underlying population.

If we wish to test the hypothesis that there is no trend in the population using confidence intervals, we use the definition of t above to find the $100(1 - \alpha)$ percent confidence interval for the slope of the line of means,

$$b - t_\alpha s_b < \beta < b + t_\alpha s_b$$

We can compute the 95% confidence interval for β by substituting the value of $t_{.05}$ with $\nu = n - 2 = 10 - 2 = 8$ degrees of freedom, 2.306, into this equation together with the observed values of b and s_b

$$.44 - 2.306 \cdot .064 < \beta < .44 + 2.306 \cdot .064 \\ .29 \text{ g/cm} < \beta < .59 \text{ gm/cm}$$

Since this interval does not contain zero, we can conclude that there is a trend in the population ($P < .05$).[†]

*This restriction is important. As discussed later in this chapter, it is possible for there to be a strong *nonlinear* relationship in the observations and for the procedures we discuss here to miss it.

[†]The 99.9% confidence interval does not contain zero either, so we could obtain the same P value (.001) as with the first method using confidence intervals as with $t = b/s_b$ earlier in this session.

Note that the interval contains the true value of the slope of the line of means, $\beta = 0.5$ g/cm.

It is likewise possible to test hypotheses about, or compute confidence intervals for, the intercept using the fact that

$$t = \frac{a - \alpha}{s_a}$$

is distributed according to the t distribution with $\nu = n - 2$ degrees of freedom. For example, since $s_a = 2.6$ g the 95% confidence interval for the intercept based on the observations in Figure 8-3B is

$$\begin{aligned} a - t_{.05} s_a < \alpha < a + t_{.05} s_a \\ -6.0 - 2.306 \cdot 2.6 < \alpha < -6.0 + 2.306 \cdot 2.6 \\ -12.0 \text{ g} < \alpha < 5.4 \text{ g} \end{aligned}$$

which includes the true intercept of the line of means, $\alpha = -8$ g.

A number of other useful confidence intervals associated with regression analysis, such as the confidence interval for the line of means, will be discussed next.

Confidence Interval for the Line of Means

There is uncertainty in the estimates of the slope and intercept of the regression line. The standard errors of the slope and the intercept, s_a and s_b , quantify this uncertainty. These standard errors are $s_a = 2.6$ g and $s_b = 0.06$ g/cm for the regression of height or weight for the Martians in the sample in Figure 8-3. Thus, the line of means could lie slightly above or below the observed regression line or have a slightly different slope. It nevertheless is likely that the line of means lies within a band surrounding the observed regression line. Figure 8-7A shows this region. It is wider at the ends than in the middle because the regression line must be straight and must go through the point defined by the means of the independent and dependent variables.

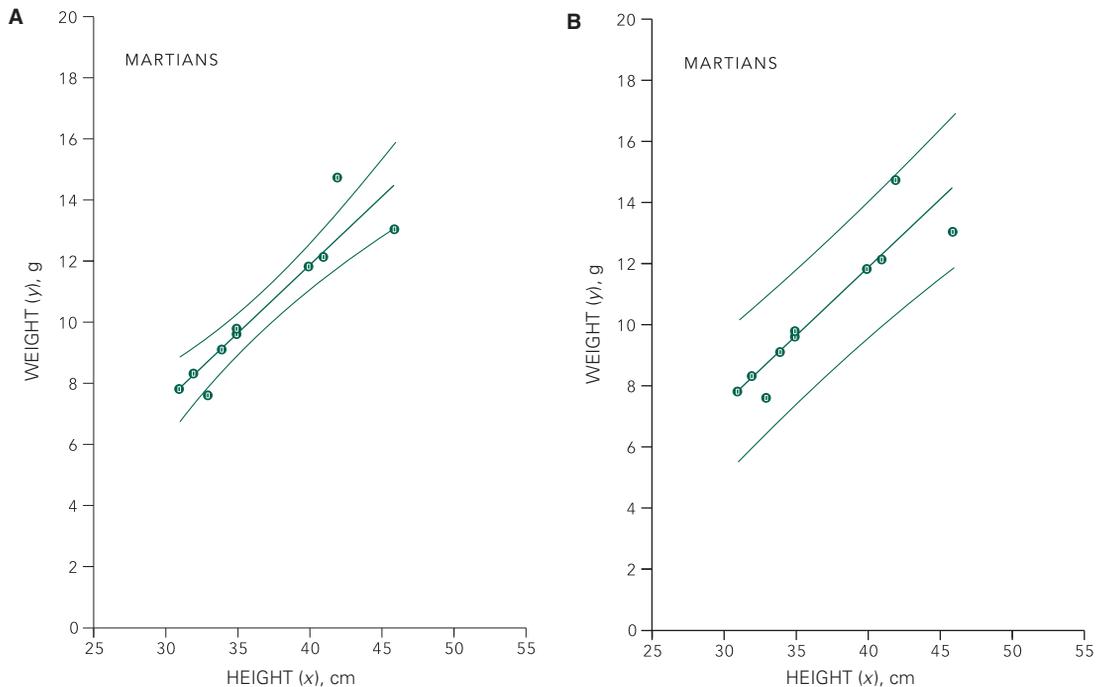


FIGURE 8-7. (A) The 95% confidence interval for the regression line relating Martian weight or height using the data in Figure 8-3. **(B)** The 95% confidence interval for an additional observation of Martian weight at a given height. This is the confidence interval that should be used to estimate true weight from height to be 95% confident that the range includes the true weight.

There is a distribution of possible values for the regression line at each value of the independent variable x . Since these possible values are normally distributed about the line of means, it makes sense to talk about the standard error of the regression line. (This is another consequence of the central-limit theorem.) Unlike the other standard errors we have discussed so far, this standard error is not constant but depends on the value of the independent variable x :

$$s_{\hat{y}} = s_{y \cdot x} \sqrt{\frac{1}{n} + \frac{(x - \bar{X})^2}{(n-1)s_x^2}}$$

Since the distribution of possible values of the regression line is normally distributed, we can compute the $100(1 - \alpha)$ percent confidence interval for the regression line with

$$\hat{y} - t_{\alpha} s_{\hat{y}} < y < \hat{y} + t_{\alpha} s_{\hat{y}}$$

in which t_{α} has $\nu = n - 2$ degrees of freedom and \hat{y} is the point on the regression line for each value of x ,

$$\hat{y} = a + bx$$

Figure 8-7A shows the 95% confidence interval for the line of means. It is wider at the ends than the middle, as it should be. Note also that it is much narrower than the range of the data because it is the confidence interval for the line of means, not the population as a whole.

It is not uncommon for investigators to present the confidence interval for the regression line and discuss it as though it were the confidence interval for the population. This practice is analogous to reporting the standard error of the mean instead of the standard deviation to describe population variability. For example, Figure 8-7A shows that we can be 95% confident that the mean weight of all 40 cm tall Martians is between 11.0 and 12.5 g. We cannot be 95% confident that the weight of any one Martian that is 40 cm tall falls in this narrow range.

Confidence Interval for an Observation

To compute a confidence interval for an individual observation, we must combine the total variability that arises from the variation in the underlying population about the line of means, estimated with, $s_{y \cdot x}$, and the variability due to uncertainty in the location of the line of means $s_{\hat{y}}$. Since the variance of a sum is the sum of the variances,

the standard deviation of the predicted value of the observation will be

$$s_{Y_{\text{new}}} = \sqrt{s_{y \cdot x}^2 + s_{\hat{y}}^2}$$

We can eliminate $s_{\hat{y}}$ from this equation by replacing it with the equation for $s_{\hat{y}}$ in the last section

$$s_{Y_{\text{new}}} = s_{y \cdot x} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{(n-1)s_x^2}}$$

This standard error can be used to define the $100(1 - \alpha)$ percent confidence interval for an observation according to

$$\hat{y} - t_{\alpha} s_{Y_{\text{new}}} < y < \hat{y} + t_{\alpha} s_{Y_{\text{new}}}$$

Remember that both \hat{y} and $s_{Y_{\text{new}}}$ depend on the value of the independent variable x .

The two curved lines around the regression line in Figure 8-7B show the 95% confidence interval for an additional observation. This band includes both the uncertainty due to random variation in the population and variation due to uncertainty in the estimate of the true line of means. Notice that most members of the sample fall in this band. It quantifies the uncertainty in using Martian height to estimate weight, and hence, the uncertainty in the true weight of a Martian of a given height. For example, it shows that we can be 95% confident that the true weight of a 40 cm tall Martian is between 9.5 and 14.0 g. This confidence interval describes the precision with which it is possible to estimate the true weight. This information is much more useful than the fact that there is a statistically significant* relationship between the Martian weight and height ($P < .001$).

Cell Phone Radiation, Reactive Oxygen Species, and DNA Damage in Human Sperm

Motivated by the human and animal studies showing that cell phone use was associated with lower sperm motility, Geoffry De Luliis and colleagues[†] conducted an experiment

* $t = b/s_b = .44/.064 = 6.875$ for the data in Figure 8-3. $t_{.001}$ for $\nu = 10 - 2 = 8$ degrees of freedom is 5.041.

[†]De Luliis GN, Newey RJ, King BV, Aitken RJ. Mobil phone radiation induces reactive oxygen species production and DNA damage in human spermatozoa *in vitro*. *PLoS One*. 2010;4(7):e6446. doi:10.1371/journal.pone.0006446.

in which they exposed normal human sperm dishes to cell phone electromagnetic signals and measured the production of intracellular reactive oxygen species (ROS) produced in cellular mitochondria that can damage DNA as well as the amount of DNA damage.

They exposed the sperm (obtained from students with no known reproductive problems or infections) to cell phone signals of varying strengths for 16 hours and investigated the relationship between the strength of the signal and the level of ROS production and DNA damage. The sperm were exposed in petri dishes maintained at a constant 21°C temperature to avoid the problem that the higher radiofrequency radiation from stronger cell phone signals would heat the sperm more, which would affect sperm function. By holding temperature constant, Di Luliis and colleagues avoided the effects of this potential confounding variable.

They sought to investigate whether there was a dose-dependent effect of cell phone exposure on the amount of ROS produced by the sperm and the level of DNA damage to the sperm.

The independent variable in their study was cell phone signal strength, measured at the specific absorption rate (SAR), for the cell phone. (SAR is the rate of absorption of an electromagnetic radiation from a cell phone by

a model designed to simulate a human head.) The dependent variables were the fraction of sperm that tested positive on a MitroSOX red test for ROS. A second question is whether there is a relationship between the level of ROS and a second dependent variable, the fraction of sperm that expressed 8-hydroxy-2'-deoxyguanosine (8-OH-dg), a marker for oxidative damage to sperm DNA.

Table 8-2 shows the data for this study.

Figure 8-8 shows the relationship between the percentage of sperm that tested positive for ROS, R , as a function of the SAR, S , together with the results of doing a linear regression on these data. Even though the slope is significantly different from zero ($P < .001$), the regression line does not provide an accurate description of the data, which shows a rapid increase in ROS generation at low SAR levels, then flattens out. *This example illustrates the importance of always looking at a plot of the data together with the associated regression line to ensure that the central assumptions of linear regression — that the line of means is a straight line and that the residuals are randomly and normally distributed around the regression line — are met.* In this case, neither assumption is satisfied so we cannot use linear regression to make statements about the relationship

■ **TABLE 8-2. Cell Phone Signal Strength and Fraction of Sperm with Reactive Oxygen Species and DNA Damage**

Cell-Phone-Specific Absorption Rate, SAR (W/kg)	Sperm with Mitochondrial ROS (%)	Sperm with DNA Damage (%)
0.4	8	5
27.5	29	18
0.0	6	3
1.0	13	8
2.8	16	10
10.1	27	15
2.8	18	5
27.5	30	13
10.1	25	15
4.3	25	7
4.3	23	8
1.0	15	4
1.0	11	3

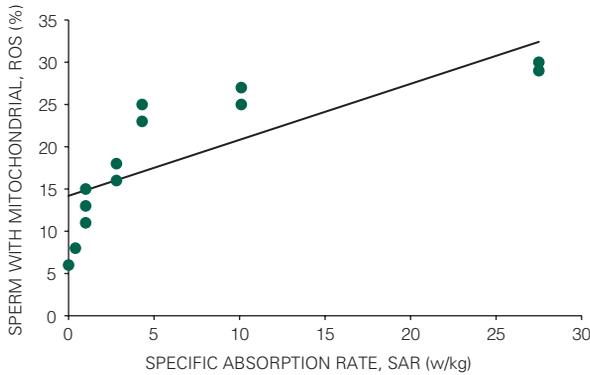


FIGURE 8-8. The fraction of sperm with positive tests for mitochondrial reactive oxygen species increases with the intensity of the electromagnetic radiation produced by the cell phone, but this increase is not linear, so linear regression cannot be used to test hypotheses about the relationship between these two variables.

apparent in these data.* Therefore, while this figure seems to indicate a strong relationship between the strength of the cell phone signal and the level of oxidative damage, we cannot make any statistical statements about the confidence we have in making such a statement using linear regression.

We are luckier about the relationship between the level of ROS and DNA damage (Fig. 8-9) where the assumptions of linear regression are satisfied. (Compare how well the regression line goes through the data compared with how it does not in Fig. 8-8.) Box 8-1 shows that the regression line has a slope of .505 with a

*Examining the relationship between ROS formation damage and SAR suggests a *saturation* exponential,

$$R = R_{\infty} \left(1 - e^{-\frac{S}{s}} \right)$$

where the two parameters are R_{∞} , the maximum fraction of sperm with testing positive for ROS, and s , the exponential rate at which ROS increases. Such a relationship would occur if the rate of sperm becoming ROS positive depends on the fraction of sperm that are not yet ROS positive. It is possible to fit a *nonlinear* equation to these data. See Glantz S, Slinker B. *Nonlinear regression. Primer of Applied Regression and Analysis of Variance*, 2nd ed. New York: McGraw-Hill; 2001:chap 11.

standard error of .105 and an intercept of $-.796\%$ with a standard error of 2.27%. We test whether the slope and intercept are significantly different from zero by computing $t = b/s_b = .505/.105 = 4.810$ and $t = a/s_a = -.796/2.27 = .351$, respectively. We compare these values of t associated with the regression with the critical value of t that defines the 95% most extreme values of the t distribution with $\nu = n - 2 = 13 - 2 = 11$ degrees of freedom (from Table 4-2), 2.201. Since the t for the slope exceeds this value, we reject the null hypothesis of no (linear) relationship between sperm ROS level and DNA damage and conclude that increased levels of ROS are associated with higher levels of DNA damage. (In fact, the value of t associated with the slope exceeds 4.437, the critical value for $P < .001$.) In contrast, the t for the intercept does not even approach the critical value, so we do not reject the null hypothesis that the intercept is zero. The overall conclusion from these two tests is that the fraction of sperm with DNA damage increases in proportion to the fraction of sperm with elevated levels of ROS. Therefore, based on this experiment, we can conclude with a high level of confidence that higher levels of ROS production in sperm mitochondria cause DNA damage ($P < .001$).

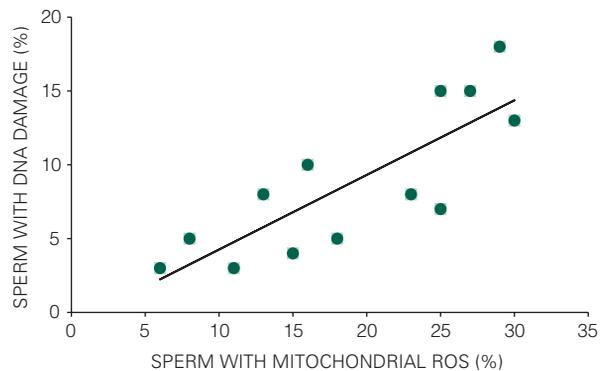


FIGURE 8-9. Sperm with higher levels of mitochondrial reactive oxygen species have higher levels of DNA damage. In contrast to the results in Figure 8-8, the data are consistent with the assumptions of linear regression, so we can use linear regression to draw conclusions about this relationship.

BOX 8-1 • Linear Regression of Fraction of Sperm with ROS as a Function of Cell Phone SAR

The first two columns of the table below present the data from Table 8-2 (last two columns), together with the square of the independent variable values and the product of the independent and dependent variables and the sums necessary to compute the linear regression.

Calculation of Linear Regression of DNA Damage

Sperm with Mitochondrial ROS (%)	Sperm with DNA Damage (%)			Fit Regression Line	Residual	Residual ²
X	Y	X^2	XY	\hat{y}	$(Y - \hat{y})$	$(Y - \hat{y})^2$
8	5	64	40	3.25	1.75	3.07
29	18	841	522	13.86	4.14	17.12
6	3	36	18	2.24	0.76	0.58
13	8	169	104	5.78	2.22	4.95
16	10	256	160	7.29	2.71	7.33
27	15	729	405	12.85	2.15	4.61
18	5	324	90	8.30	-3.30	10.91
30	13	900	390	14.37	-1.37	1.87
25	15	625	375	11.84	3.16	9.98
25	7	625	175	11.84	-4.84	23.43
23	8	529	184	10.83	-2.83	8.01
15	4	225	60	6.79	-2.79	7.76
11	3	121	33	4.76	-1.76	3.11
<u>246</u>	<u>114</u>	<u>5444</u>	<u>2556</u>			<u>102.75</u>

$n = 13$; $\bar{X} = 246/13 = 18.9\%$; $s_x = 8.11\%$.

The intercept of the regression line is

$$a = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{n(\sum X^2) - (\sum X)^2} = \frac{114 \cdot 5444 - 246 \cdot 2556}{13 \cdot 5444 - 246^2} = -.796\%$$

and the slope is

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2} = \frac{13 \cdot 2556 - 246 \cdot 114}{13 \cdot 5444 - 246^2} = .505$$

Thus, the regression equation is

$$\hat{y}(x) = -.796\% + .518 x$$

We use the regression equation to compute the predicted value of y for each observed X ; for example, for the first observation, $X = 8$,

$$\hat{y}(8) = -.796\% + .505 \cdot 8 = 3.25$$

and the associated residual is

$$(Y - \hat{y}) = 5.00 - 3.25 = 1.75$$

(continued)

BOX 8-1 • Linear Regression of Fraction of Sperm with ROS as a Function of Cell Phone SAR (Continued)

Thus, the standard error of the estimate is

$$s_{y \cdot x} = \sqrt{\frac{\sum(Y - \hat{y})^2}{n - 2}} = \sqrt{\frac{102.75}{13 - 2}} = 3.06\%$$

The standard deviation of the observed values of the independent value, s_x , is 8.11%, so the standard error of the intercept and slope are

$$s_a = \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_x^2}} = \sqrt{\frac{1}{13} + \frac{18.9^2}{(13-1)8.11^2}} = 2.27\%$$

and

$$s_b = \frac{1}{\sqrt{n-1}} \frac{s_{y \cdot x}}{s_x} = \frac{1}{\sqrt{13-1}} \frac{3.06}{11.8} = .105$$

■ HOW TO COMPARE TWO REGRESSION LINES*

The situation often arises in which one wants to compare two regression lines. There are actually three possible comparisons one might want to make:

- Test for a difference in slope (without regard for the intercepts).
- Test for a difference in intercept (without regard for the slopes).
- Make an overall test of coincidence, in which we ask if the lines are different.

The procedures for comparing two slopes or intercepts are a direct extension of the fact that the observed slopes and intercepts follow the t distribution. For example, to test the hypothesis that two samples were drawn from populations with the same slope of the line of means, we compute

$$t = \frac{\text{Difference of regression slopes}}{\text{Standard error of difference of regression slopes}}$$

*This section deals with more advanced material and can be skipped without loss of continuity. It is also possible to test for differences between three or more regression lines using techniques which are generalizations of regression and analysis of variance; see Zar JH. Comparing simple linear regression equations. *Biostatistical Analysis*, 4th ed. Upper Saddle River, NJ: Prentice-Hall; 1999:chapter 18. For a discussion of how to use multiple regression models to compare several regression lines, including how to test for parallel shifts between regression lines, see Glantz S, Slinker B. Regression with two or more independent variables. *Primer of Applied Regression and Analysis of Variance*, 2nd ed. New York: McGraw-Hill; 2001:chap 3.

or, in mathematical terms,

$$t = \frac{b_1 - b_2}{s_{b_1 - b_2}}$$

where the subscripts 1 and 2 refer to data from the first and second regression data samples. This value of t is compared to the critical value of the t distribution with $\nu = n_1 + n_2 - 4$ degrees of freedom. This test is exactly analogous to the definition of the t test to compare two sample means.

If the two regressions are based on the same number of data points, the standard error of the difference of two regression slopes is

$$s_{b_1 - b_2} = \sqrt{s_{b_1}^2 + s_{b_2}^2}$$

If there are a different number of points, use the pooled estimate of the difference of the slopes. Analogous to the pooled estimate of the variance in the t test in Chapter 4, compute a pooled estimate of the variation about the regression lines as

$$s_{y \cdot x_p}^2 = \frac{(n_1 - 2)s_{y \cdot x_1}^2 + (n_2 - 2)s_{y \cdot x_2}^2}{n_1 + n_2 - 4}$$

and use this value to compute

$$s_{b_1 - b_2} = \sqrt{\frac{s_{y \cdot x_p}^2}{(n_1 - 1)s_{x_1}^2} + \frac{s_{y \cdot x_p}^2}{(n_2 - 1)s_{x_2}^2}}$$

Likewise, to compare the intercepts of two regression lines, we compute

$$t = \frac{a_1 - a_2}{s_{a_1 - a_2}}$$

where

$$s_{a_1 - a_2} = \sqrt{s_{a_1}^2 + s_{a_2}^2}$$

if there are the same number of points for each regression equation, and we use a formula based on the pooled variance estimate if there are unequal number of points in the two regressions.

Overall Test for Coincidence of Two Regression Lines

It is also possible to test the null hypothesis that two regressions are *coincident*, that is, have the same slope and intercept. Recall that we computed the slope and intercept of the regression line by selecting the values that minimized the total sum of squared differences between the observed values of the dependent variable and the value on the line at the same value of the independent variable (residuals). The square of the standard error of the estimate, $s_{y \cdot x}$ is the estimate of this residual variance around the regression line and it is a measure of how closely the regression line fits the data. We will use this fact to construct our test by examining whether fitting the two sets of data with separate regression lines (in which the slopes and intercepts can be different) produces smaller residuals than fitting all the data with a single regression line (with a single slope and intercept).

The specific procedure for testing for coincidence of two regression lines is

- Fit each set of data with a separate regression line.
- Compute the pooled estimate of the variance around the two regression lines, $s_{y \cdot x_p}^2$, using the previous equations. This statistic is a measure of the overall variability about the two regression lines, allowing the slopes and intercepts of the two lines to be different.
- Fit all the data with one regression line, and compute the variance around this one “single” regression line, $s_{y \cdot x_s}^2$. This statistic is a measure of the overall variability observed when the data are fit by assuming that they all fall along one line of means.
- Compute the “improvement” in the fit obtained by fitting the two data sets with separate regression lines

compared to fitting them with a single regression line using

$$s_{y \cdot x_{\text{imp}}}^2 = \frac{(n_1 + n_2 - 2)s_{y \cdot x_s}^2 - (n_1 + n_2 - 4)s_{y \cdot x_p}^2}{2}$$

- The numerator in this expression is the reduction in the total sum of squared differences between the observations and regression line that occurs when the two lines are allowed to have different slopes and intercepts. It can also be computed as

$$s_{y \cdot x_{\text{imp}}}^2 = \frac{SS_{\text{res}_x} - SS_{\text{res}_p}}{2}$$

where SS_{res} are the sum of squared residuals about the regressions.

- Compute the ratio of the improvement in the fit obtained when fitting the two sets of data separately over fitting all the data with a single line with the residual variation about the regression lines when fitting the two lines separately, using the F -test statistic,

$$F = \frac{s_{y \cdot x_{\text{imp}}}^2}{s_{y \cdot x_p}^2}$$

- Compare the observed value of the F -test statistic with the critical values of F for $v_n = 2$ numerator degrees of freedom and $v_d = n_1 + n_2 - 4$ denominator degrees of freedom.

If the observed value of F exceeds the critical value of F , it means that we obtain a significantly better fit to the data (measured by the residual variation about the regression line) by fitting the two sets of data with separate regression lines than we do by fitting all the data with a single line. We reject the null hypothesis of a single line of means and conclude that the two sets of data were drawn from populations with different lines of means.

Relationship between Weakness and Muscle Wasting in Rheumatoid Arthritis

Rheumatoid arthritis is a disease in which a person's joints become inflamed so that movement becomes painful, and people find it harder to complete mechanical tasks, such as holding things. At the same time, as people age, they often lose muscle mass. As a result, P. S. Helliwell and S. Jackson*

*Helliwell PS, Jackson S. Relationship between weakness and muscle wasting in rheumatoid arthritis. *Ann Rheum Dis*. 1994;53:726–728.

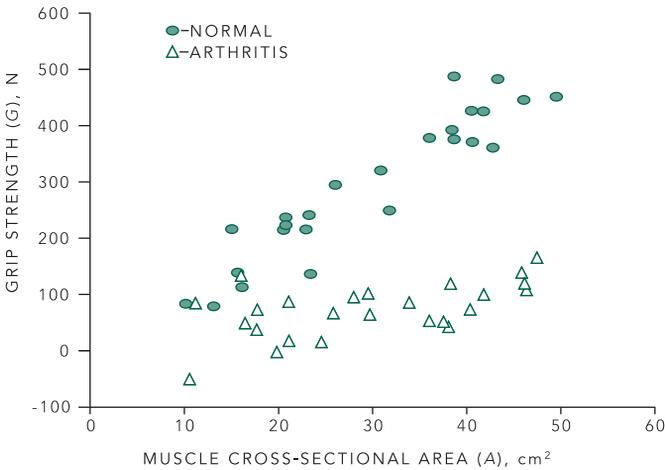


FIGURE 8-10. This plot shows the grip strength as a function of muscle cross-sectional area in 25 normal people and 25 people with arthritis. The question is: Are the relationships between these two variables the same in both groups of people?

wondered whether the reduction in grip strength noted in people who had arthritis was due to the arthritic joints or simply a reflection of a reduction in mass of muscle.

To investigate this question, they measured the cross-sectional area (in cm^2) of the forearms of a group of normal people and a group of similar people with arthritis as well as the force (in newtons) with which they could grip a test device. Figure 8-10 shows the data from such an experiment, using different symbols with the two groups of people indicated. The question is: Is the relationship between muscle cross-sectional area and grip strength different for the normal people (circles) and the people with arthritis (triangles)?

We will answer this question by first doing a test for overall coincidence of the two regressions. Figure 8-11A shows the same data as in Figure 8-10, with separate regression equations fit to the two sets of data and

Table 8-3 presents the results of fitting these two regression equations. Using the formula presented earlier, the pooled estimate of the variance about the two regression lines fit separately is

$$s_{\text{grip-area}_p}^2 = \frac{(n_{\text{normal}} - 2)s_{\text{grip-area}_{\text{normal}}}^2 + (n_{\text{arthritis}} - 2)s_{\text{grip-area}_{\text{arthritis}}}^2}{n_{\text{normal}} + n_{\text{arthritis}} - 4} = \frac{(25 - 2)45.7^2 + (25 - 2)40.5^2}{25 + 25 - 4} = 1864 \text{ N}^2$$

Next, fit all the data to a single regression equation, without regard for the group to which each person belongs; Figure 8-11B shows this result, with the results of fitting the single regression equation as the last column in Table 8-3. The total variance of the observations about the single regression line is $s_{\text{grip-area}_s}^2 = (129.1)^2 = 16,667 \text{ N}^2$.

TABLE 8-3. Comparison of the Relationship between Grip Strength and Muscle Cross-Sectional Area in Normal People and People with Arthritis (See Figs. 8-8 and 8-9)

	Normal	Arthritis	All People
Sample size n	25	25	50
Intercept a (s_a), N	-7.3 (25.3)	3.3 (22.4)	-23.1 (50.5)
Slope b (s_b), N/cm^2	10.19 (.789)	2.41 (.702)	6.39 (1.579)
Standard error of the estimate $s_{y,x}$, N	45.7	40.5	129.1

This value is larger than that observed when the two curves were fit separately. To estimate the improvement (reduction) in variance associated with fitting the two lines separately, we compute

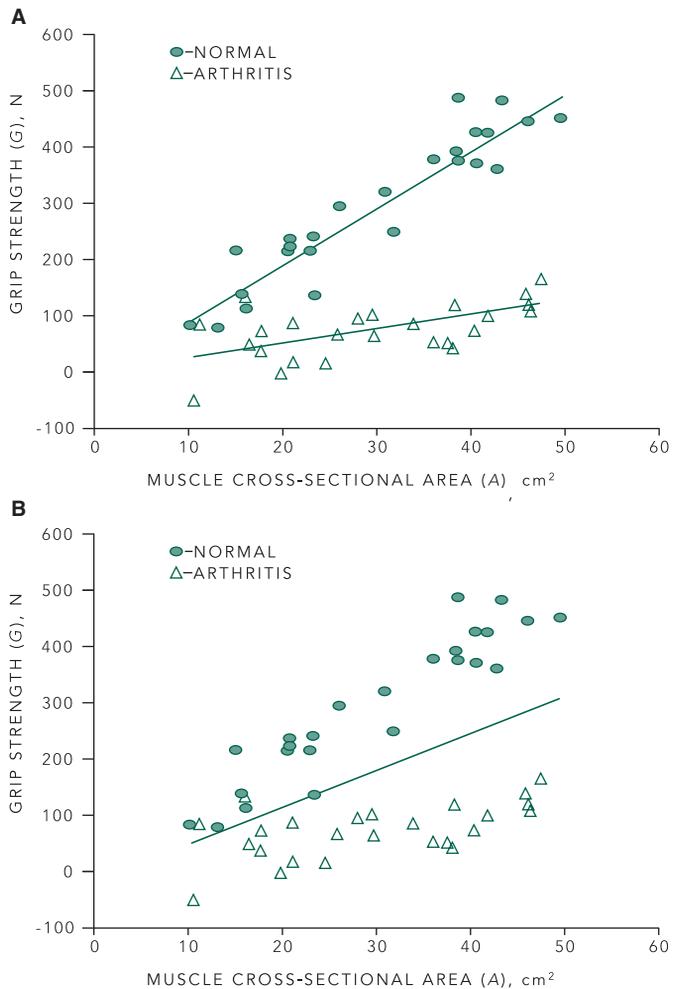
$$\begin{aligned}
 s_{\text{grip-area}_{\text{imp}}}^2 &= \frac{(n_{\text{normal}} + n_{\text{arthritis}} - 2)s_{\text{grip-area}_s}^2 - (n_{\text{normal}} + n_{\text{arthritis}} - 4)s_{\text{grip-area}_p}^2}{2} \\
 &= \frac{(25 + 25 - 2)16,667 - (25 + 25 - 4)1864}{2} \\
 &= 357,136 N^2
 \end{aligned}$$

Finally, we compare the improvement in the variance about the regression line obtained by fitting the two groups separately with that obtained by fitting them together (which yields the smallest residual variance) with the *F* test

$$F = \frac{s_{\text{grip-area}_{\text{imp}}}^2}{s_{\text{grip-area}_p}^2} = \frac{375,136}{1864} = 191.597$$

This value exceeds 5.10, the critical value of *F* for *P* < .01 with *v*_d = 2 and *v*_d = *n*_{normal} + *n*_{arthritis} - 4 = 25 + 25 - 4 = 46 degrees of freedom, so we conclude that the relationship between grip force and cross-sectional area is different for normal people and people with arthritis.

FIGURE 8-11. In order to test whether the two groups of people (normal subjects and people with arthritis) have a similar relationship between muscle cross-sectional area and grip strength, we first fit the data for the two groups separately (**A**), then together (**B**). If the null hypothesis that there is no difference between the two groups is true, then the variation about the regression lines fit separately will be approximately the same as the variation when the two sets of data are fit separately.



The next question that arises is where the difference comes from. Are the intercepts or slopes different? To answer this question, we compare the intercepts and slopes of the two regression equations. We begin with the intercepts. Since the two regressions are based on the same number of data points, we can use the results in Table 8-3 to compute the standard error, the difference in the two regression intercepts with

$$\begin{aligned} s_{a_{\text{normal}}-a_{\text{arthritis}}} &= \sqrt{s_{a_{\text{normal}}}^2 + s_{a_{\text{arthritis}}}^2} = \sqrt{25.3^2 + 22.4^2} \\ &= 33.8 N \end{aligned}$$

and

$$t = \frac{a_{\text{normal}} - a_{\text{arthritis}}}{s_{a_{\text{normal}}-a_{\text{arthritis}}}} = \frac{(-7.3) - (3.3)}{33.8} = -.314$$

which does not come near exceeding 2.013 in magnitude, the critical value of t for $P < .05$ for $v = n_{\text{normal}} + n_{\text{arthritis}} - 4 = 46$ degrees of freedom. Therefore, we do not conclude that the intercepts of the two lines are significantly different.

A similar analysis comparing the slopes yields $t = 7.367$, so we do conclude that the slopes are different ($P < .001$). Hence the increase in grip force per unit increase in cross-sectional muscle area is smaller for people with arthritis than normal people.

■ CORRELATION AND CORRELATION COEFFICIENTS

Linear regression analysis of a sample provides an estimate of how, on the average, a dependent variable changes when an independent variable changes and an estimate of the variability in the dependent variable about the line of means. These estimates, together with their standard errors, permit computing confidence intervals to show the certainty with which you can predict the value of the dependent variable for a given value of the independent variable. In some experiments, however, two variables are measured that change together but neither can be considered to be the dependent variable. In such experiments, we abandon all premise of making a statement about causality and simply seek to describe the strength of the association between the two variables. The *correlation coefficient*, a number between -1 and $+1$, is often used to quantify the strength of this association. Figure 8-12 shows that the tighter the relationship between the two variables, the closer the magnitude of r is to 1; the weaker the relationship

between the two variables, the closer r is to 0. We will examine two different correlation coefficients.

The first, called the *Pearson product-moment correlation coefficient*, quantifies the strength of association between two variables that are normally distributed like those in Figure 8-1. It, therefore, provides an alternative perspective on the same data we analyzed using linear regression. When people refer to *the* correlation coefficient, they almost always mean the Pearson product-moment correlation coefficient.

The second, called the *Spearman rank correlation coefficient*, is used to quantify the strength of a trend between two variables that are measured on an *ordinal scale*. In an ordinal scale responses can be graded, but there is no arithmetic relationship between the different possible responses. For example, Pap smears, the common test for cervical cancer, are graded according to this scale: (1) normal, (2) cervicitis (inflammation, usually due to infection), (3) mild-to-moderate dysplasia (abnormal but noncancerous cells), (4) moderate-to-severe dysplasia, and (5) cancerous cells present. In this case, a rating of 4 denotes a more serious condition than a rating of 2, but it is *not* necessarily *twice* as serious. This situation contrasts with observations quantified on an *interval scale* where there are arithmetic relationships between the responses. For example, a Martian who weighs 16 g is twice as heavy as one who weighs 8 g. Ordinal scales often appear in clinical practice when conditions are ranked according to seriousness.

The Pearson Product-Moment Correlation Coefficient

The problem of describing the strength of association between two variables is closely related to the linear regression problem, so why not simply arbitrarily make one variable dependent on the other? Figure 8-13 shows that reversing the roles of the two variables when computing the regression line results in *different* regression lines. This situation arises because in the process of computing the slope and intercept of the regression line we minimize the sum of squared deviations between the regression line and the observed values of the *dependent* variable. If we reverse the roles of the two variables, there is a different dependent variable, so different values of the regression line intercept and slope minimize the sum of squared deviations. We need a measure of association that does not require arbitrarily deciding that one of the variables is the independent variable.

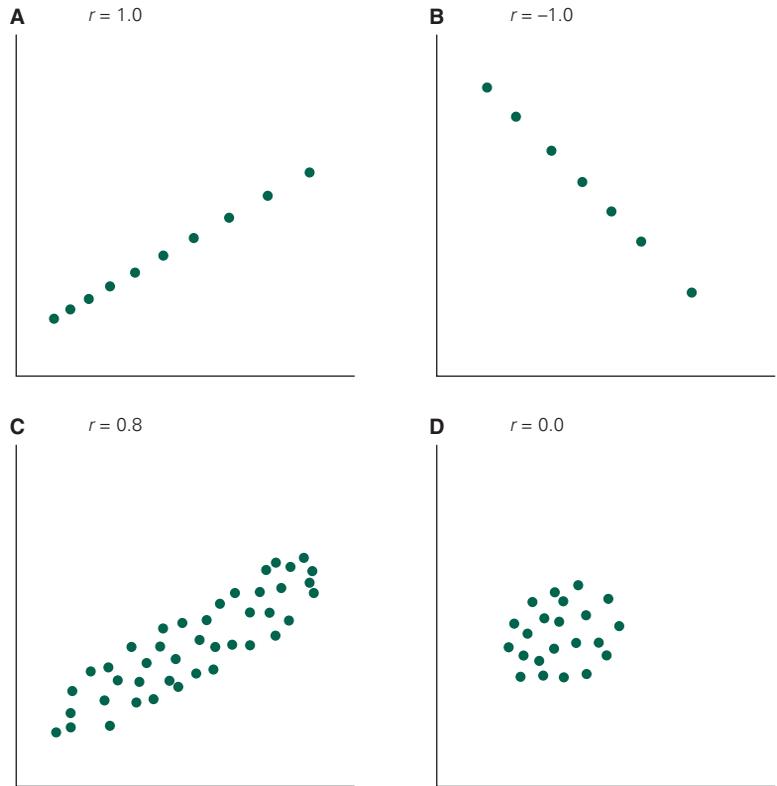


FIGURE 8-12. The closer the magnitude of the correlation coefficient is to 1, the less scatter there is in the relationship between the two variables. The closer the correlation coefficient is to 0, the weaker the relationship between the two variables.

The Pearson product-moment correlation coefficient r , defined by

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

in which the sums are over all the observed (X, Y) points, has this property. Its value does not depend on which variable we call x and y . The magnitude of r describes the *strength of the association* between the two variables, and sign of r tells the direction of this association: $r = +1$ when the two variables increase precisely linearly together (Fig. 8-12A), and $r = -1$ when one decreases linearly as the other increases (Fig. 8-12B). Figure 8-12C also shows the more common case of two variables that are correlated, though not perfectly. Figure 8-12D shows two variables that do not appear to relate to each other at all; $r = 0$.

Table 8-4 illustrates how to compute the correlation coefficient using the sample of 10 points in Figure 8-3B. (These are the same data used to illustrate the computation of the regression line in Table 8-1 and Fig. 8-5B.)

From Table 8-4, $n = 10$, $\bar{X} = \sum X/n = 369/10 \text{ cm} = 36.9 \text{ cm}$, and $\bar{Y} = \sum Y/n = 103.8/10 \text{ g} = 10.38 \text{ g}$, so $\sum(X - \bar{X})(Y - \bar{Y}) = 99.9 \text{ g} \cdot \text{cm}$, $\sum(X - \bar{X})^2 = 224.9 \text{ cm}^2$, and $\sum(Y - \bar{Y})^2 = 51.8 \text{ g}^2$. Substitute these numbers into the definition of the correlation coefficient to obtain

$$r = \frac{99.9 \text{ g} \cdot \text{cm}}{\sqrt{224.9 \text{ cm}^2 \cdot 51.8 \text{ g}^2}} = 0.925$$

To gain more feeling for the meaning to the magnitude of a correlation coefficient, Table 8-5 lists the values of the correlation coefficients for the observations in Figures 8-7 and 8-11.

The Relationship between Regression and Correlation

Obviously, it is possible to compute a correlation coefficient for any data suitable for linear regression analysis. Indeed, the correlation coefficients in Table 8-4 all were computed from the same examples we used to illustrate regression analysis. In the context of regression analysis, it is possible

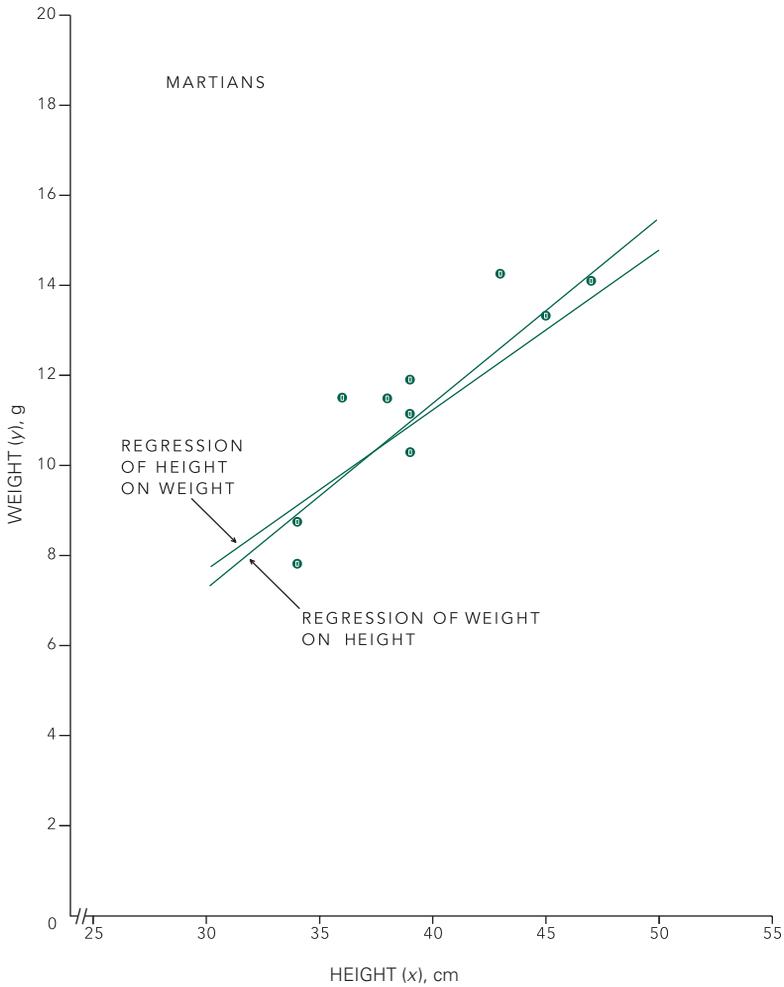


FIGURE 8-13. The regression of y on x yields a different regression line than the regression of x on y for the same data. The correlation coefficient is the same in either case.

TABLE 8-4. Computation of Correlation Coefficient for Sample in Figure 8-3B

Observed Height X (cm)	Observed Weight Y (g)	$(X - \bar{X})$ (cm)	$(Y - \bar{Y})$ (g)	$(X - \bar{X})(Y - \bar{Y})$ (cm · g)	$(X - \bar{X})^2$ (cm ²)	$(Y - \bar{Y})^2$ (g ²)
31	7.8	-5.9	-2.6	15.2	34.8	6.7
32	8.3	-4.9	-2.1	10.2	24.0	4.3
33	7.6	-3.9	-2.8	10.8	15.2	7.7
34	9.1	-2.9	-1.3	3.7	8.4	1.6
35	9.6	-1.9	-0.8	1.5	3.6	0.3
35	9.8	-1.9	-0.6	1.1	3.6	2.0
40	11.8	3.1	1.4	4.4	9.6	2.0
41	12.1	4.1	1.7	7.1	16.8	3.0
42	14.7	5.1	4.3	22.0	26.0	18.7
46	13.0	9.1	2.6	23.8	82.8	6.9
<u>369</u>	<u>103.8</u>	<u>0.0</u>	<u>0.0</u>	<u>99.9</u>	<u>224.9</u>	<u>51.8</u>

■ TABLE 8-5. Correlations between Variables in Examples

Figure	Variables	Correlation Coefficient, r	Sample Size, n
8-7	Height and weight of Martians	.925	10
8-9A	Grip force and muscle cross-sectional area in normal people	.938	25
8-9B	Area in normal and arthritic people	.581	25

to add to the meaning of the correlation coefficient. Recall that we selected the regression equation that minimized the sum of squared deviations between the points on the regression line and the value of the dependent variable at each observed value of the independent variable. It can be shown that the correlation coefficient also equals

$$r = \sqrt{1 - \frac{\text{Sum of squared deviations from regression line}}{\text{Sum of squared deviations from mean}}}$$

where the deviations are all measured for the dependent variable.

Let SS_{res} equal the sum of squared deviations (residuals) from the regression line and SS_{tot} equal the total sum of squared deviations from the mean of the dependent variable. Then

$$r = \sqrt{1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}}$$

When there is no variation in the observations about the regression line $SS_{\text{res}} = 0$, the correlation coefficient equals 1 (or -1), indicating the dependent variable can be predicted with no uncertainty from the independent variable. On the other hand, when the residual variation about the regression line is the same as the variation about the mean value of the dependent variable, $SS_{\text{res}} = SS_{\text{tot}}$, there is no trend in the data and $r = 0$. The dependent variable cannot be predicted at all from the independent variable.

The square of the correlation coefficient, r^2 , is known as the *coefficient of determination*. Since, from the preceding equation,

$$r^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

and SS_{tot} is a measure of the total variation in the dependent variable, people say that the coefficient of determination is the fraction of the total variance in the dependent variable “explained” by the regression equation. This is

rather unfortunate terminology, because the regression line does not “explain” anything in the sense of providing a mechanistic understanding of the relationship between the dependent and independent variables. Nevertheless, the coefficient of determination is a good description of how clearly a straight line describes the relationship between the two variables.

Likewise, the sum of squared deviations from the regression line, SS_{res} , is just $(n-2)s_{y-x}^2$ and the sum of squared deviations about the mean, SS_{tot} , is just $(n-1)s_y^2$. (Recall the definition of sample variance or standard deviation.) Hence the correlation coefficient is related to the results of regression analysis according to

$$r = \sqrt{1 - \frac{n-2}{n-1} \frac{s_{y-x}^2}{s_y^2}}$$

As the standard deviation of the residuals about the regression line s_{y-x} decreases with respect to the total variation in the dependent variable, quantified with s_y , the ratio s_{y-x}/s_y decreases and the correlation coefficient increases. Thus, the greater the value of the correlation coefficient, the more precisely the dependent variable can be predicted from the independent variable.

This approach must be used with caution, however, because the absolute uncertainty as described with the confidence interval is usually more informative in that it allows you to gauge the size of the uncertainty in the prediction relative to the size of effect that is of clinical or scientific importance. As Figure 8-7 showed, it is possible to have correlations well above 0.9 (generally considered quite respectable in biomedical research) and still have substantial uncertainty in the value of an additional observation for a given value of the independent value.

The correlation coefficient is also related to the slope of the regression equation according to

$$r = b \frac{s_X}{s_Y}$$

We can use the following intuitive argument to justify this relationship: When there is no relationship between the two variables under study, both the slope of the regression line and the correlation coefficient are zero.

How to Test Hypotheses about Correlation Coefficients

Earlier in this chapter we tested for a trend by testing the hypothesis that the slope of the line of means was zero using the t test

$$t = \frac{b}{s_b}$$

with $\nu = n - 2$ degrees of freedom. Since we have just noted that the correlation coefficient is zero when the slope of the regression line is zero, we will test the hypothesis that there is no trend relating two variables by testing the hypothesis that the correlation coefficient is zero with the t test

$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}}$$

with $\nu = n - 2$ degrees of freedom. While this statistic looks quite foreign, it is just another way of writing the t statistic used to test the hypothesis that $\beta = 0$.*

*To see this, recall that

$$r = \sqrt{1 - \frac{n-2}{n-1} \frac{s_{y \cdot x}^2}{s_y^2}}$$

so

$$s_{y \cdot x}^2 = \frac{n-1}{n-2} (1-r^2) s_y^2$$

Use this result to eliminate $s_{y \cdot x}$ from

$$s_b = \frac{1}{\sqrt{n-1}} \frac{s_{y \cdot x}}{s_x}$$

to obtain

$$s_b = \frac{s_y}{s_x} \sqrt{\frac{1-r^2}{n-2}}$$

Substitute this result together with $b = r(s_y/s_x)$ into $t = b/s_b$ to obtain the t test for the correlation coefficient

$$t = \frac{r(s_y/s_x)}{(s_y/s_x)\sqrt{(1-r^2)/(n-2)}} = \frac{r}{\sqrt{(1-r^2)/(n-2)}}$$

Journal Size and Selectivity

As part of an assessment of medical journals' policies regarding how the editors handle the peer review of statistical aspects of manuscripts submitted to their journals, Steven Goodman and his colleagues[†] surveyed a sample of medical journals. In addition to asking the editors about their policies on statistical review, Goodman and his colleagues also collected data on the percentage of submitted manuscripts that were ultimately accepted for publication and the size of the journals' circulation. Figure 8-12 shows data relating these two variables, which makes it possible to test whether the larger journals are more selective.

Note that rather than plotting the publication rate against the circulation, Figure 8-12 plots it against the logarithm of the circulation. The reason for this *variable transformation* is to adjust the scale so that the data more closely met the assumptions of correlation, which require that the data be scattered around a straight line. (The Spearman rank correlation coefficient, discussed in the following section, does not require making this assumption.) Variable transformations are a common tool in more advanced statistical methods to account for failures of normality or linearity.[‡] Logarithmic transformations are particularly useful when the observations span several orders of magnitude, as is the case here. This situation often arises in dose-ranging studies of drugs.

The correlation between acceptance rate and (logarithm of) journal circulation, based on the 113 journals in the sample in Figure 8-14, is 0.64. To test the null hypothesis that there is no linear relationship between acceptance rate and logarithm of journal circulation, compute

$$t = \frac{0.64}{\sqrt{(1-0.64^2)/(113-2)}} = 8.78$$

The computed value of t exceeds $t_{.001} = 3.38$ for $\nu = 113 - 2 = 111$ degrees of freedom, so we conclude that there is a correlation between journal size and selectivity ($P < .001$).

[†]Goodman SN, Altman DG, George SL. Statistical reviewing policies of medical journals. *J Gen Intern Med.* 1998;13:753–756.

[‡]For a more detailed discussion of variable transformations, see Glantz SA, Slinker BK. *Primer of Applied Regression and Analysis of Variance*, 2nd ed. New York: McGraw-Hill; 2001, 150–153, 163–166.

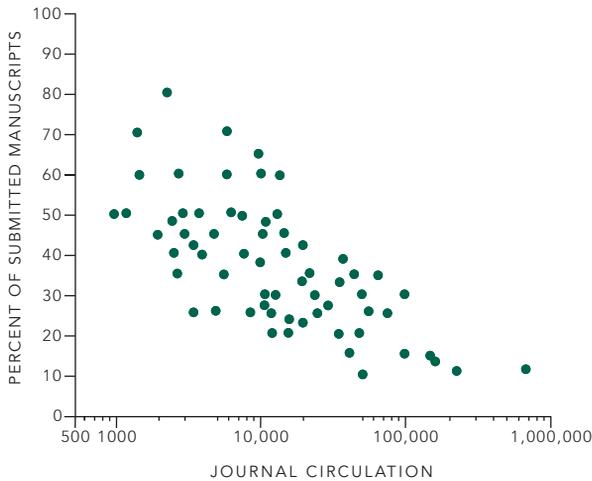


FIGURE 8-14. There appears to be a relationship between the fraction of submitted papers journals select for publication and the (logarithm of) journal circulation, with larger journals being more selective. (With kind permission from Springer Science + Business Media. Goodman SN, Altman DG, George SL. Statistical reviewing policies of medical journals. *J Gen Intern Med.* 1998;13:753-756, Fig. 1.)

Does this result *prove* that increasing circulation makes journals more selective? No. An investigator could not manipulate the size of the circulation of the sample of 113 different journals in Figure 8-14, so these data are the results of an observational rather than an experimental study. These two variables could be related to some third underlying *confounding variable* that makes both the observed variables change simultaneously. Indeed, in this case, the underlying confounding variable is probably perceived quality of the journal, with authors willing to submit their manuscripts to more competitive journals because they are more prestigious, and because of the higher quality, more people are willing to subscribe to the journal.

When interpreting the results of regression analysis, it is important to keep the distinction between observational and experimental studies in mind. When investigators can actively manipulate the independent variable and observe changes in the dependent variable, they can draw strong conclusions about how changes in the independent variable *cause* changes in the dependent variable. On the other hand, when investigators only observe the two variables changing together, they can only observe that an *association* between them in which one changes as the other changes. It is impossible to rule out the possibility that both variables are independently responding to some

third factor and that the independent variable does not causally affect the dependent variable.

■ THE SPEARMAN RANK CORRELATION COEFFICIENT

It is often desirable to test the hypothesis that there is a trend in a clinical state, measured on an ordinal scale, as another variable changes. The Pearson product-moment correlation coefficient is a parametric statistic designed to be used on data distributed normally along interval scales, so it cannot be used. It also requires that the trend relating the two variables be linear. When the sample suggests that the population from which both variables were drawn from does not meet these criteria, it is possible to compute a measure of association based on the *ranks* rather than the values of the observations. This new correlation coefficient, called the *Spearman rank correlation coefficient*, r_s , is based on ranks and can be used for data quantified with an ordinal scale.* The Spearman rank correlation coefficient is a *nonparametric* statistic because it does not require that the observations be drawn from a normally distributed population.†

The idea behind the Spearman rank correlation coefficient is simple. The values of the two variables are ranked in ascending (or descending) order, taking into account the signs of the values. For example, ranking 1, -1, and 2 (from the smallest value -1 to the largest value, 2) yields the ranks

*Another rank correlation coefficient, known as the *Kendall rank correlation coefficient* τ , can be generalized to the case in which there are multiple independent variables. For problems involving only two variables it yields conclusions identical to the Spearman rank correlation coefficient, although the value of τ associated with a given set of observations differs from the value of r_s associated with the same observations. For a discussion of both procedures, see Siegel S, Castellar NJ Jr. *Measures of association and their tests of significance. Nonparametric Statistics for the Behavioral Sciences*, 2nd ed. New York: McGraw-Hill; 1988:chap 9.

†In addition to being explicitly designed to analyze data measured on a rank scale, nonparametric methods can be used in cases where the normality assumptions that underlie the parametric methods are not met or you do not want to assume that they are met. When the assumptions of parametric methods are not met, the nonparametric methods are appropriate. When either nonparametric or parametric methods are appropriate, the nonparametric methods generally have lower power than the parametric methods. In the case of Pearson (parametric) and Spearman (nonparametric) correlations, this difference is very small. For example, sizes above 10, the power of the Spearman rank-order correlation coefficient is computed exactly the same as for the Pearson product-moment correlation, except that σ_z is computed as

$$\sigma_z = \sqrt{\frac{1.060}{n-3}} \text{ i.e., with 1.060 in the numerator instead of 1.000.}$$

■ TABLE 8-6. Computation of Spearman Rank Correlation Coefficient for Observations in Figure 8-3

Height		Weight		Difference of Rank <i>d</i>
Value (cm)	Rank*	Value (g)	Rank*	
31	1	7.7	2	-1
32	2	8.3	3	-1
33	3	7.6	1	2
34	4	9.1	4	0
35	5.5	9.6	5	0.5
35	5.5	9.9	6	-0.5
40	7	11.8	7	0
41	8	12.2	8	0
42	9	14.8	10	0
46	10	13.0	9	0

*1 = smallest value; 10 = largest value

2, 1, and 3, respectively. Next, the Pearson product-moment correlation between the *ranks* (as opposed to the observations) is computed using the same formula as before. A mathematically equivalent formula for the Spearman rank correlation coefficient that is easier to compute is

$$r_s = 1 - \frac{6 \sum d^2}{n^3 - n}$$

in which *d* is the difference of the two ranks associated with each point. The resulting correlation coefficient can then be compared with the population of all possible values it would take on if there were in fact no association between the two variables.* If the value of *r_s* associated with the data is large compared to the expected distribution of *r_s* if the null hypothesis of no relationship is true, we reject this null

hypothesis and conclude that the observations are not compatible with the hypothesis of no association between the two variables.

Table 8-6 illustrates how to compute *r_s* for the observations in Figure 8-3. Both the variables (height and weight) are ranked from 1 to 10 (since there are 10 data points), 1 being assigned to the smallest value and 10 to the largest value. When there is a tie, as there is when the height equals 35 cm, both values are assigned the mean of the ranks that would be used if there were no tie. Since the weight tends to increase as height increases, the ranks of both variables increase together. The Pearson correlation of these two lists of ranks is the Spearman rank correlation coefficient.

The Spearman rank correlation coefficient for the data in Table 8-6 is

$$r_s = 1 - \frac{6[(-1)^2 + (-1)^2 + 2^2 + 0^2 + 0.5^2 + (-0.5)^2 + 0^2 + 0^2 + 0^2 + 0^2]}{10^3 - 10} = 0.96$$

*When there are ties, *r_s* can be more accurately calculated by adjusting for ties using

$$r_s = \frac{1 - \frac{6 \sum d^2}{n^3 - n} - \frac{\sum(\tau_x^3 - \tau_x) + \sum(\tau_y^3 - \tau_y)}{2(n^3 - n)}}{\sqrt{\left[1 - \frac{\sum(\tau_x^3 - \tau_x)}{(n^3 - n)}\right] \left[1 - \frac{\sum(\tau_y^3 - \tau_y)}{(n^3 - n)}\right]}}$$

where τ_x and τ_y are the number of tied values at each value of *x* and *y*.

Table 8-7 gives various risks of making a Type I error. The observed value of *r_s* exceeds .903, the critical value for the most extreme .1% of values when there are *n* = 10 data points, so we can report that there is an association between weight and height (*P* < .001).

■ TABLE 8-7. Critical Values for Spearman Rank Correlation Coefficient*

n	Probability of Greater Value (P)								
	.50	.20	.10	.05	.02	.01	.005	.002	.001
4	.600	1.000	1.000						
5	.500	.800	.900	1.000	1.000				
6	.371	.657	.829	.886	.943	1.000	1.000		
7	.321	.571	.714	.786	.893	.929	.964	1.000	1.000
8	.310	.524	.643	.738	.833	.881	.905	.952	.976
9	.267	.483	.600	.700	.783	.833	.867	.917	.933
10	.248	.455	.564	.648	.745	.794	.830	.879	.903
11	.236	.427	.536	.618	.709	.755	.800	.845	.873
12	.217	.406	.503	.587	.678	.727	.769	.818	.846
13	.209	.385	.484	.560	.648	.703	.747	.791	.824
14	.200	.367	.464	.538	.626	.679	.723	.771	.802
15	.189	.354	.446	.521	.604	.654	.700	.750	.779
16	.182	.341	.429	.503	.582	.635	.679	.729	.762
17	.176	.328	.414	.485	.566	.615	.662	.713	.748
18	.170	.317	.401	.472	.550	.600	.643	.695	.728
19	.165	.309	.391	.460	.535	.584	.628	.677	.712
20	.161	.299	.380	.447	.520	.570	.612	.662	.696
21	.156	.292	.370	.435	.508	.556	.599	.648	.681
22	.152	.284	.361	.425	.496	.544	.586	.634	.667
23	.148	.278	.353	.415	.486	.532	.573	.622	.654
24	.144	.271	.344	.406	.476	.521	.562	.610	.642
25	.142	.265	.337	.398	.466	.511	.551	.598	.630
26	.138	.259	.331	.390	.457	.501	.541	.587	.619
27	.136	.255	.324	.382	.448	.491	.531	.577	.608
28	.133	.250	.317	.375	.440	.483	.522	.567	.598
29	.130	.245	.312	.368	.433	.475	.513	.558	.589
30	.128	.240	.306	.362	.425	.467	.504	.549	.580
31	.126	.236	.301	.356	.418	.459	.496	.541	.571
32	.124	.232	.296	.350	.412	.452	.489	.533	.563
33	.121	.229	.291	.345	.405	.446	.482	.525	.554
34	.120	.225	.287	.340	.399	.439	.475	.517	.547
35	.118	.222	.283	.335	.394	.433	.468	.510	.539
36	.116	.219	.279	.330	.388	.427	.462	.504	.533
37	.114	.216	.275	.325	.383	.421	.456	.497	.526
38	.113	.212	.271	.321	.378	.415	.450	.491	.519
39	.111	.210	.267	.317	.373	.410	.444	.485	.513
40	.110	.207	.264	.313	.368	.405	.439	.479	.507
41	.108	.204	.261	.309	.364	.400	.433	.473	.501
42	.107	.202	.257	.305	.359	.395	.428	.468	.495
43	.105	.199	.254	.301	.355	.391	.423	.463	.490
44	.104	.197	.251	.298	.351	.386	.419	.458	.484
45	.103	.194	.248	.294	.347	.382	.414	.453	.479
46	.102	.192	.246	.291	.343	.378	.410	.448	.474
47	.101	.190	.243	.288	.340	.374	.405	.443	.469
48	.100	.188	.240	.285	.336	.370	.401	.439	.465
49	.098	.186	.238	.282	.333	.366	.397	.434	.460
50	.097	.184	.235	.279	.329	.363	.393	.430	.456

Adapted from Zar JH. *Biostatistical Analysis*, 4th ed. Englewood Cliffs, NJ: Prentice-Hall; 1999, Appendix 116–117. Used by permission.

*For sample sizes greater than 50, use

$$t = \frac{r_s}{\sqrt{(1-r_s^2)/(n-2)}}$$

With $v = n - 2$ degrees of freedom to obtain the approximate P value.

In this example, of course, we could just as well have used the Pearson product-moment correlation. Had we been dealing with data measured on an ordinal scale, we would have had to use the Spearman rank correlation coefficient.

Cell Phone Radiation and Mitochondrial Reactive Oxygen Species in Sperm

We could not use linear regression to make statements about the relationship between the level of exposure to electromagnetic radiation from cell phones (measured as

the specific absorption rate, SAR) and the induced level of sperm with mitochondrial reactive oxygen species (ROS) in Figure 8-8 because the relationship was not linear. We can, however, test for a relationship between these two variables (presented without the inappropriate linear regression line in Fig. 8-5) using a Spearman rank correlation coefficient because, unlike linear regression and the Pearson product-moment correlation, it does not assume that the line of means is straight. Box 8-2 shows that the Spearman rank-order correlation for the data on cell-phone-specific

BOX 8-2 • Calculation of Spearman Rank-Order Correlation for the Data Relating Cell Phone Electromagnetic Radiation to Mitochondrial Reactive Oxygen Species

We first separately rank values of the two variables in Table 8-2 (and Figure 8-15), accounting for tied values within each variable. For example, there are three tied observations of SAR at 1.0 W/kg, which occupy ranks 3, 4, and 5. All are assigned the average rank of 4. Likewise, there are two tied values at 2.8, which occupy ranks 6 and 7, so both are assigned the average rank of 6.5.

Cell Phone Signal Strength and Fraction of Sperm with Reactive Oxygen Species

Cell-Phone-Specific Absorption Rate, SAR (W/kg)		Sperm with Mitochondrial ROS (%)		Difference of Ranks, <i>d</i>	Squared Difference of Ranks, <i>d</i> ²
Value	Rank	Value	Rank		
0.4	2	0	2	0	0.00
27.5	12.5	0.5	12	0.5	.25
0.0	1	0	1	0	0.00
1.0	4	0	4	0	0.00
2.8	6.5	0.5	6	0.5	.25
10.1	10.5	-0.5	11	-0.5	.25
2.8	6.5	-0.5	7	-0.5	.25
27.5	12.5	-0.5	13	-0.5	.25
10.1	10.5	1	9.5	1	1.00
4.3	8.5	-1	9.5	-1	1.00
4.3	8.5	0.5	8	0.5	.25
1.0	4	-1	5	-1	1.00
1.0	4	1	3	1	1.00
Sum					5.50

Next, we compute the difference in the ranks (the last column) and use these differences to compute the Spearman rank-order correlation:

$$r_s = 1 - \frac{6 \sum d^2}{n^3 - n} = 1 - \frac{6 \cdot 5.50}{13^3 - 13} = .985$$

From Table 8-8, this value of *r_s* exceeds .824, the critical value that defines the most extreme .1% of the sampling distribution of the Spearman rank-order correlation under the null hypothesis of no relationship between the two variables, so we reject this null hypothesis and conclude that there is a relationship (*P* < .001).

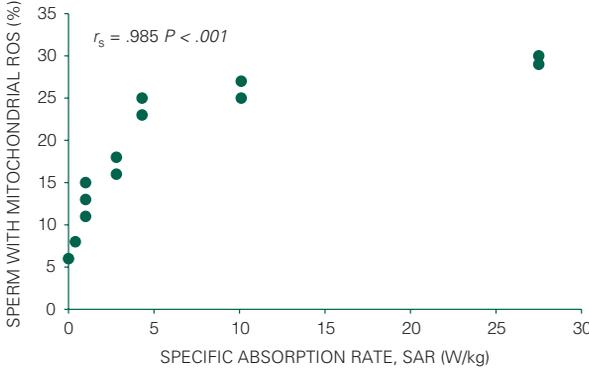


FIGURE 8-15. The fraction of sperm with mitochondrial reactive oxygen species increases nonlinearly with the level of cell phone electromagnetic radiation.

absorption rate and mitochondrial ROS in Table 8-2 is $r_s = .985$. This value exceeds .824 for $n = 13$ data points (from Table 8-7), so we can reject the null hypothesis of no relationship between these two variables with $P < .001$.

POWER AND SAMPLE SIZE IN REGRESSION AND CORRELATION

Power and sample-size computations for regression and correlation are straightforward, based on the fact that testing for a slope significantly different from zero is equivalent to testing for a correlation coefficient significantly different from zero.

The key to these computations is transforming the correlation coefficient according to

$$Z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$$

Z is normally distributed with standard deviation

$$\sigma_Z = \sqrt{\frac{1}{n-3}}$$

Thus,

$$z = \frac{Z}{\sigma_Z}$$

follows the standard normal distribution if there is no correlation between the dependent and independent variables in the underlying population. If there is a correlation in the underlying population of ρ , then

$$z = \frac{Z - Z_\rho}{\sigma_Z}$$

will be normally distributed, where

$$Z_\rho = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right)$$

We will use this fact to compute power analogously to the way we did it for the t test.*

For example, let us compute the power of a regression analysis to detect a correlation of $\rho = .9$ in the underlying population with 95% confidence based on a sample size of 10 observations. We first compute

$$Z_\rho = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right) = \frac{1}{2} \ln \left(\frac{1+.9}{1-.9} \right) = 1.472$$

and

$$\sigma_Z = \sqrt{\frac{1}{n-3}} = \sqrt{\frac{1}{10-3}} = 0.378$$

Therefore, if the actual correlation in the underlying population is .9, the distribution of z is centered on $Z_\rho / \sigma_Z = 1.472 / 0.378 = 3.894$ (Fig. 8-16; compare with Fig. 6-7).

If we use $\alpha = 0.05$ to require 95% confidence in asserting that the correlation is different from zero, then we will reject the null hypothesis when the value of z associated with the data exceeds $z_{\alpha(2)} = 1.960$, the (two-tail) value that defines the most extreme values of the normal distribution (from Table 4-1 or 6-2). This value is $1.960 - 3.894 = -1.934$ below the center of the actual distribution of z . From Table 6-2, .97 of the possible z values is to the right of -1.934 . Thus, the power of a linear regression or correlation of .9 with 95% confidence and a sample size of 10 is 97%.

This process can be reduced to a simple equation. The power of linear regression or correlation to detect a

*This fact can also be used as an alternative technique to test the hypothesis that the correlation coefficient is zero by computing the confidence interval for the observed correlation coefficient as

$$Z - z_\alpha \sigma_Z < Z_\rho < Z + z_\alpha \sigma_Z$$

then converting the upper and lower limits of Z back to correlations by inverting the transformation of r to Z .

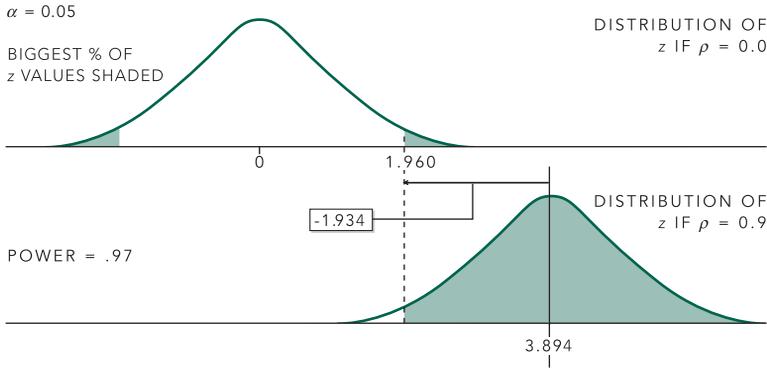


FIGURE 8-16. The power of a correlation to detect a correlation in the population of $\rho = .9$ with a sample size of 10 and 95% confidence is the area under the actual distribution of the z-test statistic above $z_{\alpha} = 1.960$. If $\rho = 0.9$ the actual distribution of z will be centered on 3.894.

correlation of ρ is the area of the standard normal distribution to the right of

$$z_{1-\beta(\text{upper})} = z_{\alpha(2)} - \frac{Z_{\rho}}{\sqrt{\frac{1}{n-3}}}$$

To obtain the sample size necessary to detect a specified correlation with a specified power to a specified level of confidence comes from solving this equation for n

$$n = \left(\frac{z_{\alpha(2)} - z_{1-\beta(\text{upper})}}{Z_{\rho}} \right)^2 + 3$$

■ COMPARING TWO DIFFERENT MEASUREMENTS OF THE SAME THING: THE BLAND–ALTMAN METHOD*

The need often arises, particularly in clinical studies, to compare two different ways of measuring the same thing, when neither method is perfect. For example, as medical technology progresses, less invasive procedures for measuring physiological parameters are developed. The question that arises in the process of developing these new techniques is: How well do they agree with older, more invasive, techniques? Similar questions arise when assessing the repeatability of a measurement: If I measure the same thing twice, how much do the measurements vary? Why not simply compute a regression equation or correlation coefficient for the two sets of observations?

*This section deals with more advanced material and can be skipped with no loss of continuity.

First, neither variable is a natural independent variable, and the choice of independent variable affects the results in a regression equation. The situation in comparing two imperfect clinical measurements of the same thing differs from the *calibration* problem that is common in laboratory science, in which one compares measured values with a known standard. For example, one could mix a known amount of salt with a known amount of distilled water to obtain a given saline concentration, then measure the salt concentrations with some device. It would then be possible to plot the actual salt concentration against the measured salt concentration to obtain a calibration curve. The standard error of the estimate would represent a good measure of uncertainty in the measurement. When comparing two imperfect clinical measurements, there is no such standard.

Second, the correlation coefficient measures the strength of agreement against the null hypothesis of no relationship. When comparing two measurements of the same thing, there will almost always be a relationship between these two measures, so the null hypothesis of no relationship that is implicit in correlation analysis makes no sense.

Third, correlation depends on the range of data in the sample. All else being equal, the wider the range of the observations, the higher the correlation. The presence of an outlier can lead to a high correlation even if there is a good deal of scatter among the rest of the observations.

J. Martin Bland and Douglas Altman[†] developed a simple descriptive technique to assess the agreement between two imperfect clinical measurements or repeatability

[†]For a more detailed discussion of the Bland–Altman method, see Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician*. 1983;32:307–317, or Bland JM, Altman DG. Statistical methods for assessing agreement between two measures of clinical measurement. *Lancet*. 1986;1(8476):307–310.

of duplicate observations. The idea is quite simple: The most straightforward measure of disagreement between the two observations is the difference, so simply compute the differences between all the pairs of observations. Next, compute the mean and standard deviation of the differences. The mean difference is a measure of the *bias* between the two observations and the standard deviation is a measure of the variation between the two observations. Finally, because both observations are equally good (or bad), our best estimate of the true value of the variable being measured is the mean of the two different observations. Plotting the difference against the mean gives an indication of whether there are any systematic differences between the two measurement techniques as a function of the magnitude of the thing being measured.

We will now illustrate the Bland-Altman method with an example.

Assessing Mitral Regurgitation with Echocardiography

The heart pumps blood around the body. The blood goes from the right heart to the lungs, where it takes up oxygen and releases waste gases, to the left heart, where it is pumped to the body, then back to the right heart. This pumping requires that there be valves inside the heart to keep the blood going in the correct direction when the heart contracts. The valve between the lungs and the left heart, known as the mitral valve, prevents blood from being pushed back into the lungs when the left heart is contracting to push the blood to the body. When this valve becomes diseased, it allows blood to be pushed back toward the lungs when the left heart contracts, a situation called *mitral regurgitation*. Mitral regurgitation is bad because it reduces the forward flow of blood from the heart to the body and also has adverse effects on the lungs.

■ TABLE 8-8. Mitral Valve Regurgitant Fraction Measured with Doppler Echocardiography and Cardiac Catheterization in 21 People

Observations		Difference	Mean
Doppler	Catheterization		
.49	.62	-.13	.56
.83	.72	.11	.78
.71	.63	.08	.67
.38	.61	-.23	.50
.57	.49	.08	.53
.68	.79	-.11	.74
.69	.72	-.03	.71
.07	.11	-.04	.09
.75	.66	.09	.71
.52	.74	-.22	.63
.78	.83	-.05	.81
.71	.66	.05	.69
.16	.34	.18	.25
.33	.50	-.17	.42
.57	.62	-.05	.60
.11	.00	.11	.06
.43	.45	-.02	.44
.11	.06	.05	.09
.31	.46	-.15	.39
.20	.03	.17	.12
.47	.50	-.03	.49
		Mean = -.03	
		SD = .12	

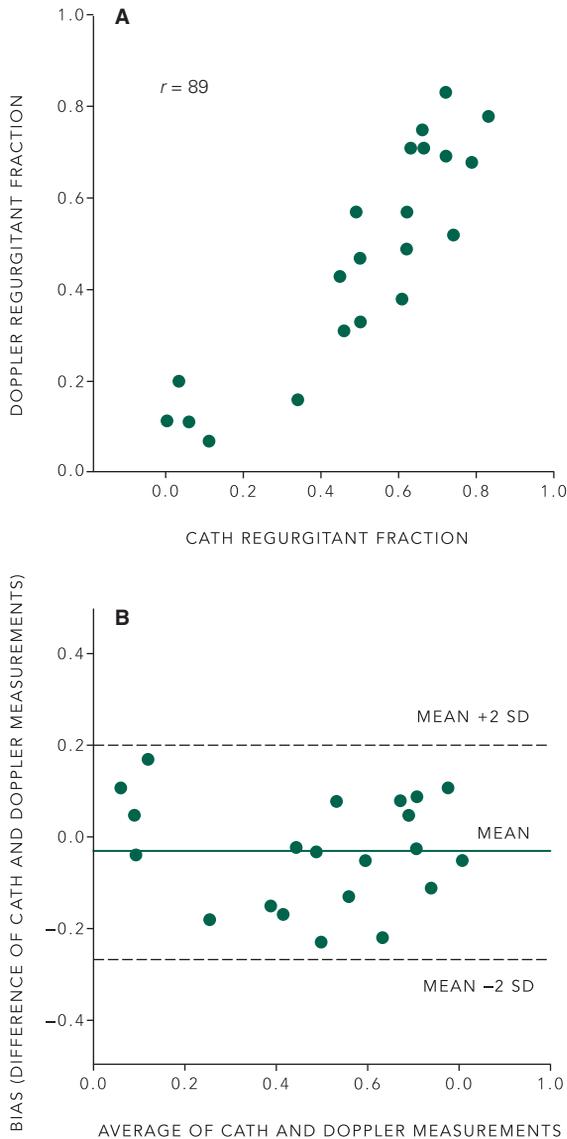


FIGURE 8-17. (A) Relationship between mitral regurgitant fraction measured with cardiac catheterization and Doppler echocardiography in 21 people. **(B)** Bland-Altman curve for the data in panel A. Note that there is little systematic difference between the two measurements.

If it gets bad enough, it becomes necessary to do open heart surgery to replace the valve. Hence, measuring the amount of mitral regurgitation is an important clinical problem.

The amount of regurgitation is quantified with the *regurgitant fraction*,

$$\text{Regurgitant fraction} = \frac{\text{Mitral flow (into the left heart)} - \text{aortic flow (out to the body)}}{\text{Mitral flow}}$$

If the mitral valve is working properly, all the mitral flow into the left heart will appear as flow out into the aorta and the regurgitant fraction will be 0. As the valve becomes more and more incompetent the regurgitant fraction will increase toward 1.

The original way to measure regurgitant fraction has been to do a cardiac catheterization, in which a small tube (called a catheter) is threaded from an artery in the person’s arm or leg into the heart; then, a chemical known as a contrast agent that appears opaque on an X-ray is injected into the heart so that the regurgitant flow can be seen in an X-ray motion picture taken while the contrast agent is being injected. This is an unpleasant, expensive, and potentially dangerous procedure.

Andrew MacIsaac and colleagues* proposed using a noninvasive procedure known as Doppler echocardiography to replace cardiac catheterization as a way to measure regurgitant fraction. Doppler echocardiography involves placing a device that sends high frequency sound waves into the heart and records the reflections on the chest of a person. This information can be used to measure flows into and out of the heart, much as weather radar measures flows of air to track storms and other weather patterns. They compared their method with traditional cardiac catheterization to assess the level of agreement between the two methods.

Table 8-8 shows the results of their study and Figure 8-17A shows a plot of the two measurements against each other, with each person in the study contributing one point. The correlation between the two methods is .89. This fact indicates reasonable agreement, but does not tell us anything about the quantitative nature of the agreement in terms of how well the two methods quantify mitral regurgitant fraction.

*MacIsaac AI, McDonald IG, Kirsner RLG, Graham SA, Gill RW. Quantification of mitral regurgitation by integrated Doppler backscatter power. *J Am Coll Cardiol.* 1994;24:690–695. Data used with permission.

Table 8-8 also shows the computations necessary to construct a Bland-Altman description of how well the two methods agree. The third column in the table represents the differences between the two determinations of regurgitant fraction for each person, and the last column is the mean of the two methods. Figure 8-17B shows a plot of the differences against the mean responses for each person. There are several points to be made from this information. First, the mean difference in regurgitant fraction between the two methods is only $-.03$, which indicates that there is little systemic difference between the two different methods. There also does not appear to be a relationship between the difference between the two observations and the mean mitral regurgitation, so each method is an unbiased estimate of the other. Next, the standard deviation of the differences is $.12$. Taking the range two standard deviations above and below the mean difference gives a measure of the extent of disagreement between the two methods, about $\pm .24$, which is more than half the entire observed range of 0 to $.83$. For example, a Doppler result of $.40$ would mean the corresponding “gold standard” catheter result could be as low as $.16$ and as high as $.64$. These results lead to the conclusion that while, on average, the two methods produce results which are related to each other, the differences for individual patients are large enough that they cannot be used interchangeably to measure mitral regurgitation.

Similar procedures could be used to quantify repeatability of two observations of the same thing by different observers or even repeat observations by the same observer.

■ MULTIPLE REGRESSION

The regression methods we developed in this chapter all involve predicting the dependent variable from one independent variable. As illustrated by the examples we have discussed (and presented in the problems), there are many problems in which such an analysis is appropriate and informative. In real life, including in biomedicine and epidemiology, the dependent (outcome) variable often depends on *several* independent variables acting simultaneously. In a way, we have already acknowledged that fact by trying to design observational studies and experiments in a way to minimize the effects of potential confounding variables. Another way to think about these

confounding variables is to think of them as *additional independent variables*.

Fortunately, the methods presented in this chapter immediately generalize to allow for such situations. To see how, let us write the simple linear regression equation for y as a function of the single independent variable x we have been studying using b_0 instead of a for the intercept and b_1 instead of b for the slope:

$$\hat{y} = b_0 + b_1x$$

It is a small logical step to consider the possibility of y depending on *two* independent variables x_1 and x_2 ,

$$\hat{y} = b_0 + b_1x_1 + b_2x_2$$

This equation is called a *multiple linear regression* (with two independent variables).

Not surprisingly, we can use the same criterion of minimizing the sum of squared differences between the observed and predicted values of the dependent variable y corresponding to the observed values of x_1 holding x_2 constant and x_2 to obtain the “best” estimates of the regression coefficients b_1 and b_2 . In this case, b_1 is an estimate of how much, on average, y changes for a unit change in x_1 and b_2 is an estimate of how much, on average, y changes for a unit change in x_2 holding x_1 constant. Put another way, b_1 is an estimate of the effect of changes in y associated with a unit change in x_1 *controlling* for the effects of x_2 (and vice versa).

Using procedures similar to those in this chapter, it is possible to estimate standard errors for the two regression coefficients and use those standard errors to test whether x_1 , x_2 or both simultaneously predict y .

In fact, we can add any number of independent predictor variables; if we want to predict y based on three independent variables, the multiple regression equation would be

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

Multiple regression methods (also called multivariate methods) are very common in biomedical research because of their ability to model the reality that outcomes often depend simultaneously on several independent variables. They are also widely available in virtually any statistical package that runs on any computer, making them accessible and easy to use. As a result, there is a great

temptation to just dump multivariate data into a computer, push “go,” and look for significant regression coefficients.

The problem with doing this is that, like the simple linear regression we discuss in this chapter, multiple linear regression is based on the assumption that the underlying population meets analogous assumptions, namely that the line of means is straight and that the residuals are normally distributed about the regression line. Unfortunately, with more than one independent variable it is impossible to simply look at a graph of the data and regression line (as we did in the sperm irradiation example in Fig. 8-8) because even if there are just two independent variables, the relationship between them and the dependent variable exists in a three-dimensional space, which is hard to visualize. (If there are three independent variables, the relationship exists in a four-dimensional space, which is impossible to draw.) In addition, multiple regression analysis requires that all the independent variables be completely *independent of each other*, a situation that rarely exists in reality. It turns out that multiple regression analysis (and other related multivariate methods) still produce reliable results when this assumption is mildly violated, but this assumption needs to be checked. Indeed, there is a whole collection of so-called *regression diagnostics* to make sure that the data are consistent with the assumptions of multiple regression analysis.

It is also possible to conduct multiple regression analysis with a qualitative dependent variable (such as presence or absence of a disease) using a technique called *logistic regression*. Logistic regression is widely used in clinical trial and epidemiological research to control for the effects of potential confounding variables in assessing the effects of multiple independent variables on the qualitative (binary, yes or no) outcome variable.

While there are many such technical details that one must attend to when doing a multiple regression analysis, the basic ideas and interpretation of the results is essentially the same as those this chapter presents. Multiple logistic regression analysis is a common technique to adjust the odds ratio for confounding factors, where the outcome variable is absence or presence of disease and the independent variable is exposure or non-exposure of risk factors in addition to the potential confounding variables.

SUMMARY

The methods described in this chapter allow us to quantify the relationship between two variables. The basic approach is the same as in earlier statistical methods: we described the nature of the underlying population, summarized this information with appropriate statistical parameters, then developed procedures for estimating these parameters and their standard errors from one or more samples. When relating two variables with regression or correlation, it is particularly important to examine a graph of the data to see that the assumptions underlying the statistical method you are using is reasonably satisfied by the data you have collected.

PROBLEMS

8-1 Plot the data and compute the linear regression of Y on X and correlation coefficient for each of the sets of observations shown in Table 8-9. In each case, draw the regression line on the same plot as the data. What stays the same and what changes? Why?

TABLE 8-9. Data for Problem 8-1

a		b		c	
X	Y	X	Y	X	Y
30	37	30	37	30	37
30	47	30	47	30	47
40	50	40	50	40	50
40	60	40	60	40	60
		20	25	20	25
		20	35	20	35
		50	62	50	62
		50	72	50	72
				10	13
				10	23
				60	74
				60	84

8-2 Plot the data and compute the linear regression of Y on X and correlation coefficient for each of the sets of observations shown in Table 8-10. In each case, draw the regression line on the same plot as the data. Discuss the results.

■ TABLE 8-10. Data for Problem 8-2

a		b	
X	Y	X	Y
15	19	20	21
15	29	20	31
20	25	30	18
20	35	30	28
25	31	40	15
25	41	40	25
30	37	40	75
30	47	40	85
60	40	50	65
		50	75
		60	55
		60	65

8-3 The plots in Box 8-3 show data from four different experiments together with the associated observations. Compute the regression and correlation coefficients for each of these four sets of data. Discuss the similarities and differences among these sets of data. Include an examination of the assumptions made in linear regression and correlation analyses.*

8-4 Polychlorinated biphenyls (PCBs) are compounds that were once used as an insulating material in electrical transformers before being banned in the United States during the 1970s because of concerns about their toxicity. Despite the ban, PCBs can still be detected in most people because they are persistent in the environment and tend to accumulate in fat tissue as animals that absorb PCBs eat other animals that have absorbed PCBs. One of the major current sources of human PCB exposure is eating fatty fish caught from contaminated waters. In the early 1980s, the husband and wife team of Joseph Jacobsen and Sandra Jacobsen[†] began a prospective study to examine the relationship between PCB levels in a group of women who ate Lake Michigan fish and the intellectual development of their children. The amount of PCBs (ng/g of fat) detected

*This example is from Anscombe FJ. Graphs in statistical analysis. *Am Stat.* 1973;27:17–21.

[†]Jacobsen J, Jacobsen S. Intellectual impairment in children exposed to polychlorinated biphenyls in utero. *N Engl J Med.* 1996;335:783–789.

in maternal milk was used as an indicator of prenatal exposure to PCBs. The Jacobsens then administered the Wechsler Intelligence Scale for Children IQ test to the children when they were 11 years old. Table 8-11 shows the results. Is there any association between maternal PCB level and the children's IQ score?

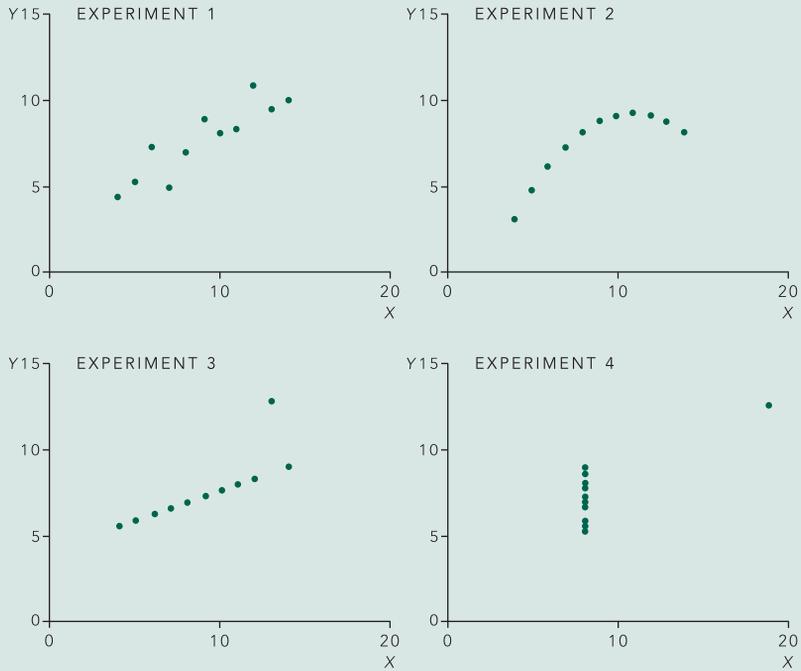
■ TABLE 8-11. Data on PCB Levels in Maternal Milk and Children's IQ

Maternal Milk PCB Level (ng/g of Fat)	Full-Scale IQ
539	116
1093	108
1665	94
476	127
550	122
999	97
974	85
1205	115
604	112
850	108
596	112
547	105
1164	95
905	108

8-5 The ability to measure hormone levels based on a blood spot (like that used by diabetics for glucose monitoring) has several advantages over measurements based on a blood draw. First, blood spots allow for repeated measurements over the time course of minutes and hours. Second, they can be collected with minimal training by a research assistant or the subject. Finally, they are easy to store in a variety of experimental conditions. The low levels of hormone in both blood spots and blood draws are currently measured by a technique called the radioimmunoassay (RIA), an assay based on binding of a radioactively labeled hormone to a specific antibody. Elizabeth Shirtcliff and colleagues[‡] used a modification of a commercially available RIA to detect estradiol (the primary estrogen found in humans) in blood spots and compared

[‡]Shirtcliff E, et al. Assaying estradiol in biobehavioral studies using saliva and blood spots: simple radioimmunoassay protocols, reliability and comparative validity. *Horm Behav.* 2000;38:137–147.

BOX 8-3 • Plots and Raw Data for Problem 8-3



Experiment 1		Experiment 2		Experiment 3		Experiment 4	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89
ΣX	99	ΣX	99	ΣX	99	ΣX	99
ΣY	82.5	ΣY	82.5	ΣY	82.5	ΣY	82.5
ΣX^2	1001	ΣX^2	1001	ΣX^2	1001	ΣX^2	1001
ΣY^2	660	ΣY^2	660	ΣY^2	660	ΣY^2	660
ΣXY	797.5	ΣXY	797.5	ΣXY	797.5	ΣXY	797.5

the results to those obtained by blood draw. How good is the agreement between the volumes measured using these two techniques? The results are shown in Table 8-12.

■ **TABLE 8-12. Estradiol Measured in Two Different Ways**

Estradiol Measurements	
Blood Spot Estradiol (pg/mL)	Blood Draw Estradiol (pg/mL)
17	25
18	29
21	24
22	33
27	35
30	40
34	40
35	45
36	58
40	63
44	70
45	60
49	70
50	95
52	105
53	108
57	95
58	90
72	130
138	200

8-6 Arteries adjust their size on a minute-to-minute basis to meet the needs of the body for blood to carry oxygen to the tissues and to remove waste products. A substantial part of this response is mediated by the one cell thick lining of the arteries known as the vascular endothelium responding to nitric oxide that the endothelium produces from the amino acid L-arginine. As part of an investigation of the effect of secondhand smoke on the ability of the endothelium to dilate arteries, Stuart Hutchison and colleagues* examined the relationship between how much segments of arteries relaxed after being exposed to differ-

ent levels of L-arginine and subjected to two different stimuli, acetylcholine and a drug, A23187 (Table 8-13). Is there a relationship between relaxation and L-arginine level in the presence of these two different relaxing agents? Is there a difference in the effects of these two drugs? (Note: To “linearize” the data, take natural logarithms of the arginine levels before doing the analysis.)

8-7 Erectile dysfunction is widely recognized to be associated with diabetes and cardiovascular disease. To investigate whether erectile dysfunction was associated with lower urinary tract infections, Wo-Sik Chung and colleagues* administered standard questionnaires to men between ages 40 and 70 years to assess the presence of lower urinary tract infections as well as questions regarding erectile function, with higher scores on the questionnaires indicating more serious problems with urinary tract infections and better erectile function, respectively (Table 8-14). Is there evidence for a relationship between urinary tract infections and erectile dysfunction?

8-8 As part of a study of the nature of cancers of the gum and lower jaw, Eiji Nakayama and his colleagues† were interested in relating the extent of cancer invasion, determined by direct histological investigation with the levels of invasion measured on a computed tomographic scan of people with cancer. They measured both variables on an ordinal scale as follows:

1. Erosive
2. Erosive and Partially Mixed
3. Mixed
4. Mixed and Partially Invasive
5. Invasive

Is there a relationship between these two ways of quantifying the extent of cancer between these two ways of assessing the disease severity? Is the relationship strong enough to use these two methods interchangeably? The data are shown in Table 8-15.

*Hutchison S, et al. Secondhand tobacco smoke impairs rabbit pulmonary artery endothelium-dependent relaxation. *Chest*. 2001;120:2004-2012.

*Chung W-S, et al. Lower urinary tract symptoms and sexual dysfunction in community-dwelling men. *Mayo Clin Proc*. 2004;79:745-749.

†Nakayama E, et al. The correlation of histological features with a panoramic radiography pattern and a computed tomography pattern of bone destruction in carcinoma of the mandibular gingival. *Oral Surg Oral Med Oral Path Oral Radiol Endod*. 2003;96:774-782.

■ TABLE 8-13. Artery Relaxation Force after Exposure to Two Different Relaxing Agents

Acetylcholine		A23187	
Arginine Level	Relaxation Force (%)	Arginine Level	Relaxation Force (%)
.02	-10	.03	-2
.03	-21	.04	-47
.1	-48	.10	-36
.5	-52	.13	-27
.6	-41	.5	-43
.7	-52	.6	-56
.9	-67	.6	-50
.9	-58	.7	-77
.9	-32	.8	-67
1.2	-58	.8	-42
1.3	-29	1.2	-60
		1.2	-36
		1.6	-68

■ TABLE 8-14. Relationship Between Urinary Tract Infections and Erectile Dysfunction

Urinary Tract Infection Score	Erectile Function Score
1	14
0	15
9	6
6	11
5	12
5	10
0	11
4	12
8	10
7	8
0	14
10	3
8	9
4	12
16	3
8	9
2	13
13	2
10	4
18	4

■ TABLE 8-15. Two Different Ways of Measuring Gum and Jaw Disorders

Histology	Computed Tomography
3	5
3	2
1	1
4	5
3	3
3	3
5	4
4	3
4	3
3	3
5	5
4	3
4	4
2	2
3	5
1	3
3	2
2	3
4	3
2	3
3	2
2	3
3	3
3	3
2	5

■ TABLE 8-16. Relationship Between Obesity and Insulin Sensitivity

Controls (No Immediate Family Member with High Blood Pressure)			Relatives (Immediate Family Member with High Blood Pressure)		
Waist/Hip Ratio, <i>R</i>	Insulin Sensitivity	Log (Insulin Sensitivity) <i>I</i>	Waist/Hip Ratio, <i>R</i>	Insulin Sensitivity	Log (Insulin Sensitivity) <i>I</i>
0.775	21.0	1.322	0.800	10.0	1.000
0.800	20.0	1.301	0.810	5.0	0.699
0.810	13.5	1.130	0.850	9.5	0.978
0.800	8.5	0.929	0.875	2.5	0.398
0.850	10.5	1.021	0.850	4.0	0.602
0.860	10.0	1.000	0.870	5.8	0.763
0.925	12.8	1.107	0.910	9.8	0.971
0.900	9.0	0.954	0.925	8.0	0.903
0.925	6.5	0.813	0.925	6.0	0.778
0.945	11.0	1.041	0.940	4.3	0.633
0.945	10.5	1.021	0.945	8.5	0.929
0.950	9.5	0.978	0.960	9.0	0.954
0.975	5.5	0.740	1.100	8.5	0.929
1.050	6.0	0.778	1.100	4.5	0.653
1.075	3.8	0.580	0.990	2.3	0.362

8-9 What is the power of the study of journal circulation and selectivity described in Figure 8-12 to detect a correlation of .6 with 95% confidence? (There are 113 journals in the sample.)

8-10 What sample size is necessary to have an 80% power for detecting a correlation between journal circulation and selectivity with 95% confidence if the actual correlation in the population is .6?

8-11 Clinical and epidemiologic studies have demonstrated an association between high blood pressure, diabetes, and high levels of lipids measured in blood. In addition, several studies demonstrated that people with high blood pressure have lower insulin sensitivity than people with normal blood pressure, and that physical fitness affects insulin sensitivity. As part of an investigation of whether there is a genetic component of the relation-

ship between high blood pressure and insulin sensitivity, Tomas Endre and colleagues* investigated the relationship between insulin sensitivity and a measure of physical fitness in two groups of men with normal blood pressure, one with immediate relatives who have high blood pressure and a similar group of men from families with normal blood pressure. They used the waist-to-hip ratio of the men as a measure of physical fitness and examined the relationship between it and insulin sensitivity index in these two groups of men (Table 8-16). Is the relationship the same in these two groups of men? (Use the logarithm of the insulin sensitivity index as the dependent variable in order to linearize the relationship between the two variables.)

*Endre T, et al. Insulin resistance is coupled to low physical fitness in normotensive men with a family history of hypertension. *J Hypertens.* 1994;12:81-88.

This page intentionally left blank

Experiments When Each Subject Receives More Than One Treatment

The procedures for testing hypotheses discussed in Chapters 3 to 5 apply to experiments in which the control and treatment groups contain *different* subjects (individuals). It is often possible to design experiments in which *each* experimental subject can be observed *before* and *after* one or more treatments. Such experiments are generally more sensitive because they make it possible to measure how the treatment *affects each individual*. When the control and treatment groups consist of different individuals, the changes due to the treatment may be masked by variability between experimental subjects. This chapter shows how to analyze experiments in which each subject is repeatedly observed under different experimental conditions.

We will begin with the *paired t test* for experiments in which the subjects are observed before and after receiving a single treatment. Then, we will generalize this test to obtain *repeated measures analysis of variance*, which permits testing hypotheses about any number of treatments whose effects are measured repeatedly in the same subjects. We will explicitly separate the total variability in the observations into three components: variability between the experimental subjects, variability in each individual subject's response, and variability due to the treatments. Like all analyses of variance (including *t tests*), these procedures require that the observations come from normally distributed populations. (Chapter 10 presents methods based on ranks that do not require this assumption.) Finally, we will develop *McNemar's test* to analyze data measured on a nominal scale and presented in contingency tables.

■ EXPERIMENTS WHEN SUBJECTS ARE OBSERVED BEFORE AND AFTER A SINGLE TREATMENT: THE PAIRED *t* TEST

In experiments in which it is possible to observe each experimental subject *before* and *after* administering a single treatment, we will test a hypothesis about the average *change* the treatment produces instead of the difference in average responses with and without the treatment. This approach reduces the variability in the observations due to differences between individuals and yields a more sensitive test.

Figure 9-1 illustrates this point. Figure 9-1A shows daily urine production in *two* samples of 10 different people each; one sample group took a placebo and the other took a drug. Since there is little difference in the mean response relative to the standard deviations, it would be hard to assert that the treatment produced an effect on the basis of these observations. In fact, *t* computed using the methods of Chapter 4 is only 1.33, which comes nowhere near $t_{.05} = 2.101$, the critical value for $v = n_{\text{pla}} + n_{\text{drug}} - 2 = 10 + 10 - 2 = 18$ degrees of freedom.

Now consider Figure 9-1B. It shows urine productions identical to those in Figure 9-1A but for an experiment in which urine production was measured in *one* sample of 10 individuals *before* and *after* administering the drug. A straight line connects the observations for each individual. Figure 9-1B shows that the drug increased urine production in 8 of the 10 people in the sample. This result suggests that the drug *is* an effective diuretic.

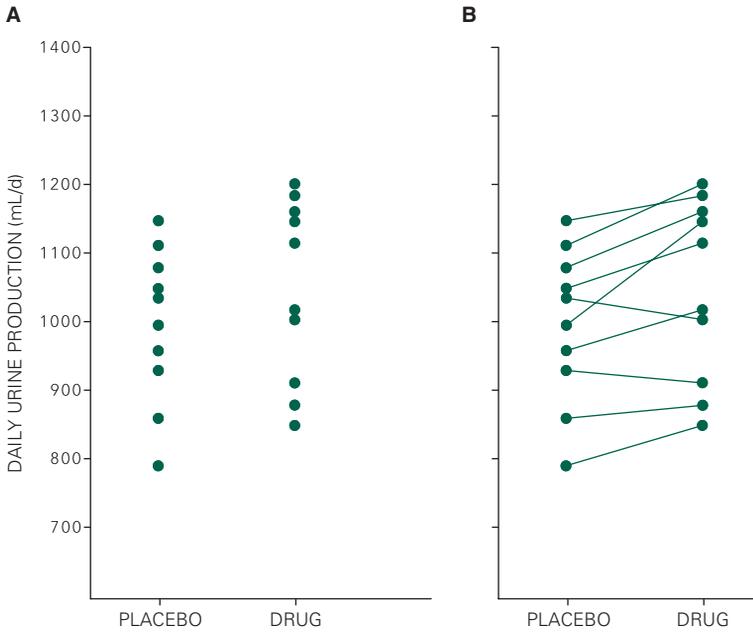


FIGURE 9-1. (A) Daily urine production in two groups of 10 different people. One group of 10 people received the placebo and the other group of 10 people received the drug. The diuretic does not appear to be effective. (B) Daily urine production in a single group of 10 people before and after taking a drug. The drug appears to be an effective diuretic. The observations are identical to those in panel A; by focusing on changes in each individual's response rather than the response of all the people taken together, it is possible to detect a difference that was masked by the between subjects variability in panel A.

By concentrating on the *change* in each individual that accompanied taking the drug (in Fig. 9-1B), we could detect an effect that was masked by the variability between individuals when different people received the placebo and the drug (in Fig. 9-1A).

Now, let us develop a statistical procedure to quantify our subjective impression in such experiments. The *paired t test* can be used to test the hypothesis that there is, on the average, no change in each individual after receiving the treatment under study. Recall that the general definition of the *t* statistic is

$$t = \frac{\text{Parameter estimate} - \text{true value of population parameter}}{\text{Standard error of parameter estimate}}$$

The parameter we wish to estimate is the average difference in response δ in *each individual* due to the treatment. If we let d equal the observed change in each individual that accompanies the treatment, we can use \bar{d} , the mean change, to estimate δ . The standard deviation of the observed differences is

$$s_d = \sqrt{\frac{\sum (d - \bar{d})^2}{n - 1}}$$

So the standard error of the difference is

$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}}$$

Therefore,

$$t = \frac{\bar{d} - \delta}{s_{\bar{d}}}$$

To test the hypothesis that there is, on the average, no response to the treatment, set $\delta = 0$ in this equation to obtain

$$t = \frac{\bar{d}}{s_{\bar{d}}}$$

The resulting value of t is compared with the critical value of $\nu = n - 1$ degrees of freedom.

To recapitulate, when analyzing data from an experiment in which it is possible to observe each individual before and after applying a single treatment:

- Compute the change in response that accompanies the treatment in each individual d .
- Compute the mean change \bar{d} and the standard error of the mean change $s_{\bar{d}}$.

- Use these numbers to compute $t = \bar{d}/s_d$.
- Compare this t with the critical value for $\nu = n - 1$ degrees of freedom, where n is the number of experimental subjects.

Note that the number of degrees of freedom, ν , associated with the paired t test is $n - 1$, less than the $2(n - 1)$ degrees of freedom associated with analyzing these data using an unpaired t test. This loss of degrees of freedom increases the critical value of t that must be exceeded to reject the null hypothesis of no difference. While this situation would seem undesirable, because of the typical biological variability that occurs between individuals this loss of degrees of freedom is virtually always more than compensated for by focusing on *differences within subjects*, which reduces the variability in the results used to compute t . All other things being equal, paired designs are almost always more powerful for detecting effects in biological data than unpaired designs.

Finally, the paired t test, like all t tests, is predicated on a normally distributed population. In the t test for unpaired observations developed in Chapter 4, responses needed to be normally distributed. In the paired t test, the differences (changes within each subject) associated with the treatment need to be normally distributed.

Cigarette Smoking and Platelet Function

Smokers are more likely to develop diseases caused by abnormal blood clots (thromboses), including heart attacks and occlusion of peripheral arteries, than nonsmokers. Platelets are small bodies that circulate in the blood and stick together to form blood clots. Since smokers experience more disorders related to undesirable blood clots than nonsmokers, Peter Levine* drew blood samples in 11 people before and after they smoked a single cigarette and measured the extent to which platelets aggregated when exposed to a standard stimulus. This stimulus, adenosine diphosphate, makes platelets release their granular contents, which, in turn, makes them stick together and form a blood clot.

Figure 9-2 shows the results of this experiment, with platelet stickiness quantified as the maximum percentage

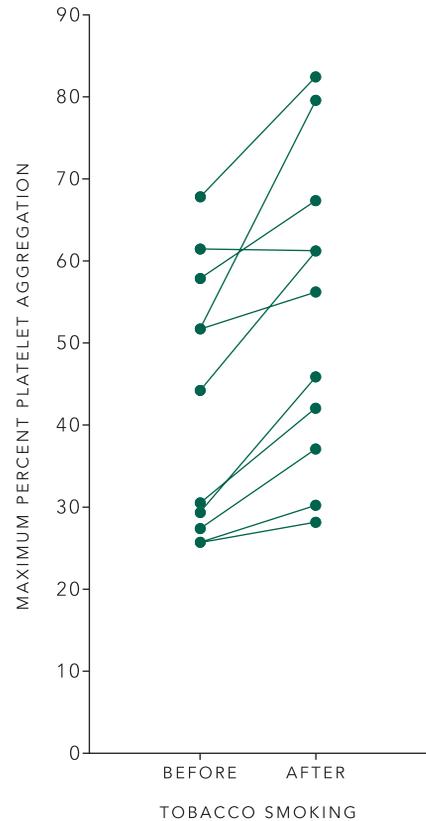


FIGURE 9-2. Maximum percentage platelet aggregation before and after smoking a tobacco cigarette in 11 people. (Adapted with permission of the American Heart Association, Inc. from Fig. 1 of Levine PH. An acute effect of cigarette smoking on platelet function: a possible link between smoking and arterial thrombosis. *Circulation*. 1973;48:619–623.)

of all the platelets that aggregated after being exposed to adenosine diphosphate. The *pair* of observations made in each individual before and after smoking the cigarette is connected by straight lines. The mean percentage aggregations were 43.1% before smoking and 53.5% after smoking, with standard deviations of 15.9% and 18.7%, respectively. Simply looking at these numbers does not suggest that smoking had an effect on platelet aggregation. This approach, however, omits an important fact about the experiment: the platelet aggregations were not measured in two different (independent) groups of people, smokers and nonsmokers, but in a single group of people who were observed both before and after smoking the cigarette.

*Levine PH. An acute effect of cigarette smoking on platelet function: a possible link between smoking and arterial thrombosis. *Circulation*. 1973;48:619–623.

In all but one individual, the maximum platelet aggregation increased after smoking the cigarette, suggesting that smoking facilitates thrombus formation. The means and standard deviations of platelet aggregation before and after smoking for all people taken together did not suggest this pattern because the variability between individuals masked the variability in platelet aggregation that was due to smoking the cigarette. When we took into account the fact that the data consisted of *pairs* of observations done before and after smoking in each individual, we could focus on the *change* in response and so remove the variability that was due to the fact that different people have different platelet-aggregation tendencies regardless of whether they smoked a cigarette or not.

The changes in maximum percent platelet aggregation that accompany smoking are (from Fig. 9-2) 2%, 4%, 10%, 12%, 16%, 15%, 4%, 27%, 9%, -1%, and 15%. Therefore, the mean change in percent platelet aggregation with smoking in these 11 people is $\bar{d} = 10.3\%$. The standard deviation of the change is 8.0%, so the standard

error of the change is $s_{\bar{d}} = 8.0/\sqrt{11} = 2.41\%$. Finally, our test statistic is

$$t = \frac{\bar{d}}{s_{\bar{d}}} = \frac{10.3}{2.41} = 4.27$$

This value exceeds 3.169, the value that defines the most extreme 1% of the *t* distribution with $\nu = n - 1 = 11 - 1 = 10$ degrees of freedom (from Table 4-1). Therefore, we report that smoking increases platelet aggregation ($P < .01$).

How convincing is this experiment that a constituent specific to *tobacco* smoke, as opposed to other chemicals common to smoke in general (e.g., carbon monoxide), or even the stress of the experiment produced the observed change? To investigate this question, Levine also had his subjects “smoke” an unlit cigarette and a lettuce leaf cigarette that contained no nicotine. Figure 9-3 shows the results of these experiments, together with the results of smoking a standard cigarette (from Fig. 9-2).

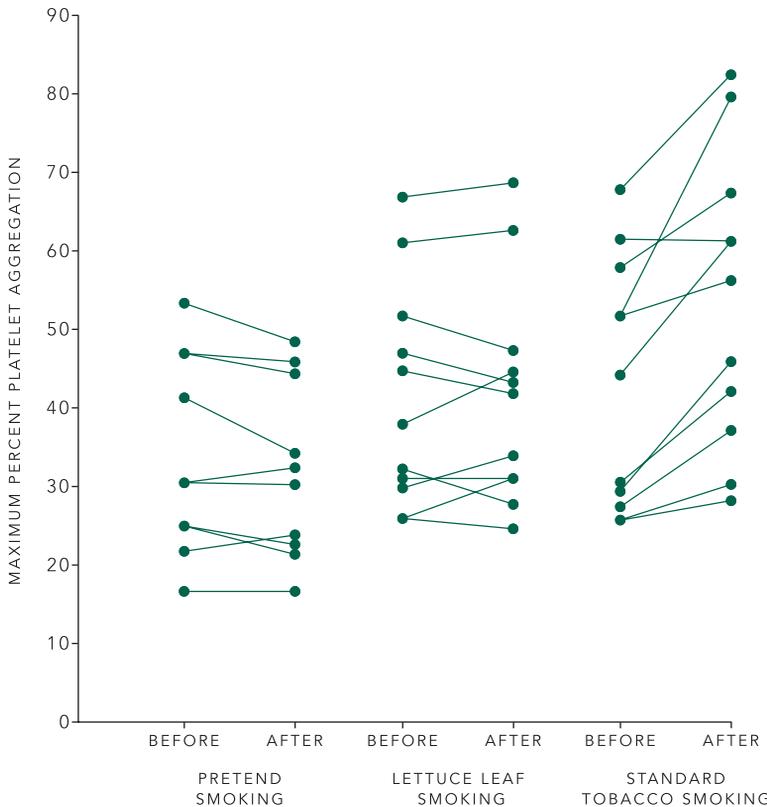


FIGURE 9-3. Maximum percentage platelet aggregation in 11 people before and after pretending to smoke (“sham smoking”), before and after smoking a lettuce-leaf cigarette that contained no nicotine, and before and after smoking a tobacco cigarette. These observations, taken together, suggest that it was something in the tobacco smoke, rather than the act of smoking or other general constituents of smoke, that produced the change in platelet aggregation. (Redrawn with permission of the American Heart Association, Inc. from Fig. 1 of Levine PH. An acute effect of cigarette smoking on platelet function: a possible link between smoking and arterial thrombosis. *Circulation*. 1973;48:619–623.)

When the experimental subjects merely pretended to smoke or smoked a non-nicotine cigarette made of dried lettuce, there was no discernible change in platelet aggregation. This situation contrasts with the increase in platelet aggregation that followed smoking a single tobacco cigarette. This experimental design illustrates an important point:

In a well-designed experiment, the only difference between the treatment group and the control group, both chosen at random from a population of interest, is the treatment.

In this experiment the treatment of interest was tobacco constituents in the smoke, so it was important to compare the results with observations obtained after exposing the subjects to non tobacco smoke. This step helped ensure that the observed changes were due to the tobacco rather than smoking in general. The more carefully the investigator can isolate the treatment effect, the more convincing the conclusions will be.

There are also subtle biases that can cloud the conclusions from an experiment. Most investigators, and their colleagues and technicians, want the experiments to support their hypothesis. In addition, the experimental subjects, when they are people, generally want to be helpful and wish the investigator to be correct, especially if the study is evaluating a new treatment that the experimental subject hopes will provide a cure. These factors can lead the people doing the study to tend to slant judgment calls (often required when collecting the data) toward making the study come out the way everyone wants. For example, the laboratory technicians who measure platelet aggregation might read the control samples on the low side and the smoking samples on the high side without even realizing it. Perhaps some psychological factor among the experimental subjects (analogous to a placebo effect) led their platelet aggregation to increase when they smoked the tobacco cigarette. Levine avoided these difficulties by doing the experiments in a *double blind* manner in which the investigator, the experimental subject, and the laboratory technicians who analyzed the blood samples did not know the content of the cigarettes being smoked until after all experiments were complete and specimens analyzed. As discussed in Chapter 2, double-blind studies are the most effective way to eliminate bias due to both the observer and experimental subject.

In *single blind* studies one party, usually the investigator, knows which treatment is being administered. This

approach controls biases due to the placebo effect but not observer biases. Some studies are also partially blind, in which the participants know something about the treatment but do not have full information. For example, the blood platelet study might be considered partially blind because both the subject and the investigator obviously knew when the subject was only pretending to smoke. It was possible, however, to withhold this information from the laboratory technicians who actually analyzed the blood samples to avoid biases in their measurements of percent platelet aggregation.

The paired *t* test can be used to test hypotheses when observations are taken before and after administering a single treatment to a group of individuals. To generalize this procedure to experiments in which the same individuals are subjected to a number of treatments, we now develop *repeated measures analysis of variance*.

To do so, we must first introduce some new nomenclature for analysis of variance. To ease the transition, we begin with the analysis of variance presented in Chapter 3, in which each treatment was applied to *different* individuals. After reformulating this type of analysis of variance, we will go on to the case of repeated measurements on the same individual.

■ ANOTHER APPROACH TO ANALYSIS OF VARIANCE*

When we developed analysis of variance in Chapter 3, we assumed that all the samples were drawn from a single population (i.e., that the treatments had no effect), estimated the variability in that population from the variability within the sample groups and between the sample groups, then compared these two estimates to see how compatible they were with the original assumption—the null hypothesis—that all the samples were drawn from a single population. When the two estimates of variability were unlikely to arise if the samples had been drawn from a single population, we rejected the null hypothesis of no effect and concluded that

*This and the following section, which develops repeated measures analysis of variance (the multitreatment generalization of the paired *t* test), are more mathematical than the rest of the text. Some readers may wish to skip this section until they encounter an experiment that should be analyzed with repeated measures analysis of variance. Despite the fact that such experiments are common in the biomedical literature, this test is rarely used. This decision leads to the same kinds of multiple *t* test errors discussed in Chapters 3 and 4 for the unpaired *t* test.

at least one of the samples represented a different population (i.e., that at least one treatment had an effect). We used estimates of the population *variance* to quantify variability.

In Chapter 8, we used a slightly different method to quantify the variability of observed data points about a regression line. We used the *sum of squared deviations* about the regression line to quantify variability. The variance and sum of squared deviations, of course, are intimately related. One obtains the variance by dividing the sum of squared deviations by the appropriate number of degrees of freedom. We now will recast analysis of variance using sums of squared deviations to quantify variability. This new nomenclature forms the basis of all forms of analysis of variance, including repeated measures analysis of variance.

In Chapter 3, we considered the following experiment: To determine whether diet affected cardiac output in people living in a small town, we randomly selected four groups of seven people each. People in the control group continued eating normally; people in the second group ate only spaghetti; people in the third group ate only steak; and people in the fourth group ate only fruit and nuts. After 1 month, each person was catheterized and his cardiac output measured. Figure 3-1 showed that diet did not, in fact, affect cardiac output. Figure 3-2 showed the results of the experiment as they would appear to you as an investigator or reader. Table 9-1 presents the same data in tabular form. The four different groups did show some variability in cardiac output. The question is: How consistent is this observed variability

with the hypothesis that diet did not have any effect on cardiac output?

Some New Notation

Tables 9-1 and 9-2 illustrate the notation we will now use to answer this question; it is required for more general forms of analysis of variance. The four different diets are called the *treatments* and are represented by the columns in the table. We denote the four different treatments with the numbers 1 to 4 (1 = control, 2 = spaghetti, 3 = steak, 4 = fruit and nuts). Seven *different* people receive each treatment. Each particular experimental subject (or, more precisely, the observation or data point associated with each subject) is represented by X_{ts} , where t represents the treatment and s represents a specific subject in that treatment group. For example, $X_{11} = 4.6$ L/min represents the observed cardiac output for the first subject ($s = 1$) who received the control diet ($t = 1$). $X_{35} = 5.1$ L/min represents the fifth subject ($s = 5$) who had the steak diet ($t = 3$).

Tables 9-1 and 9-2 also show the mean cardiac outputs for all subjects (in this case, people) receiving each of the four treatments, labeled \bar{X}_1 , \bar{X}_2 , \bar{X}_3 , and \bar{X}_4 . For example, $\bar{X}_2 = 5.23$ L/min is the mean cardiac output observed among people who were treated with spaghetti. The tables also show the variability within each of the treatment groups, quantified by the *sum of squared deviations about the treatment mean*,

Sum of squares for treatment $t =$ sum, over all subjects who received treatment t , of (value of

■ TABLE 9-1. Cardiac Output (L/min) in Four Groups of Seven People Fed Different Diets

	Treatment (Diet)			
	Control	Spaghetti	Steak	Fruit and Nuts
	4.6	4.6	4.3	4.3
	4.7	5.0	4.4	4.4
	4.7	5.2	4.9	4.5
	4.9	5.2	4.9	4.9
	5.1	5.5	5.1	4.9
	5.3	5.5	5.3	5.0
	5.4	5.6	5.6	5.6
Treatment (column) means	4.96	5.23	4.93	4.80
Treatment (column) sums of squares	0.597	0.734	1.294	1.200
Grand mean = 4.98	Total sum of squares = 4.507			

observation for subject–mean response of all individuals who received treatment t)².

The equivalent mathematical statement is

$$SS_t = \sum_s (X_{ts} - \bar{X}_t)^2$$

The summation symbol, Σ , has been modified to indicate that we sum over all s subjects who received treatment t . We need this more explicit notation because we will be summing up the observations in different ways. For example, the sum of squared deviations from the mean cardiac output for the seven people who ate the control diet ($t = 1$) is

$$\begin{aligned} SS_1 &= \sum_s (X_{1s} - \bar{X}_1)^2 \\ &= (4.6 - 4.96)^2 + (4.7 - 4.96)^2 + (4.7 - 4.96)^2 \\ &\quad + (4.9 - 4.96)^2 + (5.1 - 4.96)^2 + (5.3 - 4.96)^2 \\ &\quad + (5.4 - 4.96)^2 = 0.597(\text{L/min})^2 \end{aligned}$$

Recall that the definition of sample variance is

$$s^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

where n is the size of the sample. The expression in the numerator is just the sum of squared deviations from the sample mean, so we can write

$$s^2 = \frac{SS}{n - 1}$$

Hence, the variance in treatment group t equals the sum of squares for that treatment divided by the number of individuals who received the treatment (i.e., the sample size) minus 1:

$$s_t^2 = \frac{SS_t}{n - 1}$$

In Chapter 3, we estimated the population variance from within the groups for our diet experiment with the average of the variances computed from within each of the four treatment groups

$$s_{\text{wit}}^2 = 1/4 (s_{\text{con}}^2 + s_{\text{spa}}^2 + s_{\text{st}}^2 + s_{\text{fn}}^2)$$

In the notation of Table 9-1, we can rewrite this equation as

$$s_{\text{wit}}^2 = 1/4 (s_1^2 + s_2^2 + s_3^2 + s_4^2)$$

Now, replace each of the variances in terms of sums of squares.

$$s_{\text{wit}}^2 = \frac{1}{4} \left[\frac{\sum_s (X_{1s} - \bar{X}_1)^2}{n - 1} + \frac{\sum_s (X_{2s} - \bar{X}_2)^2}{n - 1} + \frac{\sum_s (X_{3s} - \bar{X}_3)^2}{n - 1} + \frac{\sum_s (X_{4s} - \bar{X}_4)^2}{n - 1} \right]$$

■ TABLE 9-2. Notation for One-Way Analysis of Variance in Table 9-1

	Treatment			
	1	2	3	4
	X_{11}	X_{21}	X_{31}	X_{41}
	X_{12}	X_{22}	X_{32}	X_{42}
	X_{13}	X_{23}	X_{33}	X_{43}
	X_{14}	X_{24}	X_{34}	X_{44}
	X_{15}	X_{25}	X_{35}	X_{45}
	X_{16}	X_{26}	X_{36}	X_{46}
	X_{17}	X_{27}	X_{37}	X_{47}
Treatment (column) means	\bar{X}_1	\bar{X}_2	\bar{X}_3	\bar{X}_4
Treatment (column) sums of squares	$\sum_s (X_{1s} - \bar{X}_1)^2$	$\sum_s (X_{2s} - \bar{X}_2)^2$	$\sum_s (X_{3s} - \bar{X}_3)^2$	$\sum_s (X_{4s} - \bar{X}_4)^2$
Grand mean = \bar{X}	Total sum of squares = $\sum_t \sum_s (X_{ts} - \bar{X})^2$			

or

$$s_{\text{wit}}^2 = \frac{1}{4} \left(\frac{SS_1}{n-1} + \frac{SS_2}{n-1} + \frac{SS_3}{n-1} + \frac{SS_4}{n-1} \right)$$

in which $n = 7$ represents the size of each sample group. Factor $n - 1$ out of the four expressions for variance computed from within each of the four separate treatment groups, and let $m = 4$ represent the number of treatments (diets), to obtain

$$s_{\text{wit}}^2 = \frac{1}{m} \frac{SS_1 + SS_2 + SS_3 + SS_4}{n-1}$$

The numerator of this fraction is just the total of the sums of squared deviations of the observations about the means of their respective treatment groups. Call it the *within treatments (or within groups) sum of squares* SS_{wit} . Note that the within treatments sum of squares is a measure of variability in the observations that is independent of whether or not the mean responses to the different treatments are the same.

For the data from our diet experiment in Table 9-1

$$SS_{\text{wit}} = .597 + .734 + 1.294 + 1.200 = 3.825(\text{L/min})^2$$

Given our definition of SS_{wit} and the equation s_{wit}^2 above, we can write

$$s_{\text{wit}}^2 = \frac{SS_{\text{wit}}}{m(n-1)}$$

s_{wit}^2 appears in the denominator of the F ratio associated with $v_d = m(n-1)$ degrees of freedom. Using this notation for analysis of variance, degrees of freedom are often denoted by DF rather than v , so let us replace $m(n-1)$ with DF_{wit} in the equation for s_{wit}^2 to obtain

$$s_{\text{wit}}^2 = \frac{SS_{\text{wit}}}{DF_{\text{wit}}}$$

For the diet experiment, $DF_{\text{wit}} = m(n-1) = 4(7-1) = 24$ degrees of freedom.

Finally, recall that in Chapter 2 we defined the variance as the “average” squared deviation from the mean. In this spirit, statisticians call the ratio $SS_{\text{wit}}/DF_{\text{wit}}$ the within groups *mean square* and denote it MS_{wit} . This notation is clumsy, since $SS_{\text{wit}}/DF_{\text{wit}}$ is not really a mean in the standard statistical meaning of the word and it obscures the fact that MS_{wit} is the estimate of the variance computed from within

the groups (that we have been denoting s_{wit}^2). Nevertheless, it is so ubiquitous that we will adopt it. Therefore, we will estimate the variance from within the sample groups with

$$MS_{\text{wit}} = \frac{SS_{\text{wit}}}{DF_{\text{wit}}}$$

We will replace s_{wit}^2 in the definition of F with this expression.

For the data in Table 9-1,

$$MS_{\text{wit}} = \frac{3.825}{24} = 0.159(\text{L/min})^2$$

Next, we need to do the same thing for the variance estimated from between the treatment groups. Recall that we estimated this variance by computing the standard deviation of the sample means as an estimate of the standard error of the mean, then estimated the population variance with

$$s_{\text{bet}}^2 = ns_{\bar{X}}^2$$

The square of the standard deviation of treatment means is

$$s_{\bar{X}}^2 = \frac{(\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2 + (\bar{X}_3 - \bar{X})^2 + (\bar{X}_4 - \bar{X})^2}{m-1}$$

in which m again denotes the number of treatment groups (4) and \bar{X} denotes the means of *all* the observations (which also equals the mean of the sample means when the samples are all the same size). We can write this equation more compactly as

$$s_{\bar{X}}^2 = \frac{\sum (\bar{X}_t - \bar{X})^2}{m-1}$$

so that

$$s_{\text{bet}}^2 = \frac{n \sum (\bar{X}_t - \bar{X})^2}{m-1}$$

(Notice that we are now summing over treatments rather than experimental subjects.)

The between groups variance can be written as the sum of squared deviations of the treatment means about the mean of all observations times the sample size divided by $m - 1$. Denote this sum of squares the *between groups or treatment sum of squares*

$$SS_{\text{bet}} = SS_{\text{treat}} = n \sum (\bar{X}_t - \bar{X})^2$$

The treatment sum of squares is a measure of the variability between the groups, just as the within groups sum of squares is a measure of the variability within the groups.

For the data for the diet experiment in Table 9-1

$$\begin{aligned} SS_{\text{treat}} &= n \sum_t (\bar{X}_t - \bar{X})^2 \\ &= 7[(4.96 - 4.98)^2 + (5.23 - 4.98)^2 + (4.93 - 4.98)^2 \\ &\quad + (4.80 - 4.98)^2] = 0.685(\text{L/min})^2 \end{aligned}$$

The treatment (between groups) variance appears in the numerator of the F ratio and is associated with $\nu = m - 1$ degrees of freedom; we therefore denote $m - 1$ with

$$DF_{\text{bet}} = DF_{\text{treat}} = m - 1$$

in which case

$$s_{\text{bet}}^2 = \frac{SS_{\text{bet}}}{DF_{\text{bet}}} = \frac{SS_{\text{treat}}}{DF_{\text{treat}}}$$

Just as statisticians call the ratio SS_{wit} the within groups mean square, they call the estimate of the variance from between the groups (or treatments) the *between groups (or treatment) mean square* MS_{treat} (or MS_{bet}). Therefore,

$$MS_{\text{bet}} = \frac{SS_{\text{bet}}}{DF_{\text{bet}}} = \frac{SS_{\text{treat}}}{DF_{\text{treat}}} = MS_{\text{treat}}$$

For the data in Table 9-1, $DF_{\text{treat}} = m - 1 = 4 - 1 = 3$, so

$$MS_{\text{bet}} = \frac{0.685}{3} = 0.228(\text{L/min})^2$$

We can write the F test statistic as

$$F = \frac{MS_{\text{bet}}}{MS_{\text{wit}}} = \frac{MS_{\text{treat}}}{MS_{\text{wit}}}$$

and compare it with the critical value of F for numerator degrees of freedom, DF_{treat} (or DF_{bet}), and denominator degrees of freedom, DF_{wit} .

Finally, for the data in Table 9-1

$$F = \frac{MS_{\text{treat}}}{MS_{\text{wit}}} = \frac{.228}{.159} = 1.4$$

the same value of F we obtained from these data in Chapter 3.

We have gone far a field into a computational procedure that is more complex and, on the surface, less intuitive than the one developed in Chapter 3. This approach is necessary, however, to analyze the results obtained in more complex experimental designs. Surprisingly as we will see, there are intuitive meanings which can be attached to these sums of squares and which are very important.

Accounting for All the Variability in the Observations

The sums of squares within and between the treatment groups, SS_{wit} and SS_{treat} , quantify the variability observed within and between the treatment groups. In addition, it is possible to describe the total variability observed in the data by computing the *sum of squared deviations of all observations about the grand mean \bar{X} of all the observations*, called the *total sum of squares*

$$SS_{\text{tot}} = \sum_t \sum_s (X_{ts} - \bar{X})^2$$

The two summation symbols indicate the sums over all subjects in all treatment groups. The total number of degrees of freedom associated with this sum of squares is $DF_{\text{tot}} = mn - 1$, or 1 less than the total sample size (m treatment groups times n subjects in each treatment group). For the observations in Table 9-1, $SS_{\text{tot}} = 4.507(\text{L/min})^2$ and $DF_{\text{tot}} = 4(7) - 1 = 27$.

Notice that the variance estimated from all the observations, without regard for the fact that there are different experimental groups, is just

$$\frac{\sum_t \sum_s (X_{ts} - \bar{X})^2}{mn - 1} = \frac{SS_{\text{tot}}}{mn - 1}$$

The three sums of squares discussed so far are related in a very simple way:

The total sum of squares is the sum of the between groups (treatment) sum of squares and the within groups sum of squares.

$$SS_{\text{tot}} = SS_{\text{bet}} + SS_{\text{wit}}$$

In other words, the total variability, quantified with appropriate sums of squared deviations, can be *partitioned* into two components, one due to variability between the experimental groups and another component

due to variability within the groups.* It is common to summarize all these computations in an *analysis of variance table* such as Table 9-3. Notice that the between groups and within groups sums of squares do indeed add up to the total sum of squares.

F is the ratio of MS_{bet} over MS_{wit} and should be compared with the critical value of F with DF_{bet} and DF_{wit} degrees of freedom for the numerator and denominator, respectively, to test the hypothesis that all the samples were drawn from a single population.

*To see why this is true, first decompose the amount that any given observation deviates from the grand mean, $\bar{X}_{is} - \bar{X}$, into two components, the deviation of the treatment group mean from the grand mean and the deviation of the observation from the mean of its treatment group.

$$(X_{is} - \bar{X}) = (\bar{X}_t - \bar{X}) + (X_{is} - \bar{X}_t)$$

Square both sides

$$(X_{is} - \bar{X})^2 = (\bar{X}_t - \bar{X})^2 + (X_{is} - \bar{X}_t)^2 + 2(\bar{X}_t - \bar{X})(X_{is} - \bar{X}_t)$$

and sum over all observations to obtain the total sum of squares

$$\begin{aligned} SS_{tot} &= \sum_t \sum_s (X_{is} - \bar{X})^2 \\ &= \sum_t \sum_s (\bar{X}_t - \bar{X})^2 + \sum_t \sum_s (X_{is} - \bar{X}_t)^2 + \sum_t \sum_s 2(\bar{X}_t - \bar{X})(X_{is} - \bar{X}_t) \end{aligned}$$

Since $(\bar{X}_t - \bar{X})$ does not depend on which of the n individuals in each sample are being summed over,

$$\sum_s (X_{is} - \bar{X})^2 = n(\bar{X}_t - \bar{X})^2$$

The first term on the right of the equals sign can be written as

$$\sum_t \sum_s (\bar{X}_t - \bar{X})^2 = n \sum_t (\bar{X}_t - \bar{X})^2$$

which is just SS_{bet} . Furthermore, the second term on the right of the equals sign is just SS_{wit} .

It only remains to show that the third term on the right of the equals sign equals zero. To do this, note again that $\bar{X}_t - \bar{X}$ does not depend on which member of each sample is being summed, so we can factor it out of the sum over the member of each sample, in which case

$$\sum_t \sum_s 2(\bar{X}_t - \bar{X})(X_{is} - \bar{X}_t) = 2 \sum_t (\bar{X}_t - \bar{X}) \sum_s (X_{is} - \bar{X}_t)$$

But \bar{X}_t is the mean of the n members of treatment group t , so

$$\begin{aligned} \sum_s (X_{is} - \bar{X}_t) &= \sum_s X_{is} - \sum_s \bar{X}_t = \sum_s X_{is} - n\bar{X}_t \\ &= n \left(\sum_s X_{is} / n - \bar{X}_t \right) = n(\bar{X}_t - \bar{X}_t) = 0 \end{aligned}$$

Therefore,

$$SS_{tot} = SS_{bet} + SS_{wit} + 0 = SS_{bet} + SS_{wit}$$

■ TABLE 9-3. Analysis of Variance Table for the Diet Experiment

Source of Variation	SS	DF	MS
Between groups	0.685	3	0.228
Within groups	<u>3.822</u>	<u>24</u>	0.159
Total	4.507	27	

$$F = \frac{MS_{bet}}{MS_{wit}} = \frac{0.228}{0.159} = 1.4$$

Note also that the treatment and within groups degrees of freedom also add up to the total number of degrees of freedom. This is not a chance occurrence; it will always be the case. Specifically, if there are m experimental groups with n members each,

$$DF_{bet} = m - 1; DF_{wit} = m(n - 1); DF_{tot} = mn - 1$$

so that

$$\begin{aligned} DF_{bet} + DF_{wit} &= (m - 1) + m(n - 1) \\ &= m - 1 + mn - m = mn - 1 = DF_{tot} \end{aligned}$$

In other words, just as it was possible to partition the total sum of squares into components due to between group (treatment) and within group variability, it is possible to partition the degrees of freedom. Figure 9-4 illustrates how the sums of squares and degrees of freedom are partitioned in this analysis of variance.

Now we are ready to attack the original problem, that of developing an analysis of variance suitable for experiments in which each experimental subject receives more than one treatment.

■ EXPERIMENTS WHEN SUBJECTS ARE OBSERVED AFTER MANY TREATMENTS: REPEATED MEASURES ANALYSIS OF VARIANCE

When each experimental subject receives more than one treatment it is possible to partition the total variability in the observations into three mutually exclusive components: variability between all the experimental subjects, the variability due to the treatments, and the variability within the subjects' response to the treatments. The last component of variability represents the fact that there is some random variation in how a given individual responds to a given treatment as well as measurement errors. Figure 9-5 shows

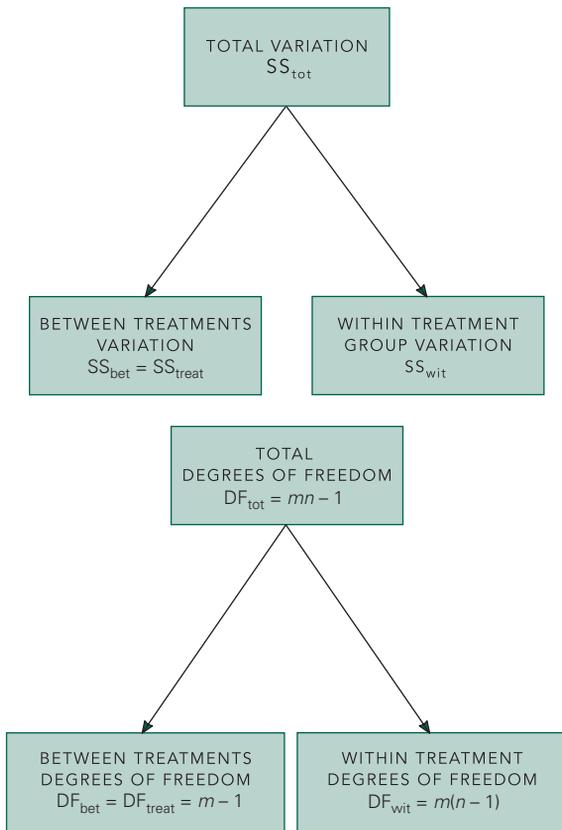


FIGURE 9-4. Partitioning of the sums of squares and degrees of freedom for a one-way analysis of variance.

this breakdown. The resulting procedure is called a *repeated measures* analysis of variance because the measurements are repeated under all the different experimental conditions (treatment) in each of the experimental subjects.*

*This chapter discusses *one-way* repeated measures analysis of variance, the simplest case. Just as the one-way analysis of variance could be generalized to two-way (and higher) analysis of variance, the one-way repeated measures analysis of variance presented in this chapter forms the basis for more general two-way (and higher) repeated measures analyses of variance. Two-way repeated measures analyses of variance are common in biomedical research in areas such as drug trials, for example, when one factor is the presence or absence of drug (and time after administering the drug) as one factor and gender as another factor. As in non-repeated measures analysis of variance, one can then test for the effects of the drug controlling for gender, gender controlling for the drug, and the drug \times gender interaction (when different genders react differently to the drug). For a detailed discussion of two-way repeated measures designs, see Glantz SA, Slinker B. *Primer of Applied Regression and Analysis of Variance*, 2nd ed. New York: McGraw-Hill; 2001 for details on how to analyze two-way and higher-order repeated measures analyses of variance.

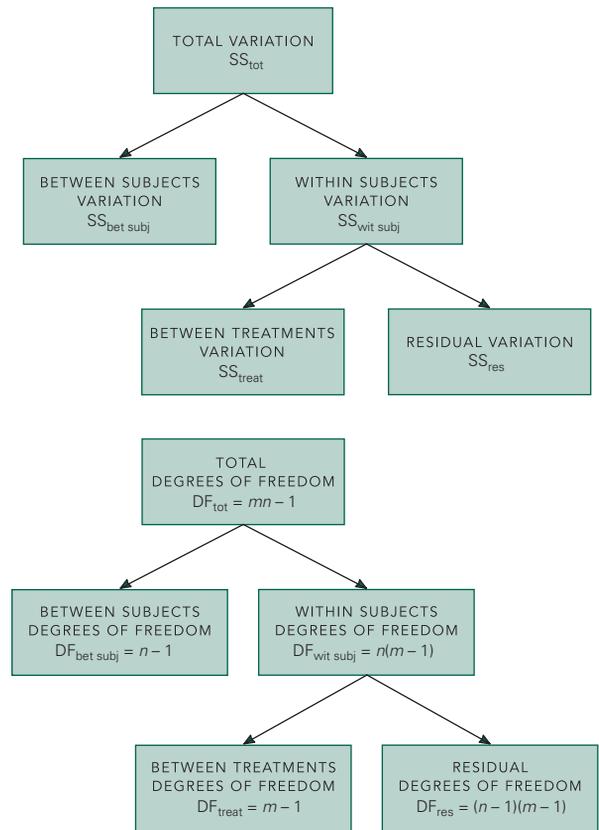


FIGURE 9-5. Partitioning of the sums of squares and degrees of freedom for a one-way repeated measures analysis of variance. Notice that this procedure allows us to concentrate on the variation within experimental subjects.

Now, let us write expressions for these three kinds of variability. As Figure 9-5 suggests, the first step is to divide the total variability into variability within subjects and between subjects.

Table 9-4 illustrates the notation we will use for repeated measures analysis of variance. (In this case, it is for an experiment in which four experimental subjects each receive three different treatments). At first glance, this table appears quite similar to Table 9-2, used to analyze experiments in which *different* subjects received each of the treatments. There is one important difference: in Table 9-4 the *same* subjects receive all the treatments. For example, X_{11} represents how the first experimental subject responded to the first treatment; X_{21} represents how the (same) first experimental subject responded to the second

■ TABLE 9-4. Notation for Repeated Measures Analysis of Variance

Experimental Subject, $n = 4$	Treatment, $m = 3$			Subject	
	1	2	3	Mean	SS
1	X_{11}	X_{21}	X_{31}	\bar{S}_1	$\sum_t (X_{t1} - \bar{S}_1)^2$
2	X_{12}	X_{22}	X_{32}	\bar{S}_2	$\sum_t (X_{t2} - \bar{S}_2)^2$
3	X_{13}	X_{23}	X_{33}	\bar{S}_3	$\sum_t (X_{t3} - \bar{S}_3)^2$
4	X_{14}	X_{24}	X_{34}	\bar{S}_4	$\sum_t (X_{t4} - \bar{S}_4)^2$
Treatment mean	\bar{T}_1	\bar{T}_2	\bar{T}_3		
Grand mean $\bar{X} = \frac{\sum_t \sum_s X_{ts}}{mn}$					$SS_{tot} = \sum_t \sum_s (X_{ts} - \bar{X})^2$

treatment. In general, X_{ts} is the response of the s th experimental subject to the t th treatment.

$\bar{S}_1, \bar{S}_2, \bar{S}_3,$ and \bar{S}_4 are the mean responses of each of the four subjects to all (three) treatments

$$\bar{S}_s = \frac{\sum_t X_{ts}}{m}$$

in which there are $m = 3$ treatments. Likewise, $\bar{T}_1, \bar{T}_2,$ and \bar{T}_3 are the mean responses to each of the three treatments of all (four) experimental subjects.

$$\bar{T}_t = \frac{\sum_s X_{ts}}{n}$$

in which there are $n = 4$ experimental subjects.

As in all analyses of variance, we quantify the total variation with the total sum of squared deviations of all observations about the grand mean. The grand mean of all the observations is

$$\bar{X} = \frac{\sum_t \sum_s X_{ts}}{mn}$$

and the total sum of squared deviations from the grand mean is

$$SS_{tot} = \sum_t \sum_s (X_{ts} - \bar{X})^2$$

This sum of squares is associated with $DF_{tot} = mn - 1$ degrees of freedom.

Next, we partition this total sum of squares into variation *within subjects* and variation *between subjects*. The variation of observations within subject 1 about the mean observed for subject 1, \bar{S}_1 , is

$$SS_{wit\ subj\ 1} = \sum_t (X_{t1} - \bar{S}_1)^2$$

Likewise, the variation in observations within subject 2 about the mean observed in subject 2 is

$$SS_{wit\ subj\ 2} = \sum_t (X_{t2} - \bar{S}_2)^2$$

We can write similar sums for the other two experimental subjects. The total variability observed within all subjects is just the sum of the variability observed within each subject

$$\begin{aligned} SS_{wit\ subjs} &= SS_{wit\ subj\ 1} + SS_{wit\ subj\ 2} + SS_{wit\ subj\ 3} + SS_{wit\ subj\ 4} \\ &= \sum_t \sum_s (X_{ts} - \bar{S}_s)^2 \end{aligned}$$

Since the sum of squares within each subject is associated with $m - 1$ degrees of freedom (where m is the number of treatments) and there are n subjects, $SS_{wit\ subjs}$ is associated with $DF_{wit\ subjs} = n(m - 1)$ degrees of freedom.

The variation between subjects is quantified by computing the sum of squared deviations of the mean response of each subject about the grand mean

$$SS_{\text{bet subjs}} = m \sum_t (\bar{S}_s - \bar{X})^2$$

The sum is multiplied by m because each subject's mean is the mean response to the m treatments. (This situation is analogous to the computation of the between-groups sum of squares as the sum of squared deviations of the sample means about the grand mean in the analysis of variance developed in the last section.) This sum of squares has $DF_{\text{bet subjs}} = n - 1$ degrees of freedom.

It is possible to show that

$$SS_{\text{tot}} = SS_{\text{wit subjs}} + SS_{\text{bet subjs}}$$

that is, that the total sum of squares can be partitioned into the within and between subjects sums of squares.*

Next, we need to partition the within subjects sum of squares into two components, variability in the observations due to the *treatments* and the *residual* variation due to random variation in how each individual responds to each treatment. The sum of squares due to the treatments is the sum of squared differences between the treatment means and the grand mean:

$$SS_{\text{treat}} = n \sum_t (\bar{T}_t - \bar{X})^2$$

We multiply by n , the number of subjects used to compute each treatment mean, just as we did above when computing the between-subjects sum of squares. Since there are m different treatments, there are $DF_{\text{treat}} = m - 1$ degrees of freedom associated with SS_{treat} .

Since we are partitioning the within subjects sum of squares into the treatment sum of squares and the residual sum of squares,

$$SS_{\text{wit subjs}} = SS_{\text{treat}} + SS_{\text{res}}$$

and so

$$SS_{\text{res}} = SS_{\text{wit subjs}} - SS_{\text{treat}}$$

The same partitioning for the degrees of freedom yields

$$\begin{aligned} DF_{\text{res}} &= DF_{\text{wit subjs}} - DF_{\text{treat}} \\ &= n(m-1) - (m-1) = (n-1)(m-1) \end{aligned}$$

Finally, our estimate of the population variance from the treatment sum of squares is

$$MS_{\text{treat}} = \frac{SS_{\text{treat}}}{DF_{\text{treat}}}$$

and the estimate of the population variance from the residual sum of squares is

$$MS_{\text{res}} = \frac{SS_{\text{res}}}{DF_{\text{res}}}$$

If the null hypothesis that the treatments have no effect is true, MS_{treat} and MS_{res} are both estimates of the same (unknown) population variance, so compute

$$F = \frac{MS_{\text{treat}}}{MS_{\text{res}}}$$

to test the null hypothesis that the treatments do not change the experimental subjects. If the hypothesis of no treatment effect is true, this F ratio will follow the F distribution with DF_{treat} numerator degrees of freedom and DF_{res} denominator degrees of freedom.

This development has been, by necessity, more mathematical than most of the explanations in this book. Let us apply it to a simple example to make the concepts more concrete.

Anti-asthmatic Drugs and Endotoxins

Endotoxin is a component of Gram-negative bacteria found in dust in both workplaces and homes. Inhaling endotoxin causes fever, chills, bronchoconstriction of airways in the lungs, and generalized bronchial hyper-responsiveness (wheezing). Prolonged endotoxin exposure is associated with chronic obstructive pulmonary disease and asthma. Olivier Michel and colleagues[†] thought that the anti-asthmatic drug, salbutamol, might

*For a derivation of this equation, see Winer BJ, Brown DR, Michels KM. Single-factor experiments having repeated measures on the same elements. *Statistical Principles in Experimental Design*, 3ed ed. New York: McGraw-Hill; 1991:chap 4.

[†]Michel O, Olbrecht J, Moulard D, Sergysels R. Effect of anti-asthmatic drugs on the response to inhaled endotoxin. *Ann. Allergy Asthma Immunol.* 2000;85:305–310.

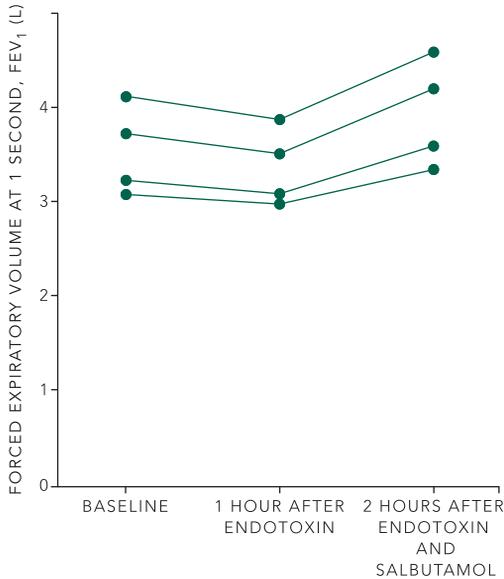


FIGURE 9-6. Forced expiratory volume at 1 second (FEV₁) in four people at baseline, 1 hour after inhaling endotoxin, and 2 hours following endotoxin and salbutamol exposure. Each individual's response is connected by straight lines. (Adapted from Table 2 and Fig. 4 of Michel O, Olbrecht J, Moulard D, Sergysels R. Effect of anti-asthmatic drugs on the response to inhaled endotoxin. *Ann Allergy Asthma Immunol.* 2000;85:305–310.)

protect against the endotoxin-induced inflammation that produces these symptoms. To test this hypothesis, they had four mildly asthmatic people breathe an aerosol containing a purified form of endotoxin and measured how many liters of air they could exhale in 1 second. This variable, known as forced expiratory volume in 1 second or FEV₁, is a measure of airway constriction. A decrease in FEV₁ indicates a higher degree of bronchoconstriction. They took three FEV₁ measurements in each person: baseline (before breathing the endotoxin), 1 hour following endotoxin inhalation, and 2 hours after each subject received an additional salbutamol treatment.

Figure 9-6 shows the results of this experiment. Simply looking at Figure 9-6 suggests that salbutamol increases FEV₁, but there are only four people in the study. How confident can we be when asserting that the drug actually reduces bronchial constriction and makes it easier to breathe? To answer this question, we perform a repeated measures analysis of variance.

Table 9-5 shows the same data as Figure 9-6, together with the mean FEV₁ observed for each of the $n = 4$ experimental subjects (people) and each of the $m = 3$ treatments (baseline, 1 hour, and 2 hours). For example, subject 2's mean response of subject 2 to all three treatments is

$$\bar{S}_2 = \frac{4.0 + 3.7 + 4.4}{3} = 4.03 \text{ L}$$

and the mean response of all four subjects to treatment 1 (baseline) is

$$\bar{T}_1 = \frac{3.70 + 4.03 + 3.0 + 3.17}{4} = 3.48 \text{ L}$$

The grand mean of all observations is $\bar{X} = 3.48 \text{ L}$ and the total sum of squares is $SS_{\text{tot}} = 2.6656 \text{ L}^2$.

Table 9-5 also includes the sum of squares within each subject; for example, for subject 2

$$\begin{aligned} SS_{\text{wit subj } 2} &= (4 - 4.03)^2 + (3.7 - 4.03)^2 + (4.4 - 4.03)^2 \\ &= 0.2467 \text{ L}^2 \end{aligned}$$

Adding the within subjects sums of squares for the four subjects in the study yields

$$\begin{aligned} SS_{\text{wit subjs}} &= 0.1800 + 0.2467 + 0.0800 + 0.1267 \\ &= 0.6334 \text{ L}^2 \end{aligned}$$

We obtain sum of squares between subjects by adding up the squares of the deviations between the subjects' means and the grand mean and multiplying by the numbers of treatments ($m = 3$, the number of observations used to compute each subject's mean response)

$$\begin{aligned} SS_{\text{bet subjs}} &= 3[(3.70 - 3.48)^2 + (4.03 - 3.48)^2 \\ &\quad + (3.00 - 3.48)^2 + (3.17 - 3.48)^2] \\ &= 2.0322 \text{ L}^2 \end{aligned}$$

(Note that $SS_{\text{wit subjs}} + SS_{\text{bet subjs}} = 0.6334 + 2.0322 = 2.6656 \text{ L}^2$, the total sum of squares, as it should.)

We obtain the sum of squares for the treatments by multiplying the squares of the differences between the treatment means and the grand mean times the number of subjects ($n = 4$, the number of numbers used to compute each mean):

$$\begin{aligned} SS_{\text{treat}} &= 4[(3.48 - 3.48)^2 + (3.20 - 3.48)^2 + (3.75 - 3.48)^2] \\ &= 0.6051 \text{ L}^2 \end{aligned}$$

■ **TABLE 9-5. Forced Expiratory Volume (L) at One Second before and after Bronchial Challenge with Endotoxin and Salbutamol Treatment**

Person (Subject)	No Drug (Baseline)	One Hour after Endotoxin	Two Hours after Endotoxin and Salbutamol	Subject	
				Mean	SS
1	3.7	3.4	4.0	3.70	0.1800
2	4.0	3.7	4.4	4.03	0.2467
3	3.0	2.8	3.2	3.00	0.0800
4	3.2	2.9	3.4	3.17	0.1267
Treatment mean	3.48	3.20	3.75		
Grand mean = 3.48					$SS_{\text{tot}} = 2.6656 \text{ L}^2$

There are $DF_{\text{treat}} = m - 1 = 3 - 1 = 2$ degrees of freedom associated with the treatments.

Finally, the residual sum of squares is

$$SS_{\text{res}} = SS_{\text{wit subj}} - SS_{\text{treat}} = 0.6334 - 0.6051 = 0.0283 \text{ L}^2$$

with

$$DF_{\text{res}} = (n - 1)(m - 1) = (4 - 1)(3 - 1) = 6$$

degrees of freedom.

Table 9-6, the analysis of variance table for this experiment, summarizes the results of all these calculations. Notice that we have partitioned the sums of squares into more components than we did in Table 9-3. (Compare these two tables with Figures 9-4 and 9-5.) We are able to do this because we made repeated measurements on the same experimental subjects.

From Table 9-6, our two estimates of the population variance are

$$MS_{\text{treat}} = \frac{SS_{\text{treat}}}{DF_{\text{treat}}} = \frac{0.6051}{2} = 0.3026 \text{ L}^2$$

and

$$MS_{\text{res}} = \frac{SS_{\text{res}}}{DF_{\text{res}}} = \frac{0.0283}{6} = 0.0047 \text{ L}^2$$

so our test statistic is

$$F = \frac{MS_{\text{treat}}}{MS_{\text{res}}} = \frac{0.3026}{0.0047} = 64.38$$

This value exceeds $F_{.01} = 10.92$, the critical value that defines the largest 1% of possible values of F with 2 and 6 degrees of freedom for the numerator and denominator. Therefore, these data permit concluding that endotoxin and salbutamol alter FEV_1 ($P < .01$).

So far we can conclude that at least one of the treatments produced a change. To isolate which one, we need

■ **TABLE 9-6. Analysis of Variance Table for One-Way Repeated Measures Analysis of FEV_1 in Endotoxin Response**

Source of Variation	SS	DF	MS
Between subjects	2.0322	3	
Within subjects	0.6334	8	
Treatments	0.6051	2	0.3026
Residual	0.0283	6	0.0047
Total	2.6656	11	

$$F = \frac{MS_{\text{treat}}}{MS_{\text{res}}} = \frac{0.3026}{0.0047} = 64.038$$

to use a multiple-comparisons procedure analogous to the Holm t test (or Holm-Sidak test or Bonferroni t test) developed in Chapter 4.

How to Isolate Differences in Repeated Measures Analysis of Variance

In Chapter 4 we conducted multiple pairwise comparisons between groups with the Holm-Sidak t test.

$$t = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{s_{\text{wit}}^2}{n_1} + \frac{s_{\text{wit}}^2}{n_2}}}$$

To use the Holm-Sidak t test to isolate differences following a repeated measures analysis of variance, we simply replace s_{wit}^2 with our estimate of the variance computed from the residual sum of squares, MS_{res} :

$$t = \frac{\bar{T}_i - \bar{T}_j}{\sqrt{\frac{MS_{\text{res}}}{n_{j1}} + \frac{MS_{\text{res}}}{n_2}}}$$

in which \bar{T}_i and \bar{T}_j represent the mean treatment responses of the pair of treatments (treatments i and j) you are comparing. The resulting value of t is compared with the critical value for DF_{res} degrees of freedom.

There are three comparisons ($k=3$) for this experiment. To compare FEV₁ 1 hour following endotoxin to FEV₁ 2 hours following endotoxin and salbutamol exposure, compute

$$t = \frac{3.20 - 3.75}{\sqrt{\frac{.0047}{2} + \frac{.0047}{2}}} = -11.346$$

To compare baseline FEV₁ to FEV₁ 2 hours following endotoxin and salbutamol exposure

$$t = \frac{3.48 - 3.75}{\sqrt{\frac{.0047}{2} + \frac{.0047}{2}}} = -5.570$$

Finally, to compare baseline FEV₁ with FEV₁ 1 hour following endotoxin exposure

$$t = \frac{3.48 - 3.20}{\sqrt{\frac{.0047}{2} + \frac{.0047}{2}}} = -5.776$$

There are 6 degrees of freedom for these comparisons. The uncorrected P values corresponding to these three comparisons are less than .0001, .001, and .001.

To keep the overall risk of erroneously reporting a difference for this family of three comparisons below 5%, we compare these P values to the Holm-Sidak t test critical P values based on $k=3$: $P_{\text{crit}} = 1 - (1 - \alpha_T)^{1/(k-j+1)}$, $1 - (1 - .05)^{1/(3-1+1)} = .0170$, $1 - (1 - .05)^{1/(3-2+1)} = .0253$ and $1 - (1 - .05)^{1/(3-3+1)} = .0500$. All three of the uncorrected P values fall below the appropriate critical P . These results allow us to conclude that endotoxin decreases FEV₁ and that subsequent administration of salbutamol reversed this effect, increasing FEV₁ to levels higher than baseline.

Power in Repeated Measures Analysis of Variance

Power is computed exactly as in a simple analysis of variance, using the within subjects variation (estimated by $\sqrt{MS_{\text{res}}}$) as the estimate of population standard deviation, σ , and the number of subjects in place of the sample size of each group, n .

■ EXPERIMENTS WHEN OUTCOMES ARE MEASURED ON A NOMINAL SCALE: McNEMAR'S TEST

The paired t test and repeated measures analysis of variance can be used to analyze experiments in which the variable being studied can be measured on an interval scale (and satisfies the other assumptions required of parametric methods). What about experiments, analogous to the ones in Chapter 5, in which outcomes are measured on a *nominal* scale? This problem often arises when asking whether or not an individual responded to a treatment or when comparing the results of two different diagnostic tests that are classified as positive or negative in the same individuals. We will develop a procedure to analyze such experiments, *McNemar's test for changes*, in the context of one such study.

p7 Antigen Expression in Human Breast Cancer

The p7 antigen has been shown to be expressed in cell lines derived from ovarian cancers but not in cell lines derived from normal tissues. In addition, expression of this antigen has been increased in ovarian cancer cells after treatment with chemotherapeutic agents. Since there are similarities between ovarian and breast cancer, Xiaowei Yang and

colleagues* wanted to study whether this antigen is present in tumor cells from women with breast cancer. They also wanted to investigate how treatment with radiation or chemotherapy affects appearance of p7, since presence of this antigen in a substantial fraction of breast cancer tumor cells has been associated with distant metastases and local recurrences. To investigate whether radiation and chemotherapy affected the expression of p7, they took tissue samples from women with breast cancer before and after they were treated and used several molecular biology techniques to test for the presence of p7.

Table 9-7 shows that four women had p7 both before and after treatment, none had p7 present before but not after treatment, 12 women who did not have p7 present before treatment but had it afterwards, and 14 had p7 neither before nor after treatment.

This table looks very much like the 2×2 contingency tables analyzed in Chapter 5. In fact, most people simply compute a χ^2 statistic from these data and look the P value up in Table 5-7. The numbers in Table 9-7 are associated with a value of $\chi^2 = 2.165$ (computed including the Yates correction for continuity). This value is well below 3.841, the value of χ^2 that defines the largest 5% of possible values of χ^2 with 1 degree of freedom. As a result, one might report “no significant difference” in the expression of p7 before and after treatment of breast cancer and conclude that treatment has no effect on the likelihood of tumor recurrence or metastasis.

There is, however, a serious problem with this approach. The χ^2 test statistic developed for contingency tables in Chapter 5 was used to test the hypothesis that the *rows and columns of the tables are independent*. In Table 9-7, the rows

and columns are *not* independent because they represent the p7 status of *the same individuals* before and after cancer treatment. (This situation is analogous to the difference between the unpaired t test presented in Chapter 4 and the paired t test presented earlier in this chapter.) In particular, the 4 women who were positive for p7 *both* before and after treatment and the 14 who were negative *both* before and after treatment do not tell you anything about whether or not breast cancer tumor cells change expression of p7 in response to radiation or chemotherapy. We need a statistical procedure that focuses on the 12 women who were negative before treatment and positive after treatment and the fact that there were no women positive before and negative after treatment.

If there was no effect of the treatment on p7 expression, we would expect half the $0 + 12 = 12$ women whose p7 status condition before and after treatment was different. In particular, we would expect $12/2 = 6$ to have been positive before treatment but not after and 6 to have been negative before but positive after treatment. Table 9-7 shows that the observed number of women who fell into each of these two categories was 0 and 12, respectively. To compare these observed and expected frequencies, we can use the χ^2 test statistic to compare these observed frequencies with the expected frequency of $12/2 = 6$.

$$\begin{aligned}\chi^2 &= \sum \frac{(|0 - E| - 1/2)^2}{E} \\ &= \frac{(|0 - 6| - 1/2)^2}{6} + \frac{(|12 - 6| - 1/2)^2}{6} = 10.083\end{aligned}$$

Notice that this computation of χ^2 includes the Yates correction for continuity because it has only 1 degree of freedom.

This value exceeds 7.879, the value of χ^2 that defines the biggest 0.5% of the possible values of χ^2 with 1 degree of freedom (from Table 5-7) if the differences in observed and expected were simply the effects of random sampling. This analysis leads to the conclusion that there *is* a difference in expression of p7 in breast cancer tumor cells after women are treated with radiation and chemotherapy ($P < .005$). This conclusion could have implications for the prognosis of these women as well as making p7 a potential target for antibody-based treatments or other types of target-based approaches.

This example illustrates that it is entirely possible to compute values of test statistics and look up P values in tables that are meaningless when the experimental

*Yang X, Groshen S, Formenti SC, Davidson N, Press MF. P7 antigen expression in human breast cancer. *Clin Cancer Res.* 2003;9:201–206.

■ **TABLE 9-7. Presence of p7 Antigen in Breast Cancer Tumor Cells Before and After Women are Treated with Radiation and Chemotherapy**

Before	After	
	Positive	Negative
Positive	4	0
Negative	12	14

design and underlying populations are not compatible with the assumptions used to derive the statistical procedure.

In sum, McNemar's test for changes consists of the following procedure:

- Ignore individuals who responded the same way to both treatments.
- Compute the total number of individuals who responded differently to the two treatments.
- Compute the expected number of individuals who would have responded positively to each of the two treatments (but not the other) as half the total number of individuals who responded differently to the two treatments.
- Compare the observed and expected number of individuals that responded to one of the treatments by computing a χ^2 test statistic (including Yates correction for continuity).
- Compare this value of χ^2 with the critical values of the χ^2 distribution with 1 degree of freedom.

This procedure yields a *P* value that quantifies the probability that the differences in treatment response are due to chance rather than actual differences in how the two treatments affect the same individuals.

■ PROBLEMS

9-1 Several epidemiological studies have shown that people who have a diet high in flavenols (which are in tea, wine, cocoa products and various fruits) have lower rates of dying from coronary artery disease. To investigate whether this effect of flavinols is mediated, at least in part, by beneficial effects on the lining of arteries known as the vascular endothelium, Christian Heiss and colleagues* recorded how much arteries expanded (dilated) in response to increases in the need for blood flow, a measure of endothelial health, in healthy people before and after one month on a diet high in flavenols (see Table 9-8). Higher values of this so-called flow-mediated dilation (FMD) indicates healthier endothelium. Did the diet lead to changes in the level of flow-mediated dilation?

*Heiss C, et al. Improvement of endothelial function with dietary flavinols is associated with mobilization of circulating angiogenic cells in patients with coronary artery disease. *J Am Coll Cardiol.* 2010;56:218–224.

■ TABLE 9-8. Flow Mediated Dilation

Person	Before Diet	After Diet
1	3.0	5.0
2	3.7	4.0
3	5.0	3.8
4	7.2	9.5
5	5.0	8.1
6	3.3	6.1
7	4.8	6.1
8	3.3	4.9
9	3.8	4.9
10	2.2	4.0
11	4.0	4.1
12	4.0	5.0
13	7.0	7.2
14	7.2	7.3

9-2 Secondhand tobacco smoke increases the risk of a heart attack. In order to investigate the mechanisms for this effect, C. Arden Pope III and his colleagues[†] studied whether breathing secondhand smoke affected autonomic (reflex) nervous system control of the heart. At rest the heart beats regularly, about once a second, but there are small beat-to-beat random fluctuations of the order of 100 milliseconds (.1 second) superimposed on the regular interval between heartbeats. This random fluctuation in the length of time between heartbeats is known as heart rate variability and quantified as the standard deviation of interbeat intervals over many beats. For reasons that are not fully understood, reductions in this heart rate variability are associated with increased risk of an acute heart attack. Pope and his colleagues measured heart rate variability in eight healthy young adults before and after they spent 2 hours sitting in the smoking lounge at the Salt Lake City Airport. Table 9-9 shows the observations on the standard deviation of the length of time between beats (in millisecond) measured over the 2 hours before and immediately after sitting in the smoking lounge. Did sitting in the smoking lounge reduce heart rate variability?

[†]Pope CA III, et al. Acute exposure to environmental tobacco smoke and heart rate variability. *Environ Health Perspect.* 2001;109:711–716.

■ **TABLE 9-9. Heart Rate Variability Before and After Spending Two Hours in a Smoking Lounge**

Experimental Subject	Standard Deviation in Beat-to-Beat Period (ms)	
	Before	After
Tom	135	105
Dick	118	95
Harry	98	80
Lev	95	73
Joaquin	87	70
Stan	75	60
Aaron	69	68
Ben	59	40

9-3 What are the chances of detecting a halving of the heart rate variability in Problem 9-2 with 95% confidence? Note that the power chart in Figure 6-9 also applies to the paired t test.

9-4 Rework Problem 9-2 as a repeated measures analysis of variance. What is the arithmetic relationship between F and t ?

9-5 In addition to measuring FEV_1 (the experiment described in conjunction with Fig. 9-6), Michel and colleagues took measurements of immune response in their subjects, including measuring the amount of C-reactive protein (CRP), a protein that is elevated when tissue is inflamed. Their results are shown in Table 9-10. Did endotoxin by itself or the combination of endotoxin and salbutamol affect CRP levels? If so, are the effects the same 1 and 2 hours after giving the bronchial challenge?

9-6 In general, levels of the hormone testosterone decrease during periods of stress. Because physical and psychological stressors are inevitable in the life of soldiers, the military is very interested in assessing stress response in soldiers. Many studies addressing this issue suffer from taking place in a laboratory setting, which may not accurately reflect the real world stresses on a soldier. To investigate the effects of stress on testosterone levels in a more realistic setting, Charles Morgan and colleagues* measured salivary testosterone levels in 12 men before and during a military training exercise. The exercise included simulated capture and interrogation modeled on American prisoner of war experiences during the Vietnam and Korean wars. Table 9-11 shows their data. What conclusions can be drawn from these observations?

9-7 What is the power of the test in Problem 9-7 to find a 100 mL change in food intake with 95% confidence?

9-8 In the fetus, there is a connection between the aorta and the artery going to the lungs called the ductus arteriosus that permits the heart to bypass the nonfunctioning lungs and circulate blood to the placenta to obtain oxygen and nourishment and dispose of wastes. After the infant is born and begins breathing, these functions are served by the lungs and the ductus arteriosus closes. Occasionally, especially in premature infants, the ductus arteriosus remains open and shunts blood around the lungs. This shunting prevents the infant from getting rid of carbon dioxide and taking in oxygen. The drug indomethacin has been used to make the ductus arteriosus close. It is very

*Morgan C, et al. Hormone profiles in humans experiencing military survival training. *Biol Psychiatry*. 2000;47:89–1901.

■ **TABLE 9-10. Effect of Endotoxin Exposure on C Reactive Protein (CRP)**

Person (Subject)	CRP (mg/dL)		
	No Drug (Baseline)	One Hour after Endotoxin	Two Hours after Endotoxin and Salbutamol
1	0.60	0.47	0.49
2	0.52	0.39	0.73
3	1.04	0.83	0.47
4	0.87	1.31	0.71

■ **TABLE 9-11. Testosterone Levels During Military Training for Capture and Interrogation**

Soldier	Testosterone (ng/dL)			
	Beginning of Training Exercise	Time of Capture	12 Hours Postcapture	48 Hours Postcapture
1	17.4	11.2	12.8	5.9
2	13.6	6.9	9.8	7.4
3	17.3	12.8	13.7	9.0
4	20.1	16.6	15.5	15.7
5	21.1	13.5	15.4	11.0
6	12.4	2.9	3.7	3.4
7	13.8	7.9	10.5	7.8
8	17.7	12.5	14.9	13.1
9	8.1	2.6	2.3	1.3
10	16.3	9.2	9.3	7.3
11	9.2	2.9	5.8	5.5
12	22.1	17.5	15.3	9.3

likely that the outcome (with or without drugs) depends on gestational age, age after birth, fluid intake, other illnesses, and other drugs the infant is receiving. For these reasons, an investigator might decide to pair infants who are as alike as possible in each of these identified variables, and randomly treat one member of each pair with indomethacin or placebo, then judge the results as improved or not improved. The findings are as shown in Table 9-12. Do these data support the hypothesis that indomethacin is no better than a placebo?

■ **TABLE 9-12. Indomethacin and Closure of Ductus Arteriosus**

		Indomethacin	
		Improved	Not Improved
Placebo	Improved	65	13
	Not improved	27	40

9-9 The data in Problem 9-8 could also be presented in the following form as shown in Table 9-13. How would these data be analyzed? If this result differs from the analysis in Problem 9-9, explain why and decide which approach is correct.

9-10 Review all original articles published in the *New England Journal of Medicine* during the last 12 months. How many of these articles present the results of experiments that should be analyzed with a repeated measures analysis of variance? What percentage of these articles actually did such an analysis? Of those that did not, how did the authors analyze their data? Comment on potential difficulties with the conclusions that are advanced in these papers.

■ **TABLE 9-13. Indomethacin and Closure of Ductus Arteriosus (Alternative Presentation of Results)**

	Improved	Not Improved
Indomethacin	92	53
Placebo	78	67

Alternatives to Analysis of Variance and the t test Based on Ranks

Analysis of variance, including the t tests, is widely used to test the hypothesis that one or more treatments had no effect on the mean of some observed variable. All forms of analysis of variance, including the t tests, are based on the assumption that the observations are drawn from normally distributed populations in which the variances are the same even if the treatments change the mean responses. These assumptions are often satisfied well enough to make analysis of variance an extremely useful statistical procedure. On the other hand, experiments often yield data that are not compatible with these assumptions. In addition, there are often problems in which the observations are measured on an *ordinal scale* rather than an interval scale and may not be amenable to an analysis of variance. This chapter develops analogs to the t tests and analysis of variance based on *ranks* of the observations rather than the observations themselves. This approach uses information about the relative sizes of the observations without assuming anything about the specific nature of the population they were drawn from.

We will begin with the nonparametric analog to the unpaired and paired t tests, the *Mann-Whitney rank-sum test*, and *Wilcoxon signed-rank test*. Then we will present the analogs of one-way analysis of variance, the *Kruskal-Wallis analysis of variance based on ranks*, and the *Friedman repeated measures analysis of variance based on ranks*.

■ HOW TO CHOOSE BETWEEN PARAMETRIC AND NONPARAMETRIC METHODS

As already noted, analysis of variance is called a *parametric* statistical method because it is based on estimates of the two population parameters, the mean and standard deviation (or variance), that completely define a normal distribution. Given the assumption that the samples are drawn from normally distributed populations, one can compute the distributions of the F or t test statistics that will occur in all possible experiments of a given size when the treatments have no effect. The critical values that define a value of F or t can then be obtained from that distribution. When the assumptions of parametric statistical methods are satisfied, they are the most powerful tests available.

If the populations the observations were drawn from are not normally distributed (or are not reasonably compatible with other assumptions of a parametric method, such as equal variances in all the treatment groups), parametric methods become quite unreliable because the mean and standard deviation, the key elements of parametric statistics, no longer completely describe the population. In fact, when the population substantially deviates from normality, interpreting the mean and standard deviation in terms of a normal distribution can produce a very misleading picture.

For example, recall our discussion of the distribution of heights of the entire population of Jupiter. The mean height

of all Jovians is 37.6 cm in Figure 2-3A and the standard deviation is 4.5 cm. Rather than being equally distributed about the mean, the population is *skewed* toward taller heights. Specifically, the heights of Jovians range from 31 to 52 cm, with most heights around 35 cm. Figure 2-3B shows what the population of heights would have been if, instead of being skewed toward taller heights, they had been normally distributed with the same mean and standard deviation as the actual population (in Figure 2-3A). The heights would have ranged from 26 to 49 cm, with most heights around 37 to 38 cm. Simply looking at Figure 2-3 should convince you that envisioning a population on the basis of the mean and standard deviation can be quite misleading if the population does not, at least approximately, follow the normal distribution.

The same thing is true of statistical tests that are based on the normal distribution. When the population the samples were drawn from does not at least approximately follow the normal distribution, these tests can be quite misleading. In such cases, it is possible to use the *ranks* of the observations rather than the observations themselves to compute statistics that can be used to test hypotheses. By using ranks rather than the actual measurements it is possible to retain much of the information about the relative size of responses without making any assumptions about how the population the samples were drawn from is distributed. Since these tests are not based on the parameters of the underlying population, they are called *nonparametric* or *distribution-free* methods.* All the methods we will discuss require only that the distributions under the different treatments have similar shapes, but there is no restriction on what those shapes are.†

When the observations are drawn from normally distributed populations, the nonparametric methods in this chapter are about 95% as powerful as the analogous parametric methods. As a result, power for these tests can be estimated by computing the power of the analogous para-

metric test. When the observations drawn from populations that are not normally distributed, nonparametric methods are not only more reliable but also more powerful than parametric methods.

Unfortunately, you can never observe the entire population. So how can you tell whether the assumptions such as normality are met, to permit using the parametric tests such as analysis of variance? The simplest approach is to plot the observations and look at them. Do they seem compatible with the assumptions that they were drawn from normally distributed populations with roughly the same variances, that is, within a factor of 2 to 3 of each other? If so, you are probably safe in using parametric methods. If, on the other hand, the observations are heavily skewed (suggesting a population such as the Jovians in Fig. 2-3A) or appear to have more than one peak, you probably will want to use a nonparametric method. When the standard deviation is about the same size or larger than the mean and the variable can take on only positive values, this is an indication that the distribution is skewed. (A normally distributed variable would have to take on negative values.) In practice, these simple rules of thumb are often all you will need.

There are two ways to make this procedure more objective. The first is to plot the observations as a *normal probability plot*. A normal probability plot has a distorted vertical scale that makes normally distributed observations plot as a straight line (just as exponential functions plot as a straight line on a semilogarithmic graph). Examining how straight the line is will show how compatible the observations are with a normal distribution. One can also construct a χ^2 statistic to test how closely the observed data agree with those expected if the population is normally distributed with the same mean and standard deviation. Since in practice simply looking at the data is generally adequate, we will not discuss these approaches in detail.‡

Unfortunately, none of these methods is especially convincing one way or the other for the small sample sizes common in biomedical research, and your choice of approach (i.e., parametric versus nonparametric) often has to be based more on judgment and preference than hard evidence.

*The methods in this chapter are obviously not the first nonparametric methods we have encountered. The χ^2 for analysis of nominal data in contingency tables in Chapter 5, the Spearman rank correlation to analyze ordinal data in Chapter 8, and McNemar's test in Chapter 9 are three widely used nonparametric methods.

†They also require that the distributions be continuous (so that ties are impossible) to derive the mathematical forms of the sampling distributions used to define the critical values of the various test statistics. In practice, however, the continuity restriction is not important, and the methods can be (and are) applied to observations with tied measurements.

‡For discussions and example of these procedures, see Zar JH. Assessing departures from the normal distribution. *Biostatistical Analysis*, 5th ed. Upper Saddle River, NJ: Prentice Hall; 2010:sec 6.6.

One informal approach is to do the analysis with both the applicable parametric *and* nonparametric methods, then compare the results. If the data are from a normal population, then the parametric method should be more sensitive (and so provide a lower P value), whereas if there is substantial nonnormality then the nonparametric method should be more sensitive (and so provide the lower P value). If the data are only slightly nonnormal, the two approaches should give similar results.

Things basically come down to the following difference of opinion: Some people think that in the *absence* of evidence that the data were *not* drawn from a normally distributed population, one should use parametric tests because they are more powerful and more widely used. These people say that you should use a nonparametric test only when there is positive evidence that the populations under study are not normally distributed. Others point out that the nonparametric methods discussed in this chapter are 95% as powerful as parametric methods when the data are from normally distributed populations and more reliable when the data are not from normally distributed populations. They also believe that investigators should assume as little as possible when analyzing their data. They therefore recommend that nonparametric methods be used *except* when there is *positive evidence* that parametric methods are suitable. At the moment, there is no definitive answer stating which attitude is preferable. And there probably never will be such an answer.

■ TWO DIFFERENT SAMPLES: THE MANN-WHITNEY RANK-SUM TEST

When we developed the analysis of variance, t test, and Pearson product-moment correlation, we began with a specific (normally distributed) population and examined the values of the test statistic associated with all possible samples of a given size that could be selected from that population. The situation is different for methods based on ranks rather than the actual observations. We will replace the actual observations with their ranks, then focus on the population of all possible combinations of ranks. Since all samples have a finite number of members, we can simply list all the different possible ways to rank the members to obtain the distribution of possible values for the test statistic when the treatment has no effect.

■ TABLE 10-1. Observations in Diuretic Experiment

Placebo (Control)		Drug (Treatment)	
Daily Urine Production (mL/d)	Rank*	Daily Urine Production (mL/d)	Rank*
1000	1	1400	6
1380	5	1600	7
1200	3	1180	2
		1220	4
	$T = 9$		

*1 = smallest; 7 = largest.

To illustrate this process but keep this list relatively short, let us analyze a small experiment in which three people take a placebo and four people take a drug that is thought to be a diuretic. Table 10-1 shows the daily urine production observed in this experiment. Table 10-1 also shows the ranks of all the observations without regard to which experimental group they fall in; the smallest observed urine production is ranked 1 and the largest one is ranked 7. If the drug affected daily urine production, we would expect the rankings in the control group to be lower (or higher, if the drug decreased urine production) than the ranks for the treatment group. We will use the sum of ranks in the smaller group (in this case, the control group) as our test statistic T . The control-group ranks add up to 9.

Is the value of $T = 9$ sufficiently extreme to justify rejecting the hypothesis that the drug had no effect?

To answer this question, we examine the *population of all possible rankings* of the seven observations divided into two groups, one with 3 individuals and one with 4, to see how likely we are to get a rank sum as extreme as that associated in Table 10-1. Notice that we are no longer discussing the actual observations but their ranks, so our results will apply to *any* experiment in which there are two samples, one containing three individuals and the other containing four individuals, regardless of the nature of the underlying populations.

We begin with the hypothesis that the drug did not affect urine production, so that the ranking pattern in Table 10-1 is just due to chance. To estimate the chances

■ TABLE 10-2. Possible Ranks and Rank Sums for Three Individuals Out of Seven

Rank							Rank Sum <i>T</i>
1	2	3	4	5	6	7	
X		X					6
X			X				7
X				X			8
X					X		9
X	X					X	10
X		X	X				8
X		X		X			9
X		X			X		10
X		X				X	11
X			X	X			10
X			X		X		11
X			X			X	12
X				X	X		12
X				X		X	13
X					X	X	14
	X	X	X				9
	X	X		X			10
	X	X			X		11
	X	X				X	12
	X		X	X			11
	X		X		X		12
	X		X			X	13
	X			X	X		13
	X			X		X	14
	X				X	X	15
		X	X	X			12
		X	X		X		13
		X	X			X	14
		X		X	X		14
		X		X		X	15
		X			X	X	16
			X	X	X		15
			X	X		X	16
			X		X	X	17
				X	X	X	18

of getting this pattern when the two samples were drawn from a single population, we need not engage in any fancy mathematics, we just *list* all the possible rankings that could have occurred. Table 10-2 shows all 35 different ways the ranks could have been arranged with three

people in one group and four in the other. The crosses indicate a person in the placebo group, and the blanks indicate a person in the treatment group. The right-hand column shows the sum of ranks for the people in the smaller (placebo) group for each possible combination

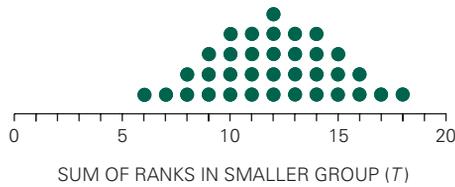


FIGURE 10-1. Sums of ranks in the smaller group for all possible rankings of seven individuals with three individuals in one sample and four in the other. Each circle represents one possible sum of ranks.

of ranks. Figure 10-1 shows the distribution of possible values of our test statistic, the sum of ranks of the smaller group T that can occur when the treatment has no effect. While this distribution looks a little like the t distribution in Figure 4-5, there is a very important difference. Whereas the t distribution is continuous and, in theory, is based on an infinitely large collection of possible values of the t test statistic, Figure 10-1 shows every possible value of the sum-of-ranks test statistic T .

Since there are 35 possible ways to combine the ranks, there is 1 chance in 35 of getting rank sums of 6, 7, 17, or 18; 2 chances in 35 of getting 8 or 16; 3 chances in 35 of getting 9 or 15; 4 chances in 35 of getting 10, 11, 13, or 14; and 5 chances in 35 of getting 12. What are the chances of getting an extreme value of T ? There is a $2/35 = .057 = 5.7\%$ chance of obtaining $T = 6$ or $T = 18$ when the treatment has no effect. We use these numbers as the critical values to define extreme values of T and reject the hypothesis of no treatment effect. Hence, the value of $T = 9$ associated with the observations in Table 10-1 is not extreme enough to justify rejecting the hypothesis that the drug has no effect on urine production.

Notice that in this case $T = 6$ and $T = 18$ correspond to $P = .057$. Since T can take on only integer values, P can take on only discrete values. As a result, tables of critical values of T present pairs of values that define the proportion of possible values nearest traditional critical P values, for example, 5% and 1%, but the exact P values defined by these critical values generally do not equal 5% and 1% exactly. Table 10-3 presents these critical values. n_S and n_B are the number of members in the smaller and larger samples group. The table gives the critical values of T that come nearest defining the most extreme 5% and 1% of all possible values of T that will occur if the treatment has no effect, as well as the exact proportion

of possible T values defined by the critical values. For example, Table 10-3 shows that 7 and 23 define the 4.8% most extreme possible values of the rank sum of the smaller of two sample groups T when $n_S = 3$ and $n_B = 6$.

The procedure we just described is the *Mann-Whitney rank-sum test*.^{*} The procedure for testing the hypothesis that a treatment had no effect with this statistic is:

- Rank all observations according to their magnitude, a rank of 1 being assigned to the smallest observation. Tied observations should be assigned the same rank, equal to the average of the ranks they would have been assigned had there been no tie (i.e., using the same procedure as in computing the Spearman rank correlation coefficient in Chapter 8).
- Compute T , the sum of the ranks in the smaller sample. (If both samples are the same size, you can compute T from either one.)
- Compare the resulting value of T with the distribution of all possible rank sums for experiments with samples of the same size to see whether the pattern of rankings is compatible with the hypothesis that the treatment had no effect.

There are two ways to compare the observed value of T with the critical value defining the most extreme values that would occur if the treatment had no effect. The first approach is to compute the exact distribution of T by listing all the possibilities, as we just did, then tabulate the results in a table such as Table 10-3. For experiments in which the samples are small enough to be included in Table 10-3 this approach gives the exact P value associated with a given set of experimental observations. For larger

^{*}There is an alternative formulation of this test that yields a statistic commonly denoted by U . U is related to T by the formula $U = T - n_S n_B + n_S(n_S + 1)/2$, where n_S is the size of the smaller sample (or either sample if both contain the same number of individuals). For a presentation of the U statistic, see Siegel S, Castellan NJ Jr. *The Wilcoxon-Mann-Whitney U test*. In: *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed. New York: McGraw-Hill; 1988:sec 6.4. For a detailed derivation and discussion of the Mann-Whitney test as developed here, as well as its relationship to U , see Mosteller F, Rourke R. *Ranking methods for two independent samples*. *Sturdy Statistics: Nonparametrics and Order Statistics*. Reading, MA: Addison-Wesley; 1973:chap 3.

■ TABLE 10-3. Critical Values (Two-Tailed) of the Mann-Whitney Rank-Sum Statistic T

n_S	n_B	Probability Levels Near			
		.05		.01	
		Critical Values	P	Critical Values	P
3	4	6,18	.057		
	5	6,21	.036		
	5	7,20	.071		
	6	7,23	.048	6,24	.024
	7	7,26	.033	6,27	.017
	7	8,25	.067		
	8	8,28	.042	6,30	.012
4	4	11,25	.057	10,26	.026
	5	11,29	.032	10,30	.016
	5	12,28	.063		
	6	12,32	.038	10,34	.010
	7	13,35	.042	10,38	.012
	8	14,38	.048	11,41	.008
	8	12,40	.016
5	5	17,38	.032	15,40	.008
	5	18,37	.056	16,39	.016
	6	19,41	.052	16,44	.010
	7	20,45	.048	17,48	.010
	8	21,49	.045	18,52	.011
6	6	26,52	.041	23,55	.009
	6	24,54	.015
	7	28,56	.051	24,60	.008
	7	25,59	.014
	8	29,61	.043	25,65	.008
	8	30,60	.059	26,64	.013
7	7	37,68	.053	33,72	.011
	8	39,73	.054	34,78	.009
8	8	49,87	.050	44,92	.010

Computed from Table A-9 of Mosteller F, Rourke R. *Sturdy Statistics: Nonparametrics and Order Statistics*. Reading, MA: Addison-Wesley; 1973.

experiments this exact approach becomes quite tedious because the number of possible rankings gets very large. For example, there are 184,756 different ways to rank two samples of 10 individuals each.

Second, when the large sample contains more than eight members, the distribution of T is very similar to the normal distribution with mean

$$T = \frac{n_S(n_S + n_B + 1)}{2}$$

and standard deviation

$$\sigma_T = \sqrt{\frac{n_S n_B (n_S + n_B + 1)}{12}}$$

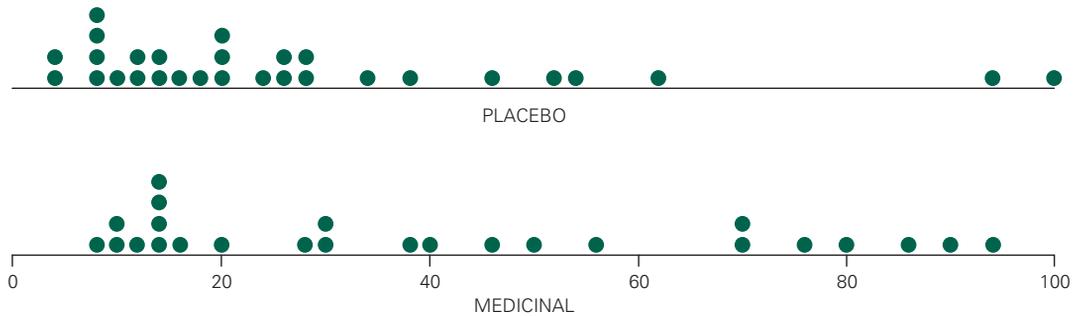


FIGURE 10-2. The level of pain reported among people with diabetic neuropathy after 12 weeks of taking a placebo or cannabis medicinal. The experimental subjects did not know which treatment they were receiving. Note that the pain distributions are not symmetrically distributed, but are skewed: most values tend to fall below about 30, but a few people experienced severe pain (high scores).

in which n_s is the size of the smaller sample.* Hence, we can transform T into the test statistic

$$z_T = \frac{T - \mu_T}{\sigma_T}$$

and compare this statistic with the critical values of the normal distribution that define the, say 5%, most extreme possible values. z_T can also be compared with the t distribution with an infinite number of degrees of freedom (Table 4-1) because it equals the normal distribution.

This comparison can be made more accurate by including a *continuity correction* (analogous to the Yates correction for continuity in Chapter 5) to account for the fact that the normal distribution is continuous whereas the rank sum T must be an integer

$$z_T = \frac{|T - \mu_T| - 1/2}{\sigma_T}$$

■ USE OF A CANNABIS-BASED MEDICINE IN PAINFUL DIABETIC NEUROPATHY

Diabetic neuropathy is a painful consequence of diabetes mellitus in which peripheral nerves are damaged, probably

*When there are tied measurements, the standard deviation needs to be reduced according to the following formula, which depends on the number of ties.

$$\sigma_T = \sqrt{\frac{n_s n_B (N+1)}{12} - \frac{n_s n_B}{12N(N+1)} \sum [(\tau_i - 1)\tau_i(\tau_i + 1)]}$$

in which $N = n_s + n_B$, τ_i = number of tied ranks in i th set of ties, the sum indicated by Σ is computed over all sets of tied ranks.

because of damage diabetes does to the small blood vessels that supply the nerves. The symptoms vary depending on the specific manifestation of the disease, but can include numbness and tingling in the extremities, uncontrollable muscle contractions and burning or electric pain. In an effort to develop better ways to control this pain, Dinesh Selvarajah and colleagues[†] conducted a prospective randomized double blind placebo controlled trial of a cannabis-based medicine to investigate whether this medicinal would effectively control the pain associated with diabetic neuropathy.

Experimental subjects were recruited from a diabetes clinic and randomly assigned to either receive the cannabis medicinal or a placebo. The experiment is *double blind* because neither the experimental subjects nor the investigators knew who was receiving the active medicinal. Including the placebo and blinding the experimental subjects was important not only to avoid the placebo effect, but also to avoid biased reporting of pain, which can be subjective. Likewise, the investigators were also blinded to the subjects' treatments to avoid biasing the recording and analysis of the pain data. The volunteers in the experiment were treated for 12 weeks, then asked to report their level of pain using a standardized questionnaire.

Figure 10-2 shows the raw data for the 29 people randomized to receive the placebo and the 24 people randomized to receive the medicinal. Even a cursory examination of the data shows that the pain responses are not normally distributed. (We discussed the data for placebo in conjunction with

[†]Selvarajah D, Emery CJ, Ghandi G, Tesfaye S. Randomized placebo-controlled double-blind clinical trial of cannabis-based medicinal product (sativex) in painful diabetic neuropathy. *Diabetes Care*. 2010;33:128–130.

■ TABLE 10-4. Diabetic Neuropathy Pain among People Treated with a Placebo and a Cannabis Medicinal

Placebo		Cannabis Medicinal	
Observation	Rank	Observation	Rank
13	16	90	50
8	6.5	10	9.5
46	39	45	38
61	44	70	45.5
28	31.5	13	16
7	4	27	30
93	51	11	11
10	9.5	70	45.5
7	4	14	19
100	53	15	20
4	1.5	13	16
16	21	75	47
23	27	50	40
33	35	30	34
18	22	80	48
51	41	40	37
26	29	29	33
19	23.5	13	16
20	25.5	9	8
54	42	7	4
19	23.5	20	25.5
37	36	85	49
13	16	55	43
8	6.5	94	52
28	31.5		
25	28		
4	1.5		
12	12.5		
12	12.5		
			<hr/> T = 737

Figure 2-11 and Box 2-1.) Because we cannot assume that the underlying populations from which the data were drawn is normally distributed, we compare these two treatment groups using the Mann-Whitney rank-sum test.

Table 10-4 shows the observed pain scores as well as the ranks of all the pain scores, without regard for which treatment each person received. All 53 people are ranked as a single group with the person with the lowest pain score ranked 1 and the highest ranked 53. In this case, two people in the placebo group are tied for the lowest pain score, 4, so each receives a rank of 1.5, the average of the first and second ranks. Because three people, two in the placebo group and one on the can-

nabis group, have the next highest pain score of 7, each receives a rank of 4, the average of the third, fourth, and fifth ranks. The person with the highest pain score, 100, who happens to also be in the placebo group, receives a rank of 53.

The cannabis medicinal group is the smaller sample, so we compute the test statistic T by summing all the ranks in that group, yielding $T = 737$. The cannabis group has $n_S = 24$ people in it and the larger placebo group, $n_B = 29$, so the mean value of T for all studies of this size is

$$\mu_T = \frac{n_S(n_S + n_B + 1)}{2} = \frac{24(24 + 29 + 1)}{2} = 648$$

and the standard deviation is

$$\sigma_T = \sqrt{\frac{n_S n_B (n_S + n_B + 1)}{2}} = \sqrt{\frac{24 \cdot 29(24 + 29 + 1)}{12}} = 55.96$$

So

$$z_T = \frac{|T - \mu_T| - \frac{1}{2}}{\sigma_T} = \frac{|737 - 648| - \frac{1}{2}}{55.96} = 1.581$$

This value is smaller than 1.960, the value of *z* that defines the most extreme 5% of the normal distribution (from Table 4-1). Hence, this study does not provide substantial evidence that the cannabis medicinal was any more or less effective than placebo in controlling pain associated with diabetic neuropathy.

■ EACH SUBJECT OBSERVED BEFORE AND AFTER ONE TREATMENT: THE WILCOXON SIGNED-RANK TEST

Chapter 9 presented the paired *t* test to analyze experiments in which each experimental subject was observed before and after a single treatment. This test required that the changes accompanying treatment be normally distributed. We now develop an analogous test based on ranks that does not require this assumption. We compute the differences caused by the treatment in each experimental subject, rank these differences according to their magnitude (without regard for sign), then attach the sign of the

difference to each rank, and, finally, sum the signed ranks to obtain the test statistic *W*.

This procedure uses information about the sizes of the differences the treatment produces in each experimental subject as well as its direction. Since it is based on ranks, it does not require making any assumptions about the nature of the population of the differences the treatment produces. As with the Mann-Whitney rank-sum test statistic, we can obtain the distribution of all possible values of the test statistic *W* by simply listing all the possibilities of the signed-rank sum for experiments of a given size. We finally compare the value of *W* associated with our observations with the distribution of all possible values of *W* that can occur in experiments involving the number of individuals in our study. If the observed value of *W* is “big,” the observations are not compatible with the assumption that treatment had no effect.

Remember that observations are ranked based on the *magnitude* of the changes *without regard for signs*, so that the differences that are equal in magnitude but opposite in sign, say -5.32 and $+5.32$, both have the same rank.

We begin with another hypothetical experiment in which we wish to test a potential diuretic on six people. In contrast to the experiments the last section described, we will observe daily urine production in each person *before* and *after* administering the drug. Table 10-5 shows the results of this experiment, together with the change in urine production that followed administering the drug in each person.

Daily urine production fell in five of the six people. Are these data sufficient to justify asserting that the drug was an effective diuretic?

■ TABLE 10-5. Effect of a Potential Diuretic on Six People

Person	Daily Urine Production (mL/d)			Rank* of Difference	Signed Rank of Difference
	Before Drug	After Drug	Difference		
1	1600	1490	-110	5	-5
2	1850	1300	-550	6	-6
3	1300	1400	+100	4	+4
4	1500	1410	-90	3	-3
5	1400	1350	-50	2	-2
6	1010	1000	-10	1	-1
					<u>W = -13</u>

*1 = smallest magnitude; 6 = largest magnitude.

■ TABLE 10-6. Possible Combinations of Signed Ranks for a Study of Six Individuals

Rank*						Sum of Signed Ranks
1	2	3	4	5	6	
-	-	-	-	-	-	-21
+	-	-	-	-	-	-19
-	+	-	-	-	-	-17
-	-	+	-	-	-	-15
-	-	-	+	-	-	-13
-	-	-	-	+	-	-11
-	-	-	-	-	+	-9
+	+	-	-	-	-	-15
+	-	+	-	-	-	-13
+	-	-	+	-	-	-11
+	-	-	-	+	-	-9
+	-	-	-	-	+	-7
-	+	+	-	-	-	-11
-	+	-	+	-	-	-9
-	+	-	-	+	-	-7
-	+	-	-	-	+	-5
-	-	+	+	-	-	-7
-	-	+	-	+	-	-5
-	-	+	-	-	+	-3
-	-	-	+	+	-	-3
-	-	-	+	-	+	-1
-	-	-	-	+	+	1
+	+	+	-	-	-	-9
+	+	-	+	-	-	-7
+	+	-	-	+	-	-5
+	+	-	-	-	+	-3
+	-	+	+	-	-	-5
+	-	+	-	+	-	-3
+	-	+	-	-	+	-1
+	-	-	+	+	-	-1
+	-	-	+	-	+	1
+	-	-	-	+	+	3
-	+	+	+	-	-	-3
-	+	+	-	+	-	-1
-	+	+	-	-	+	1
-	+	-	+	+	-	1
-	+	-	+	-	+	3
-	+	-	-	+	+	5
-	-	+	+	+	-	3
-	-	+	+	-	+	5
-	-	+	-	+	+	7
-	-	-	+	+	+	9

(continued)

■ TABLE 10-6. Possible Combinations of Signed Ranks for a Study of Six Individuals (Continued)

Rank*						Sum of Signed Ranks
1	2	3	4	5	6	
+	+	+	-	+	-	1
+	+	+	-	-	+	3
+	+	-	+	+	-	3
+	+	-	+	-	+	5
+	+	-	-	+	+	7
+	-	+	+	+	-	5
+	-	+	+	-	+	7
+	-	+	-	+	+	9
+	-	-	+	+	+	11
-	+	+	+	+	-	7
-	+	+	+	-	+	9
-	+	+	-	+	+	11
-	+	-	+	+	+	13
-	-	+	+	+	+	15
+	+	+	+	+	-	9
+	+	+	+	-	+	11
+	+	+	-	+	+	13
+	+	-	+	+	+	15
+	-	+	+	+	+	17
-	+	+	+	+	+	19
+	+	+	+	+	+	21

*Signs denote whether rank is positive or negative.

To apply the signed-rank test, we first rank the magnitudes of each observed change, beginning with 1 for the smallest change and ending with 6 for largest change. Next, we attach the sign of the change to each rank (last column of Table 10-5) and compute the sum of the signed ranks W . For this experiment, $W = -13$.

If the drug has no effect, the ranks associated with positive changes should be similar to the ranks associated with the negative changes and W should be near zero. On the other hand, when the treatment alters the variable being studied, the changes with the larger or smaller ranks will tend to have the same sign and the signed rank sum W will be a big positive or big negative number.

As with all test statistics, we need only draw the line between “small” and “big.” We do this by listing *all* 64 possible combinations of different ranking patterns, from all negative changes to all positive changes (Table 10-6).

There is one chance in 64 of getting any of these patterns by chance. Figure 10-3 shows all 64 of the signed-rank sums listed in Table 10-6.

To define a “big” value of W , we take the most extreme values of W that can occur when the treatment has no effect. Of the 64 possible rank sums, 4, or $4/64 = .0625 = 6.25\%$, fall at or beyond 19 (or -19), so we will reject the hypothesis that the treatment has no effect when the magnitude of W equals or exceeds 19 (i.e., W equals or is more negative than -19 or more positive than $+19$) with $P = .0625$.

Notice that, as with the Mann-Whitney rank-sum test, the discrete nature of the distribution of possible values of W means that we cannot always obtain P values precisely at traditional levels, such as 5%. Since the value of W associated with the observations in Table 10-5 is only -13 , these data are not sufficiently incompatible with the



FIGURE 10-3. All 64 possible sums of signed ranks for observations before and after administering a treatment to six individuals. Table 10-6 lists all the possibilities. The colored circles show that 4 out of 64 have a magnitude of 19 or more, that is, fall at or below -19 or at or above $+19$.

assumption that the treatment had no effect (that the drug is not an effective diuretic) to justify rejecting that hypothesis.

Table 10-7 presents the values of W that come closest to defining the most extreme 5% and 1% of all possible values for experiments with up to 8 subjects. For larger experiments, we use the fact that the distribution of W closely approximates a normal distribution with mean

$$\mu_w = 0$$

and standard deviation

$$\sigma_w = \sqrt{\frac{n(n+1)(2n+1)}{6}}$$

in which n equals the number of experimental subjects.

Therefore, we use

$$Z_w = \frac{W - \mu_w}{\sigma_w} = \frac{W}{\sqrt{[n(n+1)(2n+1)]/6}}$$

■ TABLE 10-7. Critical Values (Two-Tailed) of Wilcoxon W		
n	Critical Value	P
5	15	.062
	21	.032
6	19	.062
	28	.016
7	24	.046
	32	.024
8	28	.054

Data from Table A-11 of Mosteller F, Rourke R. *Sturdy Statistics: Nonparametrics and Order Statistics*. Reading, MA: Addison-Wesley; 1973.

as our test statistic. This approximation can be improved by including a continuity correction to obtain

$$Z_w = \frac{|W| - \frac{1}{2}}{\sqrt{[n(n+1)(2n+1)]/6}}$$

There are two kinds of *ties* that can occur when computing W . First, there can be no change in the observed variable when the treatment is applied, so that the difference is zero. In this case, that individual provides no information about whether the treatment increases or decreases the response variable; so it is simply dropped from the analysis, and the sample size is reduced by 1. Second, the magnitudes of the change the treatment produces can be the same for two or more individuals. As with the Mann-Whitney test, all the individuals with that change are assigned the same rank as the average of the ranks that would be used for the same number of individuals if they were not tied.*

In sum, here is the procedure for comparing the observed effects of a treatment in a single group of experimental subjects before and after administering a treatment:

- Compute the change in the variable of interest in each experimental subject.
- Rank all the differences according to their magnitude without regard for sign. (Zero differences should be dropped from the analysis with a corresponding reduction of sample

*When there are tied ranks and you use the normal distribution to compute the P value, σ_w needs to be reduced by a factor that depends on the number of ties according to the formula

$$\sigma_w = \sqrt{\frac{n(n+1)(2n+1)}{6} - \frac{\sum (\tau_i - 1)\tau_i(\tau_i + 1)}{12}}$$

in which n is the number of experimental subjects, τ_i is the number of tied ranks in the i th set of ties, and Σ indicates summation over all the sets of tied ranks.

■ TABLE 10-8. Maximum Percentage Platelet Aggregation before and after Smoking One Cigarette

Person	Before Smoking	After Smoking	Difference	Rank of Difference	Signed Rank of Difference
1	25	27	2	2	2
2	25	29	4	3.5	3.5
3	27	37	10	6	6
4	44	56	12	7	7
5	30	46	16	10	10
6	67	82	15	8.5	8.5
7	53	57	4	3.5	3.5
8	53	80	27	11	11
9	52	61	9	5	5
10	60	59	-1	1	-1
11	28	43	15	8.5	8.5
					W = 64

size. Tied ranks should be assigned the average of the ranks that would be assigned to the tied ranks if they were not tied.)

- Apply the sign of each difference to its rank.
- Add all the signed ranks to obtain the test statistic W .*
- Compare the observed value of W with the distribution of possible values that would occur if the treatment had no effect, and reject this hypothesis if W is “big.”

To further illustrate this process, let us use the Wilcoxon signed-rank test to analyze the results of an experiment we discussed in Chapter 9.

Cigarette Smoking and Platelet Function

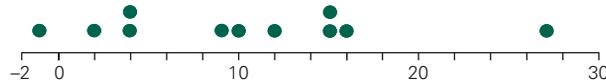
Table 10-8 reproduces the results, shown in Figure 9-2, of Levine’s experiment measuring platelet aggregation of 11 people before and after each one smokes a cigarette. Recall that increased platelet aggregation indicates a greater propensity to form blood clots (capable of caus-

ing heart attacks, strokes, and other vascular disorders). The fourth column of Table 10-8 shows the change in platelet aggregation that accompanies smoking a cigarette.

Figure 10-4 shows these differences. While this figure may not present results that preclude using methods based on the normal distribution (such as the paired t test), it does suggest that it would be more prudent to use a nonparametric method such as the Wilcoxon signed-rank test because the differences do not appear to be symmetrically distributed about the mean and more likely to be near the mean than far from it. In particular, *outliers* such as the point at 27% can bias methods based on a normal distribution.

To continue with our computation, which does not require the assumption of normally distributed changes, rank the magnitudes of each of these changes, the smallest change (1%) being ranked 1 and the largest change (27%) being ranked 11. The fifth column in Table 10-8 shows these ranks. The last column shows the same ranks with the sign of the change attached. The sum of the signed ranks W is $2 + 3.5 + 6 + 7 + 10 + 8.5 + 3.5 + 11 + 5 + (-1) + 8.5 = 64$. This value exceeds 52, the value that defines the 1.8% most extreme values of W that can occur when the treatment has no effect (from Table 10-7), so we can report that these data support the assertion that smoking increases platelet aggregation ($P = .018$).

*Note that we have developed W as the sum of *all* the signed ranks of the differences. There are alternative deviations of the Wilcoxon signed-rank test that are based on the sum of only the positively or negatively signed ranks. These alternative forms are mathematically equivalent to the one developed here. You need to be careful when using tables of the critical value W to be sure which way the test statistic was defined when the table was constructed.



CHANGE IN PLATELET AGGREGATION AFTER SMOKING A CIGARETTE (%)

FIGURE 10-4. Change in platelet aggregation after smoking a cigarette. These changes do not seem to be normally distributed, especially because of the outlier at 27%. This plot suggests that a nonparametric method, such as the Wilcoxon signed-rank test, is preferable to a parametric method, such as the paired t test, to analyze the results of this experiment.

■ EXPERIMENTS WITH THREE OR MORE GROUPS WHEN EACH GROUP CONTAINS DIFFERENT INDIVIDUALS: THE KRUSKAL-WALLIS TEST

Chapter 3 discussed experiments in which three or more different groups of experimental subjects are exposed to different treatments and the observations could be considered to come from normally distributed populations with similar variances. Now we shall develop an analogous procedure to the one-way analysis of variance (Chapter 3) based on ranks that does not require making these assumptions.

The *Kruskal-Wallis test* is a direct generalization of the Mann-Whitney rank-sum test. One first ranks all the observations *without regard for which treatment group they are in*, beginning with 1 for the smallest observation. (Ties are treated as before, that is, they are assigned the average value that would be associated with the tied observations if they were not tied.) Next, compute the rank sum for each group. If the treatments have no effect, the *large and small ranks should be evenly distributed among the different groups*, so the average rank in each group should approximate the average of all the ranks computed without regard of the grouping. The more disparity there is between observed average ranks in each group and what you would expect if the hypothesis of no treatment effect was true, the less likely we will be to accept that hypothesis. Now, let us construct such a test statistic.

For simplicity let us assume there are only three groups; then generalize the resulting equations to any number of groups when we are finished. The three different treatment groups contain n_1 , n_2 , and n_3 experimental subjects, and the rank sums for these three groups are R_1 , R_2 , and R_3 . Therefore, the mean ranks observed in the three groups are $\bar{R}_1 = R_1/n_1$, $\bar{R}_2 = R_2/n_2$, and $\bar{R}_3 = R_3/n_3$,

respectively. The average rank of all the $n_1 + n_2 + n_3 = N$ observations is the average of the first N integers

$$\bar{R} = \frac{1+2+3+\cdots+N}{N} = \frac{N+1}{2}$$

We will use the sum of squared deviations between each sample group's average rank and the overall average rank, weighted by the sizes of each group, as a measure of variability between the observations and what you would expect if the hypothesis of no treatment effect was true. Call this sum D .

$$D = n_1(\bar{R}_1 - \bar{R})^2 + n_2(\bar{R}_2 - \bar{R})^2 + n_3(\bar{R}_3 - \bar{R})^2$$

This sum of squared deviations is exactly analogous to the weighted sum of squared deviations between the sample means and grand mean that define the between-groups sum of squares in the parametric one-way repeated measures analysis of variance as developed in Chapter 9.

The distribution of possible values of D when the treatments have no effect depends on the size of the sample. It is possible to obtain a test statistic that does not depend on sample size by dividing D by $N(N+1)/12$,

$$H = \frac{D}{N(N+1)/12} = \frac{12}{N(N+1)} \sum n_i (\bar{R}_i - \bar{R})^2$$

The summation denoted with Σ is over all the treatment groups, regardless of how many treatment groups there are. It is the *Kruskal-Wallis test statistic*.

The exact distribution of H can be computed by listing all the possibilities, as we did with Mann-Whitney and Wilcoxon tests, but there are so many different possibilities that the resulting table would be huge. Fortunately, if the sample sizes are not too small, the χ^2 distribution with $\nu = k - 1$ degrees of freedom, where k is the number of treatment groups, closely approximates the distribution of

H . Hence, we can test the null hypothesis that the treatments had no effect by computing H for the observations and comparing the resulting value with the critical values for χ^2 in Table 5-7. This approximation works well in experiments with three treatment groups when each group contains at least five members and for experiments with four treatment groups when there are more than 10 individuals in the entire study. For smaller studies, consult a table of the exact distribution of the H to obtain the P value. (We do not include such a table because of its length and the relatively infrequent need for one; most intermediate statistics texts include one.)

In summary, the procedure for analyzing an experiment in which different groups of experimental subjects receive each treatment is:

- Rank each observation without regard for treatment group, beginning with a rank of 1 for the smallest observation. (Ties are treated in the same way as the other rank tests.)*
- Compute the Kruskal-Wallis test statistic H to obtain a normalized measure of how much the average ranks within each treatment group deviate from the average rank of all the observations.
- Compare H with χ^2 distribution with 1 less degree of freedom than the number of treatment groups, unless the sample size is small, in which case you must compare H with the exact distribution. If H exceeds the critical value that defines a “big” H , reject the null hypothesis that the treatment has no effect.

Now let us illustrate this procedure with an example.

Prenatal Marijuana Exposure and Child Behavior

Although most women stop using marijuana once they get pregnant, approximately 2.8% report using it during the first trimester of pregnancy and occasionally during the remainder of pregnancy. Exposure to marijuana is associated with attention deficits and impulsivity in young

children whose mothers used marijuana while pregnant, but not as much is known about the long-term effects on cognitive function. Lidush Goldschmidt and colleagues[†] designed a prospective observational study to track children whose mothers used marijuana during pregnancy. They interviewed women who came to a prenatal clinic and attempted to recruit all women who used two or more joints of marijuana per month during the first trimester of pregnancy and a random selection of other pregnant women who did not smoke marijuana. They kept in touch with these women, then evaluated temperament and behavioral characteristics of the children when they were 10 years old. One of the assessments used to address attention deficit disorder and hyperactivity was the Swanson, Noland, and Pelham (SNAP) checklist, which is a questionnaire completed by mothers.

Table 10-9 gives the SNAP scores for the 31 children in this study. It shows the ranks of each observation together with the sum of ranks and mean ranks for each of the three exposure groups. The mean rank of all 31 observations is

$$\bar{R} = \frac{1+2+3+\cdots+31}{31} = \frac{N+1}{2} = \frac{31+1}{2} = 16$$

Therefore, the weighted sum of squared deviations between the average ranks observed in each treatment group and the average of all ranks is

$$D = 13(11.23-16)^2 + 9(16.89-16)^2 + 9(22.00-16)^2 \\ = 13(-4.77)^2 + 9(0.89)^2 + 9(6.00)^2 = 626.92$$

and, so,

$$H = \frac{D}{N(N+1)/12} = \frac{626.92}{31(31+1)/12} = 7.58$$

This value exceeds 5.991, the value that defines the largest 5% of the χ^2 distribution with $\nu = k - 1 = 3 - 1 = 2$ degrees of freedom (from Tables 5-7). Therefore, we conclude that at least one of these three groups differed in hyperactivity and attention deficit ($P < .05$).

Nonparametric Multiple Comparisons

As discussed in Chapter 5, the Bonferroni, Holm and Holm-Sidak procedures are based on controlling how probabilities of false positive errors accumulate in a family

*When there are ties, the approximation between the distributions of H and χ^2 can be improved by dividing H computed above by

$$1 - \frac{\sum(\tau_i - 1)\tau_i(\tau_i + 1)}{N(N^2 - 1)}$$

where τ_i is the number of ties in the i th set of tied ranks (as before). If there are only a few ties, this correction makes little difference and may be ignored.

[†]Goldschmidt L, Day NL, Richardson GA. Effects of prenatal marijuana exposure on child behavior problems at age 10. *Neurotoxicol Tetratol*. 2000;22:325-336.

■ TABLE 10-9. Average Number of Joints Per Day (AJD)

	AJD = 0 $n_1 = 13$		0 < AJD ≤ 0.89 $n_1 = 9$		AJD > 0.89 $n_1 = 9$	
	SNAP Score	Rank	SNAP Score	Rank	SNAP Score	Rank
	7.79	4	8.84	12	8.65	11
	9.16	17	9.92	24	10.70	31
	7.34	2	7.20	1	10.24	28
	10.28	29	9.25	20	8.62	10
	9.12	15	9.45	21	9.94	25
	9.24	19	9.14	16	10.55	30
	8.40	7	9.99	26	10.13	27
	8.60	9	9.21	18	9.78	23
	8.04	5	9.06	14	9.01	13
	8.45	8				
	9.51	22				
	8.15	6				
	7.69	3				
Sum of ranks, R_t		146		152		198
Mean rank, $R_t = R_t/n_t$		11.23		16.89		22.00

of comparisons so these tests can be applied to any kind of multiple comparisons. Thus, if we reject the null hypothesis of no difference among three or more groups using a Kruskal-Wallis analysis of variance on ranks, we can do the multiple comparisons using a series of Mann-Whitney rank-sum tests, with the critical P values adjusted for the multiple comparisons using the Holm-Sidak procedure. We will do all pairwise comparisons; multiple comparisons against a single control group would proceed in an analogous way, accounting for the fact that we would have to protect the family error rate against a smaller number of comparisons.

The first step in doing all pairwise comparisons for the study of marijuana smoke exposure during pregnancy and SNAP scores for their children is do the three pairwise comparisons of the data in Table 10-9 with Mann-Whitney rank-sum tests. Table 10-10 shows these tests. Note that the observations are ranked separately in each of the three Mann-Whitney rank-sum tests.

Finally, we test the pairwise comparisons in descending order of the size of z_T , as summarized in Table 10-11. The

first comparison, no marijuana smoking compared to smoking more than an average of .89 joints/day, was statistically significant, so we can conclude that the higher level of marijuana use by a pregnant woman was associated with attention deficit and hyperactivity in her child. The results for lower use are ambiguous; there is no detectable difference between nonusers and low (less than an average of .89 joints/day) users and low users and higher users, so this study is not adequate to determine whether there is a threshold for the effect and, if so, where that threshold is. Collecting more data—which would provide more power—should resolve this ambiguity.

Interestingly, had we done multiple comparisons against a single control group rather than all pairwise comparisons, we would have concluded that children whose mothers engaged in low levels smoking were not significantly different those whose mothers did not smoke at all, while children of mothers who were heavier marijuana smokers were affected, a less ambiguous finding. (We would avoid the ambiguous finding because we would not have tested the difference between low and high

■ TABLE 10-10. Mann-Whitney Tests for Multiple Comparisons for Marijuana Exposure and Child Behavior

None vs. Low				None vs. High				Low vs. High			
SNAP Score	Rank	SNAP Score	Rank	SNAP Score	Rank	SNAP Score	Rank	SNAP Score	Rank	SNAP Score	Rank
None		Low		None		High		Low		High	
7.79	4	8.64	10	7.79	3	8.65	10	8.64	3	8.65	4
9.16	14	9.92	20	9.16	13	10.70	22	9.92	12	10.70	18
7.34	2	7.20	1	7.34	1	10.24	19	7.20	1	10.24	16
10.28	22	9.25	17	10.28	20	8.62	9	9.25	9	8.62	2
9.12	12	9.45	18	9.12	12	9.94	17	9.45	10	9.94	13
9.24	16	9.14	13	9.24	14	10.55	21	9.14	7	10.55	17
8.40	7	9.99	21	8.40	6	10.13	18	9.99	14	10.13	15
8.60	9	9.21	15	8.60	8	9.78	16	9.21	8	9.78	11
8.04	5	9.06	11	8.04	4	9.01	11	9.06	6	9.01	5
8.45	8			8.45	7						
9.51	19			9.51	15						
8.15	6			8.15	5						
7.69	3			7.69	2						
$n_B = 13$		$n_S = 9$		$n_B = 13$		$n_S = 9$		$n_B = 9$		$n_S = 9$	
$T = 126$				$T = 143$				$T = 70$			
$\mu_T = \frac{9(9 + 13 + 1)}{2} = 103.5$				$\mu_T = \frac{9(9 + 13 + 1)}{2} = 103.5$				$\mu_T = \frac{9(9 + 9 + 1)}{2} = 85.5$			
$\sigma_T = \sqrt{\frac{9 \cdot 13(9 + 13 + 1)}{12}} = 14.97$				$\sigma_T = \sqrt{\frac{9 \cdot 13(9 + 13 + 1)}{12}} = 14.97$				$\sigma_T = \sqrt{\frac{9 \cdot 9(9 + 9 + 1)}{12}} = 11.32$			
$z_T = \frac{ 126 - 103.5 - \frac{1}{2}}{14.97} = 1.470$				$z_T = \frac{ 143 - 103.5 - \frac{1}{2}}{14.97} = 2.605$				$z_T = \frac{ 70 - 85.5 - \frac{1}{2}}{11.32} = 1.325$			
$.20 < P < .10$				$P < .01$				$.20 < P < .10$			

■ **TABLE 10-11. Pairwise Comparisons of Smoking Marijuana while Pregnant and Child Behavior Using the Holm-Sidak Adjustment (Family Error Rate, $\alpha_T = 0.05$)**

Comparison	z_T	P	j	$P_{crit} = \alpha_T / (k - j + 1)$	$P < P_{crit}?$
None vs. high	2.605	<.010	1	.0170	Yes
None vs. low	1.470	<.10	2	.0253	No
Low vs. high	1.325	<.10	3	.0500	No*

*Because the second comparison is not significant, all subsequent comparisons are considered not significant.

users.) Of course, given the small size of the study, this negative conclusion could have been the result of low power.

■ **EXPERIMENTS IN WHICH EACH SUBJECT RECEIVES MORE THAN ONE TREATMENT: THE FRIEDMAN TEST**

Often it is possible to complete experiments in which each individual is exposed to a number of different treatments. This experimental design reduces the uncertainty due to variability in the responses between individuals and provides a more sensitive test of what the treatments do in a given person. When the assumptions required for parametric methods can be reasonably satisfied, such experiments can be analyzed with the repeated measures analysis of variance in Chapter 9. Now we will derive an analogous test based on ranks that does not require that the observations be drawn from normally distributed populations, the *Friedman test*.

The logic of this test is quite simple. Each experimental subject receives each treatment, so we rank *each subject's* responses to the treatments without regard for the other subjects. If the null hypothesis that the treatment has no effect is true, then, for each subject, the ranks will be randomly distributed and the sum of the ranks for each *treatment* will be similar. Table 10-12 illustrates such a case, in which five different subjects receive four treatments. Instead of the measured responses, this table contains the *ranks* within each experimental subject's responses. Hence, the treatments are ranked 1, 2, 3, and 4 separately for each subject. The bottom line in the table gives the sums of the ranks for all people receiving each treatment. These rank sums are all similar and also roughly equal to 12.5, which is the average rank, $(1 + 2 + 3 + 4) / 4 = 2.5$, times the number of subjects, 5. This table does not suggest that any of the treatments had any systematic effect on the experimental subjects.

Now consider Table 10-13. The first treatment *always* produces the greatest response in each experimental

■ **TABLE 10-12. Ranks of Outcomes for an Experiment When Five Subjects Each Receive Four Treatments**

Experimental Subject	Treatment			
	1	2	3	4
1	1	2	3	4
2	4	1	2	3
3	3	4	1	2
4	2	3	4	1
5	1	4	3	2
Rank sum R_t	11	14	13	12

■ **TABLE 10-13. Ranks of Outcomes for Another Experiment When Five Subjects Each Receive Four Treatments**

Experimental Subject	Treatment			
	1	2	3	4
1	4	1	3	2
2	4	1	3	2
3	4	1	3	2
4	4	1	3	2
5	4	1	3	2
Rank sum R_t	20	5	15	10

subjects the second treatment always produces the smallest response, and the third and fourth treatments always produce intermediate responses, the third treatment producing a greater response than the fourth treatment. The bottom line shows the column rank sums. In this case, there is a great deal of variability in the rank sums, some being much larger or smaller than five times the average rank of 12.5. Table 10-13 strongly suggests that the treatments affect the variable being studied.

All we have left to do is to reduce this subjective impression of a difference to a single number. In a way similar to that used in deriving the Kruskal-Wallis statistic, we compute the sum of squared deviations between the rank sums observed for each treatment and the rank sum that we would expect if each treatment were as likely to have any of the possible rankings. This latter number is the average of the possible ranks.

For the examples in Tables 10-12 and 10-13, there are four possible treatments, so there are four possible ranks. Therefore, the average rank is $(1 + 2 + 3 + 4)/4 = 2.5$. In general, if there are k treatments, the average rank will be

$$\frac{1+2+3+\cdots+k}{k} = \frac{k+1}{2}$$

In our example, there are five experimental subjects, so we would expect each of the rank sums to be around 5 times the average rank for each person, or $5(2.5) = 12.5$. In Table 10-12 this is the case whereas in Table 10-13 it is not. If there are n experimental subjects and the ranks are randomly distributed between the treatments, each of the rank sums should be about n times the average rank, or $n(k+1)/2$. Hence, we can collapse all this information into a single number by computing the sum of squared differences between the observed rank sums and rank sums that would be expected if the treatments had no effect.

$$S = \sum [R_t - n(k+1)/2]^2$$

in which Σ denotes the sum over all the treatments and R_t denotes the sum of ranks for treatment t .

For example, for the observations in Table 10-12, $k = 4$ treatments and $n = 5$ experimental subjects, so

$$\begin{aligned} S &= (11-12.5)^2 + (14-12.5)^2 + (13-12.5)^2 + (12-12.5)^2 \\ &= (-1.5)^2 + (1.5)^2 + (.5)^2 + (-.5)^2 = 5 \end{aligned}$$

and for Table 10-13

$$\begin{aligned} S &= (20-12.5)^2 + (5-12.5)^2 + (15-12.5)^2 + (10-12.5)^2 \\ &= (7.5)^2 + (-7.5)^2 + (2.5)^2 + (-2.5)^2 = 125 \end{aligned}$$

In the former case, S is a small number; in the latter S is a big number. The more of a pattern there is relating the ranks within each subject to the treatments, the greater the value of our test statistic S .

We could stop here and formulate a test based on S , but statisticians have shown that we can simplify the problem by dividing this sum of squared differences between the observed and expected rank sums by $nk(k+1)/12$ to obtain

$$\begin{aligned} \chi_r^2 &= \frac{S}{nk(k+1)/12} = \frac{12 \sum [R_t - n(k+1)/2]^2}{nk(k+1)} \\ &= \frac{12}{nk(k+1)} \sum R_t^2 - 3n(k+1) \end{aligned}$$

The test statistic χ_r^2 , is called *Friedman's statistic* and has the desirable property that, for large enough samples, it follows the χ^2 distribution with $\nu = k - 1$ degrees of freedom, regardless of sample size.* When there are three treatments and nine or fewer experimental subjects or four treatments with four or fewer experimental subjects each, the χ^2 approximation is not adequate, so one needs to compare χ_r^2 to the exact distribution of possible values obtained by listing all the possibilities in Table 10-14.

In summary, the procedure for using the Friedman statistic to analyze experiments in which the same individuals receive several treatments is as follows:

- Rank each observation within each experimental subject, assigning 1 to the smallest response. (Treat ties as before.)

*When there are tied measurements, χ_r^2 needs to be increased by dividing it by

$$1 - \frac{\sum_{\text{subjects}, i} \sum_{\text{ties within subjects}, j} (\tau_{ij} - 1)\tau_{ij}(\tau_{ij} + 1)}{Nk(k^2 - 1)}$$

in which τ_{ij} = number of tied ranks in the i th set of ties within the ranks for subject j , and the double sum $\Sigma\Sigma$ is computed over all ties within each subject. If there are only a few ties, this correction makes little difference and can be ignored.

■ TABLE 10-14. Critical Values for Friedman χ_r^2

$k = 3$ Treatments			$k = 4$ Treatments		
n	χ_r^2	P	n	χ_r^2	P
3	6.00	.028	2	6.00	.042
4	6.50	.042	3	7.00	.054
	8.00	.005		8.20	.017
5	5.20	.093	4	7.50	.054
	6.40	.039		9.30	.011
	8.40	.008	5	7.80	.049
6	5.33	.072		9.96	.009
	6.33	.052	6	7.60	.043
	9.00	.008		10.20	.010
7	6.00	.051	7	7.63	.051
	8.86	.008		10.37	.009
8	6.25	.047	8	7.65	.049
	9.00	.010		10.35	.010
9	6.22	.048			
	8.67	.010			
10	6.20	.046			
	8.60	.012			
11	6.54	.043			
	8.91	.011			
12	6.17	.050			
	8.67	.011			
13	6.00	.050			
	8.67	.012			
14	6.14	.049			
	9.00	.010			
15	6.40	.047			
	8.93	.010			

Data from Owen DB. *Handbook of Statistical Tables*. US Department of Energy. Reading, MA: Addison-Wesley; 1962.

- Compute the sum of the ranks observed in all subjects for each treatment.
- Compute the Friedman test statistic χ_r^2 as a measure of how much the observed rank sums differ from those that would be expected if the treatments had no effect.
- Compare the resulting value of the Friedman statistic with the χ^2 distribution if the experiment involves large enough samples or with the exact distribution of χ_r^2 in Table 10-14 if the sample is small.

Anti-asthmatic Drugs and Endotoxin

Table 10-15 reproduces the observed forced expiratory volume 1 (FEV₁) at one second in Table 9-5 that Berenson and colleagues used to study whether or not salbutamol had a protective effect on endotoxin induced bronchoconstriction. In Chapter 9, we analyzed these data with a repeated measures one-way analysis of variance. Now, let us reexamine them using ranks to avoid having to make any assumptions about the population these patients represent.

■ **TABLE 10-15. Forced Expiratory Volume at 1 Second before and after Bronchial Challenge with Endotoxin and Salbutamol**

Person (Subject)	FEV ₁ (L)					
	No Drug (Baseline)		One Hour after Endotoxin		Two Hours after Endotoxin and Salbutamol	
	Units	Rank	Units	Rank	Units	Rank
1	3.7	2	3.4	1	4.0	3
2	4.0	2	3.7	1	4.4	3
3	3.0	2	2.8	1	3.2	3
4	3.2	2	2.9	1	3.4	3
Rank sums for each group		8		4		12

Table 10-15 shows how the three treatments rank in terms of FEV₁ for each of the four people in the study. The last row gives the sums of the ranks for each treatment. Since the possible ranks are 1, 2, and 3, the average rank is $(1 + 2 + 3)/3 = 2$. Since there are four people, if the treatments had no effect, these rank sums should all be about $4(2) = 8$. Hence, our measure of the difference between this expectation and the observed data is

$$S = (8-8)^2 + (4-8)^2 + (12-8)^2 \\ = (0)^2 + (4)^2 + (4)^2 = 32$$

We convert S into χ_r^2 by dividing by $nk(k+1)/12 = 4(3)(3+1)/12 = 4$ to obtain $\chi_r^2 = 32/4 = 8.0$. Table 10-14 shows that for an experiment with $k = 3$ treatments and $n = 4$ experimental subjects there is only a $P = .042$ chance of obtaining a value of χ_r^2 as big or bigger than 8 by chance if the treatments have no effect. Therefore, we can report that endotoxin and salbutamol alter FEV₁ ($P = .042$).

Multiple Comparisons after the Friedman Test

Just as we could use the Mann-Whitney test with a Holm-Sidak (or Bonferroni or Holm) correction for multiple comparisons following a Kruskal-Wallis analysis of variance on ranks, we can use the Wilcoxon signed rank tests with a Holm-Sidak (or Bonferroni or Holm) correction for multiple comparisons following a significant Friedman repeated measures analysis of variance on ranks.

■ SUMMARY

The methods in this chapter permit testing hypotheses similar to those we tested with analysis-of-variance and t tests but do not require us to assume that the underlying populations follow normal distributions. We avoid having to make such an assumption by replacing the observations with their ranks before computing the test statistic (T , W , H , or χ_r^2). By dealing with ranks we preserve most of the information about the relative sizes (and signs) of the observations. More important, by dealing with ranks, we do not use information about the population or populations the samples were drawn from to compute the distribution of possible values of the test statistic. Instead we consider the population of all possible ranking patterns (often by simply listing all the possibilities) to compute the P value associated with the observations.

It is important to note that the procedures we used in this chapter to compute the P value from the ranks of the observations is essentially the same as the methods we have used everywhere else in this book:

- Assume that the treatment(s) had no effect, so that any differences observed between the samples are due to the effects of random sampling.
- Define a test statistic that summarizes the observed differences between the treatment groups.
- Compute all possible values this test statistic can take on when the assumption that the treatments had no effect

is true. These values define the distribution of the test statistic we would expect if the hypothesis of no effect was true.

- Compute the value of the test statistic associated with the actual observations in the experiment.
- Compare this value with the distribution of all possible values; if it is “big,” it is unlikely that the observations came from the same populations (i.e., that the treatment had no effect), so conclude that the treatment had an effect.

The specific procedure you should use to analyze the results from a given experiment depends on the design of the experiment and the nature of the data. When the data are measured on an ordinal scale or you cannot or do not wish to assume that the underlying populations follow normal distributions, the procedures developed in this chapter are appropriate.

■ PROBLEMS

10-1 Despite progresses in technique, adhesions (the abnormal connection between tissues inside the body formed during healing following surgery) continue to be a problem in abdominal surgery, such as when operating on the uterus. To see if it would be possible to reduce adhesions following uterine surgery by placing a membrane around the area of incision in the uterus, Nurullah Bülbüller and colleagues* operated on the uteruses of two groups of rats, a control group that simply received the surgery and a test group that had the membrane applied over the uterus. This bioresorbable membrane prevented the tissue of the uterus from connecting to other internal organs of the peritoneum (the inside lining of the abdomen), then was slowly absorbed by the surrounding tissue after healing was complete. They allowed the rats to heal, then sacrificed them and measured the amount of adhesions, according to the scale in Table 10-16. The scores for the two groups of rats using the different surgical techniques are in Table 10-17. Does use of the membrane affect the extent of adhesions?

*Bülbüller N, et al. Effect of a bioresorbable membrane on postoperative adhesions and wound healing. *J Reprod Med.* 2003;48:547–550.

■ **TABLE 10-16. Scale for Rating Adhesions**

Grade	Definition
0	No adhesions
1	One band between organs or between one organ and the peritoneum
2	Two bands between organs or between one organ and the peritoneum
3	More than two bands between organs or mass formed by intestines not adhering to the peritoneum
4	Organs adhering to peritoneum or extensive adhesions

■ **TABLE 10-17. Uterine Adhesions Grade Using Two Different Surgical Techniques**

Control	Bioresorbable Membrane
3	1
4	1
4	2
4	0
2	0
1	0
3	2
2	0
1	1
0	3

10-2 The inappropriate and overuse of antibiotics is a well-recognized problem in medicine. To test whether it was possible to encourage more appropriate use of antibiotics in elderly hospitalized patients, Monika Lutters and her colleagues* monitored the number of patients receiving antibiotics in a 304-bed geriatric unit at her hospital

*Lutters M, et al. Effect of a comprehensive, multidisciplinary, educational program on the use of antibiotics in a geriatric university hospital. *J Am Geriatr Soc.* 2004;52:112–116.

before any intervention, after providing information to the physicians taking care of patients in the unit, after providing pocket cards with specific therapeutic guidelines for the use of antibiotics to treat the most common need for antibiotics in these patients (urinary and respiratory tract infections) combined with weekly lectures on appropriate use of antibiotics, then while the pocket cards were continued but the lectures stopped. The number of patients in the unit receiving antibiotics was recorded on each of 12 days under each experimental condition (see Table 10-18). Did the educational interventions have any effect on the number of patients receiving antibiotics? If so, how?

■ **TABLE 10-18. Number of Patients Receiving Antibiotics (out of 304 in the Geriatric Unit)**

Baseline	Information	Pocket Cards Plus Weekly Lectures	Pocket Cards Only
55	51	50	45
54	53	51	59
57	67	52	58
54	55	50	45
59	51	53	49
57	50	52	55
67	52	64	46
80	56	52	52
55	84	53	50
55	54	51	53
56	54	52	45
65	67	45	56

10-3 Rework Problem 9-5 using methods based on ranks.

10-4 Rework Problem 9-6 using methods based on ranks.

10-5 To determine whether or not offspring of parents with type II diabetes have abnormal glucose levels compared to offspring without a parental history of type II diabetes. Gerald Berenson and colleagues* collected data on whether these offspring had different cholesterol levels. The data for 30 subjects are shown in Table 10-19. Are these data consistent with the hypothesis that these offspring differ in cholesterol levels?

10-6 People with problem gambling habits are often substance abusers; these behaviors may be connected by an underlying personality trait such as impulsivity. Nancy Petry[†] investigated whether problem gamblers would also be at higher risk for contracting HIV, since the underlying impulsivity might make problem gamblers more likely to engage in riskier sexual behavior. She administered a questionnaire known as the HIV Risk Behavior Scale (HRBS) to assess sexual risk behavior in two groups of substance abusers, those with and without problem gambling. The HRBS is an 11-item questionnaire with questions addressing drug and sex behavior and responses are coded on a six-point scale from 0 to 5, with higher values associated with riskier behavior. The results of the HRBS sex composite score are shown in Table 10-20. What do these data indicate?

10-7 Rework Problem 9-1 using the Wilcoxon signed-rank test.

*Berenson G, et al. Abnormal characteristics in young offspring of parents with non-insulin-dependent diabetes mellitus. *Am J Epidemiol.* 1996;144:962–967.

[†]Petry N. Gambling problems in substance abusers are associated with increased sexual risk behaviors. *Addiction.* 2000;95:1089–1100.

■ **TABLE 10-19. Cholesterol Levels among Children of Parents with and without Diabetes**

Offspring with a Diabetic Parent					Offspring without a Diabetic Parent				
181	183	170	173	174	168	165	163	175	176
179	172	175	178	176	166	163	174	175	173
158	179	180	172	177	179	180	176	167	176

■ TABLE 10-20. HRBS Sex Composite Score

Non-problem Gambling Substance Abusers	Problem Gambling Substance Abusers
12	14
10	15
11	15
10	16
13	17
10	15
14	15
11	14
9	13
9	13
9	14
12	13
13	12
11	

10-8 In his continuing effort to become famous, the author of an introductory biostatistics text invented a new way to test if some treatment changes an individual's response. Each experimental subject is observed before and after

■ TABLE 10-21. Computation of the G Test Statistic

Subject	Before	After	Change	Contribution
	Treatment	Treatment		
1	100	110	+ 10	+ 1
2	95	96	+ 1	+ 1
3	120	100	- 20	0
4	111	123	+ 12	+ 1

treatment, and the change in the variable of interest is computed. If this change is positive, we assign a value of +1 to that subject; if it is negative, we assign a value of zero (assume that there are never cases that remain unchanged). The soon-to-be famous G test statistic is computed by summing up the values associated with the individual subjects. For example, for the data in Table 10-21, $G = 1 + 1 + 0 + 1 = 3$. Is G a legitimate test statistic? Explain briefly. If so, what is the sampling distribution for G when $n = 4$? $n = 6$? Can you use G to conclude that the treatment had an effect in the data given above with $P < .05$? How confident can you be about this conclusion? Construct a table of critical values for G when $n = 4$ and $n = 6$.

How to Analyze Survival Data

All the methods that we have discussed so far require “complete” observations, in the sense that we know the outcome of the treatment or intervention we are studying. For example, in Chapter 5 we considered a study that compared the rate of filing advance directives in people who received in-person counseling or written instructions (Table 5-1). We compared these two groups of people by computing the expected pattern of thrombus formation in each of the two comparison groups under the null hypothesis that there was no difference in the rate of thrombus formation in the two treatment groups, then used the chi-square test statistic to examine how closely the observed pattern in the data matched the expected pattern under the null hypothesis of no treatment effect. The resulting value of χ^2 was “big,” so we rejected the null hypothesis of no treatment effect and concluded that aspirin reduced the risk of thrombus formation. In this study we knew the outcome in *all* the people in the study after a fixed length of time following treatment. Indeed, in all the methods we have considered in this book so far, we knew the outcome of the variable under study for all the individuals in the study being analyzed. There are, however, situations, in which we do not know the ultimate outcome for all the individuals in the study because the study ended before the final outcome had been observed in all the study subjects or because the outcome in some of the individuals is not known.* In addition, it would be desirable to take into account the outcomes in people who were enrolled in

the study for varying lengths of follow-up that allows for the fact that the more time that passes after treatment the more likely it is that there would be the outcome of interest. We now turn our attention to developing procedures for such data.

The most common type of study in which we have incomplete knowledge of the outcome are clinical trials or survival studies in which individuals enter the study and are followed up over time until some event—typically death or development of a disease—occurs. Since such studies do not go on forever, it is possible that the study will end before the event of interest has occurred in all the study subjects. In such cases, we have incomplete information about the outcomes in these individuals. In clinical trials it is also common to lose track of patients who are being observed over time. Thus, we would know that the patient was free of disease up until the last time that we observed them, but we do not know what happened later. In both cases, we know that the individuals in the study were event free for some length of time, but not the actual time to an

*Another reason for not having all the data would be the case of *missing data*, in which samples are lost because of experimental problems or errors. Missing data are analyzed using the same statistical techniques as complete data sets, with appropriate adjustments in the calculations to account for the missing data. For a complete discussion of the analysis of studies with missing data, see Glantz S, Slinker B. *Primer of Applied Regression and Analysis of Variance*, 2nd ed. New York: McGraw-Hill; 2001.

event. These people are *lost to follow-up*; such data are known as *censored data*.^{*} Censored data are most common in clinical trials or survival studies.

■ CENSORING ON PLUTO

The tobacco industry, having been driven farther and farther from Earth by protectors of the public health, invades Pluto and starts to promote smoking in bars. Since it is very cold on Pluto, Plutonians spend most of their time indoors and begin dropping dead from the secondhand tobacco smoke in bars. Since it would be unethical to purposely expose Plutonians to secondhand smoke, we will simply observe how long it takes Plutonians to drop dead after they begin to be exposed to secondhand smoke in bars.

Figure 11-1A shows the observations for 10 nonsmoking Plutonians selected at random and observed over the course of a study lasting for 15 Pluto months. Subjects entered the study when they started hanging out at smoky bars, and they were followed-up until they dropped dead or the study ended. As with many survival studies, individuals were recruited into the study at various times as the study progressed. Of the 10 subjects, 7 died during the period of the study (A, B, C, F, G, H, and J). As a result, we know the exact length of time that they lived after their exposure to secondhand smoke in bars. These observations are *uncensored*. In contrast, two of the Plutonians were still alive at the end of the study (D and I); we know that they lived at least until the end of the study, but do not know how long they lived after being exposed to secondhand smoke. In addition, Plutonian E was vaporized in a freak accident while on vacation before the study was completed, so was lost to follow-up. We do know, however, that these individuals lived *at least as long* as we observed them. These observations are censored.

^{*}More precisely, these observations are *right censored* because we know the time the subjects entered the study, but not when they died (or experienced the event we are monitoring). It is also possible to have *left censored* data, when the actual survival time is larger than that observed, such as when patients are studied following surgery, and the precise dates at which some patients had surgery before the beginning of the study are not known. Other types of censoring can occur when studies are designed to observe subjects until some specified fraction (say, half) die. We will concentrate on right censored data, since that is what generally comes up in biomedical studies.

Figure 11-1B shows the data in another format, where the horizontal axis is the length of time that each subject is observed after starting exposure to secondhand smoke, as opposed to calendar time. The Plutonians who died by the end of the study have a solid point at the end of the line; those that were still alive at the end of the observation period are indicated with a lighter point. Thus, we know that Plutonian A lived exactly 7 months after starting to go to a smoky bar (an uncensored observation), whereas Plutonian D lived *at least* 12 months after hanging out in a smoky bar (a censored observation).

This study has the necessary features of a clinical follow-up study:

- There is a well-defined starting time for each subject (date smoking started in this example or date of diagnosis or medical intervention in a clinical study).
- There is a well-defined end point (death in this example or relapse in many clinical studies).
- The subjects in the study are selected at random from a larger population of interest.

If all subjects were studied for the same length of time or until they reached a common end point (such as death), we could use the methods of Chapters 5 or 10 to analyze the results. These methods require researchers to assess the outcomes at a fixed time follow the intervention, then classify each subject as either having or not having the outcome of interest or not. Unfortunately, in clinical studies these situations often do not exist. The fact that the study period often ends before all the subjects have reached the end point makes it impossible to know the actual time that all the subjects reach the common end point. In addition, because subjects are recruited throughout the duration of the study, the follow-up time often varies for different subjects. These two facts require that we develop new approaches to analyzing these data that explicitly take into account the length of follow-up when assessing outcomes. The first step is to characterize the pattern of the occurrence of end points (such as death). This pattern is quantified with a *survival curve*. We will now examine how to characterize survival curves and test hypotheses about them.

■ ESTIMATING THE SURVIVAL CURVE

When discussing survival curves, one often considers death the end point—hence, the name *survival* curves—but any well-defined end point can be used. Other common end points include relapse of a disease, need for additional

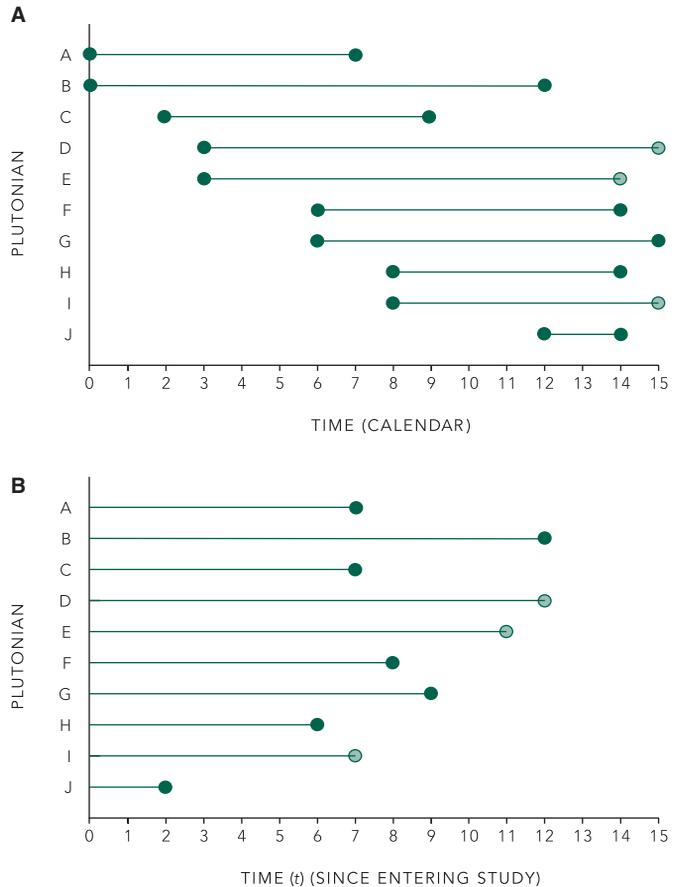


FIGURE 11-1. (A) This graph shows the observations in our study of the effect of hanging out in a smoky bar on Plutonians. The horizontal axis represents calendar time, with Plutonians entering the study at various times, when tobacco smoke invades their bars. Solid points indicate known times. Lighter points indicate the time at which observations are censored. Seven of the Plutonians die during the study (A, B, C, F, G, H, and J), so we know how long they were breathing secondhand smoke when they expired. Two of the Plutonians were still alive when the study ended at time 15 (D and I), and one (E) was lost to observation during the study, so we know that they lived at least as long as we were able to observe them, but do not know their actual time of death. **(B)** This graph shows the same data as panel A, except that the horizontal axis is the length of time each subject was observed after they entered the study, rather than calendar time.

treatment, or failure of a mechanical component of a machine. Survival curves can also be used to study the length of time to desirable events as well, such as time to pregnancy in couples having fertility problems. We will generally talk in terms of the death end point, recognizing that these other end points are also possible.

The parameter of the underlying population we seek to estimate is the *survival function*, which is the fraction of individuals who are alive at time 0 who are surviving at any given time. Specifically,

the survival function, $S(t)$, is the probability of an individual in the population surviving beyond time t .

In mathematical terms, the survival function is

$$S(t) = \frac{\text{Number of individuals surviving longer than time } t}{\text{Total number of individuals in population}}$$

Figure 11-2 shows a hypothetical survival function for a population. Note that it starts at 1 (or 100% alive) at time $t = 0$ and falls to 0% over time, as members of the population die off. The time at which half the population is alive and half is dead is called the *median survival time*.

Our goal is to estimate the survival function from a sample. Note that it is only possible to estimate the entire survival curve if the study lasts long enough for all members of the sample to die. When we are able to follow every member of a sample until all of them die, estimating the survival curve is easy: Simply compute the fraction of surviving individuals at each time someone dies. In this case, the estimate of the survival function from the data would simply be

$$\hat{S}(t) = \frac{\text{Number of individuals surviving longer than time } t}{\text{Total number of individuals in sample}}$$

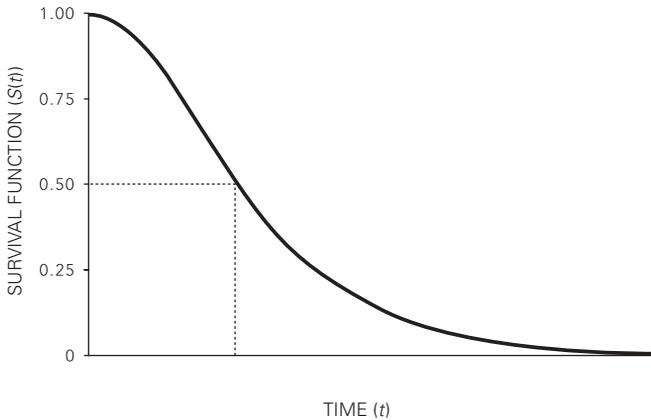


FIGURE 11-2. All population survival curves begin at 1 (100%) at time 0, when all the individuals in the study are alive, and falls to 0 as individuals die over time. The time at which 50% of the population has died is the *median survival time*.

where $\hat{S}(t)$ is the estimate of the population survival function computed from the observations in the sample.

Unfortunately, as we have already seen on Pluto, we often do not know the length of time every individual in the sample lives, so we cannot use this approach. In particular, we need a method to estimate the survival curve from real data in the presence of censoring, when we do not know the precise times of death of all the individuals in the sample. To estimate the survival function from censored data, we need to compute the probability of surviving at each time we observe a death, based on the number of individuals *known* to be surviving immediately before that death.

The first step in estimating the survival function is to list all the observations in the order of the time of death or the last available observation for each individual. Table 11-1 shows these results for the data in Figure 11-1, in the order that death or loss to follow up occurred. Uncensored observations (where the actual time of death is known) are listed before censored observations. Censored observations are indicated with a “+,” indicating that the time of death is some unknown time after the last time at which the subject was observed. For example, the first death took place (Plutonian J) at time 2, and the second death (Plutonian H) took place at time 6. Two Plutonians (A and C) died at time period 7, and one more observation (Plutonian I) *after* time 7. Thus, we know that Plutonian I lived longer than J, H, A, and C, *but we do not know how much longer*.

The second step is to estimate the probability of death within any time period, based on the number of subjects that survive to the beginning of each time period. Thus, just

before the first Plutonian (J) dies at time 2, there are 10 Plutonians alive right before J dies. Since one dies at time 2, there are $10 - 1 = 9$ survivors. Thus, our best estimate of the probability of surviving *past* time 2 *if alive just before time 2* is

$$\text{Fraction alive just before time 2 surviving past time 2} = \frac{n_2 - d_2}{n_2} = \frac{10 - 1}{10} = \frac{9}{10} = 0.900$$

where n_2 is the number of individuals alive just *before* time 2 and d_2 is the number of deaths *at* time 2. At the beginning of the time interval ending at time 2, 100% of the

■ **TABLE 11-1.** Pattern of Deaths over Time for Plutonians after Starting to Go to Smoky Bars

Plutonian	Survival Time, t_i	Number Alive at Beginning of Interval, n_i	Number of Deaths at End of Interval, d_i
J	2	10	1
H	6	9	1
A and C	7	8	2
I	7+		
F	8	5	1
G	9	4	1
E	11+		
B	12	2	1
D	12+		

■ TABLE 11-2. Estimation of Survival Curve for Plutonians after Starting to Go to Smoky Bars

Plutonian	Survival Time, t_i	Number Alive at Beginning of Interval, n_i	Number of Deaths at End of Interval, d_i	Fraction Surviving Interval, $(n_i - d_i)/n_i$	Cumulative Survival Rate, $\hat{S}(t)$
J	2	10	1	0.900	0.900
H	6	9	1	0.889	0.800
A and C	7	8	2	0.750	0.600
I	7+				
F	8	5	1	0.800	0.480
G	9	4	1	0.750	0.360
E	11+				
B	12	2	1	0.500	0.180
D	12+				

Plutonians are alive, so the estimate of the cumulative survival rate at time 2, $\hat{S}(2)$, is $1.000 \times 0.900 = 0.900$.

Next, we move to the time of the next death, at time 6. One Plutonian dies at time 6 and there are 9 Plutonians alive immediately before time 6. The estimate of the probability of surviving past time 6 if one is alive just before time 6 is

$$\text{Fraction alive just before time 6 surviving past time 6} = \frac{n_6 - d_6}{n_6} = \frac{9 - 1}{9} = \frac{8}{9} = 0.889$$

At the beginning of the time interval ending at time 6, 90% of the Plutonians are alive, so the estimate of the cumulative survival rate at time 6, $\hat{S}(6)$, is $0.900 \times 0.889 = 0.800$. Table 11-2 summarizes these calculations.

Likewise, just before time 7 there are 8 Plutonians alive and 2 die at time 7. Thus,

$$\text{Fraction alive just before time 7 surviving past time 7} = \frac{n_7 - d_7}{n_7} = \frac{8 - 2}{8} = \frac{6}{8} = 0.750$$

At the beginning of the time interval ending at time 7, 80% of the Plutonians are alive, so the estimate of the cumulative survival rate at time 7, $\hat{S}(7)$, is $0.800 \times 0.750 = 0.600$.

Up to this point, the calculations probably seem unnecessarily complex. After all, at time 7 there are 6 survivors out of 10 original individuals in the study, so why not simply compute the survival estimate as $6/10 = 0.600$? The answer to this question becomes clear after time 7, when

we encounter our first censored observation. Because of censoring, we know that Plutonian I died sometime *after* time 7, but we do not know exactly when.

The next known death occurs at time 8, when Plutonian F dies. Because of the censoring of Plutonian I, who was last observed alive at time 7, we do not know whether this individual is alive or dead at time 8. As a result, we must drop Plutonian I from the calculation of the survival function. Just before time 8, there are 5 Plutonians *known* to be alive when one dies at time 8, so, following the procedure outlined previously

$$\text{Fraction alive just before time 8 surviving past time 8} = \frac{n_8 - d_8}{n_8} = \frac{5 - 1}{5} = \frac{4}{5} = 0.800$$

At the beginning of the time interval ending at time 8, 60% of the Plutonians are known to be alive, so the estimate of the cumulative survival rate at time 8, $\hat{S}(8)$, is $0.600 \times 0.800 = 0.480$. Because of the censoring, it would be impossible to estimate the survival function based on all the Plutonians who initially entered the study.

Table 11-2 presents the remainder of the computations to estimate the survival curve. This approach is known as the *Kaplan–Meier product-limit estimate* of the survival curve. The general formula for the Kaplan–Meier product-limit estimate of the survival curve is

$$\hat{S}(t_j) = \prod \left(\frac{n_i - d_i}{n_i} \right)$$

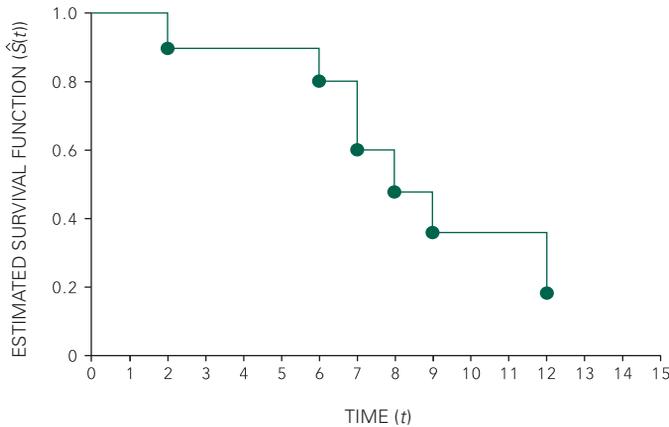


FIGURE 11-3. The survival curve for Plutonians hanging out in smoky bars, computed from the data in Table 11-1 as outlined in Table 11-2. Note that the curve is a series of horizontal lines, with the drops in survival at the times of known deaths. The curve ends at 12 months because that is the survival time of the last person known to be alive is at 12 months (Plutonian D).

where there are n_i individuals alive just before time t_i and d_i deaths occur at time t_i . The Π symbol indicates the product* taken over all the times, t_j , at which deaths occurred up to and including time t_j . (Note that the survival curve is *not* estimated at the times of censored observations because no known deaths occur at those times.) For example,

$$\hat{S}(7) = \left(\frac{10-1}{10}\right) \left(\frac{9-1}{9}\right) \left(\frac{8-2}{8}\right) = 0.600$$

Figure 11-3 shows a plot of the results. By convention, the survival function is drawn as a series of step changes, with the steps occurring at the times of known deaths. The curve ends at the time of the last observation, whether censored or not. Note that the curve, as all survival curves, begins at 1.0 and falls toward 0 as individuals die. Because one individual is still alive at the end of the study period, the data are censored and the estimated survival curve does not reach 0 during the time observations were available.

Median Survival Time

It is often desirable to provide a statistic that summarizes a survival curve with a single number. Because the survival times tend to be positively skewed, the *median survival time* is generally used. After the survival curve has been estimated, it is simple to estimate the median survival time.

The median survival time is defined to be the smallest observed survival time for which the estimated survival function is less than .5.[†]

For example, in our study of the effect of secondhand smoke on Plutonians, the median survival time is 8 months, because that is the first time at which the survival function drops below .5. (It equals .480.) If fewer than half the individuals in the study die before the end of the study, it is not possible to estimate the median survival time. Other percentiles of the survival time are estimated analogously.

Standard Errors and Confidence Limits for the Survival Curve

Like all statistics, which are based on random samples drawn from underlying populations, there is a sampling distribution of the statistic around the population parameter, in this case, the true survival function, $S(t)$. The standard deviation of the sampling distribution is estimated by the standard error of the survival function. The standard error of the estimate of the survival curve can be estimated with the following equation, known as *Greenwood's formula*:[‡]

$$s_{\hat{S}(t_j)} = \hat{S}(t_j) \sqrt{\sum \frac{d_i}{n_i(n_i - d_i)}}$$

where the summation (indicated by Σ) extends over all times, t_j , at which deaths occurred up to and including time

*The Π symbol for multiplication is used similarly to the symbol Σ for sums.

[†]An alternative approach is to connect the two observed values above and below .5 with a straight line and read the time that corresponds to $\hat{S}(t) = .5$ off the resulting line.

[‡]For a derivation of Greenwood's formula, see Collett D. *Modelling Survival Data in Medical Research*. London: Chapman and Hall; 1994, 22–26.

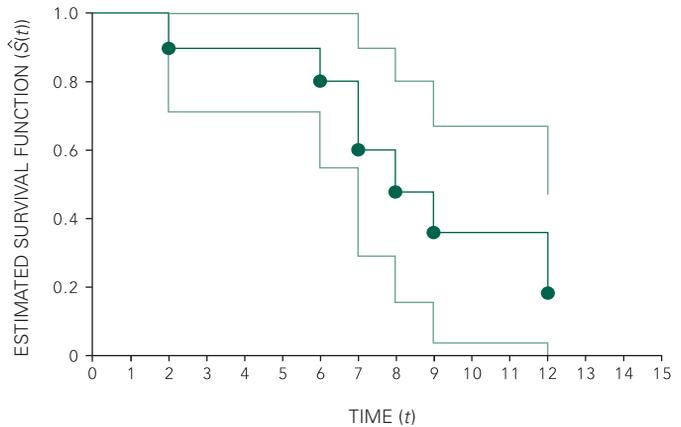


FIGURE 11-4. Survival curve for Plutonians hanging out in smoky bars, together with the 95% confidence interval (computed in Table 11-3). The upper and lower bounds of the 95% confidence interval are shown as light lines.

t_j . As with estimates of the survival curve itself, the standard error is only computed using times at which actual deaths occur. For example, the standard error for the estimated value of the survival function for the Plutonians going to smoky bars at 7 months is (using the results from Table 11-2)

$$s_{\hat{S}(7)} = .600 \sqrt{\frac{1}{10(10-1)} + \frac{1}{9(9-1)} + \frac{2}{8(8-2)}} = .155$$

Table 11-3 shows all the computations for the standard errors of the survival curve using the data in Table 11-2.

The standard error can be used to compute a confidence interval for the survival function, just as we used the standard error to compute a confidence interval for rates and proportions in Chapter 7. Recall that we defined the $100(1 - \alpha)$ percent confidence interval for a proportion to be

$$\hat{p} - z_\alpha s_{\hat{p}} < p < \hat{p} + z_\alpha s_{\hat{p}}$$

where z_α is the two-tail critical value of the standard normal distribution that defines the most α extreme values, \hat{p} is the observed proportion with the characteristic of interest, and $s_{\hat{p}}$ is its standard error. Analogously, we define the $100(1 - \alpha)$ percent confidence interval for the survival curve at time t_j to be

$$\hat{S}(t_j) - z_\alpha s_{\hat{S}(t_j)} < S(t_j) < \hat{S}(t_j) + z_\alpha s_{\hat{S}(t_j)}$$

To obtain the 95% confidence intervals, $\alpha = 0.05$, and $z_\alpha = 1.960$. Table 11-3 and Figure 11-4 show the estimated survival curve for Plutonians exposed to secondhand smoke in bars. Note the confidence interval widens as time progresses because the number of individuals remaining in the study that form the basis for the estimate of $S(t)$ falls as people die.

As with computation of the confidence intervals for rates and proportions, this normal approximation works

well when the observed values of the survival function are not near 1 or 0, in which case the confidence interval is no longer symmetric (see Fig. 7-4 and the associated discussion). As a result, applying the previous formula for values of $\hat{S}(t)$ near 1 or zero will yield confidence intervals that extend above 1 or below 0, which cannot be correct. From a pragmatic point of view, one can often simply truncate the intervals at 1 and 0 without introducing serious errors.*

■ COMPARING TWO SURVIVAL CURVES†

The end goal of much of medical practice is to prolong life, so the need to compare survival curves for groups of people receiving different treatments naturally arises in many clinical studies. We discuss how to compare the survival curves for two groups of different patients receiving different treatments. The null hypothesis we will test is that the treatments have the same effect on the pattern of survival, that

*A better way to deal with this problem is to transform the observed survival curve according to $\ln[-\ln \hat{S}(t)]$, which is not bounded by 0 and 1, compute the standard error of the transformed variable, then transform the result back into the survival function. The standard error of the transformed survival function is

$$s_{\ln[-\ln \hat{S}(t)]} = \sqrt{\frac{1}{[\ln \hat{S}(t)]^2} \sum \frac{d_i}{n_i(n_i - d_i)}}$$

The $100(1 - \alpha)$ percent confidence interval for $S(t)$ is

$$\hat{S}(t)^{\exp(-z_\alpha s_{\ln[-\ln \hat{S}(t)]})} < S(t) < \hat{S}(t)^{\exp(+z_\alpha s_{\ln[-\ln \hat{S}(t)]})}$$

†There are methods for comparing more than two survival curves that are direct generalizations of the methods discussed in this book. The computations, however, require a computer and the use of more advanced mathematical notation (in particular, matrix notation), which is beyond the scope of this text.

■ **TABLE 11-3. Estimation of Standard Error of Survival Curve and 95% Confidence Interval (CI) for Survival Curve for Plutonians after Starting to Go to Smoky Bars**

Plutonian	Survival Time, t_i	Number Alive at Beginning of Interval, n_i	Number of Deaths at End of Interval, d_i	Fraction Surviving Interval, $(n_i - d_i)/n_i$	Cumulative Survival Rate, $\hat{S}(t)$	$\frac{d_i}{n_i(n_i - d_i)}$	Standard Error, $S_{\hat{S}(t)}$	Lower 95% CI	Upper 95% CI
J	2	10	1	0.900	0.900	0.011	0.095	0.714	1.000*
H	6	9	1	0.889	0.800	0.014	0.126	0.552	1.000*
A and C	7	8	2	0.750	0.600	0.042	0.155	0.296	0.904
I	7+								
F	8	5	1	0.800	0.480	0.050	0.164	0.159	0.801
G	9	4	1	0.750	0.360	0.083	0.161	0.044	0.676
E	11+								
B	12	2	1	0.500	0.180	0.500	0.151	0.000*	0.475
D	12+								

*The computed values were truncated at 1 and 0 because the survival function cannot go above 1 or below 0.

is, that the two groups of people are drawn from the same population. If all individuals in the study are followed up for the same length of time and there are no censored observations, we could simply analyze the data using contingency tables as described in Chapter 5. If all individuals are followed up until death (or whatever the defining event is), we could compare the time to deaths observed in the different groups using nonparametric methods, such as the Mann-Whitney rank-sum test or Kruskal-Wallis analysis of variance based on ranks, described in Chapter 10. Unfortunately, in clinical studies of different treatments, these situations rarely hold. People are often lost to follow-up and the study often ends while many of the people in the study are still alive (or event free). As a result, some of the observations are censored and we need to develop appropriate statistical hypothesis testing procedures that will account for the censored data. We will use the *log rank test*.

There are three assumptions that underlie the log rank test.

1. The two samples are independent random samples.
2. The censoring patterns for the observations are the same in both samples.
3. The two population survival curves exhibit proportional hazards, so that they are related to each other according to $S_2(t) = [S_1(t)]^\psi$ where ψ is a constant called the hazard ratio.

Note that if the two survival curves are identical, $\psi = 1$. If $\psi < 1$, people in group 2 die more slowly than people in group 1, and if $\psi > 1$, people in group 2 die more quickly than people in group 1. The hazard function is the probability that an individual who has survived until time t dies at time t .^{*} Hence, the assumption of proportional hazards

^{*}The mathematical definition of the hazard function is

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\text{Probability an individual alive at time } t \text{ dies between } t \text{ and } t + \Delta t}{\Delta t}$$

The hazard function is related to the survival function according to

$$h(t) = \frac{f(t)}{S(t)}$$

where $f(t)$ is the probability density function corresponding to the failure function, $F(t) = 1 - S(t)$. The failure function begins at 0 and increases to 1, as all the members of the population die. For a discussion of these representations of the survival curve and their use, see Lee ET. *Statistical Methods for Survival Data Analysis*, 3rd ed. New York: Wiley; 2003.

means that the probability of dying at time t for individuals who have lived up to that point is a constant proportion between the two test groups.

Bone Marrow Transplantation to Treat Adult Leukemia

Acute lymphoblastic leukemia is a form of cancer in which a cancerous mutation of a lymph cell leads to greatly increased numbers of white blood cells (leukocytes). These leukemic white blood cells, however, are usually not functional in terms of the usual protections that white blood cells provide the body. At the same time, the cancerous tissue usually spreads to the bone marrow, where it interferes with the normal production of red blood cells, together with other adverse effects. The destruction of the bone marrow's ability to produce blood cells often leads to severe anemia (lack of red blood cells), which is one of the most common reasons people with this disease die.

This form of leukemia is treated through a combination of radiation and chemotherapy, which is effective in preventing recurrence in children. In adults, however, the chances of recurrence of the disease are high, even after the disease has been put into remission through chemotherapy and radiation. The chemotherapy and radiation are toxic not only to the cancer cells but also to many normal cells. In particular, at the doses used in adults, these treatments often destroy the normal bone marrow's ability to produce red blood cells. This side effect of the cancer treatment is treated by giving the person with leukemia a bone marrow transplant to reestablish function of the bone after the end of the chemotherapy and radiation. This bone marrow transplant ideally comes from a sibling who has the same type of bone marrow, a so-called *allogenic transplant*. Unfortunately, not everyone has an available sibling with matching tissue type to serve as a donor. Another option is to remove bone marrow from the person with cancer, treat the marrow with drugs in an effort to kill any residual cancer cells, preserve the "cleaned" marrow, and then inject it back into the person after the end of chemotherapy and radiation, a so-called *autologous transplant*.[†] N. Vey and

[†]Note that, because of ethical considerations and the fact that many of the people simply did not have appropriate siblings to serve as bone marrow donors, the investigators could not randomize the people in the study. They did, however, demonstrate that the two groups of people were similar in important clinical respects. This procedure is a common and reasonable way to deal with the fact that sometimes randomization is simply not possible. (See further discussion of these issues in Chapter 12.)

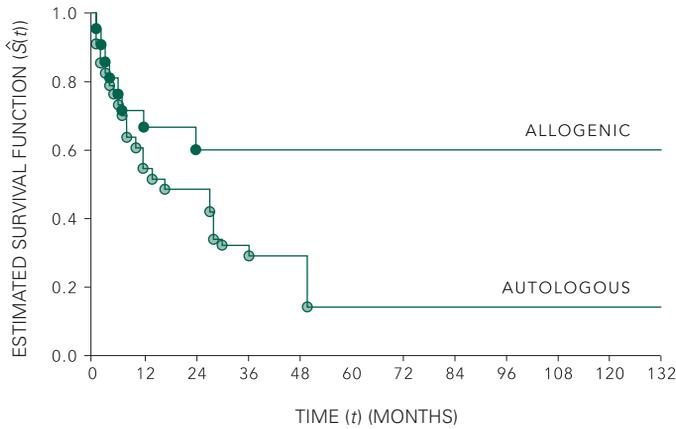


FIGURE 11-5. Survival curves for adults with leukemia who received autologous or allogenic bone marrow transplants (according to data in Table 11-4; survival curve computations are in Table 11-5). The curves extend to 132 months because that is the survival time of the last observation (even though the subsequent observations are censored immediately after that time because the study ended).

colleagues* asked the question: Is there a difference in the survival patterns of the people who receive allogenic bone marrow transplants compared to those who receive autologous transplants?

To be included in the study, patients had to have a clear diagnosis of acute lymphoblastic leukemia that involved at least 30% of their bone marrow and had to have achieved a first complete remission before receiving their bone marrow transplant. Everyone was treated using the same treatment protocols. Patients who had a willing sibling with compatible bone marrow received an allogenic transplant and the remaining people received autologous transplants. Vey and colleagues observed the two groups of people for 11 years.

Table 11-4 shows the data we seek to analyze, Table 11-5 shows the computation of the survival curves for the two groups of people, and Figure 11-5 shows the survival curves. Examining this figure suggests that an allogenic transplant from a sibling leads to better survival than an autologous transplant from the cancer patient to himself or herself. The question remains, however, whether this difference is simply due to random sampling variation. Our null hypothesis is that there is no difference in the underlying populations represented by the two treatment groups.

The first step in constructing the test statistic used in the log rank test is to consider the patterns of death in the two groups at each time a death occurs in either group.

TABLE 11-4. Time to Death (or Lost to Follow-Up) for People Receiving Autologous and Allogenic Bone Marrow Transplants

Autologous Transplant (n = 33)		Allogenic Transplant (n = 21)	
Month	Deaths or Lost to Follow-up	Month	Deaths or Lost to Follow-up
1	3	1	1
2	2	2	1
3	1	3	1
4	1	4	1
5	1	6	1
6	1	7	1
7	1	12	1
8	2	15+	1
10	1	20+	1
12	2	21+	1
14	1	24	1
17	1	30+	1
20+	1	60+	1
27	2	85+	2
28	1	86+	1
30	2	87+	1
36	1	90+	1
38+	1	100+	1
40+	1	119+	1
45+	1	132+	1
50	3		
63+	1		
132+	2		

*Vey N, Blaise D, Stoppa AM, Bouaballah R, Lafage M, Sainty D, Cowan D, Viens P, Lepeu G, Blanc AP, Jaubert D, Gaud C, Mannoni P, Camerlo J, Resbeut M, Gastaut JA, Maraninchi D. Bone marrow transplantation in 63 adult patients with acute lymphoblastic leukemia in first complete remission. *Bone Marrow Transplant.* 1994;14:383–388.

■ TABLE 11-5. Computation of the Two Survival Curves Using the Data in Table 11-4

Autologous Bone Marrow Transplant					Allogenic Bone Marrow Transplant				
Month, t_i	Number of Deaths at End of Interval d_i or Lost to Follow-up	Number Alive at Beginning of Interval, n_i	Fraction Surviving Interval, $(n_i - d_i)/n_i$	Cumulative Survival Rate, $\hat{S}_{\text{autologous}}(t)$	Month t_i	Number of Deaths at end of Interval, d_i or Lost to Follow-up	Number Alive at Beginning of Interval, n_i	Fraction Surviving Interval, $(n_i - d_i)/n_i$	Cumulative Survival Rate, $\hat{S}_{\text{allogenic}}(t)$
1	3	33	0.909	0.909	1	1	21	0.952	0.952
2	2	30	0.933	0.848	2	1	20	0.950	0.904
3	1	28	0.964	0.817	3	1	19	0.947	0.857
4	1	27	0.963	0.787	4	1	18	0.944	0.809
5	1	26	0.962	0.757	6	1	17	0.941	0.762
6	1	25	0.960	0.727	7	1	16	0.938	0.714
7	1	24	0.958	0.697	12	1	15	0.933	0.666
8	2	23	0.913	0.636	15+	1	14		
10	1	21	0.952	0.605	20+	1	13		
12	2	20	0.900	0.545	21+	1	12		
14	1	18	0.944	0.514	24	1	11	0.909	0.605
17	1	17	0.941	0.484	30+		10		
20+	1	16			60+	1	9		
27	2	15	0.867	0.420	85+	2	8		
28	1	13	0.923	0.388	86+	1	6		
30	2	12	0.833	0.323	87+	1	5		
36	1	10	0.900	0.291	90+	1	4		
38+	1	9			100+	1	3		
40+	1	8			119+	1	2		
45+	1	7			132+	1	1		
50+	3	6	0.500	0.145					
63+	1	3							
132+	2	2							

■ TABLE 11-6. Computation of Log Rank Test to Compare Survival Curves for Autologous and Allogenic Bone Marrow Transplants

Month, t_i	Autologous		Allogenic		Total		Fraction of All People who Die $d_{total,i}/n_{total,i} = f_i$	Expected Number of Autologous Deaths, $n_{autologous,i} f_i = e_i$	Observed Minus Expected Autologous Deaths, $d_{autologous,i} - e_i$	Contribution to Standard Error of U_L (see text)
	Deaths at End of Interval, $d_{autologous,i}$	Number Alive at Beginning of Interval, $n_{autologous,i}$	Deaths at End of Interval, $d_{allogenic,i}$	Number Alive at Beginning of Interval, $n_{allogenic,i}$	Deaths at End of Interval, $d_{total,i}$	Number Alive at Beginning of Interval, $n_{total,i}$				
1	3	33	1	21	4	54	0.074	2.444	0.556	0.897
2	2	30	1	20	3	50	0.060	1.800	0.200	0.691
3	1	28	1	19	2	47	0.043	1.191	-0.191	0.471
4	1	27	1	18	2	45	0.044	1.200	-0.200	0.469
5	1	26	0	17	1	43	0.023	0.605	0.395	0.239
6	1	25	1	17	2	42	0.048	1.190	-0.190	0.470
7	1	24	1	16	2	40	0.050	1.200	-0.200	0.468
8	2	23	0	15	2	38	0.053	1.211	0.789	0.465
10	1	21	0	15	1	36	0.028	0.583	0.417	0.243
12	2	20	1	15	3	35	0.086	1.714	0.286	0.691
14	1	18	0	14	1	32	0.031	0.563	0.438	0.246
17	1	17	0	13	1	31	0.032	0.548	0.452	0.248
24	0	15	1	11	1	26	0.037	0.593	-0.593	0.241
27	2	15	0	10	2	25	0.080	1.200	0.800	0.460
28	1	13	0	10	1	23	0.044	0.572	0.435	0.246
30	2	12	0	10	2	22	0.091	1.091	0.909	0.472
36	1	10	0	9	1	19	0.053	0.526	0.474	0.249
50	3	6	0	9	3	15	0.200	1.200	1.800	0.617
								Total	$U_L = 6.575$	$S_{U_L}^2 = 7.884$

Table 11-6 summarizes all the deaths actually observed in the study using the information in Table 11-5. (Censored observations are not listed in this table.) One month following the bone marrow transplantation, 3 of the 33 people who had autologous transplants died, compared to 1 of 21 of the people who had allogenic transplants. How does this pattern compare with what would be expected by chance?

There are a total of $3 + 1 = 4$ deaths out of a total of $33 + 21 = 54$ people alive before the end of month 1 in the study. Thus, $4/54 = 0.074 = 7.4\%$ of all the people died, regardless of the kind of bone marrow transplant that was received. Thus, if the type of bone marrow transplantation did not matter, we would expect that 7.4% of the 33 people who received autologous transplants, $0.074 \times 33 = 2.444$ people, to die at the end of month 1. This expected number of deaths compares to the observed 3 autologous transplant patients who died at month 1. If there is no difference between the patterns of survival between the two treatments, the observed and expected number of deaths at each time of a death should be similar for the autologous transplant patients.

To quantify the overall difference between the observed and expected number of deaths in the autologous group, we first compute the expected number of deaths at the time each death is observed in *either* group, then sum these differences up. In terms of equations, the expected number of deaths in the autologous group at time t_i is

$$e_{\text{autologous},i} = \frac{n_{\text{autologous},i} d_{\text{total}}}{n_{\text{total}}}$$

where $n_{\text{autologous},i}$ is the number of people who are known to be alive in the autologous transplant group immediately before time t_i , d_{total} is the total number of deaths in both groups at time t_i , and n_{total} is the total number of people who are known to be alive at immediately before time t_i .

Note that, while we do not explicitly include the censored observations in our summation, the censored observations do affect the results because they are included in the n s before the time at which they are censored. For example, the number of people in the allogenic transplant group known to be alive at the beginning of month 17 drops from 15 to 14 even though there were no known deaths in this group at this time because one of the patients in this group was lost to observation (censored) after month 15. The log rank test uses the censored observations up to the time that they are censored because they contribute to the number of people at risk when deaths

occur even though they do not explicitly appear in the calculations.

The first part of our test statistic is the sum of the differences between the observed and expected number of deaths in the autologous transplant group.

$$U_L = \sum (d_{\text{autologous},i} - e_{\text{autologous},i})$$

where the summation is over all the times at which anyone died in either group. For the study we are analyzing, $U_L = 6.575$ (Table 11-6). If this number is “small,” it would indicate that there is not much difference between the two survival curves; if it is “big,” we would reject the null hypothesis of no difference and report a difference in the survival associated with the two treatments.

As in earlier tests, we need to estimate the uncertainty associated with this sum to assess whether it is large. As in earlier tests, U_L follows a sampling distribution, which is approximately normally distributed, with variance*

$$s_{U_L}^2 = \sum \frac{n_{\text{autologous},j} n_{\text{allogenic},j} d_{\text{total},j} (n_{\text{total},j} - d_{\text{total},j})}{n_{\text{total},j}^2 (n_{\text{total},j} - 1)}$$

where the summation is over all times at which deaths occurred. The last column in Table 11-6 includes these computations; $s_{U_L}^2 = 7.884$ and $s_{U_L} = 2.808$. Finally, our test statistic is obtained by dividing the observed value of the test statistic by its standard error (the standard deviation of its sampling distribution).

$$z = \frac{U_L}{s_{U_L}} = \frac{6.575}{2.808} = 2.342$$

The test statistic is approximately normally distributed, so we compare its value with the critical values for the normal distribution (the last row in Table 4-1).[†] The critical value for the most extreme 2% of the normal distribution is 2.326, so we reject the null hypothesis of no difference in survival, $P < .02$. The allogenic bone marrow transplants are associated with better survival than the autologous bone marrow transplants. Bone marrow transplants from healthy siblings work better than autologous transplants from someone with leukemia to himself or herself.

*For a derivation of this result, see Collett D. *Modelling Survival Data in Medical Research*. London: Chapman and Hall;1994;40–42.

[†]Some people compute the test statistic as $U_L^2/s_{U_L}^2$. This test statistic follows the chi-square distribution with 1 degree of freedom. The results are identical to those as described in the main text.

This analysis could be done using either group; we simply use autologous transplants because it is the first group. Using the allogenic group as the reference group would have led to identical results.

The Yates Correction for the Log Rank Test

When we used the normal approximation to test for differences between two proportions in Chapters 5 and 10, we noted that, while the normal distribution is continuous, the actual sampling distribution of the test statistic will be discrete because we were analyzing counts. The Yates correction was applied to correct for the fact that simply using the normal approximation will yield P values that are slightly smaller than they should be. The situation is exactly the same for the log rank test, so many statisticians apply the Yates correction to the computation of the log rank statistic. The resulting test statistic (using the data in Table 11-6) is

$$z = \frac{|U_L| - \frac{1}{2}}{s_{U_L}} = \frac{6.575 - .500}{2.808} = 2.163$$

The value of the test statistic has been reduced from 2.342 to 2.163, and the associated P value increased to $P < .05$. The conclusion that the two types of bone marrow transplants have different effects on survival, however, remains unchanged (at $\alpha = .05$).

■ GEHAN'S TEST

The log rank test is not the only procedure available to compare two survival curves. Another procedure, known as *Gehan's test* is a generalization of the rank sum test. As discussed below, however, the log rank test is generally considered to be a superior method because Gehan's test can be dominated by a small number of early deaths. Gehan's test is computed by comparing every observation in the first treatment with every observation in the second treatment. For each comparison, score +1 if the second treatment *definitely* has a longer survival time than the first treatment, -1 if the first treatment *definitely* has a longer survival time than the second treatment, and 0 if censoring makes it impossible to say which treatment has a longer survival time for a given pair. Finally, sum up all the scores, to get U_W . A simpler way to compute U_W is to rank all the observations in time, and for each observation compute R_1 as the total number of observations whose survival time is *definitely* less than the current

observation. Likewise, let R_2 be the number of cases whose survival time is *definitely* longer than the current observation. (If the observation is censored, you do not know the actual survival time, so $R_2 = 0$.) Let $h = R_1 - R_2$. U_W equals the sum of all the h s associated with the first treatment group. The standard error of U_W equals

$$s_{U_W} = \sqrt{\frac{n_1 n_2 \sum h^2}{(n_1 + n_2)(n_1 + n_2 - 1)}}$$

Finally, the test statistic

$$z = \frac{U_W}{s_{U_W}}$$

is compared to the standard normal distribution to obtain a P value. (The Yates correction can also be applied to this test, just as with the log rank test.)

The log-rank test is also superior to Gehan's test if the assumption of *proportional hazards* is reasonably satisfied. If two survival functions exhibit proportional hazards, they will not cross.* Note that, because of random sampling variation, it is possible for the observed survival curves to cross, even if the underlying population survival functions exhibit proportional hazards.

■ POWER AND SAMPLE SIZE

As with all the other statistical hypothesis tests that we have considered, the power, $1 - \beta$, of a log rank test to detect a real difference in the survival functions for two treatments depends on the size of the difference to be detected, the false-positive risk one is willing to accept (Type I error, α), and the sample size. Conversely, the sample size required to detect a given difference depends on the power one is seeking and the false-positive risk one is willing to accept. For a given risk of Type I error and power, larger studies are required to detect smaller differences in survival.

In the interest of simplicity, we limit ourselves to estimating the power and sample size for the log rank test and assume that there are the same number of individuals in each of the test groups.† As with other statistical tests,

*A quick test for proportional hazards is to plot $\ln[-\ln \hat{S}_1(t)]$ and $\ln[-\ln \hat{S}_2(t)]$ against t . If the two lines are parallel, the assumption of proportional hazards is met.

†For a derivation of these results, see Freedman LS. Tables of number of patients required in clinical trials using the log rank test. *Stat Med.* 1982;1:121-129.

making the sample sizes equal yields the minimum total sample size to detect a given difference or, alternatively, yields the maximum power to detect a given difference for a given total sample size.

Power

Under the simplifying assumption of equal sample sizes just discussed, the power of a survival analysis with n people in each treatment group to detect an expected difference in steady state survival rates $S_1(\infty)$ and $S_2(\infty)$ for the two groups at the end of the study is

$$z_{1-\beta(\text{upper})} = z_{\alpha(2)} - \frac{1-\psi}{1+\psi} \sqrt{[2 - S_1(\infty) - S_2(\infty)]n}$$

where $z_{\alpha(2)}$ is the critical value of the normal distribution for a 2 tail test with $p = \alpha$ and $z_{1-\beta(\text{upper})}$ is the value of z that defines the upper (one tail) value of the normal distribution corresponding to $1 - \beta$, the desired power. Note that since $S_2(t) = [S_1(t)]^\psi$,

$$\psi = \frac{\ln S_2(\infty)}{\ln S_1(\infty)}$$

For example, suppose we are thinking of doing a study with 20 people in each treatment group in which we wish to detect a difference in survival from 30% to 60%. To compute the power of this study to detect this difference, we obtain $z_{\alpha(2)} = z_{.05(2)} = 1.960$ from Table 4-1 and compute

$$\psi = \frac{\ln S_2(\infty)}{\ln S_1(\infty)} = \frac{\ln .6}{\ln .3} = \frac{-.511}{-1.203} = .425$$

Therefore,

$$z_{1-\beta(\text{upper})} = 1.960 - \frac{(1-.425)}{(1+.425)} \sqrt{([2-.3-.6]20)} = .065$$

From Table 6-2, $z_{1-\beta(\text{upper})} = .065$ defines the upper .47 of the normal distribution, so the power of this study to detect the specified change is .47.

Sample Size

To compute the sample size needed to achieve a given power, we first estimate the total number of deaths (or other events we are treating as the outcome variable) that must be observed. The total number of deaths, d , required is

$$d = (z_{\alpha(2)} - z_{1-\beta(\text{upper})})^2 \left(\frac{1+\psi}{1-\psi} \right)^2$$

Once we have the required number of deaths, d , we can compute the required sample size, n , for each experimental group, given

$$n = \frac{d}{2 - S_1(\infty) - S_2(\infty)}$$

Therefore, we can estimate the sample size based on the expected survival in the two treatment groups at the end of the study.

For example, suppose we wanted to determine the sample size necessary to achieve a $1 - \beta = .80$ power in the study discussed above in which we wish to detect a difference in survival from 30% to 60% at the end of the study, with $\alpha = 0.05$ and power, $1 - \beta = .8$. From Table 4-1 $z_{\alpha(2)} = z_{.05(2)} = 1.960$ from Table 6-2, $z_{1-\beta(\text{upper})} = z_{.80(\text{upper})} = -.842$, and, as before, $\psi = .425$. Substituting into the formula for the number of deaths above,

$$\begin{aligned} d &= (z_{\alpha(2)} - z_{1-\beta(\text{upper})})^2 \left(\frac{1+\psi}{1-\psi} \right)^2 \\ &= (1.960 + .842)^2 \left(\frac{1+.425}{1-.425} \right)^2 = 48.2 \end{aligned}$$

So, we need a total of 49 deaths. To obtain this number of deaths, the number of individuals required in each of the two samples would be

$$n = \frac{d}{2 - S_1(\infty) - S_2(\infty)} = \frac{49}{2 - .3 - .6} = 44.5$$

Thus, we need 45 individuals in each group, for a total sample size of 90 individuals to achieve the desired power.

SUMMARY

This chapter developed procedures to describe outcome patterns in clinical trials where people are observed over time until a discrete event, such as death, occurs. Such trials are gaining in importance as cost pressures demand that medical treatment be demonstrated to be effective. Analysis of such data is complicated by the fact that because of the very nature of survival studies, some of the individuals in the study live beyond the end of the study and others are lost to observation because they move or die for reasons unrelated to the disease or treatment being studied. In order to construct descriptive statistics and test hypotheses about these kind of data, we used all the available information at each time an event occurred. The procedures

■ TABLE 11-7. Survival Data for People with Metastatic Lung Cancer

Month	Death or Loss to Follow-up during Month
1	1
2	1
3	3
3+	1
4	1
5	1
6	1
7	2
8	1
9	1
10+	1
11+	2
12	2
13	1
15	1
16	3
20	3
21	1
25+	1
28	1
34	1
36+	1
48+	1
56	1
62	1
84	1

we describe in this chapter can be generalized to include more complicated experimental designs in which several different treatments are being studied.* The final chapter places all the tests we have discussed in this book in context, together with some general comments on how to assess what you read and write.

*The methods we discuss in this chapter are *nonparametric* methods because they do not make any assumptions about the shape of the survival function. There are also a variety of parametric procedures that one can use when you know that the survival function follows a known functional form.

■ TABLE 11-8. Survival of People with High and Low Instrumental Activities of Daily Living

High IADL Scores		Low IADL Scores	
Month	Death or Lost to Follow-up	Month	Death or Lost to Follow-up
14	1	6	2
20	2	12	2
24	3	18	4
25+	1	24	1
28	1	26+	1
30	2	28	4
36+	1	32	4
37+	1	34+	2
38	2	36	3
42+	1	38+	3
43+	1	42	3
48	2	46+	2
48	62	47	3
		48	2
		48+	23

■ PROBLEMS

11-1 Surgery is an accepted therapeutic approach for treating cancer patients with metastases in their lungs. Philippe Girard and colleagues[†] collected data on 35 people who had metastases removed from their lungs (see Table 11-7). Estimate the survival curve and associated 95% confidence interval.

11-2 Taking care of old people on an outpatient basis is less costly than caring for them in nursing homes or hospitals, but health professionals have expressed concern about how well it is possible to predict clinical outcomes among people cared for on an outpatient basis. As part of an investigation of predictors of death in geriatric patients, Brenda Keller and Jane Potter[‡] compared survival in people aged 78.4 ± 7.2 (SD) years who scored high on the

[†]Girard P, et al. Surgery for pulmonary metastases: who are the 10-year survivors? *Cancer*. 1994;74:2791–2797.

[‡]Keller B, Potter J. Predictors of mortality in outpatient geriatric evaluation and management clinic patients. *J Gerontol*. 1994;49:M246–M251.

■ TABLE 11-9. Time to a Psychiatric Event for People Being Treated for Bipolar Disorder

Month	Combined Therapy			Valpote		
	At Risk	Events	Lost	At Risk	Events	Lost
0	110	0	0	110	0	0
3	110	14	0	110	34	0
6	96	17	0	74	18	2
9	77	10	2	56	7	0
12	67	7	0	48	3	1
15	59	4	1	42	6	3
18	53	2	2	36	3	0
21	47	4	4	29	5	4
24	36	1	7	17	0	7
27	20	0	15	6	0	11
30	3	0	17	1	0	5
33	1	0	2	0	0	1
36	0	0	1	0	0	0

Instrumental Activities of Daily Living (IADL) scale and those who scored low. Based on the survival data in Table 11-8, is there a difference in the survival patterns of these two groups of people?

11-3 What is the sample size for each experimental group to obtain .80 power using a log rank test to detect a significant difference (with $\alpha = .05$) in steady state survival rates between .40 and .20?

11-4 The BALANCE investigators who conducted the study of drug treatments for bipolar people in Problem 5-9 also collected data on when experimental subjects had emergent mood episodes for 36 months. Table 11-9 gives the observations. What is the median time to an event for the two treatment groups? Is there a difference in the time history of emergent mood events for people receiving the two treatments?

This page intentionally left blank

What Do the Data Really Show?

The statistical methods we have been discussing permit you to estimate the certainty of statements and precision of measurements that are common in the biomedical sciences and clinical practice about a population after observing a random sample of its members. To use statistical procedures correctly one needs to use a procedure that is appropriate for the study design and the scale (i.e., interval, nominal, ordinal or survival) used to record the data. All these procedures have, at their base, the assumption that the samples were selected at random from the populations of interest. If the study as conducted does not satisfy this randomization assumption, the resulting P values and confidence intervals are meaningless.

In addition to seeing that the individuals in the sample are selected at random, there is often a question of exactly what actual populations the people in any given study represent. This question is especially important and often difficult to answer when the experimental subjects are patients in academic medical centers, a group of people hardly typical of the population as a whole. Even so, identifying the population in question is the crucial step in deciding the broader applicability of the findings of any study.

■ CELL PHONES: PUTTING ALL THE PIECES TOGETHER

Taking all the information we have discussed on cell phones and sperm allows us to confidently conclude that

exposure to cell phones adversely affects sperm. We began in Chapter 3 with two human observational studies showing lower sperm motility. The first one* showed a difference between men with lower and higher cell phone use. The second study† improved upon this design by including a true control group of men that did not use cell phones at all as well as including several levels of use and finding a dose–response relationship, with greater reductions in sperm motility associated with increased levels of cell phone use. These two studies, however, were observational, leaving open the possibility that the relationships they elucidated were actually reflecting the effects of some unobserved confounding variable. Concern over confounding variables is especially acute because all the men providing the sperm samples were recruited at fertility clinics, so, even though the investigators tried to screen out men with other reasons for reproductive problems, the possibility remained that something else than exposure to cell phone radiation was causing the reduction in sperm motility.

We increased our confidence that the cell phone radiation was actually affecting sperm when we considered an

*Fejes I, Závacki Z, Szöllösi J, Koloszar S, Daru J, Kovács L, Pál A. Is there a relationship between cell phone use and semen quality? *Arch Androl.* 2005;51:385–393.

†Agarwal A, Deepinder F, Sharma RK, Ranga G, Li J. Effect of cell phone usage on semen analysis in men attending infertility clinic: an observational study. *Fertil Steril.* 2008;89:124–128.

animal experimental study* that showed that rabbits exposed to cell phone radiation had depressed sperm motility. Unlike the earlier two human studies, these results came from an experiment in which the rabbits were randomized to the different treatments and in which the investigators controlled the environment, so we can be much more confident that the results were the result of the cell phone radiation *causing* the observed changes rather than them being a reflection of some unobserved confounding variable. The issue of interspecies extrapolation, however, remains.

We addressed this issue in Chapter 8 with the experimental study that exposed sperm from normal men to controlled levels of cell phone radiation.[†] Because the investigators recruited healthy men as volunteers – not volunteers from men attending a fertility clinic – we can be more confident that the sperm were not behaving abnormally for other reasons. Because the sperm were subjected to controlled irradiation in petri dishes, the experiment avoided the possibility that other aspects of the volunteer men’s behaviors in conjunction with the cell phone use was responsible for the observed effects. The fact that there was a dose-response relationship between the strength of cell phone exposure (measured as specific absorption rate, SAR) and the induction of reactive oxygen species in the sperm, which was, in turn, related to sperm DNA damage, provides a biological mechanism for the changes observed in the original human observational studies. The problem, however, with this experimental study is that sperm in petri dishes may respond differently than sperm in men.

Thus, we are left with the situation in which we have several pieces of evidence on the effects of cell phone exposure on sperm, all of which provide some information, but none of which is definitive and above criticism. The first two studies are realistic because the data come from real people using cell phones in real situations, but they are observational and the fact that the men being studied were attending a fertility clinic could introduce

unknown confounding variables. The rabbit study was an experiment, but rabbits are not people. The study of sperm in petri dishes was also an experiment and the sperm were from normal volunteers, but the sperm were irradiated in petri dishes, not people.

The important thing to do is to consider the *evidence as a whole*. Do all the studies generally point in the same direction? Are they consistent with each other? Do the experimental studies, which almost always are conducted in artificial environments, elucidate the biological mechanisms that explain the observational studies which, while conducted in more realistic environments, suffer from the limitation that they are observational? Conversely, do the observational studies provide results consistent with one would expect based on the biology elucidated in the experiments?

The more of these questions that you can answer “yes,” the more confident that you can be in concluding that the exposure (or treatment) *causes* the outcome. In this case, we can be very confident that cell phones are causing abnormal sperm behavior.[‡]

■ WHEN TO USE WHICH TEST

We have reached the end of our discussion of different statistical tests and procedures. It is by no means exhaustive, for there are many other approaches to problems and many kinds of experiments we have not even discussed. Nevertheless, we have developed a powerful set of tools and laid the groundwork for the statistical methods needed to analyze more complex experiments. Table 12-1 shows that it is easy to place all these statistical hypothesis testing procedures this book presents into context by considering two things: the *type of experiment or observational study* used to collect the data and the *scale of measurement*.

To determine which test to use, one needs to consider the study design. Were the treatments applied to the same or different individuals? How many treatments were there? Was the study designed to define a tendency for two variables to increase or decrease together?

How the response is measured is also important. Were the data measured on an interval scale? If so, are you

*Salama N, Kishimoto T, Kanayama H. Effects of exposure to a mobile phone on testicular function and structure in adult rabbit. *Int J Androl.* 2010;33:88–94.

[†]De Iuliis GN, Newey RJ, King BV, Aitken RJ. Mobil phone radiation induces reactive oxygen species production and DNA damage in human spermatozoa *in vitro*. *PLoS One.* 2010;4(7):e6446. doi:10.1371/journal.pone.0006446.

[‡]The papers used as examples in this book contain additional information that supports this statement, as well as the larger literature on this topic.

■ **TABLE 12-1. Summary of Some Statistical Methods to Test Hypotheses**

Scale of Measurement	Study Design				
	Two Treatment Groups Consisting of Different Individuals	Three or More Treatment Groups Consisting of Different Individuals	Before and after a Single Treatment in the Same Individuals	Multiple Treatments in the Same Individuals	Association between Two Variables
Interval (and drawn from normally distributed populations*)	Unpaired <i>t</i> test (Chapter 4)	Analysis of variance (Chapter 3)	Paired <i>t</i> test (Chapter 9)	Repeated measures analysis of variance (Chapter 9)	Linear regression, Pearson product-moment correlation, or Bland-Altman analysis (Chapter 8)
Nominal	Chi-square analysis of contingency table (Chapter 5)	Chi-square analysis of contingency table (Chapter 5)	McNemar's test (Chapter 9)	Cochrane Q [†]	Relative risk or odds ratio (Chapter 5)
Ordinal [†]	Mann-Whitney rank-sum test (Chapter 10)	Kruskal-Wallis test (Chapter 10)	Wilcoxon signed-rank test (Chapter 10)	Friedman test (Chapter 10)	Spearman rank correlation (Chapter 8)
Survival time	Log rank test or Gehan's test (Chapter 11)				

*If the assumption of normally distributed populations is not met, rank the observations and use the methods for data measured on an ordinal scale.
[†]Or interval data that are not necessarily normally distributed.

satisfied that the underlying population is normally distributed? Do the variances within the treatment groups or about a regression line appear equal? When the observations do not appear to satisfy these requirements—or if you do not wish to assume that they do—you lose little power by using nonparametric methods based on ranks. Finally, if the response is measured on a nominal scale in which the observations are simply categorized, one can analyze the results using contingency tables. If the nominal dependent variable is a survival time or the data are censored, use survival analysis.

■ ISSUES IN STUDY DESIGN

Table 12-1 comes close to summarizing the lessons of this book, but there are three important things that it excludes. First, as Chapter 6 discussed, it is important to consider the power of a test when determining whether or not the failure to reject the null hypothesis of no treatment effect is likely to be because the treatment really has no effect or because the sample size was too small for the test to detect the treatment effect. Second, Chapter 7 discussed the importance of quantifying the size of the treatment effect (with confidence intervals) in addition to the certainty with which you can reject the hypothesis that the treatment had no effect (the P value). Third, one must consider how the samples were selected and whether or not there are biases that invalidate the results of any statistical procedure, however elegant or sophisticated.

It is through these more subtle aspects of the study design that authors (and the sponsors that fund the authors) can manipulate the outcomes of a research paper. Even with correct statistical calculations, an underpowered study will not detect complications in a clinical trial of a new therapy or diseases caused by an environmental toxin such as tobacco smoke or cell phone exposure.* Establishing an inappropriate comparison group can make a test drug look better or worse. When designing or

assessing a research, it is important to consider these potential biases, as well as who sponsored the work and the investigators' relationship with the sponsor.†

■ RANDOMIZE AND CONTROL

As already noted, all the statistical procedures assume that the observations represent a sample *drawn at random* from a larger population. What, precisely, does “drawn at random” mean? It means that any specific member of the population is as likely as any other member to be selected for study and, further, that in an experiment any given individual is as likely to be selected for one sample group as the other (i.e., control or treatment). The only way to achieve randomization is to use an objective procedure, such as a table of random numbers or a random number generator, to select subjects for a sample or treatment group. When other criteria are used that permit the investigator (or participant) to influence which treatment a given individual receives, one can no longer conclude that observed differences are due to the treatment rather than *biases* introduced by the process of selecting individuals for an observational study or assigning different individuals to different groups in an experimental study. When the randomization assumption is not satisfied, the logic underlying the distributions of the test statistics (F , t , χ^2 , z , r , r_s , T , W , H , or χ_r^2) used to quantify whether the observed differences between the different treatment groups are due to chance as opposed to the treatment fails and the resulting P values (i.e., estimates that the observed differences are due to chance) are meaningless.

To reach meaningful conclusions about the efficacy of some treatment, one must compare the results obtained in the individuals who receive the treatment with an appropriate *control* group that is identical to the treatment group in all respects except the treatment. Clinical studies often fail to include adequate controls. *This omission generally biases the study in favor of the treatment.*

Despite the fact that questions of proper randomization and control are really distinct statistical questions, in practice these two areas are so closely related that we will discuss them together by considering two classic examples.

*See, for example, Tsang R, Colley L, Lynd LD. Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized clinical trials. *J Clin Epidemiol.* 2009;62:609–616; Bero LA, Barnes DB. Why review articles on the health effects of passive smoking reach different conclusions. *JAMA.* 1998;279(19):1566–70; Huss A, Egger M, Huwiler-Müntener K, Rösli M. Source of funding and results of studies of health effects of mobile phone use: systematic review of experimental studies. *Environ Health Perspect.* 2007;115:1–4.

†For more details on the issue of how to detect bias and underpowered studies and estimate its effects, see Guyat GG, Rennie D, Meade MO, Cook DJ. Why study results mislead: bias and random error. In: *Users Guide to the Medical Literature*, 2nd ed. New York: McGraw-Hill; 2008: chap 5.

Internal Mammary Artery Ligation to Treat Angina Pectoris

People with coronary artery disease develop chest pain (angina pectoris) when they exercise because the narrowed arteries cannot deliver enough blood to carry oxygen and nutrients to the heart muscle and remove waste products fast enough. Relying on some anatomical studies and clinical reports during the 1930s, some surgeons suggested that tying off (ligating) the mammary arteries would force blood into the arteries that supplied the heart and increase the amount of blood available to it. By comparison with major operations that require splitting the chest open, the procedure to ligate the internal mammary arteries is quite simple. The arteries are near the skin, and the entire procedure can be done under local anesthesia.

In 1958, J. Roderick Kitchell and colleagues* published the results of a study in which they ligated the internal mammary arteries of 50 people who had angina before the operation, then observed them for 2 to 6 months: 34 of the patients (68%) improved clinically in that they had no more chest pain (36%) or fewer and less severe attacks (32%), 11 patients (22%) showed no improvement, and 5 (10%) died. On its face, this operation seems an effective treatment for angina pectoris.

In fact, even before this study was published, the widely-read popular magazine *Reader's Digest* carried an enthusiastic description of the procedure in an article entitled "New Surgery for Ailing Hearts."[†] (This article probably did more to promote the operation than the technical medical publications.)

Yet, despite the observed symptomatic relief and popular appeal of the operation, no one uses it today. Why not?

In 1959, Leonard Cobb and colleagues[‡] published the results of a double-blind randomized controlled trial of this operation. Neither the patients nor the physicians who evaluated them knew whether or not a given patient had the internal mammary arteries ligated. When the patient reached the operating room, the surgeon made the incisions necessary to reach the internal mammary arter-

ies and isolated them. At that time, the surgeon was handed an envelope instructing whether or not actually to ligate the arteries. The treated patients had their arteries ligated, and the control patients had the wound closed without touching the artery.

When evaluated in terms of subjective improvement as well as more quantitative measures, for example, how much they could exercise before developing chest pain or the appearance of their electrocardiogram, there was little difference between the two groups of people, although there was a suggestion that the control group did better.

In other words, the improvement that Kitchell and colleagues reported was a combination of observer biases and, probably more important, the placebo effect.

The Portacaval Shunt to Treat Cirrhosis of the Liver

Alcoholics often develop cirrhosis of the liver when the liver's internal structure breaks down and increases the resistance to the flow of blood through the liver. As a result, blood pressure increases and often affects other parts of the circulation, such as the veins around the esophagus. If the pressure reaches a high enough level, these vessels can rupture, causing internal bleeding and even death. To relieve this pressure, many surgeons performed a major operation to redirect blood flow away from the liver by constructing a connection between the portal artery (which goes to the liver) and the vena cava (the large veins on the other side of the liver). This connection is called a *portacaval shunt*.

Like many medical procedures, the early studies that supported this operation were completed without controls. The investigators completed the operation on people, then watched to see how well they recovered. If their clinical condition improved, the operation was considered a success. This approach has the serious flaw of not allowing for the fact that some of the people would have been fine (or died) regardless of whether or not they had the operation.

In 1966, more than 20 years after the operation was introduced, Norman Grace and colleagues[§] examined 51 papers that sought to evaluate this procedure. They examined the nature of the control group, if one was present, whether or not patients were assigned to treatment or

*Kitchell JR, Glover R, Kyle R. Bilateral internal mammary artery ligation for angina pectoris: preliminary clinical considerations. *Am J Cardiol.* 1958;1:46–50.

[†]Ratcliff J. New surgery for ailing hearts. *Reader's Dig.* 1957;71:70–73.

[‡]Cobb L, Thomas G, Dillard D, Merendino K, Bruce R. An evaluation of internal-mammary-artery ligation by a double-blind technic. *N Engl J Med.* 1959;260:1115–1118.

[§]Grace N, Muench H, Chalmers T. The present status of shunts for portal hypertension in cirrhosis. *Gastroenterology.* 1966;50:684–691.

control at random, and how enthusiastic the authors were about the operation after they finished their study. Table 12-2 shows that the overwhelming majority of investigators who were enthusiastic about the procedure did studies that failed to include a control group or included a control group that was not the result of a random assignment of patients between control and the operation. The few investigators who included controls and adequate randomization were not enthusiastic about the operation.

The reasons for biases on behalf of the operation in the studies that did not include controls—the placebo effect and observer biases—are the same as in the study of internal mammary artery ligation we just discussed.

The situation for the 15 studies with nonrandomized controls contains some of these same difficulties, but the situation is more subtle. Specifically, there *is* a control group that provides some basis of comparison; the members of the control were not selected at random, however, but assigned on the basis of the investigators' judgment. In such cases, there is often a bias to treat only patients who are well enough to respond (or occasionally, hopeless cases). This selection procedure biases the study in behalf of (or occasionally against) the treatment under study. This bias can slip into studies in quite subtle ways. For example, suppose that you are studying some treatment and decide to assign patients who are admitted to the control and treatment groups alternately in the order in which they are admitted or on alternate days of the month. It then makes it easy for the investigators to decide which group a given person will be a member of by manipulating the day or time of admission to the hospital. The investigators may not even realize they are introducing such a bias.

A similar problem can arise in laboratory experiments. For example, suppose that you are doing a study of a potential carcinogen with rats. Simply taking rats out of a cage and assigning the first 10 rats to the control group and the next 10 rats to the treatment group (or alternate rats to the two groups) will not produce a random sample because more aggressive, or bigger, or healthier rats may, as a group, stay in the front or back of the cage.

The only way to obtain a random sample that avoids these problems is *consciously to assign the experimental subjects at random* using a table of random numbers, dice, or other procedure.

Table 12-2 shows that the four randomized trials done of the portacaval shunt showed the operation to be of

■ **TABLE 12-2. Value of Portacaval Shunt According to 51 Different Studies**

Design	Degree of Enthusiasm		
	Marked	Moderate	None
No controls	24	7	1
Controls			
Not randomized	10	3	2
Randomized	0	1	3

Adapted from Table 2 of Grace ND, Muench H, Chambers TC. The present status of shunts for portal hypertension in cirrhosis. *Gastroenterology*. 1966;50:684–691. Copyright Elsevier 1966.

little or no value. This example illustrates a common pattern: *The better the study, the less likely it is to be biased in favor of the treatment.*

The biases introduced by failure to randomize the treatments in clinical trials can be substantial. For example, Kenneth Schulz and coworkers* examined 250 controlled trials and assessed how the subjects in the studies had been allocated to the different treatment groups. A well-randomized trial was one in which subjects were assigned to treatments using a table of random digits or a random number generator or some similar process. A study was deemed to have an inadequate treatment allocation procedure if the subjects were treated based on the date they entered the study (including alternating one treatment or the other), which could be subject to manipulation by the investigators or others participating in the study. The authors found that the treatments appeared 41% better in the poorly randomized studies than in the ones with careful application of strict randomization procedures.

Thus, it is very important that the randomization be conducted using a random number generator, table of random digits, or other similarly objective procedure to avoid the introduction of serious upward biases in the estimate of how well the treatment under study works.

*Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA*. 1995;273:408–412.

Is Randomization of People Ethical?

Having concluded that the randomized clinical trial is the definitive way to assess the value of a potential therapy, we need to pause to discuss the ethical dilemma that some people feel when deciding whether or not to commit someone's treatment to a random number. The short answer to this problem is that *if no one knows* which therapy is better, there is no ethical imperative to use one therapy or another.

In reality, all therapies have their proponents and detractors, so one can rarely find a potential therapy that everyone feels neutral about at the start of a trial. (If there were no enthusiasts, no one would be interested in trying it.) As a result, it is not uncommon to hear physicians, nurses, and others protesting that some patient is being deprived of effective treatment (i.e., a therapy that the individual physician or nurse believes in) simply to answer a scientific question. Sometimes these objections are well founded, but when considering them it is important to ask: *What evidence of being right does the proponent have?* Remember that uncontrolled and nonrandomized studies tend to be biased in favor of the treatment. At the time, Cobb and colleagues' randomized controlled trial of internal mammary artery ligation may have seemed unethical to enthusiasts for the surgery on the grounds that it required depriving some people of the potential benefits of the surgery. In hindsight, however, they spared the public the pain and expense of a worthless therapy.

These genuine anxieties, as well as the possible vested interests of the proponent of the procedure, must be balanced against the possible damage and costs of subjecting the patient to a useless or harmful therapy or procedure. The same holds for the randomized controlled trials of the portacaval shunt. To complete a randomized trial, it is necessary to assess carefully just *why* you believe some treatment to have an effect.

This situation is complicated by the fact that once something becomes accepted practice, it is almost impossible to evaluate it, even though it is as much a result of tradition and belief as scientific evidence, for example, the use of leeches. To return to the theme we opened this book with, a great deal of inconvenience, pain, and money is wasted pursuing diagnostic tests and therapies that are of no demonstrated value. For example, despite the fact that the provision of mammograms in younger women has become a major American industry, there is continuing debate over precisely who it helps.

Another seemingly more difficult issue is what to do when the study suggests that the therapy is or is not effective but enough cases have not yet been accumulated to reach conventional statistical significance, that is, $P = .05$. Recall (from Chapter 6) that the power of a test to detect a difference of a specified size increases with the sample size and as the risk of erroneously concluding that there is a difference between the treatment groups (the Type I error α) increases. Recall also that α is simply the largest value of P that one is willing to accept and still conclude that there is a difference between the sample groups (in this case, that the treatment had an effect). Thus, if people object to continuing a clinical trial until the trial accumulates enough patients (and sufficient power) to reject the hypothesis of no difference between the treatment groups with $P < .05$ (or $\alpha = 5\%$), all they are really saying is that they are willing to conclude that there is a difference when P is greater than .05.* In other words, they are willing to accept a higher risk of being wrong in the assertion that the treatment was effective when, in fact, it is not, because they believe the potential benefits of the treatment make it worth pursuing despite the increased uncertainty about whether or not it is really effective. Viewed in this light, the often diffuse debates over continuing a clinical trial can be focused on the real question underlying the disagreements: How confident does one need to be that the observed difference is not due to chance before concluding that the treatment really did cause the observed differences?

The answer to this question depends on personal judgment and values, not statistical methodology.

Is a Randomized Controlled Trial Always Necessary?

No. There are some occasions, such as the introduction of penicillin, when the therapy produces such a dramatic improvement that one need not use statistical tools to estimate the probability that the observed effects are due to chance.

*When one examines the data as they accumulate in a clinical trial, one can encounter the same multiple-comparisons problem discussed in Chapters 3 and 4. Therefore, it is important to use specialized techniques called *sequential testing* that account for the fact that you are looking at the data more than once.

In addition, sometimes medical realities make it impossible to do a randomized trial. For example, in Chapter 11 we considered a study of the effects of bone marrow transplants on survival among adults with leukemia. One group of people received bone marrow transplants from a tissue-matched sibling (allogenic transplants), and the other group received bone marrow removed from themselves before beginning treatment with chemotherapy and radiation for the cancer (autologous transplants). Since everyone does not have a tissue-matched sibling that could serve as a transplant donor, it was impossible to randomize the treatments. To minimize bias in this study, however, the investigators treated all the people in the study the same and carefully matched people in the two treatment groups on other characteristics that could have affected the outcome. This situation often exists in clinical studies; it is particularly important to see that the subjects in different experimental groups are as similar as possible when a strict randomization is not possible.

There are also often accidents of nature that force attentive practitioners to reassess the value of accepted therapy. For example, Ambroise Paré, a French military surgeon, followed the accepted therapy of treating gunshot wounds with boiling oil. During a battle in Italy in 1536, he ran out of oil and simply had to dress the untreated wounds. After spending a sleepless night worrying about his patients who had been deprived of the accepted therapy, he was surprised to find them “free from vehemencie of paine to have had good rest” while the conventionally treated soldiers were feverish and tormented with pain.* History does not record whether Paré then prepared a proposal to do a randomized clinical trial to study the value of boiling oil to treat gunshot wounds. Should and would it be necessary if he had made his discovery today?

■ DOES RANDOMIZATION ENSURE CORRECT CONCLUSIONS?

The randomized controlled trial is the most convincing way to demonstrate the value of a therapy. Can you

*This example is taken from Wulff HR. *Rational Diagnosis and Treatment*. Oxford: Blackwell;1976. This excellent short book builds many bridges between the ideas we have been discussing and the diagnostic therapeutic thought processes.

assume that it will always lead to correct conclusions? No.

First, as Chapter 6 discussed, the trial may involve too few patients to have sufficient power to detect a true difference.

Second, if the investigators require $P < .05$ to conclude that the data are incompatible with the hypothesis that the treatment had no effect, in the long run 5% of the “statistically significant” effects they find will be due to chance in the random-sampling process when, in actuality, the treatment had no effect, that is, the null hypothesis is correct. (Since investigators are more likely to publish positive findings than negative findings, more than 5% of the published results are probably due to chance rather than the treatments.) This means that as you do more and more tests, you will accumulate more and more incorrect statements. When one collects a set of data and repeatedly subdivides the data into smaller and smaller subgroups for comparison, it is not uncommon to “find” a difference that is due to random variation rather than a real treatment effect.

Most clinical trials, especially those of chronic disease like coronary artery disease or diabetes, are designed to answer a single broad question dealing with the effect on survival of competing treatments. These trials involve considerable work and expense and yield a great many data, and the investigators are generally interested in gleaning as much information (and as many publications) as possible from their efforts. As a result, the sample is often divided into subgroups based on various potential prognostic variables, and the subgroups are compared for the outcome variable of interest (usually survival). This procedure inevitably yields one or more subgroups of patients in whom the therapy is effective.

To demonstrate the difficulties that can arise when one begins examining subgroups of patients in a randomized controlled trial, Kerry Lee and colleagues[†] took 1073 patients who had coronary artery disease and were being treated with medical therapy at Duke University and randomly divided them into two groups. *The “treatment” was randomization*. Therefore, if the samples are representative, one would not expect any systematic differences

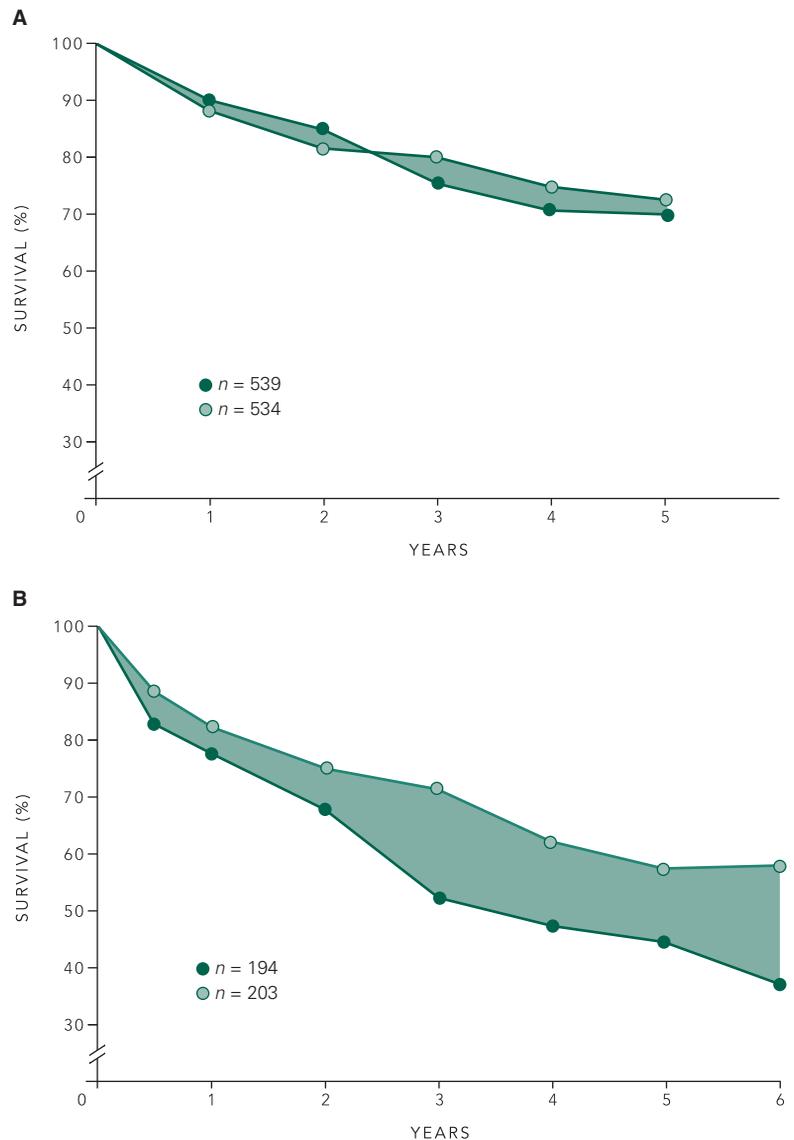
[†]Lee K, McNeer F, Starmer F, Harris P, Rosati R. Clinical judgment and statistics: lessons from a simulated randomized trial in coronary artery disease. *Circulation*. 1980;61:508–515.

between the two groups. Indeed, when they compared the two groups with respect to age, sex, medical history, electrocardiographic findings, number of blocked coronary arteries, or whether or not the heart exhibited a normal contraction pattern, using the methods this book describes, they found no significant differences between the two groups, except in the left ventricular contraction pattern. This result is not surprising, given that the two

groups were created by randomly dividing a single group into two samples. Most important, there was virtually no difference in the pattern of survival in the two groups (Fig. 12-1A). So far, this situation is analogous to a randomized clinical trial designed to compare two groups receiving different therapies.

As already noted, after going to all the trouble of collecting such data, investigators are usually interested in

FIGURE 12-1. (A) Survival over time of 1073 people with medically treated coronary artery disease who were randomly divided into two groups. As expected, there is no detectable difference. **(B)** Survival in two subgroups of the patients shown in panel A who have three-vessel disease and abnormal left ventricular function. The two different groups were selected at random and received the same medical treatment. The difference is statistically significant ($P < .025$) if one does not include a Bonferroni correction for the fact that many hypotheses were tested even though the only treatment was randomization into two groups. Survival curves appear smooth because of the large number of deaths in all cases. (Data for panel A from the text of Lee K, McNeer J, Starmer C, Harris P, Rosati R. Clinical judgment and statistics: lessons from a simulated randomized trial in coronary artery disease. *Circulation*. 1980;61:508–515, and personal communication with Dr. Lee. Panel B is reproduced from Fig. 1 of the same paper with permission from American Heart Association, Inc.)



examining various subgroups to see whether any finer distinctions can be made that will help the individual clinician deal with each individual patient according to the particular circumstances to the case. To simulate this procedure, Lee and colleagues subdivided (the technical statistical term is *stratified*) the 1073 patients into six subgroups depending on whether one, two, or three coronary arteries were blocked and whether or not the patient's left ventricle was contracting normally. They also further subdivided these six groups into subgroups based on whether or not the patient had a history of heart failure. They analyzed the resulting survival data for the 18 subgroups (6 + 12) using the techniques discussed in Chapter 11. This analysis revealed, among others, a statistically significant ($P < .025$) difference in survival between the two groups of patients who had three diseased vessels and an abnormal contraction pattern (Fig. 12-1B). How could this be? After all, *randomization was the treatment*.

This result is another aspect of the multiple-comparisons problem. Without counting the initial test of the global hypothesis that the survival in the two original sample groups is not different, Lee and colleagues completed 18 different comparisons on the data. The chances of obtaining a statistically significant result with $P < .05$, by chance is for these 18 comparisons is $\alpha_T = 1 - (1 - .05)^{18} = .60$. The result in Figure 12-1B is an example of this fact. When the total patient sample in a clinical trial is subdivided into many subgroups and the treatment compared within these subgroups, the results of these comparisons need to be interpreted very cautiously, especially when the P values are relatively large (say, around .05, as opposed to being around .001).*

This problem is not simply theoretical. Isabelle Boutron and colleagues[†] examined 72 randomized controlled trials published in December 2006 that had a clearly identified primary outcome showing statistically nonsignificant results for that outcome. They found that about two-thirds of the papers had “spun” the results to highlight that the experimental treatments were beneficial in some way by focusing on subgroup comparisons or

restricting the analysis to a subset of the population or other dubious interpretations of the data.

This exercise illustrates an important general rule for all statistical analysis: design the experiment to *minimize the total number of statistical tests of hypotheses that need to be computed*.

■ PROBLEMS WITH THE POPULATION

In most laboratory experiments and survey research, including marketing research and political polling, it is possible to define and locate the population of interest clearly and then arrange for an appropriate random sample. In contrast, in clinical research, the sample generally has to be drawn from patients and volunteers at medical centers who are willing to participate in the project. This fact can make the interpretation of the study in terms of the population as a whole quite difficult.

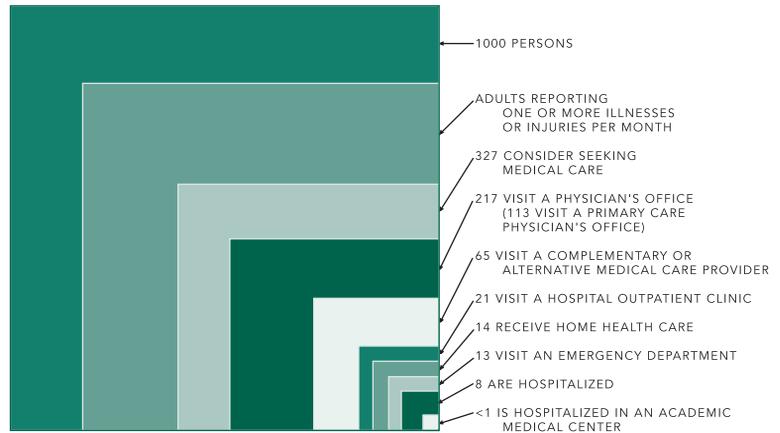
People who either attend clinics or are hospitalized at university medical centers are not really typical of the population as a whole or even the population of sick people. Figure 12-2 shows that, of 1000 people in the United States, only eight are admitted to a hospital in any given month, and *less than one* is referred to an academic medical center. It is often that one person who is available to participate in a clinical research protocol. Sometimes the population of interest consists of people with the arcane or complex problems that lead to referral to an academic medical center; in such cases, a sample consisting of such people can be considered to represent the relevant population. However, as Figure 12-2 makes clear, a sample of people drawn (even at random) from the patients at a university medical center can hardly be considered to be representative of the population as a whole. This fact must be carefully considered when evaluating a research report to decide just what population (i.e., whom) the results can be generalized to.

In addition to the fact that people treated at academic medical centers do not really represent the true spectrum of illness in the community, there is an additional difficulty due to the fact that hospitalized patients do not represent a random sample of the population as a whole. It is not uncommon for investigators to complete studies of the association between different diseases based on hospitalized patients (or patients who seek medical help as outpatients). In general, different diseases lead to different rates of hospitalization (or physician consultation). Unless extreme care is taken in analyzing the results of

*One approach to dealing with this problem would be to treat the secondary tests (along with the main hypothesis) as a family of comparisons and use the Holm-Sidak procedure to determine if it was appropriate to conclude that any observed differences were statistically significant.

[†]Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*. 2010;303:2058–2064.

FIGURE 12-2. Estimates of the number of people in the United States who report illnesses and receive various forms of health care. Less than 1 in 1000 is hospitalized in an academic medical center. (Redrawn with permission from Fig. 2 of Green LA, Fryer GE Jr, Yawn BP, Lanier D, Dovey SM., The ecology of medical care revisited. *N Engl J Med.* 2001;344:2021–2025.)



such studies to ensure that there are comparable rates of all classes of disease, any apparent association (or lack of association) between various diseases and symptoms is as likely to be due to the differential rates at which patients seek help (or die, if it is an autopsy study) as to a true association between the diseases. This problem is called *Berkson's fallacy*, after the statistician who first identified the problem.

■ HOW YOU CAN IMPROVE THINGS

Using statistical thinking to reach conclusions in clinical practice and the biomedical sciences amounts to much more than memorizing a few formulas and looking up *P* values in tables. Like all human endeavors, applying statistical procedures and interpreting the results requires insight—not only into the statistical techniques but also into the clinical or scientific question to be answered. As discussed in Chapter 1, these methods will continue to increase in importance as economic pressures grow for evidence that diagnostic procedures and therapies actually are worth the cost both to the individual patient and to society at large. Statistical arguments play a central role in many of these discussions.

Even so, the statistical aspects of most medical research are supervised by investigators who only have heard of *t* tests (and, perhaps, contingency tables) regardless of the nature of the experimental design and the data. Since the investigators themselves best know what they are trying to establish and are responsible for drawing the conclusions, they should take the lead in the analysis of the data.

Unfortunately, this task often falls to a laboratory technician or statistical consultant who does not really understand the question at hand or the data collected.

This problem is aggravated by the fact that investigators often go into the clinic or laboratory and collect data before clearly thinking out the specific question they wish to answer. As a result, after the data are collected and the investigators begin searching for a *P* value (often under the pressure of the deadline for submitting an abstract to a scientific meeting), they run into the fact that *P* values are associated with statistical *hypothesis tests* and that in order to test a hypothesis, you need a hypothesis to test.

As discussed earlier in this chapter, the hypothesis (as embodied in the type of experiment or observational study) combined with the scale of measurement determines the statistical method to be used. Armed with a clearly stated hypothesis, it is relatively straightforward to design an observational study or experiment and determine the method of statistical analysis to be applied before one starts collecting the data. The simplest procedure is to make up the table that will contain the data before collecting it, assume that you have the numbers, and then determine the method of analysis. This exercise will ensure that after going to the trouble and expense of actually collecting the data, it will be possible to analyze it.

While this procedure may seem obvious, very few people follow it. As a result, problems often arise when the time comes to compute the prized *P* value, because the study design does not fit the hypothesis—which is finally verbalized when a feisty statistician demands

it—or the design does not fit into the paradigm associated with one of the established statistical hypothesis tests. (This problem is especially acute when dealing with more complex study designs.) Faced with a desperate investigator and the desire to be helpful, the statistical consultant will often try to salvage things by proposing an analysis of a subset of the data, suggesting the use of less powerful methods, or suggesting that the investigator use his or her data to test a different hypothesis (i.e., ask a different question). While these steps may serve the short-term goal of getting an abstract or manuscript out on time, they do not encourage efficient clinical and scientific investigation. These frustrating problems could be easily avoided if investigators simply thought about how they were going to analyze their data at the *beginning* rather than the end of the process. Unfortunately, most do not.

When evaluating the strength of an argument for or against some treatment or scientific hypothesis, what should you look for? The investigator should clearly state*

- *The hypothesis being examined (preferably, as the specific null hypothesis to be analyzed statistically).*
- *The data used to test this hypothesis and the procedure used to collect them (including the randomization procedure).*
- *The population the samples represent.*
- *The statistical procedure used to evaluate the data and reach conclusions.*
- *The power of the study to detect a specified effect, particularly if the conclusion is “negative.”*

The closer a paper or oral presentation comes to meeting this standard, the more aware the authors are of the statistical issues in what they are doing and the more confident you can be of their conclusions.

One should immediately be suspicious of a paper that says nothing about the procedures used to obtain “*P* values” or that includes meaningless statements such as “standard statistical procedures were used.”

Likewise—particularly for studies funded by an organization with a strong financial interest in the outcome—such

as a pharmaceutical or tobacco company—take care to ensure that the conclusions stated in the paper are, in fact, consistent with the results in the paper and that they are not being “spun” to support the sponsor’s interests.[†]

Finally, the issues of ethics and scientific validity, especially as they concern human and animal subjects, are inextricably intertwined. Any experimentation that produces results that are misleading or incorrect as a result of avoidable methodologic errors—statistical or otherwise—is unethical. It needlessly puts subjects in jeopardy by not taking every precaution to protect them against unnecessary risk of injury, discomfort, and, in the case of humans, inconvenience. In addition, significant amounts of time and money can be wasted trying to reproduce or refute erroneous results. Alternatively, these results might be accepted without further analysis and adversely affect not only the work of the scientific community, but also the treatment of patients in the future.

Of course, a well-designed, properly analyzed study does not automatically make an investigator’s research innovative, profound, or even worth placing subjects at risk as part of the data collection process. However, even for important questions, it is clearly not ethical to place subjects at risk to collect data in a poorly designed study when this situation can be avoided easily by a little technical knowledge (such as that included in this book) and more thoughtful planning.

How can you help improve the situation?

Do not let people get away with sloppy statistical thinking or biased study design any more than you would permit them to get away with sloppy clinical or scientific thinking. Write letters to the editor. Ask questions in class, rounds, and meetings. When someone answers that they do not know how or where *P* came from, ask them how they can be certain that their results mean what they say. The answer may well be that they cannot.

Most important, if you decide to contribute to the fund of scientific and clinical knowledge, take the time and care to do it right.

*Many journals have moved to formalize the reporting of results from randomized controlled trials. For a widely accepted standard, see Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gøtzsche PC, Lang T, CONSORT Group (Consolidated Standards of Reporting Trials). The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Int Med.* 2001;134:663–694.

[†]In addition to the discussion of spinning results earlier in this chapter, Probs. 5-5 and 5-6 provided examples of spinning. See also Tong EK, Glantz SA. Constructing “sound science” and “good epidemiology”: tobacco, lawyers, and public relations firms. *Am J Public Health.* 2001;91:1749–1757.

Computational Forms

■ TO INTERPOLATE BETWEEN TWO VALUES IN A STATISTICAL TABLE

If the value you need is not in the statistical table, it is possible to estimate the value by *linear interpolation*. For example, suppose you want the critical value of a test statistic, C , corresponding to ν degrees of freedom, and this value of degrees of freedom is not in the table. Find the values of degrees of freedom that are in the table that bracket ν , denoted a and b . Determine the fraction of the way between a and b that ν lies, $f = (\nu - a)/(b - a)$. Therefore, the desired critical value is $C = C_a + f(C_b - C_a)$, where C_a and C_b are the critical values that correspond to a and b degrees of freedom.

A similar approach can be used to interpolate between two P values at a given degrees of freedom. For example, suppose you want to estimate the P value that corresponds to $t = 2.620$ with 20 degrees of freedom. From Table 4-1 with 20 degrees of freedom $t_{.01} = 2.845$ and $t_{.02} = 2.528$, $f = (2.620 - 2.845)/(2.528 - 2.845) = 0.7098$, and $P = .01 + .07098 \times (.02 - .01) = .0171$.

■ VARIANCE

$$s^2 = \frac{\sum X^2 - (\sum X)^2/n}{n-1}$$

■ ONE-WAY ANALYSIS OF VARIANCE

These formulas can be used for equal or unequal sample sizes.

Given Sample Means and Standard Deviations

For treatment group t : n_t = size of sample, \bar{X}_t = mean, s_t = standard deviation. There are a total of k treatment groups.

$$\begin{aligned} N &= \sum n_t \\ SS_{\text{wit}} &= \sum (n_t - 1)s_t^2 \\ \nu_{\text{wit}} &= DF_{\text{wit}} = N - k \\ s_{\text{wit}}^2 &= \frac{SS_{\text{wit}}}{DF_{\text{wit}}} \\ SS_{\text{bet}} &= \sum n_t \bar{X}_t^2 - \frac{(\sum n_t \bar{X}_t)^2}{N} \\ \nu_{\text{bet}} &= DF_{\text{bet}} = k - 1 \\ s_{\text{bet}}^2 &= \frac{SS_{\text{bet}}}{DF_{\text{bet}}} \\ F &= \frac{s_{\text{bet}}^2}{s_{\text{wit}}^2} \end{aligned}$$

Given Raw Data

Subscript t refers to treatment group; subscript s refers to experimental subject.

$$\begin{aligned} C &= (\sum_t \sum_s X_{ts})^2 / N \\ SS_{\text{tot}} &= \sum_t \sum_s X_{ts}^2 - C \\ SS_{\text{bet}} &= \sum_t \frac{(\sum_s X_{ts})^2}{n_t} - C \\ SS_{\text{wit}} &= SS_{\text{tot}} - SS_{\text{bet}} \end{aligned}$$

Degrees of freedom and F are computed as above.

■ UNPAIRED *t* TEST

Given Sample Means and Standard Deviations

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$$

where

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{n_1 + n_2}{n_1 n_2 (n_1 + n_2 - 2)} [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}$$

$$v = n_1 + n_2 - 2$$

Given Raw Data

Use

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{n_1 + n_2}{n_1 n_2 (n_1 + n_2 - 2)} \left[\frac{\sum X_1^2 - \frac{(\sum X_1)^2}{n_1}}{+ \sum X_2^2 - \frac{(\sum X_2)^2}{n_2}} \right]}$$

in the equation for *t* above.

■ 2 × 2 CONTINGENCY TABLES (INCLUDING YATES CORRECTION FOR CONTINUITY)

The contingency table is

A	B
C	D

Chi Square

$$\chi^2 = \frac{N(|AD - BC| - N/2)^2}{(A + B)(C + D)(A + C)(B + D)}$$

where $N = A + B + C + D$.

McNemar's Test

$$\chi^2 = \frac{(|B - C| - 1)^2}{B + C}$$

where *B* and *C* are the numbers of people who responded to only one of the treatments.

Fisher Exact Test

Interchange the rows and columns of the contingency table so that the smallest observed frequency is in position *A*. Compute the probabilities associated with the resulting table, and all more-extreme tables obtained by reducing *A* by 1 and recomputing the table to maintain the row and column totals until *A* = 0. Add all these probabilities to get the first tail of the test. If either the two-row sums or two-column sums are equal, double the resulting probability to obtain the two-tail *P* value. Otherwise, to obtain the second tail of the test, identify the smallest of elements *B* or *C*. Suppose that it is *B*. Reduce *B* by 1 and compute the probability of the associated table. Repeat this process until *B* has been reduced to 0. Identify those tables with probabilities equal to or less than the probability associated with the original observations. Add these probabilities to the first-tail probabilities to obtain the two-tail value of *P*. All the tables computed by varying *B* may not have probabilities below that of the original table; those that do not do not contribute to *P*.

Table A-1 lists values of *n!* for use in computing the Fisher exact test. For larger values of *n*, use a computer or logarithms as $P = \text{antilog} [(\log 9! + \log 14! + \log 11! + \log 12!) - \log 23! - (\log 1! + \log 14! + \log 11! + \log 12!)]$, using tables of log factorials available in handbooks of mathematical tables.

■ LINEAR REGRESSION AND CORRELATION

$$SS_{\text{tot}} = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

$$SS_{\text{reg}} = b \left(\sum XY - \frac{\sum X \sum Y}{n} \right)$$

$$s_{y \cdot x} = \sqrt{\frac{SS_{\text{tot}} - SS_{\text{reg}}}{n - 2}}$$

$$r = \frac{\sqrt{SS_{\text{reg}}}}{\sqrt{SS_{\text{tot}}}} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sqrt{(\sum X^2 - n\bar{X}^2)(\sum Y^2 - n\bar{Y}^2)}}$$

■ TABLE A-1. Values of $n!$ for $n = 1$ to $n = 20$

n	$n!$
0	1
1	1
2	2
3	6
4	24
5	120
6	720
7	5040
8	40320
9	362880
10	3628800
11	39916800
12	479001600
13	6227020800
14	87178291200
15	1307674368000
16	20922789888000
17	355687428096000
18	6402373705728000
19	121645100408832000
20	2432902008176640000

■ REPEATED MEASURES ANALYSIS OF VARIANCE

There are k treatments and n experimental subjects.

$$A = \frac{(\sum_t \sum_s \sum_{ts})^2}{kn}$$

$$B = \sum_t \sum_s X_{ts}^2$$

$$C = \frac{(\sum_t \sum_s X_{ts})^2}{n}$$

$$D = \frac{\sum_s (\sum_t X_{ts})^2}{k}$$

$$SS_{\text{treat}} = C - A$$

$$SS_{\text{res}} = A + B - C - D$$

$$DF_{\text{treat}} = k - 1$$

$$DF_{\text{res}} = (n - 1)(k - 1)$$

$$F = \frac{SS_{\text{treat}}/DF_{\text{treat}}}{SS_{\text{res}}/DF_{\text{res}}}$$

■ KRUSKAL-WALLIS TEST

$$H = \frac{12}{N(N+1)} \sum \left(\frac{R_t^2}{n_t} \right) - 3(N+1)$$

where $N = \sum n_t$.

■ FRIEDMAN TEST

$$x_r^2 = \frac{12}{nk(k+1)} \sum R_t^2 - 3n(k+1)$$

where there are k treatments and n experimental subjects and R_t is the sum of ranks for treatment t .

This page intentionally left blank

Statistical Tables and Power Charts

Statistical Tables

Table 3-1. Critical Values of F Corresponding to $P < .05$ and $P < .01$

Table 4-1. Critical Values of t (Two-Tailed)

Table 4-4. Holm-Sidak Critical P Values for Individual Comparisons to Maintain a 5% Family Error Rate ($\alpha_T = .05$)

Table 5-5. Critical Values for the χ^2 Distribution

Table 6-2. Critical Values of t (One-Tailed)

Table 8-7. Critical Values for Spearman Rank Correlation Coefficient

Table 10-3. Critical Values (Two-Tailed) of the Mann-Whitney Rank-Sum T

Table 10-7. Critical Values (Two-Tailed) of Wilcoxon W

Table 10-14. Critical Values for Friedman χ_r^2

Power Charts for Analysis of Variance

This page intentionally left blank

■ TABLE 3-1. Critical Values of *F* Corresponding to *P* < .05 (Lightface) and *P* < .01 (Boldface)

<i>v_d</i>	<i>v_n</i>																							
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	249	250	251	252	253	253	254	254	254
	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6082	6106	6142	6169	6208	6234	6261	6286	6302	6323	6334	6352	6361	6366
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44	19.45	19.46	19.47	19.47	19.48	19.48	19.49	19.49	19.50
	98.49	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.41	99.42	99.43	99.44	99.45	99.46	99.47	99.48	99.48	99.49	99.49	99.49	99.50	99.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.56	8.54	8.54	8.53
	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05	26.92	26.83	26.69	26.60	26.50	26.41	26.35	26.27	26.23	26.18	26.14	26.12
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.66	5.65	5.64	5.63
	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.54	14.45	14.37	14.24	14.15	14.02	13.93	13.83	13.74	13.69	13.61	13.57	13.52	13.48	13.46
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36
	16.26	13.27	12.06	11.39	10.97	10.67	10.45	10.29	10.15	10.05	9.96	9.89	9.77	9.68	9.55	9.47	9.38	9.29	9.24	9.17	9.13	9.07	9.04	9.02
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67
	13.74	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.60	7.52	7.39	7.31	7.23	7.14	7.09	7.02	6.99	6.94	6.90	6.88
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24	3.23
	12.25	9.55	8.45	7.85	7.46	7.19	7.00	6.84	6.71	6.62	6.54	6.47	6.35	6.27	6.15	6.07	5.98	5.90	5.85	5.78	5.75	5.70	5.67	5.65
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.93
	11.26	8.65	7.59	7.01	6.63	6.37	6.19	6.03	5.91	5.82	5.74	5.67	5.56	5.48	5.36	5.28	5.20	5.11	5.06	5.00	4.96	4.91	4.88	4.86
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72	2.71
	10.56	8.02	6.99	6.42	6.06	5.80	5.62	5.47	5.35	5.26	5.18	5.11	5.00	4.92	4.80	4.73	4.64	4.56	4.51	4.45	4.41	4.36	4.33	4.31
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55	2.54
	10.04	7.56	6.55	5.99	5.64	5.39	5.21	5.06	4.95	4.85	4.78	4.71	4.60	4.52	4.41	4.33	4.25	4.17	4.12	4.05	4.01	3.96	3.93	3.91
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.45	2.42	2.41	2.40
	9.65	7.20	6.22	5.67	5.32	5.07	4.88	4.74	4.63	4.54	4.46	4.40	4.29	4.21	4.10	4.02	3.94	3.86	3.80	3.74	3.70	3.66	3.62	3.60
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.35	2.32	2.31	2.30
	9.33	6.93	5.95	5.41	5.06	4.82	4.65	4.50	4.39	4.30	4.22	4.16	4.05	3.98	3.86	3.78	3.70	3.61	3.56	3.49	3.46	3.41	3.38	3.36
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.55	2.51	2.46	2.42	2.38	2.34	2.32	2.28	2.26	2.24	2.22	2.21
	9.07	6.70	5.74	5.20	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.85	3.78	3.67	3.59	3.51	3.42	3.37	3.30	3.27	3.21	3.18	3.16
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13
	8.86	6.51	5.56	5.03	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.70	3.62	3.51	3.43	3.34	3.26	3.21	3.14	3.11	3.06	3.02	3.00
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07
	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.56	3.48	3.36	3.29	3.20	3.12	3.07	3.00	2.97	2.92	2.89	2.87
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01
	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.61	3.55	3.45	3.37	3.25	3.18	3.10	3.01	2.96	2.98	2.86	2.80	2.77	2.75
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96
	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.45	3.35	3.27	3.16	3.08	3.00	2.92	2.86	2.79	2.76	2.70	2.67	2.65
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19	2.15	2.11	2.07	2.04	2.00	1.98	1.95	1.93	1.92
	8.28	6.01	5.09	4.58	4.25	4.01	3.85	3.71	3.60	3.51	3.44	3.37	3.27	3.19	3.07	3.00	2.91	2.83	2.78	2.71	2.68	2.62	2.59	2.57
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31	2.26	2.21	2.15	2.11	2.07	2.02	2.00	1.96	1.94	1.91	1.90	1.88
	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30	3.19	3.12	3.00	2.92	2.84	2.76	2.70	2.63	2.60	2.54	2.51	2.49

v_n = degrees of freedom for numerator; *v_d* = degrees of freedom for denominator.

(continued)

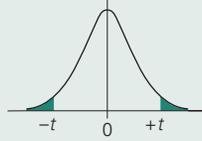
■ TABLE 3-1. Critical Values of F Corresponding to $P < .05$ (Lightface) and $P < .01$ (Boldface) (Continued)

v_d	v_n																							
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.90	1.87	1.85	1.84
	8.10	5.85	4.94	4.43	4.10	3.87	3.71	3.56	3.45	3.37	3.30	3.23	3.13	3.05	2.94	2.86	2.77	2.69	2.63	2.56	2.53	2.47	2.44	2.42
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20	2.15	2.09	2.05	2.00	1.96	1.93	1.89	1.87	1.84	1.82	1.81
	8.02	5.78	4.87	4.37	4.04	3.81	3.65	3.51	3.40	3.31	3.24	3.17	3.07	2.99	2.88	2.80	2.72	2.63	2.58	2.51	2.47	2.42	2.38	2.36
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23	2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.80	1.78
	7.94	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12	3.02	2.94	2.83	2.75	2.67	2.58	2.53	2.46	2.42	2.37	2.33	2.31
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20	2.14	2.10	2.04	2.00	1.96	1.91	1.88	1.84	1.82	1.79	1.77	1.76
	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.14	3.07	2.97	2.89	2.78	2.70	2.62	2.53	2.48	2.41	2.37	2.32	2.28	2.26
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13	2.09	2.02	1.98	1.94	1.89	1.86	1.82	1.80	1.76	1.74	1.73
	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.25	3.17	3.09	3.03	2.93	2.85	2.74	2.66	2.58	2.49	2.44	2.36	2.33	2.27	2.23	2.31
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16	2.11	2.06	2.00	1.96	1.92	1.87	1.84	1.80	1.77	1.74	1.72	1.71
	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.21	3.13	3.05	2.99	2.89	2.81	2.70	2.62	2.54	2.45	2.40	2.32	2.29	2.23	2.19	2.17
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.10	2.05	1.99	1.95	1.90	1.85	1.82	1.78	1.76	1.72	1.70	1.69
	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.17	3.09	3.02	2.96	2.86	2.77	2.66	2.58	2.50	2.41	2.36	2.28	2.25	2.19	2.15	2.13
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13	2.08	2.03	1.97	1.93	1.88	1.84	1.80	1.76	1.74	1.71	1.68	1.67
	7.68	5.49	4.60	4.11	3.79	3.56	3.39	3.26	3.14	3.06	2.98	2.93	2.83	2.74	2.63	2.55	2.47	2.38	2.33	2.25	2.21	2.16	2.12	2.10
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12	2.06	2.02	1.96	1.91	1.87	1.81	1.78	1.75	1.72	1.69	1.67	1.65
	7.64	5.45	4.57	4.07	3.76	3.53	3.36	3.23	3.11	3.03	2.95	2.90	2.80	2.71	2.60	2.52	2.44	2.35	2.30	2.22	2.18	2.13	2.09	2.06
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05	2.00	1.94	1.90	1.85	1.80	1.77	1.73	1.71	1.68	1.65	1.64
	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.08	3.00	2.92	2.87	2.77	2.68	2.57	2.49	2.41	2.32	2.27	2.19	2.15	2.10	2.06	2.03
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09	2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.64	1.62
	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.06	2.98	2.90	2.84	2.74	2.66	2.55	2.47	2.38	2.29	2.24	2.16	2.13	2.07	2.03	2.01
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07	2.02	1.97	1.91	1.86	1.82	1.76	1.74	1.69	1.67	1.64	1.61	1.59
	7.50	5.34	4.46	3.97	3.66	3.42	3.25	3.12	3.01	2.94	2.86	2.80	2.70	2.62	2.51	2.42	2.34	2.25	2.20	2.12	2.08	2.02	1.98	1.96
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05	2.00	1.95	1.89	1.84	1.80	1.74	1.71	1.67	1.64	1.61	1.59	1.57
	7.44	5.29	4.42	3.93	3.61	3.38	3.21	3.08	2.97	2.89	2.82	2.76	2.66	2.58	2.47	2.38	2.30	2.21	2.15	2.08	2.04	1.98	1.94	1.91
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03	1.98	1.93	1.87	1.82	1.78	1.72	1.69	1.65	1.62	1.59	1.56	1.55
	7.39	5.25	4.38	3.89	3.58	3.35	3.18	3.04	2.94	2.86	2.78	2.72	2.62	2.54	2.43	2.35	2.26	2.17	2.12	2.04	2.00	1.94	1.90	1.87
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.96	1.92	1.85	1.80	1.76	1.71	1.67	1.63	1.60	1.57	1.54	1.53
	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.91	2.82	2.75	2.69	2.59	2.51	2.40	2.32	2.22	2.14	2.08	2.00	1.97	1.90	1.86	1.84
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00	1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51
	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.88	2.80	2.73	2.66	2.56	2.49	2.37	2.29	2.20	2.11	2.05	1.97	1.94	1.88	1.84	1.81
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99	1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.60	1.57	1.54	1.51	1.49
	7.27	5.15	4.29	3.80	3.49	3.26	3.10	2.96	2.86	2.77	2.70	2.64	2.54	2.46	2.35	2.26	2.17	2.08	2.02	1.94	1.91	1.85	1.80	1.78
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.50	1.48
	7.24	5.12	4.26	3.78	3.46	3.24	3.07	2.94	2.84	2.75	2.68	2.62	2.52	2.44	2.32	2.24	2.15	2.06	2.00	1.92	1.88	1.82	1.78	1.75
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97	1.91	1.87	1.80	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46
	7.21	5.10	4.24	3.76	3.44	3.22	3.05	2.92	2.82	2.73	2.66	2.60	2.50	2.42	2.30	2.22	2.13	2.04	1.98	1.90	1.86	1.80	1.76	1.72
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96	1.90	1.86	1.79	1.74	1.70	1.64	1.61	1.56	1.53	1.50	1.47	1.45
	7.19	5.08	4.22	3.74	3.42	3.20	3.04	2.90	2.80	2.71	2.64	2.58	2.48	2.40	2.28	2.20	2.11	2.02	1.96	1.88	1.84	1.78	1.73	1.70

50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95	1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44
	7.17	5.06	4.20	3.72	3.41	3.18	3.02	2.88	2.78	2.70	2.62	2.56	2.46	2.39	2.26	2.18	2.10	2.00	1.94	1.86	1.82	1.76	1.71	1.68
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.86	1.81	1.75	1.70	1.65	1.59	1.56	1.50	1.48	1.44	1.41	1.39
	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50	2.40	2.32	2.20	2.12	2.03	1.93	1.87	1.79	1.74	1.68	1.63	1.60
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89	1.84	1.79	1.72	1.67	1.62	1.56	1.53	1.47	1.45	1.40	1.37	1.35
	7.01	4.92	4.08	3.60	3.29	3.07	2.91	2.77	2.67	2.59	2.51	2.45	2.35	2.28	2.15	2.07	1.98	1.88	1.82	1.74	1.69	1.62	1.56	1.53
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88	1.82	1.77	1.70	1.65	1.60	1.54	1.51	1.45	1.42	1.38	1.35	1.32
	6.96	4.88	4.04	3.56	3.25	3.04	2.87	2.74	2.64	2.55	2.48	2.41	2.32	2.24	2.11	2.03	1.94	1.84	1.78	1.70	1.65	1.57	1.52	1.49
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85	1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.30	1.28
	6.90	4.82	3.98	3.51	3.20	2.99	2.82	2.69	2.59	2.51	2.43	2.36	2.26	2.19	2.06	1.98	1.89	1.79	1.73	1.64	1.59	1.51	1.46	1.43
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.87	1.84	1.78	1.73	1.66	1.61	1.56	1.50	1.46	1.39	1.37	1.32	1.28	1.25
	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.40	2.34	2.23	2.15	2.03	1.95	1.86	1.76	1.70	1.61	1.56	1.48	1.42	1.38
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75	1.69	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.00
	6.63	4.60	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.24	2.18	2.07	1.99	1.87	1.79	1.69	1.59	1.52	1.41	1.36	1.25	1.15	1.00

v_n = degrees of freedom for numerator; v_d = degrees of freedom for denominator.

Reproduced from Snedecor GW, Cochran WG. *Statistical Methods*, 8th ed. Copyright © 1989. Reproduced with the permission of John Wiley & Sons, Inc.

TABLE 4-1. Critical Values of t (Two-Tailed)


ν	Probability of Greater Value (P)								
	0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
1	1.000	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	1.415	1.895	2.365	2.998	3.449	4.029	4.785	5.408
8	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
31	0.682	1.309	1.696	2.040	2.453	2.744	3.022	3.375	3.633
32	0.682	1.309	1.694	2.037	2.449	2.738	3.015	3.365	3.622
33	0.682	1.308	1.692	2.035	2.445	2.733	3.008	3.356	3.611
34	0.682	1.307	1.691	2.032	2.441	2.728	3.002	3.348	3.601
35	0.682	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591
36	0.681	1.306	1.688	2.028	2.434	2.719	2.990	3.333	3.582
37	0.681	1.305	1.687	2.026	2.431	2.715	2.985	3.326	3.574
38	0.681	1.304	1.686	2.024	2.429	2.712	2.980	3.319	3.566

■ TABLE 4-1. Critical Values of *t* (Two-Tailed) (Continued)

<i>v</i>	Probability of Greater Value (<i>P</i>)								
	0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
39	0.681	1.304	1.685	2.023	2.426	2.708	2.976	3.313	3.558
40	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
42	0.680	1.302	1.682	2.018	2.418	2.698	2.963	3.296	3.538
44	0.680	1.301	1.680	2.015	2.414	2.692	2.956	3.286	3.526
46	0.680	1.300	1.679	2.013	2.410	2.687	2.949	3.277	3.515
48	0.680	1.299	1.677	2.011	2.407	2.682	2.943	3.269	3.505
50	0.679	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
52	0.679	1.298	1.675	2.007	2.400	2.674	2.932	3.255	3.488
54	0.679	1.297	1.674	2.005	2.397	2.670	2.927	3.248	3.480
56	0.679	1.297	1.673	2.003	2.395	2.667	2.923	3.242	3.473
58	0.679	1.296	1.672	2.002	2.392	2.663	2.918	3.237	3.466
60	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
62	0.678	1.295	1.670	1.999	2.388	2.657	2.911	3.227	3.454
64	0.678	1.295	1.669	1.998	2.386	2.655	2.908	3.223	3.449
66	0.678	1.295	1.668	1.997	2.384	2.652	2.904	3.218	3.444
68	0.678	1.294	1.668	1.995	2.382	2.650	2.902	3.214	3.439
70	0.678	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435
72	0.678	1.293	1.666	1.993	2.379	2.646	2.896	3.207	3.431
74	0.678	1.293	1.666	1.993	2.378	2.644	2.894	3.204	3.427
76	0.678	1.293	1.665	1.992	2.376	2.642	2.891	3.201	3.423
78	0.678	1.292	1.665	1.991	2.375	2.640	2.889	3.198	3.420
80	0.678	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
90	0.677	1.291	1.662	1.987	2.368	2.632	2.878	3.183	3.402
100	0.677	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
120	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
140	0.676	1.288	1.656	1.977	2.353	2.611	2.852	3.149	3.361
160	0.676	1.287	1.654	1.975	2.350	2.607	2.846	3.142	3.352
180	0.676	1.286	1.653	1.973	2.347	2.603	2.842	3.136	3.345
200	0.676	1.286	1.653	1.972	2.345	2.601	2.839	3.131	3.340
∞	0.6745	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.0902	3.2905
Normal	0.6745	1.2816	1.6449	1.9600	2.3263	2.5758	2.8070	3.0902	3.2905

Adapted from Zar JH. *Biostatistical Analysis*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall; 1984, 484–485:table B.3, by permission of Pearson Education, Inc., Upper Saddle River, NJ.

■ TABLE 4-4. Holm-Sidak Critical P Values for Individual Comparisons to Maintain a 5% Family Error Rate ($\alpha_T = .05$)

Comparison Number (j)	Total Number of Comparisons (k)														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.0500	.0253	.0170	.0127	.0102	.0085	.0073	.0064	.0057	.0051	.0047	.0043	.0039	.0037	.0034
2		.0500	.0253	.0170	.0127	.0102	.0085	.0073	.0064	.0057	.0051	.0047	.0043	.0039	.0037
3			.0500	.0253	.0170	.0127	.0102	.0085	.0073	.0064	.0057	.0051	.0047	.0043	.0039
4				.0500	.0253	.0170	.0127	.0102	.0085	.0073	.0064	.0057	.0051	.0047	.0043
5					.0500	.0253	.0170	.0127	.0102	.0085	.0073	.0064	.0057	.0051	.0047
6						.0500	.0253	.0170	.0127	.0102	.0085	.0073	.0064	.0057	.0051
7							.0500	.0253	.0170	.0127	.0102	.0085	.0073	.0064	.0057
8								.0500	.0253	.0170	.0127	.0102	.0085	.0073	.0064
9									.0500	.0253	.0170	.0127	.0102	.0085	.0073
10										.0500	.0253	.0170	.0127	.0102	.0085
11											.0500	.0253	.0170	.0127	.0102
12												.0500	.0253	.0170	.0127
13													.0500	.0253	.0170
14														.0500	.0253
15															.0500

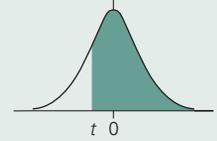
$P_{\text{crit}} = 1 - (1 - \alpha_T)^{1/(k-j+1)}$.

■ TABLE 5-5. Critical Values for the χ^2 Distribution

v	Probability of Greater Value P							
	.50	.25	.10	.05	.025	.01	.005	.001
1	.455	1.323	2.706	3.841	5.024	6.635	7.879	10.828
2	1.386	2.773	4.605	5.991	7.378	9.210	10.597	13.816
3	2.366	4.108	6.251	7.815	9.348	11.345	12.838	16.266
4	3.357	5.385	7.779	9.488	11.143	13.277	14.860	18.467
5	4.351	6.626	9.236	11.070	12.833	15.086	16.750	20.515
6	5.348	7.841	10.645	12.592	14.449	16.812	18.548	22.458
7	6.346	9.037	12.017	14.067	16.013	18.475	20.278	24.322
8	7.344	10.219	13.362	15.507	17.535	20.090	21.955	26.124
9	8.343	11.389	14.684	16.919	19.023	21.666	23.589	27.877
10	9.342	12.549	15.987	18.307	20.483	23.209	25.188	29.588
11	10.341	13.701	17.275	19.675	21.920	24.725	26.757	31.264
12	11.340	14.845	18.549	21.026	23.337	26.217	28.300	32.909
13	12.340	15.984	19.812	22.362	24.736	27.688	29.819	34.528
14	13.339	17.117	21.064	23.685	26.119	29.141	31.319	36.123
15	14.339	18.245	22.307	24.996	27.488	30.578	32.801	37.697
16	15.338	19.369	23.542	26.296	28.845	32.000	34.267	39.252
17	16.338	20.489	24.769	27.587	30.191	33.409	35.718	40.790
18	17.338	21.605	25.989	28.869	31.526	34.805	37.156	42.312
19	18.338	22.718	27.204	30.144	32.852	36.191	38.582	43.820
20	19.337	23.828	28.412	31.410	34.170	37.566	39.997	45.315
21	20.337	24.935	29.615	32.671	35.479	38.932	41.401	46.797
22	21.337	26.039	30.813	33.924	36.781	40.289	42.796	48.268
23	22.337	27.141	32.007	35.172	38.076	41.638	44.181	49.728
24	23.337	28.241	33.196	36.415	39.364	42.980	45.559	51.179
25	24.337	29.339	34.382	37.652	40.646	44.314	46.928	52.620
26	25.336	30.435	35.563	38.885	41.923	45.642	48.290	54.052
27	26.336	31.528	36.741	40.113	43.195	46.963	49.645	55.476
28	27.336	32.020	37.916	41.337	44.461	48.278	50.993	56.892
29	28.336	33.711	39.087	42.557	45.722	49.588	52.336	58.301
30	29.336	34.800	40.256	43.773	46.979	50.892	53.672	59.703
31	30.336	35.887	41.422	44.985	48.232	52.191	55.003	61.098
32	31.336	36.973	42.585	46.194	49.480	53.486	56.328	62.487
33	32.336	38.058	43.745	47.400	50.725	54.776	57.648	63.870
34	33.336	39.141	44.903	48.602	51.966	56.061	58.964	65.247
35	34.336	40.223	46.059	49.802	53.203	57.342	60.275	66.619
36	35.336	41.304	47.212	50.998	54.437	58.619	61.581	67.985
37	36.336	42.383	48.363	52.192	55.668	59.893	62.883	69.346
38	37.335	43.462	49.513	53.384	56.896	61.162	64.181	70.703
39	38.335	44.539	50.660	54.572	58.120	62.428	65.476	72.055
40	39.335	45.616	51.805	55.758	59.342	63.691	66.766	73.402
41	40.335	46.692	52.949	56.942	60.561	64.950	68.053	74.745
42	41.335	47.766	54.090	58.124	61.777	66.206	69.336	76.084
43	42.335	48.840	55.230	59.304	62.990	67.459	70.616	77.419
44	43.335	49.913	56.369	60.481	64.201	68.710	71.893	78.750
45	44.335	50.985	57.505	61.656	65.410	69.957	73.166	80.077
46	45.335	52.056	58.641	62.830	66.617	71.201	74.437	81.400
47	46.335	53.127	59.774	64.001	67.821	72.443	75.704	82.720
48	47.335	54.196	60.907	65.171	69.023	73.683	76.969	84.037
49	48.335	55.265	62.038	66.339	70.222	74.919	78.231	85.351
50	49.335	56.334	63.167	67.505	71.420	76.154	79.490	86.661

Adapted from Zar JH. *Biostatistical Analysis*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall; 1984, 479–482:table B.1, by permission of Pearson Education, Inc., Upper Saddle River, NJ.

■ TABLE 6-2. Critical Values of t (One-Tailed)



Probability of Larger Value (Upper Tail)										
	.995	.99	.98	.975	.95	.90	.85	.80	.70	.60
Probability of Smaller Value (Lower Tail)										
ν	.005	.01	.02	.025	.05	.10	.15	.20	.30	.40
2	-9.925	-6.965	-4.849	-4.303	-2.920	-1.886	-1.386	-1.061	-0.617	-0.289
4	-4.604	-3.747	-2.999	-2.776	-2.132	-1.533	-1.190	-0.941	-0.569	-0.271
6	-3.707	-3.143	-2.612	-2.447	-1.943	-1.440	-1.134	-0.906	-0.553	-0.265
8	-3.355	-2.896	-2.449	-2.306	-1.860	-1.397	-1.108	-0.889	-0.546	-0.262
10	-3.169	-2.764	-2.359	-2.228	-1.812	-1.372	-1.093	-0.879	-0.542	-0.260
12	-3.055	-2.681	-2.303	-2.179	-1.782	-1.356	-1.083	-0.873	-0.539	-0.259
14	-2.977	-2.624	-2.264	-2.145	-1.761	-1.345	-1.076	-0.868	-0.537	-0.258
16	-2.921	-2.583	-2.235	-2.120	-1.746	-1.337	-1.071	-0.865	-0.535	-0.258
18	-2.878	-2.552	-2.214	-2.101	-1.734	-1.330	-1.067	-0.862	-0.534	-0.257
20	-2.845	-2.528	-2.197	-2.086	-1.725	-1.325	-1.064	-0.860	-0.533	-0.257
25	-2.787	-2.485	-2.167	-2.060	-1.708	-1.316	-1.058	-0.856	-0.531	-0.256
30	-2.750	-2.457	-2.147	-2.042	-1.697	-1.310	-1.055	-0.854	-0.530	-0.256
35	-2.724	-2.438	-2.133	-2.030	-1.690	-1.306	-1.052	-0.852	-0.529	-0.255
40	-2.704	-2.423	-2.123	-2.021	-1.684	-1.303	-1.050	-0.851	-0.529	-0.255
60	-2.660	-2.390	-2.099	-2.000	-1.671	-1.296	-1.045	-0.848	-0.527	-0.254
120	-2.617	-2.358	-2.076	-1.980	-1.658	-1.289	-1.041	-0.845	-0.526	-0.254
∞	-2.576	-2.326	-2.054	-1.960	-1.645	-1.282	-1.036	-0.842	-0.524	-0.253
Normal	-2.576	-2.326	-2.054	-1.960	-1.645	-1.282	-1.036	-0.842	-0.524	-0.253

■ TABLE 6-2. Critical Values of *t* (One-Tailed) (Continued)

Probability of Larger Value (Upper Tail)										
.50	.40	.30	.20	.15	.10	.05	.025	.02	.01	.005
Probability of Smaller Value (Lower Tail)										
.50	.60	.70	.80	.85	.90	.95	.975	.98	.99	.995
0	0.289	0.617	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925
0	0.271	0.569	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604
0	0.265	0.553	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707
0	0.262	0.546	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355
0	0.260	0.542	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169
0	0.259	0.539	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055
0	0.258	0.537	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977
0	0.258	0.535	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921
0	0.257	0.534	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878
0	0.257	0.533	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845
0	0.256	0.531	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787
0	0.256	0.530	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750
0	0.255	0.529	0.852	1.052	1.306	1.690	2.030	2.133	2.438	2.724
0	0.255	0.529	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704
0	0.254	0.527	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660
0	0.254	0.526	0.845	1.041	1.289	1.658	1.980	2.076	2.358	2.617
0	0.253	0.524	0.842	1.036	1.282	1.645	1.960	2.054	2.326	2.576
0	0.253	0.524	0.842	1.036	1.282	1.645	1.960	2.054	2.326	2.576

■ TABLE 8-7. Critical Values for Spearman Rank Correlation Coefficient*

n	Probability of Greater Value (P)								
	.50	.20	.10	.05	.02	.01	.005	.002	.001
4	.600	1.000	1.000						
5	.500	.800	.900	1.000	1.000				
6	.371	.657	.829	.886	.943	1.000	1.000		
7	.321	.571	.714	.786	.893	.929	.964	1.000	1.000
8	.310	.524	.643	.738	.833	.881	.905	.952	.976
9	.267	.483	.600	.700	.783	.833	.867	.917	.933
10	.248	.455	.564	.648	.745	.794	.830	.879	.903
11	.236	.427	.536	.618	.709	.755	.800	.845	.873
12	.217	.406	.503	.587	.678	.727	.769	.818	.846
13	.209	.385	.484	.560	.648	.703	.747	.791	.824
14	.200	.367	.464	.538	.626	.679	.723	.771	.802
15	.189	.354	.446	.521	.604	.654	.700	.750	.779
16	.182	.341	.429	.503	.582	.635	.679	.729	.762
17	.176	.328	.414	.485	.566	.615	.662	.713	.748
18	.170	.317	.401	.472	.550	.600	.643	.695	.728
19	.165	.309	.391	.460	.535	.584	.628	.677	.712
20	.161	.299	.380	.447	.520	.570	.612	.662	.696
21	.156	.292	.370	.435	.508	.556	.599	.648	.681
22	.152	.284	.361	.425	.496	.544	.586	.634	.667
23	.148	.278	.353	.415	.486	.532	.573	.622	.654
24	.144	.271	.344	.406	.476	.521	.562	.610	.642
25	.142	.265	.337	.398	.466	.511	.551	.598	.630
26	.138	.259	.331	.390	.457	.501	.541	.587	.619
27	.136	.255	.324	.382	.448	.491	.531	.577	.608
28	.133	.250	.317	.375	.440	.483	.522	.567	.598
29	.130	.245	.312	.368	.433	.475	.513	.558	.589
30	.128	.240	.306	.362	.425	.467	.504	.549	.580
31	.126	.236	.301	.356	.418	.459	.496	.541	.571
32	.124	.232	.296	.350	.412	.452	.489	.533	.563
33	.121	.229	.291	.345	.405	.446	.482	.525	.554
34	.120	.225	.287	.340	.399	.439	.475	.517	.547
35	.118	.222	.283	.335	.394	.433	.468	.510	.539
36	.116	.219	.279	.330	.388	.427	.462	.504	.533
37	.114	.216	.275	.325	.383	.421	.456	.497	.526
38	.113	.212	.271	.321	.378	.415	.450	.491	.519
39	.111	.210	.267	.317	.373	.410	.444	.485	.513
40	.110	.207	.264	.313	.368	.405	.439	.479	.507
41	.108	.204	.261	.309	.364	.400	.433	.473	.501
42	.107	.202	.257	.305	.359	.395	.428	.468	.495
43	.105	.199	.254	.301	.355	.391	.423	.463	.490
44	.104	.197	.251	.298	.351	.386	.419	.458	.484
45	.103	.194	.248	.294	.347	.382	.414	.453	.479
46	.102	.192	.246	.291	.343	.378	.410	.448	.474
47	.101	.190	.243	.288	.340	.374	.405	.443	.469
48	.100	.188	.240	.285	.336	.370	.401	.439	.465
49	.098	.186	.238	.282	.333	.366	.397	.434	.460
50	.097	.184	.235	.279	.329	.363	.393	.430	.456

*For sample sizes greater than 50, use

$$t = \frac{r_s}{\sqrt{(1-r_s^2)/(n-2)}}$$

With $\nu = n - 2$ degrees of freedom to obtain the approximate P value.

Adapted from Zar JH. *Biostatistical Analysis*, 4th ed. Englewood Cliffs, NJ: Prentice-Hall; 1999, Appendix 116–117. Used by permission.

■ TABLE 10-3. Critical Values (Two-Tailed) of the Mann-Whitney Rank-Sum *T*

<i>n_S</i>	<i>n_B</i>	Probability Levels Near			
		.05		.01	
		Critical Values	<i>P</i>	Critical Values	<i>P</i>
3	4	6,18	.057		
	5	6,21	.036		
	5	7,20	.071		
	6	7,23	.048	6,24	.024
	7	7,26	.033	6,27	.017
	7	8,25	.067		
	8	8,28	.042	6,30	.012
	8	9,27	.077		
4	4	11,25	.057	10,26	.026
	5	11,29	.032	10,30	.016
	5	12,28	.063		
	6	12,32	.038	10,34	.010
	7	13,35	.042	10,38	.012
	8	14,38	.048	11,41	.008
	8	15,37	.078	12,40	.016
	8	16,36	.109		
5	5	17,38	.032	15,40	.008
	5	18,37	.056	16,39	.016
	6	19,41	.052	16,44	.010
	7	20,45	.048	17,48	.010
	8	21,49	.045	18,52	.011
6	6	26,52	.041	23,55	.009
	6	27,51	.065	24,54	.015
	7	28,56	.051	24,60	.008
	7	29,55	.075	25,59	.014
	8	29,61	.043	25,65	.008
	8	30,60	.059	26,64	.013
	8	31,59	.083		
7	7	37,68	.053	33,72	.011
	8	39,73	.054	34,78	.009
8	8	49,87	.050	44,92	.010

Computed from Table A-9 of Mosteller F, Rourke R. *Sturdy Statistics: Nonparametrics and Order Statistics*. Reading, MA: Addison-Wesley; 1973.

■ **TABLE 10-7. Critical Values (Two-Tailed) of Wilcoxon W**

<i>n</i>	Critical Value	<i>P</i>
5	15	.062
6	21	.032
	19	.062
7	28	.016
	24	.046
8	32	.024
	28	.054

Data from Table A-11 of Mosteller F, Rourke R. *Sturdy Statistics: Nonparametrics and Order Statistics*. Reading, MA: Addison-Wesley; 1973.

■ **TABLE 10-14. Critical Values for Friedman χ_r^2**

<i>k</i> = 3 Treatments			<i>k</i> = 4 Treatments		
<i>n</i>	χ_r^2	<i>P</i>	<i>n</i>	χ_r^2	<i>P</i>
3	6.00	.028	2	6.00	.042
4	6.50	.042	3	7.00	.054
	8.00	.005		8.20	.017
5	5.20	.093	4	7.50	.054
	6.40	.039		9.30	.011
	8.40	.008	5	7.80	.049
6	5.33	.072		9.96	.009
	6.33	.052	6	7.60	.043
	9.00	.008		10.20	.010
7	6.00	.051	7	7.63	.051
	8.86	.008		10.37	.009
8	6.25	.047	8	7.65	.049
	9.00	.010		10.35	.010
9	6.22	.048			
	8.67	.010			
10	6.20	.046			
	8.60	.012			
11	6.54	.043			
	8.91	.011			
12	6.17	.050			
	8.67	.011			
13	6.00	.050			
	8.67	.012			
14	6.14	.049			
	9.00	.010			
15	6.40	.047			
	8.93	.010			

Data from Owen DB. *Handbook of Statistical Tables*. US Department of Energy. Reading, MA: Addison-Wesley; 1962.

■ POWER CHARTS FOR ANALYSIS OF VARIANCE*

*These charts are adapted from Pearson ES, Hartley HO. Charts for the power function for analysis of variance tests, derived from the non-central F distribution. *Biometrika*. 1951;38:112–130.

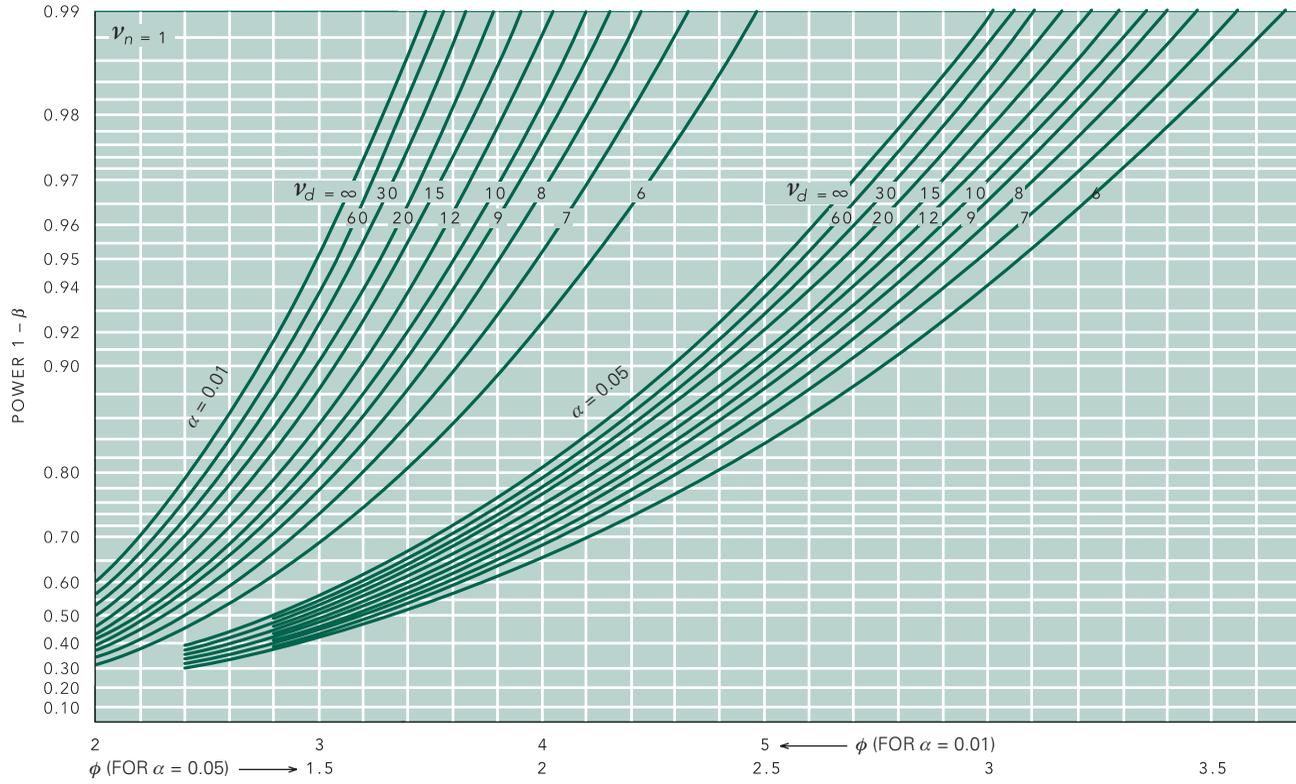


FIGURE B-1.

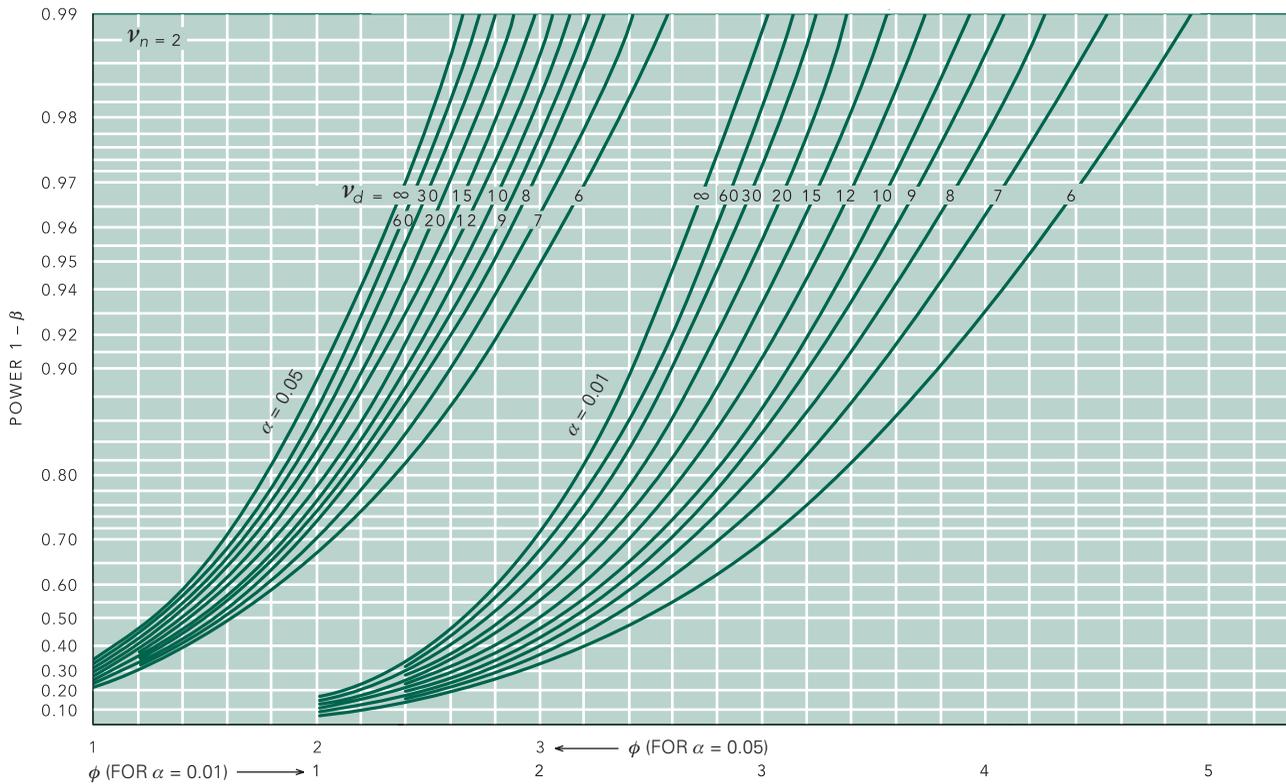


FIGURE B-2.

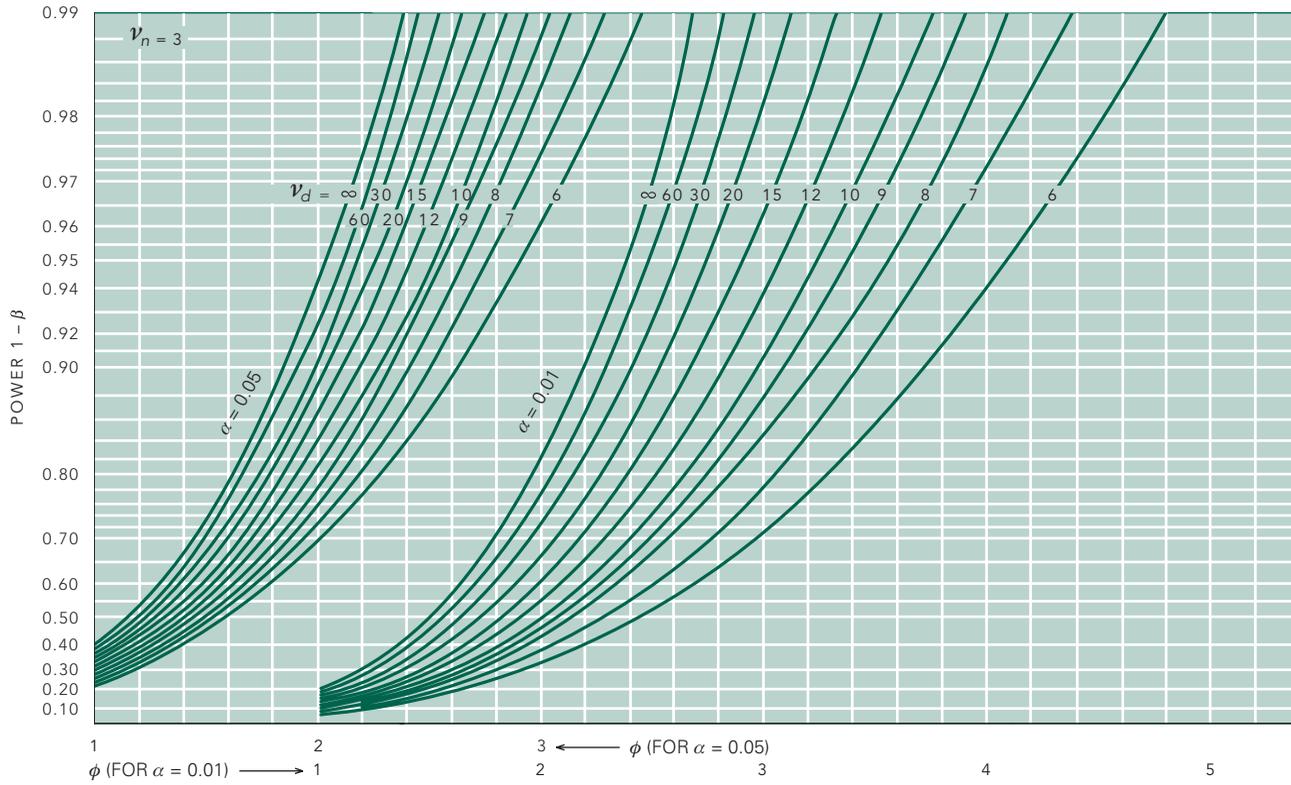


FIGURE B-3.

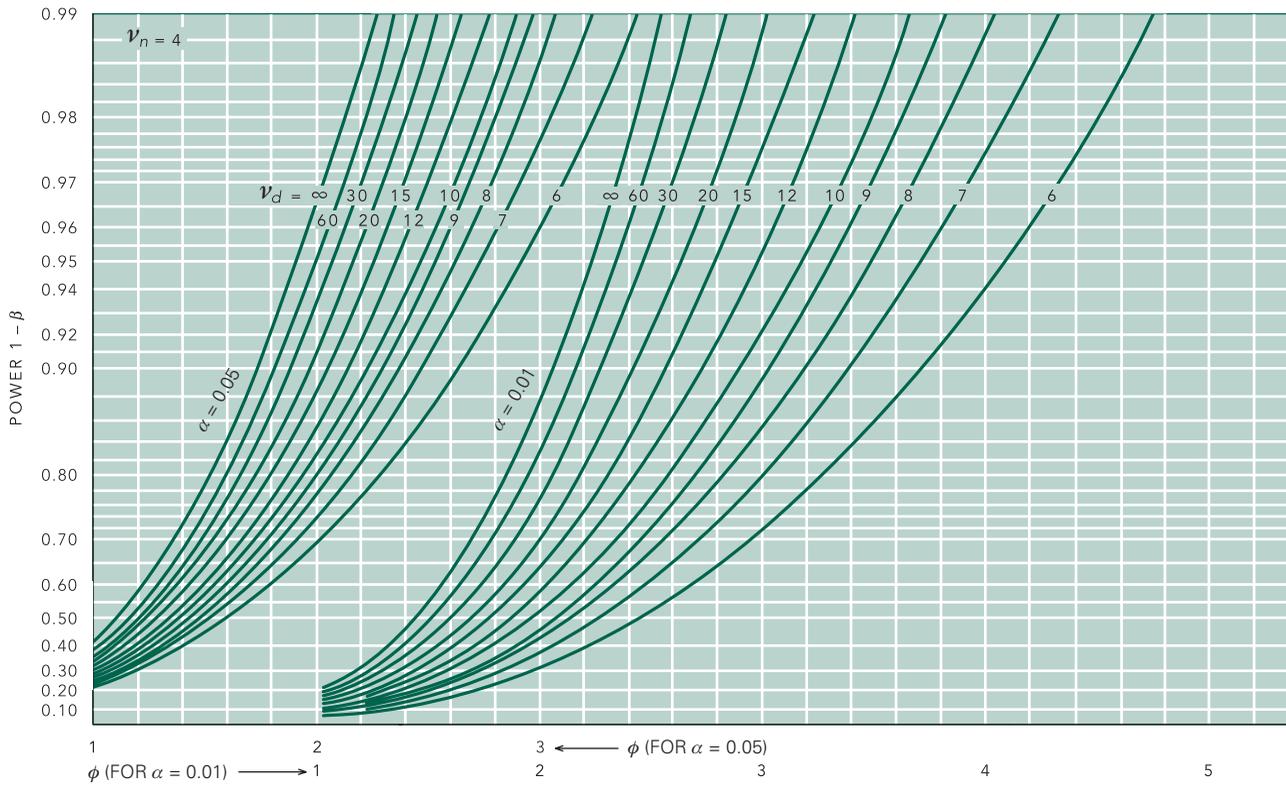


FIGURE B-4.

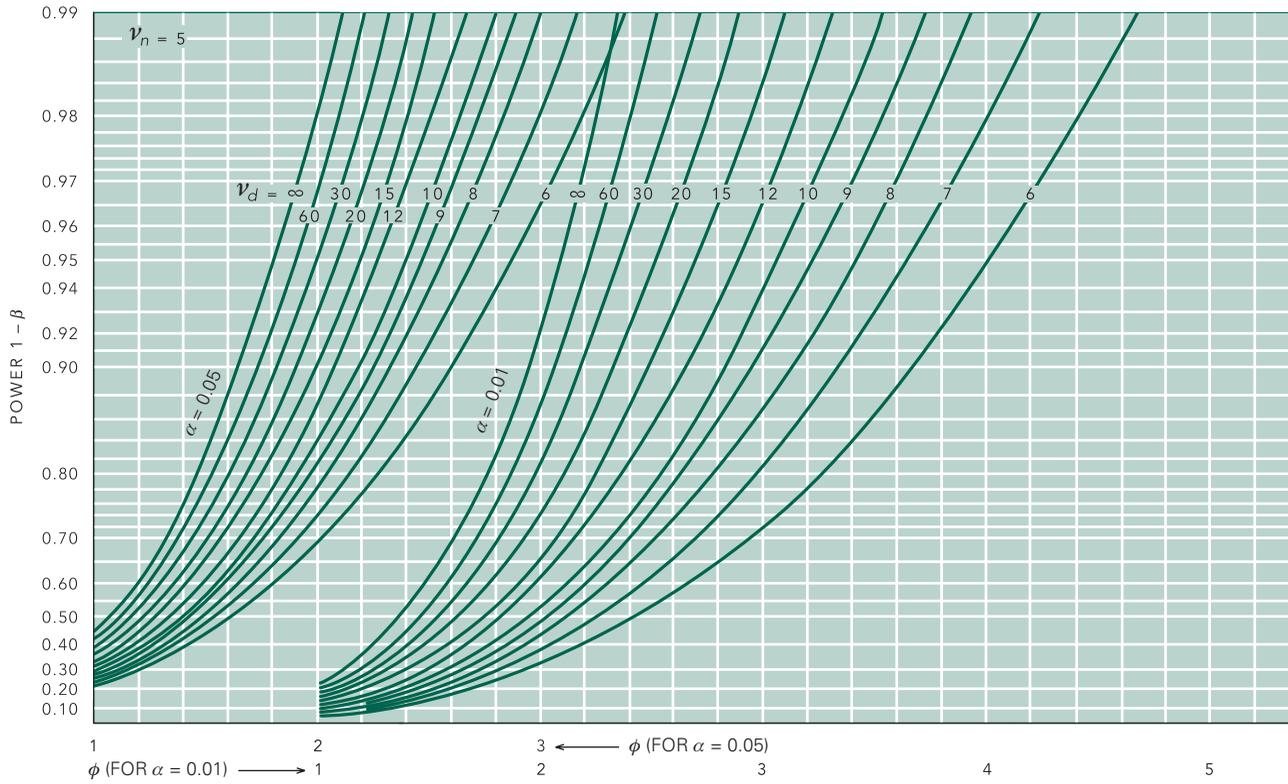


FIGURE B-5.

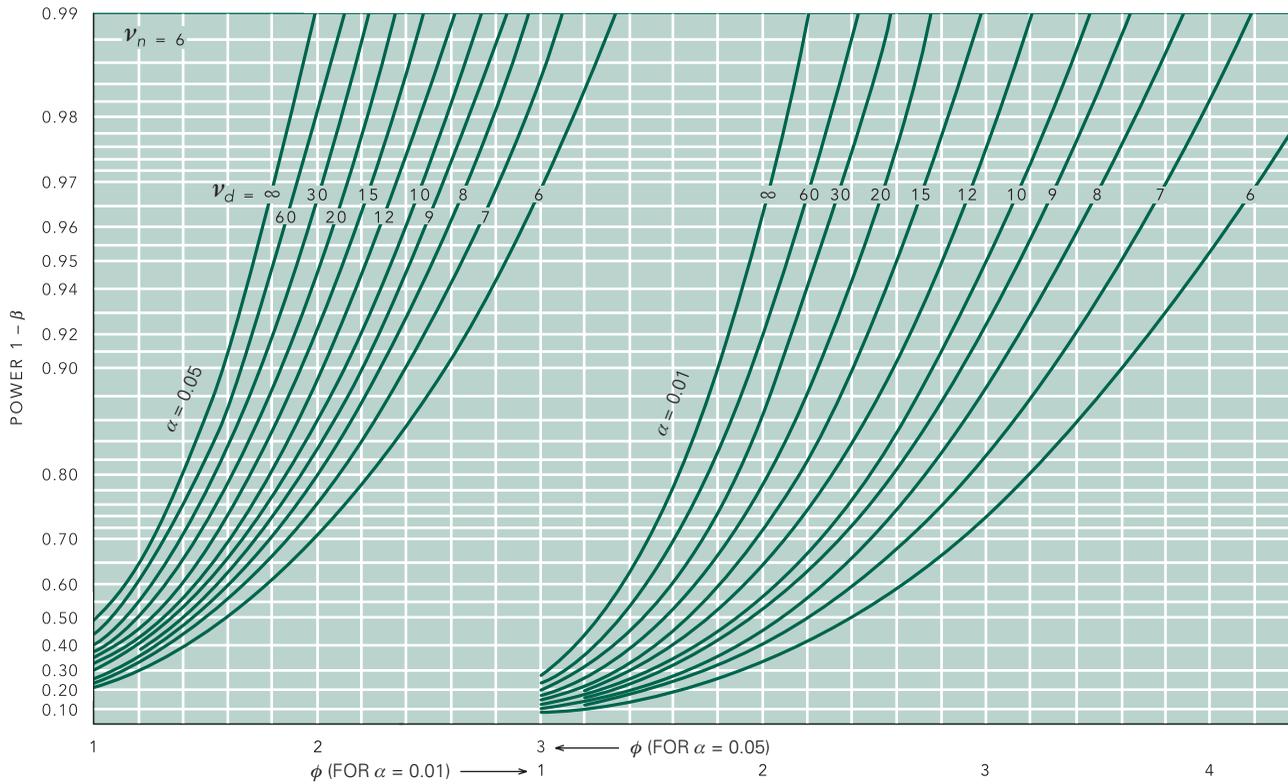


FIGURE B-6.

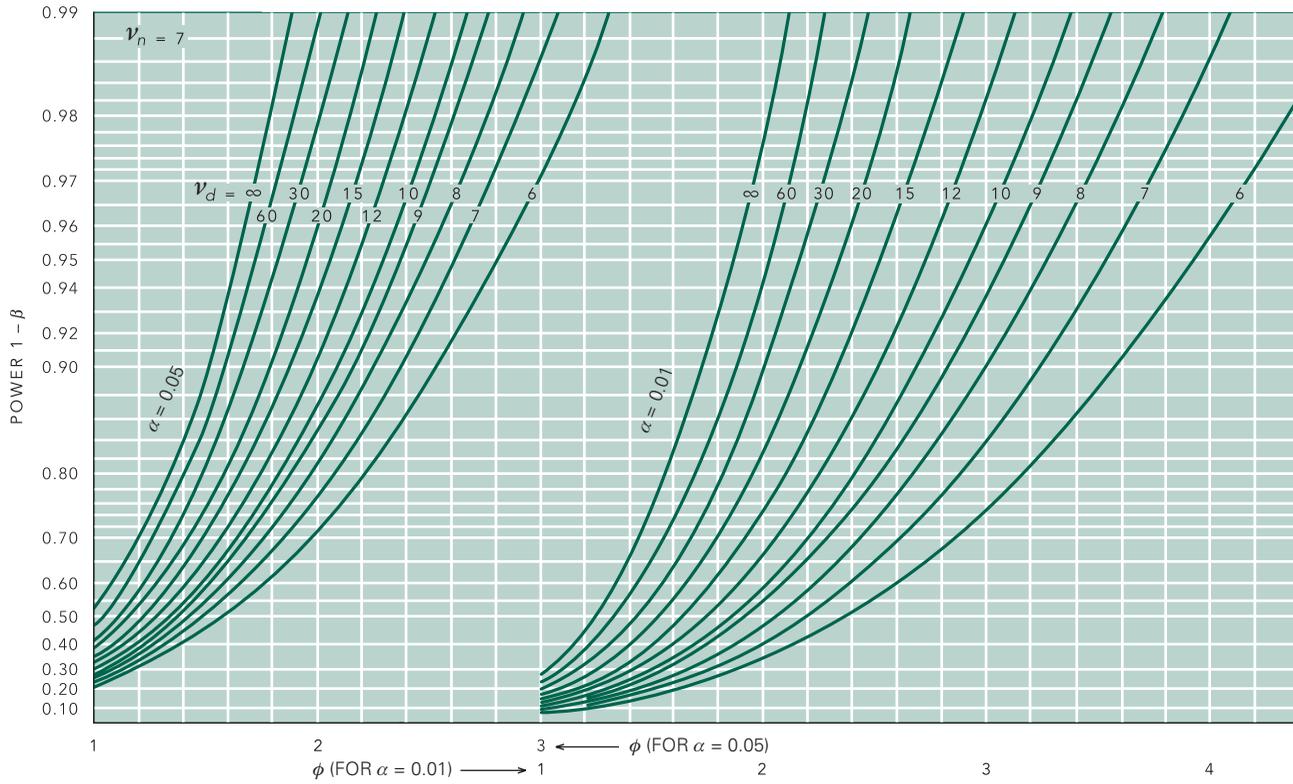


FIGURE B-7.

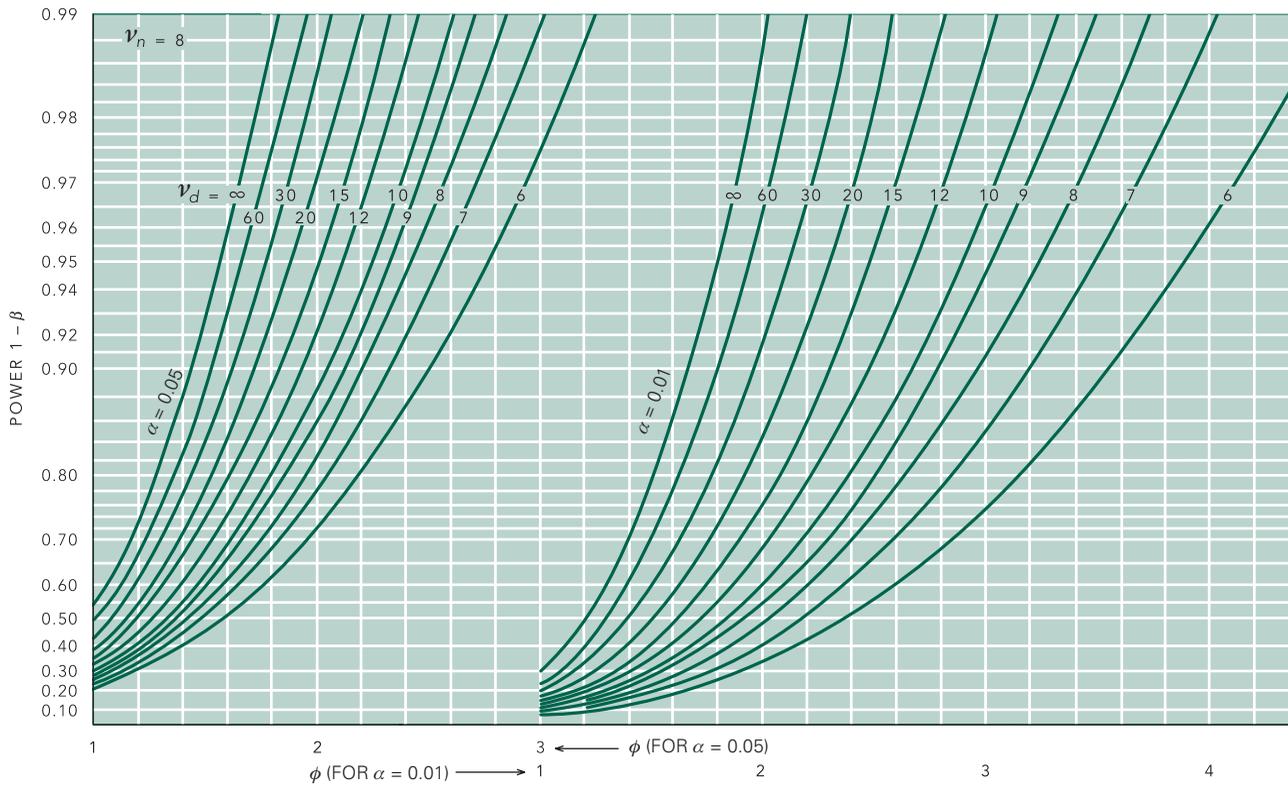


FIGURE B-8.

This page intentionally left blank

Answers to Exercises

2-1 The mean is the sum of the observations divided by the number of observations, 24: $965/24 = 40.2$. To find the median we list the observation in order, then select the $(50/100)(24 + 1) = 12.5$ th point, which is the average of the 12th and 13th observations, $(29 + 30)/2 = 29.5$. The standard deviation is the square root of the sum of the squared differences between the observations and the sample mean divided by the sample size minus 1, 29.8. The 25th percentile is the $(25/100)(24 + 1) = 6.25$. Thus, the 25th percentile is between the 6th and 7th observation, which we average to obtain $(13 + 13)/2 = 13$. Likewise, the 75th percentile is $(75/100)(24 + 1) = 18.75$, so we average the 18th and 19th observations to obtain $(70 + 70)/2 = 70$. The fact that the median is very different from the mean (29.5 versus 40.2) and not located roughly equidistant between the top and bottom quartile indicates that the data were probably not drawn from a normal distribution. (If the data were symmetrically distributed about the median, we could have further checked for normality by computing the 2.5th, 16th, 84th and 97.5th percentiles and comparing them with values 2 and 1 standard deviations below and above the mean, as described in Fig. 2-10.)

2-2 Mean = 61,668, median = 13,957, standard deviation = 117,539, 25th percentile = 8914, 75th percentile = 63,555, mean - 0.67 standard deviations = -17,083, mean + 0.67 standard deviations = 140,419. These data appear not to be drawn from a normally distributed population for several reasons. (1) The mean and median are very different. (2) All the observations are (and have to be, since you

cannot have a negative viral load) greater than zero and the standard deviation is larger than the mean. If the population were normally distributed, it would have to include negative values of viral load, which is impossible. (3) The relationship between the percentiles and numbers of standard deviations about the mean are different from what you would expect if the data were drawn from a normally distributed population.

2-3 Mean = 4.30, median = 4.15, standard deviation = 0.67, 25th percentile = 5.25, 75th percentile = 4.79, mean - 0.67 standard deviations = 3.85, mean + 0.67 standard deviations = 4.75. These data appear to be drawn from a normally distributed population on the basis of the comparisons in the answer to Prob. 2-2.

2-4 Mean = 1709, median = 1750, standard deviation = 825, 25th percentile = 825, 75th percentile = 2400, mean - 0.67 standard deviations = 1157, mean + 0.67 standard deviations = 2262. These data appear to be drawn from a normally distributed population on the basis of the comparisons in the answer to Prob. 2-1.

2-5 There is 1 chance in 6 of getting each of the following values: 1, 2, 3, 4, 5, and 6. The mean of this population is 3.5.

2-6 The result is a sample drawn from the distribution of all means of samples of size 2 drawn from the population described in Problem 2-4. Its mean is an estimate of the

population mean, and its standard deviation is an estimate of the standard error of the mean of samples of size 2 drawn from the population in Problem 2-4.

3-1 $F = 8.92$, $v_n = 1$, $v_d = 28$. These observations are not consistent with the null hypothesis that there is no difference in the average rate of ATP production in the two groups; we conclude that the rate of ATP production depends on insulin resistance ($P < .01$).

3-2 $F = 64.13$, $v_n = 4$, $v_d = 995$. Mean forced midexpiratory flow is not the same, on the average, in all of the experimental groups studied ($P < .01$).

3-3 $F = 131,700/10,601 = 12.42$ with $v_n = 2 - 1 = 1$ and $v_d = 2(21 - 1) = 40$. From the table of critical values in Table 3-1, $P < .01$.

3-4 Yes. $F = 4997/135 = 37.01$ with 3 numerator and 156 denominator degrees of freedom. Comparing this value with the table of critical values in Table 3-1, $P < .01$.

3-5 $F = 2.15$, $v_n = 1$, $v_d = 98$. This value of F is not large enough to reject the hypothesis that there is no difference in vertebral bone density between similarly aged men and women who have had vertebral bone fractures.

3-6 $F = 3.450$, $v_n = 3$, $v_d = 96$. Health professionals in at least one unit experience more burnout than those in the others ($P < .02$).

3-7 $F = 95.79$, $v_n = 3$, $v_d = 57$. At least one strain of mouse differs in response to estrogen ($P < .01$).

3-8 No. $F = 1.11$, $v_n = 4$, $v_d = 130$, which does not approach the critical value of F that defines the upper 5% of possible values under the null hypothesis of no difference among the groups, 2.37. Therefore, we cannot reject the null hypothesis that all these samples were drawn from the same population.

4-1 $s^2 = (11.6^2 + 8.8^2)/2 = 106$ and $t = (54.4 - 58.2)/\sqrt{106/29 + 106/24} = -1.338$ with $v = 29 + 24 - 2 = 51$ degrees of freedom. From Table 4-1, $.20 < P < .10$, so there is not a detectable difference in ages in the two groups.

4-2 Yes. $t = 5.916$ with 138 degrees of freedom. $P < .001$.

4-3 $t = (555 - 394)/\sqrt{65^2/21 + 65^2/21} = 8.026$ with $v = 2(21 - 1) = 40$ degrees of freedom. From Table 4-1, $P < .001$.

4-4 Because standard errors of the mean are reported, they first need to be converted to standard deviations by multiplying each by the square root of the sample size, yielding standard deviations of 10.5 and 10.4, respectively, before computing $t = -.154$ with 260 degrees of freedom. $P > .50$. There is no detectable difference in the ages.

4-5 Problem 3-1: $t = 2.986$, $v = 40$, $P < .01$; Problem 3-3: $t = 3.525$, $v = 40$, $P < .001$; Problem 3-5: $t = -1.467$, $v = 98$, $P < .01$. In both these cases, we can reject the null hypothesis of no difference between the groups. $t^2 = F$.

4-6 People who work in a smoky environment and light smokers form one subgroup; each of the other groups are distinct subgroups. Here are the results of the pairwise comparisons using a Holm-Sidak t test (with $v = 995$) with 1 = nonsmokers in smoke free environment, 2 = worked in smoky environment, 3 = light smokers, 4 = moderate smokers, 5 = heavy smokers.

Comparison	P	P _{crit}	P < .05?
1 vs. 5	<.001	.005	Yes
1 vs. 4	<.001	.006	Yes
2 vs. 5	<.001	.006	Yes
1 vs. 3	<.001	.007	Yes
3 vs. 5	<.001	.009	Yes
1 vs. 2	<.001	.010	Yes
2 vs. 4	<.001	.013	Yes
3 vs. 4	<.001	.017	Yes
4 vs. 5	.018	.025	Yes
2 vs. 3	.212	.050	No

4-7 All the groups have worse lung function than the non-smokers breathing clean air (the control group).

Comparison	P	P _{crit}	P < .05?
5 vs. 1	<.001	.013	Yes
4 vs. 1	<.001	.017	Yes
3 vs. 1	<.001	.025	Yes
2 vs. 1	<.001	.050	Yes

4-8 The control group (no exposure) and the low exposure group are not detectably different from each other. The Medium and High use men have lower sperm viability than the Control/Low exposure men and each other. Hence, there are three subgroups in the sperm viability: (1) Control and Low exposure, (2) Medium exposure, and (3) High exposure.

Comparison	<i>P</i>	<i>P</i> _{crit}	<i>P</i> < .05?
Control vs. High	<.001	.009	Yes
Low vs. High	<.001	.010	Yes
Control vs. Medium	<.001	.013	Yes
Medium vs. High	<.001	.017	Yes
Low vs. Medium	<.001	.025	Yes
Control vs. Low	.126	.050	No
<i>v</i> = 156			

4-9 There are only three comparisons since all three groups are being compared against the Control group of men who do not use cell phones. In this case, we conclude that sperm viability in the Low use men is not detectably different from Control nonusers, whereas the Medium users and High users have significantly less viable sperm than the Control users. Note that we cannot make any statements about the differences or lack of differences between the three groups of men who use cell phones.

Comparison	<i>P</i>	<i>P</i> _{crit}	<i>P</i> < .05?
Control vs. high	<.001	.017	Yes
Control vs. medium	<.001	.025	Yes
Control vs. low	.126	.050	No
<i>v</i> = 156			

4-10 The lowest burnout rate is in the Hemophilia service and the highest is in Internal Medicine. These burnout rates are significantly different from each other by Holm-Sidak *t* tests. None of the other units are significantly different from each other, including the Internal Medicine and Infectious Disease, the unit with the second highest burnout rate, which creates an ambiguity in interpreting the results. Sometimes such ambiguities arise in multiple comparison testing.

Comparison	<i>P</i>	<i>P</i> _{crit}	<i>P</i> < .05?
Med vs. Hem	.004	.009	Yes
ID vs. Hem	.013	.010	No
Onc vs. Hem	.034	.013	No
IM vs. Onc	.426	.017	No
Med vs. ID	.682	.025	No
ID vs. Onc	.698	.050	No
<i>v</i> = 96			

4-11 a No, **b** No, **c** No, **d** Yes.

5-1 Yes. $\chi^2 = 1.247$, *v* = 1, *P* = .264; no.

5-2 Violent suicide: $\chi^2 = 1.380$, Yates corrected $\chi^2 = 0.870$, *v* = 1, *P* > 0.25; suicide under the influence of alcohol: $\chi^2 = 18.139$, Yates corrected $\chi^2 = 16.480$, *v* = 1, *P* < .001; BAC > = 150 mg/dL: $\chi^2 = 19.204$, Yates corrected $\chi^2 = 17.060$, *v* = 1, *P* < .001; suicide during weekend: $\chi^2 = 4.850$, Yates corrected $\chi^2 = 4.020$, *v* = 1, *P* < .05; parental divorce: $\chi^2 = 5.260$, Yates corrected $\chi^2 = 4.340$, *v* = 1, *P* < .05; parental violence: $\chi^2 = 9.870$, Yates corrected $\chi^2 = 8.320$, *v* = 1, *P* < .01; parental alcohol abuse: $\chi^2 = 4.810$, Yates corrected $\chi^2 = 3.890$, *v* = 1, *P* < .05; paternal alcohol abuse: $\chi^2 = 5.630$, Yates corrected $\chi^2 = 4.570$, *v* = 1, *P* < .05. The key factors seem to be suicide under the influence of alcohol, BAC ≥ 150 mg/dL, suicide during weekend, parental divorce, parental violence, parental alcohol abuse, and paternal alcohol abuse. Despite the high confidence we can have in reporting these differences, they probably are not stark enough to be of predictive value in any given adolescent.

5-3 There is a possibility that the fact that the families declined to be interviewed reflects a systematic difference between the 106 suicides that were included and the ones that we excluded. One way to investigate whether this situation leads to biases would be to compare what is known about the families that granted interviews and ones that did not (using variables such as age, socioeconomic status, gender of victim) to see if there were any systematic differences. If there were no differences, the lack of interviews is probably not a problem. If there are differences, the lack of interviews could bias the conclusions of the analysis.

5-4 For the three groups, $\chi^2 = 21.176$, $\nu = 2$, $P < .001$, so there is evidence that at least one group differs in the number of remissions. Comparing just nefazodone and psychotherapy yields:

	Remission	No Remission
Nefazodone	36	131
Psychotherapy	41	132

$\chi^2 = 0.220$, Yates corrected $\chi^2 = 0.120$, $\nu = 1$, $P > .6$. Comparing nefazodone to nefazodone and psychotherapy yields:

	Remission	No Remission
Nefazodone	36	131
Nefazodone and psychotherapy	75	104

$\chi^2 = 15.488$, $\nu = 1$, $P < .001$. Comparing psychotherapy with nefazodone and psychotherapy yields:

	Remission	No Remission
Psychotherapy	41	132
Nefazodone and psychotherapy	75	104

The Holm-Sidak table for the pairwise comparisons is

Comparison	χ^2	P	P_{crit}	$P < .05?$
N vs. N and P	15.488	<.001	.017	Yes
P vs. N and P	12.378	<.001	.025	Yes
N vs. P	.120	>.6	.050	No
$\nu = 156$				

Nefazodone alone and psychotherapy alone have similar performance, which differs from nefazodone and psychotherapy combined.

5-5 There is an association between funding source and whether or not the study concluded that smoking restrictions harmed the hospitality industry. $\chi^2 = 71.861$, $\nu = 1$,

$P < .001$. (Without the Yates correction $\chi^2 = 75.871$.) Note that a χ^2 test is appropriate for these data even though the *observed* counts are less than 5 because the *expected* counts all exceed 5. It is the expected, not observed, counts that determine whether or not the χ^2 test can be used. The odds ratio for a study concluding that smoking restrictions was supported by the tobacco industry or one of its allies is $OR = (29 \times 60)/(2 \times 2) = 435$.

5-6 Being supported by a single drug company was not associated with the results of the meta-analysis ($\chi^2 = 1.301$, $\nu = 1$, $P > .25$) but was associated with the conclusions that were presented ($\chi^2 = 5.369$, $\nu = 1$, $P < .025$), suggesting that there were no biases in the conduct of the meta-analysis, but there was in how the results were presented. (χ^2 values without the Yates correction are .912 and 6.501.)

5-7 For honorary authors among all journals, the contingency table is

Journal	No Honorary Authors	Articles with Honorary Authors
<i>American Journal of Cardiology</i>	115	22
<i>American Journal of Medicine</i>	87	26
<i>American Journal of Obstetrics and Gynecology</i>	111	14
<i>Annals of Internal Medicine</i>	78	26
<i>Journal of the American Medical Association</i>	150	44
<i>New England Journal of Medicine</i>	112	24

$\chi^2 = 11.026$, $\nu = 5$, $.05 < P < .10$, so we do not reject the null hypothesis that the rate of honorary authorship does not vary among journals. Since we did not reject the null hypothesis based on all the journals, there is no need to subdivide the table between small and large circulation journals. (This negative conclusion should be taken as tentative, since the critical value of χ^2 for $P = .05$ is 11.070, which the data just misses.) Overall, 156 of 809 articles (19%) included honorary authors.

For ghost authors among all journals, the contingency table is

Journal	No Ghost Authors	Articles with Ghost Authors
<i>American Journal of Cardiology</i>	124	13
<i>American Journal of Medicine</i>	98	15
<i>American Journal of Obstetrics and Gynecology</i>	112	13
<i>Annals of Internal Medicine</i>	88	16
<i>Journal of the American Medical Association</i>	180	14
<i>New England Journal of Medicine</i>	114	22

$\chi^2 = 8.331$, $\nu = 5$, $.25 < P < .10$, so we do not reject the null hypothesis that the rate of ghost authorship does not vary among journals. Since we did not reject the null hypothesis based on all the journals, there is no need to subdivide the table between small and large circulation journals. Overall, 93 of 809 articles (11%) had ghost authors.

5-8 $\chi^2 = 4.880$, Yates corrected $\chi^2 = 4.450$, $\nu = 1$, $P < .05$; yes.

5-9 This is a prospective study so we can compute a relative risk. Treating combination therapy as the treatment condition and valproate alone as the control condition, $n_{TD} = 59$, $n_T = 110$, $n_{CD} = 76$, and $n_C = 110$, so the relative risk of an emergent mood episode for people being treated with combination therapy compared to valproate alone is $RR = (59/110)/(76/110) = .78$. People being treated with combination therapy were less likely to have an episode than people being treated with valproate alone. To see if this difference is larger than expected by chance, we compute χ^2 for the 2×2 contingency table:

Therapy	Emergent Episode	
	Yes	No
Combination	59	51
Valproate	76	34

For this table, $\chi^2 = 4.908$, $\nu = 1$, $P < .05$, so the difference is statistically significant. (Without the Yates correction $\chi^2 = 5.541$.) The absolute risk reduction is $.69 - .54 = .15$, so the number needed to treat to prevent one emergent episode is $1/.15 = 6.7$ or 7 people.

5-10 $\chi^2 = 8.8124$, $\nu = 1$, $P < .005$. She would not reach the same conclusion if she observed the entire population because the sample would not be biased by differential admission rates.

5-11 $OR = 1.40$, $\chi^2 = 14.122$, $\nu = 1$; $P < .001$. Smoking significantly increases the odds of developing renal cell cancer.

5-12 $OR = 0.74$, $\chi^2 = 4.556$, $\nu = 1$; $P = .03$. Stopping smoking significantly reduces the risk of renal cell cancer.

5-13 $RR = 0.58$, $\chi^2 = 127.055$, $\nu = 1$; $P < .001$. Hormone replacement therapy is associated with a reduction in risk of death compared with nonusers.

5-14 $RR = 1.00$, $\chi^2 = .002$, $\nu = 1$; $P = .962$. Past use of hormone replacement therapy did not affect the risk of death compared to never users.

6-1 $\phi = \delta/\sigma = 25/35 = .7$. From Figure 6-9, the power is .60.

6-2 From Figure 6-9, we would need 35 people in each diet group.

6-3 From Figure 6-9, $\phi = .9$. Using a standard deviation of 35 mg/dL, $\delta = 31.8$ mg/dL.

6-4 The power is 93% based on a difference in bone density of 14, which is 20% of 70.3.

6-5 Twenty people in each group, based on a difference of 21, which is 30% of 70.3.

6-6 Power = .80.

6-7 From Problem 3-3, use the standard deviation for the normal people, 121 W, as the estimate of σ . The sample size per group is $n = 21$. For a 50 W change $\phi = \delta/\sigma = 50/121 = .4$. From Figure 6-9, the power is .26. For a 100 W change, the power is .74.

6-8 $n = 37$ per sample.

6-9 The desired pattern of the responses is

Antibiotic	Remission	No Remission	Total
Nefazodone	.107	.215	.322
Psychotherapy	.111	.222	.333
Both	.172	.172	.345
Total	.390	.609	1.000

$\phi = 2.6$, $v_n = (3 - 1)(2 - 1) = 2$, so from Figure 6-10, power = .98.

6-10 $N \approx 367$.

7-1 95% confidence interval: 1233 to 2185 ng/g; 90% confidence interval: 1319 to 2100 ng/g.

7-2 95% confidence interval for the difference: .72 to 3.88 $\mu\text{mol/g}$ of muscle/min. Since this interval does not include 0, we reject the null hypothesis of no difference ($P < .05$).

7-3 95% confidence intervals: Anesthetic gel: 0 to .17; Placebo: .08 to .32; Difference $-.28$ to $+.04$. We cannot reject the null hypothesis of no difference in effect between the placebo and the anesthetic gel. This is the same conclusion we reached in Problem 5-1.

7-4 The 95% confidence interval for the difference is 120 to 201 meters. Because the confidence interval includes zero, we cannot reject the null hypothesis of no difference with $P < .05$. In Problem 4-3, we rejected the null hypothesis of no difference with $P < .001$, which would be the same as checking if the 99.9% confidence interval excludes 0. The 99.9% confidence interval for the difference extends from $161 - 3.2905 \times 20.7 = 92.8$ to $161 + 3.2905 \times 20.7 = 229.1$, which does not include zero, so we could have obtained the same level of confidence in rejecting the null hypothesis using confidence intervals as we did using a t test in Chapter 4.

7-5 The standard error of the proportion for the 49 studies funded by a single drug company is .071 so the 95% confidence interval for the proportion with positive results extends from $.55 - 1.960 \times .071 = .41$ to $.55 + 1.960 \times .071 = .69$. For the 75 studies funded in other ways, the standard error of the proportion is .055, so the 95% confidence interval extends from $.65 - 1.960 \times .055 = .54$ to $.65 + 1.960 \times .055 = .76$.

7-6 95% confidence interval for 90% of the population: -518 to 3936 ng/g lipid; 95% confidence interval for 95%

of the population: -930 to 4349 ng/g lipid. The negative numbers at the lower ends of the confidence intervals are possible members of the actual populations; these negative numbers reflect the conservative nature of this computation based on small sample sizes.

7-7 $s_{\ln} \text{OR} = \sqrt{1/29 + 1/2 + 1/2 + 1/60} = 1.025$ so the 95% confidence interval for the odds ratio is

$$e^{\ln 435 - 1.96 \times 1.025} < \text{OR} < e^{\ln 435 + 1.96 \times 1.025}$$

$$e^{4.066} < \text{OR} < e^{8.084}$$

$$58 < \text{OR} < 3242$$

The 95% confidence interval does not include 1, so we can reject the null hypothesis that the funding source does not affect the conclusions drawn from the meta-analyses.

7-8 $\text{OR} = 1.40$. The 95% confidence interval is from 1.18 to 1.66, which does not include 1. Therefore, we conclude that smoking significantly increases the risk of renal cell cancer.

7-9 $\text{OR} = 0.74$. The 95% confidence interval is from .57 to .96, which does not include 1. Therefore, we conclude that stopping smoking significantly reduces the odds of developing renal cell cancer.

7-10 $\text{RR} = 0.61$. The 95% confidence interval is from .56 to .66, which does not include 1. Therefore, we conclude that hormone replacement therapy reduces the risk of death.

7-11 $\text{RR} = 1.00$. The 95% confidence interval is from .94 to 1.07, which includes 1. Therefore, we cannot conclude that past use of hormone replacement therapy affects the risk of death.

8-1 a: $a = 3.00$, $b = 1.30$, $r = .792$; **b:** $a = 5.10$, $b = 1.24$, $r = .941$; **c:** $a = 5.60$, $b = 1.23$, $r = .973$. Note that as the range of data increases, the correlation coefficient increases.

8-2 a: $a = 24.3$, $b = .36$, $r = .561$; **b:** $a = 0.5$, $b = 1.15$, $r = .599$. Part **a** illustrates the large effect one outlier point can have on the regression line. Part **b** illustrates that even though there are two different and distinct patterns in the data, this is not reflected when a single regression line is drawn through the data. This problem illustrates why it is important to look at the data before computing regression lines through it.

8-3 $a = 3.0$, $b = 0.5$, $r = .82$ for all four experiments, despite the fact that the patterns in the data differ from experiment to experiment. Only data from experiment 1 satisfies the assumption of linear regression analysis.

8-4 Yes. As maternal milk PCB levels increase, children's IQ at 11 years of age falls; the slope is $-.021$ (standard error $.00754$, so $t = -2.785$ with 12 degrees of freedom; $P < .05$). The Pearson product-moment correlation, r , is $-.63$ (also $P < .05$). You could also have tested the hypothesis of no relationship with a Spearman rank-order correlation, which would yield $r_s = -.610$ ($P < .05$).

8-5 Use the Bland-Altman method to compare the two methods of estradiol measurement. The mean difference is -25.9 pg/mL and the standard deviation of the differences is 19.4 pg/mL. These results suggest that there is not particularly good agreement between the two methods, with the blood spot yielding lower results and a substantial amount of variability in the results of the two methods in comparison with the magnitude of the observations.

8-6 These regression results are computed after conducting the regressions of relaxation force as the dependent variable against \ln (arginine level) as the independent variable:

	Slope	Intercept	$s_{y \cdot x}$	P
Acetylcholine	-7.85	-50.5	13.80	.024
A23187	-10.3	-57.1	15.03	.009
Common estimate	-9.03	-54.0	14.08	.001

To do the overall test of coincidence, we compute

$$s_{y \cdot x_p}^2 = \frac{(11-2)13.80^2 + (13-2)15.10^2}{11+13-4} = 211.40$$

and

$$s_{y \cdot x_{imp}}^2 = \frac{(11+13-2)14.16^2 - (11+13-4)211.40}{2} = 91.56$$

so that $F = 91.56/211.40 = .433$ with $v_n = 2$ and $v_d = 20$, which does not even approach the critical value of 3.49 required to reject the null hypothesis of no difference with $P < .05$. Therefore, we cannot reject the null hypothesis that there is no difference between the two relationships; given the very small value of F , we can be reasonably confident in concluding that the two different stimuli have similar effects on force levels (arterial relaxation).

8-7 There is a significant relationship. $r_s = .912$, $n = 20$, $P < .001$.

8-8 $r_s = .472$, $n = 25$, $P = .018$ (including tie adjustment). There is a significant relationship between these two different ways of measuring the extent of cancer, but the correlation is weak enough that they cannot be used interchangeably for clinical purposes. Results without tie adjustment: $r_s = .402$, $n = 25$, $P = .047$.

8-9 Power = .999.

8-10 $n = 20$, so this study could have been done with a smaller sample size than was actually used.

8-11 To answer the question, we fit linear regressions to the two groups of men, then do an overall test of coincidence. For controls $I = -1.77R + 2.59$, $r = -0.800$, $s_{\text{slope}} = 0.369$, $s_{\text{intercept}} = 0.336$, $s_{I \times R} = 0.125$, $n = 15$. For relatives: $I = -0.18R + 0.932$, $r = -0.075$, $s_{\text{slope}} = 0.651$, $s_{\text{intercept}} = 0.932$, $s_{I \times R} = 0.219$, $n = 15$. For common regression: $I = -1.09R + 1.88$, $r = -0.432$, $s_{\text{slope}} = 0.441$, $s_{\text{intercept}} = 0.405$, $s_{I \times R} = 0.211$, $n = 30$. Overall test of coincidence: $F = 6.657$ with $v_n = 2$ and $v_d = 26$; $P < .01$; the relationships are different. Test for difference in slopes: $t = -2.137$, $v = 26$, $P < .05$. Test for difference in intercepts: $t = 2.396$, $v = 26$, $P < .05$. Therefore, the slopes and intercepts of the two lines are significantly different. The relationship between physical fitness and insulin index is different in these two groups of men.

9-1 The mean difference is 1.18 with a standard error of .32. $t = -3.668$ with $v = 14 - 1 = 13$ degrees of freedom. From Table 4-1, $P < .005$. The 95% confidence interval for the difference is from .49 to 1.87.

9-2 There is a significant difference. $t = 6.160$ with $v = 7$, $P < .001$.

9-3 $\delta = 9$ ms (half the 18 ms difference observed in Prob. 9-2) and $\sigma = 8.3$ ms, the standard deviation of the differences before and after breathing secondhand smoke so the noncentrality parameter $\phi = 9/8.3 = 1.1$. From the power chart in Figure 6-9, the power is .75.

9-4 $F = 37.94$, $v_n = 1$, $v_d = 7$, $P < .01$. $F = t^2$.

9-5 $F = 0.519$, $v_n = 2$, $v_d = 6$. This value falls far short of 5.14, the critical value that defines the greatest 5% of possible values of F in such experiments. Thus, we do not have sufficient evidence to conclude that there are differences in C reactive protein over time ($P > .50$).

9-6 There are significant differences between the different experimental conditions ($F = 50.77$, $v_n = 3$, $v_d = 33$). Multiple comparisons using the residual mean square and Holm-Sidak t test show that testosterone levels are greater before capture than at any time after. In addition, testosterone levels after 48 hours of capture are decreased compared to time of capture and 12 hours post-capture, which do not differ.

9-7 By McNemar's test: $\chi^2 = 4.225$, $v = 1$, $P < .05$. No; indomethacin is significantly better than placebo.

9-10 When the data are presented in this format, they are analyzed as a 2×2 contingency table. $\chi^2 = 2.402$, $v = 1$, $P < .10$, so there is no significant association between drug and improvement of shunting. This test, in contrast to the analysis in Prob. 9-8, failed to detect an effect because it ignores the paired nature of the data, and so is less powerful.

10-1 $z_T = 2.080$, $P < .05$; there is a significant difference in the level of adhesions between the two groups. (Adjusting for ties, $z_T = 2.121$, $P < .05$.)

10-2 A Kruskal-Wallis test yields $H = 15.161$ with $v = 3$, $P = .002$. There is a significant difference among the treatments. The table below shows the results of pairwise comparisons using Mann-Whitney tests with a Holm-Sidak correction with a 5% family error rate.

Comparison	z_T	P	j	$P_{crit} = \alpha_T / (k-j+1)$	$P < P_{crit}?$
Baseline vs. C&L	3.609	<.001	1	.0102	Yes
Info vs. C&L	1.993	.046	2	.0127	No
Info vs. Cards	1.674	.094	3	.0170	No*
Baseline vs. Info	1.588	.112	4	.0253	No*
Cards vs. C&L	.289	.773	5	.0500	No*

*Because the second comparison is not significant, all subsequent comparisons are considered not significant.

This analysis yields ambiguous results in that it shows that none of the interventions are significantly different from

each other while the baseline differs from the cards plus the lectures, but none of the other interventions.

10-3 Problem 9-5: Endotoxin and salbutamol did not affect CRP levels ($\chi_r^2 = 1.5$, $k = 3$, $n = 4$, $P > .05$).

10-4 Capture produced significant differences in testosterone levels ($\chi_r^2 = 27.3$, $v = 3$, $P < .001$). The table below shows the results of pairwise comparisons using Wilcoxon signed-rank tests with a Holm-Sidak correction with a 5% family error rate.

Comparison	z_w	P	j	$P_{crit} = \alpha_T / (k-j+1)$	$P < P_{crit}?$
Begin vs. capture	3.039	0.002	1	.0085	Yes
Begin vs. 12 h	3.039	0.002	2	.0102	Yes
Begin vs. 24 h	3.039	0.002	3	.0127	Yes
12 h vs. 24 h	2.961	0.003	4	.0170	Yes
Capture vs. 12 h	1.863	0.062	5	.0253	No
Capture vs. 24 h	1.627	0.103	6	.0500	No*

*Because the second comparison is not significant, all subsequent comparisons are considered not significant.

The testosterone levels increase significantly between the beginning of the study to after capture. From capture through the end of the experiment the levels are not detectably different (although there is a detected difference between 12 and 24 hours).

10-5 $T = 195.0$, $n_S = 15$, $n_B = 15$; $z_T = 1.535$ and $P > 10$. They do not appear to have different cholesterol levels.

10-6 The Mann-Whitney rank-sum test yields $z_T = 3.870$ ($P < .001$), so problem gambling substance abusers exhibit riskier sexual behavior than non-problem gambling substance abusers.

10-7 $W = -91.0$ with $n = 14$, so compute

$$\sigma_w = \sqrt{14(14+1)(2 \times 14+1)/6} = 31.86$$

so $z_w = (|-91| - 1/2) / 31.86 = 2.841$ and $P < .005$.

10-8 Yes, G is a legitimate test statistic. The sampling distribution of G when $n = 4$:

G	Possible Ways to Get Value	Probability
0	1	1/16
1	4	4/16
2	6	6/16
3	4	4/16
4	1	1/16

When $n = 6$:

G	Possible Ways to Get Value	Probability
0	1	1/64
1	6	6/64
2	15	15/64
3	20	20/64
4	15	15/64
5	6	6/64
6	1	1/64

G cannot be used to conclude that the treatment in the problem had an effect with $P < .05$ because the two most extreme possible values (i.e., the two tails of the sampling distribution of G), 0 and 4, can occur $1/16 + 1/16 = 1/8 = 0.125 = 12.5\%$ of the time, which exceeds 5%. G can be used for $n = 6$, where the extreme values, 0 and 6, occur $1/64 + 1/64 = 2/64 = .033\%$ of the time, so the (two-tail) critical values (closest to 5%) are 1 and 6.

11-1 Here is the survival curve in tabular form:

Month	Cumulative Survival, $\hat{S}(t)$	Standard Error	95% Confidence Interval	
			Lower	Upper
1	0.971	0.028	0.916	1.000
2	0.943	0.039	0.866	1.000
3	0.857	0.059	0.741	0.973
4	0.828	0.064	0.702	0.953
5	0.798	0.068	0.664	0.932
6	0.768	0.072	0.628	0.909
7	0.709	0.078	0.557	0.861
8	0.680	0.080	0.524	0.836
9	0.650	0.082	0.490	0.810
12	0.582	0.086	0.413	0.751
13	0.548	0.088	0.376	0.719
15	0.513	0.089	0.340	0.687
16	0.411	0.088	0.237	0.584
20	0.308	0.084	0.144	0.472
21	0.274	0.081	0.115	0.433
28	0.235	0.079	0.081	0.389
34	0.196	0.075	0.049	0.342
56	0.130	0.073	0.000	0.273
62	0.065	0.059	0.000	0.180
84	0.000	0.000	0.000	0.000

The median survival time is 16 months.

11-2 The survival curves for the two groups are:

High IADL Score		Low IADL Score	
Month	Survival, $\hat{S}_{Hi}(t)$	Month	Survival, $\hat{S}_{Lo}(t)$
14	0.988	6	0.967
20	0.963	12	0.934
24	0.925	18	0.867
28	0.913	24	0.85
30	0.887	28	0.782
38	0.861	32	0.714
48	0.834	36	0.643
		42	0.584
		47	0.522
		48	0.48

Use the log rank test to compare the two survival curves. The sum of the differences between expected and observed number of survivals at each time is -13.243 ; the standard error of the differences is 3.090 , so $z = -4.285$ (or -4.124 with the Yates correction). We conclude that there are significant differences in survival between these two groups of people, $P < .001$.

11-3 $\psi = \ln .4 / \ln .2 = .569$ so the number of required deaths is

$$d = (1.960 + .842)^2 \left(\frac{1 + .569}{1 - .569} \right)^2 = 278.9$$

and the number of people in each group would need to be $n = 278.9 / (2 - .4 - .2) = 199$.

11-4 The survival curves and calculations for the log rank test are in the table below. The median time to event for combined therapy was 21 months and 9 months for valpone therapy. So $U_L = -15.497$ and $s_{U_L}^2 = 27.654$ and $z = (|-15.497| - 1/2) / \sqrt{27.654} = 2.852$, $P < .005$. The combined therapy produced better results than valpone alone.

Month	Survival Curves		Log Rank Test					
	Combined Therapy	Valpote	Total		Fraction with Events	Expected Number of Events	Difference	Contribution to $s_{U_L}^2$
			At Risk	Events				
0	1.000	1.000	220	0	0.000	0.000	0.000	0.000
3	0.873	0.691	220	48	0.218	24.000	-10.000	9.425
6	0.718	0.523	170	35	0.206	19.765	-2.765	6.873
9	0.625	0.457	133	17	0.128	9.842	0.158	3.642
12	0.560	0.429	115	10	0.087	5.826	1.174	2.240
15	0.522	0.368	101	10	0.099	5.842	-1.842	2.211
18	0.502	0.337	89	5	0.056	2.978	-0.978	1.150
21	0.459	0.279	76	9	0.118	5.566	-1.566	1.897
24	0.447	0.279	53	1	0.019	0.679	0.321	0.218
27	0.447	0.279	26	0	0.000	0.000	0.000	0.000
30	0.447	0.279	4	0	0.000	0.000	0.000	0.000
33	0.447		1	0	0.000	0.000	0.000	0.000
36			0	0	0.000	0.000	0.000	0.000
						Total	-15.497	27.654

Index

The *n* after a page number indicates footnote.

A

- Adenosine triphosphate (ATP), 44
- Adhesions, 226
- Adult leukemia, bone marrow transplant for treatment of, 237–242, 254
- Advance directives, 72, 73, 77–78, 79–80, 139
- AIDS, 25, 46, 227–228
- Allogenic versus autologous transplants, 237–242, 254
- Analysis of variance:
 - assumptions, 27–28, 34, 205
 - between groups sums of squares, 192–139, 193
 - between groups variance, 32, 193
 - Bonferroni *t* test to isolate differences, 62–63, 67,
 - class of procedures, 27
 - computational formulas, 259
 - degrees of freedom, 38, 193
 - Dunnett's test to isolate differences, 64*n*,
 - examples, 38–44, 56–59, 190–193
 - F*, 34–38, 192–193, 194
 - general approach, 27–30, 189–190
 - Holm *t* test to isolate differences, 64–65, 67, 200
 - Holm-Sidak test to isolate differences, 65–67
 - limitations, 205
 - mean squared, 193
 - method based on ranks (*See* Kruskal-Wallis test)
 - multiple comparisons procedures, 62–67, 87, 88, 200, 219–222, 225
 - notation in terms of sums of squares, 190–193
 - null hypothesis, 27, 189
 - one way, 34–38, 259
 - paired *t* test, 181–187
 - parametric method, 29–30
 - partitioning sums of squares, 193–194
 - power, 116, 117
 - power function, 277–285
 - repeated measures (*See* Repeated measures analysis of variance)
 - sample size, 116
 - single-factor, 34–38
 - Student-Newman-Keuls test to isolate differences, 64*n*,
 - t* test, 56, 59–60, 186–187
 - table, 194
 - total sum of squares, 193
 - treatment of sum of squares, 190–193
 - Tukey test to isolate for differences, 64*n*,
 - Two-way, 44
 - unpaired *t* test (*See t* test, unpaired)
 - when to use, 205
 - within groups sum of squares, 192
 - within groups variance, 31
 - (*See also t* test, paired; *t* test, unpaired)
- Anesthesia, 95
- Angina pectoris, 251
- Anti-asthmatic drugs and endotoxin, 197–200, 224–225
- Antibiotics:
 - inappropriate use, 226–227
 - prescribing, 136
- Area under the curve (to define critical value of test statistic), 34–38
- Arterial function, 181
- Arthritis, rheumatoid, 161–164
- Association:
 - versus causality, 144–145
 - and correlation, 164
 - and regression, 143, 144–145
 - (*See also* Pearson product-moment correlation coefficient; Spearman rank correlation coefficient)

- Assumptions:
 for analysis of variance, 27–28, 34–38, 205
 for Bernoulli trials, 78
 for confidence intervals, 126
 for paired t test, 186, 205
 for unpaired t test, 205
- Authorship, 96–97
 ghost authors, 96
 honorary authors, 96
- Autologous versus allogenic transplants, 237–242, 254
- Autopsy data, 98*n*
- Average (*See* Mean)
- B**
- Bayes' Rule, 70*n*, 70
- Bayesian decision-making:
 approach 70
 interpretation of P values, 70*n*
- Berkson's fallacy, 257
- Bernoulli trials, 78
- Best fit (*See* Linear regression)
- Beta error (*See* Type II error)
- Between groups sum of squares, 192–193
- Between groups variance, 32
- Bias:
 on behalf of treatment, 5
 and control group, 250, 252
 definition, 12
 due to observers, 187–189
 due to poor design, 5
 examples of, 12–13
 placebo effect, 12–13
 and randomization, 29, 250, 252–253
 in a sample, 5, 12–13
 in selection process, 252
 sources, 12–13
- Binomial distribution, 77*n*, 137, 137*n*
- Bipolar disorder, 97–98, 245
- Bland-Altman test:
 calibration, assessed with, 174–177
 correlation, contrasted with, 174–175
 description, 174–175
 example, 175–177
 when to use, 249
- Blinding, 12, 13
- Boiling oil, for gunshot wounds, 254
- Bone marrow transplant for treatment of adult leukemia, 237–242, 254
- Bonferroni t test (or Bonferroni correction):
 basis of multiple comparison procedure, 62–63, 64
 compared with exact Type I error, 65
 control group, 67
 definition, 62–63
 examples, 63–64, 87,
 versus Holm-Sidak test, 65–67
 versus Holm t test, 64–65
 inequality, 62, 65
 for Mann-Whitney rank sum test, 219–220
 multiple comparisons, 62–63
 rejective criterion, 64–65
 subdividing contingency table, 87
 tests of accumulating data, 253*n*
 for unpaired t test, 62
- Breast cancer, 94–95, 140, 200–202
- Burnout, 45
- C**
- Cancer:
 breast, 94–95, 140, 200–202
 lung, 13–14, 244
 mouth, 181
 renal, 98
- Cannabis for pain control in diabetic neuropathy, 23, 71, 211–213
- Case
- Case-control study:
 computation, 93–94
 definition, 93
 example, 94–95
 identifying, 93
 and odds ratio, 93–94
 versus prospective study, 93–94
- Causality:
 versus association, 144–145
 in observational versus experimental study, 41
 and regression, 144–145
- Cell phones and sperm function, 38–42, 56–59, 63–64, 156–159, 172–173, 247–248
- Censored data:
 definition, 230
 to estimate survival curve, 230–234
 left censoring versus right censoring, 230*n*

- Central limit theorem:
 and confidence intervals, 137*n*
 implications, 18
 and proportions, 76
 and regression, 137*n*, 154, 156
 statement, 18
- Chi-square:
 analysis of contingency table, 80–86, 85, 201, 249, 260
 Bonferroni correction for multiple comparisons, 87
 contrasted with McNemar test, 201–202
 distribution with one degree of freedom, 83–84
 how to use, 87
 and odds ratio, 93–94
 power, 121
 relationship with z for comparing proportions, 82–83
 and relative risk, 91–92
 restrictions on use, 83–84
 sample size, 121
 table of critical values, 85
 to test for normal distribution, 206
 to test paired data on nominal scale, 201
 test statistic, 81–85
 ties, 219*n*
 used with Friedman's statistic, 223
 Yates continuity correction (*See* Yates correction, for continuity)
- Cigarette smoking:
 and platelet function, 187–189, 217
 secondhand smoke, 44–45
- Chocolate, 202
- Chronic obstructive pulmonary disease, 45, 71–72, 113–114
- Clinical trial:
 Vs. epidemiological studies, 91, 92
 negative conclusions, 133, 134
 outcome measures, 84
 prospective studies, 14, 91–93
 source of censored data, 230
 stratified sampling, 12
 between two nominal variables, 91
 (*See also* Randomized trial)
- Coefficient of determination, 167
- Cohort studies, 92*n*
- Collecting data:
 goals for, 7
 stratification, 12
 ways to, 13
- Comparative effectiveness research, 2
- Complication rates, 137–138
- Confidence, 129–130
- Confidence interval:
 assumption of normal distribution, 126
 definition, 125, 127
 dependence on experimental design, 250
 dependent on sample, 127, 127–129
 for difference of population means, 126, 126*n*, 132
 for difference of proportions, 132–133
 examples, 127–129, 133, 136, 154–155
 for intercept, 155
 for line of means, 155–156
 for mean, 126, 126*n*, 132
 meaning, 126, 129–130
 for an observation in regression, 156
 for odds ratio, 94*n*, 138–140
 for population, 132
 for power, 126*n*
 for proportion, 136–138
 for regression, 155–156
 for relative risk, 94*n*, 138–140
 for slope, 154
 for survival curve, 234–235
 to test hypothesis, 125, 130–132, 155–156
- Confounding variable:
 control for, 14*n*
 definition, 13
 example, 13–14,
 in observational studies, 13–14, 57
- Contingency table:
 chi-square versus Fisher exact test, 87–91
 and comparison of observed proportions, 86–87
 definition, 80
 degrees of freedom, 83–84
 examples, 81–87
 with more than two treatments or outcomes, 86–87
 and odds ratio, 93–94
 for paired data, 201
 power, 121
 and relative risk, 91–92
 restrictions on use of chi-square, 83–84
 sample size, 121
 subdividing, 87
 summary of procedure, 87
 when to use, 249, 250
 (*See also* Chi-square; Fisher exact test; McNemar test; Odds ratio; Relative risk)

Continuity correction:

- effect, 84
- for Mann-Whitney rank sum test, 211
- need for, 79
- for odds ratio, 93–95
- for relative risk, 91–92
- for 2×2 contingency table, 84
- for Wilcoxon signed rank test, 216
- for z statistic, 79
- (*See also* Yates correction, for continuity)

Control:

- control event rate, 92
- group, 40
- and odds, 93–94
- versus prospective study, 93–94
- and relative risk, 91

Control event rate, 92

Control group:

- and bias, 4, 4*n*, 12, 250, 252
- necessity, 189, 250

Coronary artery disease, 254–256

Correlation coefficient:

- and coefficient of determination, 167
- computation, 165–168
- general characteristics, 164–165
- nonparametric (*See* Spearman rank correlation coefficient)
- strength of association, 143
- (*See also* Pearson product-moment correlation coefficient; Spearman rank correlation coefficient)

Cost:

- of medical care (*See* Medical care, costs)
- of statistical errors, 5, 257–258

Critical values:

- of chi-square, table, 85
- computation, 67
- of F , table, 35–37
- of Friedman's statistic, table, 224
- of Mann-Whitney rank sum statistic, T , table, 210
- normal (one-tail), table, 108–109
- one versus two-tailed, 107
- of Spearman rank correlation coefficient, table, 171
- of t , (two-tailed), table, 57
- of t test (one-tail), table, 108–109
- of Wilcoxon signed-rank test statistic, W , table, 216

DData, collecting (*See* Collecting data)

Dartmouth Atlas of Health Care, 1

Decision making:

- Bayesian, 70
- clinical, 69–70
- meta-analysis, 134–136
- statistical, 69–70

Degrees of freedom:

- analysis of variance, 38, 192–193
- for contingency table, 83–84
- for linear regression, 154, 168
- for paired t test, 186
- partitioning, 193–194
- purpose, 38
- for repeated-measures analysis of variance, 196
- and sums of squares, 193
- for unpaired t test, 56
- and variance, 193

Dependent variable, 144

Depression, 95–96

Descriptive statistics, 24

Diabetes, 44, 124, 179–181, 183, 227

- and erectile dysfunction, 181–182
- neuropathy, 23–25, 71, 211–213

Dioxin, 97

Distribution:

- chi-square, 82–83
- F , 34–38, 38
- normal (*See* Normal distribution)
- parameters, 8–9
- of population, 8–9
- shape, 9
- skewed, 19–20, 21
- t , 55
- T , 205–206
- W , 215–217

Distribution-free method, 30*n*, 206(*See also* Nonparametric method)

DNA damage, 156–159, 172–173

Double-blind study:

- and bias, 13, 211, 189, 250
- definition, 211
- example, 211–213
- mechanism to carrying out, 211

Dunnett's test, 64*n*

- E**
- Echocardiography to assess mitral regurgitation, 175–177
- Effectiveness of medical procedures, 4–5
- Endothelial function, 181
- Endotoxins, 197–200, 203, 224–225
- Epidemiological study:
 versus clinical trials, 91, 92
 measures of association, 91
 versus prospective studies, 91, 92, 92*n*
- Erectile dysfunction, 181
- Errors:
 in medical journals, 2
 Type I (*See* Type I errors)
 Type II (*See* Type II errors)
 in use of statistics, 4–5
- Estimating proportion from sample, underlying assumptions, 78
- Estrogen, 46
- Ethical implications:
 and poorly designed studies, 258
 of randomization, 253
- Experimental study, versus observational study, 41
- Evaluation of therapy, 2, 253
- Evidence synthesis, 247–248
- Examples:
 adenosine triphosphate (ATP), 44
 adhesions following surgery, 226
 advanced directives for homeless people, 72, 73, 77–78, 79–80, 139
 AIDS, 25, 46, 227–228
 analgesia, 95
 angina pectoris, 251
 anti-asthmatic drugs and endotoxins, 197–200, 224–225
 antibiotic prescribing, 136
 antibiotics, 226–227
 articles with statistical errors, 4–5, 89–91,
 authorship, 96–97
 bias, 12–13
 boiling oil for gunshot wounds, 254
 bone marrow transplant for adult leukemia, 237–242, 254
 breast cancer, 94–95, 140, 200–202
 burnout, 45
 cannabis and diabetic neuropathy pain, 25, 211–213
 cell phones and sperm function, 42, 156–159, 172–173, 247–248
 chronic obstructive pulmonary disease, 45, 113–114
 cost of treating elderly, 244–245
 depression, 95–96
 diabetes, 44, 179–182, 183, 227
 diabetic neuropathy pain, 25, 211–213
 dioxin, 97
 effect of seeing smoking in movies on smokers' brains, 43
 endothelial function, 181
 endotoxins, 197–200, 203, 224–225
 erectile dysfunction, 181
 fallacious conclusions from autopsy data, 98*n*
 glucose levels, 227
 heart disease, 13–14, 134–136
 heart rate variability, 203
 HIV, 25, 46, 227–228
 homeless people, 72, 73, 77–78, 79–80
 hormone replacement therapy, 99, 179–182, 203
 hypothermia in low birth weight infants, 119–120, 133, 139
 insulin resistance in physical fitness, 183
 internal mammary artery ligation, 251
 journal size and selectivity, 168–169
 low birth weight infants, 119–120, 133, 139
 lung cancer, 13–14, 244
 marijuana, 219–222
 measuring heart size, 176–177
 medical investigators, 18
 menopause, 99
 mitral regurgitation assessed with echocardiography, 175–177
 mouth cancer, 181
 observational study, 38–41
 pet birds, 13–14
 population, 7–9
 portacaval shunt, 251–252
 of random samples, 10–12, 15–18
 relationship between weakness and muscle wasting in rheumatoid arthritis, 161–164
 secondhand smoke:
 and arterial function, 181
 and breast cancer, 94–95, 140
 and heart disease, 13–14, 134–136, 145*n*, 217
 and heart rate variability, 203
 and lung cancer, 13–14
 and lung function, 45
 and pet birds, 13–14

Examples (*continued*)

- smoking:
 - and platelet function, 187–189, 217
 - and renal cell cancer, 98
 - sperm function and cell phones, 116–117, 156–159, 172–173, 247–248
 - suicide in adolescents, 95
 - surgery to treat lung cancer, 244
 - testosterone, 203
 - vertebral fracture, 124
- (*See also specific statistical tests*)

Expected frequency in contingency table, 81–83

Experiment, 13

Experimental design:

- common errors, 5
- control group, 189

Experimental study:

- compared to observational study, 13, 14–15
- definition, 14
- example, 14–15
- role of pilot study, 122

F

F, 34–38

- distribution, 33, 34, 38
- examples, 38–44
- observational study, 38, 40, 41
- for one-way analysis of variance, 38
- power function, 277–285
- to reject null hypothesis, 27, 32–33, 34
- for repeated measures analysis of variance, 197
- table of critical values, 35–37
- in terms of mean of squares, 193
- in terms of sums of squares, degrees of freedom, 193
- to treat overall coincidence of two regression lines, 161
- variance ratio, 32–33

Factorial, 89, 261

False negative, 104

False positive, 104

Fisher exact test, 87–91, 260

Fisher and 5% *P* value, 70–71

5% *P* value, 70–71

Flavenols, 202

Friedman test:

- and chi-square for large samples, 223
- examples, 223, 224–225
- formula, 223, 261
- general approach, 222–224

- multiple comparisons, 225
- summary procedure, 223–224
- table of critical values, 224

G

Gaussian distribution (*See* Normal distribution)

Gehan's test:

- compared with log rank test, 242, 242*n*
- definition, 242
- when to use, 249
- Yates correction for, 242

Glucose levels, 227

Greenwood's formula for standard error of survival curve, 234

H

Hazard function, 237, 237*n*

Hazard ratio, 237

Heart disease:

- meta-analysis, 134–136
- and secondhand smoke, 13–14, 134–136

Heart rate variability, 203

Histogram, 23

HIV, 25, 46, 227–228

Hochberg's test, 64*n*

Holm *t* test:

- against a single-control, 67
- versus Bonferroni *t* test, 64–65
- definition, 64
- examples, 64–65
- versus Holm-Sidak test, 65–67
- multiple comparisons, 64–65
- power, 65
- procedure, 64
- rejective criterion, 64–65
- repeated-measures analysis of variance, 200

Holm-Sidak test:

- definition, 65
- compared to Bonferroni *t* test, 65–67
- compared to Holm *t* test, 65–67

Homeless people, 72, 73, 77–78, 79–80, 139

Hormones, 179–182, 203

Hospitalized patients, 256

Hypertension and insulin sensitivity, 183

Hypothermia and low birth weight infants, 71, 119–120, 133, 139

Hypothesis test:

definition, 3
 dependency on experimental design and measurement scale, 248
 examples, 3–4, 79–80
 identification in medical journals, 5, 256
 limitations, 257–258
 minimize number of tests, 256
 need to have hypothesis to test, 257–258
 one-tailed versus two-tailed, 55
 for Pearson product-moment correlation coefficient, 164–165
 for proportions, 78–79
 for regression, 154–155
 for Spearman rank correlation coefficient, 169–173
 test of significance, 4, 27
 using confidence interval, 125, 130–132, 154–155
 (See also Statistical significance; and specific statistical procedures)

I

Independent Bernoulli trial, 78
 Independent variable, 144
 Insulin sensitivity and hypertension, 183
 Intercept:
 comparing, 160–161
 of line of means, 145–147
 of regression line, 150
 test that notes zero, 154
 (See also Linear regression)
 Interval scale:
 characteristics, 164
 definition, 73
 and hypothesis testing procedure, 248
 Interpolation, 259
 Journal size and selectivity, 168–169

K

Kaplan-Meier product-limit estimate of survival curve, 233
 Kendall rank correlation coefficient, 169*n*
 Kruskal-Wallis test:
 and chi-square distribution, 218–219
 example, 219–222
 formula, 261
 mean rank, 219
 multiple comparisons, 219–222
 outline of procedure, 219
 ties, 219, 219*n*
 when to use, 249

L

Least squares analysis (See Linear regression)
 Leukemia, bone marrow transplant to treat, 237–242, 254
 Line of means:
 confidence interval, 155–156
 definition, 145–147
 intercept, 146–147
 residual variation, 145–147
 slope, 145
 Linear least squares analysis (See Linear regression)
 Linear regression:
 association versus causality, 145–146
 best straight line, 148–149
 comparison:
 of intercepts, 160
 of two regression lines, 160–164
 of two slopes, 160–161
 confidence interval, 155–156
 criteria for best fit, 148–149
 degrees of freedom, 154
 dependent variable, 144–147
 effect of interchanging dependent and independent variables, 164–165
 to estimate how much one variable changes with another variable, 143
 examples, 3, 149–150
 formulas, 149, 260
 hypothesis test, 152, 154
 independent variable, 144–147
 least squares, 149
 line of means, 145–147
 to make predictions, 144
 multiple, 177–178
 nonlinear relationships, 154*n*
 notation contrasted with Type I and II errors, 145*n*
 null hypothesis, 154
 overall test of coincidence, 161
 parametric procedure, 143
 population, 143–147
 regression line, 149
 slope and correlation coefficient, 164–165
 standard error:
 of the estimate, 151–152
 of the intercept, 151–154
 of the slope, 151–154
 variability about the line of means, 144–147
 when to use, 249
 Linear relationship (See Linear regression)

Low birth weight infants and hypothermia, 71, 82, 119–120

Log rank test:

- compared with Gehan's test, 242
- power of, 242–243
- and proportional hazards, 237, 242, 242*n*
- sample size for, 242–243
- when to use, 249
- Yates correction for, 242

Logarithmic transformation, 168

Lost to follow-up, 230

Lung cancer, treated with surgery, 244, 201–202
when to use, 249

M

Male fertility, 46

Mammary artery ligation to treat angina pectoris, 251

Mann-Whitney rank sum test:

- continuity correction, 211
- examples, 209–211
- for large samples, 209, 210–213
- logic, 207
- normal approximation, 210–211
- summary of procedure, 209
- T* test statistic, 207
- table of critical values, 210
- U* test statistic, 209*n*
- when to use, 249

Marijuana, 219–222 (*See also* Cannabis)

McNemar test:

- contrasted with chi-squared test for contingency table, 201–202
- example, 200–202
- formula, 260
- nonparametric method, 206, 206*n*
- for paired data measured on nominal scale, 200
- purpose, 185

summary of procedure Mean:

- line of (*See* Line of means)
- parameters, 9
- for percentiles of normal distribution, 22
- of population, 7, 9
- sample, 10, 15, 17–18

Mean square, 192

- in repeated-measures analysis of variance, 197

Measurements:

- bias, 12–13
- blinded, 12–13

Median:

- calculation of, 20–22
- definition, 20
- percentiles, 19–22
- of population, 7, 20–22
- of sample, 19–22

Median survival time, estimating, 234

Medical care:

- costs:
 - and biostatistics, 1–2
 - for elderly, 244–245
 - due to inaccurate medical literature, 5
 - inappropriate prescriptions, 136
 - magnitude, 1
 - and outcome, 84
 - tests and pharmaceuticals, use of therapies, 257
- hospitalized patients, 256
- role of clinicians, 2

Medical journals:

- accuracy, 5
- authorship, 96–97
- and bias, 5
- common statistical errors in articles, 122–124
- consequences of error, 5
- errors and inaccuracies, 4, 5
- how to improve, 257–258
- information should be provided about statistical methods, 257–258
- to keep informed, 4
- lack of hypothesis test procedures, 5
- letters to the editor, 258
- medical investigators, 18, 122–123
- quality of evidence, 136
- quality of statistical analysis, 2
- randomized clinical trials, 123–124
- reviews, 4–5
- selectivity, 168–169

Medical literature (*See* Medical journals)

Meta-analysis, 96–97, 133–136, 134*n*

Methods based on ranks:

- general approach, 205–206
- (*See also* Nonparametric method; *and specific statistical procedures*)

Mitral regurgitation, echocardiography for assessment of, 174–177

Mouth cancer, 181

- Movies, effect of smoking on brain function, 43
- Multiple comparison procedure:
 against a single control group, 67
 for analysis of variance, 62–67
 based on ranks, 219–222
 based on unpaired t test, 62–67
 Bonferroni t test, 62–63
 critical value computation, 64–65
 definition, 62
 examples, 63–64, 87–89
 Hochberg test, 64*n*
 Holm-Sidak procedure, 65–67
 Holm t test, 64–65
 incorrectly done with t test, 60–62
 power, 67
 rejective criterion, 65
 repeated analysis of accumulating data, 253, 253*n*
 for repeated measures analysis of variance, 200
 for repeated measures based on ranks, 225
 retrospective analysis of randomized control trials, 254–256
 for subdividing contingency tables, 87
 Tukey test, 64*n*
- Multiple regression, 177–178
- Muscle wasting in rheumatoid arthritis, 161–164
- N**
- N , 10
- $n!$, 89*n*, 261
- Natural logarithm, 138*n*
- Negative result:
 contrasting with proving no effect, 133
 interpretation, 123–124
 of randomized trials, 123
 (*See also* Power; Type II error)
- Newman-Keuls test (*See* Student-Newman-Keuls test)
- No significant difference, meaning, 101, 103
 (*See also* Negative result; Power)
- Nominal scale:
 definition, 73
 hypothesis tests, 248–250
- Nominal variables, 91–95
- Noncentrality parameter:
 for analysis of variance, 116
 for chi-square, 121
 for contingency table, 121
 definition, 112, 116
 for t test, 112, 114
- Nonlinearity relationship, difficulties for linear regression, 154, 178–179
- Nonparametric method:
 chi-square, 81–87, 206
 for contingency tables, 80
 contrast with parametric method, 206–207
 decision to use, 206–207
 Friedman test, 222–224
 general approach based on ranks, 207–211
 Kruskal-Wallis statistic, 218–219
 and McNemar's test, 206*n*
 Mann-Whitney rank sum test, 207–211
 methods based on ranks, 225–226
 multiple comparisons, 219–222
 versus parametric method, 29–30, 250
 power, 206
 Spearman rank correlation, 169–172, 206*n*
 Wilcoxon signed-rank test, 213–217
- Normal distribution:
 approximation, importance of, 10
 and central limit theorem, 18
 to compute confidence interval for proportions, 132–133
 and confidence intervals for the mean, 132
 critical values (one-tail), table, 108–109
 definition, 10
 described by mean and standard deviation, 10, 22–23, 17–18
 equation, 10
 to estimate proportions from samples, 75–76
 hypothesis testing procedures, 248–250
 for Mann-Whitney rank sum test, 209
 to obtain, 26
 and parametric method, 29–30, 205–206
 percentiles, 19–23
 population, 22–23
 and power, 118–119
 required for paired t test, 185–187
 and t distribution, 78–79, 133*n*
 table, 108–109
 test for normality, 10, 205–207
 transforming data, 25
 for Wilcoxon signed-rank test, 213–217

Normal probability graph paper, 206

Null hypothesis:

- for analysis of variance, 27–28
- for contingency table, 80–81
- definition, 27
- needs to be stated in articles, 258
- for odds ratio, 94
- for power, 118
- for regression, 154
- relationship to *P* value, 67–69
- for relative risk, 92, 94
- role in hypothesis testing, 103–104
- for *t* test, 52–53

O

Observational study:

- advantages, 41
- compared with experiment,
 - 13, 14, 41
- confounding variables, 13–14, 38
- definition, 13
- example, 14, 38–43
- limitations, 41

Odds ratio:

- and case-control study, 93–94
- chi-square, 94
- computation, 93, 94
- confidence intervals, 94*n*, 138–139
- example, 94–95, 140
- formula, 93
- interpretation, 91–92
- natural logarithm, 138
- null hypothesis, 94
- power for, 121
- relative risk, comparison with, 94*n*
- sample size for, 121
- standard error, 138–139

One-tail test, 107

One-tail value (*See* Normal distribution)

One-tail versus two-tailed tests, 55

One-way analysis of variance

(*See* Analysis of variance)

Ordinal scale:

- correlation, 164
- definition, 164
- example, 164
- hypothesis tests, 249, 249*n*

Outcome variable, 84

P

P value:

- and the Bayesian approach, 69–70
- definition, 67–69
- dependence on experimental design, 247, 250
- and ethics of randomized trials, 253
- from *F* distribution, 34
- 5% cutoff, origin of, 70–71
- highly prized, 27
- meaning, 60, 67–69, 70–71, 250, 253
- for nonparametric methods, 209
- unadjusted, 68
- (*See also* Power)

Paired observations (*See* Friedman statistics; Repeated-measures analysis of variance; *t* test, paired; Wilcoxon signed-rank test)

Paired *t* test (*See* *t* test, paired)

Patient Protection and Affordable Care Act, 1

Parameters:

- for linear regression, 144–147
- of population, 8

Parametric method:

- analysis of variance, 29–30
- contrast with nonparametric methods, 206–207
- decision to use, 205–206
- linear regression, 143
- Pearson product-moment correlation, 164–165
- requires normal distribution, 29–30

Partitioning sums of squares and degrees of freedom:

- for analysis of variance, 193–194, 195
- for repeated-measures analysis of variance, 194–197

Passive smoking and breast cancer, 94–95

PCBs, health effects, 25, 179

Pearson product-moment correlation coefficient:

- definition, 164
- formula, 260
- hypothesis test, 168
- no explicit dependent or independent variable,
 - 164–165
- and regression, 164, 165–168
- and regression slope, 167
- related to sum of squared deviations about regression
 - line, 165–168
- when to use, 249

Percentile:

- calculation of, 20–22
- definition, 20
- median, 20–22

- for population, 7, 20–22
- for sample, 20–23, 35
- skewed distribution, 21
- test for normal distribution, 22–23
- Pet birds, 13–14
- Pilot studies, 122
- Placebo effect:
 - and bias, 12
 - blinding protocol to minimize, 211
 - definition, 2, $2n$, 23
 - examples, $4n$, 91, 251
 - after internal mammary artery ligation surgery, 251
 - after portavacal shunt, 251–252
- Planning for experiments, power to estimate sample size, 112–113
- Polychlorinated biphenyls (PCBs), health effects, 25, 179
- Pooled variance estimate:
 - for confidence interval for difference of mean, 125–126
 - definition, 52
 - for proportion, 78–79
 - for regression lines, 160–161
- Poor supervision of statistical analysis, 257
- Population:
 - average squared deviation from the mean, 9
 - bias, 12–13
 - confidence interval, 140–142
 - difficulties in identifying, 256–257
 - distribution, 7–8, 9
 - examples, 3, 7–8, 9
 - limited sample, 7
 - for mean, 15, 132
 - meaning, 7
 - measure of dispersion about the mean, 9
 - median, 7, 20–23
 - normal distribution, 10, 20–21, 15
 - parameter, confidence interval for, 140
 - percentiles, 7, 19–23
 - of possible rankings, 207
 - power of test, 110–112
 - and proportion, 74
 - random (*See* Random sample)
 - range, 22–23
 - for regression, 144–147
 - sample, 10, 257–258 (described in journal article)
 - skewed, 19
 - standard deviation, 7, 9–10, 15
 - unobserved, 10
 - variability, 7–9
 - variance, 9–10
- Portacaval shunt:
 - bias in uncontrolled trials, 251–252
 - definition, 251
 - and placebo effect, 252
- Posterior probability, 70
- Power:
 - analysis of variance, 116
 - comparing proportions, 118–119
 - computation, 116
 - and confidence interval, $126n$
 - of contingency table, 121
 - for correlation, 173–174
 - definition, 104
 - to determine sample size, 112, 114, 115
 - examples, 116–117, 121–122
 - factors that determine, 105
 - and hypothesis testing, 115
 - of linear regression, 173–174
 - and log rank test, 242–243
 - multiple comparisons, 104
 - noncentrality parameter, 116
 - of nonparametric methods, 206
 - normal distribution, table, 108–109
 - not considered in published studies, 122–124
 - null hypothesis, 118
 - and odds ratio, 121
 - one-tail value, 108–109
 - and population variability, 110–112
 - power function of t test, 110–111, 112
 - practical problems in use, 122
 - purpose, 101
 - randomized trial, 123
 - and relative risk, 121
 - and repeated-measures analysis of variance, 200
 - role of pilot study, 122
 - and sample size, 101, 112, 118–119, 121
 - of t test, 105, 112–113
 - of tests for rates and proportions, 118–121
 - and treatment effects, 108–110
 - and Type I error, 104–106, 108
 - and Type II error, 105, 108
 - with unequal sample sizes, 112–113
 - and Yates correction, $119n$
- Prior probability, 70

Probability:

- Bayes' rule, 70, 70*n*
- posterior, 70
- prior, 70
- in a random sample, 10

Prediction:

- accuracy and confidence interval for an observation
 - in regression, 156
- with linear regression, 143

Process variable, 84**Proportion:**

- of population, 74
- and power, 118–119

Proportional hazards:

- assumption for log rank test, 237, 242
- computing power and sample size for log rank test, 242–243
- definition, 237

Prospective study:

- versus case-control studies, 93
- and chi-square, 92
- clinical trial, 91
- definition, 14, 91
- difficulties, 93
- versus epidemiological studies, 91–93, 93*n*
- example, 91–92
- and relative risk, 91–92

Q

Quality of evidence, 136

Quality of life, 84

R

r (*See* Correlation coefficient)

Random number generator, 10–11

Random numbers, table, 11

Random sample:

- bias, 12
- definition, 10
- example, 10–11
- frame, 11–12
- mean, 15
- population, characteristics of, 15
- probability of, 10
- procedure, 11–12
- selection (*See* Random number generator)
- simple, 11–12
- stratified, 12

Randomization:

- describe procedure in journal articles, 258
- ethical implications, 253
- to ensure correct conclusion, 253–254
- meaning, 250
- necessity, 248, 250
- procedure, 252
- to reduce bias, 252
- table of random numbers, 11

Randomized clinical trial (*See* Randomized trial)

Randomized trial:

- and bias, 252
- common patterns, 252
- definition, 14
- examples, 14–15, 79–80, 211, 250, 253–256
- method of choice to evaluate therapy, 14–15, 14*n*
- with negative results, 123–124
- versus nonrandomized trial, 14–15
- pilot studies, 122
- power of, 123–124
- practical aspects, 29, 250
- repeated analysis, 123
- sample size, 112
- (*See also* Randomization)

Rank order correlation coefficient (*See* Spearman rank correlation coefficient)

Ranks:

- to construct hypothesis test, 206, 207–209
- ranking procedure, summary of methods, 225–226
- (*See also* Nonparametric method)

Rates and proportions:

- approximate confidence interval for, 136
- confidence interval for difference, 132–133
- exact confidence interval, 137–138
- (*See also* Chi-square; Contingency table; Odds ratio; Relative risk)

Reactive oxygen species, 156–159, 172–173

Reader's Digest, 251

Regression (*See* Linear regression)

Regression lines, comparison of two, 160–164

Rejection criterion, 65, 224–225

Relative risk:

- chi-squared, 92
- computation, 91–92
- confidence intervals, 138
- control, 91–92
- example, 91–92, 139

formula, 91
 interpretation, 91–92
 null hypothesis, 92, 94
 odds ratio, $94n$
 power for, 121
 and prospective studies, 91–92
 sample size for, 121
 standard error, 138
 treatment, 91
 and 2×2 contingency table, 92
 Repeated-measures analysis of variance:
 analysis of variance, table, 196
 degrees of freedom, 195, 196–197
 examples, 197–200, 224–225
 formulas, 261
 general procedure, 195
 grand mean, 196
 and Holm t test, 200
 mean squares, 193, 196–197
 multiple comparisons, 200
 notation, 196
 power, 200
 purpose, 185, 194–195
 relationship to paired t test, 189
 sample size, 192
 between subject sum of squares, 194–197
 total sum of squares, 193, 196–197
 when to use, 249
 within subjects sums of squares, 193, 195, 197, 198
 (*See also* Friedman statistic; Wilcoxon signed rank test)
 Residual sum of squares of linear regression, 167
 Rheumatoid arthritis, 161–164

S

Sample:

and bias, 5, 12–13
 bias due to patient population, 247
 to compute confidence, 127, 127–129
 definition, 10
 to estimate proportion, 75–78
 limited, 7
 for linear regression, 148–150
 mean, 15
 and population, 10, 256–257, 258
 and power, 101, 110–111
 random (*See* Random sample)
 standard deviation, 15

Sample size:

for analysis of variance, 116
 comparing proportions, 118–119, 120
 computation, 120–121
 and confidence intervals, 137
 for contingency table, 121
 for correlation, 173–174
 to detect treatment effect, 101, 108–110
 formula, 120
 and log rank test, 242
 noncentrality parameter, 112
 for odds ratio, 121
 for regression, 173–174
 for relative risk, 121
 small, in most studies, 123
 and power, 101, 112, 118–119, 121
 for t test, 112

Sampling frame, 11

Scale of measurement:

interval, 73
 nominal, 73
 ordinal, 164, 205
 relationship to hypothesis testing procedure,
 248

Schizophrenia, 46

SD (*See* Standard deviation)

Secondhand smoke:

and arterial function, 181
 and breast cancer, 94–95, 140
 and confounding variables, 13–14
 and heart disease, 13–14, 134–136,
 $145n$, 217
 and heart rate variability, 203
 and lung cancer, 13–14
 and lung function, 44–45
 meta-analysis, 134–136
 and pet birds, 13–14

SEM (standard error of the mean) and central limit theorem, 17–18

Sequential analysis, $123n$

Single blinded study, 13, 189

(*See also* Double-blinded study)

Single factor analysis of variance (*See* Analysis of variance)

Size of treatment effect:

and power, 108–110, 112, 114
 worth detecting, 122

Skewed distribution, 19–20, 21

- Slope:
 comparison of two regression slopes with t statistic, 160–161
 hypothesis test, 154
 of line of means, 144–147
 of regression line, 149
- Sloppy thinking, 257–258
- Smoking in movies, and smokers' brains, 43
- SNK test (*See* Student-Newman-Keuls test)
- Spearman rank correlation coefficient:
 description, 169–170
 examples, 169–173
 formula, 170
 nonparametric method, 206*n*
 versus Pearson product-moment correlation coefficient, 169–172
 table of critical values, 171
 when to use, 249
- Sperm function and cell phones, 38–42, 56–59, 63–64, 116–117, 156–159, 172–173, 247–248
- Spinal fractures, 45–46
- Standard deviation (SD):
 and confidence interval for population from sample observation, 140–142
 contrasted with standard error of the mean, 18
 of a difference or a sum, 51–52
 estimated from sample, 10, 15
 population, 7, 9, 74–75
 in population with or without given attribute, 76–77
 about regression line, 150–151, 167
- Standard error:
 contrasted with standard deviation, 17–18, 142
 cost of, 4–5, 257–258
 definition, 15–16
 to describe variability in data, 18
 deviation of formula, 52*n*
 to estimate:
 definition, 152
 between groups variance, 32
 formula, 18
 of the intercept, definition, 152–153
 of the mean (SEM), and central limit theorem, 18
 of odds ratio, 138
 and Pearson product-moment correlation coefficient, 167
 and population range, 142
 of a proportion, 76–77
 of regression coefficients, 151–154
 of relative risk, 138
 of the slope, definition, 152–154
 of survival curve for Greenwood's formula, 234
 (*See also* standard error of specific statistic)
- Statistical significance:
 contrast with scientific or clinical significance, 125
 definition, 101
 dependence on sample size, 125–126
 distinction with proving no effect, 101–102
 and ethics, 258
 Fisher and, 70
 5% P value, origin of, 70–71
 lack of, 123, 253
 origin, 70–71
 (*See also* Null hypothesis; P value)
- Statistical tables (*See* Critical values)
- Statisticians, feisty, 257–258
- Step-down procedure, 64
- Stratification, 12 (*See also* Random sample)
- Strength of association (*See* Correlation coefficient)
- Student-Newman-Keuls (SNK) test, 64*n*
- Student's t test, 49–51
 (*See also* t test, paired; t test, unpaired)
- Study design, issues, 250
- Suicide, 95
- Sum of squares:
 to define F , 194
 and degrees of freedom, 192–193
 between groups, 192–193
 within groups, 192
 about regression line, 167
 total, 193–194
 for treatment, 190–191
 and variance, 190, 198–199
 (*See also* Analysis of variance; Repeated-measures analysis of variance)
- Surgery:
 adhesions, 226
 internal mammary artery ligation, 251
 for treatment of lung cancer, 244
- Survival curves:
 compared with Gehan's test, 242
 compared with log rank test, 237–242
 comparison of two, 235–237
 estimating, 230–234
 Kaplan-Meier product-limit estimate, 233
 median survival time, 231, 234
 proportional hazards, 237, 242
 standard error for, 234–235, 235*n*

Survival functions:
 definition, 231
 (*See also* Survival curves)

T

T (*See* Mann-Whitney rank sum test)

t distribution:

development, 51–56
 and normal distribution, 132–133
 one versus two-tails, 107

t statistic:

to compare two sample means (*See t* test, unpaired)
 general definition, 56, 186
 and intervals for difference of means, 125–126
 meaning, 55
 and normal distribution, 79, 132–133
 table of critical values, to test for changes (*See t* test, paired)

t test:

analysis of variance, 59–60
 assumptions, 205
 to compare regression intercepts, 161
 to compare regression lines, 160–161
 to compare regression slopes, 160–161
 effect of population variability, 101
 effect of sample size, 101
 Holm-Sidak test, 65–67
 misuse, 60–62, 122–124
 one-tail, table of critical values, 108–109
 one versus two-tails, 107–108
 paired:
 assumptions, 205
 common errors of use, 189
 definition, 185
 degrees of freedom, 187
 examples, 187–189
 purpose, 185, 189
 and repeated-measures analysis of variance, 189
 when to use, 249
 (*See also* Wilcoxon signed-rank test)
 power of function, 115
 size of treatment effect, 101, 108–110
 for slopes, 160–161
 unpaired:
 as analysis of variance, 59–60
 to analyze experiments for data collected before and after treatment in same subjects, 185
 assumptions, 49–50, 187, 205

definitions, 49, 185
 degrees of freedom, 56, 187
 differences within subjects, 187
 effect of the sample size, 49–50, 54–56
 examples, 56–59, 101–104
 formula, 52, 260
 general approach, 49–51
 misuse, 60–62
 most common procedure in medical literature, 49
 for multiple comparisons, 62–67
 nonparametric analog, 206
 null hypothesis, 62
 one-tail versus two-tailed, 55
 power, 105, 112–113
 samples drawn from different populations, 102–104
 summary of procedure, 178
 unequal sample sizes, 56–59
 when to use, 249
 (*See also* Mann-Whitney rank sum test)

Tables, statistical (*See Critical values*)

Test(s):

accuracy, 4
 approximation, 23
 of hypothesis (*See* Hypothesis test)
 for normality (*See* Normal distribution)
 purpose, 248–250

Test statistic:

chi-square, 81–84
 definition, 78
F, 32–33
 purpose, 32–33, 248–250
r, 171
t, 50, 57
 (*See also specific statistical tests*)

Testosterone, 203

Ties:

and chi-square, 219*n*
 and Friedman test, 223–224
 and Kruskal-Wallis statistic, 219, 219*n*
 and Mann-Whitney rank sum test, 209
 and Spearman rank correlation coefficient, 170
 and Wilcoxon signed-rank test, 216, 216*n*

Tolerance limit, 140*n*

(*See also* Confidence interval, for population)

Total sum of squares:

in analysis of variance, 193
 in linear regression, 167

Treatment sum of squares, 190–191, 192–193

Treatments:

- definition, 190
- in relative risk, 92

Trend (*See* Linear regression)

True negative, 105

True positive, 104–105

Two-tailed test:

- critical values, table, 57–58
- versus one tailed tests, 55, 107–108

Type I error:

- and confidence intervals, 127, 130–132
- definition, 104–105
- ethical implications, 253
- in medical literature, 122
- notation contrasted with regression, 145*n*
- and power, 105–107, 108
- and Type II error, 105, 108

Type II error:

- and confidence intervals, 127, 130
- definition, 105
- notation contrasted with regression, 145*n*
- and power, 105–107, 108
- and Type I error, 105, 108

U*U* (*See* Mann-Whitney rank sum test)

Uncontrolled trials, 250

V

Variable transformation, 168

Variability of population, 7–10

Variance:

- basis of all forms of analysis of variance, 189–190
- effect of power, 112–113
- estimated from sums of squares and degrees of freedom, 189–195

formula, 259

about line of means, 145–147

population, 9–10

about regression line, 150–151

Variance ratio (*See* *F*)**W**

Wilcoxon signed-rank test:

- continuity correction, 216
- general approach, 213–215
- normal approximation for large numbers, 215–26
- summary of procedure, 216–217
- table of critical values, 216
- ties, 216–217, 216*n*
- when to use, 249

Within groups sum of squares, 192

Within-groups variance, 31

Y

Yates correction:

- For chi-square, 84–86
- for continuity, 79, 84, 202
- for Gehan's test, 242
- for log rank test, 242
- and power, 119*n*
- (*See also* Continuity correction)

Z*z* (*See* Normal distribution)*z* test:

- to compare sample proportions, 78–79
- continuity correction for, 79
- examples, 80–81
- power, 118–119
- and two-tailed, critical values, 118–119