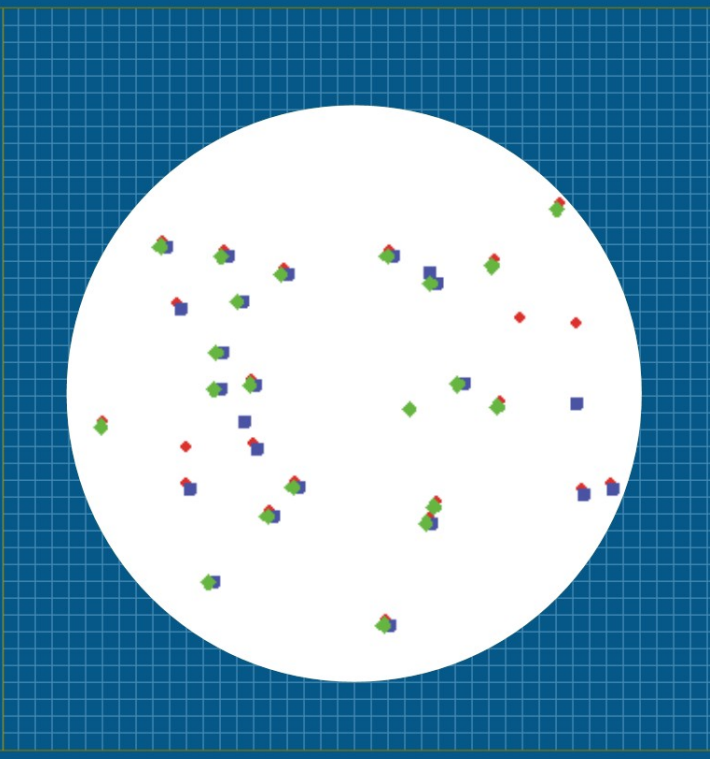
THIRD EDITION

RESAMPLING METHODS

A PRACTICAL GUIDE TO DATA ANALYSIS



PHILLIP I. GOOD

Birkhäuser

Phillip I. Good

Resampling Methods

A Practical Guide to Data Analysis

Third Edition

Birkhäuser
Boston • Basel • Berlin

Phillip I. Good
205 West Utica Avenue
Huntington Beach, CA 92648
U.S.A.

Cover design by Alex Gerasev.

AMS Subject Classifications: 05A05, 46N30, 62-01, 62-07, 62B15, 62E15, 62F40, 62G09, 62F03, 62H12, 62H15, 62H17, 65C05, 65C60, 93C57

Library of Congress Cataloging-in-Publication Data
Good, Phillip I.
Resampling methods : a practical guide to data analysis / Phillip I. Good.– 3rd ed.
p. cm.
Includes bibliographical references and indexes.
ISBN 0-8176-4386-9 (acid-free paper)
1. Resampling (Statistics) I. Title.

QA278.8.G66 2005
519.5'4–dc22 2005048191

ISBN-10 0-8176-4386-9 eISBN 0-8176-4444-X Printed on acid-free paper.
ISBN-13 978-0-8176-4386-7

Printed in the United States of America. (KeS/EB)

9 8 7 6 5 4 3 2 1 SPIN 11371397

www.birkhauser.com

Contents

Preface to the Third Edition

As with previous editions of *Resampling Methods*, this text is aimed at the practitioner or student, whether he or she is primarily interested in anthropological, archaeological, astronomical, biological, business, economic, medical, meteorological, psychological, or sociological applications.

Greater value is provided through four sets of changes with respect to the previous edition:

1. Procedures are now grouped by application; a prefatory chapter guides you to the appropriate reading matter.
2. Program listings and screen shots accompany each resampling procedure. Whether you program in C++, CART®, Box Sampler© (an Excel add-in), EViews, MatLab®, *R*, Resampling Stats®, SAS® macros, S-Plus®, or Stata®, you'll find the program listings and screen shots you need to put each resampling procedure into practice.
3. Notation has been simplified and, where possible, eliminated.
4. A glossary and answers to selected exercises are included.

To simplify putting these ideas to work, we've posted program code for you to download and apply at http://www.springeronline.com/0-8176-4386-9.

As the backgrounds of our readers vary, we've accompanied the first use of each statistical term with a brief definition. For those in need of a more thorough introduction to statistics, I recommend the purchase of *Introduction to Statistics via Resampling Methods and R/S-PLUS* (Wiley, 2005). For a quick brush-up, see *Common Errors in Statistics (and How to Avoid Them)*, (Wiley, 2003). A third companion text, *Permutation, Parametric, and Bootstrap Tests for Testing Hypotheses* (3rd ed., Springer, 2005) provides a rigorous mathematical foundation absent from the present text whose focus is on application, not theory.

My thanks to the many reviewers, including Dominic Lusinchi, Mikko Mönkkönen, Melanie Duffin, S. R. Millis, R. Goedman, Linus Svensson, and Douglas Matheson as well as the students in my on-line classes for statistics.com and my face-to-face classes at UCLA.

Huntington Beach, CA *Phillip I. Good*
July 2005

Preface to the Second Edition

Intended for class use or self-study, this text aspires to introduce statistical methodology to a wide audience, simply and intuitively, through resampling from the data at hand.

The resampling methods—permutations, cross-validation, and the bootstrap—are easy to learn and easy to apply. They require no mathematics beyond introductory high-school algebra, yet are applicable in an exceptionally broad range of subject areas.

Introduced in the 1930s, the numerous, albeit straightforward, calculations resampling methods require were beyond the capabilities of the primitive calculators then in use. And they were soon displaced by less powerful, less accurate approximations that made use of tables. Today, with a powerful computer on every desktop, resampling methods have resumed their dominant role and table lookup is an anachronism.

Physicians and physicians in training, nurses and nursing students, business persons, business majors, research workers and students in the biological and social sciences will find here a practical and easily-grasped guide to descriptive statistics, estimation, and testing hypotheses.

For advanced students in biology, dentistry, medicine, psychology, sociology, and public health, this text can provide a first course in statistics and quantitative reasoning.

For industrial statisticians, statistical consultants, and research workers, this text provides an introduction and day-to-day guide to the power, simplicity, and versatility of the bootstrap, cross-validation, and permutation tests.

For mathematics majors, this text will form the first course in statistics to be followed by a second course devoted to distribution theory.

Hopefully, all readers will find my objectives are the same as theirs: *To use quantitative methods to characterize, review, report on, test, estimate, and classify findings*.

If you're just starting to use statistics in your work, begin by reading Chapters 1 to 5 which cover descriptive statistics, sampling, hypothesis testing, and estimation, along with portions of Chapter 6 with its coverage of the essential, but challenging, ideas of significance level, sample size, and power, and Chapters 7 on contingency tables and/or 8 on experimental design, depending upon your interests. Recurrent themes—for example, the hospital data considered in the exercises for Chapters 1 and 5 to 8—tie the material together and provide a framework for self-study and classes.

For a one-quarter short course, I took the students through Chapter 1 (we looked at, but did not work through Section 1.2 on charts and graphs), Chapter 2, letting the

students come up with their own examples and illustrations, Chapter 3 on hypothesis testing, Chapter 4, and Sections 5.1, 5.2, and 5.5 on bootstrap estimation. One group wanted to talk about sample size (Sections 6.2–6.4), the next about contingency tables (Sections 7.1, 7.3, 7.5).

Research workers, familiar with the material in Chapters 1 and 2, should read Chapters 3, 5 and 6, and then any and all of Chapters 7 to 9 according to their needs. If you have data in hand, turn first to Chapter 10 whose expert system will guide you to the appropriate sections of the text.

A hundred or more exercises included at the end of each chapter plus dozens of thought-provoking questions will serve the needs of both classroom and self-study. C++ algorithms, Stata, and SC code and a guide to off-the-shelf resampling software are included as appendixes. The reader is invited to download a self-standing IBM-PC program from http://users.oco.net/drphilgood/resamp.htm that will perform most of the permutation tests and simple bootstraps described here. The software is self-guiding, so if you aren't sure what method to use, let the program focus and limit your selection. To spare you and your students the effort of retyping, I've included some of the larger data sets, notably the hospital data (Section 1.8) along with the package.

My thanks to Symantek, Design Science, Cytel Software, Stata, and Salford Systems without whose GrandView® outliner, Mathtype© equation generator, StatXact®, Stata®, and CART® statistics packages this text would not have been possible.

I am deeply indebted to my wife Dorothy, to Bill Sribney for his contributions, to Aric Agmon, Barbara Heller, Tim Hesterberg, David Howell, John Kimmel, Alson Look, Lloyd S. Nelson, Dan Nordlund, Richard Saba, Lazar Tenjovic, and Bill Teel for their comments and corrections, and to the many readers of the first edition who encouraged me to expand and revise this text.

Huntington Beach, CA *Phillip I. Good*
November 2001

Preface to the First Edition

Intended for class use or self-study, this text aspires to introduce statistical methodology—estimation, hypothesis testing, and classification—to a wide audience, simply and intuitively, through resampling from the data at hand.

The resampling methods—permutations, cross-validation, and the bootstrap—are easy to learn and easy to apply. They require no mathematics beyond introductory high-school algebra, yet are applicable in an exceptionally broad range of subject areas.

Introduced in the 1930s, their numerous, albeit straightforward and simple calculations were beyond the capabilities of the primitive calculators then in use; they were soon displaced by less powerful, less accurate approximations that made use of tables. Today, with a powerful computer on every desktop, resampling methods have resumed their dominant role and table lookup is an anachronism.

Physicians and physicians in training, nurses and nursing students, business persons, business majors, research workers and students in the biological and social sciences will find here a practical guide to descriptive statistics, classification, estimation, and testing hypotheses.

For advanced students in biology, dentistry, medicine, psychology, sociology, and public health, this text can provide a first course in statistics and quantitative reasoning.

For industrial statisticians, statistical consultants, and research workers, this text provides an introduction and day-to-day guide to the power, simplicity, and versatility of the bootstrap, cross-validation, and permutation tests.

For mathematics majors, this text will form the first course in statistics to be followed by a second course devoted to distribution theory.

Hopefully, all readers will find my objectives are the same as theirs:

To use quantitative methods to characterize, review, report on, test, estimate, and classify findings.

If you're just starting to use statistics in your work, begin by reading Chapters 1 to 5 which cover descriptive statistics, cause and effect, sampling, hypothesis testing, and estimation, portions of Chapter 6 with its coverage of the essential, but challenging ideas of significance level, sample size and power, Chapter 10 on classification, and Chapters 7 on contingency tables and/or 8 on experimental design, depending upon your interests. Recurrent themes—for example, the hospital data considered in the

exercises for Chapters 1–2 and 5–8—tie the material together and provide a framework for self-study and classes.

For a one-quarter short course, I took the students through Chapter 1 (we looked at, but did not work through Section 1.2 on charts and graphs), Chapter 2, letting the students come up with their own examples and illustrations, Chapter 3 on hypothesis testing (3.1–3.3 and 3.5–3.7), Chapter 4 (reviewing Section 2.4), and Sections 5.1, 5.2, and 5.5 on bootstrap estimation. One group wanted to talk about sample size (Sections 6.2–6.4), the next about contingency tables (Sections 7.1, 7.3, 7.5).

Research workers, familiar with the material in Chapters 1 and 2, should read Chapters 3, 5 and 6, and then any and all of Chapters 7–11 according to their needs. If you have data in hand, turn first to Chapter 12 whose expert system will guide you to the appropriate sections of the text.

A hundred or more exercises included at the end of each chapter plus dozens of thought-provoking questions will serve the needs of both classroom and self-study. C++ algorithms, Stata, SPlus, SC and SAS code and a guide to off-the-shelf resampling software are included as appendixes. The reader is invited to download a self-standing IBM-PC program that will perform most of the permutation tests and simple bootstraps described here. The software is self-guiding, so if you aren't sure what method to use, let the program focus and limit your selection. To obtain your copy of the software, follow the instructions on my home page http\\users.oco.net\drphilgood. To spare you and your students the effort of retyping, I've included some of the larger data sets, notably the hospital data (Section 1.8) and the birth data (Section 10.9) along with the package.

My thanks to Symantek, Cytel Software, Stata, and Salford Systems without whose GrandView® outliner, Mathtype© equation generator, StatXact®, Stata®, and CART® statistics packages this text would not have been possible.

I am deeply indebted to Bill Sribney and Jim Thompson for their contributions, to John Kimmel and Lloyd S. Nelson for reading and commenting on portions of the manuscript, and to the many readers of *Permutation Tests*, my first text, who encouraged me to reach out to a wider audience.

Huntington Beach, Fullerton, San Diego,
San Ramon, and Torrance, CA
March 1999

Phillip I. Good

Which Chapter Should I Read?

Haven't decided what software to use? Read Chapter 1.

Need interval estimates for the parameters of a population (mean, variance, percentiles)? Read Chapter 2.

Need to compare two populations? Two techniques? Before and after effects of treatment? Read Chapter 3.

Looking for guidelines on choosing the best statistical procedure? Or uncertain about the relation between tests and confidence intervals? Read Chapter 4.

Need to compare three or more populations? Three or more treatments, doses, or techniques? Read Chapter 5.

Need to analyze data that fall into categories? Or are recorded in the form of a contingency table? Or are the result of surveys? Read Chapter 6.

Need to compensate or control for multiple factors? Read Chapter 7.

Need to describe the relationships among two or more variables? To develop models for use in prediction or classification? Read Chapters 8 and 9.

1

Software for Resampling

In the absence of high-speed computers and the supporting software, the resampling methods, indeed almost all statistical procedures, would merely be interesting theoretical exercises. Though introduced in the 1930s, the numerous, albeit straightforward calculations required by the resampling methods—the bootstrap, cross-validation, and rearrangements—were beyond the capabilities of the primitive calculators then in use. They were soon displaced by less powerful, less accurate parametric approximations that made use of tables. Today, with a powerful computer on every desktop, resampling methods have resumed their dominant role and table lookup is an anachronism.

But not without the software! There are two approaches: One can program it oneself in a computer language such as C++, *R*, Resampling Stats, or SAS macros, or make use of a menu-driven program such as Excel, *S*-Plus, Stata, StatXact, or Testimate. One might well make use of both types of software: Menu-driven programs because of their convenience, and a programming language because sooner or later you'll encounter a new application beyond a menu-driven program's capabilities. I make use of all of these programs in my work as a statistical consultant.

1.1 Box Sampler©: An Excel Add-In

Box Sampler is an Excel add-in. To use, load Excel, then make use of the Box Sampler pull-down menu on Excel's menu bar. You can download this add-in without charge from http://www.introductorystatistics.com/escout/tools/boxsampler.htm. To assist you in using the program, you'll find full documentation at
http://www.introductorystatistics.com/escout/BSHelp/Main.htm.

To Excel's existing capabilities, Box Sampler adds the ability to simulate random sampling from various distributions, draw uncorrected bootstrap samples, or generate random rearrangements.

If your data is already in Excel format and you are familiar with Excel's capabilities, then Box Sampler might be the way to go. If you want more graphic and model building capabilities, then read on.

1.2 C++

C++ is a do-it-all-yourself programming language. In contrast to the other computer languages described here, C++ is a *compiler*. This means it is demanding of the programmer's time, but yields highly efficient programs. Not surprisingly, it is the language of choice for teams of programmers developing software for commercial distribution. Libraries of C++ routines are available that provide pull-down menus and a Windows-like interface.

1.3 CART®

A highly flexible, menu-driven program for creating decision trees for classification and regression. Download a trial version from www.salford-systems.com.

1.4 EViews

The makers of this program designed for the analysis of economic and other time series may not be able to tell a permutation test from a bootstrap, but if your objective is to test series components and residuals for independence, then this is the program to buy. Includes routines to @permute and @resample from matrices. Order a 30-day trial version from **http://www.eviews.com/.**

1.5 MatLab®

Can be used with some difficulty to program your own bootstrap and permutation routines. Thanks to Melanie Duffin for the code included in this text. For more information consult http://www.mathworks.com/products/index.html?ref=fp.

1.6 R

R is a do-it-yourself programming language specifically designed for use by statisticians. This means that functions like mean(), quantiles(), binom(), glm(), plot(), sample(), and tree() are precompiled and ready for immediate use. Moreover, it is easy to add your own statistical functions. You can download *R* for free from http://www.cran.r-project.org/ and obtain manuals and directions to self-help bulletin boards from the same Internet address. Many user-written precompiled libraries of resampling functions are available for download.

1.7 Resampling Stats®

Resampling Stats is a do-it-yourself programming language aimed at the beginner new to statistics and to programming languages. Make an error in coding and the source of the

error is immediately displayed. It has built-in functions for bootstrap, rearrangements, and generating random samples from various distributions. But its lack of graphics, regression capabilities, and most parametric procedures limits long-term use.

Resampling Stats is also available as an Excel add-in with somewhat greater ease of use particularly for matched pairs and stratified data.
http://resample.com/content/software/standalone/index.shtml

1.8 SAS®

SAS is a highly expensive menu-driven program best known for its ability to build customized tables and graphs. Widely available among corporate users, resampling methods can be added via a macro language. The cumbersome nature of this language, a giant step backward from the C++ in which SAS is written, makes it extremely difficult and time consuming to write and debug your own resampling methods. SAS Proc MulTTest is recommended for obtaining accurate significance levels for multiple tests.

1.9 S-PLUS®

S-Plus is a menu-driven variant of *R* aimed at corporate users. It's expensive but includes more built-in functions, the ability to design and analyze sequential trials, a huge library of resampling methods, plus a great deal of technical support. Check it out at http://www.insightful.com/contactus/request_cd.asp

1.10 Stata®

Stata is a comprehensive menu-driven program that allows you to generate bootstrap confidence intervals for the results of virtually any statistical procedure from population means to linear, logistic, and quantile regression coefficients. The latest version allows you to perform permutation tests for virtually all situations including, unfortunately, some tests, such as multifactor analysis and multiple regression, for which permutation methods are not recommended. http://stata.com

1.11 Statistical Calculator

Like *R*, the Statistical Calculator is an extensible statistical environment, supplied with over 1200 built-in (compiled C) and external (written in SC's C-like language) routines. It includes permutation-based methods for contingency tables (chi-square, likelihood, Kendall S, Theil U, kappa, tau, odds ratio), one- and two-sample inference for both means and variances, correlation, and multivariate analysis (MV runs test, Boyett/Schuster, Hotelling's T^2). Includes ready-made bootstrap routines for testing homoscedacity, detecting multimodality, plus general bootstrapping and jackknifing facilities.
http://www.mole-soft.demon.co.uk/

1.12 StatXact®

Menu-driven StatXact is the package to purchase for the analysis of contingency tables with categorical or ordered data. Includes power and sample size calculations. www.cytel.com.

1.13 Xlminer®

Excel-based program for developing classification and regression trees, and for market-basket analysis. Includes a variety of data-mining algorithms. Download a trial copy from www.resample.com/xlminer/.

1.14 Miscellaneous

The programs in this group are all extremely limited in scope. On the plus side, all include one or more resampling methods not present in most "full-function" software.

Blossom Statistical Analysis Package©

Blossom is an interactive program utilizing multiresponse permutation procedures (MRPP) for grouped data, agreement of model predictions, circular distributions, goodness of fit, least absolute deviation, and quantile regression. Online manual in HTML format contains many worked-through examples.
http://www.fort.usgs.gov/products/software/blossom/blossom.asp

Ctree©

A macro-filled Excel spreadsheet useful for a first look at decision trees. Download without charge from
http://www.geocities.com/adotsaha/CTree/CtreeinExcel.html

GoodStats©

GoodStats provides permutation tests for two-sample comparison, correlation, k-sample comparison with ordered or unordered populations, plus it is the only software available using synchronized permutations for analysis of 2xK designs. DOS only. Windows version in preparation. Download without charge from
http://mysite.verizon.net/res7sf1o/GoodStat.htm

NPC Test©

Menu-driven NPC Test is the only statistics program on the market today that provides for the Pesarin omnibus multivariate test. A demonstration version may be downloaded from http://www.methodologica.it

2

Estimating Population Parameters

In this chapter, after a brief review of estimation procedures, you'll learn to use the bootstrap to estimate the precision of an estimate and to obtain confidence intervals for population parameters and statistics. Computer code is provided to help you put these resampling methods into practice.

2.1 Population Parameters

Population parameters and statistics[1] include all of the following:

- *Percentiles* of the population's *frequency distribution.*
- Central values such as the *arithmetic mean*, the *geometric mean*, and the *median*.
- Measures of dispersion such as the *variance*, the *standard deviation*, and the *interquartile deviation*.

If you are already familiar with the terms in *italics*, you can skip ahead to Section 2.2. But for the benefit of those for whom it's been a long time since their last statistics class, we define these terms and comment on their applications in what follows.

2.1.1 Frequency Distribution and Percentiles

The *cumulative distribution function* $F[x]$ is the probability that an observation drawn at random from a population will be less than or equal to x. The kth percentile of a population is that value P_k such that $F[P_k] = k$ percent. There are many economic, health, and political applications where our primary interest is in these percentiles or *quantiles* as they are sometimes called.

[1] A statistic is any characteristic of a population; a parameter determines the values of all of a population's statistics. A single parameter determines all the statistics of a Poisson distribution; a normal distribution's characteristics are determined by two independent parameters.

2.1.2 Central Values

Some would say that the center of a distribution is its 50th percentile or *median*, that value P_{50} such that half the population has equal or smaller values and half the population has greater values. Others, particularly physicists, would say that it is the *arithmetic mean* $\overline{X}$ or balance point, such that the sum of the deviations about that value is 0. In symbols, we would write $\sum (X - \overline{X}) = 0$. A second useful property of the arithmetic mean is that the sum of the squared deviations about any value K, $\sum (X - K)^2$, is a minimum when we choose $K = \overline{X}$.

Which parameter ought you to use? We'd suggest the median and here are two reasons why:

1. When a cumulative distribution function is symmetric as in Figure 2.1, the arithmetic mean and the median have the same value. But even then, one or two outlying observations, typographical errors for example, would radically affect the value of the mean, while leaving the value of the median unaffected. For example, compare the means and medians of the following two sets of observations: 1, 2, 4, 5, 6, 7, 9 and 1, 2, 4, 5, 6, 7, 19.
2. When a cumulative distribution function is highly skewed as with incomes and house prices, the arithmetic mean can be completely misleading. A recent LA Times featured a great house in Beverly Park at $80 million U.S. A single luxury house like that has a large effect on the mean price of homes in its area. The median house price is far more representative than the mean, even in Beverly Hills.

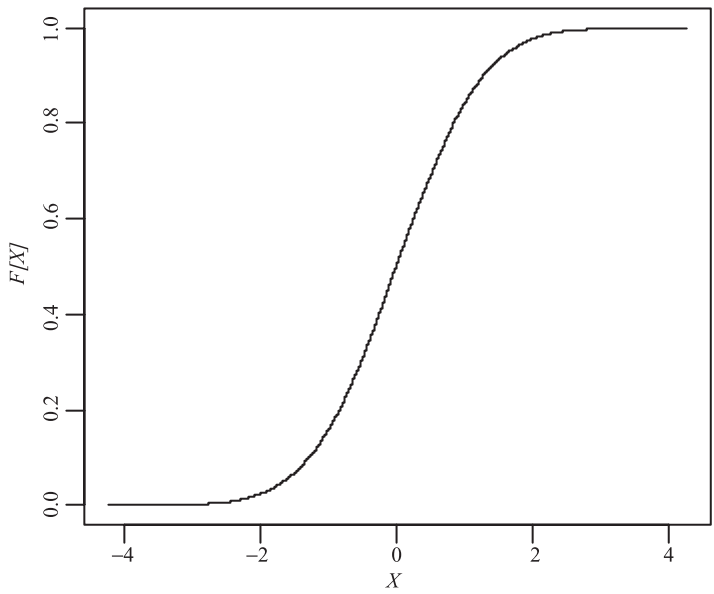


Fig. 2.1. Symmetric cumulative distribution function $F[X]$ of a normally distributed random variable X.

The *geometric mean* is the appropriate choice of central value when we are expressing changes in percentages, rather than absolute values. For example, if in successive months the cost of living was 110%, 105%, 110%, 115%, 118%, 120%, and 115% of the value in the base month, the geometric mean would be $(1.1*1.05*1.1*1.15*1.18*1.2*1.15)^{1/7}$.

Accuracy and Precision

Let us suppose Robin Hood and the Sheriff of Nottingham engage in an archery contest. Each is to launch three arrows at a target 50 meters (half a soccer pitch) away. The Sheriff launches first and his three arrows land one atop the other in a dazzling display of shooting *precision*. Unfortunately, all three arrows penetrate and fatally wound a cow grazing peacefully in the grass nearby. The Sheriff's *accuracy* leaves much to be desired.

2.1.3 Measures of Dispersion

The mean or median only tells part of the story. Measure my height using a tape measure three times and you'll probably come up with readings of 69″, 70″, and 71″. Measure my diastolic blood pressure three times and you might come up with readings as diverse as 80 mm, 85 mm, and 90 mm. Clearly, the latter set of readings is far more dispersed than the first.

One measure of dispersion is the population *variance*, the sum of the squared deviations about the population mean, $\sum (X - \overline{X})^2$. Unfortunately, the units in which the variance is defined are the squares of the units in which the original observations were recorded. In consequence, the *standard deviation* defined as the square root of the variance is more commonly used. The standard deviation has the same units as the original observations. The ratio of the standard deviation to the mean (known as *the coefficient of variation*) is dimensionless and independent of the units in which the original observations were recorded.

If the frequency distribution we are studying has the *normal distribution* depicted in Figure 2.1, then the probability that an observation will lie within one 1.64 standard deviations of the mean is 90%. If the distribution is not normal, nor has some other well-tabulated standard form, then knowledge of the standard deviation is far less informative. The *interquartile deviation* tells us the length of the interval between the 25th and the 75th percentile. In other cases, we might want to know the values of the 25th and 75th percentile or of the 10th and the 90th. Figure 2.2 depicts a box and whiskers plot of 22 observations in which we can see at a glance the values of the 25th and 75th percentiles (the box), the median (the bar inside the box), and the minimum and maximum (the ends of the whiskers).

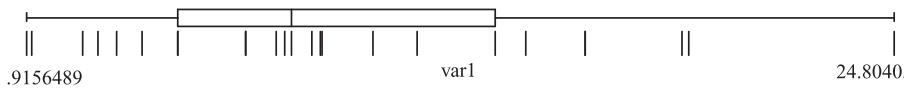


Fig. 2.2. Box plot superimposed on a strip chart of the actual observations.

2.2 Samples and Populations

By far the best way to establish the values of *population parameters*—the population mean, its dispersion, and its various percentiles is to examine each and every member of the population, make the appropriate measurement or measurements, and record their values. There are at least three reasons why this might not be possible:

1. The method of examination is destructive as is the case when we need to open a can to check for contamination or burst a condom to measure tensile strength.
2. The method of examination is prohibitively expensive or time-consuming or both.
3. The population is hypothetical in part; for example, when we test a new infant formula, we want to extrapolate our findings to all the children yet to be born.

Our alternative is to select a *sample* of the members of the population, and to use the sample *statistics* to estimate the population parameters. The recommended *point* or *single-value estimate* in most cases is the *plug-in* estimate: For the population mean, use the sample mean; for the population median, use the sample median; and to estimate a proportion in the population, use the proportion in the sample. These plug-in estimates have the virtue of being *consistent*, that is, as the sample size grows, the plug-in estimate becomes progressively closer to the population value. In many instances, particularly for measures of location, plug-in estimates are *unbiased*, that is, they are closer on the average to the population value than to any other value.

The exception that proves the rule is the population variance. The recommended estimate is $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ while the plug-in estimate is $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

The first of these formulas is recommended because if we were to take a large number of samples each of size n, the mean or *mathematical expectation* of the estimate would be the population variance. While both estimates are consistent, only the recommended estimate is unbiased.

But the sample is not the population. If we had taken a different sample, we might well have recorded quite different values. The best way to find out how different and thus how precise our estimates are is by taking repeated samples from the same population. But if we could have afforded this many samples, we ought to have taken a single very large sample in the first place, as estimates from larger samples are always more precise. Note: Unbiased estimates from larger samples are also more accurate.

2.2.1 The Bootstrap

One practical alternative, known as the *bootstrap*, is to treat the original sample of values as a stand-in for the population and to resample from it repeatedly, *with replacement*, computing the desired estimate each time.

Consider the following set of 22 observations on the heights of sixth-grade students, measured in centimeters and ordered from shortest to tallest. Note that the median of the sample, a plug-in estimate of the median of the population from which it is drawn, is 153.5 cm.

137.0	138.5	140.0	141.0	142.0	143.5	145.0	147.0	148.5	150.0	153.0
154.0	155.0	156.5	157.0	158.0	158.5	159.0	160.5	161.0	162.0	167.5

Suppose we record each student's height on an index card, 22 index cards in all. We put the cards in a big hat, shake them up, pull one out and make a note of the height recorded on it. We *return the card to the hat* and repeat the procedure for a total of 22 times till we have a second sample, the same size as the original. Note that we may draw a specific student's card several times as a result of using this method of *sampling with replacement*.

As an example, our first bootstrap sample, which I've arranged in increasing order of magnitude for ease in reading, might look like this:

138.5	138.5	140.0	141.0	141.0	143.5	145.0	147.0	148.5	150.0	153.0
154.0	155.0	156.5	157.0	158.5	159.0	159.0	159.0	160.5	161.0	162.0

Several of the values have been repeated as we are sampling with replacement. The minimum of this sample is 138.5 cm, higher than that of the original sample, the maximum at 162.0 cm is less, while the median remains unchanged at 153.5 cm.

137.0	138.5	138.5	141.0	141.0	142.0	143.5	145.0	145.0	147.0	148.5
148.5	150.0	150.0	153.0	155.0	158.0	158.5	160.5	160.5	161.0	167.5

In this second bootstrap sample, we again find repeated values; this time the minimum, maximum, and median are 137.0 cm, 167.5 cm, and 148.5 cm, respectively.

The bootstrap can be used to determine the precision of any estimator. For example, the variance of our sample of heights is 76.7 cm^2. The variances of 100 bootstrap samples drawn from our sample range between 47.4 cm^2 and 115.6 cm^2 with a mean of 71.4 cm^2. They provide a feel for what might have been had we sampled repeatedly from the original population. The resulting values from our 100 bootstrap samples are summarized in Figure 2.3.

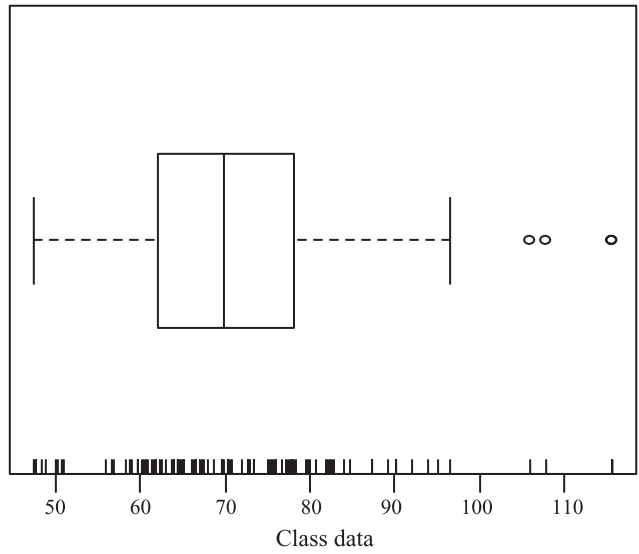


Fig. 2.3. Boxplot and strip chart of variances of 100 bootstrap samples.

2.2.2 Programming the Bootstrap

C++ Code

In the following listing, the line

```
for (j=0; j<n; j++) Y[j]=X[random(n)];
```

selects a random sample from X with replacement. Replace the call to compute_statistic (Y) with a call to the function that computes the statistic whose precision you wish to estimate.

```
#include <stdlib.h>
get_data();              //put the variable of interest
                              in the first n elements
                              of the array X[].
randomize();             //initializes random number
                              generator
for(i=0; i<100; i++){
    for(j=0; j<n; j++) Y[j]=X[random(n)];
    Z[i]=compute_statistic(Y);
   //compute the statistic for the array Y and store it
        in Z compute_stats(Z);
    }
```

EViews

```
create u 1 22
series x

x.fill 141,156.5,162,159,157,143.5,154,158,140,142,150,
     148.5,138.5,161,153,145,147,158.5,160.5,167.5,155,
     137

!rep=50 ' number of replications
vector(!rep) var_boot
For !i = 1 to !rep
         X.resample XB
         var_boot(!i) = @var(XB)
Next

scalar q5=@quantile(var_boot,.05) '5% quantile of the
     bootstrap
scalar q95=@quantile(var_boot,.95) '95% quanitle of the
     bootstrap
```

Gauss

(This routine only generates a single resample.)
$n = \text{rows(Y)}$
$U = \text{rnds}(n, 1, \text{integer seed}))$;
$I = \text{trunc}(n^*U + \text{ones}(n, 1, 1))$;
Ystar $= Y[I, .]$;

MatLab

```
%
%  Bootstrap Method
% This function takes N bootstrap samples with
% replacement of the original data and stores the
% calculation of the desired statistic in the returned
% array.
%
% bootstrap(data,N)
% data is the original array of data to be sampled.
% N is the number of bootstrap samples to be taken.
%

%to be ran before executing the function:
%clear all;
%data=[141,156.5,162,159,157,143.5,158,140,142,150,
148.5,138.5,161,153,145,147,158.5,160.5,167.5,155,137];
%to run: bootstrap(data,100)

function [stat]=bootstrap(data,N)
    rand('state', sum(100*clock));
    %reset random generator to different state at each
        time

    n=length(data);
    stat=zeros(1,N); %initialize array to be returned
        of bootstrap estimations
    sample=zeros(1,n); %initialize array to hold
        bootstrap sample

    for i=1:N
        choose=round(((n-1)*rand(1,n))+1);
        %choose an array length n of random #'s between
            [1,n]
        %the sampled data will come from the values at
            these random indices.
```

```
        %fill the sample array with values at
             randomized indices
        for j=1:n
            sample(j)=data(choose(j));
        end;
        stat(i)=mean(sample); %fill stat array with
             bootstrap estimations
    end;
```

R

```
#This R program selects 50 bootstrap samples from the
     classroom data
#and then produces a boxplot and stripchart of their
     variances.
class=c(141,156.5,162,159,157,143.5,154,158,140,142,150,
     148.5,138.5,161,153,145,147,158.5,160.5,167.5,155,
     137)
#record group size
n = length(class)
#set number of bootstrap samples
N =50
stat = numeric(N)  #create a vector in which to store
                        the results
                   #the elements of the vector will be
                        numbered from 1 to N
#Set up a loop to generate a series of bootstrap
     samples
for (i in 1:N){
     #bootstrap sample counterparts to observed samples
          are denoted with "B"
   classB= sample (class, n, replace=T)
   stat[i] = var(classB)
   }
boxplot (stat)
stripchart(stat)
```

Resampling Stats

```
'The following Resampling Stats program yields a
     histogram of median values derived from 100
'bootstrap samples of the classroom data.
DATA (141 156.5 162 159 157 143.5 154 158 140 142 150
     148.5 138.5 161 153 145 147 158.5 160.5 167.5 155
     137) A
```

```
MEDIAN A med_orig
REPEAT 100
  SAMPLE 21 A B
  MEDIAN B med
  SCORE med scrboard
END
HISTOGRAM scrboard
```

SAS Code

```
%macro boot(    /* Bootstrap resampling analysis */
 data=,         /* Input data set, not a view or a
                   tape file. */
 samples=200,   /* Number of resamples to generate. */
 residual=,     /* Name of variable in the input data
                   set that contains residuals; may
                   not be used with SIZE= */
 equation=,     /* Equation (in the form of an
                   assignment statement) for computing
                   the response variable */
 size=,         /* Size of each resample; default is
                   size of the input data set.
                   The SIZE= argument may not be
                   used with BALANCED=1 or with a
                   nonblank value for RESIDUAL= */
 balanced=,     /* 1 for balanced resampling; 0 for
                   uniform resampling. By default,
                   balanced resampling is used unless
                   the SIZE= argument is specified,
                   in which case uniform resampling
                   is used. */
 random=0,      /* Seed for pseudorandom numbers. */
 stat=_numeric_,/* Numeric variables in the OUT= data
                   set created by the %ANALYZE macro
                   that contain the values of
                   statistics for which you want to
                   compute bootstrap distributions. */
 id=,           /* One or more numeric or character
                   variables that uniquely identify
                   the observations of the OUT=data
                   set within each BY group. No ID
                   variables are needed if the
                   OUT= data set has only one
                   observation per BY group.
                   The ID variables may not be
                   named _TYPE_, _NAME_, or _STAT_ */
```

```
biascorr=1,     /* 1 for bias correction;
                   0 otherwise */
alpha=.05,      /* significance (i.e., one minus
                   confidence) level for confidence
                   intervals; blank to suppress normal
                   confidence intervals */
print=1,        /* 1 to print the bootstrap estimates;
                   0 otherwise. */
chart=1         /* 1 to chart the bootstrap
                   resampling distributions;
                   0 otherwise. */
);

%if %bquote(&data)= %then %do;
 %put ERROR in BOOT: The DATA= argument must be
      specified.;
 %goto exit;
%end;

%global _bootdat; %let _bootdat=&data;

%local by useby;
%let useby=0;

%global usevardf vardef;
%let usevardf=0;

*** compute the actual values of the statistics;
%let vardef=DF;
%let by=;
%analyze(data=&data,out=_ACTUAL_);
%if &syserr>4 %then %goto exit;

*** compute plug-in estimates;
%if &usevardf %then %do;
 %let vardef=N;
 %analyze(data=&data,out=_PLUGIN_);
 %let vardef=DF;
 %if &syserr>4 %then %goto exit;
%end;

%if &useby=0 %then %let balanced=0;

%if %bquote(&size)^= %then %do;
 %if %bquote(&balanced)= %then %let balanced=0;
 %else %if &balanced %then %do;
```

```
  %put %cmpres(ERROR in BOOT: The SIZE= argument
            may not be used with BALANCED=1.);
  %goto exit;
 %end;
 %if %bquote(&residual)^= %then %do;
  %put %cmpres(ERROR in BOOT: The SIZE= argument
            may not be used with RESIDUAL=.);
  %goto exit;
 %end;
%end;
%else %if %bquote(&balanced)= %then %let balanced=1;

*** find number of observations in the input data set;
%global _nobs;
data _null_;
 call symput('_nobs',trim(left(put(_nobs,12.))));
 if 0 then set &data nobs=_nobs;
 stop;
run;
%if &syserr>4 %then %goto exit;

%if &balanced %then
 %bootbal(data=&data,samples=&samples,
      random=&random,print=0);

%else %if &useby %then
 %bootby(data=&data,samples=&samples,
    random=&random,size=&size,print=0);

%if &syserr>4 %then %goto exit;

%if &balanced | &useby %then %do;
 %let by=_sample_;
 %analyze(data=BOOTDATA,out=BOOTDIST);
%end;

%else
 %bootslow(data=&data,samples=&samples,
      random=&random,size=&size);

%if &syserr>4 %then %goto exit;

%if &chart %then %do;
 %if %bquote(&id)^= %then %do;
  proc sort data=BOOTDIST; by &id; run;
  proc chart data=BOOTDIST(drop=_sample_);
```

```
      vbar &stat;
      by &id;
    run;
  %end;
  %else %do;
   proc chart data=BOOTDIST(drop=_sample_);
     vbar &stat;
    run;
   %end;
  %end;

%bootse(stat=&stat,id=&id,alpha=&alpha,
     biascorr=&biascorr,print=&print)

%exit:;

%mend boot;
```

S-Plus

Download the S+Resample package.
boot = bootstrap(urdata, median)
boxplot(boot)
plot(boot, 0*boot)

Stata

Once the height data is entered, the following line of code invokes the bootstrap to produce a 95% confidence interval for the interquartile deviation, plus a point estimate of its bias.

bstrap "summarize height,detail" (r(p75)-r(p25)), reps(100) nobc nonormal

2.2.3 Estimating Bias

When an estimate is inaccurate or *biased*, we would like an estimate of its *bias*, that is, the amount by which its expected value differs from the quantity to be estimated. The bootstrap can also help us here. Recall that while the variance of our sample of sixth-graders' heights is 76.7 cm^2, the mean of the variances of 100 bootstrap samples drawn from is 71.4 cm^2. Thus our original estimate of the population variance would appear to be biased upward by 5.3 cm^2.

More generally, let $E(X)$ denote the expected or mean value of a random variable X. An estimate $\theta[X]$ based on a sample is also a random variable; we define the bias of $\theta[X]$ as $b = E(\theta[X] - \theta)$ where θ is the population parameter we are trying to estimate. A bootstrap estimate for the bias b of $\theta[X]$ is given by $b^* = \Sigma_i(\theta_i^* - \theta[X])/k$ where θ_i^* is the ith bootstrap sample estimate of θ for $1 \leq i \leq k$.

2.2.4 An Example

The intent of a small-scale clinical trial was to show the FDA that a product produced at a new plant was "equivalent" to the product produced at the old plant. In this crossover trial, each of eight patients received in random order each of the following:

- Patch containing hormone that was manufactured at the old site
- Patch containing hormone that was manufactured at the new site
- Patch without hormone (placebo) that was manufactured at the new site.

To establish equivalence, the FDA requires that $|\theta/\mu| \leq 0.20$ where $\theta = E(\text{new}) - E(\text{old})$ and $\mu = E(\text{old}) - E(\text{placebo})$.

The natural estimate for θ is the average of the old-patch hormone level in patients minus the average of the new-patch hormone level in patients. Similarly, the natural estimate of μ is the average of the new-patch hormone level in patients minus the average of the placebo hormone level in patients. The plug-in estimate, which is the ratio of these two estimates, is the natural estimate of the ratio.

Unfortunately, such an estimate is biased, both because it is a ratio and because the same factor $E(\text{old})$ appears in both the numerator and the denominator. How large is the bias in the present case?

Table 2.1. Patch Data Summary

Subject	Old-Placebo	New-Old
1	8406	−1200
2	2342	2601
3	8187	−2705
4	8459	1982
5	4795	−1290
6	3516	351
7	4796	−638
8	10238	−2719
average	6342	−452.3

The plug-in estimate for θ/μ for the data in Table 2.1 is -0.07 which is considerably less in absolute value than the FDA's criteria of 0.20. But this estimate has a potentially large bias. Efron and Tibshirani [1993; chapter 10] generated 400 bootstrap samples and found the bootstrap estimate of bias to be only 0.0043.

Applying this bias adjustment, we still get an estimate that is considerably less than 0.20 in absolute value. Regardless, the bootstrap is notoriously unreliable for small samples and we would be ill advised to draw conclusions without additional data.

2.3 Confidence Intervals

The problem with *point estimates* as our study of precision reveals is that we will always be in error (unless we can sample the entire population), albeit by some vanishingly

small amount as sample size increases. The solution is an *interval estimate* or *confidence interval* where we can have confidence that the true value of the population *functional*[2] we are attempting to estimate lies between some minimum and some maximum value with a pre-specified probability.

For example, to obtain a 90% confidence interval for the variance of sixth-graders' heights, we might exclude 5% of the bootstrap values from each end of Figure 2.3. The result is a confidence interval whose lower bound is 52 cm^2 and whose upper bound is 95 cm^2. Note that our original point estimate of 76.7 cm^2 is neither more nor less likely than any other value in the interval [52,95].

2.3.1 Limitations of the Percentile Bootstrap Confidence Interval

Almost immediately, two questions arise about the confidence interval we just derived:

1. Is it accurate? That is, is it correct to say that 90% of the time we expect this interval to cover the true value of the population variance? Or is it really a smaller percentage?
2. Is it efficient? That is, is the interval no wider than it should be, so that the probability it includes false values of the population variance is as small as possible?

Unfortunately, the answer for the percentile bootstrap confidence interval in both cases is "No." The balance of this chapter is devoted to providing improved interval estimates.

2.3.2 The Bias-Corrected Bootstrap Confidence Interval

The bias-corrected BC interval due to Efron and Tibshirani [1986] represents a substantial improvement over the percentile bootstrap. The idea behind these intervals comes from the observation that percentile bootstrap intervals are most accurate when the estimate is symmetrically distributed about the true value of the parameter and the tails of the estimate's distribution drop off rapidly to zero. In other words, when the estimate has an almost-normal distribution.

Suppose θ is the parameter we are trying to estimate, $\hat{\theta}$ is the estimate, and we are able to come up with a monotone increasing transformation t such that $t(\hat{\theta})$ is normally distributed about $t(\theta)$. We could use this normal distribution to obtain an unbiased confidence interval, and then apply a back-transformation to obtain an almost-unbiased confidence interval.

The method is not as complicated as it reads because we don't actually have to go through all these steps, merely agree that we could if we needed to. The resulting formula *is* complicated.[3] But it is already incorporated in several computer programs as described in the next section. The form in which it is incorporated is a further refinement due to Efron [1987] known as the *bias-corrected-and-accelerated* or BC_α *bootstrap*,

[2] See the Glossary for a definition of this and all other italicized terms whose definitions are not provided in the surrounding text.

[3] See Chapter 14 of Efron and Tibshirani [1993].

a form that offers the further advantage that the original sample need not be as large to obtain the same degree of accuracy.

Even with these modifications, we do not recommend the use of the bootstrap with samples of fewer than 100 observations. Simulation studies suggest that with small sample sizes, the coverage is far from exact and the endpoints of the intervals vary widely from one set of bootstrap samples to the next. For example, Tu and Zhang [1992] report that with samples of size 50 taken from a normal distribution, the actual coverage of an interval estimate rated at 90% using the BC_α bootstrap is 88%. When the samples are taken from a mixture of two normal distributions (a not uncommon situation with real-life data sets) the actual coverage is 86%. With samples of only 20 in number, the actual coverage is 80%.

More serious when trying to apply the bootstrap is that the endpoints of the resulting interval estimates may vary widely from one set of bootstrap samples to the next. For example, when Tu and Zhang drew samples of size 50 from a mixture of normal distributions, the average of the left limit of 1000 bootstrap samples taken from each of 1000 simulated data sets was 0.72 with a standard deviation of 0.16, the average and standard deviation of the right limit were 1.37 and 0.30 respectively.

2.3.3 Computer Code: The BC_α Bootstrap

R

Make sure you are connected to the Internet and then type

- `install.packages("boot")`

The installation which includes downloading, unzipping, and integrating the new routines is done automatically. The installation needs to be done once and once only. But each time before you can use any of the boot library routines, you'll need to load the supporting functions into computer memory by typing

- `library(boot)`

We'll need to employ two functions from the boot library.

The first of these functions **boot (Data, Rfunction, number)** has three principal arguments. **Data** is the name of the data set you want to analyze, **number** is the number of bootstrap samples you wish to draw, and **Rfunction** is the name of an *R* function you must construct separately to generate and hold the values of existing *R* statistics functions such as **median** or **var** whose value you want a bootstrap interval estimate of. For example,

```
•     f.median<- function(y,id) {
+     median( y[id])
+     }
```

where *R* knows id will be a vector of form 1:n. Then

```
•   boot.ci(boot(classdata, f.median, 400), conf = 0.90)
```

will calculate a 90% confidence interval for the median of the classdata based on 400 simulations.

Note: The first argument in boot() can be a vector or a data frame. The latter would be employed, for example, if one wanted a BCa confidence interval for the difference in means of two samples.

SAS

A macro may be downloaded from
http://ftp.sas.com/techsup/download/stat/jackboot.html

S-Plus

boot = bootstrap(data, median)
limits.bca(boot)

Stata

Once the height data is entered, the following line of code invokes the bootstrap to produce a 95% confidence interval for the interquartile deviation, plus a point estimate of its bias.

bstrap "summarize height, detail" (r(p75)-r(p25)), reps(100) nonormal nopercentile

2.3.4 The Bootstrap-t

Suppose the hypothesis we wish to test is that $\theta = \theta_0$, and that $\hat{\theta}$ is our estimator. The accuracy of bootstrap confidence intervals for testing purposes can be improved by using the distribution of the differences $\hat{\theta}^* - \hat{\theta}$ rather than the distribution of the differences $\hat{\theta}^* - \theta_0$ according to Hall and Wilson [1991].

Now suppose $\hat{\sigma}$ is an estimate of the standard error of $\hat{\theta}$ and $\hat{\sigma}^*$ the estimate of the standard error of $\hat{\theta}^*$ computed for the bootstrap sample. To reduce the width of our confidence intervals (thus decreasing the probability of a Type II error), Hall and Wilson [1991] propose we scale each of the differences by $\hat{\sigma}^*$. Instead of basing our confidence interval estimates on the distribution of the differences $\hat{\theta}^* - \hat{\theta}$, we base them on the distribution of the Studentized differences $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$. The resulting confidence intervals are known as bootstrap-t confidence intervals.

The problem with this approach is that unless θ is a linear function of population means, there are few if any explicit formulas for estimating $\hat{\sigma}$. The solution is to take a set of second level bootstrap samples from each bootstrap sample and use these to estimate $\hat{\sigma}^*$, just as we use the variability in the first level $\hat{\theta}^*$ to estimate $\hat{\sigma}$.

Taking 500 bootstrap samples as well as 50 bootstrap samples from each to estimate its standard error would require 25,500 bootstrap samples in total. Since the resulting burden falls upon the computer, this is not really a problem.

Tibshirani [1988] found that both the accuracy and the precision of bootstrap-t confidence intervals could be improved by use of a variance stabilizing transformation. The problem is to determine what the correct transformation would be.

You may recall from an earlier statistics course being told to use the arc sine transformation for binomial variables, the square root for Poisson observations, and the log transformation when looking for percentage changes. The theoretical finding on which all these transformations are based is that if X is a random variable with mean θ and standard deviation $s(\theta)$, the desired transformation function $g(x)$ will have as its derivate $1/s(x)$.

Suppose we use the bootstrap to estimate g. The result is a drastic savings in computation time, for now that the variance is stabilized, we need only estimate s or, equivalently, $\hat{\sigma}$ once. Taking 50 bootstrap samples from the original sample to estimate the transformation function g and 50 more to estimate the standard error, plus 500 bootstrap samples, would require only 3,000 bootstrap samples in total. Of course, we need add in the time required to perform the transformation and back transformation.

If you are familiar with *R*, then you may calculate the variance-stabilized bootstrap-t with the following code:

```
class=c(141,156.5,162,159,157,143.5,154,158,140,142,
     150,148.5,138.5,161,153,145,147,158.5,160.5,167.5,
     155,137)
# download the bootstrap package from ttp:
     //rweb.stat.umn.edu/R/library/bootstrap/R/
#install manually, for example, source
     ("/temp/Documents/RData/bootstrap")
     library(bootstrap)
# find an 80% CI for the population variance
results=boott(class, var, VS=TRUE, v.nbootg=50,
     v.nbootsd=50, v.nboott=200, perc=c(.10,.90))
results[1]
$confpoints
     0.1    0.9
[1,] 63.39514 111.1910
```

2.3.5 Parametric Bootstrap

When we know the form of the population distribution, the use of the *parametric bootstrap* to obtain interval estimates may prove advantageous either because the parametric bootstrap provides more accurate answers than textbook formulas or because no textbook formulas exist. The parametric bootstrap is particularly valuable when the form of the distribution is known and confidence intervals are required for statistics such as the variance or the 90th percentile that depend heavily on values in the tails.

Suppose we know the observations come from a normal distribution and want an interval estimate for the 95th percentile. We would draw repeated bootstrap samples from the original sample, use the mean and variance of the bootstrap sample to estimate the parameters of a normal distribution, and then derive a bootstrap estimate of the 95th percentile from tables of an $N(0, 1)$ distribution.

MatLab

```
%
% Parametric Bootstrap
% This function calculates desired estimations
% using the parametric bootstrap method.
% This problem uses the exponential distribution to
% calculate an array of IQR''s of the bootstrapped data.
%

% To be ran before the function:
% clear all;
%
class=[141,156.5,162,159,157,143.5,154,158,140,142,150,
     148.5,138.5,161,153,145,147,158.5,160.5,167.5,155,
     137];
% To run function:
% parametric(class,100)

function [stat]=parametric(data,N)
    rand('state', sum(100*clock)); %reset random
                                        generator to a
                                        different state
    n=length(data);
    stat=zeros(1,N);
    samp=zeros(1,n);

    for i=1:N
        choose=round(((n-1)*rand(1,n))+1); %randomize
                                            indices for
                                            bootstrap
                                            sample
        for j=1:n
            samp(j)=data(choose(j)); %set bootstrap
                                      sample from
                                      choose
                                      indices
        end
        lambda=1/mean(samp);
        stat(i)=log(1/3)/(-lambda); %IQR for an
                                    exponential
                                    function
    end
```

R

```
#The following R program fits an exponential
```

```
     distribution to the data set A
#Then uses a parametric bootstrap to get a 90%
     confidence interval for the IQR of the population
     from which the data set A was taken.
#n=length(A)
#create a vector in which to store the IQR's
IQR = numeric(1000)
#Set up a loop to generate the 1000 IQR's
for (i in 1:1000) {
    bA=sample (A, n, replace=T)
    IQR[i]  = qexp(.75,1/mean(bA))-qexp(.25, 1/mean(bA))
    }
quantile (IQR , probs = c(.05,.95))
```

Resampling Stats

```
'The following program fits an exponential distribution
     to the data set A
'Then uses a parametric bootstrap to get a 90%
     confidence interval for the IQR of the 'population
     from which the data set A was taken.
MAXSIZE board 1000
MEAN A x_bar
REPEAT 1000
   EXPONENTIAL 100 x_bar X
 PERCENTILE X (25 75) P
 MAX P max
 MIN P min
  SUBTRACT max min IQR
  SCORE IQR board
END
BOXPLOT board
PERCENTILE board (5 95) k
PRINT x_bar
PRINT k
```

Stata

Once the height data is entered, the following line of code invokes the parametric bootstrap with a normal distribution to produce a 95% confidence interval for the interquartile deviation.

bstrap "summarize height, detail" $(r(p75) - r(p25))$, reps(100) nobc nopercentile

2.3.6 Smoothing the Bootstrap

If you haven't kept up your math, this section can be challenging. Two alternatives suggest themselves: 1) read through the next few sections quickly to get a feel for the

type of problems that can be solved and come back to them if and when you need them; or 2) work through the algebra step by step substituting real data for symbols.

An inherent drawback of the bootstrap, particularly with small samples, lies in the discontinuous nature of the *empirical distribution function*.[4] Presumably, our samples come from a continuous or near-continuous distribution. Figure 2.4 illustrates the distinction. The jagged curve is the empirical distribution function of the following set of 12 examination scores: 81, 81, 81, 68, 63, 73, 68, 56, 70, 45, 54, 44. The smooth curve that passes through it was obtained by replacing the original observations with their Studentized[5] equivalents, $z_i = (x_i - \overline{x})/s$, then plotting the distribution function of a normally distributed random variable $N(\overline{x}, s^2)$.

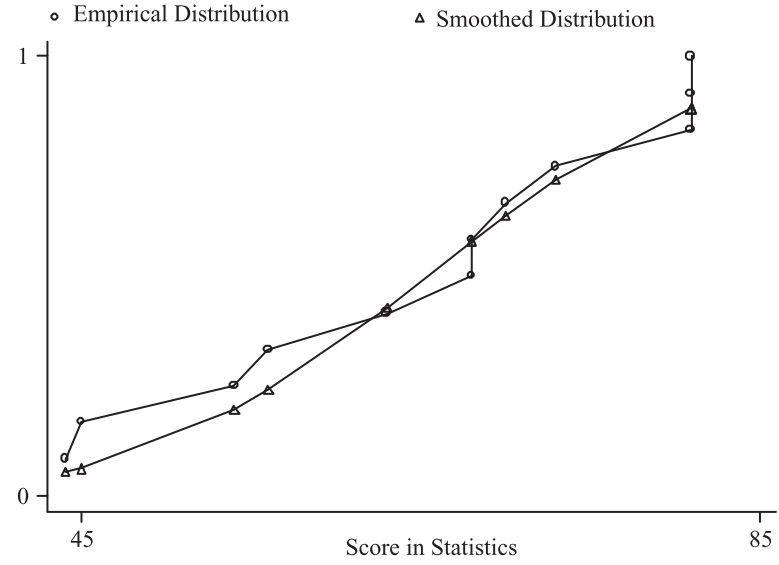


Fig. 2.4. Cumulative distribution of examination scores. The jagged curve o–o–o is the empirical (observed) distribution function.

To obtain an improved bootstrap interval for a median, for example, we would modify our bootstrap procedure as follows:

1. Sample $z_1, z_2, \ldots, z_n$ with replacement from the original observations $x_1, x_2, \ldots, x_n$.
2. Compute their mean $\overline{z}$ and the plug-in estimate of the population variance $\hat{\sigma}^2 = \sum_{i=1}^{n}(z_i - \overline{z})^2/n$.
3. Set

$$x_i^* = \overline{z} + \frac{(z_i - \overline{z} + h\varepsilon_i)}{\sqrt{1 + h^2/\hat{\sigma}^2}},$$

[4] See Glossary at the end of this text.

[5] So-called because of its similarity in form to Student's t.

where the ε_i are drawn from an $N(0, 1)$ distribution,[6] and h is a smoothing constant, chosen by trial and error.

4. Use $x_1^*, x_2^*, \ldots, x_n^*$ to estimate the median.

To obtain a smoothed bootstrap estimate of the distribution of a correlation coefficient, we use a similar procedure:

1. We take a bootstrap sample $(x, y)_1^*, (x, y)_2^*, \ldots, (x, y)_n^*$ with replacement from the original set of simultaneous observations on two *variables* $(x, y)_1, (x, y)_2, \ldots, (x, y)_n$.
2. Following Hall and Wilson [1991], we compute the bivariate mean $(\overline{x}^*, \overline{y}^*)$ of the bootstrap sample, and its covariance matrix Σ^*.
3. We draw n pairs (ε_i, η_i) from the bivariate normal distribution with mean $(\overline{x}^*, \overline{y}^*)$ and covariance matrix Σ^*, and use them to form

$$x_i^* = \overline{x}^* + \frac{(x_i - h\varepsilon_i)}{\sqrt{1 + h^2/\sigma_x^{2*}}}, \quad y_i^* = \overline{y}^* + \frac{(y_i - hn_i)}{\sqrt{1 + h^2/\sigma_y^{2*}}}.$$

4. Finally, we use $(x_1^*, y_1^*), (x_2^*, y_2^*), \ldots, (x_n^*, y_n^*)$ to estimate the correlation coefficient.

Computing a Smoothed Bootstrap Estimate of the Median

Sample Calculation with S-Plus
After downloading the resample library, at
www.insightful.com/downloads/libraries sbootstrap(data, median)

Sample Calculation with Stata
. bsample _N

$$\overline{z} = 149 \text{ and } \hat{\sigma}^2 = 64.18$$

. invnorm (uniform)
If we take $h = 0.5$, then

$$x_1^* = 148.9 + \frac{143.5 - 148.9 + 0.5\varepsilon_1}{\sqrt{1 + 0.25/64.18}} = 143.54$$

2.3.7 Importance Sampling and the Tilted Bootstrap

The idea behind importance sampling is to reduce the variance of the estimate and, thus, the number of bootstrap samples necessary to achieve a given level of confidence. This technique is of particular value with estimates such as the sample variance and the 90th percentile that are strongly dependent upon outlying values. Instead of giving

[6] That is, a normal distribution with mean 0 and variance 1.

equal probability to each of the observations in the original sample, greater weight is given to some of the observations and less weight to others.

Suppose we are testing the hypothesis that the population mean is 0 and our original sample contains the observations −2, −1, −0.5, 0, 2, 3, 3.5, 4, 7, 8. Bootstrap samples containing 7 and 8 are much more likely to have large means, so instead of drawing bootstrap samples such that every observation has the same probability 1/10 of being included, we weight the sample so that larger values are more likely to be drawn, selecting −2 with probability 1/55 for example, −1 with probability 2/55, and so forth, with the probability of selecting 8 being 9/55th's. Let $I(S^* > 0) = 1$ if the mean of the bootstrap sample is greater than 0 and 0 otherwise. The standard estimate of the proportion of bootstrap samples greater than zero is given by the formula

$$\frac{1}{B}\sum_{b=1}^{B} I(S_b^* > 0),$$

where B is the number of bootstrap samples. Since the elements of our bootstrap samples were selected with unequal probability, we have to use a weighted estimate of the proportion

$$\frac{1}{B}\sum_{b=1}^{B} I(S_b^* > 0)\frac{\Pi_{i=1}^{n}\pi_i^{n_{ib}}}{(1/10)^n},$$

where π_i is the probability of selecting the ith element of the original sample of size n and n_{ib} is the number of times the ith element appears in the bth bootstrap sample of size n. Efron and Tibshirani [1993, p. 355] found a seven-fold reduction in variance using importance sampling with samples of size ten.

2.3.8 Iterated Bootstrap

Rephrasing the two principles that were enunciated earlier, we want confidence intervals that are:

a) Unbiased in the sense that they are more likely to contain the correct value of the functional being estimated than any specific incorrect value;
b) Narrow so that they are less likely to contain incorrect values.

Running any of the programs provided earlier in this chapter, it is easy to see that we can reduce the width of a confidence interval by lowering the degree of required confidence: a 90% confidence interval is narrower than a 95% confidence interval, and an 80% confidence interval is narrower still.

We can also narrow the width of a confidence interval by increasing the size of the original sample. In fact, if we could afford to examine every member of the population, we could reduce the interval to a point. If you've taken a prior course in statistics, then you'll recall that a 95% confidence interval for the mean of a normal distribution is given by the formula $\overline{X} \pm 1.96s/\sqrt{n}$ where s is the standard deviation of the sample. If you want to halve the size of such a confidence interval, you have to increase the

sample size from 100 to 400. Such confidence intervals, whose width for large samples is proportional to $n^{-1/2}$ are said to be *first-order exact*.

The percentile bootstrap confidence interval is also first-order exact. BC_α intervals are second-order exact in that their width for large samples is proportional to n^{-1}.

Bootstrap iteration can improve the accuracy of bootstrap confidence intervals, that is can make it more likely that they will cover the true value of the population functional with the claimed degree of confidence. They can also increase the rate at which the width decreases from second to third order.

Let I be a $1-\alpha$ level confidence interval for θ whose actual coverage $\pi_\theta[\alpha]$ depends upon both the true value of θ and the hoped-for coverage $1-\alpha$. In most cases $\pi_\theta[\alpha]$ will be larger than $1-\alpha$. Let α' be the value of α for which the corresponding confidence interval I' has probability $1-\alpha$ of covering the true parameter. We can't solve for α' and I' directly. But we can obtain a somewhat more accurate confidence interval I^* based on α^* the estimate of α' obtained by replacing θ with its plug-in estimate and the original sample with a bootstrap sample. Details of the procedure are given in Martin [1990, pp. 113–114]. The iterated bootstrap, while straightforward, is computationally intensive. Suppose the original sample has 30 observations, and we take 300 bootstrap samples from each of 300 bootstrap samples. That's $30 \times 300 \times 300 = 2{,}700{,}000$ values, plus an equal number of calculations![7]

2.4 Summary

In this chapter, you learned and quickly mastered the use of the primitive or percentile bootstrap to obtain estimates of the precision of estimates. You then learned a variety of methods including the BC_α bootstrap, the bootstrap-t, the tilted bootstrap, the iterative bootstrap, and bootstrap smoothing for improving the accuracy of confidence intervals. Computer code was provided to aid in putting these methods into practice.

2.5 To Learn More

The bootstrap has its origins in the seminal work of Jones [1956], McCarthy [1969], Hartigan [1969,1971], Simon [1969], and Efron [1979,1982]. One of its earliest applications to real-world data may be found in Makinodan et al. [1976]. For further examples of the wide applicability of the bootstrap method, see Chernick [1999] and Chapter 7 of Efron and Tibshirani [1993]. These latter authors comment on when and when not to use bias-corrected estimates. Nontechnical descriptions may be found in Diaconis and Efron [1983], Efron and Tibshirani [1991], Lunneborg [1985], and Rasmussen [1987].

Other general texts and articles of interest include Efron and Tibshirani [1986, 1991], Mooney and Duval [1993], Stine [1990], and Young [1994]. Bias-corrected-and-accelerated bootstrap confidence intervals and a computationally rapid approximation

[7] A third iteration involving some 800 million numbers will further improve the interval's accuracy.

known as the ABC method are described in Efron and Tibshirani [1993, Chapter 14]. For deriving improved two-sided confidence intervals, see the discussion by Loh following Hall [1988; pp. 972–976]. See, also, DiCiccio and Romano [1988], Tibshirani [1988], and Efron and DiCiccio [1992].

The smoothed bootstrap was introduced by Efron [1979] and Silverman [1981], and is considered at length in Silverman and Young [1987], Praska Rao [1983], DiCiccio, Hall, and Romano [1989], and Falk and Reiss [1989]. For discrete distributions, the approach of Lahiri [1993] is recommended. Hårdle [1991] provides S-Plus computer algorithms to implement the smoothed bootstrap. The weighted bootstrap is analyzed by Barbe and Bertail [1995] and Gleason [1988]. Importance sampling for bootstrap tail probabilities is discussed in Johns [1988], Hinkley and Shi [1989], Do and Hall [1991], and Phipps [1997].

Hall [1986], Beran [1987], Hall and Martin [1988], and Hall [1992] describe iterative methods, known respectively as bootstrap pivoting and bootstrap inverting, that provide third-order accurate confidence intervals. Loh [1987, 1991] describes a bootstrap calibration method that yields confidence intervals that in some circumstances are fourth-order accurate.

Potential flaws in the bootstrap approach are considered by Schenker [1985], Wu [1986], Dicicio and Romano [1988], Efron [1988, 1992], Knight [1989], and Gine and Zinn [1989]. Canty et al. [2000] provide a set of diagnostics for detecting and dealing with potential error sources.

2.6 Exercises[8]

1. Use 50–100 bootstrap samples to obtain 80% confidence intervals for the standard deviation and median of the heights of my sample of 22 sixth-graders. How do these values compare with the standard deviation and the median of the original sample?
2. If every student in my sixth-grade class grew five inches overnight, what would the mean, median, and variance of their new heights have been?
   If I'd measured their heights in inches rather than centimeters, what would the mean, median, and variance of their heights have been? (Assume that 2.54 cm = 1 inch.)
   (Hint: Step away from your computer—you won't need it—and think about your answer.)
3. Health care costs are still out of control and there is wide variation in the costs from region to region. The Community Care Network which provides billing services for several dozen PPO's recorded the following cost data for a single dermatological procedure, debridement:

   CIM $198 200.2 242
   HAJ 83.2 193.6
   HAV 197
   HVN 93

[8] Answers to selected exercises will be found in a separate section at the end of this volume.

LAP 105

MBM 158.4 180.4 160.6 171.6 170 176 187

VNO 81.2 103.0 93.8

VPR 154 228.8 180.4 220 246.4 289.7 198 224.4

Obtain 90% confidence intervals for the 80th and the 20th percentile.

4. Suppose you wanted to estimate the population variance. Most people would recommend you divide the sum of squares about the sample mean by the sample size minus 1, while others would say, too much work, just use the plug-in estimate. What do you suppose the difference is? (Need help deciding? Try this experiment: Take successive bootstrap samples of size 11 from the sixth-grade height data and use them to estimate the variance of the entire class.)

5. "Use the sample median to estimate the population median and the sample mean to estimate the population mean." This sounds like good advice, but is it? Use the technique described in the previous exercise to check it out.

6. Below is the billing data from four Swedish hospitals. Obtain confidence intervals for the median and standard deviation of the bills at each hospital. Use as many methods for obtaining the CI's as are practical with the software at your disposal.

Hospital 1

64877, 21152, 11753, 1834, 3648, 12712, 11914, 14290, 17132,
7030, 23540, 5413, 4671, 39212, 7800, 10715, 11593, 3585,
12116, 8287, 14202, 4196, 22193, 3554, 3869, 3463, 2213,
3533, 3523, 10938, 3533, 3523, 10938, 17836, 3627, 30346,
2673, 3703, 28943, 8321, 19686, 18985, 2243, 4319, 3776,
3668, 11542, 14582, 9230, 7786, 7900, 7886, 67042, 7707,
18329, 7968, 5806, 5315, 11088, 6966, 3842, 13217, 13153,
8512, 8328, 207565, 2095, 18985, 2143, 7976, 2138, 15313,
8262, 9052, 8723, 4160, 7728, 3721, 18541, 7492, 18703,
6978, 10613, 15940, 3964, 10517, 13749, 24581, 3465, 11329,
7827, 3437, 4587, 14945, 23701, 61354, 3909, 14025, 21370,
4582, 4173, 4702, 7578, 5246, 3437, 10311, 8103, 11921,
10858, 14197, 7054, 4477, 4406, 19170, 81327, 4266, 2873,
7145, 4018, 13484, 7044, 2061, 8005, 7082, 10117, 2761,
7786, 62096, 11879, 3437, 17186, 18818, 4068, 10311, 7284,
10311, 10311, 24606, 2427, 3327, 3756, 3186, 2440, 7211,
6874, 26122, 5243, 4592, 11251, 4141, 13630, 4482, 3645,
5652, 22058, 15028, 11932, 3876, 3533, 31066, 15607, 8565,
25562, 2780, 9840, 14052, 14780, 7435, 11475, 6874, 17438,
1596, 10311, 3191, 37809, 13749, 6874, 6874, 2767, 138133

Hospital 2

4724, 3196, 3151, 5912, 7895, 19757, 21731, 13923, 11859,
8754, 4139, 5801, 11004, 3889, 3461, 3604, 1855

Hospital 3

4181, 2880, 5670, 11620, 8660, 6010, 11620, 8600, 12860,
21420, 5510, 12270, 6500, 16500, 4930, 10650, 16310, 15730,
4610, 86260, 65220, 3820, 34040, 91270, 51450, 16010, v6010,
15640, 49170, 62200, 62640, 5880, 2700, 4900, 55820, 9960,
28130, 34350, 4120, 61340, 24220, 31530, 3890, 49410, 2820,
58850, 4100, 3020, 5280, 3160, 64710, 25070

Hospital 4

10630, 81610, 7760, 20770, 10460, 13580, 26530, 6770, 10790,
8660, 21740, 14520, 16120, 16550, 13800, 18420, 3780, 9570,
6420, 80410, 25330, 41790, 2970, 15720, 10460, 10170, 5330,
10400, 34590, 3380, 3770, 28070, 11010, 19550, 34830, 4400,
14070, 10220, 15320, 8510, 10850, 47160, 54930, 9800, 7010,
8320, 13660, 5850, 18660, 13030, 33190, 52700, 24600, 5180,
5320, 6710, 12180, 4400, 8650, 15930, 6880, 5430, 6020,
4320, 4080, 18240, 3920, 15920, 5940, 5310, 17260, 36370,
5510, 12910, 6520, 5440, 8600, 10960, 5190, 8560, 4050,
2930, 3810, 13910, 8080, 5480, 6760, 2800, 13980, 3720,
17360, 3770, 8250, 9130, 2730, 18880, 20810, 24950, 15710,
5280, 3070, 5850, 2580, 5010, 5460, 10530, 3040, 5320,
2150, 12750, 7520, 8220, 6900, 5400, 3550, 2640, 4110,
7890, 2510, 3550, 2690, 3370, 5830, 21690, 3170, 15360,
21710, 8080, 5240, 2620, 5140, 6670, 13730, 13060, 7750,
2620, 5750, 3190, 2600, 12520, 5240, 10260, 5330, 10660,
5490, 4140, 8070, 2690, 5280, 18250, 4220, 8860, 8200,
2630, 6560, 9060, 5270, 5850, 39360, 5130, 6870, 18870,
8260, 11870, 9530, 9250, 361670, 2660, 3880, 5890, 5560,
7650, 7490, 5310, 7130, 5920, 2620, 6230, 12640, 6500,
3060, 2980, 5150, 5270, 16600, 5880, 3000, 6140, 6790,
6830, 5280, 29830, 5320, 7420, 2940, 7730, 11630, 9480,
16240, 2770, 6010, 4410, 3830, 3280, 2620, 12240, 4120,
5030, 8010, 5280, 4250, 2770, 5500, 7910, 2830, 11940,
9060, 20130, 10150, 6850, 10160, 7970, 12960, 31550

7. In answering the preceding question, did you use the same size for the bootstrap sample from each of the four sets of observations? Would there be an advantage to taking a bootstrap sample that was larger or smaller than the original sample? Why or why not?

3

Comparing Two Populations

Perhaps the second most common application of statistics lies in the comparison of two populations or, more accurately, the comparison of two populations on the basis of two samples, one from each population.

Will a new medication reduce blood pressure at a lower dosage than the current most popular treatment?

Will a sales letter using color enhancements attract more customers than one limited to black and white?

While the bootstrap methods used in the previous chapter can be adapted for these comparisons, they are neither as exact nor as powerful as the permutation methods you will now study.

3.1 A Laboratory Experiment

Shortly after I received my doctorate in statistics,[1] I decided that if I really wanted to help bench scientists apply statistics I ought to become a scientist myself. So back to school[2] I went to learn all about physiology and aging in cells raised in Petri dishes.

I soon learned there was a great deal more to an experiment than the random assignment of subjects to treatments. In general, 90% of the effort was spent in mastering various arcane laboratory techniques, 9% in developing new techniques to span the gap between what had been done and what I really wanted to do, and a mere 1% on the experiment itself. But the moment of truth came finally—it had to if I were to publish and not perish—and I succeeded in cloning human diploid fibroblasts in eight culture dishes: Four of these dishes were filled with a conventional nutrient solution and four held an experimental "life-extending" solution to which vitamin E had been added.

I waited three weeks with my fingers crossed—there is always a risk of contamination with cell cultures—but at the end of this test period three dishes of each type

[1] From the University of California at Berkeley.

[2] The Wistar Institute, Philadelphia, PA and the W. Alton Jones Cell Science Center in Lake Placid, NY.

had survived. My technician and I transplanted the cells, let them grow for 24 hours in contact with a radioactive label, and then fixed and stained them before covering them with a photographic emulsion.

Ten days passed and we were ready to examine the autoradiographs. Two years had elapsed since I first envisioned this experiment and now the results were in: I had the six numbers I needed.

"I've lost the labels," my technician said as she handed me the results.

"What!?" Without the labels, I had no way of knowing which cell cultures had been treated with vitamin E and which had not.

"121, 118, 110, 34, 12, 22." I read and reread these six numbers over and over again. If the first three counts were from treated colonies and the last three were from untreated, then I had found the fountain of youth. Otherwise, I really had nothing to report.

3.2 Analyzing the Experiment

How had I reached this conclusion? Let's take a second, more searching look at the problem of the missing labels. First, we identify the hypothesis and the alternative(s) of interest.

I wanted to assess the life-extending properties of a new experimental treatment with vitamin E. To do this, I had divided my cell cultures into two groups: one grown in a standard medium and one grown in a medium containing vitamin E. At the conclusion of the experiment and after the elimination of several contaminated cultures, both groups consisted of three independently treated dishes.

My *null hypothesis* was that the growth potential of a culture would not be affected by the presence of vitamin E in the media: All the cultures would have equal growth potential. The *alternative* of interest was that cells grown in the presence of vitamin E would be capable of many more cell divisions.

Under the null hypothesis, the labels "treated" and "untreated" provide no information about the outcomes: The observations are expected to have more or less the same values in each of the two experimental groups. If they do differ it should only be as a result of some uncontrollable random fluctuation. Thus, if this null or no-difference hypothesis were true, I was free to exchange the labels.

The next step in the permutation method is to choose a test statistic that discriminates between the hypothesis and the alternative. The statistic I chose was the sum of the counts in the group treated with vitamin E. If the alternative hypothesis is true, most of the time this sum ought to be larger than the sum of the counts in the untreated group. If the null hypothesis is true, that is, if it doesn't make any difference which treatment the cells receive, then the sums of the two groups of observations should be approximately the same. One sum might be smaller or larger than the other by chance, but most of the time the two shouldn't be all that different.

The third step in the permutation method is to compute the test statistic for each of the possible relabelings and compare these values with the value of the test statistic as the data was labeled originally. Fortunately, I'd kept a record of the treatments independent of my technician. In fact, I'd deliberately not let my technician know which cultures were

which in order to ensure she would give them equal care in handling. As it happened, the first three observations my technician showed me—121, 118, and 110—were those belonging to the cultures that received vitamin E. The value of the test statistic for the observations as originally labeled is $349 = 121 + 118 + 110$.

I began to rearrange (permute) the observations, randomly reassigning the six labels, three "treated" and three "untreated," to the six observations. For example: treated, 121 118 34, and untreated, 110 12 22. In this particular rearrangement, the sum of the observations in the first (treated) group is 273. I repeated this step till all $\binom{6}{3} = 20$ distinct rearrangements had been examined.[3]

	First Group	Second Group	Sum of First Group
1.	121 118 110	34 22 12	349
2.	121 118 34	110 22 12	273
3.	121 110 34	118 22 12	265
4.	118 110 34	121 22 12	262
5.	121 118 22	110 34 12	261
6.	121 110 22	118 34 12	253
7.	121 118 12	110 34 22	251
8.	118 110 22	121 34 12	250
9.	121 110 12	118 34 22	243
10.	118 110 12	121 34 22	240
11.	121 34 22	118 110 12	177
12.	118 34 22	121 110 12	174
13.	121 34 12	118 110 22	167
14.	110 34 22	121 118 12	166
15.	118 34 12	121 110 22	164
16.	110 34 12	121 118 22	156
17.	121 22 12	118 110 34	155
18.	118 22 12	121 110 34	152
19.	110 22 12	121 118 34	144
20.	34 22 12	121 118 110	68

The sum of the observations in the original vitamin-E-treated group, 349, is equaled only once and never exceeded in the 20 distinct random relabelings. If chance alone is operating, then such an extreme value is a rare, only one-time-in-20 event. If I reject the null hypothesis and embrace the alternative that the treatment is effective and responsible for the observed difference, I only risk making an error and rejecting a true hypothesis one in every 20 times.

In this instance, I did make just such an error. I was never able to replicate the observed life-promoting properties of vitamin E in other repetitions of this experiment. Good statistical methods can reduce and contain the probability of making a bad decision, but they cannot eliminate the possibility.

[3] Determination of the number of relabelings, "6 choose 3" in the present case, is considered in Section 3.7.1.

3.3 Some Statistical Considerations

The preceding experiment, simple though its analysis may be, raises a number of concerns that are specifically statistical in nature. How do we go about framing a hypothesis? If we reject a hypothesis, what are the alternatives? Why must our conclusions always be in terms of probabilities rather than certainties? We consider each of these issues in turn.

3.3.1 Framing the Hypothesis

The hypothesis that we tested in this example, that cell cultures raised in medium to which vitamin E has been added have no particular advantage in growth potential over cultures raised in ordinary medium, is what statisticians term a *null hypothesis*. It is based on the assumption the various samples tested are all drawn from the same hypothetical population, or, at least, from populations for which the variable we observe all have the same distribution. Thus, the null hypothesis that the distribution of the heights of sixth-grade boys is the same as the distribution of the heights of sixth-grade girls clearly is false, while the null hypothesis that distribution of the heights of sixth-grade boys in Orange County is the same as the distribution of the heights of sixth-grade boys in neighboring Los Angeles County may well be true.

The majority of hypotheses considered in this text are null hypotheses, but they are not the only possible hypotheses. For example, we might want to test whether all cultures treated with vitamin E would show an increase of at least three generations in life span. As it is difficult to test such a hypothesis, we normally would perform an initial *transformation* of the data, subtracting three from each of the observations in the vitamin-E-treated sample. Then, we would test the null hypothesis that there is no difference between the observations from the original untreated sample and the transformed observations.[4]

3.3.2 Hypothesis Versus Alternative

Whenever we reject a hypothesis, we accept an *alternative*.[5] In the present example, when we rejected the null hypothesis, we accepted the alternative that vitamin E had a beneficial effect. Such a test is termed *one-sided*. A two-sided test would also guard against the possibility that vitamin E had a detrimental effect. A two-sided test would reject for both extremely large and extremely small values of our test statistic. See Section 3.4.3 for a more extensive discussion. In more complex situations, such as those considered in Chapter 7, there can be many different tests depending on which alternatives are of greatest interest.

3.3.3 Unpredictable Variation

In reviewing the possible outcomes of this experiment, the phrase “most of the time” was used repeatedly. As we saw in Chapter 2, something that is “predominantly” true

[4] See Exercise 6 at the end of this chapter.

[5] Or we may conclude we need more data before we can reach a decision.

may still be false on occasion. We expect a *distribution* of outcomes. Vitamin E indeed may have a beneficial effect, but all cultures may not be equally affected. Too many other factors can intervene. Any conclusions we draw must be probabilistic in nature. Always, we risk making an error and rejecting a true hypothesis some percentage of the time.[6]

We say that we make a *Type I error* when we accept an alternative hypothesis and yet the null or primary hypothesis is true. We say that we make a *Type II error* when we accept the primary hypothesis, yet an alternative is true. Before we analyze data, we establish a set of values for the test statistic for which we will reject the primary hypothesis known as *the rejection region*. Its complement, the set of values for the test statistic for which we will accept the primary hypothesis is known as *the acceptance region*. The boundaries separating these regions are chosen so that the *significance level*, defined as the probability of making a Type I error, will be less than some fixed value. 5% and 1% are among the most frequent choices for a significance level.

Once this choice is made, the power of the test is also determined. *Power* of a test is defined as the probability of rejecting the hypothesis when a specific alternative is true. Thus the power is 1 minus the probability of making a Type II error. Note that the power depends on the alternative.[7]

After we analyze the data, we will obtain a p-value that depends upon the samples. If the p-value is less than or equal to the significance level we will reject the primary hypothesis, otherwise we will accept it.

3.3.4 Some Not-so-Hidden Assumptions

The p-values we obtain by examining all possible rearrangements will be *exact*, not approximations, *only* if the following is true:

> *Under the null hypothesis, we can rearrange the labels on the observations without affecting the underlying distribution of possible outcomes.*

This latter requirement deserves a second look. Under the null hypothesis, each possible rearrangement is equally likely. In our vitamin E example, rearrangement

$$7. \qquad 121\ 118\ 12 \qquad 110\ 34\ 22 \qquad \text{sum} = 251$$

is as likely as our initial outcome

$$1. \qquad 121\ 118\ 110 \qquad 34\ 22\ 12 \qquad \text{sum} = 349$$

under the null hypothesis. But if, to the contrary, vitamin E has an age-extending effect, then our initial outcome is more probable than any other rearrangement including rearrangement 7.

By *exact*, we mean that when we state that a p-value is 5%, it is exactly that. By contrast, p-values obtained from the chi-square distribution in the analysis of contingency

[6] A possible exception might arise when we have hundreds of thousands of observations. There might be a chance even then, albeit a very small one, of our making an error.

[7] Further comments on the relationship of power to the alternative, to the significance level, and to sample size will be found in the next chapter.

tables with small numbers of observations can be crude misleading approximations; a stated 5% might correspond to an actual 20% (see Chapter 6). And when a t-test is used to compare two samples of five observations taken from a distribution that is not normal, a stated p-value of 5% may correspond to a true value of 7% or 8%.

The true p-values of tests based on *parametric methods*,[8] such as the t-test, the chi-square, and the F-test, depend on the underlying distribution of the observations and are exact only for specific distributions. By contrast, the p-values of tests based on permutation methods do *not* depend on the underlying distribution of the observations and are termed *distribution-free*.

As we saw in Chapter 2, bootstrap confidence intervals are inaccurate. As we will show in Section 4.3, this means that tests based on the bootstrap are not exact. Theoreticians tell us that regardless of the underlying distributions, parametric, permutation, and bootstrap tests of hypotheses are (almost) exact for samples in excess of 100,000 in number. Good news for some, I guess.

The power of permutation tests is quite high. For small samples, a permutation test may be the most powerful test available. For very large samples, a permutation test will be as powerful as the most powerful parametric test.

3.3.5 More General Hypotheses

Our permutation technique is applicable whenever we can freely exchange labels under the null hypothesis; this is obviously the case when testing the hypothesis that both samples are drawn from the same population. But suppose our hypothesis is somewhat different, that we believe, for example, that a new gasoline additive will increase mileage by at least 20 miles per tank. (Otherwise, why pay the big bucks for the additive?)

In order to perform a permutation test, the hypothesis must always be cast in null form. We could swap our primary hypothesis and our null alternative before performing the analysis. That is, we can test the hypothesis: "the additive is not effective" against the alternative "additive will increase mileage by 20 miles per tank."

But in this case, there is a second possible approach. Suppose that before using the additive, we recorded 295, 320, 329, and 315 miles per tank of gasoline under various traffic conditions. With the additive, we recorded 330, 310, 345, and 340 miles per tank. To perform the analysis, we could begin by subtracting 20 from each of the post-additive figures. Our two transformed samples are

295 320 329 315

305 290 325 320

Our revised hypothesis is that the transformed mileages will be approximately the same, that is, "the additive increases performance by 20 mpg," versus the alternative that the post-additive transformed mileages will be less than the others, that is, "the increase in mileage is strictly less than 20 mpg." Which approach you choose will depend upon which hypothesis/alternative pair comes closest to your original intentions.

[8] See Glossary.

Resampling Terminology

Like too many professionals, statisticians have developed an extensive terminology with words and phrases whose meanings are almost but not quite the same as other similar terms. For example, the *permutation test* makes use of random rearrangements, not the more numerous permutations. A *rank test* is a permutation test applied to the ranks of observations rather than their original values.

The permutation test, like the *nonparametric bootstrap* is a *distribution-free test*. Such tests are often referred to as *nonparametric tests*. Along with the *parametric bootstrap*, they are called *resampling methods* because they involve resampling limited to or based on an existing set of observations. The resampling methods also include *decision trees* (see Chapter 9) as well as the *jackknife* (see Chapter 8) and its variants.

3.4 Computing the p-Value

Each time you analyze a set of data via the permutation method, you will follow the same five-step procedure:

1. Analyze the problem—identify the alternative(s) of interest.
2. Choose a test statistic that best distinguishes between the alternative and the null hypothesis. In the case of the two-sample comparison, the sum of the observations in the sample from the treated population is the obvious choice as it will be large if the alternative is true and entirely random otherwise.
3. Compute the test statistic for the original labeling of the observations.
4. Rearrange the labels, then compute the test statistic again. Repeat until you obtain the distribution of the test statistic for all possible rearrangements or for a large random sample thereof.
5. Set aside one or two tails of the resulting distribution as your rejection region. In the present example with a one-sided alternative, we would set aside the largest values as our rejection region. If the sum of the treatment observations for the original labeling of the observations is included in this region then we will reject the null hypothesis.

For small samples, it would seem reasonable (and a valuable educational experience) to examine all possible rearrangements of labels. But for larger samples, even for samples with as few as six or seven values, complete enumeration may be impractical. While there are only 20 different ways we can apply two sets of three labels, there are 5542 different ways we can apply two sets of six labels. Though 5542 is not a challenge for today's desktop computers, if you are writing your own programs you'll find it simpler and almost as accurate to utilize *the Monte Carlo method* described next.

3.4.1 Monte Carlo

In a Monte Carlo (named for the Monaco casino), we use the computer to generate a random rearrangement of the labels (control and treated; new and old). Suppose

that **p** is the unknown probability that the value of the test statistic for the rearranged values will be as or more extreme than our original value. Then if we generate **n** random rearrangements, the number of values that are as or more extreme than our original value will be a binomial random variable $B(n, p)$ with mean np and variance $np(1-p)$. Our estimate of p based on this random sample of n rearrangements will have an expected value of p and a standard deviation of $[p(1-p)/n]^{1/2}$.

Suppose p is actually 4% and we examine 400 rearrangements, then 95% of the time we can expect our estimate of p to lie between .04–1.96[.04*.96/400]$^{1/2}$ and .04 + 1.96[.04*.96/400]$^{1/2}$ or 2% to 6%. Increasing the number of rearrangements to 6400, we would get a p-value that is accurate to within half of 1%.

3.4.2 Program Code

To compare teaching methods, 20 school children were randomly assigned to one of two groups. The following are the test results:

conventional	65	79	90	75	61	85	98	80	97	75
new	90	98	73	79	84	81	98	90	83	88

Are the two teaching methods equivalent in result?

C++

```
int Choose (int lo, int hi)
{
       int z = rand()%(hi - lo +1) + lo;
       return (z);
}

//Pack all the observations in a single vector Data.
//Determine n[1]=#conventional and n[2]=#New.

float CompTwo(float *X){
  float sum =0, temp;
   int k;
   for (int i=0; i< n[1]; ++i){
          k=Choose (i, n[1]+n[2]-1);
          temp = *(X+k); *(X+k)=*(X+i); *(X+i)= temp;
          sum =sum +temp;
     }
   return (sum);
   }

float stat0=0, stat;
int cnt;
for (int i =0; i < n[1]; ++i)stat0=stat0 +Data[i];
cnt=0;
```

```
for (int i =0; i < MC; ++i){
    stat=Comp2(Data);
    if (stat >=stat0) cnt++;
   }
float p=float(cnt)/MC;
cout << p << endl;
```

EViews

```
!n = 400 ' number of rearrangements to be examined
create u 1 10
series conventional
conventional.fill 65,79,90,75,61,85,98,80,97,75
series new
new.fill 90,98,73,79,84,81,98,90,83,88
!sumorig =@sum(new)
!count = 0

range 1 20 ' expande worfile space
' place data into a new data series A
series a = new
smpl 11 20
a = conventional(-10)
smpl @all

for !i  = 1 to !n
          ' draw from A without replacement 10 elements
          a.resample(outsmpl=1 10,permute) boot
          if (@sum(boot)<=!sumorig) then
                 !count=!count+1
          endif
next
scalar pvalue=!count/!n
show pvalue
```

Excel – Using Resampling Statistics for Excel

Place the two sets of observations in adjoining columns. Outline them and use the S or Shuffle command.

MatLab

```
% Monte Carlo for Calculating p-value
% This function takes N permuted samples of the
% original data, without replacement and calculated
% the desired statistic on the permutted sample.  It
```

```
% then counts how many cases as or more extreme
% occured to calculate the p-value for the test.
%
% montecarlo(samp1,samp2,NMCS)
% samp1, samp2 are the two arrays of data to be tested
% NMCS is the number of Monte Carlo Simulations to be
% performed.
%


%To be ran before executing the function:
%clear all;
%conventional=[65,79,90,75,61,85,98,80,97,75];
%new=[90,98,73,79,84,81,98,90,83,88];
%to run: montecarlo(conventional,new,1000)


function [p]=montecarlo(samp1,samp2,NMCS)
    rand('state',sum(100*clock)); %reset random
                                        generator to a
                                        different state
    n=length(samp1);
    m=length(samp2);
    data=cat(2,samp1,samp2); %concatinate data arrays

    stat0=sum(samp1); %calculate test statistic
    cnt=0; %intialize count for p calculation

    sample=zeros(1,n); %initialize array to hold
         permutation

    for i=1:NMCS
        choose=randperm(n+m); %randomize indices
        for j=1:n
            %fill sample with random data, up to first
                 n elements
            sample(j)=data(choose(j));
        end;
        stat1=sum(sample);
        if stat1<=stat0
            cnt=cnt+1;
        end;
    end;
    p=cnt/NMCS; %p-value
```

R

```
N=400  #number of rearrangements to be examined
  conventional =c(65, 79, 90, 75, 61, 85, 98, 80,
       97, 75)
  new  = c (90, 98, 73, 79, 84, 81, 98, 90, 83, 88)
n=length (new)
sumorig = sum(new)
cnt= 0  #zero the counter
#Stick both sets of observations in a single vector
A = c(new, conventional)
for (i in 1:N){
    D= sample (A,n)
    if (sum(D) <= sumorig)cnt=cnt+1
   }
cnt/N  #pvalue

[1] 0.9025
```

Resampling Statistics

```
DATA (65 79 90 75 61 85 98 80 97 75) A
SIZE A n
SUM A sumorig
DATA (90 98 73 79 84 81 98 90 83 88) B
LET cnt=0
CONCAT A B C
REPEAT 400
  SHUFFLE C D
  TAKE D 1,n E
  SUM E sumperm
  ' we count only rearrangements in the lower tail
  IF sumperm <= sumorig
  LET cnt=cnt+1
  END
END
DIVIDE cnt 400 pvalue
PRINT pvalue
```

S-Plus

Using S+Resample, we may use the R code or do:
perm = permutationTest2(new,sum,data2 = Other,alternative="less")
perm # this prints the p-value, among other things
plot(perm) # plot shows the relationship between the observed value and the null distribution

STATA

Enter the classification as a separate variable.
input score method

```
                score method
  1.  65   0
  2.  79   0
  3.  ..   ..
 10.  75   0
 11.  90   1
 12.  98   1
```

and so forth

permute score "sum score if method" sum=r(sum), reps(1000) left nowarn
command: sum score if method
statistic: sum = r(sum)
permute var: score
Monte Carlo permutation statistics Number of obs = 20
Replications = 1000

T	T(obs)	c	n	p=c/n	SE(p)	[95% Conf.	Interval]
sum	864	895	1000	0.8950	0.0097	.8743232	.9133152

Note: confidence interval is with respect to p=c/n
Note: c = #{T <= T(obs)}

3.4.3 One-Sided Versus Two-Sided Test

The preceding section provided program code for a number of one-sided tests. These are the appropriate tests to use when we test whether a specific population parameter has the same value for different populations against the one-sided alternative that the parameter's value is greater for the second population. To fix ideas, suppose the parameter in question is the population mean and that the test will be based upon the values of the two sample means. If $\bar{X} < \bar{Y}$ we count the proportion of rearrangements for which $\bar{X}$ is less than or equal to its original value and reject the null hypothesis if this proportion is less than or equal to our predetermined significance level. If $\bar{X} \geq \bar{Y}$ we automatically accept the null hypothesis.

On the other hand, if we wish to test the null hypothesis the population mean of X is identical with the population mean of Y against the two-sided alternative that the population mean is unequal, we would proceed as follows:

- if $\bar{X} < \bar{Y}$ we count the proportion of rearrangements for which $\bar{X}$ is less than or equal to its original value and reject the null hypothesis if twice this proportion is less than or equal to our predetermined significance level.

- if $\bar{X} > \bar{Y}$ we count the proportion of rearrangements for which $\bar{X}$ is greater than or equal to its original value and reject the null hypothesis if twice this proportion is less than or equal to our predetermined significance level.

The resultant p-value correctly reflects our intent to reject if the value of our test statistic proves to be either extremely large or extremely small with respect to the distribution of all possible values that might be obtained by random relabeling.

3.5 Matched Pairs

We can increase the power of our tests by eliminating or reducing unwanted sources of variation. One of the best ways to eliminate a source of variation and subsequent errors is through the use of *matched pairs*. Each subject in one group is matched as closely as possible by a subject in the treatment group. If a 45-year-old black male hypertensive is given a blood-pressure lowering pill, then we give a second similarly built 45-year-old black male hypertensive a placebo.

In performing the corresponding permutation tests, we rearrange labels on a block-by-block basis. Suppose that each subject in our experiment served as his/her own control and we had exactly three subjects; there would be 8 possible rearrangements. For example,

Original Results			
Subject	1	2	3
Control	17	19	14
Treated	21	21	17

Rearrangement One			
Subject	1	2	3
Control	21	19	14
Treated	17	21	17

Rearrangement Two			
Subject	1	2	3
Control	17	21	14
Treated	21	19	17

Rearrangement Three			
Subject	1	2	3
Control	21	21	14
Treated	17	19	17

With n matched pairs, there would be 2^n possible rearrangements. Using as our test statistic the sum of the control observations, we would reject only if this sum when calculated for our observations with their original labels fell in a tail of the permutation distribution.

EViews

```
create u 8
series New
new.fill 48722, 28965, 36581, 40543, 55423, 38555,
     31778, 45643
series standard
standard.fill 46555, 28293,  37453, 38324,  54989,
     35687,  32000, 43289
```

```
!N=400
!sumorig =@sum(new)
!count = 0
for !i  = 1 to !N
     series stat = @recode(rnd>=0.5,standard,new)
     if (@sum(stat)<=!sumorig) then
          !count=!count+1
     endif
next
scalar pvalue=!count/!N
show pvalue
```

Excel – Using Resampling Statistics for Excel

Place the two sets of observations in adjoining columns. Outline them and use the S or Shuffle command with the "Shuffle Within Rows" option checked.

MatLab

```
%
% Matched Pairs
% This function tests two samples by using matched
% pairs.
% The function chooses between the ith element of the
% two samples, computes and compares the test
% statistics to calculate the p-value.
% match(samp1,samp2,N)
% samp1,samp2 are arrays containing data to be tested/
% N is the number times to perform mathed pairs.
%

% Before executing function:
% clear all;
% new=[48722,28965,36581,40543,55423,38555,31778,
     45643];
% standard=[46555,28293,37453,35324,54989,35687,
     32000,43289];
% To run function:
% match(new,standard,1000)

function [p]=match(samp1,samp2,N)
    rand('state', sum(100*clock)); %reset random
                                        generator to a
                                        different state
    n=length(samp1);
```

```
stat0=sum(samp1); %test statistic
samp=zeros(1,n);  %initialize sample array
cnt=0;

for i=1:N
    choose=rand(1,n);
    for j=1:n  %will choose between ith element
                    of samp1 or samp2
        if choose(j)<.5
            samp(j)=samp1(j);
        else
            samp(j)=samp2(j);
        end
    end
    stat1=sum(samp); %test statistic of sample
    if stat1>=stat0
        cnt=cnt+1;
    end
end
p=cnt/N; %p-value
```

R Code

```
New = c(48722, 28965, 36581, 40543, 55423, 38555,
     31778, 45643)
Standard=c(46555, 28293,  37453, 38324,  54989,
     35687,  32000, 43289)
Diff=New-Standard
N=400   #number of rearrangements to be examined
sumorig = sum(Diff)
n=length(Diff)
stat = numeric (n)
cnt= 0  #zero the counter
for (i in 1:N){
       for (j in  1:n) stat[j]=ifelse(runif(1) < 0.5,
            Diff[j], -Diff[j])
     if (sum(stat) >= sumorig)cnt=cnt+1
   }
cnt/N  #one-sided p-value
    [1] 0.032
```

Stata

```
drop _all
input new stand
48722 46555
```

```
28965 28293
36581 37453
40543 38324
55423 54989
38555 35687
31778 32000
45643 43289
end

set seed 1234
local reps 400
tempvar which
quietly gen `which' = 0
sum new, meanonly
scalar sumorig = r(sum)
local cnt 0
forval i = 1/`reps' {
quietly replace `which' = cond(uniform() < 0.5, new,
     stand)
sum `which', meanonly
if r(sum) <= scalar(sumorig) {
local ++cnt
}
}
di `cnt'/`reps'
exit
```

3.6 Unequal Variances

A slight modification of the preceding assumptions yields a problem whose exact solution has eluded statisticians for decades. Suppose we cannot be certain that the variances of the two populations from which the samples are drawn are the same. Can we still test for the equality of the means?

We cannot use the permutation approach because the labels are no longer exchangeable under the null hypothesis. A similar barrier exists to the use of Student's t. For even if the means of the two populations were equal, an observation in the first sample could be distinguished from one in the second sample because it comes from a population with a different variance.

We *can* use the bootstrap since its use requires only the equality of the population means under the null hypothesis rather than the equality of the distributions. We can obtain a bootstrap test from a confidence interval based on the difference of the two sample means. The "trick" is to draw two separate bootstrap samples each time, one from each of the original samples.

EViews

```
create u 1 9
series treatmt
 treatmt.fill 94, 38, 23, 197, 99, 16, 141
series control
 control.fill 52, 10, 40, 104, 51, 27, 146, 30, 46
' Find observed difference
scalar obsdif = @mean(treatmt) - @mean(control)
!N = 1000
vector(!N) stat
for !i = 1 to !n
 treatmt.resample(outsmpl=1 7) treatmtB
 control.resample(outsmpl=1 9) controlB
 stat(!i)  = @mean(treatmtB) - @mean(controlB)
next
scalar quant05=@quantile(stat,.05)
scalar quant95=@quantile(stat,.95)
```

MatLab

```
%
% Unequal Variances
% Calculates difference of means
% using the bootstrap method, with replacement.
% uneqvar(samp1,samp2,N)
% samp1,samp2 are arrays of data being tested.
% N is the number of times to perform bootstrap
%

% clear all;
% samp1=[94,38,23,197,99,16,141];
% samp2=[52,10,40,104,51,27,146,30,46];
% to run function:
% uneqvar(samp1,samp2,100)

function [stat]=uneqvar(samp1,samp2,N)
    rand('state', sum(100*clock));
    n=length(samp1);
    m=length(samp2);
    samp1B=zeros(1,n);
    samp2B=zeros(1,m);
    stat=zeros(1,N);

    for i=1:N
        for j=1:n
```

```
            samp1B(j)=samp1(round((n-1)*rand+1));
        end
        for j=1:m
            samp2N(j)=samp2(round((m-1)*rand+1));
        end
        stat(i)=mean(samp1B)-mean(samp2B);
    end
```

R

```
#Two Samples
#Efron & Tibshirani, 1993, p. 11
#The observed difference in survival times between
#treatment mice and control mice is 30.63 days.
#Determine a 90% confidence interval around this
#estimate using the percentile bootstrap.
treatmt = c(94, 38, 23, 197, 99, 16, 141)
     #treatment group
control = c(52, 10, 40, 104, 51, 27, 146, 30, 46)
     #control group
n = length (treatmt)
m = length (control)
#Find observed difference
obsdif = mean(treatmt) - mean (control)
#We want to determine whether obsdif is too large to
     have occurred soley by chance
N = 1000
stat = numeric(N)  #create a vector in which to store
     the results
for (i in 1:N){
#bootstrap sample counterparts to observed samples are
     denoted with "B"
 treatmtB = sample (treatmt, replace =T)
 controlB = sample (control, replace =T)
 stat [i]  = mean(treatmtB) - mean (controlB)
  }
quantile (stat, c(0.05, 0.95))
#If the interval does not include obsdif, reject the
     null hypothesis.
```

Resampling Stats

```
'Two Samples
'Efron & Tibshirani, 1993, p. 11
'The observed difference in survival times between
     treatment mice
```

```
'and control mice is 30.63 days.  Determine a 90%
     confidence interval
'around this estimate.  Employ the boostrap-t.
DATA (94 38 23 197 99 16 141) treatmt
     'treatment group
DATA (52 10 40 104 51 27 146 30 46) control
     'control group
'Record group sizes
SIZE treatmt n
SIZE control m
'Find observed difference
MEAN treatmt tmean
MEAN control cmean
LET obsdif = tmean-cmean
'We want to determine whether obsdif is too large to
     have occurred
' soley by chance
'compute std of observed diff
VARIANCE treatmt vt
VARIANCE control vc
LET den = (((n-1)*vt+(m-1)*vc))*(1/n+1/m)/(n+m-2)
SQRT den std
LET t = obsdif/std
'Bootstrap
REPEAT 1000
'bootstrap sample counterparts to observed samples are
     denoted with "$"
  SAMPLE n treatmt treatmt$
  SAMPLE m control control$
  'find the numerator for first Hall-Wilson correction
  MEAN treatmt$ tmean$
  MEAN control$ cmean$
  'parentheses in next statement are essential
  LET dif$ = (tmean$-cmean$)-obsdif
  'find the denominator for second Hall-Wilson
       correction
  VARIANCE treatmt$ vt$
  VARIANCE control$ vc$
  LET den2 = (((n-1)*vt$+(m-1)*vc$))*(1/n+1/m)/(n+m-2)
  SQRT den2 den
  LET stat = dif$/den
  SCORE stat board
END
'rescale to use as CI for difference in population
     means
MULTIPLY std board board
```

```
ADD obsdif board board
HISTOGRAM board
PERCENTILE board (5 95) interval
PRINT interval obsdif
'If the interval does not include zero, reject the
     null hypothesis.
```

Stata

```
A dummy "treat" variable is used to distinguish the
     two samples
enabling a stratified bootstrap to be used
drop _all
input treat value
1 94   1 38  1 23  1 197  1 99  1 16  1 141
0 52   0 10  0 40  0 104  0 51  0 27  0 146 0 30 0 46
end

capture program drop mydiff
program mydiff, rclass
args treat value
sum ‘value’ if ‘treat’, meanonly
return scalar mean_treat = r(mean)
sum ‘value’ if !‘treat’, meanonly
return scalar mean_contr = r(mean)
return scalar diff = return(mean_treat) -
     return(mean_contr)
end

set seed 1234
bootstrap "mydiff treat value" r(diff),
     strata(treat) nowarn
exit

command:      mydiff treat value
statistic: _bs_1      = r(diff)

Bootstrap statistics            Number of obs = 16
                                Number of strata = 2
                                Replications = 50

-------------------------------------------------------
Variable| Reps Observed Bias   Std. Err.   [95% Conf.
                                           Interval]
--------+----------------------------------------------
  _bs_1 | 50   30.63492 1.508255 30.15216  -29.95811
```

```
                                                    91.22795(N)
          |                                          -28.19048
                                                    85.42857(P)
          |                                          -28.69841
                                                    85.12698(BC)
-------------------------------------------------------------
Note: N  = normal
      P  = percentile
      BC = bias-corrected
```

3.6.1 Underlying Assumptions

At this point we are in a position to distinguish among the various testing procedures—parametric, permutation, and bootstrap—on the basis of their underlying assumptions.

To use any of the tests, the observations must be independent.

To use either a permutation or a parametric test, the observations must be exchangeable.

To use a parametric test, the observations must all come from a distribution of pre-determined form.

3.7 Comparing Variances

Precision is essential in a manufacturing process. Items that are too far out of tolerance must be discarded. An entire production line can be brought to a halt if too many items exceed (or fall below) designated specifications. With some testing equipment, such as that used in hospitals, precision can be more important than accuracy. Accuracy can always be achieved through the use of standards with known values, while a lack of precision may render an entire sequence of tests invalid.

There is no shortage of methods to test the hypothesis that two samples come from populations with the same inherent variability, but few can be relied on. Many methods promise an error rate (significance level) of 5% but in reality make errors as frequently as 8% to 20% of the time. Other methods for comparing variances have severe restrictions.

For example, a permutation test based on the ratio of the sample variances is appropriate only if the means of the two populations are the same or we know their values. If the means are the same, then the labels are exchangeable under the null hypothesis. And if we know the value of the means, then we can make a preliminary transformation that makes the labels exchangeable.

We can derive a permutation test for comparing variances that is free of these restrictions if instead of working with the original observations, we replace them with the differences between successive order statistics and then permute the labels. The test statistic proposed by Aly [1990] is

$$\delta = \sum_{i=1}^{m-1} i(m-i)(X_{(i+1)} - X_{(i)}),$$

where $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(m)}$ are the order statistics of the first sample. That is, $X_{(1)}$ is the smallest of the observations in the first sample (the minimum), $X_{(2)}$ is the second smallest and so forth, up to $X_{(m)}$ the maximum.

To illustrate the application of Aly's statistic, suppose the first sample consists of the measurements 121, 123, 126, 128.5, 129 and the second sample of the measurements 153, 154, 155, 156, 158. $X_{(1)} = 121$, $X_{(2)} = 123$ and so forth.

Set $z_{1i} = X_{(i+1)} - X_{(i)}$ for $i = 1, \ldots, 4$. In this instance, $z_{11} = 123 - 121 = 2$, $z_{12} = 3$, $z_{13} = 2.5$, $z_{14} = 0.5$.

The original value of Aly's test statistic is $8 + 18 + 15 + 2 = 43$. To compute the test statistic for other arrangements, we also need to know the differences $z_{2i} = Y_{(i+1)} - Y_{(i)}$ for the second sample; $z_{21} = 1$, $z_{22} = 1$, $z_{23} = 1$, $z_{24} = 2$.

Only certain exchanges are possible. Rearrangements are formed by first choosing either z_{11} or z_{21}, next either z_{12} or z_{22}, and so forth until we have a set of four differences.

One possible rearrangement is $\{2, 1, 1, 2\}$ which yields a value of $\delta = 24$. There are $2^4 = 16$ rearrangements in all, of which only one $\{2, 3, 2.5, 2\}$ yields a more extreme value of the test statistic than our original observations. With two out of 16 rearrangements yielding values of the statistic as or more extreme than the original, we should accept the null hypothesis. Better still, given the limited number of possible rearrangements, we should gather more data before we make a decision.[9]

MatLab

```
%
% Variance test using Aly's statistic
% This function takes in two data arrays
% and tests their variance using Aly's statistic.
%
% vartest(samp1,samp2,N)
% samp1, samp2 are arrays of data to be compared
% N is the number of times to permute
%

%
% To be run before executing the program
% clear all;
% samp1=[129,123,126,128.5,121];
% samp2=[153,154,155,156,158];
% To run program:
% vartest(samp1,samp2,6400)
%

function [p]=vartest(samp1,samp2,N)
    n=length(samp1);
    m=length(samp2);
```

[9] How much data? See Chapter 5.

```
    diff1=diff(samp1); % get difference vector
                           of samp1
    diff2=diff(samp2); % get difference vector
                           of samp2
    aly0=aly(diff1);   % compute original test
                           statistic
    cnt=0;

    samp=zeros(1,n);

    for i=1:N
        samp=choose(diff1,diff2); % finds sampled
                                      array
        aly1=aly(samp);           % compute aly
                                      statistic of
                                      sample
        if aly1<=aly0
            cnt=cnt+1;
        end
    end
    p=cnt/N;  % p-value


% this function finds the difference vector used in
     aly's calculation
function [d]=diff(samp)
    s=sort(samp);
    n=length(samp);
    d=zeros(1,n-1);

    for i=1:(n-1)
        d(i)=s(i+1)-s(i);
    end

% this function computes the aly statistic with the
     difference vector
function [aly]=aly(diff)
    aly=0;
    n=length(diff);
    m=n+1;
    for i=1:n
        aly=aly+(i*(m-i)*diff(i));
    end

% this function chooses between the ith element if the
     to difference functions.
```

```
function [c]=choose(samp1,samp2)
    rand('state',sum(100*clock));
    n=length(samp1);
    c=zeros(1,n);
    for i=1:n
        if rand<.5
            c(i)=samp1(i);
        else
            c(i)=samp2(i);
        end
    end
```

R Code for Aly's Test Statistic

```
diff=function(samp){
 s=sort(samp)
 l= length(samp)
 d=1:(l-1)
 for(k in 2:l){
  d[k-1]=s[k]-s[k-1]
  }
 return(d)
 }

aly=function(samp){
 stat=0
 l=length(samp)
 for (k in 1:l)
 stat=stat+k*(l+1-k)*samp[k]
 return(stat)
 }

vartest=function(samp1,samp2, NMonte){
 d1=diff(samp1)
 d2=diff(samp2)
 l=length(d1)
 stat0=aly(d1)
 pd=d1
 cnt=0
 for(j in 1:NMonte){
   r=rbinom(l,1,.5)
   for (k in 1:l)pd[k]=ifelse(r[k],d1[k],d2[k])
   if (aly(pd)>=stat0)cnt=cnt+1
  }
 return(cnt/NMonte) #one-sided p-value
 }
```

```
x1 = c(129, 123, 126, 128.5, 121)
y1 = c(153, 154, 155, 156, 158)

vartest(x1,y1,1600)
[1] 0.20125
```

A word of caution: If we are performing a two-sided test, and the p-value, cnt/Nmonte, obtained from the preceding program is less than 0.5, we need to double it, to be consistent with our intention to reject if the test statistic proved to be either extremely large or extremely small. If the p-value is greater than 0.5, this suggests that the second sample is less dispersed than the first. Consequently, we will need to rerun vartest() after modifying the appropriate line of its listing to read

```
if (aly(pd)<=stat0)cnt=cnt+1.
```

Again, we would need to double the resulting p-value.

3.7.1 Unequal Sample Sizes

If our second sample is larger than the first, we have to resample in two stages. Suppose m observations are in the first sample and n in the second, where $m < n$. Select a random subset of m values $\{Y_{*i}, i = 1, \ldots, m\}$ without replacement from the n observations in the second sample. Compute the order statistics $Y_{*(1)} \leq Y_{*(2)} \leq \ldots$, their differences $\{z_{*2i}\}$, and the value of Aly's measure of dispersion for the 2^m possible rearrangements of the combined sample $\{\{z_{1i}\}, \{z_{*2i}\}\}$. Repeat this procedure for all possible subsets of the second sample, combine all the permutation distributions into one, and compare Aly's measure for the original observations with the combined distribution.

Should the total number of calculations appear prohibitive were all subsets used, then use only a representative random sample of them.

EViews Test of Equality of Variance Between Series

```
create u 1 5

series x1
x1.fill 129, 123, 126, 128.5, 121
series y1
y1.fill 153, 154, 155, 156, 158

group a x1 y1

show a.testbtw(var)
```

Test for Equality of Variances Between Series
Date: 03/10/05 Time: 15:49
Sample: 1 5
Included observations: 5

Method	df	Value	Probability
F-test	(4, 4)	3.243243	0.2809
Siegel–Tukey		0.000000	1.0000
Bartlett	1	1.165467	0.2803
Levene	(1, 8)	2.795647	0.1331
Brown–Forsythe	(1, 8)	1.877778	0.2078

Category Statistics

Variable	Count	Std. Dev.	Mean Abs. Mean Diff.	Mean Abs. Median Diff.	Mean Tukey-Siegel Rank
X1	5	3.464102	2.800000	2.700000	5.400000
Y1	5	1.923538	1.440000	1.400000	5.600000
All	10	15.87460	2.120000	2.050000	5.500000

Bartlett weighted standard deviation: 2.801785

3.8 To Learn More

The permutation tests were introduced by Pitman [1937, 1938] and Fisher [1935] and have seen wide application in agriculture [Eden and Yates 1933; Higgins and Noble 1993; Kempthorne 1952], anthropology [Valdes-Perez and Pericliev 1999], aquatic science [Quinn 1987; Ponton and Copp 1997], archaeology [Klauber 1971; Berry, Kvamme, and Mielke 1980, 1983], astronomy [Zucker and Mazeh 2003], atmospheric science [Adderley 1961; Tukey, Brillinger, and Jones 1978; Gabriel and Feder 1969], biology [Daw et al. 1998; Howard 1981], biotechnology [Vanlier 1996; Xu and Li 2003], botany [Mitchell-Olds 1987; Ritland and Ritland 1989], chemistry [vanKeerberghen et al. 1991], climatology [Hisdal et al. 2001], clinical trials [Berger 2000; Grossman et al. 2000; Gail, Tan, and Piantadosi 1988; Howard 1981], cybernetics [Valdes-Perez 1995], diagnostic imaging [Arndt et al. 1996; Raz, Zheng, Ombao, Turetsky 2003], ecology [Busby 1990; Cade 1997; Pollard, Lakhand and Rothrey 1987; Prager and Hoenig 1989], genetics [Thaler-Neto, Fries and Thaller 2000; Gonzalez et al. 2002, Levin 1977; Karlin and Williams 1984; North et al. 2003; Varga and Toth 2003], law [Gastwirht 1992], education [Gliddentracy and Greenwood 1997; Gliddentracy and Parraga 1996], medicine [Feinstein 1973; Tsutakawa and Yang 1974], meteorology [Gabriel 1979; Tukey 1985], neurology [Burgess and Gruzelier 2000; Ford, Colom, and Bland 1989], ornithology [Cade and Hoffman 1993], pharmacology [Plackett and Hewlett 1963], physiology [Boess et al. 1990; Faris and Sainsbury, 1990; Zempo et al. 1996], psychology [Antretter, Dunkel, and Haring 2000; Hollander and Pena 1988; Hollander and Sethuraman 1978; Kazdin 1976, 1980], radiology [Milano, Maggi, and del Turco 2000; Hossein-Zadeh, Ardekani, and Soltanian-Zadeh 2003; Raz et al. 2003], sociology [Tsuji 2000], software engineering [Laitenberger et al. 2000], theology [Witzum, Rips, and Rosenberg 1994], toxicology [Farrar and Crump 1988, 1991], virology [Good 1979] and zoology [Adams and Anthony 1996; Jackson 1990]. Texts dealing with

their application include Bradley [1968], Edgington [1995], Maritz [1996], Noreen [1989], Manly [1997], and Good [2004]. See, also, the review by Barbella, Denby, and Glandwehr [1990]. Early articles include Pearson [1937], Wald and Wolfowitz [1944].

3.9 Exercises

1. How was the analysis of the cell culture experiment described in Section 3.1 affected by the loss of two of the cultures due to contamination? Suppose these cultures had escaped contamination and given rise to the observations 90 and 95; what would be the results of a permutation analysis applied to the new, enlarged data set consisting of the following cell counts?

Treated	121	118	110	90
Untreated	95	34	22	12

   a. Solve this problem by determining both the total number of possible rearrangements and the number of rearrangements that are as or more extreme than the one observed.
   b. Solve this problem by running a Monte Carlo analysis.
2. In the preceding example, what would the result have been if you had used as your test statistic the difference between the sums of the first and second samples? the difference between their means? the sum of the squares of the observations in the first sample? the sum of their ranks?
3. My Acura Integra seems to be less fuel-efficient each year (or maybe it's because I'm spending more time on city streets and less on freeways). Here are my miles per gallon result—when I remembered to record them—for 2002 and 2004:

2002	34.1	32.3	31.7	33.0	29.5	32.8	31.0	32.9		
2004	30.1	30.1	29.5	28.5	29.9	30.6	29.3	32.4	30.5	30.0

   Is my Acura getting less fuel-efficient with each passing year?
4. Cognitive dissonance theory suggests that when people experience conflicting motives or thoughts, they will try to reduce the dissonance by altering their perceptions. For example, college students were asked to perform a series of repetitive, boring tasks. They were then paid either $1 or $20 to tell the next student that the tasks were really interesting and fun. In private, they were asked to rate their own feelings about the tasks on a scale from 0 (dumb, dull, and duller) to 10 (exciting, better than sex).
   Those who received $20 for lying assigned ratings of 3, 1, 2, 4, 0, 5, 4, 5, 1, 3. Those who received only $1, appeared to rate the tasks more favorably, 4, 8, 6, 9, 3, 6, 7, 10, 4, 8. Is the difference statistically significant? If you aren't a psychologist and would like to know what all this proves, see Festinger and Carlsmith [1959].
5. Babies can seem awfully dull for the first few weeks after birth. To us it appears that all they do is nurse and wet, nurse and wet. Yet in actuality, their brains are incredibly active. (Even as your brain, dulled by the need to wake up every few hours during the night to feed the kid, has gone temporarily on hold.) Your child is

learning to see, to relate nervous impulses received from two disparate retinas into concrete visual images. Consider the results of the following deceptively simple experiment in which an elaborate apparatus permits the experimenter to determine exactly where and for how long an infant spends in examining a triangle. (See Salaptek and Kessen, 1966, for more details and an interpretation.)

Subject	Corners	Sides
Tim A	10	7
Bill B	16	10
Ken C	23	24
Kath D	23	18
Misha E	19	15
Carl F	6	18
Hillary G	12	11
Holly H	18	4

a. Is there a statistically significant difference in viewing times between the sides and corners of the triangle?

b. Did you (and should you) use the same statistic as you used to analyze the cognitive dissonance data?

6. How would you go about testing the hypothesis that the fuel additive described in Section 3.3.5 increases mileage by 10%?

7. To compare teaching methods, 10 school children were first taught by conventional methods, tested, and then taught comparable material by an entirely new approach. The following are the test results:

conventional	65	79	90	75	61	85	98	80	97	75
new	90	98	73	79	84	81	98	90	83	88

Are the two teaching methods equivalent in result?
How does this experiment and your subsequent analysis differ from that described in Section 3.4.2?

8. 144 hours after mice were inoculated with Herpes virus type II the following virus titers were observed in their vaginas (see Good, 1979 for complete details):

Saline controls	10000	3000	2600	2400	1500
Treated with antibiotic	9000	1700	1100	360	1

a. Does treatment have an effect?

b. Most authorities would suggest using a logarithmic transformation before analyzing this data because of the exponential nature of viral growth. Repeat your analysis after taking the logarithm of each observation. Is there any difference? Compare your results and interpretations with those of Good [1979].

9. You can't always test a hypothesis. A marketing manager would like to show that an intense media campaign just before Christmas resulted in increased sales. Should he compare this year's sales with last year's? What would the null hypothesis be? And what would be some of the alternatives?

10. An economist developed the following model for personal consumption expenditures C as a function of per capita disposable income D:

$$C[i] = 444 + O.84D[i] + e[i],$$

where the model errors $e[i]$ are independent and identically symmetrically distributed with a mean of zero. Test the hypothesis that the model is correct, that is, the expected value of $C = 444 + O.84D$ against the alternative that the expected value of $C > 444 + O.84D$ with the aid of the following data:

$$D = 6036, 6113, 6271, 6378, 6727, 7027, 7280$$
$$C = 5561, 5579, 5729, 5855, 6099, 6362, 6607.$$

4

Choosing the Best Procedure

In Chapter 3, you were introduced to some practical, easily computed tests of hypotheses that utilized resampling methods. But are they the best tests one can use? And are they always appropriate? What is their relation to the confidence intervals derived in Chapter 2? In this chapter, we consider the assumptions that underlie statistical tests and confidence intervals and look at some of their formal properties: Type I errors and significance level; power, Type II errors, and estimation losses; and robustness to failure of assumptions.

4.1 Why You Need to Read This Chapter

In our example of the missing labels in Chapter 3, we introduced a statistical test based on the random assignment of labels to treatments, a permutation test. We showed this test provided a significance level of 5%, an exact significance level, not an approximation. The test we derived is valid under very broad assumptions. The data could have been drawn from a symmetric normal distribution (see Figure 4.1) or they could have come from some quite different distribution like the exponential that is skewed (Figure 4.2). All that is required for our permutation test to be valid is that under the null hypothesis the distribution from which the data in the treatment group are drawn be the same as the one from which the untreated sample is taken.

This freedom from reliance on numerous assumptions is a big plus. The fewer the assumptions, the fewer the limitations and the broader the potential applications of a test. But before statisticians introduce a test into their practice, they need to know a few more things about it:

How powerful a test is it? That is, how likely is it to pick up actual differences between treated and untreated populations? Is this test as or more powerful than the test they are using currently?

How robust is the new test? That is, how insensitive is it to violations of the underlying assumptions and conditions of an experiment?

What if data are missing as is the case in so many of the practical experiments we perform? Will missing data affect the significance level?

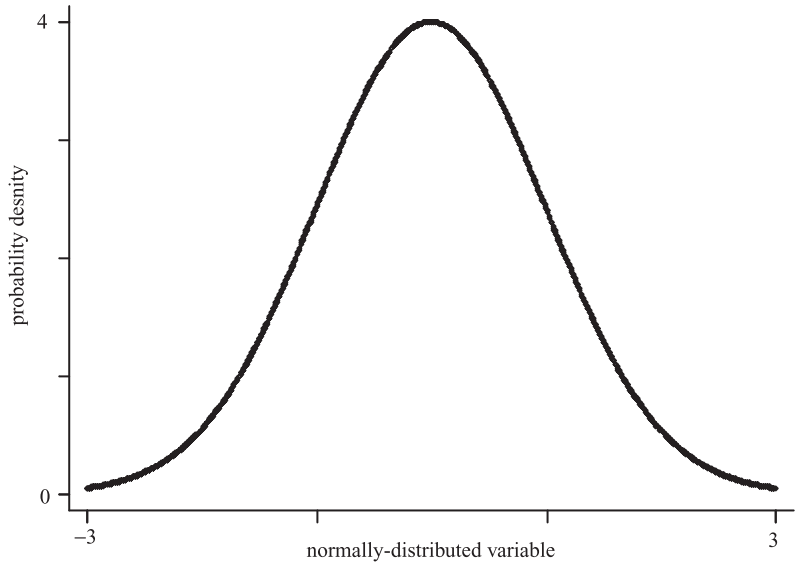


Fig. 4.1. Bell-shaped symmetric curve of a normally distributed population.

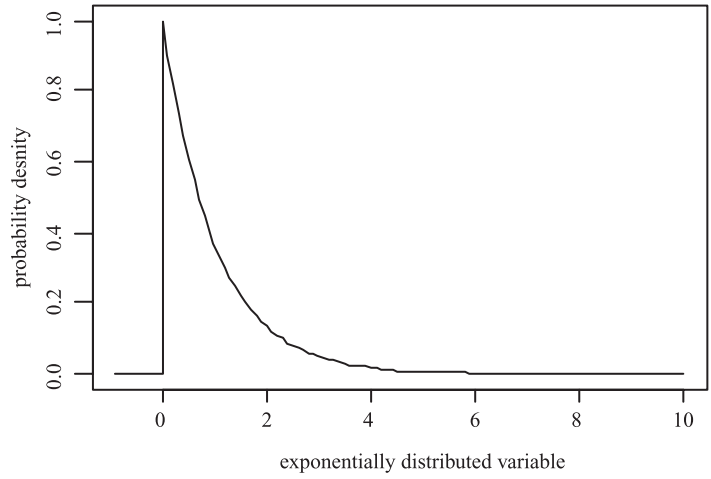


Fig. 4.2. Skewed curve of an exponentially distributed population.

What are the effects of extreme values or outliers? Can we extend the method to other, more complex experimental designs in which there are several treatments at several different levels and several simultaneous observations on each subject?

The balance of this chapter provides a theoretical basis for the answers. And we consider a series of other questions that frequently arise in practice: What is the relationship between confidence intervals and tests of hypotheses? Is it possible for the confidence intervals around each of two means to overlap, yet the difference between the means

to be statistically significant? If multiple tests are performed, will the significance level remain unchanged?

4.2 Fundamental Concepts

The purpose of the present section is merely to serve as a refresher. We expand on the practical implications of these fundamental concepts in subsequent sections.

4.2.1 Two Types of Error

Because variation is inherent in nature, we are bound to make errors when we draw inferences from experiments and surveys, particularly if chance hands us a completely unrepresentative sample. When I toss a coin in the air six times, I can get three heads and three tails, but I can also get six heads. This latter event is less probable, but it is not impossible. Does the best team always win?

Figure 4.3 depicts the results of an experiment in which two groups were each given a "painkiller." The first group got buffered aspirin, the second group received a new experimental drug. Each participant provided a subjective rating of the effects of the drug. The ratings ranged from "got worse," to "much improved," depicted below on a scale of 0 to 5. Take a close look at this figure. Does the new drug represent an improvement over aspirin?

Some of those who took the new experimental drug do seem to have done better. But not everyone. Are the differences we observe in Figure 4.3 simply the result of chance? Or do they represent a true treatment effect? If it's just a chance effect and we opt in favor of the new drug, we've made a *Type I error*. We also make an error, a *Type II error,* if we decide there is no difference and the new drug really is better. These decisions and the effects of making them are summarized in Table 4.1.

We distinguish the two types of error because they have quite different implications. For example, suppose we are a manufacturer hoping to bring to market a new drug with the potential for curing a life-threatening disease. We test the drug for possible dangerous side effects, detect them, and determine not to market the drug. Now, suppose

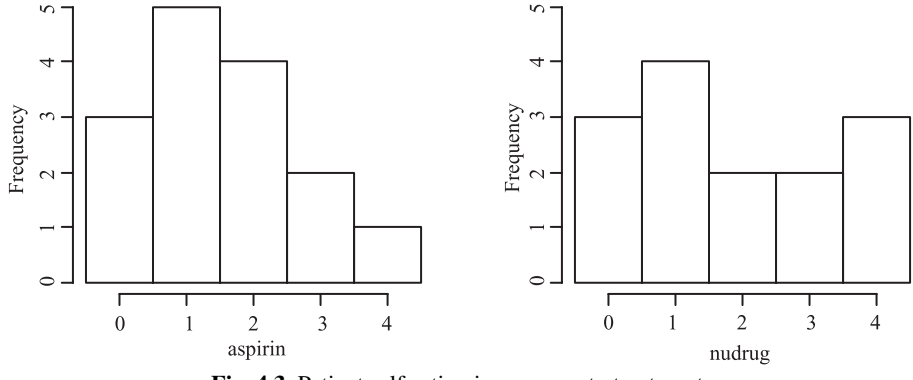


Fig. 4.3. Patient self-rating in response to treatment.

Table 4.1. Decision Making Under Uncertainty

	Our Decision	
The Facts	No Effect	Effect Present
No Effect		Type I error
Effect Present	Type II error	

we are wrong and the new drug is completely safe. Quick, what type of error have we made?[1]

Our losses in this instance are evident. We've wasted the money we've spent so far in testing and developing our drug. And we've forgone potential profits. Not incidentally, the public has suffered in being denied a potential cure.

A Type II error, consisting of the failure to detect a dangerous side effect, would have even more disastrous results. Not only would we continue to pour money into the development of a dangerous drug, foregoing investing in far more rewarding efforts, but down the line, if and when patients die, there are certain to be lawsuits.

4.2.2 Losses

A problem is a statistical one when the investigator is not in a position to say that an observation will take a specific value, but only that there is a probability between 0 and 1 that it might take it. A statistical problem is defined by four elements:

1. The observations,
2. The distribution of the variables in the populations from which the observations are drawn,
3. The set of possible decisions one can make on analyzing the observations,
4. The *loss* expressed in dollars, lives, or some other quantifiable measure, that results when we make a particular decision. This loss will depend both upon the decision we make and upon the underlying state of nature.

In this text, we've so far limited ourselves to two-sided decisions in which either we accept a hypothesis H and reject an alternative K, or we reject the hypothesis H and accept the alternative K. But an all-or-nothing loss function is the exception, not the rule.

For example, suppose you've developed a new drug to relieve anxiety and are investigating its side effects. Does it raise blood pressure? You do a study and find the answer is no, your drug raises systolic blood pressure only an average of 1 mm. What is the cost to the average patient? Not much, negligible.

Now, suppose your new drug actually raises blood pressure an average of 10 mm. What is the cost to the average patient? to the entire potential patient population? to your company in lawsuits? One thing is sure, the cost of making a Type II error and the resultant losses you will be subject to depend on the magnitude of that error.

Typically, our *losses* depend on some function of the difference between the true (but unknown) value of the parameter we are trying to estimate and our best guess

[1] A Type I error because we mistook a chance result for a real effect.

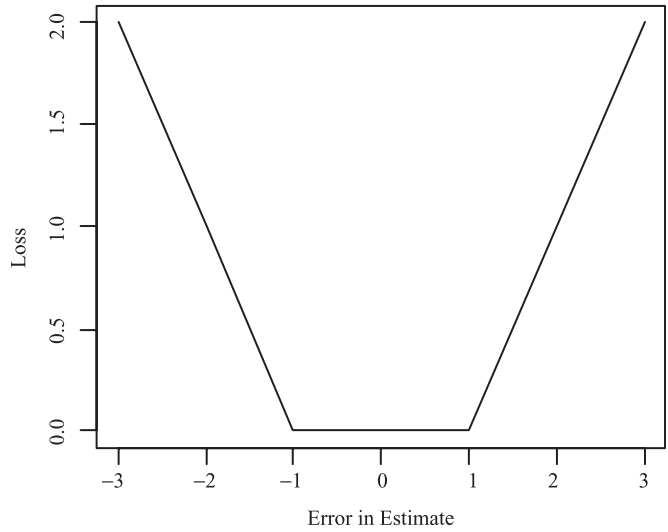


Fig. 4.4. Loss as a function of error in estimate.

(point estimate) of its value. Typical loss functions increase as the difference between the actual value and our estimate increases. The loss may be proportional to the absolute value of this difference or to its square, it may be a continuously increasing function of the difference, or it may increase in steps. Figure 4.4 depicts a loss function that is zero unless the error in estimation exceeds 1.

4.2.3 Significance Level and Power

In selecting a statistical method, statisticians work with two closely related concepts, significance level and power. The *significance level* of a test is the allowable upper limit on the probability of making a Type I error. We should always establish its value *before* we perform an analysis.

The *power* of a test is the probability of detecting a deviation from the hypothesis, of deciding on the alternative when the alternative is the correct choice. Thus, it is the probability of *not* making a Type II error. Power $= 1 - \Pr\{\text{Type II error}\}$.

The ideal statistical test would never incur a Type I error and would have a power of 1 or 100%. But unless we are all-knowing, this ideal cannot be realized. In practice, we fix our significance level at the largest value we feel comfortable with, and choose a statistic that maximizes or comes closest to maximizing the power.

The power of a test depends on all of the following:

1. The magnitude of the effect (the signal)
2. The amount of variation in the data (the noise)
3. The significance level
4. The sample size
5. The method used for testing.

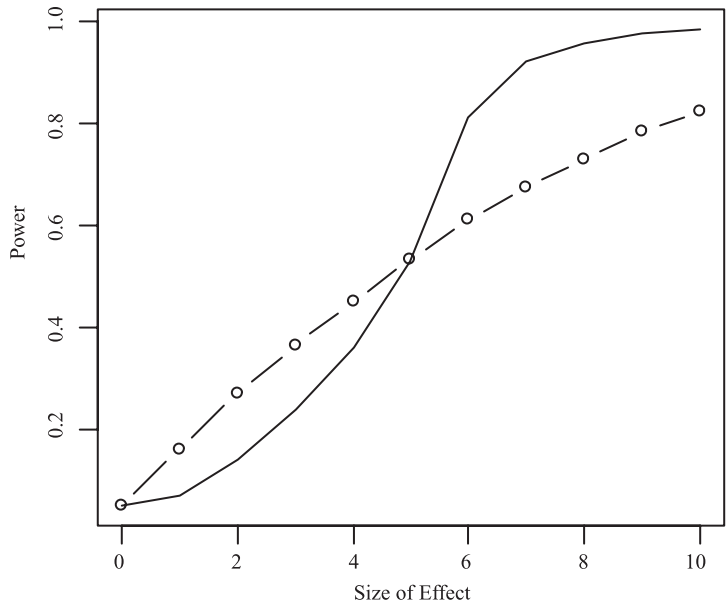


Fig. 4.5. Comparing power curves of two hypothesis tests.

To increase the power of our tests we can

1. Increase the sensitivity of our methods of measurement
2. Reduce the amount of variation in the data (see the next chapter for suggestions)
3. Allow a higher frequency of Type I errors
4. Increase the sample size
5. Use the most powerful statistical method for testing.

The ideal method will be the one that yields a test that is most powerful (for a given significance level and sample size) against all possible alternatives. Unfortunately, a testing procedure that is optimal for one value of the parameter might not be optimal for another. This situation is illustrated in Figure 4.5 where two power curves cross each other. In Section 4.4, we consider conditions under which a uniformly most powerful test might exist.

4.2.4 What Significance Level Should I Use?

Choice of significance level and power are determined by the environment in which you work.

Most scientists simply report the observed p-value leaving it to their peers to decide for themselves what weight should be given the results.

A manufacturer preparing to launch a new product line or a pharmaceutical company conducting a research for promising compounds typically adopts a three-way decision procedure: If the observed p-value is less than 1%, they go forward with the project.

If the p-value is greater than 20%, they abandon it. And if the p-value lies in the gray area in between, they arrange for additional surveys or experiments.

A regulatory commission like the FDA that is charged with oversight responsibility must work at a fixed significance level, typically 5%. In the case of unwanted side effects, the FDA may also require a certain minimum power. The choice of a fixed significance level ensures consistency in both result and interpretation as the agency reviews the findings from literally thousands of tests.

4.3 Confidence Intervals

There is a direct connection between confidence intervals and the acceptance regions of our tests. Recall that the *acceptance region*, $A(\theta_o)$ of a test is the set of values of the test statistic $T[X]$ for which we would accept the hypothesis H: $\theta = \theta_o$. Its complement is called the rejection region. A *confidence interval*, $S(X)$ is a set of values of the parameter θ for which given the set of observations $X = \{x_1, x_2, \ldots, x_n\}$ and the statistic $T[X]$ we would accept the corresponding hypothesis.

The boundaries of a confidence interval are random variables that depend upon the sample X. The boundaries of acceptance regions are fixed. p-values are random variables that depend upon the sample X; significance levels are fixed values.

Suppose $A(\theta')$ is a $1 - \alpha$ level acceptance region for testing the hypothesis $\theta = \theta'$, that is, we accept the hypothesis $\theta = \theta'$ if our test statistic T belongs to the acceptance region $A(\theta')$ and reject it otherwise. Suppose $S(X)$ is a $1 - \alpha$ level confidence interval for θ based on the set of observations $X = \{x_1, x_2, \ldots, x_n\}$. Then, $S(X)$ consists of all the parameter values $\theta*$ for which $T[X]$ belongs to the acceptance region $A(\theta*)$.

$\Pr\{S(X)$ includes θ_o when $\theta = \theta_o\} = \Pr\{T[X] \varepsilon A(\theta_o)$ when $\theta = \theta_o\} \geq 1 - \alpha$.

Suppose our hypothesis is $\theta = 0$ and we observe X. Then we accept this null hypothesis if and only if our confidence interval $S(X)$ includes the value 0.

Note that we may start with a confidence interval and use it to test a variety of hypotheses or start with a set of tests and use them to construct confidence intervals. One-sided hypotheses yield one-sided confidence intervals and vice versa.

To illustrate these points, suppose that we want to test a one-sided hypothesis concerning the proportion of the population p who favor our candidate. We ask five individuals chosen at random whom they favor. The result is a binomial random variable with 5 trials and probability p of favoring our candidate.[2]

The chart in Figure 4.6 can be used either to obtain an upper bound at the 5% significance level for the acceptance region for the hypothesis that p is less than or equal to some predetermined value or to obtain a lower bound for the 95% confidence interval for p. To obtain an upper bound for the acceptance region, find the value p on the horizontal axis corresponding to the hypothesis you are testing, go to the nearest point on the curve connecting the points on the diagram, then go left to the nearest point on the vertical axis. To obtain a lower confidence bound, start on the vertical axis with

[2] That is, it will be a binomial variable if the answers we receive are independent of one another and each individual has the same probability p of favoring our candidate. In the next chapter, we revisit this problem and consider methods to ensure the validity of these assumptions.

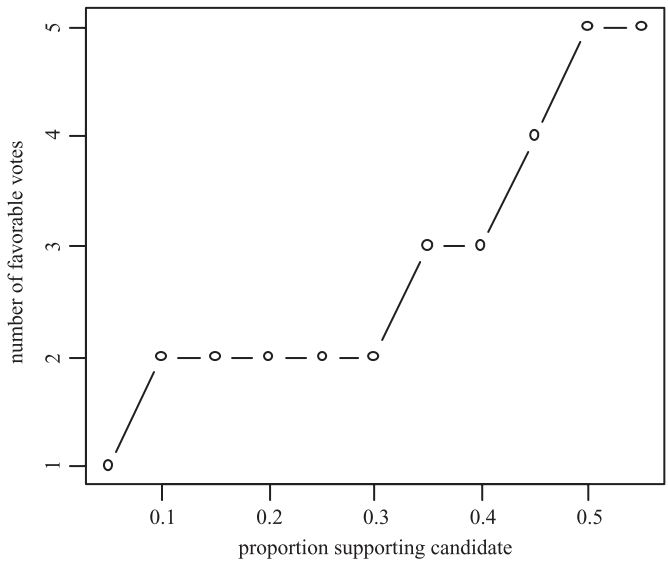


Fig. 4.6. Chart equating 95% confidence bounds and acceptance regions for a binomial.

the number of favorable votes observed, go to the nearest point on the curve connecting the points, then drop a vertical line to the horizontal axis.

4.3.1 Interpreting the Confidence Interval

No value in the confidence interval is more likely than any other value. Nor can we be sure that our confidence interval covers the true value. The level of confidence that we cite refers to the long-run behavior of our procedures. In the long run, an 80% confidence interval will cover the true value of the parameter at least 80% of the time.

That is, it will do so if the underlying statistical procedure is exact. Confidence intervals based on permutation tests are exact only if when the null hypothesis is true the observations are exchangeable. Confidence intervals based on parametric tests are exact only if the observations are independent and come from the same pre-specified parametric distribution.

4.3.2 Multiple Tests

When we construct a K% confidence interval, it means we believe that, over time, K% of the confidence intervals we construct will contain the true value of the parameter. When we test hypotheses at the $(100-K)\%$ significance level, it means we believe that over a long period of time, we will reject the hypothesis when it is is true in $(100-K)\%$ of the cases. The probability that we will not make *any* Type I errors in N independent constructions or tests is on the order of K raised to the Nth power.

The preceding statement is true whether we are concerned with the analysis of a single experiment or of a lifetime's worth of experiments. If $K = 95\%$ or 0.95, then

the probability that we won't reject a single hypothesis in error in 20 independent experiments is $(0.95)^{20}$ or 36% considerably less than the 95% associated with any single test or confidence interval.

4.4 Which Test Should Be Used?

The type of test to be employed depends on the nature of the underlying data and the distribution of values in the population(s) from which the data is drawn. In this section, we consider in turn the various types of data and how they ought to be approached, how to recognize data that comes from specific parametric distributions, distribution-free procedures, transformations that can make our data more amenable to analysis, and brief guidelines to procedure selection.

4.4.1 Types of Data

Ordinal Statistics such as the minimum, maximum, median, and other quantiles make sense only if the data is *ordinal*, that is, if it can be ordered from smallest to largest. Clearly height, weight, number of voters, and blood pressure are ordinal. So are the answers to survey questions such as "How do you feel about President Bush?" which range in intensity from "He's O.K," to "Hanging's too good for him."

Ordinal data can be subdivided into metric and nonmetric data.

Metric Data like heights and weights can be added and subtracted. We can compute the mean as well as the median of such *metric* data. We can further subdivide metric data into observations like time that can be measured on a *continuous* scale and counts such as "buses per hour" that are *discrete*. Chapters 2 and 3 are devoted to the analysis of metric data.

But what is the average of "he's destroying our country" and "he's no worse than any other politician?" Such preference data is ordinal in that it may be ordered, but it is *not* metric.

Many times, in order to analyze nonmetric ordinal data, statisticians will impose a metric on it—assigning, for example, weight 5 to "Bush is destroying our country" and weight 1 to "Bush is no worse than any other politician." Such analyses are suspect, for another observer using a different set of weights might get a quite different answer.

Categorical The answers to other survey questions are not so readily ordered. For example, "What is your favorite color?" Oops, bad example, as we can associate a metric wavelength with each color. Consider instead the answers to "What is your favorite breed of dog?" or "What country do your grandparents come from?" The answers to these questions fall into nonordered categories. Pie charts and box plots are used to display such *categorical* data and contingency tables, considered in Chapter 6, are used to analyze it. A scatter plot of categorical data would not make sense.

4.4.2 Assumptions

Every hypothesis test and estimation procedure is based on certain assumptions. For example, that the sample was selected at random from the population, and the observations

are independent of one another. These assumptions must be satisfied in order to achieve a desired significance level.

Underlying the *bootstrap* are two assumptions: i) if the null hypothesis is true then all observations in the sample(s) come from populations that have the same value of the parameter that is being estimated and ii) the observations are independent of one another.

Underlying all *permutation tests* is the assumption that if the null hypothesis is true, all observations in the sample (or subsample) are exchangeable. Observations are exchangeable if they are either

a. independent and identically distributed,
b. identically normally distributed with the same mutual correlations,
c. obtained by random sampling without replacement from the same finite population.

Sometimes a simple transformation will ensure observations are exchangeable, a topic we return to in Section 4.4.4.

Parametric procedures require not only that the observations be independent and identically distributed, but that this distribution have a specific form, such as the binomial, Poisson, normal, or exponential. As parametric procedures often lead to the most powerful tests, guidelines for recognizing common parametric distributions are provided in the next section.

4.4.3 Recognizing Common Parametric Distributions

Binomial The number of successes in n independent trials such that the probability of success is the same in each trial has the *binomial distribution*. The number of binomial trials required to reach a specific number of successes has the *negative binomial distribution*. Confidence intervals can be obtained by reference to these distributions. If the probability of success varies on the basis of other known variables, a general linear model or logistic regression may be more appropriate for use in estimating population parameters. Methods for comparing two sets of binomial outcomes are given in Section 5.1.

Poisson Suppose we count the frequency with which a specific event occurs in a given time interval or a given area. If the counts in nonoverlapping intervals are independent of one another, and the probability that more than one event will occur in an extremely small interval is vanishingly small, then such counts will have a *Poisson distribution* and should be analyzed by parametric means. To compare two Poisson distributions use the binomial.

Normal If an observation is the cumulative result of a large number of factors that each make only a small contribution to the total, it will have a *normal distribution*. Thus, the mean of a large number of observations taken from any distribution will have a near-normal distribution. This same comment does not apply to other sample statistics.

Normally distributed data should be analyzed by parametric means in most cases.

Most real-world data comes from a mixture of normal distributions. To make such data amenable to analysis, an attempt should be made to resolve the mixture into its components, by blocking, for example. (See Chapter 5.)

The distribution of the ratio of two normal distributions is a heavy-tailed far-from-normal distribution. Data concerning ratios should only be analyzed using resampling methods.

Time to Event If an event comes about as a result of a single failure—a Poisson event—the time between events will have an *exponential distribution*. If an event can only come about as a result of multiple failures, the time between events will have a *chi-square distribution*. Such data should be analyzed by parametric means. The times at which subjects are lost to a study, die, or fail to respond to treatment are often treated as *counting processes*, for which the data is recorded in pairs (n, t) with n a count and t a point in time. Counting processes may be analyzed by either parametric or resampling methods.

4.4.4 Transformations

We may need to transform our data before we analyze it for any or all of the following reasons:

1. Permit testing a null hypothesis (permutation tests only)
2. Obtain a more symmetric distribution
3. Equalize variances (and, hopefully, the distributions),
4. Reduce the impact of outliers.

Permutation tests can be used only to test a null hypothesis. But this requirement need not limit our ability to use permutations. Depending on the specifics of the testing problem, we may obtain the desired result merely by switching hypothesis and alternative or via simple subtraction. For example, suppose we want to test the hypothesis that the expected value of Y is greater than $AX + B$ against the alternative that the expected value is less than or equal to $AX + B$. After making I pairs of observations $(y_1, x_1), (y_2, x_2), \ldots, (y_I, x_I)$, we form a set of transformed observations $z_i = y_i - Ax_i - B$, for $i = 1, \ldots, I$ and then form a one-sided permutation test of the null hypothesis that the expected value of Z is 0 against the alternative that it is greater than zero.

Recall that more accurate results can be obtained if the bootstrap distribution is symmetric and that to perform a one-sample permutation test the observations must come from a symmetric distribution. If we anticipate changes measured in percentage rather than absolute terms, a preliminary logarithmic transformation will often produce a more symmetric distribution.

The key issue with permutation tests is whether, under the null hypothesis of no differences among the various experimental or survey groups, we can exchange the labels on the observations without affecting the results. Suppose we have taken steps to ensure our sampling method is representative, and that our observations are independent of one another. Certain equalizing transformations then can help us to ensure the variances (and, hopefully, the distributions) of the variables we are comparing are the same.

If the variances are expected to be proportional to the mean, take the logarithms of the observations. If the standard deviations are expected to be proportional to the mean as is the case with Poisson variables, take the square roots of the observations.

With proportions, such as the proportion of binomial trials that result in success, use the arcsin transformation. If there is no obvious transformation, then ranks may be employed.

Not infrequently, we encounter values among our observations that seem to stand out from the rest, a "19" for example when all the rest of the observations lay between 1 and 6. Is the "19" a typographical error? Should it be 1.9 instead, or simply 1, or 9? When we can't be sure, replacing the observations by their ranks will at least reduce the impact of exceptional values, while not reducing their effect completely in case they're real. Replacing observations by their ranks does not affect the significance level if done routinely and has the same effect with large samples on the power of a two-sample comparison as reducing the sample size from 100 to between 86 and 96 [Bradley 1968].

We can obtain a still more powerful test if we have reason to suspect that most of the observations—the few excepted—come from a distribution of a specific form. Before we start, we replace the original observations by the corresponding percentiles of that distribution. For example, if we think the observations are from a normal distribution, we would replace all the observations, including the outliers, by the corresponding normal scores. The resulting hypothesis test will be more powerful against shift alternatives from a normal distribution than would one based on ranks.

4.4.5 Distribution-Free Tests

The significance level and power of a test depend upon how the variables we observe are distributed. Does the population distribution follow a bell-shaped normal curve with the most frequent values in the center? or is the distribution something quite different? To protect our interests, we may need to require that the Type I error be less than or equal to some predetermined value for all possible distributions. When applied correctly, with the assumption of exchangeability satisfied, permutation tests always have this property. The significance level of a test based on the bootstrap or a parametric test is dependent on the underlying distribution.

In practice, we seldom know the distribution of a variable or its variance. Suppose we wish to test the hypothesis "X has mean 0." This compound hypothesis includes several simple hypotheses such as "X is normal with mean 0 and variance 1," "X is normal with mean 0 and variance 2," and "X is a gamma distribution with mean 0 and four degrees of freedom."

A test is said to be *exact* with respect to a compound hypothesis if the probability of making a Type I error is the same for each and every one of the possibilities that make up the hypothesis. A test is said to be *conservative* if the Type I error never exceeds a pre-specified level. Obviously, an exact test is conservative though the reverse may not be true.

The importance of an exact test cannot be overestimated, particularly a test that is exact regardless of the underlying distribution. If a test that is nominally at level 5% is actually at level 10%, we may be in trouble before we start: If 10%, the risk of a Type I error is greater than we are willing to bear. If 2%, then our test is suboptimal, and we can improve on it by enlarging its rejection region.

A test is said to be unbiased and of level α providing its power function $\beta(\theta)$ satisfies the following two conditions:

1. $\beta(\theta)$ is conservative; that is, $\beta(\theta) \leq \alpha$ for every value of the parameter θ that satisfies our hypothesis.
2. $\beta(\theta) > \alpha$ for every value of the parameter θ that is an alternative to the hypothesis.

In other words, a test is unbiased if it is more likely to reject a false hypothesis than a true one. A confidence interval is said to be unbiased if the correct value of the parameter is more likely to be covered by it than an incorrect value. As we shall see in the next section, unbiased tests lead to unbiased confidence intervals.

Faced with some new experimental situation, our objective always is to derive a uniformly most powerful unbiased test if one exists. Permutation tests correctly applied are exact, unbiased, and distribution-free.

4.4.6 Which Test?

If the observations come from a distribution of known form, it is safe and preferable to use a parametric test. Otherwise, if you can exchange the labels on the observations without affecting the results, it is safe and preferable to use a permutation test. As the permutation test is distribution free and is almost as efficient and powerful as the most powerful parametric test if one exists, use a permutation test if uncertain about the distribution of the underlying values.

Use of the bootstrap is generally reserved for two situations:

1. Confidence intervals for statistics that are not distribution parameters.
2. When neither a permutation nor a parametric test is available or applicable.

4.5 Summary

In this chapter, you learned that power, sample size, and significance level are interrelated. You learned that your choice of test statistic will depend on the hypothesis, the alternative, the loss function, and the type of test you employ. You learned the value of exact unbiased tests. You learned the assumptions underlying and the differences among test of hypotheses based on the bootstrap, the permutation test, and parametric distributions.

4.6 To Learn More

A formal discussion of risk theory is found in Lehmann [1986] and Good [2004]. For a further discussion of exchangeability, see Lehmann [1986, p. 231], Koch [1982], and Draper et al. [1993]. Examples of power calculations for resampling methods are given in Oden [1991], Keller-McNulty and Higgens [1987], and Hall and Titterington [1989]. For a discussion of the relative advantages of variable versus fixed significance levels, see Kempthorne [1966].

See Good and Hardin [2003] for a further discussion of the problems associated with post hoc hypotheses and multiple tests.

To analyze counting processes, download the R survival5 library or make use of Stata's many functions for survival analysis. For a discussion of applicable resampling methods, see Andersen et al. [1993] and Therneau and Grambsch [2000].

Singh [1981] was the first to demonstrate the advantages of the bootstrap over large-sample parametric approximations. For other large-sample results, see Good [2004].

4.7 Exercises

1. **a.** Sketch the power curve (power versus difference in means) for one of the two-sample comparisons described in Chapter 3. (You already know two of the values for each power curve. What are they?)
   b. Using the same set of axes, sketch the power curve of a test based on a much larger sample.
2. Suppose that without looking at the data you
   i) always reject; ii) always accept; or iii) use a chance device so as to reject with probability 16.6%.
   For each of these three tests, determine the power and the significance level.
   Are any of these three tests exact? unbiased?
3. True or False? Tests should be designed so that
   a. the risk of making a Type I error is low;
   b. the probability of making a Type II error is high;
   c. the null hypothesis is likely to be rejected;
   d. economic consequences of a decision are considered.
4. True or False? The expected loss from using a particular statistical procedure will depend on
   a. the probability the procedure will lead to a wrong decision;
   b. the losses associated with a wrong decision.
5. Many statistics packages print out the results of several tests—for example, in the two-sample comparison, SAS prints out the results of both the t-test and the Wilcoxon. What effect might this have on the resultant significance level?
6. Suppose you have two potentially different radioactive isotopes with half-life parameters λ_1 and λ_2, respectively. You gather data on the two isotopes and, taking advantage of a uniformly most powerful unbiased permutation test, you reject the null hypothesis that the half lives are the same in favor of the one-sided alternative that the first isotope has a longer half life. What are you or the person you are advising going to do about it? Will you need an estimate of the ratio of the two half lives? What estimate will you use?
7. Review some of the hypotheses you tested in the past. Distinguish your actions after the test was performed from the conclusions you reached. (In other words, did you do more testing? rush to publication? abandon a promising line of research?) What losses were connected with your actions? Should you have used a higher/lower significance level? Should you have used a more powerful test or taken more or fewer observations? And, if you used a parametric test like Student's t or Welch's z, were all the assumptions for these tests satisfied?

8. Late for a date, you dash out of the house, slipping a bill into your wallet. Was it a $2 bill or a $20? Too late now. At the movie theater, your friend slips a $2 bill into your hand, "for my share." (A share of what? the popcorn?) You put this bill away in your wallet, too. As you approach the ticket window, you decide to perform a simple statistical test. You'll take a quick look at one of the bills. If it's a $2 bill, you'll accept the hypothesis that both bills are $2. Otherwise, you'll reject this hypothesis. What is the significance level of your test? the power?

9. Given a choice of a permutation test based on the original scores, a permutation test based on ranks, a bootstrap, and the best and most appropriate parametric test, which would you use in the following situations and why?
In cases a–e, you want to test a hypothesis about the median of a population:

- **a.** You have 10 observations from a normal distribution.
- **b.** You have 10 observations from an almost-normal distribution.
- **c.** You have 10 observations from an exponential distribution.
- **d.** You have 10 observations from an almost-exponential distribution.
- **e.** You have 50 observations from an almost-exponential distribution.

In cases f–h, you want to compare two populations.

- **f.** You have two samples of size 10 taken from normal distributions with the same variance and want to compare medians.
- **g.** You have two samples of size 10 taken from similarly shaped distributions and want to compare medians. One problem: all the data are in the hundreds, except for one observation which is 10.1. Or is that decimal point a flaw in the paper?
- **h.** You have two samples of size 25 and want to compare variances.

10.
- **a.** Your lab has been gifted with a new instrument offering 10 times the precision of your present model. How might this affect the power of your tests? their significance level? the number of samples you'll need to take?
- **b.** A directive from above has loosened the purse strings so you now can take larger samples. How might this affect the power of your tests? their significance level? the precision of your observations? the precision of your results?
- **c.** A series of lawsuits over those silicon implants you thought were harmless has totally changed your company's point of view. How might this affect the power of your tests? their significance level? the precision of your observations? the number of samples you'll take? the precision of your results?

11. Turn to the literature of your field and consider some of the estimates that have been made. Which do you feel is the most appropriate loss function in each case, the absolute difference between the estimate and the true value or the square of this difference?
If hypotheses were tested, were all test assumptions satisfied?

12. Imagine your company, a manufacturer of industrial supplies, plans to launch a new product to sell to existing customers. Development costs would be approximately one million dollars. Your anticipated profit is $1000 per unit sold. You have a customer base of 10,000. What percentage would you need to sell to break even? Of course the actual percentage you sell might be greater or less than this amount. If you decide to manufacture the product, what sort of losses might you experience? If you decide against development, what profits might you be surrendering?

You decide to do a survey of your customers and use this survey to estimate the percentage you will sell. If p is your estimate and π the actual value of this percentage, construct a graph of losses in dollars versus the estimate error $\pi - p$.

13. Your company policy is to insist on a 99% compliance policy in the components you purchase. That is, no more than 1% of the units can be out of compliance. How many units would you need to examine to be sure of rejecting a lot shipped to you with 2% defective components? Examining all of them would get the job done, but there must be a less expensive way. Make a chart for $n = 100, 200$, and so forth, in which you list i) the maximum number of tested units that can be out of compliance), ii) the resulting significance level when 99% of the components are actually in compliance, and iii) the power when only 98% of the units are actually in compliance.

14. Take a second look at the two approaches to comparing teaching methods described in Chapter 3. Which approach ought to yield the more powerful test?

15. Formulate hypotheses and alternatives for comparing the billing practices of the four hospitals considered in Exercise 6 of Chapter 2. What will you compare: means? medians? variances? frequency distributions?

16. The seeming paradox of overlapping confidence intervals is often raised on statistics mailing lists and on-line bulletin boards. The difference between two sample means was found to be significant, yet the confidence intervals for the individual means overlap. Explain this seeming paradox by relating hypothesis tests to confidence intervals and vice versa.

5

Experimental Design and Analysis

Failing to account for or balance extraneous factors can lead to major errors in interpretation. In this chapter, you learn to block or measure all factors that are under your control and to utilize random assignment to balance the effects of those you cannot. You learn to design experiments to investigate multiple factors simultaneously, thus obtaining the maximum amount of information while using the minimum number of samples.

5.1 Separating Signal from Noise

Have you ever tried to hear a baseball game late at night on the radio when the station's signal kept fading in and out? Or pick up a distant TV station with rabbit ears? Or try to understand what the waiter is saying about the evening's specials when some jerk at the next table is yammering on his cell phone?

All the problems we examine in this text would be straightforward if it weren't for the noise (variation) life has imposed on the signal (relationships) that we're trying to make sense of. Fortunately, in many cases—not all—properly applied statistics can reduce the noise and sharpen the picture. They also can help us avoid mistaking noise for signal.

Eliminating or reducing extraneous variation is the first of several preventive measures we use each time we design an experiment or survey. We strive to conduct our experiments in a biosphere with atmosphere and environment totally under our control. And when we can't—which is almost always the case—we record the values of the extraneous variables to use them either as blocking units or as covariates.

5.1.1 Blocking

Although the significance level of a permutation test may be distribution free, its power depends on the underlying distribution. The more variable our observations, the less the power of our tests and our ability to detect the alternative. A major source of this variation is the lack of homogeneity in the underlying populations. Some soils like

that of my back yard consist of nonporous infertile clay, while others across the street boast of a rich sandy loam. College classes in the United States, Canada, and Australia that once consisted solely of white upper-class males 17–23 years of age are now a heterogeneous mixture of many races, genders, and age groups. One way to reduce the variance of our experiments and surveys is to *block* them, that is, to subdivide the population into more homogeneous subpopulations and to take separate independent samples from each.

Suppose you were designing a survey on the effect of income level on the respondent's attitude toward compulsory pregnancy. Obviously, the views of men and women differ markedly on this controversial topic. To reduce the variance and increase the power of your tests, block the experiment, interviewing and reporting on men and women separately. A physician would want to block by gender in a medical study, and probably by age, race, concurrent medications, and pre-existing conditions as well. An agronomist would want to distinguish among clay, sand, and sandy-loam soils.

Whenever a population can be subdivided into distinguishable subpopulations, you can reduce the variance of your observations and increase the power of your statistical tests by blocking or stratifying your sample.

Suppose we have agreed to divide our sample into two blocks—one for men, one for women. If this were an experiment, rather than a survey, we would then assign subjects to treatments separately and independently within each block. In a study that involves two treatments and 10 experimental subjects, four men and six women, we would first assign the men to treatment and then the women. We could assign the men in any of 4 choose 2 or 6 ways and the women in any of 6 choose 3 or 20 ways, for a total of $6 \times 20 = 120$ possible random assignments in all.

When we come to analyze the results of our experiment, we use the permutation approach to ensure that we analyze in the way the experiment was designed. Our test statistic is a natural extension of that used for the two-sample case: The sum of the sums of the observations in the first sample of each block. In symbols $\sum_{j=1}^{\text{blocks}} \sum_{i=1}^{n[1]} X_{j1i}$ where X_{j1i} is the ith observation in the first sample in the jth block.

The design need not be balanced, as this test statistic is a sum of several independent sums. Unequal sample sizes resulting from missing data or an inability to complete one or more portions of the experiment will affect the analysis only in the relative weights assigned to each subgroup.

Blocking is applicable to any number of subgroups; the extreme case in which every pair of observations forms a distinct subgroup is matched pairs, which we studied in Section 3.5.

5.1.2 Analyzing a Blocked Experiment

We consider two quite different cases in this section:

1. Combining data to obtain improved estimates.
2. Comparing samples from two populations.

How Many Controls?

Every experiment involves at least two groups of subjects—those that took the drug and those who took the placebo (the control group), or those that took the new drug versus those that took the old.

For the self-assured, the smug, and the person with something to sell, controls may be an unnecessary luxury. But the day I started at the Upjohn Company, my boss John R. Schultz recommended I use twice as many subjects in each control group as the number devoted to experimental treatment, and that I use two types of controls, positive and negative.

For a study on Motrin, the positive control was aspirin (the best treatment at the time) and the negative or neutral control was harmless filler used to give aspirin tablets their compact shape. The reason we used so many subjects in each control group? "Life is full of surprises," said John, "you leave work one day whistling, the next day you're back with a head cold. Most of the time these negative effects have nothing to do with the treatment. By using many control subjects, you ensure the normal wear and tear of ordinary life will be detected and accounted for and won't be falsely associated with the treatment you're trying to investigate."

Further proof of John's wisdom came several years later with the controversy over silicon implants. Women who had the implants suffered from a wide variety of complaints. Dow Corning, the manufacturer, didn't have the data on control groups to show that such complaints might be pure coincidence and was forced to pay millions of dollars in compensation.

Combining Data to Obtain Improved Estimates

Occasionally, we find ourselves in a position to combine data from several sources; for example, after we install a new more precise measuring instrument. In this latter case, we would want to give greater weight to the more recent and more precise observations. Importance sampling provides the answer. In the following *R* listing adapted from the code of Section 2.2.2, the weight given to each sample is inversely proportional to its standard deviation.

```
#Combine and weight samples to obtain a CI for
     population median
first=c(3.87, 4.48, 5.00, 3.60, 3.89, 2.50, 4.72, 3.15,
     2.94, 4.65, 4.59, 3.60, 3.67, 3.16)
second=c(8.67, 2.64, 4.01, 4.31, 2.75, 2.44, 2.95,
     7.86, 3.36, 6.97, 1.97, 5.31)
#record group sizes
n1 = length(first)
n2=length(second)
n=n1+n2
#record standard deviations
s1=sqrt(var(first))
s2=sqrt(var(second))
```

```
ratio= (n1/s1)/(n1/s1 + n2/s2)
#set number of bootstrap samples
N =100
stat = numeric(N)  #create a vector in which to store
                        the results
                   #the elements of the vector will be
                        numbered from 1 to N
#Set up a loop to generate a series of bootstrap
     samples
for (i in 1:N){
    choice = sort(runif (n,0,1))
    cnt=0
    while (choice[cnt+1]<ratio) cnt=cnt+1
    #bootstrap sample counterparts to observed samples
         are denoted with "B"
    firstB= sample (first, cnt, replace=T)
    secondB=sample(second,n-cnt,replace=T)
    stat[i] = median(c(firstB,secondB))

    }
    quantile(stat,c(.05,.95))
```

Comparing Samples from Two Populations

For C++, *R*, and Resampling Stats, comparing samples from two populations is merely a question of embedding a second loop inside the original loop of Section 3.4.2. For example, in *R*

```
for (k in 1:MC){
    stat=0
    for (j in 1:blocks){
          D= sample (A[j],n)
          stat=stat+sum(D)
     }
    if (stat <= sumorig)cnt=cnt+1
    }
cnt/MC  #pvalue
```

With *Stata*, one needs to enter the block variable along with the other data, then make use of the stratified **permute** command as follows:

```
block   regimen   wtloss
man     diet      50
man     diet      25
man     diet      30
man     exer      22
man     exer      34
```

```
man     exer      28
wom     diet      30
wom     diet      20
wom     exer      28
wom     exer      25
wom     exer      24

.  encode regimen, gen(treat)
.  permute wtloss "sum wtloss if treat==1" sum=r(sum),
     reps(400) strata(block)  nowarn

command:      sum wtloss if treat==1
statistic:  sum   = r(sum)
permute var:  wtloss

Monte Carlo permutation statistics   Number of obs = 11
Number of strata =  2
Replications     =  400

T    T(obs)  c  n   p=c/n  SE(p)  [95% Conf. Interval]
sum   155   117 400  0.2925 0.0227 .248349   .3397525

Note:  confidence interval is with respect to p=c/n
Note:  c = #{T >= T(obs)}
```

5.1.3 Measuring Factors We Can't Control

Many of the factors in an experiment will be beyond our control. Rainfall in an agricultural experiment is one example. If we are studying the effects of advertising on the sale of beer, a sudden heat wave or an unexpected drop in temperature will markedly affect our results. But we can measure rainfall and temperature, and use our observations either to block or to correct our results.

In Section 5.3.2, we show how the effects of rainfall and fertilizer could be incorporated in a model, and treatment comparisons made on the basis of differences among the coefficients of the models corresponding to each treatment.

Experimental Design

LIST factors you feel may influence the outcome of your experiment.
BLOCK factors under your control. You may want to use some of these factors to restrict the scope of your experiment, e.g., eliminate all individuals under 18 and over 60.
MEASURE remaining factors.
RANDOMLY assign units to treatment within each block.

5.1.4 Randomization

Imagine you are all prepared for an experiment to find a new and better fertilizer for growing tomatoes. You've set up the experiment indoors under artificial lights in a climate-controlled chamber so you can keep total control over duration of daylight, moisture, and temperature. Since you are hunting for a general-purpose fertilizer, you've put out three separate sets of trays containing sandy loam, a sand-clay mix, and a clay soil respectively. Now, it just remains to put the plants in the soil. Careful . . . Tomato plants are not all alike and there are big differences among seedlings. Studies have shown there is a natural tendency to plant the tall seedlings first and save the runts for last. This is not a good idea if all the runts end up in one tray and all the big sturdy plants in another.

An alternate plan would be to deal the seedlings out like cards, the first for tray 1, the second for tray 2, and so forth down the line. Alas, this method would still seem to put the better plants into the first tray.

When we analyze this experiment using permutation methods, we assume that each assignment of labels to treatments is equally likely. Only if we assign the plants to the different trays completely at random will this assumption be fulfilled. If the plants assigned to tray 1 prove in the end to be larger on the average than those in tray 2, this should be strictly a matter of chance, not the result of a specific bias on our part.

In short, you must *randomly assign subjects to treatment whenever you don't or can't control all the factors in an experiment*.

Our tomato study might use one of the following randomization schemes:

1. Choose a random rearrangement of the integers, 1 to m, corresponding to the m trays. Place the first m plants in the trays in the order specified. Choose a second random rearrangement, a third, and so on, until all the plants have been placed in the trays.
2. Throw an m-sided die and place the first plant in the indicated tray. Continue to throw the die until all the plants have been placed. If the die indicates that an already full tray has been selected, just ignore it and throw again. (See also Exercise 4.)

5.2 k-Sample Comparison

When samples from more than two populations are under consideration, we may wish to test for any and all differences among the populations or our interest may be in specific ordered alternatives.

5.2.1 Testing for Any and All Differences

Suppose we are trying to choose one of three brands of fertilizer to use on our company's fields. Prices vary and if there are no real differences in the crop yield resulting from the use of one fertilizer rather than another, we'd just as soon purchase the cheapest brand. Our model is that crop yield is a function of the fertilizer used and a random component that is the cumulative result of a large number of factors such as minute

differences in the soil surrounding the various plants that we simply cannot control. In symbols, our additive model is that

$$X_{ij} = f_i + z_{ij},$$

where X_{ij} is the yield of the jth planting treated with the ith fertilizer, f_i is the effect of the ith fertilizer, and the experimental errors $\{z_{ij}\}$ all come from the same probability distribution whose mean is zero.

The sum of squares of the deviations about the grand mean may be analyzed into two sums,

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{\cdot\cdot})^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2 + \sum_{i=1}^{I} n_i(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2,$$

the first of which represents the within-treatment sum of squares and the second the between-treatment sum of squares.

For testing the null hypothesis that the effects of each fertilizer are the same, against the alternative that represents an improvement over the others, only the between-treatment sum is of interest. Examining this sum for factors that are the same for all permutations (and thus would be a waste of time to calculate), we see may neglect the grand mean. Only the sum $F_2 = \sum_{i=1}^{I} n_i(\bar{X}_{i\cdot})^2 = \sum_{i=1}^{I}(\sum_{j=1}^{n_i} X_{ij})^2/n_i$ changes as we rearrange the observations among levels and it is this sum whose permutation distribution we use to test the hypothesis.

F_2 emphasizes large deviations from the mean; if we wish to attribute equal value to all deviations, we should use as our test statistic $F_1 = \sum_{i=1}^{I} n_i|\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}|$. We may also want to substitute medians for means, or replace the original observations by their ranks or some other transformation as described in Chapter 4.

5.2.2 Analyzing a One-Way Table

Suppose we wish to compare the effects of three brands of fertilizer on crop yield recorded as follows:

FastGro = c(27, 30, 55, 71, 18)
NewGro = c(38, 12, 72)
WunderGro = c(75,76,54)

C++

```
//set global variables
//n[0]=0, n[1]=n1, n[2]=n1+n2, ..., n[columns+1]
     =#observations=N
int columns, n[columns+1], N;
float data[N];
//compute statistic for original observations
float f2Orig(float *data){
```

```
    float Bsum =0, sum, temp;
    int k, m=0;
    for (int j=0; j<columns; ++j){
         sum = 0;
         for (int i=n[j]; i< n [j+1]; ++i)sum =sum
              + *(data+i);
         Bsum=Bsum +sum*sum;
    }
    return (Bsum);
  }

//select observations and compute statistic in a
     single pass
float f2(float X){
  float Bsum =0, sum, temp;
  int k, m=0;
  for (int j=1; j<columns; ++j){
    sum = 0;
    for (int i=n[j]; i< n [j+1]; i++){
        k=Choose (m, N-1);
        temp = X[k]; X[k]= X[m]; X[m]= temp;
        sum =sum +temp;
        m++;
    }
    Bsum=Bsum +sum*sum;
    }
    return (Bsum);
  }

S = f2Orig(data);
for (int i =0; i < MC; ++i)if (f2(data) >=S) cnt++;
float p=float(cnt)/MC;
cout << p << endl;
```

R

```
F1=function(size,data){
#size is a vector containing the sample sizes
#data is a vector containing all the data in the same
     order as the sample sizes
    stat=0
    start=1
    grandMean = mean(data)
    for (i in 1:length(size)){
        groupMean = mean(data[seq(from = start, length
             = size[i])])
```

```
        stat = stat + abs(groupMean - grandMean)
        start = start + size[i]
        }
    return(stat)
}
```

We use this function repeatedly in the following R program:

```
# One-way analysis of unordered data via a Monte Carlo
size = c(4,6,5,4)
data = c(FastGro, NewGro, WunderGro)
f1=F1(size,data)
#Number MC of simulations determines precision of
     p-value
MC = 1600
cnt = 0
for (i in 1:MC){
    pdata = sample (data)
    f1p=F1(size,pdata)
# counting number of rearrangements for which F1
     greater than or equal to original
    if (f1 <= f1p) cnt=cnt+1
    }
cnt/N
```

Stata

```
*** program to calculate F2 due to Lynn Markham
capture program drop f2test
program define f2test, rclass
  tempvar xijsq sumxj obsnum f2 bcount
  sort brand
  qui by brand: gen 'obsnum'=_n
  qui by brand: gen 'bcount'=_N
  egen 'sumxj'=sum(growth) , by(brand)
  replace 'sumxj'=. if 'obsnum'>1
  gen 'xijsq'=('sumxj'*'sumxj')/'bcount'
  egen 'f2'=sum('xijsq')
  return scalar F2='f2'[1]
end
```

5.2.3 Ordered Alternatives

Suppose we want to assess the effect on crop yield of hours of sunlight, observing the yield for I different levels of sunlight, with n_i observations at each level. The null hypothesis would be that sunlight has no effect on production while our ordered alternative would be that yield is an increasing function of sunlight.

The statistics F_1 and F_2 we made use of earlier offer protection against a broad variety of alternatives including a) the more sunlight the better, b) a little sunlight is good but too much is bad, and so forth. As a result, they may not provide a most powerful test against the ordered alternative that is our specific interest.

Suppose we suspect that yield is a specific function $f[d]$ of the number of days d of sunlight. If we have measured crop yield on land that received $d_1, d_2, \ldots, d_I$ days of sunlight, the optimal statistic for distinguishing between the null hypothesis and our ordered alternative would be the Pitman correlation $\sum_{i=1}^{I} f[d_i] \sum_{j=1}^{n_i} y_{ij}$ where y_{ij} is the yield of the jth plot receiving the ith level of sunlight.

It is the optimal choice for test statistic because while all values are equally likely under the null hypothesis, large values are more likely if the ordered alternative is true.

Frank, Trzos, and Good [1978] studied the increase in chromosome abnormalities and micronucleii as the dose of various known mutagens was increased. Their object was to develop an inexpensive but sensitive biochemical test for mutagenicity that would be able to detect even marginal effects. Thus, they were more than willing to trade the global protection offered by the F-test for a statistical test that would be sensitive to ordered alternatives.

Let us apply the Pitman approach to the data collected by Frank et al. shown in Table 5.1. As the anticipated effect is proportional to the logarithm of the dose, we take $f[\text{dose}] = \log[\text{dose} + 1]$. (Adding a 1 to the dose keeps this function from blowing up at a dose of zero.)

Table 5.1. Micronuclei in Polychromatophilic Erythrocytes and Chromosome Alternations in Bone Marrow of CY Treated Mice

Dose (mg/kg)	Number of Animals	Micronucleii per 200 cells	Breaks per 25 cells
0	4	0 0 0 0	0 1 1 2
5	5	1 1 1 4 5	0 1 2 3 5
20	4	0 0 0 4	3 5 7 7
80	5	2 3 5 11 20	6 7 8 9 9

There are four dose groups; the original data for breaks may be written in the form

$$0\,1\,1\,2 \quad 0\,1\,2\,3\,5 \quad 3\,5\,7\,7 \quad 6\,7\,8\,9\,9$$

As $\log[0 + 1] = 0$, the value of the Pitman statistic for the original data is

$$0 + 11^* \log[6] + 22^* \log[21] + 39^* \log[81] = 112.1.$$

The only larger values are associated with the small handful of rearrangements of the form

$$\begin{array}{llll} 0\,0\,1\,2 & 1\,1\,2\,3\,5 & 3\,5\,7\,7\,6 & 7\,8\,9\,9 \\ 0\,0\,1\,1 & 1\,2\,2\,3\,5 & 3\,5\,7\,7\,6 & 7\,8\,9\,9 \\ 0\,0\,1\,1 & 1\,2\,2\,3\,3 & 5\,5\,7\,7\,6 & 7\,8\,9\,9 \\ 0\,0\,1\,2 & 1\,1\,2\,3\,3 & 5\,5\,7\,7\,6 & 7\,8\,9\,9 \\ 0\,1\,1\,2 & 0\,1\,2\,3\,3 & 5\,5\,7\,7\,6 & 7\,8\,9\,9 \end{array}$$

0 1 1 2	0 1 2 3 5	3 5 6 7 7	7 8 9 9
0 0 1 2	1 1 2 3 5	3 5 6 7 7	7 8 9 9
0 0 1 1	1 2 2 3 5	3 5 6 7 7	7 8 9 9
0 0 1 1	1 2 2 3 3	5 5 6 7 7	7 8 9 9
0 0 1 2	1 1 2 3 3	5 5 6 7 7	7 8 9 9
0 1 1 2	0 1 2 3 3	5 5 6 7 7	7 8 9 9

A statistically significant ordered dose response ($\alpha < 0.001$) has been detected. The micronucleii also exhibit a statistically significantly dose response when we calculate the permutation distribution of S with $f[i] = \log[\text{dose}_i + 1]$.

A word of caution; if we use some function of the dose other than $f[\text{dose}] = \log[\text{dose} + 1]$, we might not observe a statistically significant result. Our choice of a test statistic must always make practical as well as statistical sense.

5.2.4 Calculating Pitman Correlation

R

```
#One-way analysis of ordered data via a Monte Carlo
dose <- c(0, 0, 0, 0, 5, 5, 5, 5, 5, 20, 20, 20, 20,
     80, 80, 80, 80, 80)
breaks <- c (0, 1, 1, 2, 0, 1, 2, 3, 5, 3, 5, 7, 7, 6,
     7, 8, 9, 9)
temp <- 1 +dose
logdose <- log(dose)
rho0 <- cor(logdose, breaks)
#Number N of simulations determines precision of
     p-value
N <- 400
cnt<- 0
for (i in 1:N){
    D <- sample (breaks)
    rho <- cor(D, logdose)
# counting correlations larger than original by chance
    if (rho0 <= rho ) cnt<-cnt+1
    }
cnt/N                       #pvalue
```

Resampling Stats

```
'One-way analysis of ordered data via a Monte Carlo
DATA (0 0 0 0 5 5 5 5 5 20 20 20 20 80 80 80 80 80)
     dose
DATA (0 1 1 2 0 1 2 3 5 3 5 7 7 6 7 8 9 9) breaks
ADD 1 dose temp
LOG temp logdose
CORR logdose breaks rho0
```

```
'N determines precision of p-value
COPY 400 N
REPEAT N
   SHUFFLE breaks B
   CORR logdose B rho
   SCORE rho scrboard
END
'Is this a one- or two-sided test?
COUNT scrboard>=rho0 extremes
LET pvalue =extremes/N
PRINT pvalue
```

Resampling Stats for Excel

Outline the two columns you wish to shuffle. When completing the Matrix Shuffle form, specify "Shuffle within Columns." Compute Excel's **Correl()** function repeatedly.

Stata

```
gen logdose = log(dose+1)
permute breaks "corr logdose breaks", teststat=r(rho)
```

5.2.5 Effect of Ties

Seemingly, ties can complicate the determination of the significance level. Because of ties, each of the rearrangements noted in the preceding example might actually have resulted from several distinct reassignments of subjects to treatment groups and must be weighted accordingly. To illustrate this point, suppose we put tags on the 1's in the original sample

$$0\,1^*\,1\#\,2 \quad 0\,1\,2\,3\,5 \quad 3\,5\,7\,7 \quad 6\,7\,8\,9\,9$$

The rearrangement

$$0\,0\,1\,2 \quad 1\,1\,2\,3\,5 \quad 3\,5\,7\,7 \quad 6\,7\,8\,9\,9$$

corresponds to the three reassignments

$$\begin{array}{llll} 0\,0\,1\;\;2 & 1^*\,1\#\,2\,3\,5 & 3\,5\,7\,7 & 6\,7\,8\,9\,9 \\ 0\,0\,1^*\,2 & 1\;\;1\#\,2\,3\,5 & 3\,5\,7\,7 & 6\,7\,8\,9\,9 \\ 0\,0\,1\#\,2 & 1\;\;1^*\,2\,3\,5 & 3\,5\,7\,7 & 6\,7\,8\,9\,9 \end{array}$$

The 18 observations are divided into four dose groups containing 4, 5, 4, and 5 observations, respectively, so that there are $\binom{18}{4\;5\;4}$ possible reassignments of observations to dose groups. Each reassignment has probability $1/\binom{18}{4\;5\;4}$ of occurring, so the probability of the rearrangement

$$0\,0\,1\,2 \quad 1\,1\,2\,3\,5 \quad 3\,5\,7\,7 \quad 6\,7\,8\,9\,9$$

is $3/\binom{18}{4\;5\;4}$.

To determine the significance level when there are ties, weight each distinct rearrangement by its probability of occurrence. This weighting is done automatically if you use Monte Carlo sampling methods.

5.3 Balanced Designs

Suppose we want to assess the simultaneous effects on crop yield of hours of sunlight and rainfall. We determine to observe the crop yield X_{ijm} for I different levels of sunlight, $i = 1$ to I, and J different levels of rainfall, $j = 1$ to J, and to make M observations at each factor combination ij, $m = 1$ to M. We adopt as our model relating the dependent variable crop yield (the effect) to the independent variables of sunlight and rainfall (the causes)

$$X_{ijm} = \mu + s_i + r_j + (sr)_{ij} + \varepsilon_{ijm}.$$

In this model, terms with a single subscript like s_i, the effect of sunlight, are called *main effects*. Terms with multiple subscripts like $(sr)_{ij}$, the residual and nonadditive effect of sunlight and rainfall, are called interactions. The residuals $\{(sr)_{ij}\}$ represent that portion of crop yield that cannot be explained by the independent variables alone. To ensure the residuals are exchangeable so that permutation methods can be applied, the experimental units must be assigned at random to treatment.

When we have multiple factors, we must also have multiple test statistics. In the preceding example, we require two separate tests and test statistics for the main effects of rainfall and sunlight, plus a test for their interaction. Will we be able to find statistics that measure a single intended effect without confounding it with a second unrelated effect? Will the several test statistics be independent of one another?

The answer is yes to both questions only if the design is balanced, that is, if there are equal numbers of observations in each subcategory. Moreover, only symmetric permutations (defined in Section 5.3.4) will ensure the test statistics are independent of one another. In an unbalanced design, main effects will be confounded with interactions, so the two cannot be tested separately, a topic we return to with a bootstrap solution in Section 5.6.

5.3.1 Main Effects

In a k-way analysis with equal sample sizes M in each category, we assess the main effects using essentially the same statistics we would use for randomized blocks. To simplify our calculations, the overall average μ is chosen so that $\Sigma_i s_i = 0$, $\Sigma_j r_j = 0$, $\Sigma_i (sr)_{ij} = 0$, and $\Sigma_j (sr)_{ij} = 0$.

If we have only two levels of sunlight, then our test statistic for the effect of sunlight was shown in Chapter 3 to be the sum of all observations at the first level of sunlight.

If we have more than two levels of sunlight, our test statistic may be any of the three test statistics described in Section 5.2, the distinction being that we need to sum over the blocks for the various levels of rainfall as in Section 5.1.1. For example, to compute the Pitman correlation for the main effect of sunlight, we would use the statistic

$$\sum_{j=1}^{J}\sum_{i=1}^{I} f[d_i]\sum_{m=1}^{n_{ij}} y_{ijm}.$$

To obtain the permutation distributions of the test statistics for the effect of sunlight, we permute the observations independently in each of the J blocks determined by a specific level of rainfall.

5.3.2 Analyzing a Two-Way Table

In this second example, we apply the permutation method to determine the main effects of sunlight and fertilizer on crop yield using the data from the two-factor experiment depicted in Table 5.2. As there are only two levels of sunlight in this experiment, we use the sum of the sums of the observations in the first sample of each block to test for the main effect.

Table 5.2. Effect of Sunlight and Fertilizer on Crop Yield

			FERTILIZER		
S			Lo	Med	High
U	Lo	5	15	21	
N		10	22	29	
L		8	18	25	
I					
G	Hi	6	25	55	
H		9	32	60	
T		12	40	48	

For the original observations, this sum is $23 + 55 + 75 = 153$. One possible rearrangement is shown in Table 5.3 in which we have interchanged the two observations marked with an asterisk, the 5 and 6. The new value of our test statistic is 154.

Table 5.3. Effect of Sunlight and Fertilizer Data Rearranged

	Lo	Med	High
Lo	6*	15	21
	10#	22	29
	8	18	25
Hi	5*	25	55
	9#	32	60
	12	40	48

As can be seen by a continuing series of straightforward hand calculations, the test statistic for the main effect of sunlight is as small or smaller than it is for the original observations in only 8 out of the $\binom{6}{3}^3 = 8000$ possible rearrangements. For example,

it is smaller when we swap the 9 of the Hi-Lo group for the 10 of the Lo-Lo group (the two observations marked with the pound sign #). We conclude the effect of sunlight is statistically significant.

The computations for the main effect of fertilizer are more complicated—we must examine $\binom{9}{3}^2$ rearrangements, and compute the statistic F_1 for each. We use F_1 rather than the Pitman correlation because of the possibility that too much fertilizer—the "High" level—might actually suppress growth. Only a computer can do this many calculations quickly and correctly, so we adapted one of the programs in Section 5.2.2 to make them. The estimated significance level is 0.001 and we conclude that this main effect, too, is statistically significant.

In this last example, each category held the same number of experimental subjects. If the numbers of observations were unequal, our main effect would have been confounded with one or more of the interactions (see Section 5.6). In contrast to the simpler designs we studied in Chapter 3, missing data will affect our analysis.

5.3.3 Testing for Interactions

In the preceding analysis of main effects, we assumed the effect of sunlight was the same regardless of the levels of the other factors. But this may not always be the case. Of what value is sunlight to a plant if there is not enough fertilizer in the ground to support growth? Or vice versa, of what value is fertilizer if there is insufficient sunlight?

Sunlight and fertilizer interact; we cannot simply add their effects. Our model should and does include an interaction term $(sf)_{ij}$:

$$X_{ijm} = \mu + s_i + f_j + (sf)_{ij} + \varepsilon_{ijm}.$$

The presence of the constant term μ allows us to simplify the analysis by setting $\Sigma_i s_i = \Sigma_j f_j = \Sigma_i (sf)_{ij} = \Sigma_j (sf)_{ij} = 0$, so that the various subscripted terms represent deviations from an overall average.

Suppose we were to eliminate row and column effects by subtracting the row and column means from the original observations as in Table 5.4,

$$X'_{ijm} = X_{ijm} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...}.$$

The pattern of plus and minus signs in this table of residuals suggests that fertilizer and sunlight affect crop yield in a superadditive or synergistic fashion. Note the minus signs associated with the mismatched combinations of a high level of sunlight and a low level of fertilizer and a low level of sunlight with a high level of fertilizer.

An obvious statistic to test the hypothesis that the interactions terms are zero for all levels of sunlight and fertilizer would appear to be $W_{IJ} = \Sigma_i \Sigma_j (\Sigma_m X'_{ijm})^2$.

Unfortunately, the labels on the deviates X'_{ijm} are not exchangeable so we cannot permute them. The expected value of X'_{ijm} is $\varepsilon_{ijk} - \bar{\varepsilon}_{i..} - \bar{\varepsilon}_{.j.} + \varepsilon_{...}$. Thus the deviates are correlated and the correlation between two deviates in the same row or column has a different value than the correlation between deviates in different rows or columns.

Table 5.4. Effect of Sunlight and Fertilizer on Crop Yield

Testing for Nonadditive Interaction				
		FERTILIZER		
S		Lo	Med	High
U	LO	4.1	−2.1	−11.2
N		9.1	4.1	−3.2
L		7.1	0.1	−7.2
I				
G	HI	−9.8	−7.7	7.8
H		−7.8	−0.7	12.8
T		−3.8	7.2	0.8

5.3.4 Synchronized Rearrangements

Let's take a closer look at the confounding between main effects and interactions. Figure 5.1a depicts a two-factor experimental design with two levels of each factor. Within each row and column, the interaction terms sum to zero because of the way we've defined them. If we were to compare the expected values of the row sums, the difference would depend solely on the row effects if any. Similarly, any differences in the expected values of the column sums would depend solely on the column effects if any.

Fig. 5.1a. A 2 × 2 experimental design.

The situation is quite different in Figure 5.1b where row and row-column interactions are clearly confounded.

Fig. 5.1b. An unsynchronized rearrangement of a 2 × 2 experimental design.

In Figure 5.1c, the row sum of the interaction terms is still zero, so that once more any differences in the expected values of the row sums would depend solely on the row

Fig. 5.1c. A synchronized rearrangement of a 2×2 experimental design.

effects if any. By restricting ourselves to similar synchronized rearrangements, we can avoid confounding interactions with main effects.

Note that the original set of all possible rearrangements $\binom{4m}{m\ m\ m}$ in number is now subdivided into a set of synchronized rearrangements, $\sum_{k=0}^{m} \binom{m}{k}^4$ in number, for testing row effects, two sets of similar size for testing column effects and interactions respectively, and the unsynchronized permutations. Only the original design of Figure 5.1a is common to more than one set of synchronized observations, so that by restricting ourselves to one of these sets each time, we obtain independent exact tests for row effects, column effects, and interactions.

If there are only two observations per cell as depicted in the diagram, this means that from the original set of 2520 rearrangements, we have only three sets of 18 synchronized arrangements that can be used for testing purposes.

5.3.5 A Worked-Through Example

The following C++ code can be used to obtain the desired tests for a 2×2 balanced design with n observations per cell. Assume data is stored in the vector X.

```
void Compute_Bal2 (float *X, n){
      float S1=0, S2=0, S12=0, s1=0, s2=0, s12=0, *Y;
      Y = new float[4*n];
      int begin, chng, chng2, cnt1=0, cnt2=0, cnt12=0;
/* compute the row, column, and interaction statistics
     for the original sample */
      Stats2x2 (X, &S1, &S2, &S12);
      for (int i =0; i < MC; ++i){
            s1=0; s2=0; s12=0;
            for (int i=0; i< 4*n; ++i) Y[i]=X[i];
      //compute row main effect
            chng= ShuffleR (Y, &s1);
            RearrangR(Y, &s1, chng);
      //compute interaction
            chng2=ShuffleC (Y, &s2, chng);
            RearrangC (Y, &s2, chng, chng2);
            for (begin =0; begin < n; ++begin)
            s12 = s12 + *(Y + begin) + *(Y + 3*n +
                 begin);
      //compute column main effect
```

```
            for (int i=0; i< 4*n; ++i) Y[i]=X[i];
            chng2=ShuffleC (Y, &s2,0);
            RearrangC (Y, &s2, 0, chng2);
            if (s1 >= S1)cnt1++;
            if (s2 >= S2)cnt2++;
            if (s12 >= S12)cnt12++;
      }
      float p1=float(cnt1)/MC, p2=float(cnt2)/MC,
           p12=float(cnt12)/MC;
      cout << p1 << "  " << p2 << "  " << p12;
}

void Stats2x2 (float *X, float *S1, float *S2, float
     *S12)
{
      for (int begin =0; begin < n; ++begin){
            *S1 = *S1 + *(X+ begin) + *(X+n+begin);
            *S2 = *S2 + *(X+ begin) + *(X+2*n+begin);
            *S12 = *S12 + *(X+ begin) + *(X+3*n+begin);
      }
}

int ShuffleR (float *X, float *s1)
      {
      //interchange elements in 1st column between rows
      int chng=0, z, v, i;
      float temp1, temp2;
      for (i =0; i < n; ++i){
      z=Choose (chng, 2*n-1-chng);

            if (z>n-1){
            v = z-n;
                  z = Choose (chng, n-1);
                  temp1 = *(X+z);
                  *(X+z) = *(X+chng);
                  temp2= *(X+2*n+chng+v);
                  *(X+2*n+chng+v) = *(X+2*n+ chng);
         *(X+chng)= temp2;
            *(X+2*n + chng) = temp1;
                  chng++;
                  }
         }
      //sum contents of R1C1
      for (i =0; i < n; ++i)*s1 = *s1 + *(X+i);
   //cout << " s1=" << *s1;
return (chng);
}
```

```
void RearrangR (float *X, float *s1, int chng)
      {
//interchange elements in 2nd column between rows
      float temp1, temp2;
   int begin, z1, z2;
      for (begin =0; begin< chng; ++begin){
            z1=Choose (begin, n-1);
            z2=Choose (begin, n-1);
            temp1= *(X+n+z1);
            temp2= *(X+3*n+z2);
            *(X+n+z1)= *(X+n+ begin);
            *(X+3*n+ z2)= *(X+ 3*n+ begin);
            *(X+n+ begin)=temp2;
            *(X+3*n+ begin)=temp1;
            *s1 =*s1 + temp2;
            }
      //sum remaining contents of R1C2
       for (begin =chng; begin < n; ++begin)*s1= *s1
            + *(X+n+begin);
       }

int ShuffleC (float *X, float *s2, int chng)
      {
   //interchange elements in 1st row between columns
      int chngC=0, z, v, gap, i;
      float temp1, temp2;
      for (i =chng; i < n; ++i){
      gap = chng+chngC; // gap contains previously
           changed elements
      z=Choose (gap, 2*n-1-gap);
            if (z>n-1){
            v=z-n;
         z = Choose (gap, n-1);
         temp1 = *(X+z);
         *(X+z) = *(X+chng);
         temp2= *(X+ n + gap + v);
                  *(X+ n + gap + v) = *(X+ n + gap);
         *(X+chng)= temp2;
            *(X+ n + gap) = temp1;
         chngC++;
                  }
   }
   //sum contents of R1C1
   *s2=0;
      for (i =0; i < n; ++i)*s2 = *s2 + *(X+i);
return (chngC);
}
```

```
void RearrangC (float *X, float *s2, int chng, int
     chngC)
      {
      int z1, z2, begin;
   float temp1, temp2;
      for (begin =chng; begin< chngC + chng; ++begin){
            z1=Choose (begin, n-1);
            z2=Choose (begin, n-1);
            temp1= *(X+2*n+ z1);
            temp2= *(X+3*n+ z2);
            *(X+2*n+ z1)= *(X+2*n+ begin);
            *(X+3*n+ z2)= *(X+3*n+ begin);
            *(X+2*n+ begin)=temp2;
            *(X+3*n+ begin)=temp1;
            }
      for (int begin = 0; begin < n; ++begin)*s2= *s2
           + *(X+2*n+ begin);
      }
```

To test your compiled program, consider the data displayed in Table 5.5 taken from Hettmansperger [1984]. Survival times are measured for each of two different poisons and two different treatments and the experiment is replicated four different times.

Table 5.5a. Survival Times Following Treatment

	T1	T2
P1	31, 45, 46, 43	45, 71, 66, 62
P2	22, 21,18, 23	30, 36, 31, 33

Table 5.5b. Residuals After Subtracting Treatment and Poison Mean Effects

	T1	T2
P1	−12.3, 1.7, 2.7, −0.3	−13.9, 12.1, 7.1, 3.1
P2	3.1, 2.1, −0.9, 4.1	−4.6, 1.4, −3.6, −1.6

There is a significant difference between the poisons at the 5% level. There is a marginally significant difference between the treatments, the effect of treatment being most pronounced for the first poison. The interactions are marginally significant at the 5% level also, and the effects of treatment should be reported separately for the two poisons.

5.4 Designing an Experiment

All the preceding results are based on the assumption that the assignment of treatments to plots (or subjects) is made at random. While it might be convenient to fertilize our

plots as shown in Table 5.6, the result could be a systematic bias. For example, suppose there is a gradient in dissolved minerals from east to west across the field.

Table 5.6. Systematic Assignment of Fertilizer Levels to Plots

Hi	Med	Lo
Hi	Med	Lo
Hi	Med	Lo

The layout adopted in Table 5.7, obtained with the aid of a computerized random number generator, reduces but does not eliminate the effects of this hypothetical gradient. Because this layout was selected at random, the exchangeability of the error terms and, hence, the exactness of the corresponding permutation test are ensured. Unfortunately, the layout of Table 5.6 with its built-in bias can also result from a random assignment; its selection is neither more nor less probable than any of the other possibilities.

Table 5.7. Random Assignment of Fertilizer Levels to Plots

Hi	Med	Lo
Lo	Med	Lo
Hi	Hi	Med

What can we do to avoid such an undesirable event? In the layout of Table 5.8, known as a *Latin Square*, each fertilizer level occurs once and once only in each row and in each column; if there is a systematic gradient of minerals in the soil, then this layout ensures the gradient will have almost equal impact on each of the three treatment levels. It will have an almost equal impact even if the gradient extends from northeast to southwest rather than from east to west or north to south. I use the phrase "almost equal" because a gradient effect may still persist. The design and analysis of Latin Squares is described in the next section.

Table 5.8. Latin Square Assignment of Fertilizer Levels to Plots

Hi	Med	Lo
Lo	Hi	Med
Med	Lo	Hi

5.4.1 Latin Square

The Latin Square is one of the simplest examples of an experimental design in which the statistician takes advantage of some aspect of the model to reduce the overall sample size. A Latin Square is a three-factor experiment in which each combination of factors occurs once and once only. We can use a Latin Square like that of Table 5.8 to assess the effects of soil composition on crop yield.

In Table 5.8, suppose that Factor 1: gypsum concentration, say, is increasing from left to right and Factor 2 is increasing from top to bottom (or from North to South). Note that the third factor, treatment, occurs in combination with the other two in such a way that each combination of factors—row, column, and treatment—occurs once and once only.

Because of this latter restriction, there are only 12 different ways in which we can assign the varying factor levels to form a 3×3 Latin Square. Among the other 11 designs are the following where the varying levels of the third factor are denoted by the capital letters A, B, and C,

Design 2

	1	2	3
1	A	C	B
2	B	A	C
3	C	B	A

and

Design 3

	1	2	3
1	C	B	A
2	B	A	C
3	A	C	B

Assume we begin our experiment by selecting one of these 12 designs at random and planting our seeds in accordance with the indicated conditions.

Because there is only a single replication of each factor combination in a Latin Square, we cannot estimate the interactions. The Latin Square is appropriate only if we feel confident in assuming the effects of the various factors are completely *additive*, that is, the interaction terms are zero.

Our model for the Latin Square is

$$X_{kji} = \mu + q_k + r_j + s_i + \varepsilon_{kji},$$

with the effects as always representing deviations from an average, so that the sums of the main effects are zero. As always in a permutation analysis, we assume the labels on the errors $\{\varepsilon_{kji}\}$ are exchangeable. Our null hypothesis is that the additive effects of the various levels of sunlight $\{s_i\}$ are all zero. If we assume an ordered alternative, K: $s_1 < s_2 < s_3$, our test statistic for the main effect is similar to the correlation statistic: $\sum_{i=1}^{3}(i-1)(\bar{X}_{i..} - \bar{X}_{...})$ or, equivalently, after eliminating the grand mean $\bar{X}_{...}$ which is invariant under permutations, $R' = \bar{X}_A - \bar{X}_C$.

We evaluate this test statistic both for the observed design and for each of the 12 possible Latin Square designs that might have been employed in this particular experiment. We reject the hypothesis of no treatment effect only if the test statistic R for the original observations is an extreme value.

For example, suppose we employ the Latin Square of Table 5.8 and observe

21	28	17
14	27	19
13	18	23

Then $\bar{X}_A - \bar{X}_C = 58 - 57 = 1$. Had we employed Design 2, then $R' = 71 - 65 = 6$. With Design 3, our test statistic $57 - 58 = -1$.

We see from the permutation distribution obtained in this manner that 1, the value of our test statistic for the design actually employed in the experiment, is an average value, not an extreme one. We accept the null hypothesis and conclude that increasing the treatment level from A to B to C does not significantly increase the yield.

Designing and Analyzing a Latin Square

H: mean/median the same for all levels of each treatment.
K: means/medians are different for at least one level.

Assumptions:

1) Observations are exchangeable if the hypothesis is true.
2) Treatment effects are additive (not synergistic or antagonistic).

Procedure:

List all possible Latin Squares for the given number of treatment levels.
Assign one design at random and use it to perform the experiment.
Choose a test statistic (R, F_1, or F_2).
Compute the statistic for the design you used.
Compute the test statistic for all the other possible Latin Square designs.
Determine from the resultant permutation distribution whether the original value of the test statistic is an extreme one.

5.5 Determining Sample Size

Power is an increasing function of sample size, as we saw in Chapter 4, so we should take as large a sample as we can afford, providing the gain in power from each new observation is worth the expense of gathering it.

For one- and two-sample comparisons of means and proportions, a number of commercially available statistics packages can help us with sample-size determination. A typical calculation using Stata is provided in what follows. Note that to make the determination with Stata or any other program, we need to specify both the alternative of interest and the desired significance level.

To simulate the sampling process for a known distribution like the normal, we first choose a uniformly distributed random number between 0 and 1, the same range of values taken by the distribution function, then we look this number up as an entry in a table of the inverse of the normal distribution.

Programming in Stata, for example, we write **invnorm(uniform())** and repeat this command for each element in the untreated sample. If our alternative is that the population comes from a population with mean 5 and standard deviation 2, we would write $5 + 2$*invnorm(uniform()). We repeat this command for each observation in the treated sample.

Consider a potential study of the effect of oral contraceptives on the blood pressure of women ages 35–39 where it is known that the study values for oral contraceptive users average 133 ± 15, while nonusers average 127 ± 18. The Stata command

. sampsi 133 127, alpha(0.05) power(0.8) sd1(15) sd2(18)

yields the following output:

Estimated sample size for two-sample comparison of means

Test $H_0 : m_1 = m_2$, where m_1 is the mean in population 1 and m_2 is the mean in population 2

Assumptions:

alpha $= 0.0500$ (two-sided)
power $= 0.8000$
$m_1 = 133$
$m_2 = 127$
$sd_1 = 15$
$sd_2 = 18$
$n_2/n_1 = 1.00$

Estimated required sample sizes:

$n_1 = 120$
$n_2 = 120$

If we aren't sure about the underlying distribution, we draw a bootstrap sample with replacement from the empirical distribution. We compute the test statistic for the sample, and note whether we accept or reject. We repeat the entire process 50 to 400 times (50 times when just trying to get a rough idea of the correct sample size, 400 times when closing in on the final value). The number of rejections divided by the number of simulations provides us with an estimate of the power for the specific experimental design and our initial sample sizes. If the power is still too low, we increase the sample sizes and repeat the preceding simulation process.

5.5.1 Group Sequential Designs

Recent developments include group sequential designs in which testing is performed only after groups of observations have been collected, for example, after every 6 months in a clinical trial. Figure 5.2 illustrates the use of S+SeqTrial, from http://www.insightful.com, for designing a trial to compare binomial proportions in a treatment and control group. The null hypothesis is $p = 0.4$ in both groups, and the alternative hypothesis is $p = 0.45$ in the treatment group. An "O'Brien–Fleming" design is to be employed, with a total of four analyses (three "interim analyses" and a final analysis).

The resultant output (see sidebar) begins with the call to the "seqDesign" function that you would use if working from the program's command line rather than using the menu interface. The null hypothesis is that Theta (the difference in survival probability between the two groups) is 0.0, and the alternative hypothesis is that Theta is at least

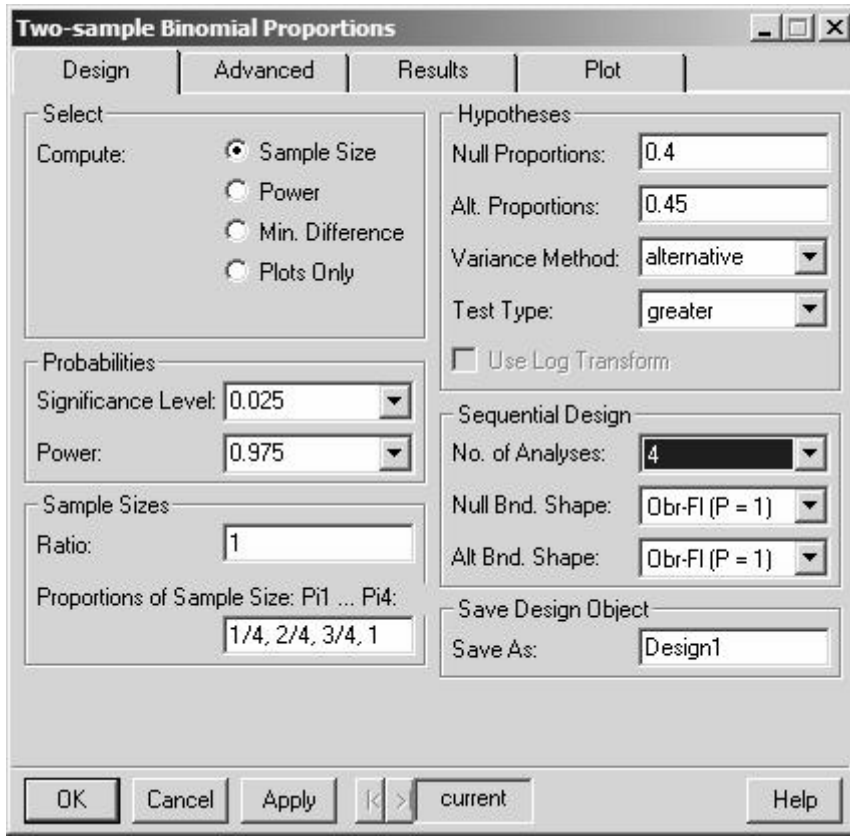


Fig. 5.2. Group-sequential design menu, in S+SeqTrial.

0.05. The last section indicates the stopping rules: After 1565 observations (split roughly equally between the two groups) we should analyze the interim results. At the first analysis, if the treatment group has a survival probability which is 10% greater than the control group, we stop early and reject the null hypothesis; if the treatment group is doing 5% worse we also stop early, and accept the null hypothesis (at this point it appears that our treatment is actually killing people; there is little point in continuing the trial). Any ambiguous result, in the middle, causes us to collect more data. At the second analysis time the decision boundaries are narrower, with lower and upper boundaries 0% and 5%; stop and declare success if the treatment group is doing 5% better, stop and give up if the treatment group is doing at all worse. The decision boundaries at the third analysis time are even narrower, and at the final time (6260 total observations) they coincide; at this point we make a decision one way or the other.

Thus, we may need to examine anywhere from 1565 to 6260 subjects before reaching a conclusion. A fixed-sample trial would require 6000 subjects, regardless.

```
   *** Two-sample Binomial Proportions Trial ***

Call:
seqDesign(prob.model = "proportions", arms = 2,
     null.hypothesis
   = 0.4, alt.hypothesis = 0.45, ratio = c(1., 1.),
        nbr.analyses
   = 4, test.type = "greater", power = 0.975, alpha =
        0.025, beta
   = 0.975, epsilon = c(0., 1.), display.scale =
        seqScale(scaleType = "X"))

PROBABILITY MODEL and HYPOTHESES:
   Two arm study of binary response variable
   Theta is difference in probabilities (Treatment -
        Comparison)
   One-sided hypothesis test of a greater alternative:
          Null hypothesis : Theta <= 0 (size  = 0.025)
   Alternative hypothesis : Theta >= 0.05
                                          (power = 0.975)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                               a      d
   Time 1 (N= 1565.05) -0.0500 0.1000
   Time 2 (N= 3130.09)  0.0000 0.0500
   Time 3 (N= 4695.14)  0.0167 0.0333
   Time 4 (N= 6260.18)  0.0250 0.0250
```

5.6 Unbalanced Designs

Imbalance in the design will result in the confounding of main effects with interactions. Suppose the observations in a two-factor experimental design are distributed as in the following diagram:

Mean 0	Mean 2
Mean 2	Mean 0

There are no main effects in this example—both row means and both column means have the same expectation, but there is a clear interaction represented by the two nonzero off-diagonal elements.

If the design is balanced, with equal numbers per cell, the lack of significant main effects and the presence of a significant interaction should and will be confirmed by our analysis. But suppose the design is not in balance, that for every 10 observations in the

first column, we have only one observation in the second. Because of this imbalance, when we use the statistic S, we will uncover a false "row" effect that is actually due to the interaction between rows and columns. The main effect is said to be *confounded* with the interaction.

If a design is unbalanced as in the preceding example, we cannot test for a "pure" main effect or a "pure" interaction. But we may be able to test for the combination of a main effect with an interaction by using the same statistic (F_1, F_2, or Pitman correlation) that we would use to test for the main effect alone. This combined effect will not be confounded with the main effects of other unrelated factors.

5.6.1 Multidimensional Contingency Tables

Whether or not the design is balanced, the methods of Chapter 6 can be applied to multifactor designs whose outcomes are either categorical or of the yes/no variety. Table 5.9 contains the results of an experiment by Plackett and Hewlett [1963] in which milkweed bugs were exposed to various levels of two insecticides. At issue is whether the two drugs act independently.

Table 5.9. Deaths of Milkweed Bugs Exposed to Various Levels of Two Insecticides

Dose B	0			0.2		
Dose A	0	0.05	0.07	0	0.05	0.07
Died		9	22	5	27	27
Survived		39	26	43	21	21

Although death due to a variety of spontaneous and background causes could be anticipated, no attempt was made to actually measure this background—the cell corresponding to a zero dose of each drug is empty. The resultant design is an unbalanced one. Still, a solution is possible via the bootstrap [Währendorf and Brown 1980].

An underlying biological assumption is that the dose threshold above which a given insecticide is toxic varies from insect to insect. Suppose we form a pair of bootstrap samples. The first sample we construct in two stages: first, we draw an observation at random from the sample of 48 milkweed bugs treated with 0.05 units of the first insecticide alone. If by chance we select one of the 39 survivors, then we draw from the sample of 48 bugs treated with 0.2 units of the second insecticide alone. Otherwise, we record a "death."

Of course, we don't actually perform the drawing but simulate it through the use of a random number generator. If this number is greater than 9/48, the insect lives to be treated a second time, otherwise it dies.

The second bootstrap sample we select with replacement from the 27 killed and 21 survivors in the sample treated with both insecticides. We repeat the process 50–200 times, each time comparing the number of survivors in the two bootstrap samples. If the two insecticides act independently, the numbers should be comparable.

5.6.2 Missing Combinations

If an entire factor-combination is missing, we may not be able to estimate or test any of the effects. One very concrete example is an unbalanced design I encountered in the 1970s when I worked with Makinodan et al. [1976] to study the effects of age on the mediation of the immune response. They measured the anti-SBRC response of spleen cells derived from C57BL mice of various ages. In one set of trials, the cells were derived entirely from the spleens of young mice, in a second, they came from the spleens of old mice, and in a third they came from mixtures of the two.

Let $X_{i,j,k}$ denote the response of the kth sample taken from a population of type i, j ($i = 1 = j$: controls; $i = 2$, $j = 1$: cells from young animals only; $i = 1$, $j = 2$: cells from old animals only; $i = 2 = j$: mixture of cells from old and young animals). We assume that for lymphocytes taken from the spleens of young animals,

$$X_{2,1,k} = \mu + \alpha + e_{2,1,k},$$

for the spleens of old animals,

$$X_{1,2,k} = \mu - \alpha + e_{1,2,k},$$

and for a mixture of p spleens from young animals and $(1-p)$ spleens from old animals, where $0 \leq p \leq 1$,

$$X_{2,2,k} = p(\mu + \alpha) + (1 - p)(\mu - \alpha) - \gamma + e_{2,2,k},$$

where the $e_{2,2,k}$ are independent values.

Makinodan knew beforehand that there would be differences in the results for old and young animals, that is, $\alpha > 0$. He also knew that the errors $e_{i,j,k}$ were unlikely to have normal distributions.

His primary interest was the possible interaction between the cells from the different age groups. If the interaction term γ were 0, then one could infer the two cell populations did not interact. $\gamma < 0$ meant there were excess lymphocytes in young populations, while $\gamma > 0$ suggests the presence of suppressor cells in the spleens of older animals.

But what statistic and what statistical method are we to use to do the test? If the design were balanced, or we could be sure that the effect μ in the absence of lymphocytes was 0, then the statistic of choice would be

$$S = |\bar{X}_{22.} - p\bar{X}_{21.} - (1 - p)\bar{X}_{12.}|.$$

But the design is not balanced, with the result that the main effects in which we are not interested are confounded with the interaction with which we are.

Fortunately, another resampling method, the bootstrap, can provide a solution: Draw an observation at random and with replacement from the set $\{x_{10k}\}$; label it x^*_{10}. Similarly, draw the bootstrap observations x^*_{01} and x^*_{11} from the sets $\{x_{01k}\}$ and $\{x_{11k}\}$. And let

$$\gamma^* = p\bar{X}^*_{2,1} + (1 - p)\bar{X}^*_{1,2} - \bar{X}^*_{2,2}.$$

Repeat this resampling procedure several hundred times, obtaining a bootstrap estimate γ^* of the interaction each time you resample. Use the resultant set of bootstrap estimates to obtain a confidence interval for γ. If 0 belongs to this confidence interval, accept the hypothesis of additivity; otherwise reject.

Mean DPFC Response

Effect of pooled old BC3FL spleen cells on the anti-SRBC response of indicator pooled BC3FL spleen cells. Data extracted from Makinodan et al. [1976]. Bootstrap analysis.

Young Cells	Old Cells	$1/2 + 1/2$
5640	1150	7100
5120	2520	11020
5780	900	13065
4430	50	
7230		

Bootstrap sample 1:	5640	900	11020	4480
Bootstrap sample 2:	5780	1150	11020	4090
Bootstrap sample 3:	7230	1150	7100	1280
. . . . .				
. . . . .				
Bootstrap sample 100:	5780	2520	7100	1200

The result for this set of samples is a 90% CI that stretches from -7815 to $+1060$ so that we are unable to draw any firm conclusions. But Makinodan et al. [1976] conducted many replications of this experiment for varying values of p with comparable results; they could feel confident in concluding that $\gamma < 0$ showing that young spleens have an excess of lymphocytes.

5.7 Summary

In this chapter, you learned the principles of experimental design: to block or measure all factors under your control, to randomize with regard to factors that are not.

You learned to analyze balanced k-way designs for main effects, and balanced two-by-two designs for both main effects and interactions. You learned to use the Latin Square to reduce sample size and to use bootstrap methods when designs are not balanced.

5.8 To Learn More

For more on the principles of experimental design, see Fisher [1935], Kempthorne [1955], Wilk and Kempthorne [1956, 1957], Scheffe [1959], Maxwell and Cole [1991]. Further sample-size guidelines are provided in Shuster [1993].

Permutation tests have been applied to a wide variety of experimental designs including analysis of variance [Kempthorne 1955, 1966, 1969, 1977; Jin 1984; Soms 1985; Diggle, Lange, and Benes 1991; Ryan, Tracey, and Rounds 1996, Pesarin 1999], clinical trials [Lachin 1988a,b], covariance [Peritz, 1982], crossovers [Shen and Quade 1986], factorial [Loughin and Noble 1997], growth curves [Foutz, Jensen, and Anderson 1985; Zerbe 1979a,b], matched pairs [Peritz 1985; Welch 1987, 1988; Rosenbaum 1988; Good 1991, Zumbo 1996], randomized blocks [Wilk 1955], restricted randomization [Smythe 1988], and sequential clinical trials [Wei 1988; Wei, Smythe and Smith 1986]. Bradbury [1987] compares parametric with randomization tests. Mapleson [1986] applied the bootstrap to the analysis of clinical trials.

The theory of symmetric rearrangements and weak exchangeability may be found in Pesarin [2001] and Good [2003].

The bootstrap approach to the data of Table 5.9 was suggested by Wåhrendorf and Brown [1980]. See, also, Romano [1988].

5.9 Exercises

1. Design an experiment.
- **a.** List all the factors that might influence the outcome of your experiment.
- **b.** Write a model in terms of these factors.
- **c.** Which factors are under your control?
- **d.** Which of these factors will you use to restrict the scope of the experiment?
- **e.** Which of these factors will you use to block?
- **f.** Which of the remaining factors will you neglect initially, that is, lump into your error term?
- **g.** How will you deal with each of the remaining covariates?
- **h.** How many subjects/items will you observe in each subcategory?
- **i.** Write out two of the possible assignments of subjects to treatment.
- **j.** How many possible assignments are there in all?

2. A known standard was sent to six laboratories for testing.
- **a.** Are the results comparable among laboratories?

	Laboratory					
Date	A	B	C	D	E	F
1/1	221.1	208.8	211.1	208.3	221.1	224.2
1/2	224.2	206.9	198.4	214.1	208.8	206.9
1/3	217.8	205.9	213.0	209.1	211.1	198.4

- **b.** The standard was submitted to the same laboratories the following month. Are the results comparable from month to month?

	Laboratory					
Date	A	B	C	D	E	F
2/1	208.8	211.4	208.9	207.7	208.3	214.1
2/2	212.6	205.8	206.0	216.2	208.8	212.6
2/3	213.3	202.5	209.8	203.7	211.4	205.8

3. Potted tomato plants in groups of six were maintained in one of two levels of artificial light and two levels of water. What effects, if any, did the different levels have on yield?

Light	Water	Yield	Light	Water	Yield
1	1	12	2	1	16
1	1	8	2	1	12
1	1	8	2	1	13
1	2	13	2	2	19
1	2	15	2	2	16
1	2	16	2	2	17

4. **a.** Are the two methods of randomization described in Section 5.1.4 equivalent?
b. Suppose you had a six-sided die and three coins. How would you assign plots to one of four treatments? three rows and two columns? eight treatments?

5. Do the four hospitals considered in Exercise 6 of Chapter 2 have the same billing practices? Hint: Block the data before analyzing.

6. Four containers, each with 10 oysters, were randomly assigned to four stations each kept at a different temperature in the wastewater canal of a power plant. The containers were weighed before treatment and after one month in the water. Were there statistically significant differences among the stations? Can these data be analyzed by k-sample comparison methods?

Trt	Initial	Final	Trt	Initial	Final
1	27.2	32.6	3	28.9	33.8
1	31.0	35.6	3	23.1	29.2
1	32.0	35.6	3	24.4	27.6
1	27.8	30.8	3	25.0	30.8
2	29.5	33.7	4	29.3	34.8
2	27.8	31.3	4	30.2	36.5
2	26.3	30.4	4	25.5	30.8
2	27.0	31.0	4	22.7	25.9

7. You can increase the power of a statistical test in three ways: a) making additional observations, b) making more precise observations, c) adding covariates. Discuss this remark in the light of your own experimental efforts.

8. A pregnant animal was accidentally exposed to a high dose of radioactivity. Her seven surviving offspring (one died at birth) were examined for defects. Three tissue samples were taken from each animal and examined by a pathologist. What is the sample size?

9. Show that if T' is a monotonic increasing function of T, that is, T' always increases when T increases, then a test based on the permutation distribution of T' will accept or reject only if a permutation test based on T also accepts or rejects.

10. Should your tax money be used to fund public television? When a random sample of adults were asked for their views on a 9-point scale (1 is very favorable and 9 is totally opposed) the results were as follows:

3, 4, 6, 2, 1, 1, 5, 7, 4, 3, 8, 7, 6, 9, 5.

The first 10 of these responses came from the City, and the last five came from the Burbs. Are there significant differences between the two areas? Given that two-thirds of the population live in the City, provide a point estimate and confidence interval for the mean response of the entire population. Hint: modify the test statistic so as to weight each block proportionately.

11. Take a third look at the experimental approaches described in Section 3.4.2 and Exercise 7 of Chapter 3. What might be the possible drawbacks of each? How would you improve on the methodology?

12. Your company is considering making a take-over bid for a marginally profitable line of retail stores. Some of the stores are real moneymakers; others could be a major drain on your own company's balance sheet. You still haven't made up your mind, when you get hold of the following figures based on a sample of 15 of the stores:

Store Size	Percent Profit
Small	7.0, 7.3, 6.5, 6.4, 7.5
Medium	7.3, 5.0, 5.5, 5.2, 6.8
Large	5.7, 5.1, 5.9, 9.2, 9.5

Have these figures helped you make up your mind? Write up a brief report for your CEO summarizing your analysis.

13. Using the insecticide data in Table 5.9, test for independence of action using only the zero and highest dose level of the first drug. Can you devise a single test that would utilize the data from all dose levels simultaneously?

14. Aflatoxin is a common and undesirable contaminant of peanut butter. Are there significant differences among the following brands?

Snoopy	0.5	7.3	1.1	2.7	5.5	4.3
Quick	2.5	1.8	3.6	5.2	1.2	0.7
Mrs. Good's	3.3	1.5	0.4	4.8	2.2	1.0

The aflatoxin content is given in ppb.

15. In how many different ways can we assign 9 subjects to 3 treatments, given equal numbers in each sample? What if we began with 6 men and 3 women and wanted to block our experiment?

16. Unequal sample sizes? Take a second look at the data of Section 5.6.2. We seem to have three different samples with three different sample sizes, each drawn from a continuous domain of possible values. Or do we? From the viewpoint of the bootstrap, each sample represents our best guesstimate of the composition of the larger population from which it was drawn. Each hypothetical population appears to consist of only a finite number of distinct values, a different number for each of the different populations. Discuss this seeming paradox.

17. Without thinking through the implications, you analyze your data from a matched pairs experiment as if you had two independent samples and obtain a significant result. Consequently, you decide not to waste time analyzing the data correctly. Right or wrong? [Hint: Were the results within each matched pair correlated? What if one but not both of the observations in a matched pair were missing?]

6

Categorical Data

In many experiments and in almost all surveys, many if not all the results fall into categories rather than being measurable on a continuous or ordinal scale: male vs. female, black vs. Hispanic vs. oriental vs. white, in favor vs. against vs. undecided. The corresponding hypotheses concern proportions: "Blacks are as likely to be Democrats as they are to be Republicans." Or, "the dominant genotype 'spotted shell' occurs with three times the frequency of the recessive." On occasion, for example when survey responses are recorded on a Likert scale, hypotheses may concern relative valuations. In this chapter, you learn to test hypotheses concerning categorical and ordinal data.

6.1 Fisher's Exact Test

As an example, suppose on examining the cancer registry in a hospital, we uncover the following data that we put in the form of a 2×2 *contingency table* (Table 6.1).

The 9 denotes the nine males who survived, the 1 denotes the remaining male who died, and so forth. The four marginal totals or *marginals* are 10, 14, 13, and 11. The total number of men in the study is 10; the total number of women is 14, and so forth.

We see in this table an apparent difference in the survival rates for men and women: Only 1 of 10 men died following treatment, but 10 of the 14 women failed to survive. Is this difference statistically significant?

The answer is yes. Let's see why, using the same line of reasoning that Fisher advanced at the annual Christmas meeting of the Royal Statistical Society in 1934. (After Fisher's talk was concluded, incidentally, a second speaker compared Fisher's talk to "the braying of the Golden Ass." I hope you will take more kindly to my own explanation.) Contingency Table 6.1 has several fixed elements:

- the total number of men in the survey, 10,
- the total number of women, 14,
- the total number who died, 11,
- and the total number who survived, 13.

These totals are immutable; no swapping of labels will alter the total number of individual men and women or bring back the dead. But these totals do not determine the contents of the table.

Suppose that before completing the table, we only had a list of patients identified by name and sex, but could not tell without further investigation which patients were still alive and which were dead. Suppose now I were to hand you 11 labels with the word “dead” and 12 labels with the word "alive." Under the null hypothesis, you are allowed to distribute these labels to the patients independently of their sex.

The result might have been the original Table 6.1 or either of the two Tables 6.2 or 6.3 whose marginals are identical with those of our original table.

Table 6.1.

	Survived	Died	Total
Men	9	1	10
Women	4	10	14
Total	13	11	24

Table 6.2.

	Survived	Died	Total
Men	10	0	10
Women	3	11	14
Total	13	11	24

Table 6.3.

	Survived	Died	Total
Men	8	2	10
Women	5	9	14
Total	13	11	24

The first of these tables makes a strong case for the superior fitness of the male, stronger even than our original observations. In the second table, the survival rates for men and women are more alike than they were in our original table.

These tables are not equally likely, not even under the null hypothesis. Table 6.2 could have arisen in any of 13 choose 3 ways, Table 6.3 in any of $\binom{13}{8}\binom{11}{2}$ ways.

Fisher would argue that if the survival rates were the same for both sexes, then each of the redistributions of labels to subjects, that is, each of the N possible contingency tables with these same four fixed marginals, is equally likely, where

$$N = \sum_{x=0}^{10} \binom{13}{x}\binom{11}{10-x} = \binom{24}{10}.$$

How did we get this value for N? The component terms are taken from the *hypergeometric* distribution:

$$\sum_{x=0}^{t} \binom{m}{x}\binom{n}{t-x} \Big/ \binom{m+n}{t}, \tag{6.1}$$

where n, m, t, and x occur as the indicated elements in the following 2×2 contingency Table 6.4

Table 6.4.

	Category 1	Category 2	Total
Category A	x	$M - x$	m
Category B	$t - x$		n
Total	t		$m + n$

If men and women have the same probability of surviving, then all tables with the marginals m, n, t are equally likely, and $\sum_{k=0}^{t-x} \binom{m}{t-k} \binom{n}{k}$ are as or more extreme.

In our example, $m = 13$, $n = 11$, $x = 9$, and $t = 10$, so that $\binom{14}{10} \binom{10}{1}$ of the N tables are as extreme as our original table and $\binom{14}{11} \binom{10}{0}$ are more extreme. The resulting sum is still only a very small fraction of the total N, so we conclude that a difference in survival rates as extreme as the difference we observed in our original table is very unlikely to have occurred by chance. We reject the hypothesis that the survival rates for the two sexes are the same and accept the alternative that, in this instance at least, males are more likely to profit from treatment.

6.1.1 Computing Fisher's Exact Test

It would be difficult to find statistical software that does *not* include Fisher's Exact Test. Thus, only the *R* code is provided in this section. To execute most of the other procedures described in this chapter, it is necessary to download a trial copy of StatXact from `http://www.cytel.com/Downloads/Default.asp`.

R

To simplify the programming, assume that the smallest marginal is in the first row and the smallest cell frequency is located in the first column, and that we have the actual cell frequencies

- `data =c(f11, f12, f21, f22)`
- `m = data[2] + data[4]`
- `n = data [1] + data [3]`
- `t = data[1] + data[2]`
- `ntab=0`
- `for (k in 0:data[1]) ntab = ntab + comb(m,t-k)*comb (n,k)`
- `ntab/comb(m+n,t) # prints the p-value for Fisher's Exact Test`

```
where fact = function(n)prod (1:n)
and comb = function (n,t) fact(n)/(fact(t)*fact(n-t)).
```

In S-PLUS, we would substitute **`choose()`** for **`comb()`**.

6.1.2 One-Tailed and Two-Tailed Tests

In the preceding example, we tested the hypothesis that survival rates do not depend on sex against the alternative that men diagnosed as having cancer are likely to live longer than women similarly diagnosed. We rejected the null hypothesis because only a small fraction of the possible tables were as extreme as the one we observed initially. This is an example of a one-tailed test. Or is it? Wouldn't we have been just as likely to reject the null hypothesis if we had observed a table similar to Table 6.5?

Table 6.5.

	Survived	Died	Total
Men	0	10	10
Women	13	1	14
Total	13	11	24

Of course, we would. In determining the significance level in the present example, we must add together the total number of tables that lie in either of the two extremes or tails of the permutation distribution.

McKinney et al. [1989] reviewed some 70 plus articles that appeared in six medical journals. In over half these articles, Fisher's Exact Test was applied improperly. Either a one-tailed test had been used when a two-tailed test was called for or the authors of the paper simply hadn't bothered to state which test they had used.

When you design an experiment, decide at the same time whether you wish to test your hypothesis against a two-sided or a one-sided alternative. A two-sided alternative dictates a two-tailed test; a one-sided alternative dictates a one-tailed test.

As an example, suppose we decide to do a follow-on study of the cancer registry to confirm our original finding that men diagnosed as having tumors live significantly longer than women similarly diagnosed. In this follow-on study, we have a one-sided alternative. Thus, we would analyze the results using a one-tailed test rather than the two-tailed test we applied in the original study.

> Warning: McKinney et al. [1989] report that more than half the published articles that apply Fisher's Exact Test fail to make the correct distinction between one-tailed and two-tailed tests. Don't you make the same mistake!

6.1.3 The Two-Tailed Test

Unfortunately, it is not obvious which tables should be included in the second tail. Is Table 6.5 as extreme as Table 6.2? We need to define a test statistic to serve as a basis of comparison. One commonly used measure is the Pearson χ^2 statistic defined for the 2×2 contingency table after eliminating elements that are invariant under permutations as $[x - tm/(m+n)]^2$. This statistic is proportional to the square of the difference between what one would expect to observe if the survival rates are the same, that is, $tm/(m+n)$ and the frequency x actually observed. For Table 6.1, this statistic is approximately 13, for Table 6.5, it is approximately 29. We leave it to you to do the

computations to show that using the chi-square criteria Table 6.6 is more extreme than Table 6.1, but Table 6.7 is not.

Table 6.6.

	Survived	Died	Total
Men	1	9	10
Women	12	2	14
Total	13	11	24

Table 6.7.

	Survived	Died	Total
Men	2	8	10
Women	11	3	14
Total	13	11	24

The Pearson statistic is far from being our only choice; among the others are the likelihood ratio statistic

$$x \log[xtm/(m+n)]$$

and Fisher's statistic

$$-2\log(h[y]) - \log[2.51(m+n)^{-3/2}(mnt)^{1/2}(m+n-t)^{1/2}],$$

where $h[y]$ is the proportion of all tables with the same marginals that have precisely the same four entries. For very large samples with a large number of observations in each cell, all three statistics lead to the same conclusion.

6.1.4 When to Accept

A problem with any of the methods we've used so far is that we're very unlikely to observe 0.05 or 0.01 exactly. If the number of observations is small, p-values may jump from 0.040 (as in Table 6.3) to 0.185 (as in Table 6.8) as a result of a single additional case.

Table 6.8.

	Survived	Died	Total
Men	7	3	10
Women	6	8	14
Total	13	11	24

What is the appropriate criteria for rejection of the null hypothesis? Limiting ourselves to 4% of the tables means we may accept when we should reject. Rejecting for 18% means we are rejecting far more often than we should. There are at least six solutions:

1. Deliberately err on the conservative side. See Boschloo [1970] and McDonald, Davis, and Miliken [1977] for some slight improvements on this approach.
2. Randomize on the boundary. If you get a p-value of 0.185 and the next closest value would have been 0.040, let the computer choose a random number between 0 and 1 for you. If this number is less than (0.05–0.04)/(0.185–0.040), reject the hypothesis at the 5% level, accept it otherwise. This is a great technique for abstract mathematics, but I don't recommend it for use in practice.
3. Use the mid-p value. Let p be equal to half the probability of the table you actually observe plus all of the probability of more extreme results.
4. Present your audience with the data and the p-value you calculated. Let them make up their own minds whether it is a significant result or not. See Kempthorne [1977].
5. Make use of a back-up statistic; see Streitberg and Roehmel [1990] and Cohen and Sackrowitz [2003].
6. Conduct a sensitivity analysis, Dupont [1986]. Add a single additional case to one of the cells (say the cell that has the most observations already, so your addition will have the least percentage impact). Does the p-value change appreciably? This approach is particularly compelling if you are presenting statistical evidence in a courtroom as it turns impersonal percentages into individuals; see Good [2001].

6.1.5 Is the Sample Large Enough?

In the last chapter, we reviewed the close relation between significance level, power, and sample size. A question arises, particularly when we do *not* reject the null hypothesis as to whether the sample was large enough. Of course, we ought to have addressed this question *before* we collected the data.

Table 6.9.

	Drug A	Drug B
Response	5	9
No Response	5	1

Consider Table 6.9 in which we've recorded the results of a comparison of two drugs. It seems obvious that Drug B offers significant advantages over Drug A. A chi-square analysis by parametric means in which the value of the chi-squared statistic is compared with a table of the chi-square distribution yields an erroneous p-value of 3%. But Fisher's Exact Test yields a one-sided p-value of only 7%.

Actually, we were quite fortunate in getting a p-value this small. Suppose the probability of a response with Drug A really is 50% and the probability of a response with Drug B is 0.9; using the Stata command, `sampsi 0.5 0.9, n1(10) r(1)`, we learn that the power of Fisher's Exact Test to detect this alternative with this small a sample size is only 29%!

How large a sample size would we need to obtain a power of 80% for comparing proportions of 0.5 and 0.9? According to StatXact, we'd need at least 24 observations in a balanced design.

```
Using Stata to Estimate the Power of Fisher's Exact
     Test

. sampsi 0.5 0.9, n1(10) r(1)
Estimated power for two-sample comparison of
     proportions
Test Ho: p1 = p2, where p1 is the proportion in
                              population 1
                    and p2 is the proportion in
                              population 2
Assumptions:
         alpha = 0.0500  (two-sided)
            p1 = 0.5000
            p2 = 0.9000
sample size n1 =      10
            n2 =      10
         n2/n1 =    1.00

Estimated power:
         power = 0.2907
```

6.2 Odds Ratio

In most instances, we won't be satisfied with merely rejecting the null hypothesis but will want to make a more powerful statement such as "men are twice as likely as women to get a good-paying job," or "women under thirty are twice as likely as men over 40 to receive an academic appointment."

For the survival data of Table 6.1, the odds ratio $\frac{\pi_w}{1-\pi_w} / \frac{\pi_m}{1-\pi_m}$ is 0.044 with a 90% confidence interval using the method of Gart [1970] extending from 0.002 to 0.44.

In the discrimination case of Fisher versus Transco Services of Milwaukee [1992], the plaintiffs claimed that Transco was 10 times as likely to fire older employees. Can we support this claim with statistics? The Transco data are in Table 6.10.

Table 6.10.

	Transco Employment	
Outcome	Young	Old
Fired	13	1
Retained	13	11

Let π_y denote the probability of firing a young person and π_o the probability of firing an older person. We want to go beyond testing the null hypothesis $\pi_y = \pi_o$ to determine a confidence interval for the odds ratio $\frac{\pi_o}{1-\pi_o} / \frac{\pi_y}{1-\pi}$. We turn for aid to StatXact, a statistical package whose emphasis is the analysis of categorical and ordinal data. Choose in turn Statistics, Two Binomials, and "CI. Odds Ratio" from successive StatXact menus.

(See Figures 6.1 and 6.2.) Based on the results depicted in the accompanying sidebar, we can tell the judge that older workers were fired at a rate at least 1.65 times the rate at which younger workers were discharged.

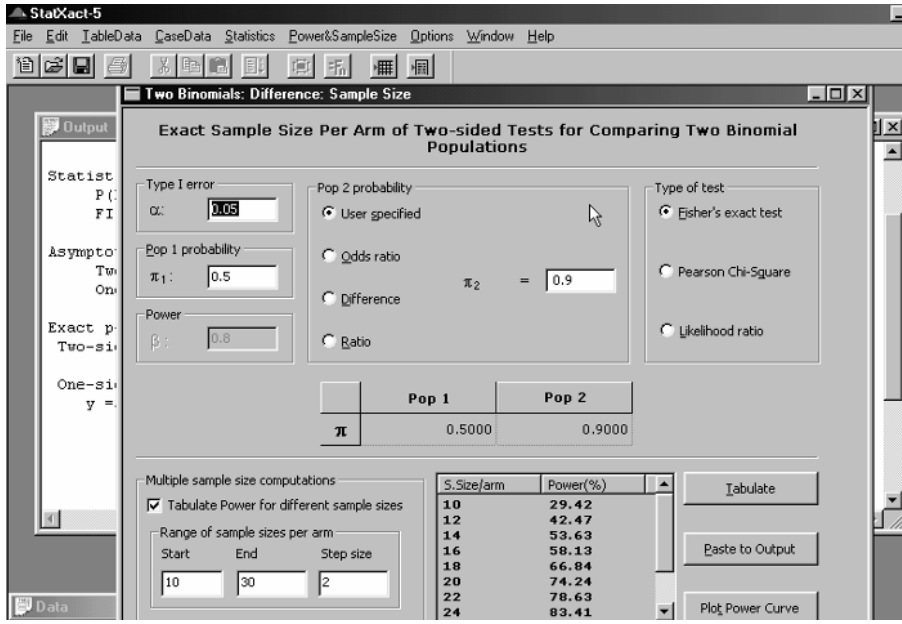


Fig. 6.1. Using StatXact to determine required sample size.

```
        ODDS RATIO OF TWO BINOMIAL PROPORTIONS

StatXact Output
Datafile: C:\EXAMPLES\TRNSCO.CY3
Statistic based on the observed 2 by 2 table :

  Binomial proportion for column <young > :
       pi_1 = 0.04000
  Binomial proportion for column <Old   > :
       pi_2 = 0.3704
                    (pi_2 )/( 1-pi_2 )
  Odds Ratio      = ------------------- = 14.12
                    (pi_1 )/( 1-pi_1 )
Results:
Method                      P-value      95.00%
                           (2-sided)  Confidence Interval
Asymp (Mantel-Haenszel)     0.0157      (1.649,    120.9)
Exact                       0.007145    (1.649,    637.5)
```

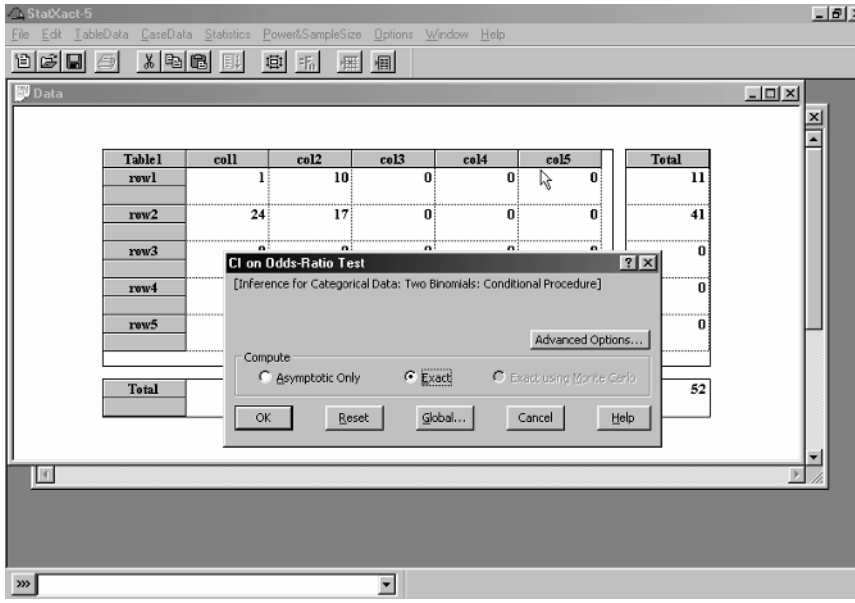


Fig. 6.2. Using StatXact to determine a confidence interval for an odds ratio.

6.2.1 Stratified 2×2's

In trying to develop a cure for a relatively rare disease, we face the problem of having to gather data from a multitude of test centers, each with its own set of procedures and its own way of executing them. Before we can combine the data, we must be sure the odds ratios across the test centers are approximately the same. Consider the set of results in Table 6.11 obtained by the Sandoz drug company and reproduced with permission from the StatXact manual. One of the cites, number 15, stands out from the rest. But is the difference statistically significant?

Similar problems are encountered in studies where test subjects may use one of several different test apparatuses or be examined by one of several different examiners. We cannot combine results from the different machines or the different technicians who operated them until we've performed an initial test of equivalence.

With 22 contingency tables, the number of computations needed to examine all rearrangements is in the billions. Fortunately, StatXact utilizes several time-saving linear programing algorithms, including the one introduced in Mehta, Patel, and Senchaudhuri [1988], to obtain a Monte Carlo estimate of the significance level. We pull down menus Statistics, Stratified 2×2 Tables, and Homogeneity of Odds Ratios. The estimated p-value of 0.013, just a fraction greater than 1%, tells us it would be unwise to combine the results from the different sites.

The output of this program provides us with one more important finding: Displayed above the Monte Carlo estimate of the exact p-value, 0.01237, is the asymptotic or large-sample approximation based on the chi-square distribution. Its value, 0.0785, is many times larger than the correct value, and relying on this so-called approximation would have led us to a completely different and erroneous conclusion.

Table 6.11. Sandoz Drug Data

Test Site	New Drug		Control Drug	
	Response	#	Response	#
1	0	15	0	15
2	0	39	6	32
3	1	20	3	18
4	1	14	2	15
5	1	20	2	19
6	0	12	2	10
7	3	49	10	42
8	0	19	2	17
9	1	14	0	15
10	2	26	2	27
11	0	19	2	18
12	0	12	1	11
13	0	24	5	19
14	2	10	2	11
15	0	14	11	3
16	0	53	4	48
17	0	20	0	20
18	0	21	0	21
19	1	50	1	48
20	0	13	1	13
21	0	13	1	13
22	0	21	0	21

```
          TEST FOR HOMOGENEITY OF ODDS RATIOS

              [18  2 × 2 informative tables]

StatXact Output
Datafile: C:\EXAMPLES\SANDOZ.CY3

Observed Statistics:

    BD: Breslow and Day Statistic = 25.78
    ZE: Zelen Statistic = 9.481e-009

Asymptotic p-value: (based on Chi-Square distribution
     with 17 df)

   Pr { BD .GE. 25.78 } = 0.0785

Monte Carlo estimate of p-value:

  Pr { ZE .GE. 9.481e-009 } = 0.0127
  99.00% Confidence Interval = (0.0119, 0.0135)
```

6.3 Exact Significance Levels

The preceding result is not an isolated one. Asymptotic approximations are to be avoided except with very large samples.[1] Table 6.12 contains data on oral lesions in three regions of India derived from Gupta et al. [1980] by Mehta and Patel. We want to test the hypothesis that the location of oral lesions is unrelated to geographical region. Possible test statistics include Freeman–Halton (see Section 6.4), chi-square, and the log-likelihood ratio $\Sigma\Sigma f_{ij} \log[f_{ij} f_{..}/f_{i.} f_{.j}]$.

We may calculate the exact significance levels of these test statistics by deriving their permutation distributions or use asymptotic approximations obtained from tables of the chi-square statistic. Table 6.13 taken from the StatXact manual compares the various approaches.

Table 6.12. Oral Lesions in Three Regions of India

Site of Lesion	Kerala	Gujarat	Andh
Labial Mucosa	0	1	0
Buccal Mucosa	8	1	8
Commissure	0	1	0
Gingiva	0	1	0
Hard Palate	0	1	0
Soft Palate	0	1	0
Tongue	0	1	0
Floor of Mouth	1	0	1
Alveolar Ridge	1	0	1

Table 6.13. Three Tests of Independence

Statistic	Chi-Square	Freeman–Halton	Log-Likelihood
Exact p-value	0.0269	0.0101	0.0356
Tabulated p-value	0.1400	0.2331	0.1060

The exact significance level varies from 1% to 3.5% depending on which test statistic we select. Tabulated p-values based on asymptotic approximations vary from 11% to 23%. Using the Freeman–Halton statistic, the permutation test tells us the differences among regions are significant at the 1% level; the large-sample approximation says no, they are insignificant even at the 20% level. The permutation test is correct. The large-sample approximation is grossly in error. With so many near-zero entries in the original contingency table, the chi-square large-sample approximation is not appropriate.

6.4 Unordered $r \times c$ Contingency Tables

With a computer at hand, the principal issue in the analysis of a contingency table with more than two rows and two columns is deciding on an appropriate test statistic. Halter

[1] Not too surprising, as "asymptotic" in this context means "as the sample size grows infinitely large."

Do It Yourself?

You could write a computer program to perform the tests described in this chapter, one that would select from the set of all tables with a given set of marginals to generate the permutation distribution, but there's a more efficient way. Branch-and-bound algorithms developed by Mehta and Patel [1980,1983] use a network approach to enumerate only those tables that have a more extreme value of the test statistic than the original. An outline of their method is given in Section 14.5 of Good [2004]. StatXact is the only version of these algorithms that is commercially available at present, and in this chapter, you learn how to use StatXact to do each of the needed tasks.

[1969] showed we can find the probabilities of any individual $r \times c$ contingency table through a straightforward generalization of the hypergeometric distribution given in Equation 6.1. An $r \times c$ contingency table consists of a set of frequencies

$$\{f_{ij},\ 1 \le i \le r; \quad 1 \le j \le c\}$$

with row marginals $\{f_{i.},\ 1 \le i \le r\}$ and column marginals $\{f_{.j},\ 1 \le j \le c\}$.

Suppose once again we have mixed up the labels. To make matters worse, this time every item/subject is to be assigned both a row label ($f_{1.}$ of which are labeled row 1, $f_{2.}$ of which are labeled row 2 and so forth) and a column label. Let P denote the probability with which a specific table assembled at random will have these exact frequencies.

$P = Q/R$ where

$$Q = \prod_{i=1}^{r} f_{i.}! \prod_{j=1}^{c} f_{.j}! f_{..}! \text{ and } R = \prod_{i=1}^{r} \prod_{j=1}^{c} f_{ij}!$$

An obvious extension of Fisher's Exact Test is the Freeman and Halton [1951] test based on the proportion p of tables for which P is greater than or equal to P_o for the original table.

It's not as obvious this extension offers any protection against the alternatives of interest. Just because one table is less likely than another under the null hypothesis does not mean it is going to be more likely under the alternatives of interest to us. Consider the 1×3 contingency table $\boxed{f_1}\boxed{f_2}\boxed{f_3}$ which corresponds to the multinomial with probabilities $p_1 + p_2 + p_3 = 1$; the table whose entries are 1, 2, 3 argues more in favor of the null hypothesis $p_1 = p_2 = p_3$ than of the ordered alternative $p_1 > p_2 > p_3$.

The classic statistic for independence in a contingency table with r rows and c columns is the Pearson chi-square statistic

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} (f_{ij} - E[f_{ij}])^2 / E[f_{ij}],$$

where $E(f_{ij})$ is the number of observations in the ijth category one would expect on theoretical grounds.

With very large samples, this statistic has the chi-square distribution with $(r-1)(c-1)$ degrees of freedom. But in most practical applications, the chi-square distribution is only an approximation to the distribution of this statistic and is notoriously inexact for small and unevenly distributed samples.

The permutation statistic based on the proportion of tables for which the Pearson chi-square statistic is greater than or equal to its value for the original table provides an exact test and possesses all the advantages of the original chi-square. The distinction between the two approaches is that with the chi-square test as it is described in most textbooks we look up the significance level in a table, while with the permutation statistic, we derive the significance level from the permutation distribution. With large samples, the two approaches are equivalent, as the permutation distribution converges to the tabulated distribution; see Bishop, Fienberg, and Holland [1975; chapter 14].

To obtain the results displayed in Table, choose from StatXact's main menu first "Statistics," and then "Unordered $R \times C$ Table 6.13." One then has a choice of "Pearson's Chi-Square", "Likelihood ratio", or "Fisher-Freeman-Halton."

6.4.1 Test of Association

Regardless of which statistic we choose, these three permutation tests have one of the original chi-square test's disadvantages: while they offer global protection against a wide variety of alternatives, they offer no particular protection against any single one of them. Row and column categories are treated equivalently and no attempt is made to distinguish between cause and effect. To address this deficiency, Goodman and Kruskal [1954] introduce an asymmetric measure of association for nominal scale variables called tau τ which measures the proportional reduction in error obtained when one variable, the "cause" or independent variable, is used to predict the other, the "effect" or dependent variable.

Assuming the independent variable determines the row,

$$\tau = \frac{\sum_j \max_i f_{ij} - \max_i f_i}{\max_{ij} f_{..} - \max_i f_{i.}}.$$

$\tau = 0$ when the variables are independent; $\tau = 1$ when for each category of the independent variables all observations fall into exactly one category of the dependent. These points are illustrated in the following 2×3 tables:

3	6	9
6	12	18

$\tau = 0$

18	0	0
0	36	0

$\tau = 1$

3	6	9
12	18	6

$\tau = 0.057$

To obtain this latter result from StatXact, select first "Statistics," and then "Nominal Response" (at the foot of the Statistics menu). A permutation test of independence in

```
            StatXact Measures of Association

       Nominal Response:Goodman-Kruskal Tau Test


GOODMAN AND KRUSKAL TAU
Coefficient estimates based on 54 observations.

   Coefficient   Estimate    ASE1    95% Conf.Interval
   G-K tau(C|R)   0.0571    0.0457   (0.0000, 0.1467)

Asymptotic p-value based on Chi-Square distribution
     with 2 df
         Pr { Statistic .GE. Observed  } =      0.0484

Exact p-value
         Pr { Statistic .GE. Observed  } =      0.0714
```

this latter table is based upon the proportion of tables with the same marginals for which τ is less than 0.057.

Cochran's Q provides an alternate test for independence. Suppose we have I experimental subjects on each of whom we administer J tests. Let $y_{ij} = 1$ or 0 denote the outcome of the jth test on the ith patient, e.g., if the test is positive, y is set to 1 and is set to 0 otherwise. Define $R_i = jy_{ij}$; $Cj = iy_{ij}$;

$$Q = \frac{\sum_j (C_j - C)^2}{R - \sum_i R_i^2}$$

Which Test?

The Data Are in Categories and the Categories Can't Be Ordered

There are exactly two rows and two columns.
Use Fisher's Exact Test.
There are more than two rows and at least two columns.
You want to test whether the relative frequencies are the same in each row and in each column.
Use the Freedman–Halton Test or use chi-square.
You want to test whether the column frequencies depend on the row.
Use τ or Q.

6.4.2 Causation Versus Association

A significant value for any of the above statistics only means that the variables are associated. It does not mean that there is a cause and effect relationship between them. They may both depend on a third variable omitted from the study.

Regrettably, the converse is also true. A third omitted variable may also result in two variables appearing to be dependent when the opposite is true. Consider Table 6.14 whose entries provide an example of what is termed Simpson's paradox:

Table 6.14. Sexes Combined

	control	treated
Alive	6	20
Dead	6	20

We don't need a computer program to tell us the treatment has no effect on the death rate. Or does it? Consider the tables, 6.15 and 6.16, that result when we examine the men and women separately:

Table 6.15. Men Only

	control	treated
Alive	4	8
Dead	3	5

Table 6.16. Women Only

	control	treated
Alive	2	12
Dead	3	15

In the first of these tables, treatment reduces the male death rate from 0.43 to 0.38. In the second, from 0.6 to 0.55. Both sexes show a reduction, albeit neither reduction is statistically significant, yet the combined population does not. Resolution of this paradox is accomplished by avoiding a knee-jerk response to statistical significance when association is involved. One needs to think deeply about underlying cause and effect relationships before analyzing data. Thinking about cause and effect relationships in the preceding example might have led us to thinking about possible sexual differences, and to employing the analytical techniques described in Section 6.2.

6.5 Ordered Statistical Tables

Suppose we have conducted a survey in which respondents were asked to rate themselves on a discrete ordinal scale. See, for example, Table 6.17, which summarizes data gathered by Graubard and Korn [1987] in a survey of maternal alcohol consumption.

In contrast to data that are measured on a continuous basis, ties with ordinal and categorical data are inevitable, the rule, not the exception. Moreover, it is uncertain what value we ought to assign to each of the ordered categories. Among the leading choices for a scoring method are the following:

Table 6.17. Data gathered by Graubard and Korn [1987]

	Maternal Alcohol Consumption (drinks/day)					
	0	< 1	1 to 2	3–5	> 6	Total
Malformation Absent	17066	14464	788	126	37	
Malformation Present	48	38	5	1	1	

i) the category number: 1 for the first category, 2 for the second and so forth,
ii) the midrank scores,
iii) scores determined by the domain expert—a biologist, a physician, a physiologist.

To show how such scores might be computed, consider the following 1×2 contingency table:

	Alcohol Consumption	
Drinks/day	0	1 or 2
Frequency	3	5

The category or equidistant scores are 1 and 2. The ranks of the 8 observations are 1 through 3, and 4 through 8, so that the midrank score of those in the first category is 2, and in the second 6. While user-chosen scores, based on alcohol consumption, might be 0 and 1.5.

To analyze grouped data such as that in Table 6.17 using *R*, we proceed as follows:

```
MalAbs=c(17066,  14464, 788, 126, 37)
MalPres=c(48, 38, 5, 1, 1)
NumCat=5
Score = c(1: NumCat)
Samp1=c(rep(Score[1], MalAbs[1]))
for (i in 2: NumCat)Samp1 = c(Samp1, rep(Score[i],
     MalAbs[i]))
#Create Samp2 in similar fashion, then analyze as
     in Section 3.4.2.

#If you wish to use midrank scores instead, insert the
     following code
Score=numeric(NumCat)
TotFreq=MalAbs +MalPres
Total = 0
for (i in 1: NumCat){
     Score[i]= Total +TotFreq[i]/2
     Total = Total +TotFreq[i]
     }
```

6.5.1 More Than Two Rows and Two Columns

Two cases need be considered. The first when the columns but not the rows of the table may be ordered (the other variable being purely categorical), and the second when both columns and rows can be ordered.

Singly Ordered Tables

Our approach parallels that of Section 5.2 in which we describe a k-sample comparison of metric data. Our test statistic is

$$F_2 = \sum (S_i - \overline{S})^2 \quad \text{where} \quad S_i = \sum g[j] f_{ij}.$$

As in the case of the $2 \times c$ table our problem is in deciding on the appropriate scoring function $g[]$.

Table 6.18 provides tumor regression data for five chemotherapy regimes. After questioning an oncologist (the domain expert), we learned that a partial response corresponds to approximately two years in remission (about 100 weeks) and a complete response to an average of three years (150 weeks). Subsequently, we assigned scores of 0, 100, and 150 to the ordered response categories.

Table 6.18. Response to Chemotherapy

	None	Partial	Complete
CTX	2	0	0
CCNU	1	1	0
MTX	3	0	0
CTX + CCNU	2	2	0
CTX + CCNU + MTX	1	1	4

To use StatXact to analyze the tumor data, we first click on TableData in the main menu, click on Settings, then enter Column Scores 0, 100, and 150. To execute the analysis, we select in turn Statistics, Singly Ordered $R \times C$ Table, ANOVA with Arbitrary Scores, and Exact test method. The results are tabulated below.

```
ANOVA TEST [that the 5 rows are identically distributed]

  Datafile: C:\EXAMPLES\TUMOR.CY3
  Statistic based on the observed data :
     The Observed Statistic =    7.507
  Asymptotic p-value: (based on Chi-square distribution
       with 4 df )
    Pr { Statistic .GE.  7.507 } =  0.0747
  Monte Carlo estimate of p-value :
    Pr { Statistic .GE.  7.507 } =  0.0444
    99.00% Confidence Interval  = ( 0.0350, 0.0538)
```

Although our estimated significance level is less than 0.0444, a 99% confidence interval for this estimate, based on a sample of 3200 possible rearrangements, does include values greater than 0.05. We can narrow this confidence interval by sampling additional rearrangements. With a Monte Carlo of 10,000 sample tables, our estimate of the p-value is 0.0434 with a 99% confidence interval of (0.0382, 0.0486). Of course, the calculations take three times as long. When I first perform a permutation test, I

use as few as 400 to 1600 simulations. If the results are equivocal as they are in this example, then and only then will I run 10,000 simulations.

Doubly Ordered Tables

In an $r \times c$ contingency table conditioned on fixed marginal totals, the outcome depends on the $(r-1)(c-1)$ odds ratios

$$\phi_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}},$$

where π_{ij} is the probability of an individual being classified in row i and column j.

In a 2×2 table, conditional probabilities depend on a single odds ratio and hence one- and two-tailed tests of association are easily defined. In an $r \times c$ table there are potentially $n(r-1)(c-1)$ sets of extreme values, two for each of the $(r-1)(c-1)$ odds ratios. An omnibus test for no association, e.g., χ^2, might have as many as 2^n tails.

Following Patefield [1982], we consider tests of the null hypothesis of no association between row and column categories $\phi_{ij} = 1$ for all i, j against the alternative of a positive trend $\phi_{ij} \geq 1$ for all i, j.

The principal test statistic considered by Patefield, also known as the linear-by-linear association test, is

$$\Lambda = \Sigma\Sigma f_{ij} r_i c_j,$$

where $\{r_i\}$ and $\{c_j\}$ are user-chosen row and column scores.[2]

6.6 Summary

In this chapter, you were introduced to the concept of a contingency table with fixed marginals, and shown you could test against a wide variety of general and specific alternatives by examining the resampling distribution of the appropriate test statistic. Among the test statistics you considered were Fisher's Exact, Freedman–Halton, Pearson's Chi-Square, Tau, Q, Pitman's correlation, and linear-by-linear association. These latter two statistics are to be used when you can take advantage of an ordering among the categories.

6.7 To Learn More

Excellent introductions to the analysis of contingency tables may be found in Agresti [1990, 1992], and in the StatXact manual authored by Mehta and Patel. Major advances in analysis by resampling means have come about through the efforts of Gail and Mantel [1977], Mehta and Patel [1983], Mehta, Patel, and Senchaudhuri [1988], Baglivo, Olivier, and Pagano [1988], and Smith, Forester, and McDonald [1996].

[2] This statistic is actually just another form of Mantel's U, perhaps the most widely used of all multivariate statistics. See Chapter 7.

Berkson [1978], Basu [1979], Haber [1987], and Mielke and Berry [1992] examine Fisher's Exact Test. The power of the Freeman–Halton statistic in the $r \times 2$ case is studied by Krewski, Brennan, and Bickis [1984]. Details of the calculation of the distribution of Cochran's Q under the assumption of independence are given in Patil [1975]. For a description of some other, alternative statistics for use in $r \times c$ contingency tables, see Nguyen [1985].

To study a 2×2 table in the presence of a third covariate, see Bross [1964] and Mehta, Patel, and Gray [1985].

6.8 Exercises

1. A total of 40,500 babies born in 1981 in the United States died before they were 28 days old. Of these babies, 30,000 were white, and 10,500 were nonwhite. Comment on the hypothesis that black kids have a better chance to survive in North America than white kids.
2. A preliminary poll conducted well before the 2004 U.S. presidential election yielded the following results: Among Men, Bush 49%, Kerry 49%. Among Women, Bush 40%, Kerry 60%. Are the differences between the sexes significant?
3. Referring to Table 6.11, if Sandoz excluded cite 15 from their calculations, could they safely combine the data from the remaining cites?
4. a. Will encouraging your child promote his or her intellectual development? A sample of 100 children and their mothers were observed and the children's IQs tested at 6 and 12 years. Before examining the data, b. Do you plan to perform a one-tailed or two-tailed test? c. What is the significance level of your proposed test?

	Mothers Encourage Schoolwork		
	Rarely	Sometimes	Never
IQ Increased	8	15	27
IQ Decreased	30	9	11

5. Does 1, 2-dichloroethane induce tumors? Consider the following data evaluated by Gart et al. [1986].

	Tumor	No Tumor
Treated	15	21
Control	2	35

6. McGill's hockey team won 11 of its 15 games last season, while your team only won 8 of its 14. Can you use Fisher's Exact Test to analyze this data? Would your answer be the same under all circumstances? What if McGill and your team compete in the same league?
7. Referring to the literature of your own discipline, see if you can find a case where an $r \times 2$ table with at least one entry smaller than 7 gave rise to a borderline p-value using the traditional chi-square approximation. Reanalyze this table using resampling methods. Did the authors use a one-tailed or a two-tailed test? Was their choice appropriate?

8. Show that once you have selected $(r-1)(c-1)$ of the entries in a contingency table with r rows and c columns the remainder of the entries are determined. (Hint: Solve in turn for 2×3, $2 \times c$, and $r \times c$ tables.)

9. Holmes and Williams [1954] studied tonsil size in children to verify a possible association with the virus *S. pyrogenes*. Do you feel there is an association? How many rows and columns in the following contingency table? Which, if any, of the variables is ordered?

Tonsil Size by Whether Carrier of *S. pyrogenes*			
	Not Enlarged	Enlarged	Greatly Enlarged
Noncarrier	497	560	269
Carrier	19	29	24

10. Based on the following table of results, would you conclude that treatment A is superior?

	A	B
Marked Improvement	7	3
Moderate	15	9
Slight	16	14
No change	13	21
Worse	1	5

11. Does dress make the woman?

	Time to First Promotion (months)		
	Long	Average	Short
Poorly Dressed	12	8	4
Well Dressed	18	25	20
Very Well Dressed	13	19	31

7

Multiple Variables and Multiple Hypotheses

The value of an analysis based on simultaneous observations on several variables such as height, weight, blood pressure, and cholesterol level is that it can be used to detect subtle changes that might not be detectable, except with very large, prohibitively expensive samples, were we to consider only one variable at a time.

Any of the resampling procedures can be applied in a multivariate setting providing we can either

a) find a single-valued test statistic that can stand in place of the multivalued vector of observations
b) combine the p-values associated with the various univariate tests into a single p-value.

Let us consider each of these approaches in turn along with the generalized quadratic form and the challenging problems of repeated measures and multiple hypotheses.

Increase the Sensitivity of Your Experiments

1. Take more observations
2. Take more precise observations
3. Block your experiments
4. Observe multiple variables

7.1 Single-Valued Test Statistic

Suppose a small college has found that the best predictor of academic success is a weighted combination of SAT scores, adjusted GPA, and the number of extracurricular activities. Then, it would use this same weighted combination when testing hypotheses concerning, say, the relative merits of in-state and out-of-state students.

In the absence of such domain specific knowledge, Hotelling's T^2, a straightforward generalization of Student's t may be used. Some notation is necessary, alas. We use bold type to denote a vector $\mathbf{X}$ whose components $(X_1, X_2, \ldots, X_J)$ have expected values $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_J)$. X_1 might be a student's GPA, X_2 the same student's SAT score and so forth. If we have collected vectors of observations for n students, then let $\bar{\mathbf{X}}$ denote the corresponding vector of sample means, and $\mathbf{V}$ the matrix whose ijth component is the covariance of X_i and X_j.

In the one-sample case, Hotelling's T^2 is defined as $(\bar{\mathbf{X}} - \mu)V^{-1}(\bar{\mathbf{X}} - \mu)^T$.

In the two-sample case, Hotelling's T^2 is defined as $(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)V^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T$ where the ijth component of $\mathbf{V}$ is estimated from the combined sample by the formula

$$V_{ij} = \frac{1}{n_1 + n_2 - 2} \sum_{g=1}^{2} \sum_{k=1}^{n_g} (x_{gik} - \bar{x}_{gi.})(x_{gjk} - \bar{x}_{gj.}).$$

This statistic weighs the contribution of individual variables and pairs of variables in inverse proportion to their covariances. This has the effect of rescaling each variable so that the most weight is given to those variables that can be measured with the greatest precision and those that can provide information not provided by the others.

While the significance level of the univariate Student's t is almost exact for not-far-from-normally distributed observations, the significance level of the multivariate Hotelling's T^2 is sensitive to departures from normality and a resampling approach is recommended [Davis 1982]. For the purpose of resampling, each vector of observations on an individual subject is treated as a single indivisible entity. When we relabel, we relabel on a subject-by-subject basis so that all observations on a single subject receive the same new label.

As with all permutation tests we proceed in three steps:

1. We compute the test statistic for the original observations,
2. We compute the test statistic for all relabelings,
3. We determine the percentage of relabelings that lead to values of the test statistic that are as or more extreme than the original value.

Note that we are forced to compute the covariance matrix $\mathbf{V}$ and, more time consuming, its inverse for each new rearrangement. To reduce the number of computations, Wald and Wolfowitz [1944] suggest a slightly different statistic W that is a monotonic function of T^2. That is, W is large when T^2 is large and vice versa:

$$\text{Let} \quad U_i = \frac{1}{n_1 + n_2} \sum_{g=1}^{2} \sum_{k=1}^{n_g} x_{gik}$$

$$c_{ij} = \sum_{g=1}^{2} \sum_{k=1}^{n_g} (x_{gik} - U_i)(x_{gjk} - U_j).$$

Let C be the matrix whose components are the c_{ij}. Then

$$W = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)C^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T.$$

As always, we ought to let the computer do the calculations for us. In this example, our objective is to see whether there are significant differences between foreign and domestic automobiles in terms of price, miles per gallon, headroom, and turning radius. The following is a subset of the Stata auto.dta. (The full data set has 74 observations and many more variables.)

price	mpg	hdrm	turn	foreign
4099	22	2.5	40	Domestic
4749	17	3.0	40	Domestic
3799	22	3.0	35	Domestic
4816	20	4.5	40	Domestic
7827	15	4.0	43	Domestic
5788	18	4.0	43	Domestic
6295	23	2.5	36	Foreign
9735	25	2.5	34	Foreign
6229	23	1.5	35	Foreign
4589	35	2.0	32	Foreign
5079	24	2.5	34	Foreign
8129	21	2.5	38	Foreign

Entering the Stata command.

.permute foreign "hotelling price mpg turn headroom, by (foreign)" r(T2), reps(400)

Yields the following output:

```
command:      hotelling price mpg turn headroom,
                   by(foreign)
statistic:    _pm_1      = r(T2)
permute var:  foreign

Monte Carlo permutation statistics Number of obs = 74
                                   Replications  = 400

T     T(obs)   c   n  p=c/n  SE(p) [95% Conf.Interval]

_pm_1 62.52314 0 400 0.0000 0.0000         0  .0091798

Note: confidence interval is with respect to p=c/n
Note: c = #{T >= T(obs)}
```

The output of Stata tells us that of 400 rearrangements, none yielded a value of Hotelling's T^2 as large as the value associated with the original observations. $p < 0.01$.

A second example of the computation of Hotelling's T^2 is given in Section 7.3.4.

7.1.1 Applications to Repeated Measures

When we do a bioequivalence study, we replace a set of discrete interdependent values with a "smooth" curve. This curve is derived in one of two ways: 1) by numerical

analysis, 2) by modeling. The first way yields a set of coefficients, the second a set of parameter estimates. Either the coefficients or the estimates may be treated as if they were the components of a multivariate vector and the methods of the previous section applied to them.

Here is an elementary example: Suppose you observe the time course of a drug in the urine over a period for which a linear model would be appropriate. Suppose further that the chief virtue of your measuring system is its low cost so that measurement errors are significant. Consequently, you take a series of measurements on each patient about half an hour apart and then use a regression technique (see Chapter 8) to derive a best-fitting curve for each patient. Following Zerbe and Walker [1977], we replace each set of measurements taken over time with the corresponding pair of regression coefficients (intercept and slope). The problem is reduced to that of a bivariate comparison between the two treatment groups and we use the methods described in the preceding section to obtain the permutation distribution of Hotelling's T^2.

The preceding is just one of the many possible experiments in which we study the development of a process over a period of time, the growth of a tumor or the gradual progress of a cure. If our observations are made by sacrificing different groups of animals at different periods of time, then time is simply another variable in the analysis that we may treat as a covariate using the methods of Chapters 8 and 9. But if all our observations are made on the same subjects, then the multiple observations on a single individual will be interdependent. And all the observations on a single subject must be treated as a single multivariate vector.

We may ask at least three questions about the response profiles:

1. Are the response profiles the same for the various treatments?
2. Are the response profiles parallel?
3. Are the response profiles at the same level?

A "yes" answer to question 1 implies "yes" answers to questions 2 and 3, but we may get a "yes" answer to 2 even when the answer to 3 is "no".

One simple test of parallelism entails computing the successive differences in value from time point to time point and then applying the methods of the preceding section to these differences. Of course, this approach is applicable only if the observations on both treatments were made at identical times.

To circumvent this limitation and to obtain a test of the narrower hypothesis of equivalent response profiles, Koziol et al. [1981] use an approach based on ranks:

Suppose there are n_g subjects in the gth treatment group and n_{gt} observations were made on these subjects at time t. Let R_{gjt} be the rank of the observation on the jth subject among these n_{gt} values.

If luck is with us so that all subjects remain with us to the end of the experiment, then $n_{gt} = n_g$ for all t and each treatment group, and we may adopt a test statistic first proposed by Puri and Sen [1966]

$$L = \sum_g n_g \bar{\mathbf{R}}_g V^{-1} \bar{\mathbf{R}}_g^T,$$

where $\bar{\mathbf{R}}_g$ is a $1 \times T$ vector whose tth component is $\bar{R}_{gt}$ and V is a $T \times T$ covariance matrix whose stth component is $\Sigma_g \Sigma_k R_{gkt} R_{gks}$.

7.1.2 An Example

Higgins and Noble [1993] analyze an experiment whose goal was to compare two methods of treating beef carcasses in terms of their effect on pH measurements of the carcasses taken over time. Treatment level B is suspected to induce a faster decay of pH values. Formally, we wish to test a hypothesis of no difference between the treatments against the alternative that $X_{B[t]}$ is stochastically smaller than $X_{A[t]}$, for some time t and no greater at other times.

Observed data are:

$t =$	0	1	2	3	4	5
A1	6.81	6.16	5.92	5.86	5.80	5.39
A2	6.68	6.30	6.12	5.71	6.09	5.28
A3	6.34	6.22	5.90	5.38	5.20	5.46
A4	6.68	6.24	5.83	5.49	5.37	5.43
A5	6.79	6.28	6.23	5.85	5.56	5.38
A6	6.85	5.51	5.95	6.06	6.31	5.39
B1	6.64	5.91	5.59	5.41	5.24	5.23
B2	6.57	5.89	5.32	5.41	5.32	5.30
B3	6.84	6.01	5.34	5.31	5.38	5.45
B4	6.71	5.60	5.29	5.37	5.26	5.41
B5	6.58	5.63	5.38	5.44	5.17	5.62
B6	6.68	6.04	5.62	5.31	5.41	5.44

Although normality of these observations may be assumed, the variances and covariances surely vary with time so that the two-way ANOVA model is not appropriate. Instead, we may proceed as follows: First, we standardize the observations, subtracting the baseline value at $t = 0$ from each one. At each time point, the resulting differences are exchangeable. Treating each time point separately, the resulting p-values are as follows:

$T = 1$	$T = 2$	$T = 3$	$T = 4$	$T = 5$
0 01056	0 000127	0 000309	0 000395	0 06803

Using Fisher's nonparametric combination rule, $F = -2\log[\Pi_i p_i] = -2\Sigma_i \log[p_i]$, the combined p-value for the global hypothesis is 0.000127. We can conclude that decay of treatment B is faster than that of A, even though at the last time point, $T = 5$, substantially the same distribution of pH values is observed ($p = 0068$).

7.2 Combining Univariate Tests

The methods described in this section have the advantage that they apply to continuous, ordinal, or categorical variables or to any combination thereof. They can be applied to one-, two- or k-sample comparisons. As in the preceding section, suppose we have made a series of exchangeable vector-valued observations on K subjects, each vector

consisting of the values of J variables. The first variable might be a 0 or 1 according to whether or not the kth seedling in the ith treatment group germinated, the second might be the height of the kth seedling, the third the weight of its fruit, the fourth a subjective appraisal of fruit quality, and so forth. With each variable is associated a specific type of hypothesis test, the type depending on whether the observation is continuous, ordinal, or categorical and whether it is known to have come from a distribution of specific form. Let $\mathbf{T}_o$ denote the vector of single-variable test statistics derived from the original unpermuted matrix of observations. These might include differences of means or weighted sums of the total number germinated, or any other statistic one might employ when testing just one variable at a time. When we rearrange the treatment labels on the observation vectors we obtain a new vector of single-variable test statistics $\mathbf{T}_\pi$.

In order to combine the various tests, we need to reduce them all to a common scale. Proceed as follows:

1. Generate S permutations of X and thus obtain S vectors of univariate test statistics.
2. Rank each of the single-variable test statistics separately. Let R_{ij} denote the rank of the test statistic for the jth variable for the ith permutation when it is compared to the value of the test statistic for the jth variable for the other $S-1$ permutations including the values for the original test statistics.
3. Combine the ranks of the various univariate tests for each permutation using Fisher's omnibus method:

$$U_i = -\sum_{j=1}^{J} \log\left[\frac{S + 0.5 - R_{ij}}{S+1}\right]; \quad i = 1, \ldots, S.$$

4. Reject the null hypothesis only if the value of U for the original observations is an extreme value.
   This straightforward, yet powerful method is due to Pesarin [1990].

7.3 The Generalized Quadratic Form

7.3.1 Mantel's U

Mantel's $U = \Sigma\Sigma a_{ij} b_{ij}$ is perhaps the most widely used of all multivariate statistics because of its broad range of applications. By appropriately restricting the values of a_{ij} and b_{ij}, the definition of Mantel's U can be seen to include several of the standard measures of correlation including those usually attributed to Pearson, Pitman, Kendall, and Spearman [Hubert 1985]. In Mantel's original formulation [Mantel 1967], a_{ij} is a measure of the difference in time between items i and j, while b_{ij} is a measure of the spatial distance. As an example, suppose t_i is the day on which the ith individual in a study came down with cholera and (x_i, y_i) are the coordinates of her position in space (for example, five blocks west and two blocks north of city center). For all i, j set $a_{ij} = 1/(t_i - t_j)$ and $b_{ij} = 1/\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$.

A large value for U would support the view that cholera spreads by contagion from one household to the next. How large is large? As always, we compare the value of U

for the original data with the values obtained when we fix the i's but permute the j's as in $\pi[U] = \Sigma\Sigma a_{ij} b_{i\pi[j]}$.

7.3.2 Example in Epidemiology

An ongoing fear of many parents is that something in their environment—asbestos or radon in the walls of their house, or toxic chemicals in their air and groundwater—will affect their offspring. Table 7.1 is extracted from data collected by Siemiatycki and McDonald [1972] on congenital neural-tube defects. Eyeballing the gradient along the diagonal of this table one might infer that births of ancephalic infants occur in clusters. We could test this hypothesis statistically using the methods of Chapter 6 for ordered categories, but a better approach since the exact time and location of each event is known is to use Mantel's U. The question arises as to which measures of distance and time we should employ. Mantel [1967] reports striking differences between one analysis of epidemiologic data in which the coefficients are proportional to the differences in position and a second approach (which he recommends) to the same data in which the coefficients are proportional to the reciprocals of these differences.

Table 7.1. Incidents of Pairs of Ancephalic Infants

	Months Apart		
km apart	< 1	$1 << 2$	$2 << 4$
< 1	39	101	235
< 5	53	156	364
< 25	211	652	1516

Using Mantel's approach, a pair of infants born 5 kilometers and 3 months apart contribute $\frac{1}{3} * \frac{1}{5} = \frac{1}{15}$ to the statistic. Summing up the contributions from all pairs, then repeating the summing process for a series of random rearrangements, Siemiatycki and McDonald conclude the clustering of ancephalic infants is not statistically significant. You can confirm their result by entering the data into StatXact, using TableData, Settings to specify the row weights (2, 0.4, 0.08) and the column weights (2, 0.67, 0.33), then selecting Statistics, Doubly Ordered $R \times C$ Table, Linear-by-linear, and Exact from the menus to obtain a p-value of 0.18.

7.3.3 Further Generalization

Mantel's U is quite general in its application. The coefficients need not correspond to space and time. In a completely disparate application in sociology [Hubert and Schultz 1976], observers studied k distinct variables in each of a large number of subjects. Their object was to test a specific sociological model for the relationships among the variables. The $\{a_{ij}\}$ in Mantel's U were elements of the $k \times k$ sample-correlation matrix while the $\{b_{ij}\}$ were elements of an idealized or theoretical correlation matrix derived from the model. A large value of U supported the model; a small value would have ruled against it.

7.3.4 The MRPP Statistic

The MRPP or multiresponse permutation procedure [Mielke 1979] has been applied to applications as diverse as the weather and the spatial distribution of archaeological artifacts. The MRPP uses the permutation distribution of between-object distances to determine whether a classification structure has a nonrandom distribution in space or time. As we show in the next section, the MRPP method is flexible enough that it is readily adapted to obtain a permutation version of Hotelling's T^2 comparison of multivariate medians.

An example of the application of the MRPP arises in the assignment of antiquities (artifacts) to specific classes based on their spatial locations in an archaeological dig. Presumably, the kitchen tools of primitive man—woks and Cuisinarts—should be found together, just as a future archaeologist can expect to find TV, VCR, and stereo side by side in a neolithic living room.

Following Berry et al. [1980, 1983], suppose we have a collection of N artifacts within a site, and that the artifacts can be divided into G distinct groups (dishes, electronic devices, knick-knacks, and not-yet-classified), with n_g artifacts tentatively assigned to the gth class. Let D_g denote the average distance between artifacts within the gth class. The test statistic is the weighted within-class average of these distances, $D = \Sigma_g n_g D_g / N$. The permutation distribution associated with D is taken over all allocations of the N artifacts to the G classes with n_g assigned to each class.

7.3.5 An Example: Blue Grouse Migration Data

The distance and elevation change (in meters) for male and female blue grouse (*Dendragapus obscurus*) migrating from where they were marked on their breeding range to their winter range are given in Table 7.2 taken from Cade and Hoffman [1993]. Generally the males seem to migrate farther and higher than the females.

Table 7.2. Blue Grouse Migration Data

DIST	ELEV	SEX
800	0	F
6400	503	M
7800	488	M
10500	457	M
10600	610	M
11800	183	M
12500	549	M
14100	549	M
17400	671	M
29400	427	M
100	0	F
100	76	F
400	198	F
700	91	F

1300	213	F
5200	320	F
4400	-61	F
6700	305	F
8800	518	F
15100	360	F
28000	760	F

Blossom

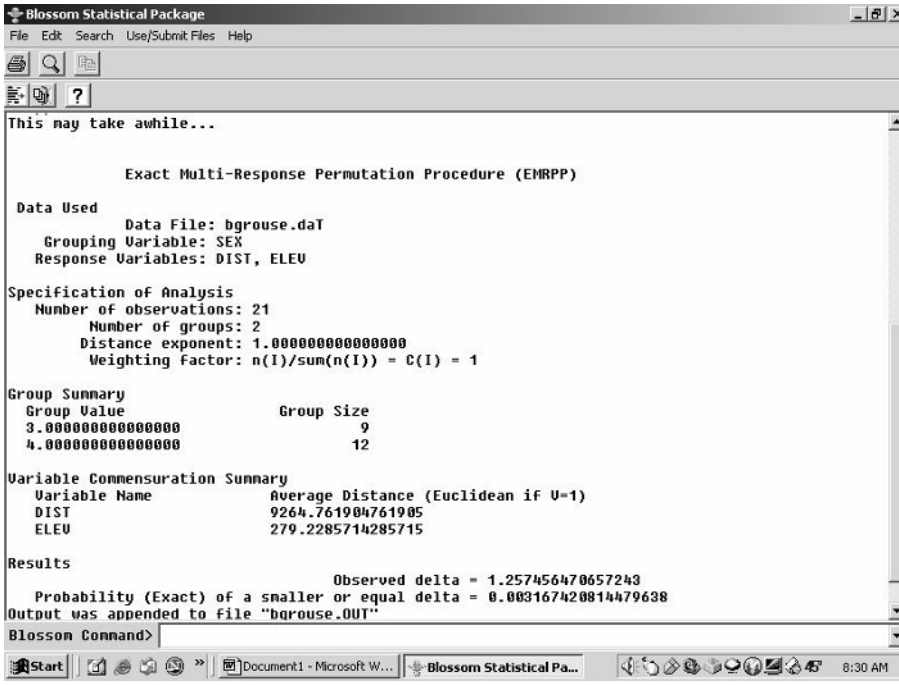
To test gender differences in both distance and elevation using the parametric version of Hotelling's T^2 yields a p-value of 0.033. To obtain a p-value based on the permutation distribution of Hotelling's T^2, you would issue the following commands while in Blossom:

```
USE BGROUSE.DAT
>MRPP DIST ELEV * SEX/HOT V = 2 C = 2 EXACT
```

This latter command bases its calculations on the square of the Euclidean distances. Mielke and Berry [1999] showed that one can obtain a more powerful test by using the absolute value of the Euclidean distance via the Blossom command:

```
>MRPP DIST ELEV * SEX/EXACT
```

yielding the following output:



7.4 Multiple Hypotheses

One of the difficulties with clinical trials and other large-scale studies is that frequently so many variables are under investigation that one or more of them is practically guaranteed to be significant by chance alone. If we perform 20 tests at the 5% or 1/20 level, we expect at least one significant result on the average. If the variables are related (and in most large-scale medical and sociological studies the variables have complex interdependencies), the number of falsely significant results could be many times greater.

A resampling procedure outlined by Troendle [1995] allows us to work around the dependencies. Suppose we have measured k variables on each subject, and are now confronted with k test statistics. To make these statistics comparable, we need to standardize them and render them dimensionless, dividing each by its respective L_1 norm or by its standard error. For example, if one variable, measured in centimeters, takes values like 144, 150, 156 and the other, measured in meters, takes values like 1.44, 1.50, 1.56, we might divide each of the first set of observations by 4, and each of the second set by 0.04.

Next, we order the standardized statistics by magnitude, that is, from smallest to largest. We also reorder and renumber the corresponding hypotheses. The probability that at least one of these statistics will be significant by chance alone at the 5% level is $1 - (1 - 0.05)^k$. But once we have rejected one hypothesis (assuming it was false), there will only be $k - 1$ true hypotheses to guard against rejecting.

1. Focusing initially on the largest of the k test statistics, repeatedly resample the data, (with or without replacement), to determine the p-value.
2. If this p-value is less than the pre-determined significance level, then accept this hypothesis as well as all the remaining hypotheses.
3. Otherwise, reject the corresponding hypothesis, remove it from further consideration, and repeat steps 1, 2, and 3.

7.4.1 Testing for Trend

Suppose you conduct a small study to test the effect of a drug on 15 subjects. The subjects receive 0 mg, 1 mg, and 2 mg of the drug, and the presence or absence of ten different side effects is noted for each subject.

```
data Drug;
   input Dose$ SideEff1-SideEff10;
   datalines;
  0MG  0   0   1   0   0   1   0   0   0   0
  0MG  0   0   0   0   0   0   0   0   0   1
  0MG  0   0   0   0   0   0   0   0   1   0
  0MG  0   0   0   0   0   0   0   0   0   0
  0MG  0   1   0   0   0   0   0   0   0   0
  1MG  1   0   0   1   0   1   0   0   1   0
  1MG  0   0   0   1   1   0   0   1   0   1
  1MG  0   1   0   0   0   0   1   0   0   0
  1MG  0   0   1   0   0   0   0   0   0   1
```

```
1MG  1   0   1   0   0   0   0   1   0   0
2MG  0   1   1   1   0   1   1   1   0   1
2MG  1   1   1   1   1   1   0   1   1   0
2MG  1   0   0   1   0   1   1   0   1   0
2MG  0   1   1   1   1   0   1   1   1   1
2MG  1   0   1   0   1   1   1   0   0   1
;
```

The data is taken from the SAS manual at http://support.sas.com/onlinedoc/913/docMainpage.jsp. Contrary to the statements made there, we shall suppose that the idea of testing for trend was conceived *prior* to our examining the data. Otherwise, our stated significance level is a gross underestimate. For while SAS PROC MULTTEST corrects for our making tests on the 10 side effects simultaneously, it does not correct for our having focused after-the-fact on the most prominent of the many possible contrasts among the means of the three dose groups.

```
proc multtest perm nsample=2000 seed=41287 notables
     pvals;
   class Dose;
   test ca(SideEff1-SideEff10/perm=2500);
   contrast 'Trend' 0 1 2;
 run;

The SAS System  12:29 Thursday, September 30, 2004 1

                The Multtest Procedure

                  Model Information

Test for discrete variables          Cochran-Armitage
Exact permutation distribution used  Everywhere
Tails for discrete tests             Two-tailed
Strata weights                       None
P-value adjustment                   Permutation
Number of resamples                  2000
Seed                                 41287


                Contrast Coefficients

                                Dose

  Contrast            0MG           1MG           2MG

  Trend                 0             1             2
```

p-Values

Variable	Contrast	Raw	Permutation
SideEff1	Trend	0.1006	0.5205
SideEff2	Trend	0.3337	0.9430
SideEff3	Trend	0.1228	0.6665
SideEff4	Trend	0.0240	0.1555
SideEff5	Trend	0.0806	0.2420
SideEff6	Trend	0.1139	0.6325
SideEff7	Trend	0.0173	0.0995
SideEff8	Trend	0.1006	0.5205
SideEff9	Trend	0.3337	0.9430
SideEff10	Trend	0.3536	0.9735

While the original p-values suggested that at least two of the side effects had a significant dose response at the 5% significance level, correcting for the multiple tests by permutation means reveals no significant differences at that level.

7.5 Summary

In this chapter, you learned the essentials of multivariate analysis for two-sample comparisons and applied them to repeated measures on the same subject. You learned how to detect clustering in time and space and to validate clustering models. You used the generalized quadratic form in its several guises including Mantel's U and Mielke's multiresponse permutation procedure (MRPP) to work through applications in archaeology, epidemiology, and ornithology. And you learned how to combine the results of multiple, simultaneous analyses.

7.6 To Learn More

Blair et al. [1994], Mielke and Berry [1999], and van-Putten [1987] review alternatives to Hotelling's T^2; Boyett and Shuster [1977] consider its medical applications. Extensions to other experimental designs are studied by Pesarin [1997; 2001] and Barton and David [1961].

Hayasaka and Nichols [2004] use Pesarin's combining function to analyze brain image data.

Mantel's U has been rediscovered frequently, often without proper attribution (see Whaley, 1983). Empirical power comparisons between MRPP rank tests and with other rank tests are made by Tracy and Tajuddin [1986] and Tracy and Khan [1990].

The generalized quadratic form has seen widespread application in anthropology [Williams-Blangero 1989], archaeology [Klauber 1971, 1975], ecology [Bryant 1977; Douglas and Endler 1982; Highton 1977; Levin 1977; Mueller and Altenberg 1985; Royaltey, Astrachen, and Sokal 1975; Ryman et al. 1980; Syrjala 1996], earth science [Mielke 1991], education [Schultz and Hubert 1976], epidemiology [Alderson

and Nayak 1971; Fraumeni and Li 1969; Glass and Mantel 1969; Glass et al. 1971; Klauber and Mustacchi 1970; Kryscio et al. 1973; Mantel and Bailar 1970; Merrington and Spicer 1969; Siemiatycki and McDonald 1972; Smith and Pike 1976], forestry [Cade 1997], geography [Cliff and Ord 1971, 1981; Hubert, Golledge, and Costanzo 1982; Hubert et al. 1984], management science [Graves and Whinston 1970], meteorology [Wong, Chidambaram, and Mielke 1983], ornithology [Cade and Hoffman 1993], paleontology [Marcus 1969], psychology [Hubert and Schultz 1976], sociology [Hubert and Baker 1977, 1978], and systematics [Dietz 1983; Gabriel and Sokal 1969; Selander and Kaufman 1975; Sokal 1979]. Siemiatycki [1978] considered various refinements.

Blair, Troendle, and Beck [1996], Troendle [1995], and Westfall and Young [1993] expand on the use of permutation methods to analyze multiple hypotheses; also, see the earlier work of Shuster and Boyett [1979], Ingenbleek [1981], and Petrondas and Gabriel [1983]. Simultaneous comparisons in contingency tables are studied by Passing [1984]. For additional insight into the analysis of repeated measures see Zerbe and Murphy [1986].

7.7 Exercises

1. You are studying a new tranquilizer you hope will minimize the effects of stress. The peak effects of stress manifest themselves between 5 and 10 minutes after the stressful incident, depending on the individual. To be on the safe side, you've made observations at both the 5- and 10-minute marks.

Subject	Prestress	5-minute	10-minute	Treatment
A	9.3	11.7	10.5	Brand A
B	8.4	10.0	10.5	Brand A
C	7.8	10.4	9.0	Brand A
D	7.5	9.2	9.0	New drug
E	8.9	9.5	10.2	New drug
F	8.3	9.5	9.5	New drug

How would you correct for the prestress readings? Is this a univariate or a multivariate problem? List possible univariate and multivariate test statistics. Perform the permutation tests and compare the results.

2. Show that Pitman's correlation is a special case of Mantel's U.

3. Sixteen seedlings were planted in two trays, eight per tray. The first tray contained a new fertilizer but was otherwise the same as the second tray. The heights of the surviving plants were compared after two weeks. Use Pesarin's test to analyze the results.

Tray 1:	5", 4", 7", 7.5", 6", 8"
Tray 2:	5", 4.5", 3", 6"

4. You wish to test whether a new fuel additive improves gas mileage and ride quality in both stop-and-go and highway situations. Taking 12 vehicles, you run them first on a highway-style track and record the gas mileage and driver's comments. You then repeat on a stop-and-go track. You empty the fuel tanks and refill, this time including the additive, and again run the vehicles on the two tracks.
The following data was supplied in part by the Stata Corporation. Use Hotelling's T^2 to test whether the additive affects gas mileage on the two tracks. Then use Pesarin's combination method to test whether the additive affects either gas mileage or ride quality.

Id	bmpg1	ampg1	rqi1	bmpg2	ampg2	rqi2
1	20	24	0	19	23.5	1
2	23	25	0	22	24.5	1
3	21	21	1	20	20.5	0
4	25	22	0	24	20.5	−1
5	18	23	1	17	22.5	1
6	17	18	−1	16	16.5	−1
7	18	17	0	17	16.5	0
8	24	28	1	23	27.5	0
9	20	24	0	19	23.5	1
10	24	27	0	22	25.5	0
11	23	21	0	22	20.5	0
12	19	23	1	18	22.5	1

bmpg1 track 1 before additive
ampg1 track 1 after additive
rqi1 ride quality improvement track 1
bmpg2 track 2 before additive
ampg2 track 2 after additive
rqi2 ride quality improvement track 2

8

Model Building

In this chapter, we will be concerned with using the values of one variable to predict the value of another. We will show the resampling methods may be used to estimate the values of model parameters, to test their significance, to estimate prediction errors, and to aid in validating the resulting model.

8.1 Picturing Relationships

Picturing a relationship in physics is easy. A funeral procession travels along the freeway at a steady 55 miles an hour, so that when we plot its progress on a graph of distance traveled versus time, the points all fall along a straight line as in Figure 8.1.

A graph of my own progress when I commute to work looks a lot more like Figure 8.2; this is because I occasionally go a little heavy on the gas pedal, while at other times traffic slows me down. Put a highway patrol car in the lane beside me and a graph of my progress would look much like that of the funeral procession. The underlying pattern is the same in each case—the distance traversed increases with time. But in real life, random fluctuations in traffic result in accelerations and decelerations in the curve.

The situation is similar but much more complicated when we look at human growth as in Figure 8.3: A rapid rate of growth for the first year, slow steady progress for the next 10 to 14 years and then, almost overnight it seems, we're a different size. Inflection points in the growth curve are different for different individuals. My ex-wife was $5'6''$ tall at the end of grade six, and $5'7''$ tall when she married me. I was $4'10''$ at the end of grade six, $5'$ at the beginning of grade ten, and $5'10''$ when she married me.[1] The pattern was the same for both of us, but the timing of our growth spurts differed.

These differences among individuals are why we need a way to characterize both average behavior and variation. A single number just won't do.

Before I launch an investigation, I start thinking about how and in what form I will report my results. I try to imagine the relationships I will be depicting. Suppose, for example, I were planning to investigate the relationship between education and income.

[1] Alas, I'm $5'9''$ today. Tant pis.

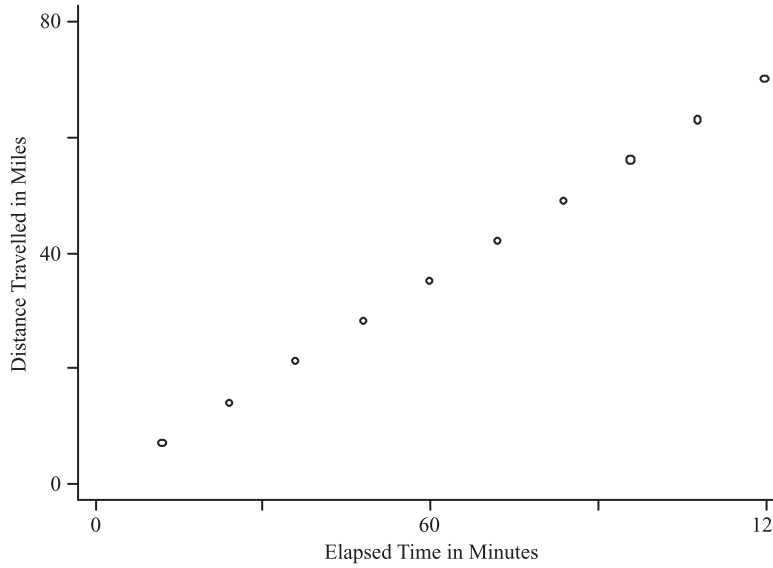


Fig. 8.1. Graphing the progress of a funeral procession.

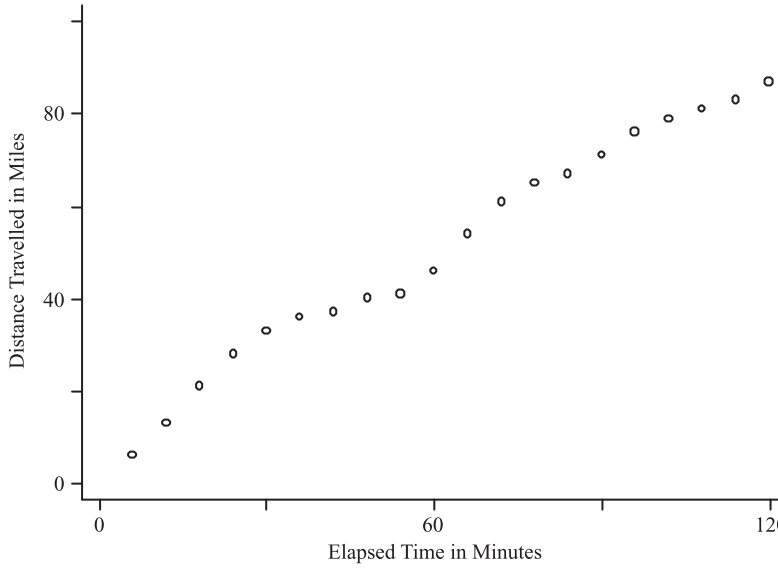


Fig. 8.2. Graphing my progress through freeway traffic.

What variables should I use to quantify this relationship? Years of education? Annual income at age 30? Total lifetime income?

To gain further insight, I'll pretend I've done a preliminary survey of several hundred individuals. I write down guesstimates of average responses and transfer these guesstimates to a two-way plot as in Figure 8.4a.

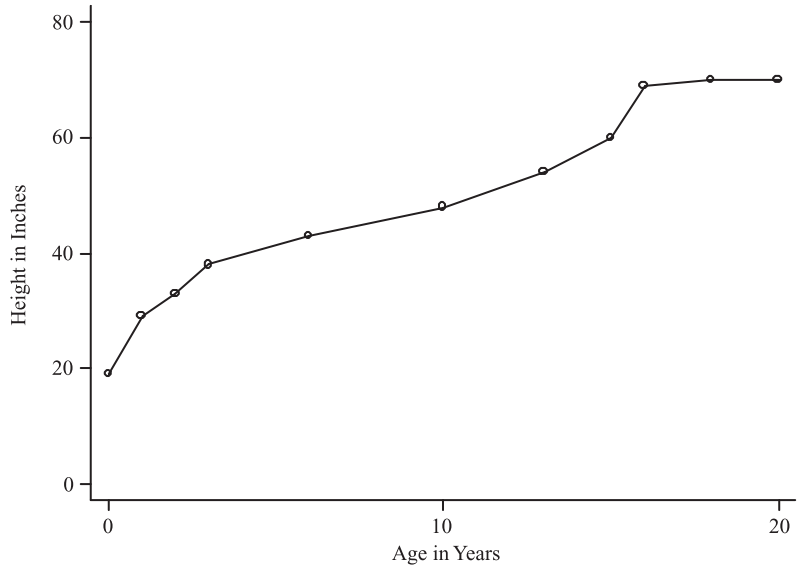


Fig. 8.3. Growth curve of Donny Travaglia.

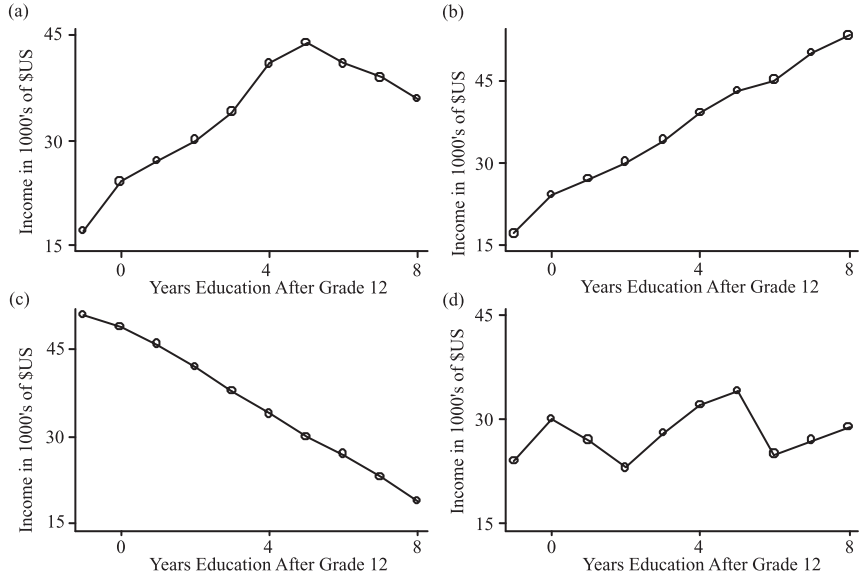


Fig. 8.4. Various trend lines for income as a function of years of education: a) nonlinear, b) rising linear, c) falling linear, d) no association.

A positive but nonlinear relationship is depicted in Figure 8.4a with income rising as one completes high school and college and falling off again with each year beyond the first in graduate school. Other possibilities, such as those depicted in Figure 8.4b,

c, and d, include a positive linear relationship, a negative linear relationship, and no relationship between income and education.

My next step is to begin to quantify this relationship in the form of an equation. Figures 8.1, 8.2, 8.4b, c, and the initial rising portion of Figure 8.4a all have the same underlying linear form: $Y = a + bX$, where Y is the dependent variable—income or distance traveled—X is the so-called "independent" variable—years of education or time— and a and b are the to-be-estimated intercept and slope of the line, respectively.

Suppose for example, we were to write $I = \$20{,}000 + \$5{,}000E$ where I stands for income and E for years of education after high school. This relationship is depicted in Figure 8.4b. Among its implications is that without college ($E = 0$) average annual income is \$20,000, while with college completed ($E = 4$) average income is doubled or \$40,000.

8.2 Unpredictable Variation

If our model were perfect, then all the points and all future observations would lie exactly on a straight line. We might be able to improve the fit to the existing values by writing the dependent variable, income in the previous example, as a function of several different variables or *predictors* $X_1, X_2, \ldots, X_n$ each representing some characteristic which might influence future income, but it's unlikely, again, that the fit would be perfect. Even with time-tested models, of the kind studied in freshman science laboratories, the observations just won't cooperate. Always there seems to be a portion we can't explain, the result of observer error, or DNA contamination, or a hundred other factors we did not think of measuring or were unable to measure.

For simplicity, let's represent all the different explanatory variables by the single letter X, and again suppose that even with all these additional variables included in our model, we still aren't able to predict Y exactly. A small fraction ε of each observation continues to defy explanation. We can write a functional of Y (Y's mean or median or 25th percentile) as a mixture of deterministic and stochastic (random) components,

$$F(Y) = bX + \varepsilon,$$

where X represents the variables we know about, b is a vector of constants used to apportion the effects of the individual variables that make up X, and ε denotes a random fluctuation, the part of the relationship we can't quite pin down or attribute to any specific cause.

8.2.1 Building a Model

Imagine you are the proud owner of Fawlty Towers and have just succeeded in booking the International Order of Arcadians and Porcupine Fanciers for a weekend conference. In theory you ought to prepare to serve as many meals as the number of registrants, but, checking with your fellow hotel owners, you soon discover that with no-shows and non-diners you can get by with a great many less. Searching through the records of the former owners of your hotel, you come up with the data in Table 8.1. You convert these numbers to a graph, Figure 8.5, and discover what looks almost like a linear relationship between the meals you need to prepare and the number of guests in your hotel!

Developing and Testing a Model

1. Begin with the Reports.
   Picture Relationships.
2. Convert Your Pictures to Formal Models.
   with both Deterministic and Stochastic Elements.
3. Plan Your Sampling Method.

Define the Population of Interest.
Ensure
a) Sample is representative.
b) Observations are independent.

Table 8.1. Registrants and Servings at Fawlty Towers

Registrants	Maximum servings	Registrants	Maximum servings
289	235	339	315
391	355	479	399
482	475	500	441
358	275	160	158
365	345	319	305
561	522	331	225

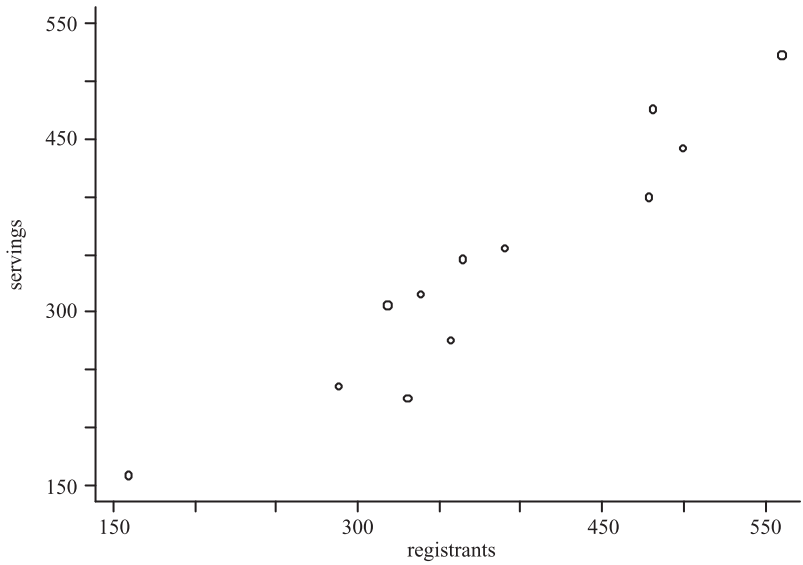


Fig. 8.5. Servings versus conference registrants at Fawlty Towers.

8.2.2 Bivariate Dependence

Before we begin to construct a model, perhaps we ought to ask whether the apparent association we observe in Table 8.1 between registrants and servings is statistically significant. We can use the Pitman correlation method introduced in Section 5.2.3 to verify our conjecture. If $\{(r_1, m_1), (r_2, m_2), \ldots, (r_n, m_n)\}$ is a set of n pairs of observations on registrants and servings, then to test for association between them we need to look at the permutation distribution of the Pitman correlation $S = \sum_{i=1}^{n} r_i m_i$.

R

```
#Calculating a p-value via a Monte Carlo
armspan <- c(139, 140, 141, 142.5, 143.5)
height <- c (137, 138.5, 140, 141, 142)
rho0 <- cor(armspan, height)
cnt<- 0
for (i in 1:400){
    D <- sample (armspan)
    rho <- cor(D, height)
    # counting correlation larger than original
         by chance
    if (rho0 <= rho ) cnt<-cnt+1
    }
cnt/400                    #pvalue
```

Resampling Stats

```
'Calculating a p-value via a Monte Carlo
DATA (139 140 141 142.5 143.5) armspan
DATA (137 138.5 140 141 142) height
CORR armspan height rho0
REPEAT N
  SHUFFLE height H
  CORR armspan H rho
  SCORE rho scrboard
END
COUNT scrboard>=rho0 extremes
LET pvalue =extemes/N
PRINT pvalue
```

Stata

```
permute armspan "corr armspan height" teststat=r(rho)
```

Test for Bivariate Dependence

H: The observations within each pair are independent of one another.
K: The observations within each pair are all positively correlated or are all negatively correlated.

Assumptions:

1. (x_i, y_i) denotes the members of the ith pair of observations.
2. The labels of the $\{x_i\}$ are exchangeable as are those of the $\{y_i\}$.

Test statistic:
$S = \Sigma y_i x_i$. Reject the hypothesis of independence if the value of S for the observations before they are rearranged is an extreme value in the permutation distribution.

8.2.3 Confidence Interval for the Correlation Coefficient

We will get exactly the same result in our tests if we replace the Pitman correlation statistic by the Pearson correlation statistic:

$$\rho = \frac{Cov(XY)}{\sigma_X \sigma_Y}.$$

The covariance when two variables X and Y are observed simultaneously is

$$Cov(XY) = \Sigma_{i=1}^{n}(x - \bar{x})(y - \bar{y})/(n - 1).$$

If X and Y are independent and uncorrelated then $\rho = 0$. If X and Y are totally and positively dependent, for example, if $X = 2Y$, then $\rho = 1$. In most cases of dependence $0 < |\rho| < 1$.

If we knew the standard deviations of X and Y, we could use them to help obtain a confidence interval for ρ by the method of permutations. But we only have estimates of the standard deviations. The bootstrap provides an obvious solution providing we remember our sample consists of paired observations.

R

```
armspan <- c(139, 140, 141, 142.5, 143.5)
height <- c (137, 138.5, 140, 141, 142)
n = length(armspan)
#collect all variables in a single frame so as to
     sample as a unit
data=cbind(armspan,height)
#set number of bootstrap samples
N =400
stat = numeric(N)  #create a vector in which to store
     the results
```

```
for(i in 1:N){
   ind=sample(n,n, replace=T)
   boot= data[ind,]
   stat[i]=cor(boot[,1],boot[,2])
   }
quantile(stat,prob=c(0.05,0.95))
```

Stata

```
bootstrap "corr armspan height" r(rho), reps(1000)

Monte Carlo permutation statistics  Number of obs = 5
                                    Replications =1000

T         | T(obs)   c    n  p=c/n SE(p)   [95% Conf.
                                            Interval]

teststat | .986465 13 1000  0.0130 0.0036 .0069396
                                             .0221278

Note: confidence interval is with respect to p=c/n
Note: c = #{|T| >= |T(obs)|}
```

8.2.4 But Does the Model Make Sense?

Does your model have a cause and effect basis? Far too often, we let our computer and our software do the thinking for us. Just because your software says that a statistically significant relationship exists does not mean that the relationship is genuine. Perhaps, there is a third or fourth variable that is directly responsible for the observed changes. The height and weight of a child normally increase as the child grows older, but would you use a child's height to predict its weight? In the case of Fawlty Towers, it is easy to see there is a direct relationship between the number of guests and the number of meals consumed, so we can feel free to move forward.

Or can we? We may say the relationship between the number of meals consumed and the number of guests is linear, but is it? Before you continue with your reading, please do the first exercise at the end of this chapter. A wide variety of different models may provide a fit to the data in hand, but none may be truly correct.

One final warning: The registrants for whom we have data in Table 8.1 vary in number from 160 to 561 per day. It would be foolish to assume that any model we develop would be applicable outside this range. It may be that Fawlty's met the needs of additional diners by adding tables to an already overcrowded dining room or by extending the serving hours from four in the afternoon to well past ten. Fewer registrants might mean that everyone dines in, while more might mean that guests don't even bother to see if there is a table before going elsewhere in search of food.

8.2.5 Estimating the Parameters

Many methods exist for estimating the parameters of a relationship. As discussed in Section 4.2, your choice of method will depend upon the purpose of your estimation procedure and the losses you may be subjected to should mistakes be made.

In most instances, your purpose will be to use historical data to predict the future. This is just what the owner of Fawlty Towers is trying to do. Seeing that a relationship of the form Meals $= a + b^*$ Registrants appears to exist, he wants to determine values of a and b so that when he attempts to predict the number of meals required in the future, he will minimize the losses associated with producing too few or too many meals.

The first question is whether he wants to predict the expected number of meals required or the upper 90th or 95th percentile of that number. The latter might be the case if, unlike John Fawlty, the owner of the inn wants to keep to a minimum the number of guests he turns away hungry. Predicting the expected number of meals is generally done by a method called linear regression. Predicting quantiles is done by a method called quantile regression. In the following sections, we consider each approach in turn.

8.3 Linear Regression

The two most popular linear regression methods for estimating model coefficients are referred to as ordinary-least-squares (OLS) and least-absolute-deviation (LAD) goodness of fit, respectively. Because they are popular, a wide selection of computer software is available to help us do the calculations.

With *least-squares* goodness of fit, we seek to minimize the sum

$$\sum_i (M_\iota - a - bR_i)^2,$$

where M_i denotes the number of meals served and R_i the number of registrants on the ith occasion. With the LAD method, we seek to minimize

$$\sum_i |M_\iota - a - bR_i|.$$

Those who've taken calculus, know the OLS minimum is obtained when

$$\sum_i (M_\iota - a - bR_i)b = 0 \text{ and } \sum_i (M_\iota - a - bR_i) = 0,$$

that is, when

$$b = \frac{\text{Covariance}(RM)}{\text{Variance}(R)} = \frac{\Sigma(R_i - \bar{R})(M_i - \bar{M})}{\Sigma(R_i - \bar{R})^2}$$

and

$$a = \bar{M} - b\bar{R}.$$

Finding the LAD minimum is more complicated and requires linear programming.

For Fawlty Towers, the OLS method yields the following equation:

$$\text{Meal Servings} = 0.94^* \text{ Registrants} - 20.6.$$

To test whether the coefficients are significantly different from zero, we use the more accurate permutation approach; for confidence intervals, we use the bootstrap.

R

```
#obtain LAD regression coefficients and test slope to
     see if greater than zero
library("quantreg")
Guests =c (289,391,482,358,365,561,339,479,500,160,
     319,331)
Meals = c(235,355,475,275,345,522,315,399,441,158,
     305,225)
N=400
f = coef(rq(formula = Meals ~ Guests))
names(f)=NULL
stat0=f[2]
cnt=0
for(i in 1:N){
   guestP=sample(Guests)
   fp= coef(rq(formula = Meals ~ guestP))
   names(fp)=NULL
   if (fp[2] > = stat0)
   cnt=cnt+1
   }
f

cnt/N

#obtain bootstrap confidence intervals for LAD
     regression coefficients
library("quantreg")
Guests =c (289,391,482,358,365,561,339,479,500,160,
     319,331)
Meals = c(235,355,475,275,345,522,315,399,441,158,
     305,225)
n = length(Guests)
data=cbind(Guests,Meals)
#set number of bootstrap samples
N =400
stat = numeric(N)  #create a vector in which to
     store the results
for(i in 1:N){
   ind=sample(n,n, replace=T)

   guestP= data[ind,]
   fp= coef(rq(formula = Meals ~ guestP))
```

```
    stat[i]= fp[2]
    }
quantile(stat,prob=c(0.05,0.95))
```

SAS

```
// code uses the wrapper method of David L. Cassell

%macro rand_gen(
  indata=_last_,
  outdata=outrand,
  depvar=y,
  numreps=1000,
  seed=0);
  /* Get size of input dataset into macro variable
       &NUMRECS */
  proc sql noprint;
     select count(*) into :numrecs from &INDATA;
     quit;
  /* Prepare for sorting by generating random
       numbers */
   data __temp_1;
      retain seed &SEED ; drop seed;
    set &INDATA;
    do replicate = 1 to &NUMREPS;
     call ranuni(seed,rand_dep);
     output;
     end;
   run;
proc sort data=__temp_1;
by replicate rand_dep;
run;
data &OUTDATA ;
   array deplist{ &NUMRECS } _temporary_ ;
   set &INDATA(in=in_orig)
   __temp_1(drop=rand_dep);
   if in_orig then do;
     replicate=0;
     deplist{_n_} = &DEPVAR ;
     end;
  else &DEPVAR = deplist{ 1+ mod(_n_,&NUMRECS) };
  run;
%mend rand_gen;

%rand_gen(indata=nudata,outdata=outrand,
depvar=Meals,numreps=1600,seed=12345678)
```

```
data nudata;
input Guests Meals;
datalines;
289 235
391 355
482 475
358 275
365 345
561 522
339 315
479 399
500 441
160 158
319 305
331 225
;
proc glm data=outrand noprint outstat=outstat1;
by replicate;
model Meals = Guests;
run;

%rand_anl(randdata=outstat1,
where=_source_='Guests' and _Type_='SS3',
testprob=prob,testlabel=Model F test)

%macro rand_anl(
  randdata=outrand,
  where=,
  testprob=probf,
  testlabel=F test,);
  data _null_;
  retain pvalue numsig numtot 0;
  set &RANDDATA end=endofile;
  %if "&WHERE" ne ""
 %then where &WHERE %str(;) ;
if Replicate=0 then pvalue = &TESTPROB ;
else do;
   numtot+1;
   numsig + ( &TESTPROB < pvalue );
end;
if endofile then do;
    ratio = numsig/numtot;
    put "Randomization test for &TESTLABEL "
    %if "&WHERE" ne "" %then "where &WHERE";
    " has significance level of "
    ratio 6.4 ;
```

```
 end;
 run;
%mend rand_anl;
```

Stata

permute Meals "regress Meals Guests" _b, reps(400) left

```
command:      regress Meals Guests
statistics: b_Guests = _b[Guests]
b_cons    = _b[_cons]
permute var:  Meals

Monte Carlo permutation statistics  Number of obs = 12
Replications = 400

T         |  T(obs)   c   n   p=c/n [95% Conf.Interval]
b_Guests  | .9393529  400 400 1.0000 .9908202 1
b_cons    | -20.55001  0  400 0.0000   0      .0091798

Note: confidence intervals are with respect to p=c/n
Note: c = #{T <= T(obs)}
```

. bootstrap "regress Meals Guests" _b, reps(400) nonormal nopercentile

```
command:      regress Meals Guests
statistics: b_Guests   = _b[Guests]
b_cons    = _b[_cons]

Bootstrap statistics         Number of obs = 12
Replications = 400

Variable | Reps  Observed   Bias     Std. Err.
                              [95% Conf. Interval]

b_Guests | 400    .9393529 .0072056  .0921974
                                .8209208  1.190321 (BC)
b_cons   | 400 -20.55001 -3.715155 38.81548
                              40.0799    24.55453 (BC)

Note: BC = bias-corrected
```

8.3.1 Other Regression Methods

As can be seen from the ease with which one can adapt these programs from LAD to OLS regression, we can employ resampling methods with logistic regression, quantile regression, and even nonlinear univariate regression. Indeed, as the resampling methods are independent of the functional relationship between the predictor(s) and the effect, they are applicable to equations of the general form $G[y, x] = 0$.

8.4 Improving the Model

Of course, there are discrepancies, also known as *residuals*, from any model, as can be seen from Figure 8.6 and as you note in Table 8.2. Is there an explanation for these discrepancies? 331 registrants, 290 meals predicted and only 225 meals served! Well, one thing you forgot to mention to the officers of the IOAPF before you signed them up is that for most of the year the weather out Fawlty Towers way is dreadful and it is unpleasant the rest of the time.[2] Perhaps, bad weather is the explanation. Digging deeper into the previous owners' records you come up with the expanded Table 8.3 in which you've assigned numerical values to the weather conditions ranging from 1-Sunny to 5-Hurricane.

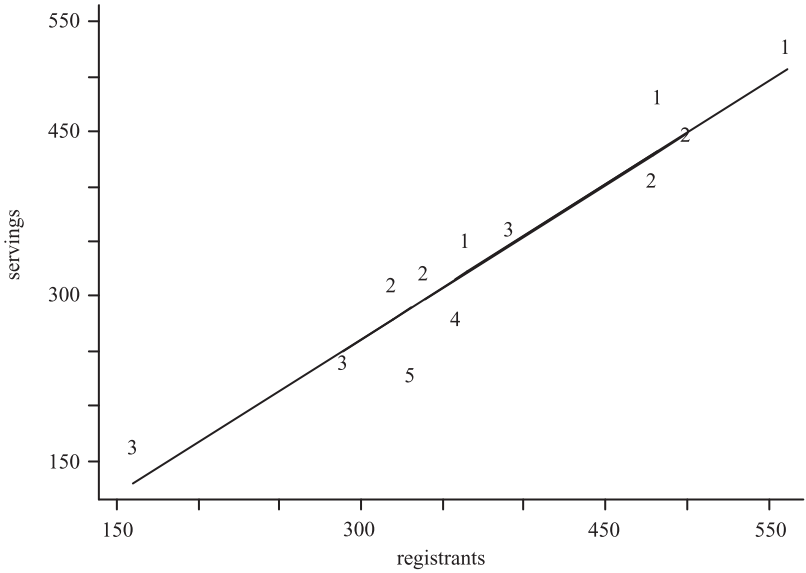


Fig. 8.6. Regression of servings on registrants at Fawlty Towers. (Numbers refer to severity of weather conditions.)

Adding a term to account for the weather to your model and applying multipredictor OLS techniques yields the following formula:

$$\text{Servings} = 0.78^{*}\text{Registrants} - 27^{*}\text{Weather} + 104.$$

The residuals as you can see from the revised Table 8.4 are much smaller, though, of course, the fit still is not perfect.

[2] The pictures you showed them of Fawlty Towers, taken on your own Hawaiian vacation, didn't help the situation. What you didn't realize and will be forced to deal with later when you come up 100 meals short is that the weather in Nova Scotia where most of those Arcadians live is awful for most of the year, also.

Table 8.2. Model I vs. Reality

	Servings		
Registrants	Actual	Model I	Residual
289	235	251	−15.9
391	355	347	8.3
482	475	432	42.8
358	275	316	−40.7
365	345	322	22.7
561	522	506	15.6
339	311	298	17.1
479	399	429	−30.4
500	441	449	−8.1
160	158	130	28.3
319	305	279	25.9
331	225	290	−65.4

Table 8.3. Registrants and Servings at Fawlty Towers

Registrants	Servings	Weather	Registrants	Servings	Weather
289	235	3	339	315	2
391	355	3	479	399	2
482	475	1	500	441	2
358	275	4	160	158	3
365	345	1	319	305	2
561	522	1	331	225	5

Table 8.4. Model II vs. Reality

	Servings		
Registrants	Actual	Model II	Residual
289	235	250	−14.5
391	355	329	25.6
482	475	455	20.2
358	275	277	−1.6
365	345	363	−18.1
561	522	517	5.3
339	311	316	−0.7
479	399	425	−26.4
500	441	442	−0.9
160	158	148	9.6
319	305	300	4.9
331	225	228	−3.4

8.4.1 Testing for Significance in Multipredictor Regression

Alas, when more than one variable is used as a predictor, permutation tests for the individual regression coefficients are not exact.[3] The bootstrap procedures described

[3] The reasons are similar to those discussed in Section 5.3.3.

in this chapter and in Chapter 2 cannot be relied on when the sample is too small. Increasing the number of bootstrap samples will not remedy the situation when there is inadequate information to begin with. SAS reports at
http://ftp.sas.com/techsup/download/stat/jackboot.html
that when attempts were made to find a 95% confidence interval for $R^{**}2$ in a linear regression with 20 observations and 10 predictors, the bootstrap distribution was not even close to the true sampling distribution. The bootstrap BCa interval was extremely short and did not contain the true value.

We may still use the bootstrap to obtain confidence intervals for the coefficients subject to the caveats raised in Section 8.2.4. (These same objections apply—though they are seldom discussed—to confidence intervals based on parametric assumptions.) In particular, if we need n^2 observations to ensure that the distribution of one predictor in the sample is sufficiently close to its distribution in the population, then if we have k different independent predictor variables, we will need $(n^2)^k$ multivariate observations in the original sample to ensure similar accuracy.

Fortunately, in most practical cases, our predictor variables are not independent but correlated and we can ensure a good fit between the sample distribution and the population distribution with a much smaller sample though still of order n^{2j} where $1 < j < k$.

A further problem arises when we must decide which *interaction* terms to include in the model. An example of an interaction would be a term of the form aX_1X_2. Interaction corrects for shifts in the relationships among the variables as the values of the variables change. An excellent example is the effect of age on the relationship between a person's height and weight. During the growth period between 1 and 14 years, height and weight are closely correlated. But over the age of 20, the only body dimension that increases with weight is width. A possible model relating height H to weight W would be $H = a_0 + a_1WA$ where A is a variable that takes various values depending on age.

If we have k predictors, we will have $k(k-1)$ interaction terms of the form X_iX_j, $k(k-1)(k-2)$ of the form $X_iX_jX_m$, and so forth. If we're not careful, we'll end up with more terms than we have observations with which to estimate them. Regression methods alone cannot make the determination.

The computer code of Section 8.3 can be used to test for significance by replacing the single predictor regression commands with multivariable ones. In Stata, for example, one merely needs to write

. bootstrap "regress Meals Guests Weather" _b, reps(400) nonormal noper-centile

8.4.2 Comparing Two Regression Lines

A question that often arises in practice is whether two regression lines based on two sets of independent observations have the same slope.[4] Suppose we can assume

$$y_{ij} = a_i + bx_{ij} + \varepsilon_{ij} \quad \text{for } i = 1, 2; \quad j = 1, \ldots, n_i,$$

[4] This is quite different from the case of repeated measures where a series of dependent observations are made on the same subjects (Good 2004, pp. 181–184) over a period of time.

where the errors $\{\varepsilon_{ij}\}$ are exchangeable; then

$$\bar{y}_{i.} = a_i + b\bar{x}_{i.} + \bar{\varepsilon}_{i.}.$$

Define $y' = \frac{1}{2}(\bar{y}_1 - \bar{y}_2)$; $x' = \frac{1}{2}(\bar{x}_1 - \bar{x}_2)$; $\varepsilon' = \frac{1}{2}(\bar{\varepsilon}_1 - \bar{\varepsilon}_2)$; $a' = \frac{1}{2}(a_1 + a_2)$. Define $y'_{1i} = y_{1i} - y'$ for $i = 1$ to n_1 and $y'_{2i} = y_{2i} + y'$ for $i = 1$ to n_2.

Define $x'_{1i} = x_{1i} - x'$ for $i = 1$ to n_1 and $x'_{2i} = x_{2i} + x'$ for $i = 1$ to n_2.

Then $y'_{ij} = a' + bx'_{ij} + \varepsilon'_{ij}$ for $i = 1, 2$; $j = 1, \ldots, n_i$.

Two cases arise. If the original values of the predictors were the same for both sets of observations, that is, if $x_{1j} = x_{2j}$ for all j, then the errors $\{\varepsilon'_{ij}\}$ are exchangeable and we can apply the method of matched pairs as in Section 3.5. Otherwise, we need to proceed as follows: First, estimate the two parameters a' and b by least-squares means. Use them to derive the transformed observations $\{y'_{ij}\}$. Then test the hypothesis that $b_1 = b_2$ using a two-sample comparison as in Section 3.4.2. If the original errors were from a symmetric distribution and were exchangeable, then the errors $\{\varepsilon'_{ij}\}$ though not independent are exchangeable and this test is exact.

Alternately, we may know two curves are parallel, but suspect they are not coincident. This problem is similar to that of the two-sample comparison, the difference being that we wish to increase the power of the test by correcting for the effects of various covariates. This problem also serves to illustrate some of the major differences between the permutation and the bootstrap approach.

Our solution by resampling methods requires us to take observations from the two populations for the same set of values of the covariates $\{X_i; i = 1, \ldots, n\}$. Given that,

$$Y_{1i} = a_1 + f[X_i] + \eta_{1i} \text{ for } I = 1, \ldots, n$$
$$Y_{2i} = a_2 + f[X_i] + \eta_{2i} \text{ for } I = 1, \ldots, n,$$

where the $\{\eta_{ji}\}$ are independent random values each of whose expected value is zero.

To test the null hypothesis that $a_1 = a_2$, our statistic for the permutation test is that for matched pairs, $S = \Sigma(Y_{1i} - Y_{2i})$, where we perform n independent permutations, one for each pair. The permutation test is exact if we can assume the $\{\eta_{ji}\}$ are identically distributed.

The statistic for the bootstrap derived by Hall and Hart [1990] is

$$S = \left[\sum_{j=0}^{n-1}\left(\sum_{i=j+1}^{j+m} D_i\right)^2\right]\left[n\sum_{i=1}^{n-1}(D_{i+1} - D_i)^2\right]^{-1},$$

where

$$D_i = Y_{1i} - Y_{2i} - n^{-1}\sum_{i=1}^{n}(Y_{1i} - Y_{2i}) \text{ for } 1 \le i \le n;$$
$$D_i = D_{i-n} \text{ for } n + 1 \le i \le n + m,$$

m equals the integer part of np, with $0 \le p \le 1$. The complexities of this statistic are occasioned by the need to Studentize to obtain asymptotically exact significance levels and the introduction of m necessary to the proof that the results are asymptotically exact. The bootstrap test requires no additional assumption beyond the original one of the independence of the errors $\{\eta_{ji}\}$.

8.4.3 Prediction Error

For most hotel owners, John Fawlty excepted, a single figure would not be enough; a prediction interval is needed. Averages are often exceeded and the sight of six or seven conference participants fighting over a single remaining pork cutlet can be a bit unsettling. In this section, we bootstrap not once, but twice to obtain the desired result.

Consider the more general regression problem where given a vector X of observations, we try to predict Y. Suppose we've already collected a set of independent identically distributed observations $w = \{w_i = (x_i, m_i), i = 1, \ldots, n\}$ taken from the multidimensional distribution function F. The data in Table 8.1, with X the number of registrants and M the number of meals, provides one example. Given a new value for the number of registrants x_{n+1} we use our regression equation based on w to derive an estimate $m[x_{n+1}, w]$ of the number of meals we can expect to serve. Any difference between the estimate and the actual number of meals served could result in a loss. This loss could be something as simple as the cost of preparing an unnecessary meal or it could be far more complicated including the cost of a replacement meal obtained from a nearby hotel, and the costs of lawsuits brought by thoroughly irritated porcupine fanciers.

The prediction error $L(m_{n+1}, m[x_{n+1}, w])$ should be averaged over all possible outcomes (x_{n+1}, m_{n+1}) we might draw from F in the future. Call this error $E[w, F]$. We don't know what F is, but we do have an estimate, F' based on the data w we've collected so far. The *apparent error rate* is the average of the losses over all the pairs of values (x_i, m_i) we have observed:

$$E[w, F'] = \Sigma_i L(m_i, m[x_i, w])/n.$$

To estimate the true error, we treat the observations as if they were the population, and a series of bootstrap samples from the original sample as if they were the observations. Each bootstrap sample consists of pairs $\{w_i^*\}$ drawn without replacement from w. Our plug-in estimate of error based on a bootstrap sample is

$$E[w^*, F'] = \Sigma_i L(m_i, m[x_i, w^*])/n.$$

To obtain this estimate, we must compute the regression coefficients a second time using w^* in place of w in our calculations, although we continue to average the errors over the original observations. As results based on a single bootstrap sample can be quite misleading, we have to repeat the process of sampling, regression, and error determination several hundred times, then take the average of our estimates. I'm not complaining for, like you, I've got a computer to do the work. Our final estimate of prediction error is the average of the plug-in estimates over the B bootstrap samples,

$$E[w^*, F^b] = \Sigma^B \Sigma_i L(m_i, m[x_i, w^*])/nB.$$

We use the distribution of errors from these same calculations to obtain an interval estimate of the prediction error.

Estimating Prediction Error I

Determine the loss function L.
Repeat the following several hundred times:
Draw a bootstrap sample w^* from the pairs of observation $w = \{(x_i, m_i)\}$.
Compute the regression coefficients based on this sample w^*.
Using these coefficients, compute the losses $L(m_i, m[x_i, w^*])$ for each pair in the original sample.
Compute the mean of these losses.
Compute the mean of these means averaged over all the bootstrap samples.

8.4.4 Correcting for Bias

When we estimate a population mean using the sample mean, the result is unbiased, that is, the mean of the means of all possible samples taken from a population is the population mean. In the majority of cases, including the preceding example, estimates of error based on bootstrap samples are biased and tend to underestimate the actual error. If we can estimate this bias by bootstrapping a second time, we can improve on our original estimate.

This bias, which Efron and Tibshirani [1993] call the *optimism*, is the difference $E[w, F] - E[w, F^b]$ which we estimate as before by using the observations in place of the population and the bootstrap samples in place of the observations,

$$E[w^*, F^b] - E[w, F^{b*}] = \Sigma^B \Sigma_i \{L(m_i, m[x_i, w^*]) - L(m_i^*, m[x_i^*, w^*])\}/nB.$$

$L(m_i, m[x_i, w^*])$ uses the observations from the original sample and the coefficients derived from the bootstrap; $L(m_i^*, m[x_i^*, w^*])$ uses the bootstrap observations and the coefficients derived from the bootstrap. Our corrected estimate of error is the apparent error rate plus the optimism or

$$E[w, F^b] + E[w^*, F^b] - E[w, F^{b*}].$$

8.5 Validation

How can we be confident of our ability to predict the future when all we have to work with is the past? Before applying a model we need to *validate* it. Ideally, we would have some independent source of verification we could draw on. But in most instances, we will have no alternative but to make use of the same data to validate the model that we used to develop it. To do so, we need either develop a *metric* and determine its permutation distribution, *bootstrap*, or apply some other form of *cross-validation*. In what follows we consider each of these approaches in turn.

8.5.1 Metrics

With the availability of more and more powerful computers, scientists are addressing problems involving hundreds of variables such as weather prediction and economic

Estimating the Prediction Error II

Determine the loss function L.
Solve for the regression coefficients using the original observations.
Compute apparent error rate:

$$E[w, F'] = \Sigma_i L(m_i, m[x_i, w])/n.$$

Repeat the following 160 times:
Choose a bootstrap sample w^* with replacement.
Solve for regression coefficients using the bootstrap sample.
Use these coefficients to determine apparent error for original observations

$$E[w^*, F^{b*}]$$

and the optimism

$$E[w^*, F^b] - E[w, F^{b*}].$$

Compute the average optimism.
Compute the corrected error rate.

modeling. In many instances, the computer has taken the place of a laboratory. Where else could you emulate a thousand years of stellar evolution or study over and over the effects of turbulence on an aircraft landing?

Whether the data are real or simulated, we need to analyze our results. To distinguish between close and distant, between a good fitting model and a poor one, we first need a metric, and then a range of typical values for that metric. A metric m defined on two points x and y, has the following properties:

$$\begin{aligned} m(x, y) &\geq 0, \\ m(x, x) &= 0, \\ m(x, y) &\leq m(x, z) + m(z, y), \end{aligned}$$

which corresponds to most of the common ways in which we measure distance. The third property of a metric, for example, simply restates that the shortest distance between two points is a straight line.

A good example is the standard Euclidian metric used to measure the distance between two points x and y whose coordinates in three dimensions are (x_1, x_2, x_3) and (y_1, y_2, y_3)

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}.$$

This metric can be applied even when x_i is not a coordinate in space but the value of some variable like blood pressure or laminar flow or return on equity. Hotelling's T^2, defined in Chapter 7, is a example of such a metric.

As in Chapter 7, the next step is to establish the range of possible values that such a metric might take by chance, using either the bootstrap or permutations of the combined set of theoretical and observed values. If the value of this metric for the original observations is an extreme one, we should reject our model and begin again.

Nearest Neighbors

Suppose we have a set of N observations and a second set of N simulated values. For each point in the set of theoretical values, Set 1, record the number of points among its three nearest neighbors (nearest in terms of the Euclidian metric) which are also in Set 1; sum the number of points so recorded. Call this sum S_0.

To determine whether S_0 is an extreme value, compute S for rearrangements of the combined group of theoretical and observed values. In all but the original and one other arrangement, Set 1 will consist of a mixture of observed and theoretical values. If these values are significantly different from one another, fewer nearest neighbors will be in the same set and S will be smaller than S_0.

Structured Exploratory Data Analysis

Karlin and Williams [1984] use permutation methods in a structured exploratory data analysis (SEDA) of familial traits. A SEDA has four principal steps:

1) The data are examined for heterogeneity, discreteness, outliers, and so forth, after which they may be adjusted for covariates and the appropriate transform applied.
2) A collection of summary SEDA statistics are formed from ratios of functionals as in the example that follows.
3) The SEDA statistics are computed for the original family trait values and for reconstructed family sets formed by permuting the trait values within or across families.
4) The values of the SEDA statistics for the original data are compared with the resulting permutation distributions.

As one example of a SEDA statistic, consider the OBP, the Offspring-Between-Parent SEDA statistic:

$$\frac{\sum_i^N \sum_j^{J_i} |O_{ij} - (M_i + F_i)/2|}{\sum_i^N |M_i - F_i|}.$$

In family $i = 1, \ldots, I$, F_i and M_i are trait values of the father and mother (cholesterol levels in their blood, for example), while O_{ij} is the corresponding trait value of the jth child of those same parents, $j = 1, \ldots, J_i$.

To evaluate the permutation distribution of the OBP, consider all permutations in which the children are kept together in their respective family units, while we either

a) randomly assign to them a father and (separately) a mother, or
b) randomly assign to them an existing pair of spouses.

The second of these methods preserves the spousal interaction. Which method a geneticist might choose will depend upon the alternative(s) of interest.

It would be difficult if not impossible to derive the distribution of this statistic by mathematical analysis. To obtain the permutation distribution for the OBP statistic, we merely need to substitute its formula for the statistic in our sample programs in Chapter 3.

Goodness of Fit

A metric can always be derived. Before we begin to develop our model, divide the data at random into two parts, one of which will be used for model development and estimation, the other for validation. Our goodness-of-fit metric is

$$G = \frac{\sum_{k\in\{\text{validation}\}}(Y_{\text{observed}} - Y_{\text{predicted}})^2}{\sum_{k\in\{\text{estimation}\}}(Y_{\text{observed}} - Y_{\text{predicted}})^2},$$

where the summation in the numerator is taken over all the observations in the validation data set and the summation in the denominator is taken over all the observations in the estimation data set.

For two reasons, this ratio will almost always be larger than unity:

1. The estimation data set, not the validation set, was used to choose the variables that went into the model and to decide whether to use the original values of the variables or to employ some sort of transformation.
2. The estimation data set, not the validation set, was used to estimate the values of the model coefficients.

Divide the original data set into two parts at random a second time, but use the estimation set only to calculate the values of the coefficients. Use the same model you used before, that is, if $\log[X]$ was used in the original model, use $\log[X]$ in this new one. Compute G a second time.

Repeat this resampling process several hundred times. If the original model is appropriate for prediction purposes, it will provide a relatively good fit to most of the data sets; if not, the goodness-of-fit statistic for our original estimation set will be among the largest of the values, since its denominator will be among the smallest.

8.5.2 Cross-Validation

Five techniques of cross-validation are in general use:

K-fold, in which we subdivide the data into K roughly equal-sized parts, then repeat the modeling process K times, leaving one section out each time for validation purposes,

Leave-one-out, an extreme example of K-fold, in which we subdivide into as many parts as there are observations. We leave one observation out of our classification procedure, and use the remaining $n - 1$ observations as a training set. Repeating this procedure n times, omitting a different observation each time, we arrive at a figure for the number and percentage of observations classified correctly. A method that requires

Guidelines for Model Building

Your objective in modeling is prediction, not goodness of fit:

- Use the minimum number of predictors—over-fitted models are numerically unstable.
- Use automated methods only to produce several apparently good models that can be investigated further.
- Include only predictors with which a plausible cause and effect relationship can be established.
- Replace multiple highly correlated (collinear) predictors with an average or some other linear combination. This procedure recommends itself when a coefficient that should be positive (negative) has the opposite sign.
- When investigating models for several similar products, select the model that fits all the products reasonably well, rather than trying to find the best model for each product. (Any exception should be justified on a cause and effect basis.)

this much computation would have been unthinkable before the advent of inexpensive readily available high-speed computers. Today, at worst, we need step out for a cup of coffee while our desktop completes its efforts.

Jackknife, an obvious generalization of the leave-one-out approach, where the number left out can range from one observation to half the sample.

Delete–d, where we set aside a random percentage d of the observations for validation purposes, use the remaining $100 - d\%$ as a training set, then average over 100 to 200 such independent random samples. The goodness-of-fit statistic is an example of delete–50.

The *bootstrap* is a fifth, recommended alternative that we discuss at length in Section 8.4.3.

We may use any of these methods to determine which of several competing models to adopt. For the data from Fawlty Towers, let us use the delete–50% cross-validation method, dividing our sample into two equal parts at random. We'll use the first part of the data to estimate the model parameters and the second part to check the accuracy of the resultant model or models.

For example, we might use the six pairs of observations (289, 235), (391, 355), (358, 275), (339, 315), (160, 159), (319, 305) to estimate the regression coefficients for each of the competing models, and use the residuals when we try to fit the remaining six pairs to determine the prediction errors.

We repeat this process 20 or 30 times—selecting six points, estimating the regression coefficients, computing the prediction errors. Our model of choice is the one that minimizes the average prediction error.

8.5.3 Using the Bootstrap for Model Validation

In the late 1970's, Peter Gregory observed 155 chronic hepatitis patients at Stanford Hospital and made 20 observations on each one summarizing medical histories, physical

examinations, X rays, liver function tests, and biopsies. He used this data to develop a prediction rule for use in identifying patients at high risk.

Subsequently, Gail Gong [1986] applied the bootstrap to the data Gregory had gathered. Not surprisingly, the stepwise regression method gave rise to a somewhat different set of predictors for each bootstrap sample. Surprisingly, only four of the predictors were common to each of the 100 models she developed. The conclusions to be drawn from this are two in number. First, any regression model (linear or nonlinear) should be regarded with a great deal of skepticism. Second, the bootstrap is invaluable for differentiating amongst the essential and inessential components of a model.

8.6 Summary

In this chapter, you reviewed the steps in model development beginning with descriptive statistics such as scatter plots and correlations to identify relationships among variables. You also considered some of the possible caveats. You learned a number of methods for ascertaining the statistical significance of model coefficients and for deriving the associated confidence intervals. You learned it is essential to distinguish between goodness of fit and prediction, and were provided with a variety of cross-validation resampling methods to tie the two together.

8.7 To Learn More

Regression is a rich and complex topic. Ryan [1997] provides a review of the many methods of estimation. Mosteller and Tukey [1977] and Good and Hardin [2003] document the many pitfalls. An excellent example of LAD regression is given in Cade and Richards [1996]. A gentle introduction to quantile regression is provided by Cade and Noon [2003].

Attempts at formulating exact permutation tests for the coefficients of multipredictor regression are summarized by Anderson and Legendre [1999]. Good [2000; pp. 127–130] cites the lack of exchangeability as responsible for the deviations from the declared significance level.

The resampling approach to model validation has been used in archeology [Berry et al. 1980], business administration [Carter and Catlett 1987; Chhikara 1989; Thompson, Bridges, and Ensor 1992], chemistry [Penninckx et al. 1996], DNA sequencing (Doolittle [1981] and Karlin et al. [1983] adopt permutation methods while Hasegawa, Kishino, and Yano [1988] apply the bootstrap), ecology [Solow 1990], medicine [Arndt et al. 1996; Bullmore et al. 1996; Titterington et al. 1981], nuclear power generation [Dubuisson and Lavison 1980], and physics [Priesendorfer and Barnett 1983]. For an example of the bootstrap's application in a nonlinear regression, see Shimbukaro et al. [1984].

Mielke [1986] reviews alternate metrics. Marron [1987], Hjorth [1994], and Stone [1974] provide a survey of parameter selection and cross-validation methods. Techniques for model validation also are reviewed in Shao and Tu [1995; p. 306–313] who show that the delete–50% method, first proposed by Geisser [1975], is far superior to delete–1.

8.8 Exercises

1. Using the data of Table 8.1, determine whether there is a statistically significant correlation between meals served and the following variables:
   - **a.** Registrants (R)
   - **b.** Log(R)
   - **c.** Sin (R/500)
2. Characterize the following relationships as positive linear, negative linear, or nonlinear:
   - **a.** Distance from the top of a bathtub to the surface of the water as a function of time when you fill the tub, then pull the plug.
   - **b.** Same problem, only now you turn on the taps full blast.
   - **c.** Sales as a function of your advertising budget.
   - **d.** Blood pressure as you increase the dose of a blood-pressure lowering medicine.
   - **e.** Electricity use as a function of the day of the year.
   - **f.** Number of bacteria at the site of infection as you increase the dose of an antibiotic.
   - **g.** Size of an untreated tumor over time.
3. Which is the cause and which the effect?
   - **a.** overpopulation and poverty
   - **b.** highway speed limits and number of accidents
   - **c.** cases of typhus and water pollution
4. In Table 5.1, breaks appear to be related to log dose in accordance with the formula breaks $= a + b \log[\text{dose} + 0.01]$.
   - **a.** estimate b
   - **b.** what does a represent? How would you estimate a?
5. Suppose you wanted to predict the sales of automobile parts; which observations would be critical? Number of automobiles in service? Number of automobiles under warranty? Number of automobiles more than a given number of years in service? M2, the total money supply that is readily available in cash, savings and checking accounts, and so forth? Retail sales? Consumer confidence? Weather forecasts? Would your answer depend on whether you owned a dealership or an independent automobile parts chain? (By the way, in the studies Mark Kaiser and I completed, weather was the most significant variable. Go figure.)
6. Compute and compare the apparent error rates of the two Fawlty Towers prediction models.
7. Use the delete-50% method to compare the two Fawlty Towers prediction models.
8. Suppose you've collected the prices of a number of stocks over a period of time and have developed a model with which to predict their future behavior. How would you go about validating your model?
9. Growth rates for chicks whose calorie-deficient diet was supplemented with vitamin B are summarized below by log-dose and sex. Is the dose response the same for the two sexes?

	0.301	0.602	0.903	1.204	1.505
males	17.1,14.3,21.6	24.5,20.6,23.8	27.7,31.0,29.4	28.6,34.2,37.3	33.3,31.8,40.2
females	18.5,22.1,15.3	23.6,26.9,20.2	24.3,27.1,30.1	30.3,33.0,35.8	32.6,36.1,30.5

10. Hosmer and Lemeshow [1989] provide data on 189 births at a U.S. hospital. How would you go about predicting a low (defined as less than 2.5 kilograms)?
The last value in each row in the table below is the birth weight in grams. Reading across each row the variables are observation number, low birth weight? (no 0, yes 1), age of mother in years, mother's weight in pounds at last menstrual period, race (white 1, black 2, other 3), smoked during pregnancy? (no 0, yes 1), number of previous premature labors, history of hypertension (no 0, yes 1), has uterine irritability (no 0, yes 1), number of physician visits in the first trimester, and birth weight in grams.

4,1,28,120,3,1,1,0,1,0,709	10,1,29,130,1,0,0,0,1,2,1021
11,1,34,187,2,1,0,1,0,0,1135	13,1,25,105,3,0,1,1,0,0,1330
15,1,25,85,3,0,0,0,1,0,1474	16,1,27,150,3,0,0,0,0,0,1588
17,1,23,97,3,0,0,0,1,1,1588	18,1,24,128,2,0,1,0,0,1,1701
19,1,24,132,3,0,0,1,0,0,1729	20,1,21,165,1,1,0,1,0,1,1790
22,1,32,105,1,1,0,0,0,0,1818	23,1,19,91,1,1,2,0,1,0,1885
24,1,25,115,3,0,0,0,0,0,1893	25,1,16,130,3,0,0,0,0,1,1899
26,1,25,92,1,1,0,0,0,0,1928	27,1,20,150,1,1,0,0,0,2,1928
28,1,21,200,2,0,0,0,1,2,1928	29,1,24,155,1,1,1,0,0,0,1936
30,1,21,103,3,0,0,0,0,0,1970	31,1,20,125,3,0,0,0,1,0,2055
32,1,25,89,3,0,2,0,0,1,2055	33,1,19,102,1,0,0,0,0,2,2082
34,1,19,112,1,1,0,0,1,0,2084	35,1,26,117,1,1,1,0,0,0,2084
36,1,24,138,1,0,0,0,0,0,2100	37,1,17,130,3,1,1,0,1,0,2125
40,1,20,120,2,1,0,0,0,3,2126	42,1,22,130,1,1,1,0,1,1,2187
43,1,27,130,2,0,0,0,1,0,2187	44,1,20,80,3,1,0,0,1,0,2211
45,1,17,110,1,1,0,0,0,0,2225	46,1,25,105,3,0,1,0,0,1,2240
47,1,20,109,3,0,0,0,0,0,2240	49,1,18,148,3,0,0,0,0,0,2282
50,1,18,110,2,1,1,0,0,0,2296	51,1,20,121,1,1,1,0,1,0,2296
52,1,21,100,3,0,1,0,0,4,2301	54,1,26,96,3,0,0,0,0,0,2325
56,1,31,102,1,1,1,0,0,1,2353	57,1,15,110,1,0,0,0,0,0,2353
59,1,23,187,2,1,0,0,0,1,2367	60,1,20,122,2,1,0,0,0,0,2381
61,1,24,105,2,1,0,0,0,0,2381	62,1,15,115,3,0,0,0,1,0,2381
63,1,23,120,3,0,0,0,0,0,2410	65,1,30,142,1,1,1,0,0,0,2410
67,1,22,130,1,1,0,0,0,1,2410	68,1,17,120,1,1,0,0,0,3,2414
69,1,23,110,1,1,1,0,0,0,2424	71,1,17,120,2,0,0,0,0,2,2438
75,1,26,154,3,0,1,1,0,1,2442	76,1,20,105,3,0,0,0,0,3,2450
77,1,26,190,1,1,0,0,0,0,2466	78,1,14,101,3,1,1,0,0,0,2466
79,1,28,95,1,1,0,0,0,2,2466	81,1,14,100,3,0,0,0,0,2,2495
82,1,23,94,3,1,0,0,0,0,2495	83,1,17,142,2,0,0,1,0,0,2495
84,1,21,130,1,1,0,1,0,3,2495	85,0,19,182,2,0,0,0,1,0,2523
86,0,33,155,3,0,0,0,0,3,2551	87,0,20,105,1,1,0,0,0,1,2557
88,0,21,108,1,1,0,0,1,2,2594	89,0,18,107,1,1,0,0,1,0,2600

91,0,21,124,3,0,0,0,0,0,2622 92,0,22,118,1,0,0,0,0,1,2637
93,0,17,103,3,0,0,0,0,1,2637 94,0,29,123,1,1,0,0,0,1,2663
95,0,26,113,1,1,0,0,0,0,2665 96,0,19,95,3,0,0,0,0,0,2722
97,0,19,150,3,0,0,0,0,1,2733 98,0,22,95,3,0,0,1,0,0,2751
99,0,30,107,3,0,1,0,1,2,2750 100,0,18,100,1,1,0,0,0,0,2769
101,0,18,100,1,1,0,0,0,0,2769 102,0,15,98,2,0,0,0,0,0,2778
103,0,25,118,1,1,0,0,0,3,2782 104,0,20,120,3,0,0,0,1,0,2807
105,0,28,120,1,1,0,0,0,1,2821 106,0,32,121,3,0,0,0,0,2,2835
107,0,31,100,1,0,0,0,1,3,2835 108,0,36,202,1,0,0,0,0,1,2836
109,0,28,120,3,0,0,0,0,0,2863 111,0,25,120,3,0,0,0,1,2,2877
112,0,28,167,1,0,0,0,0,0,2877 113,0,17,122,1,1,0,0,0,0,2906
114,0,29,150,1,0,0,0,0,2,2920 115,0,26,168,2,1,0,0,0,0,2920
116,0,17,113,2,0,0,0,0,1,2920 117,0,17,113,2,0,0,0,0,1,2920
118,0,24,90,1,1,1,0,0,1,2948 119,0,35,121,2,1,1,0,0,1,2948
120,0,25,155,1,0,0,0,0,1,2977 121,0,25,125,2,0,0,0,0,0,2977
123,0,29,140,1,1,0,0,0,2,2977 124,0,19,138,1,1,0,0,0,2,2977
125,0,27,124,1,1,0,0,0,0,2922 126,0,31,215,1,1,0,0,0,2,3005
127,0,33,109,1,1,0,0,0,1,3033 128,0,21,185,2,1,0,0,0,2,3042
129,0,19,189,1,0,0,0,0,2,3062 130,0,23,130,2,0,0,0,0,1,3062
131,0,21,160,1,0,0,0,0,0,3062 132,0,18,90,1,1,0,0,1,0,3062
133,0,18,90,1,1,0,0,1,0,3062 34,0,32,132,1,0,0,0,0,4,3080
135,0,19,132,3,0,0,0,0,0,3090 136,0,24,115,1,0,0,0,0,2,3090
137,0,22,85,3,1,0,0,0,0,3090 138,0,22,120,1,0,0,1,0,1,3100
139,0,23,128,3,0,0,0,0,0,3104 140,0,22,130,1,1,0,0,0,0,3132
141,0,30,95,1,1,0,0,0,2,3147 42,0,19,115,3,0,0,0,0,0,3175
143,0,16,110,3,0,0,0,0,0,3175 144,0,21,110,3,1,0,0,1,0,3203
145,0,30,153,3,0,0,0,0,0,3203 146,0,20,103,3,0,0,0,0,0,3203
147,0,17,119,3,0,0,0,0,0,3225 148,0,17,119,3,0,0,0,0,0,3225
149,0,23,119,3,0,0,0,0,2,3232 150,0,24,110,3,0,0,0,0,0,3232
151,0,28,140,1,0,0,0,0,0,3234 154,0,26,133,3,1,2,0,0,0,3260
155,0,20,169,3,0,1,0,1,1,3274 156,0,24,115,3,0,0,0,0,2,3274
159,0,28,250,3,1,0,0,0,6,3303 160,0,20,141,1,0,2,0,1,1,3317
161,0,22,158,2,0,1,0,0,2,3317 162,0,22,112,1,1,2,0,0,0,3317
163,0,31,150,3,1,0,0,0,2,3321 164,0,23,115,3,1,0,0,0,1,3331
166,0,16,112,2,0,0,0,0,0,3374 167,0,16,135,1,1,0,0,0,0,3374
168,0,18,229,2,0,0,0,0,0,3402 169,0,25,140,1,0,0,0,0,1,3416
170,0,32,134,1,1,1,0,0,4,3430 172,0,20,121,2,1,0,0,0,0,3444
173,0,23,190,1,0,0,0,0,0,3459 174,0,22,131,1,0,0,0,0,1,3460
175,0,32,170,1,0,0,0,0,0,3473 176,0,30,110,3,0,0,0,0,0,3544
177,0,20,127,3,0,0,0,0,0,3487 179,0,23,123,3,0,0,0,0,0,3544
180,0,17,120,3,1,0,0,0,0,3572 181,0,19,105,3,0,0,0,0,0,3572
182,0,23,130,1,0,0,0,0,0,3586 183,0,36,175,1,0,0,0,0,0,3600
184,0,22,125,1,0,0,0,0,1,3614 185,0,24,133,1,0,0,0,0,0,3614
186,0,21,134,3,0,0,0,0,2,3629 187,0,19,235,1,1,0,1,0,0,3629
188,0,25,95,1,1,3,0,1,0,3637 189,0,16,135,1,1,0,0,0,0,3643
190,0,29,135,1,0,0,0,0,1,3651 191,0,29,154,1,0,0,0,0,1,3651

192,0,19,147,1,1,0,0,0,0,3651 193,0,19,147,1,1,0,0,0,0,3651
195,0,30,137,1,0,0,0,0,1,3699 196,0,24,110,1,0,0,0,0,1,3728
197,0,19,184,1,1,0,1,0,0,3756 199,0,24,110,3,0,1,0,0,0,3770
200,0,23,110,1,0,0,0,0,1,3770 201,0,20,120,3,0,0,0,0,0,3770
202,0,25,241,2,0,0,1,0,0,3790 203,0,30,112,1,0,0,0,0,1,3799
204,0,22,169,1,0,0,0,0,0,3827 205,0,18,120,1,1,0,0,0,2,3856
206,0,16,170,2,0,0,0,0,4,3860 207,0,32,186,1,0,0,0,0,2,3860
208,0,18,120,3,0,0,0,0,1,3884 209,0,29,130,1,1,0,0,0,2,3884
210,0,33,117,1,0,0,0,1,1,3912 211,0,20,170,1,1,0,0,0,0,3940
212,0,28,134,3,0,0,0,0,1,3941 213,0,14,135,1,0,0,0,0,0,3941
214,0,28,130,3,0,0,0,0,0,3969 215,0,25,120,1,0,0,0,0,2,3983
216,0,16,95,3,0,0,0,0,1,3997 217,0,20,158,1,0,0,0,0,1,3997
218,0,26,160,3,0,0,0,0,0,4054 219,0,21,115,1,0,0,0,0,1,4054
220,0,22,129,1,0,0,0,0,0,4111 221,0,25,130,1,0,0,0,0,2,4153
222,0,31,120,1,0,0,0,0,2,4167 223,0,35,170,1,0,1,0,0,1,4174
224,0,19,120,1,1,0,0,0,0,4238 225,0,24,116,1,0,0,0,0,1,4593

9

Decision Trees

How can we tell whether an incoming letter is spam without opening it? Or whether the stranger on the phone is a telemarketer? To what genus should a newfound species be assigned? In making these decisions, what factors should be considered? Which are most meaningful?

I've told my wife, who works in a hospital, to turn back immediately should she ever hear creepy music while walking down a hallway. I reached this conclusion by abstracting from the many slasher movies I've watched in which creepy music always precedes an attack. No music, no danger. Observation on a single variable leads to an immediate decision.

On the other hand, it takes 22 points of comparison for the FBI to match a fingerprint with one of the millions of prints on file in its electronic database. Again, a decision tree is involved in which the answers to a series of questions lead finally to a decision. What questions should we ask to help us make our decisions? Which answers will be most meaningful? In this chapter, you learn how to construct decision trees and study their use in classification and as an alternative to regression in model development. You also learn that the ambiguities inherent in stepwise regression are still present with decision trees and that validation again is essential.

9.1 Classification

The sepal length, sepal width, petal length, and petal width of 150 iris plants were recorded by Anderson [1935] and are reproduced in Table 9.1. The observations may be downloaded from http://stat.bus.utk.edu/Stat579/iris.txt. Our first clues to the number of subpopulations or categories of iris, as well as to the general shape of the underlying frequency distribution, come from consideration of the histogram of Figure 9.1. A glance suggests the presence of at least two species, though because of the overlap of the various subpopulations it is difficult to be sure. Three species actually are present as shown in Figure 9.2.

Proceeding a variable at a time (Figure 9.3), CART constructs a classification tree (Figure 9.4) for the iris data consisting of two nodes. First, CART establishes that petal length is the most informative of the four variables and the tree is split on the

Table 9.1. The Iris Data

Species	sepal length	sepal width	petal length	petal width	Species	sepal length	sepal width	petal length	petal width
1	5.1	3.5	1.4	0.2	2	6.6	3	4.4	1.4
1	4.9	3	1.4	0.2	2	6.8	2.8	4.8	1.4
1	4.7	3.2	1.3	0.2	2	6.7	3	5	1.7
1	4.6	3.1	1.5	0.2	2	6	2.9	4.5	1.5
1	5	3.6	1.4	0.2	2	5.7	2.6	3.5	1
1	5.4	3.9	1.7	0.4	2	5.5	2.4	3.8	1.1
1	4.6	3.4	1.4	0.3	2	5.5	2.4	3.7	1
1	5	3.4	1.5	0.2	2	5.8	2.7	3.9	1.2
1	4.4	2.9	1.4	0.2	2	6	2.7	5.1	1.6
1	4.9	3.1	1.5	0.1	2	5.4	3	4.5	1.5
1	5.4	3.7	1.5	0.2	2	6	3.4	4.5	1.6
1	4.8	3.4	1.6	0.2	2	6.7	3.1	4.7	1.5
1	4.8	3	1.4	0.1	2	6.3	2.3	4.4	1.3
1	4.3	3	1.1	0.1	2	5.6	3	4.1	1.3
1	5.8	4	1.2	0.2	2	5.5	2.5	4	1.3
1	5.7	4.4	1.5	0.4	2	5.5	2.6	4.4	1.2
1	5.4	3.9	1.3	0.4	2	6.1	3	4.6	1.4
1	5.1	3.5	1.4	0.3	2	5.8	2.6	4	1.2
1	5.7	3.8	1.7	0.3	2	5	2.3	3.3	1
1	5.1	3.8	1.5	0.3	2	5.6	2.7	4.2	1.3
Species	sepal length	sepal width	petal length	petal width	Species	sepal length	sepal width	petal length	petal width
1	5.4	3.4	1.7	0.2	2	5.7	3	4.2	1.2
1	5.1	3.7	1.5	0.4	2	5.7	2.9	4.2	1.3
1	4.6	3.6	1	0.2	2	6.2	2.9	4.3	1.3
1	5.1	3.3	1.7	0.5	2	5.1	2.5	3	1.1
1	4.8	3.4	1.9	0.2	2	5.7	2.8	4.1	1.3
1	5	3	1.6	0.2	3	6.3	3.3	6	2.5
1	5	3.4	1.6	0.4	3	5.8	2.7	5.1	1.9
1	5.2	3.5	1.5	0.2	3	7.1	3	5.9	2.1
1	5.2	3.4	1.4	0.2	3	6.3	2.9	5.6	1.8
1	4.7	3.2	1.6	0.2	3	6.5	3	5.8	2.2
1	4.8	3.1	1.6	0.2	3	7.6	3	6.6	2.1
1	5.4	3.4	1.5	0.4	3	4.9	2.5	4.5	1.7
1	5.2	4.1	1.5	0.1	3	7.3	2.9	6.3	1.8
1	5.5	4.2	1.4	0.2	3	6.7	2.5	5.8	1.8
1	4.9	3.1	1.5	0.2	3	7.2	3.6	6.1	2.5
1	5	3.2	1.2	0.2	3	6.5	3.2	5.1	2
1	5.5	3.5	1.3	0.2	3	6.4	2.7	5.3	1.9
1	4.9	3.6	1.4	0.1	3	6.8	3	5.5	2.1
1	4.4	3	1.3	0.2	3	5.7	2.5	5	2
1	5.1	3.4	1.5	0.2	3	5.8	2.8	5.1	2.4
1	5	3.5	1.3	0.3	3	6.4	3.2	5.3	2.3
1	4.5	2.3	1.3	0.3	3	6.5	3	5.5	1.8
1	4.4	3.2	1.3	0.2	3	7.7	3.8	6.7	2.2

Table 9.1. *Continued*

Species	sepal length	sepal width	petal length	petal width	Species	sepal length	sepal width	petal length	petal width
1	5	3.5	1.6	0.6	3	7.7	2.6	6.9	2.3
1	5.1	3.8	1.9	0.4	3	6	2.2	5	1.5
1	4.8	3	1.4	0.3	3	6.9	3.2	5.7	2.3
1	5.1	3.8	1.6	0.2	3	5.6	2.8	4.9	2
1	4.6	3.2	1.4	0.2	3	7.7	2.8	6.7	2
1	5.3	3.7	1.5	0.2	3	6.3	2.7	4.9	1.8
1	5	3.3	1.4	0.2	3	6.7	3.3	5.7	2.1
2	7	3.2	4.7	1.4	3	7.2	3.2	6	1.8
2	6.4	3.2	4.5	1.5	3	6.2	2.8	4.8	1.8
2	6.9	3.1	4.9	1.5	3	6.1	3	4.9	1.8
2	5.5	2.3	4	1.3	3	6.4	2.8	5.6	2.1
2	6.5	2.8	4.6	1.5	3	7.2	3	5.8	1.6
2	5.7	2.8	4.5	1.3	3	7.4	2.8	6.1	1.9
2	6.3	3.3	4.7	1.6	3	7.9	3.8	6.4	2
2	4.9	2.4	3.3	1	3	6.4	2.8	5.6	2.2
2	6.6	2.9	4.6	1.3	3	6.3	2.8	5.1	1.5
2	5.2	2.7	3.9	1.4	3	6.1	2.6	5.6	1.4
2	5	2	3.5	1	3	7.7	3	6.1	2.3
2	5.9	3	4.2	1.5	3	6.3	3.4	5.6	2.4
2	6	2.2	4	1	3	6.4	3.1	5.5	1.8
2	6.1	2.9	4.7	1.4	3	6	3	4.8	1.8
2	5.6	2.9	3.6	1.3	3	6.9	3.1	5.4	2.1
2	6.7	3.1	4.4	1.4	3	6.7	3.1	5.6	2.4
Species	sepal length	sepal width	petal length	petal width	Species	sepal length	sepal width	petal length	petal width
2	5.6	3	4.5	1.5	3	6.9	3.1	5.1	2.3
2	5.8	2.7	4.1	1	3	5.8	2.7	5.1	1.9
2	6.2	2.2	4.5	1.5	3	6.8	3.2	5.9	2.3
2	5.6	2.5	3.9	1.1	3	6.7	3.3	5.7	2.5
2	5.9	3.2	4.8	1.8	3	6.7	3	5.2	2.3
2	6.1	2.8	4	1.3	3	6.3	2.5	5	1.9
2	6.3	2.5	4.9	1.5	3	6.5	3	5.2	2
2	6.1	2.8	4.7	1.2	3	6.2	3.4	5.4	2.3
2	6.4	2.9	4.3	1.3	3	5.9	3	5.1	1.8

classification rule "petal length less than or equal to 2.450." Node 2 is split on the rule "petal width less than or equal to 1.750." CART concludes the sepal measurements do not contribute significant additional information.

As shown in Table 9.2, 96% are correctly classified! A moment's thought dulls the excitement; this is the same set of observations we used to develop our selection criteria, so naturally we were successful. It's our success with new plants whose species we don't know in advance that is the true test.

Not having a set of yet-to-be-classified plants on hand, CART makes use of the existing data in a K-fold validation as described in Section 8.5.2. The original group of

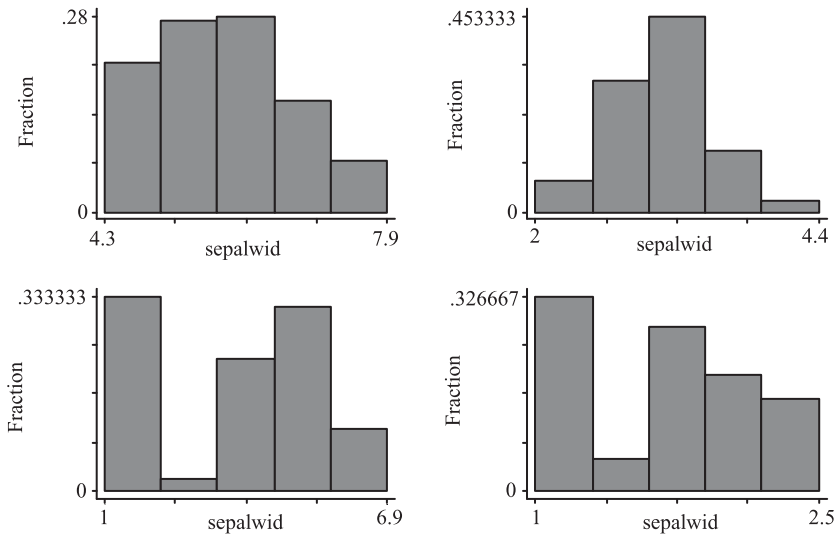


Fig. 9.1. Sepal and petal measurements of 150 iris plants.

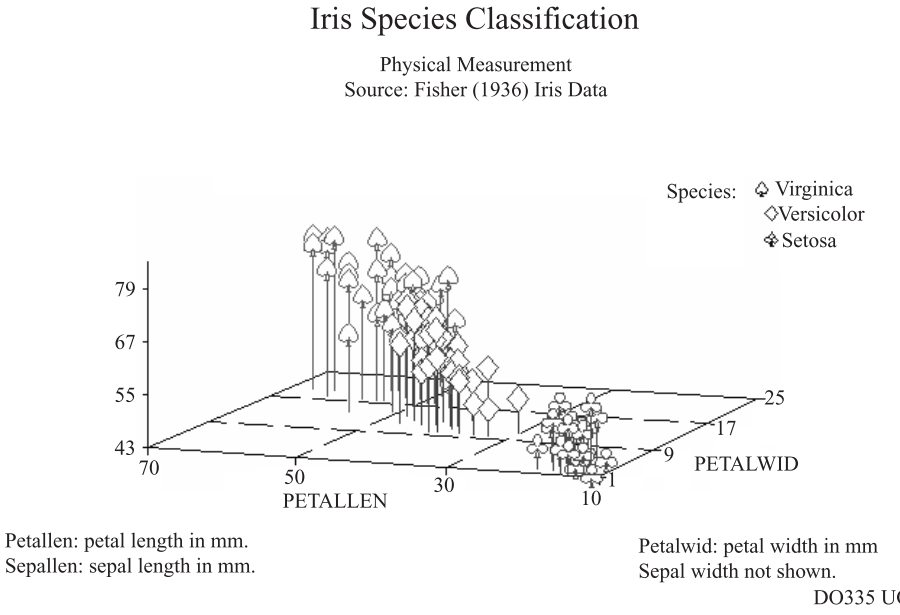


Fig. 9.2. Representing three variables in two dimensions. Iris species representation derived with the help of SAS/Graph.

150 plants is divided into 10 subgroups, each consisting of 15 plants. The binary-tree-fitting procedure is applied 10 times in succession, each time omitting one of the 10 groups, each time applying the tree it developed to the omitted group. Only 10 of the

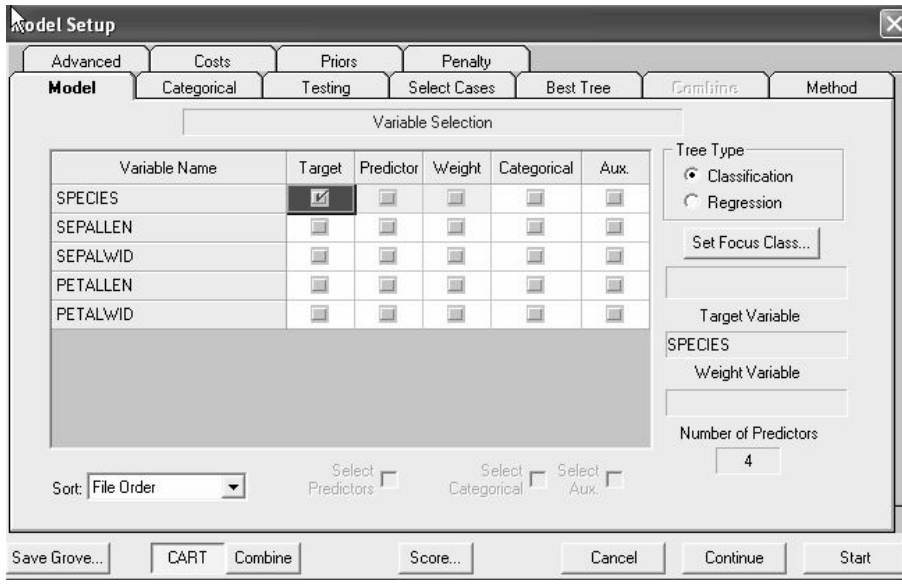


Fig. 9.3. Setting up the CART model for the iris data.

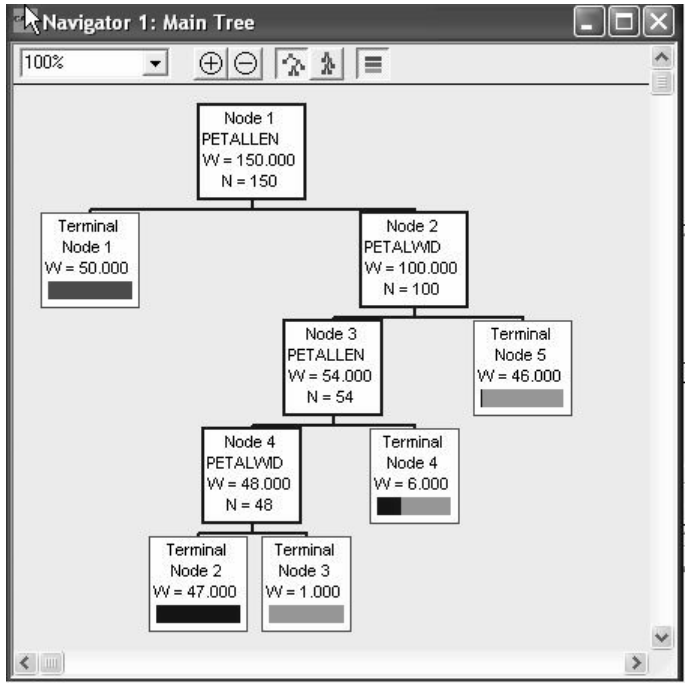


Fig. 9.4. Decision tree for iris classification.

Table 9.2. Learning Sample Classification Table

Actual Class	Predicted Class 1	2	3	Actual Total
1	50.000	0.000	0.000	50.000
2	0.000	49.000	1.000	50.000
3	0.000	5.000	45.000	50.000
Pred. Tot.	50.000	54.000	46.000	150.000
Correct	1.000	0.980	0.900	
Success Ind.	0.667	0.647	0.567	
Tot. Correct	0.960			

Table 9.3. Cross-Validation Classification

	Cross-Validation				Learning Sample		
Class	Prior Prob.	N	Mis-Classified	Cost	N	Mis-Classified	Cost
1	0.333	50	0	0.000	50	0	0.000
2	0.333	50	5	0.100	50	1	0.020
3	0.333	50	5	0.100	50	5	0.100
Tot	1.000	150	10		150	6	

150 plants or 7.5% are misclassified (Table 9.3), a 92.5% success rate, suggesting that in future samples consisting of irises of unknown species, the system of classification developed here would be successful 92.5% of the time.

We would have obtained similar results had we used the following *R* commands after installing and loading the "tree" library:

- `# Species must be a factor`
- `iris.tr=tree(Species ~ SL+SW+PL+PW)`
- `plot(iris.tr); text(iris.tr,srt=90)`

SAS PROC TREE has a misleading name as it does not yield a decision tree, but rather makes use of the nearest-neighbor procedure decribed by Wong and Lane [1983] to form clusters.

Excel users can form decision trees by downloading Ctree**,** a macro-filled Excel spreadsheet, from
http://www.geocities.com/adotsaha/CTree/CtreeinExcel.html

9.2 Consumer Survey

The objective of a study of consumers' attitudes, interests, and opinions was to direct an advertising campaign for a new model of automobile toward the most likely customers. Those surveyed were questioned on their attitudes toward risk, foreign-made products, product styling, spending habits, emissions, pollution, self-image, family, and so forth. A final question concerned the potential customer's attitude toward purchasing the product itself. All responses were tabulated on a nine-point Likert scale.

In *R* format, the principle results were as follows:

Purchase = c(6, 9, 8, 3, 5, 1, 3, 3, 7, 4, 2, 8, 6, 1, 3, 6, 1, 9, 9, 7, 9, 2, 2, 8, 8, 5, 1, 3, 7, 9, 3, 6, 9, 8, 5, 4, 8, 9, 6, 2, 8, 5, 6, 5, 5, 3, 7, 6, 4, 5, 9, 2, 8, 2, 8, 7, 9, 4, 3, 3, 4, 1, 3, 6, 6, 5, 2, 4, 2, 8, 7, 7, 6, 1, 1, 9, 4, 4, 6, 9, 1, 6, 9, 6, 2, 8, 6, 3, 5, 3, 6, 8, 2, 5, 6, 7, 7, 5, 7, 6, 3, 5, 8, 8, 1, 9, 8, 8, 7, 5, 2, 2, 3, 8, 2, 2, 8, 9, 5, 6, 7, 4, 6, 5, 8, 4, 7, 8, 2, 1, 7, 9, 7, 5, 5, 9, 9, 9, 7, 3, 8, 9, 8, 4, 8, 5, 5, 8, 4, 3, 7, 1, 2, 1, 1, 7, 5, 5, 1, 4, 1, 2, 9, 7, 6, 9, 9, 6, 5, 4, 3, 6, 6, 4, 5, 7, 2, 6, 5, 6, 3, 8, 2, 5, 3, 4, 2, 3, 8, 3, 9, 1, 3, 1, 2, 5, 1, 5, 6, 7, 1, 1, 1, 4, 4, 8, 4, 7, 4, 4, 2, 6, 6, 6, 7, 2, 9, 4, 1, 9, 3, 5, 7, 2, 2, 8, 9, 2, 4, 1, 7, 1, 3, 6, 2, 6, 2, 8, 4, 4, 1, 1, 2, 2, 8, 3, 3, 3, 1, 1, 6, 8, 3, 7, 5, 9, 8, 3, 5, 6, 3, 4, 6, 1, 1, 5, 6, 6, 9, 6, 9, 9, 6, 7, 3, 8, 4, 2, 6, 4, 8, 3, 3, 6, 4, 4, 9, 5, 6, 4, 5, 3, 3, 2, 5, 9, 5, 1, 3, 4, 3, 6, 8, 1, 5, 3, 4, 8, 2, 5, 3, 2, 3, 2, 5, 8, 3, 1, 6, 3, 7, 8, 9, 2, 3, 5, 7, 7, 3, 7, 3, 9, 2, 9, 3, 9, 2, 8, 9, 5, 1, 9, 9, 1, 8, 7, 1, 4, 9, 3, 4, 9, 1, 3, 9, 1, 5, 2, 7, 9, 6, 5, 7, 4, 6, 1, 4, 2, 7, 5, 4, 5, 9, 5, 5, 5, 2, 4, 1, 8, 7, 9, 6, 8, 1, 5, 9, 9, 9, 9, 1, 3, 3, 7, 2, 5, 6, 1, 5, 8)

Fashion = c(5, 6, 8, 2, 5, 2, 5, 1, 7, 3, 5, 5, 4, 3, 3, 5, 6, 3, 4, 3, 4, 6, 4, 6, 3, 6, 5, 4, 6, 5, 5, 3, 4, 4, 4, 3, 6, 2, 3, 4, 4, 4, 5, 2, 3, 4, 5, 5, 6, 4, 5, 5, 6, 3, 4, 4, 5, 8, 4, 5, 6, 4, 2, 5, 3, 6, 2, 3, 2, 5, 3, 5, 4, 4, 5, 4, 6, 6, 5, 8, 2, 6, 5, 6, 4, 7, 4, 5, 5, 3, 6, 6, 4, 5, 5, 4, 4, 4, 4, 3, 5, 3, 3, 5, 4, 4, 5, 7, 6, 6, 4, 4, 5, 5, 2, 2, 7, 5, 1, 6, 5, 4, 7, 7, 6, 5, 6, 3, 2, 4, 5, 3, 9, 4, 4, 6, 6, 6, 9, 4, 4, 3, 3, 3, 2, 4, 4, 5, 4, 6, 6, 3, 3, 3, 5, 4, 4, 5, 4, 6, 3, 4, 6, 3, 4, 6, 4, 5, 4, 3, 3, 6, 4, 3, 3, 4, 3, 1, 4, 5, 5, 6, 2, 6, 6, 5, 5, 3, 9, 3, 3, 1, 1, 4, 3, 3, 3, 7, 6, 6, 4, 4, 1, 3, 5, 5, 4, 6, 4, 5, 5, 4, 6, 5, 6, 2, 4, 4, 3, 8, 5, 3, 6, 5, 3, 5, 3, 3, 5, 3, 2, 2, 3, 5, 5, 5, 1, 6, 5, 1, 5, 4, 4, 3, 6, 4, 4, 5, 5, 4, 5, 5, 3, 7, 4, 7, 6, 1, 5, 4, 4, 4, 3, 3, 5, 4, 7, 4, 6, 7, 6, 4, 6, 3, 4, 4, 2, 6, 3, 6, 5, 2, 2, 5, 3, 4, 4, 4, 3, 2, 4, 6, 4, 6, 5, 6, 2, 4, 2, 3, 6, 2, 6, 5, 6, 4, 4, 4, 6, 5, 5, 1, 4, 5, 5, 4, 4, 2, 3, 6, 5, 5, 2, 2, 5, 2, 5, 4, 3, 8, 3, 6, 3, 4, 3, 6, 4, 3, 4, 2, 5, 6, 4, 5, 5, 6, 4, 6, 5, 4, 3, 8, 2, 5, 5, 3, 2, 3, 5, 4, 3, 4, 3, 5, 2, 3, 1, 4, 4, 6, 6, 6, 6, 6, 6, 4, 4, 3, 4, 4, 3, 3, 5, 4, 4, 5, 4, 6, 8, 3, 3, 5, 4, 5, 4, 5, 4, 4, 6, 6)

Gamble = c(5, 4, 7, 4, 5, 4, 3, 3, 3, 6, 2, 6, 5, 4, 5, 5, 2, 7, 6, 6, 6, 4, 2, 8, 4, 4, 3, 3, 4, 5, 4, 3, 4, 6, 5, 4, 8, 9, 7, 3, 6, 4, 6, 6, 5, 3, 4, 6, 5, 4, 5, 3, 7, 3, 8, 5, 7, 5, 3, 4, 7, 4, 4, 5, 4, 6, 1, 4, 4, 9, 5, 4, 6, 4, 4, 5, 5, 5, 6, 6, 4, 4, 8, 7, 4, 5, 3, 3, 5, 3, 4, 5, 3, 5, 6, 6, 6, 5, 7, 4, 2, 3, 7, 6, 6, 4, 8, 4, 6, 3, 4, 4, 5, 8, 3, 3, 4, 5, 5, 5, 4, 5, 1, 6, 8, 5, 6, 4, 4, 6, 5, 7, 6, 5, 6, 7, 7, 6, 6, 4, 7, 6, 6, 5, 7, 6, 6, 5, 2, 5, 5, 4, 3, 3, 4, 6, 4, 4, 4, 3, 5, 2, 6, 4, 4, 6, 7, 6, 5, 4, 4, 7, 4, 7, 8, 5, 4, 5, 5, 3, 2, 4, 3, 6, 2, 3, 6, 5, 2, 4, 6, 2, 2, 1, 4, 3, 5, 5, 4, 6, 2, 3, 3, 5, 3, 6, 5, 6, 5, 3, 4, 6, 5, 5, 5, 3, 7, 7, 2, 6, 4, 2, 7, 2, 6, 3, 6, 2, 5, 3, 7, 3, 4, 2, 3, 7, 3, 6, 3, 7, 2, 4, 4, 4, 8, 3, 4, 4, 3, 1, 4, 7, 5, 4, 5, 8, 4, 6, 4, 6, 4, 4, 5, 4, 2, 6, 5, 5, 7, 2, 7, 4, 5, 6, 5, 3, 3, 2, 5, 3, 6, 1, 5, 6, 6, 5, 8, 6, 6, 5, 6, 4, 4, 6, 4, 7, 4, 4, 5, 4, 3, 7, 8, 1, 4, 4, 7, 4, 5, 4, 5, 1, 3, 4, 4, 4, 5, 3, 5, 5, 4, 7, 6, 3, 6, 4, 6, 5, 3, 4, 5, 7, 4, 5, 4, 5, 3, 7, 6, 4, 6, 4, 8, 3, 4, 2, 5, 5, 5, 4, 5, 6, 3, 5, 8, 4, 5, 2, 5, 4, 5, 6, 3, 3, 3, 1, 5, 3, 4, 7, 4, 4, 6, 4, 3, 5, 3, 4, 4, 8, 6, 7, 4, 6, 4, 5, 4, 6, 8, 7, 2, 5, 4, 7, 4, 5, 6, 4, 6, 6)

Ozone = c(4, 7, 7, 3, 4, 2, 3, 4, 2, 5, 4, 6, 2, 3, 6, 3, 7, 8, 7, 5, 5, 5, 5, 8, 5, 5, 5, 1, 4, 6, 6, 9, 2, 6, 3, 4, 6, 6, 7, 5, 6, 6, 6, 4, 3, 5, 6, 5, 5, 3, 5, 4, 5, 6, 8, 3, 8, 5, 6, 4, 4, 4, 3, 7, 8, 5, 3, 6, 6, 8, 6, 4, 5, 6, 4, 3, 6, 6, 3, 5, 5, 6, 7, 6, 6, 7, 3, 4, 5, 5, 6, 5, 3, 3, 5, 6, 3, 4, 6, 5, 5, 6, 6, 5, 9, 8, 5, 5, 4, 8, 4, 3, 5, 5, 4, 6, 8, 8, 4, 7, 5, 9, 2, 2, 5, 2, 7, 7, 2, 4, 4, 6, 3, 7, 7, 4, 3, 6, 3, 6, 6, 7, 3, 5, 5, 4, 3, 6, 4, 5, 6, 5, 5, 4, 7, 5, 5, 2, 4, 7, 5, 5, 5, 4, 6, 5, 5, 5, 7, 5, 3, 6, 5, 6, 6, 4, 4, 2, 6, 6, 4, 8, 3, 5, 3, 3, 5, 5, 6, 5, 7, 4, 1, 3, 4, 6, 4, 3, 8, 5, 2, 7, 1, 5, 3,

7, 5, 4, 3, 7, 4, 2, 8, 7, 4, 3, 6, 7, 6, 6, 7, 9, 9, 3, 7, 6, 6, 4, 5, 6, 6, 4, 6, 5, 7, 5, 4, 6, 5, 6, 5, 5, 5, 4, 4, 6, 9, 3, 3, 2, 5, 5, 5, 7, 3, 6, 4, 5, 7, 5, 4, 5, 5, 6, 6, 7, 4, 4, 4, 4, 2, 7, 4, 5, 4, 4, 5, 3, 6, 4, 7, 6, 4, 6, 5, 4, 5, 5, 4, 5, 7, 1, 3, 8, 6, 7, 5, 5, 5, 4, 5, 6, 5, 3, 5, 2, 3, 3, 4, 3, 3, 5, 5, 7, 7, 5, 6, 6, 6, 4, 7, 5, 7, 5, 8, 7, 7, 4, 5, 6, 6, 4, 9, 8, 5, 6, 6, 4, 4, 5, 4, 6, 3, 5, 4, 5, 8, 6, 6, 5, 3, 6, 7, 4, 7, 5, 4, 3, 6, 4, 6, 6, 4, 5, 5, 3, 7, 4, 6, 7, 3, 5, 6, 4, 9, 6, 3, 5, 7, 4, 5, 3, 7, 3, 3, 6, 6, 4, 6, 6, 6, 5, 5, 9, 4, 3, 6, 3, 4, 6)

Pollution = c(5, 7, 7, 4, 5, 1, 3, 6, 3, 5, 5, 6, 2, 2, 7, 2, 6, 7, 7, 5, 5, 5, 6, 8, 6, 4, 5, 1, 4, 6, 6, 9, 3, 6, 3, 6, 4, 6, 8, 6, 6, 5, 6, 5, 3, 3, 8, 7, 7, 3, 4, 5, 5, 6, 8, 3, 8, 5, 6, 5, 4, 6, 5, 7, 7, 6, 4, 6, 5, 8, 5, 6, 6, 6, 4, 4, 5, 6, 5, 6, 5, 6, 8, 5, 5, 6, 5, 4, 5, 6, 5, 6, 3, 4, 6, 6, 5, 6, 6, 5, 4, 6, 8, 4, 9, 7, 6, 4, 5, 9, 4, 4, 4, 5, 4, 5, 7, 8, 3, 7, 7, 7, 2, 2, 5, 3, 5, 7, 4, 5, 5, 7, 5, 5, 6, 5, 4, 7, 4, 7, 7, 7, 4, 5, 5, 5, 3, 6, 5, 5, 7, 6, 4, 6, 7, 4, 6, 4, 4, 7, 6, 6, 7, 4, 6, 5, 6, 5, 6, 5, 4, 6, 6, 5, 6, 5, 6, 3, 6, 6, 5, 7, 4, 3, 3, 4, 6, 5, 5, 6, 6, 5, 2, 4, 4, 6, 2, 3, 6, 5, 4, 6, 2, 5, 2, 8, 4, 5, 4, 7, 5, 1, 7, 5, 6, 4, 5, 7, 7, 5, 6, 8, 8, 5, 7, 5, 5, 6, 6, 5, 6, 4, 7, 5, 6, 5, 4, 5, 5, 6, 7, 5, 6, 4, 6, 4, 8, 4, 4, 2, 4, 5, 6, 6, 4, 5, 6, 4, 6, 6, 4, 5, 4, 7, 6, 7, 5, 5, 5, 4, 2, 6, 3, 5, 3, 4, 5, 3, 5, 4, 7, 7, 4, 6, 5, 5, 4, 6, 4, 6, 6, 1, 4, 7, 5, 8, 3, 6, 4, 4, 4, 6, 7, 6, 5, 3, 3, 5, 4, 4, 4, 6, 5, 7, 6, 4, 7, 6, 6, 4, 8, 4, 8, 6, 7, 8, 8, 4, 5, 4, 6, 5, 7, 7, 5, 6, 6, 5, 6, 6, 4, 6, 6, 5, 4, 7, 6, 6, 5, 6, 3, 4, 9, 5, 6, 5, 3, 5, 6, 4, 6, 4, 4, 6, 6, 3, 7, 5, 5, 7, 4, 6, 6, 4, 9, 6, 5, 6, 7, 4, 5, 3, 5, 4, 3, 7, 6, 6, 6, 6, 5, 6, 6, 9, 6, 3, 6, 3, 5, 7).

With R a tree is grown by binary recursive partitioning using the response in the specified formula and choosing splits from the terms of the right-hand side. Numeric variables are divided into $X\langle a$ and $X\rangle a$; the levels of an unordered factor are divided into two nonempty groups. The split that maximizes the reduction in impurity is chosen, the data set split, and the process repeated. Splitting continues until the terminal nodes are too small or too few to be split. Figure 9.5 is an example of a tree we constructed using the data from this consumer survey and the following R commands:

- `buy.tr=tree(Purchase ~ Fashion+Ozone+Gamble+Pollution)`
- `plot(buy.tr); text(buy.tr,srt=90)`

Some of the decisions seem to be made on the basis of a single predictor—Gamble—others utilize the values of two predictors—Gamble and Fashion—and still others utilize the values of Gamble, Fashion, and Ozone, the equivalent of three separate linear regression models. In each case, the Pollution variable is discarded as of little predictive value.

The output from this example is a *regression tree*, that is, each *terminal node* has a numerical value. This is misleading in the present case where our objective is to pinpoint the most likely sales prospects.

Suppose instead that we begin by grouping our customer attitudes into categories. Purchase attitudes of 1, 2, or 3 indicate low interest, 4, 5, and 6 indicate medium interest, 7, 8, and 9 indicate high interest. To convert them using R, we enter

- ```
  Pur.cat = cut(Purchase,breaks=c(0,3,6,9),
      labels = c(''lo'',''med'',''hi''))
  ```
  ```

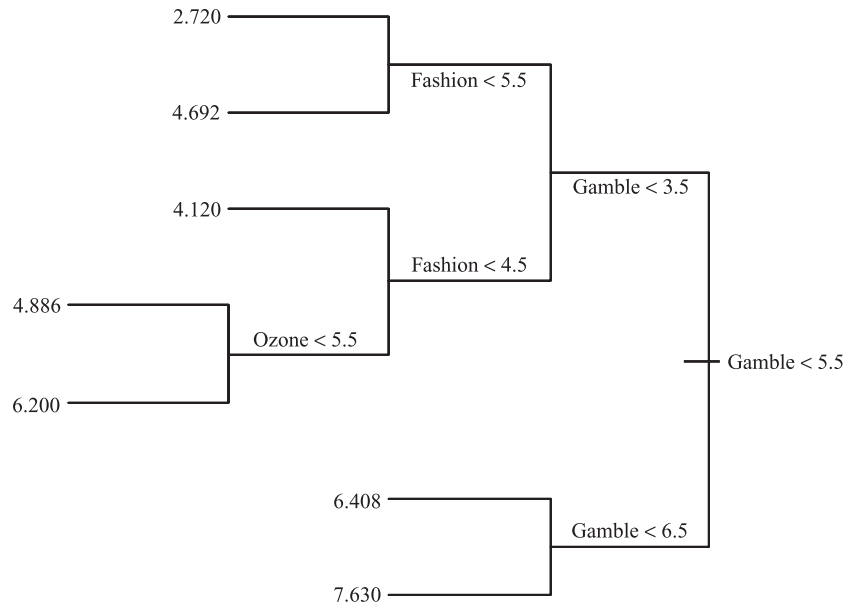


Fig. 9.5. Labeled regression tree for predicting purchase attitude.

Now, we can construct the classification tree depicted in Figure 9.6.

- `catt.tr=tree(Pur.cat~Fashion+Ozone+Gamble)`
- `plot(catt.tr,type=''u''); text(catt.tr,srt=90)`

Many of the branches of this tree appear redundant. If we have already classified a prospect, there is no point in additional questions. We can remove the excess branches and create a display similar to Figure 9.7 with the following *R* code:

- `tmp=prune.tree(catt.tr, best=4)`
- `plot(tmp); text(tmp)`

Note that CART would have pruned the tree automatically.
These classifications are actually quite tentative; typing the command

- `print(tmp)`

will display the nodes in this tree along with the proportions of correct and incorrect assignments.

```
node), split, n, deviance, yval, (yprob)
   * denotes terminal node

1) root 400 878.70 med ( 0.3225 0.3375 0.3400 )
    2) Gamble < 5.5 275 567.60 lo
         ( 0.1782 0.4618 0.3600 )
```

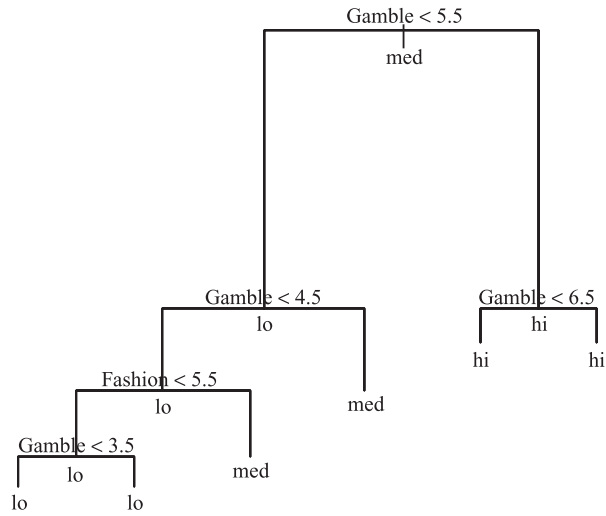


Fig. 9.6. Labeled classification tree for predicting buyer prospect attitude.

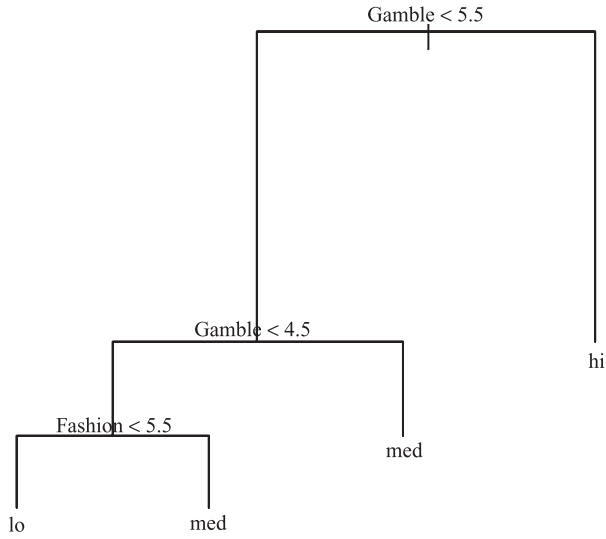


Fig. 9.7. Pruned classification tree for predicting prospect attitude.

```
4) Gamble < 4.5 192 372.30 lo
     ( 0.1615 0.5677 0.2708 )
8) Fashion < 5.5 160 281.00 lo
     ( 0.1187 0.6438 0.2375 ) *
9) Fashion > 5.5 32  66.77 med
     ( 0.3750 0.1875 0.4375 ) *
```

```
5) Gamble > 4.5 83 163.50 med
     ( 0.2169 0.2169 0.5663 ) *
3) Gamble > 5.5 125 205.50 hi
     ( 0.6400 0.0640 0.2960 ) *
```

The three fractional values within parentheses refer to the proportions in the data set that were assigned to the "hi," "low," and "med" categories, respectively. For example, in the case when Gamble > 5.5, 64% are assigned correctly by the classification scheme, while just over 6% who are actually poor prospects (with scores of 3 or less) are assigned in error to the high category.

9.3 Trees Versus Regression

Recall that the basis of the regression techniques discussed in Chapter 8 was a desire to minimize the sum of the potential losses $\sum_i L(\hat{\theta}_i - \theta_i)$. With regression techniques, we are restricted to estimates that take the linear form $\hat{\theta}_i = AX_i$ and loss functions L that can be minimized by least-squares or linear programming techniques. With decision trees, we are able to specify loss functions of arbitrary form. Moreover, if we have some prior knowledge of the distribution of the various classifications in the population at large (as opposed to what may be a quite different distribution in the sample at hand), we can take advantage of this knowledge to derive an improved estimate.

Figure 9.8 depicts the model setup in CART for predicting low birth weight (LOW) on the basis of the mother's AGE and RACE, whether she smoked during pregnancy (SMOKE), had a history of hypertension (HT), uterine irritability (UI), how often she

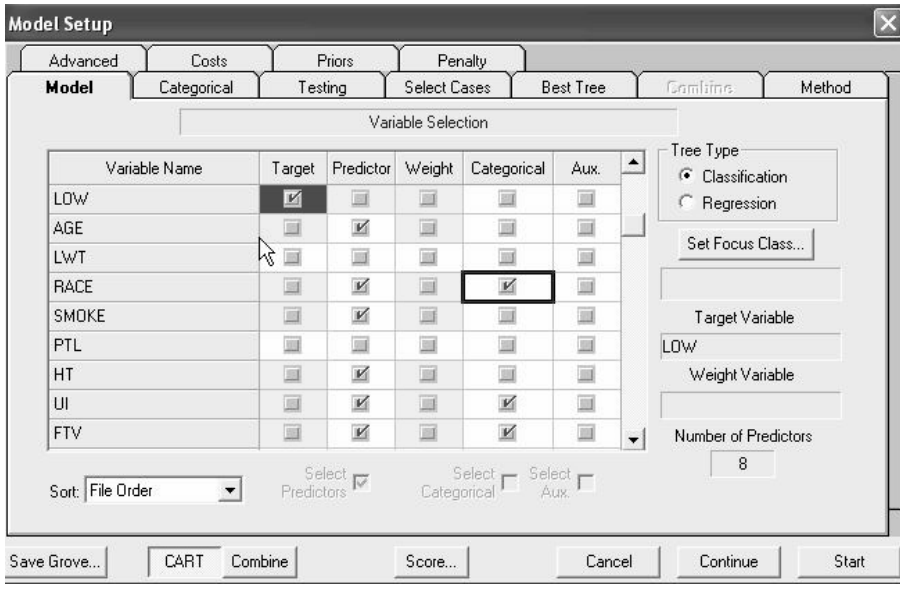


Fig. 9.8. Model setup in CART for predicting low birth weight.

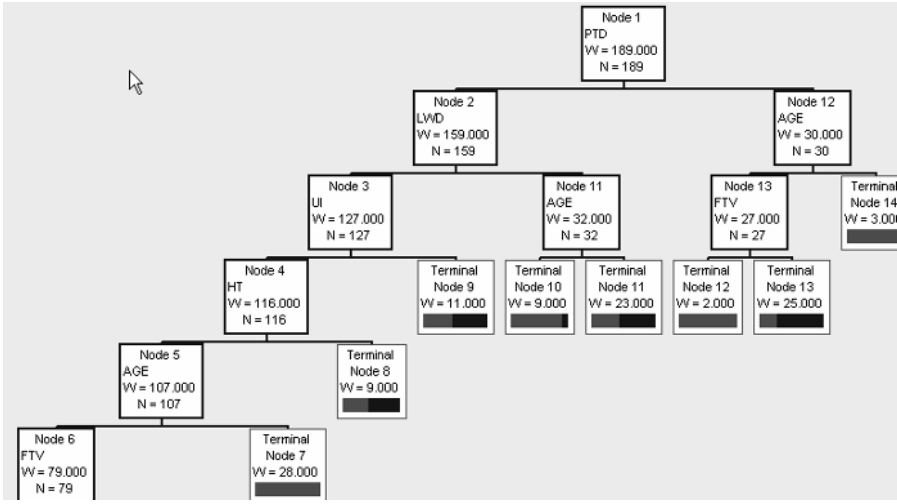


Fig. 9.9. Decision tree for equal costs equal frequency in population.

visited a physician during the first trimester (FTV), and so forth. The CART defaults are that the costs of misclassification are the same for both categories and that low and high birth weights are equally likely outcomes. The resulting decision tree is displayed in Figure 9.9.

Note that in contrast to continuous predictors like petal length and age for which decision trees send a case left if [split-variable] <= [split-value], with categorical predictors like race and histories, a case goes left if [split-variable] = [level1, level2, level(n)].

Figure 9.10 depicts the model setup when a higher cost is associated with predicting a normal birth weight when the actual birth weight is low, and Figure 9.11 depicts the resulting decision tree.

The frequencies of the various classications are seldom equal, nor are the sample frequencies necessarily representative of the population. To reduce the overall error rate, we may specify the actual frequencies as shown in Figure 9.12. The resulting decision tree is shown in Figure 9.13.

9.3.1 Head-to-Head Comparison

In this section, we fit several models to the same data set, first by regression means and then via a decision tree. The data consists of the median value (MV) of owner-occupied homes in about 500 U.S. census tracts in the Boston area, along with several potential predictors including

- CRIM per capita crime rate by town,
- ZN proportion of residential land zoned for lots over 25,000 sq.ft,
- INDUS proportion of nonretail business acres per town,
- CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise),

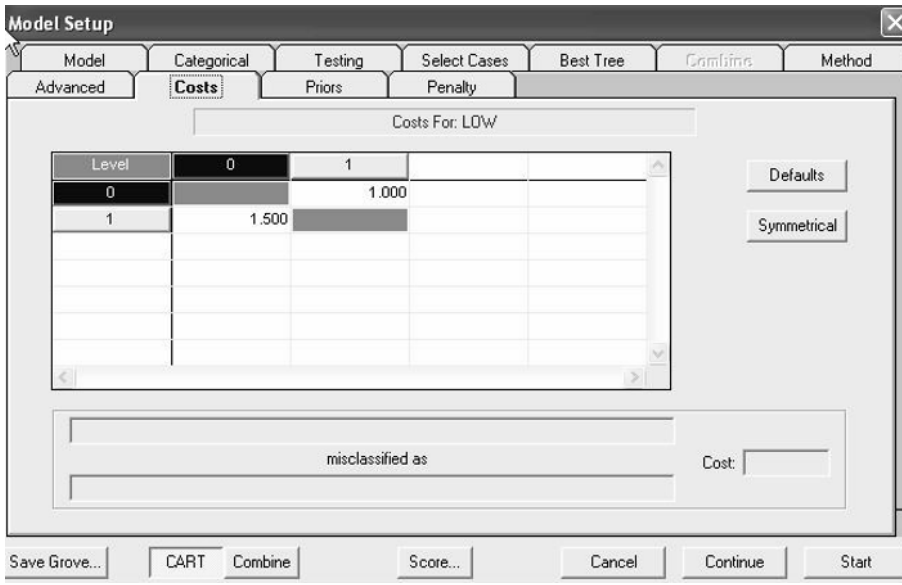


Fig. 9.10. Model setup in CART for unequal error costs.

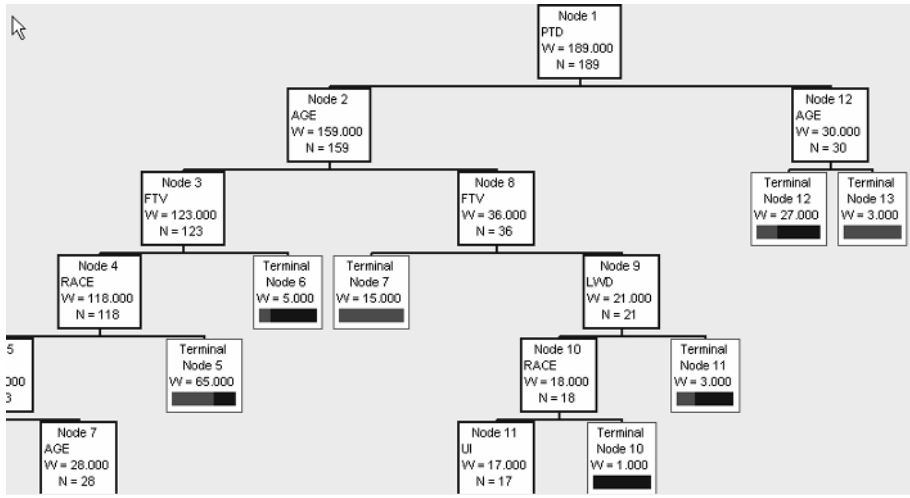


Fig. 9.11. Decision tree for unequal costs, equal frequency in population.

- NOX nitric oxides concentration (parts per 10 million),
- RM average number of rooms per dwelling,
- AGE proportion of owner-occupied units built prior to 1940,
- DIS weighted distances to five Boston employment centers,
- RAD index of accessibility to radial highways,

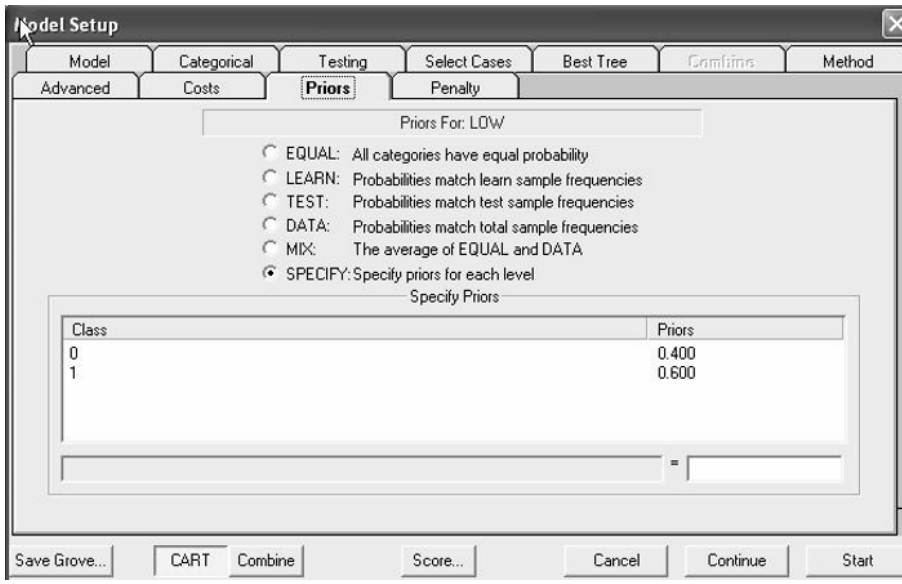


Fig. 9.12. Model setup in CART for unequal frequencies in population.

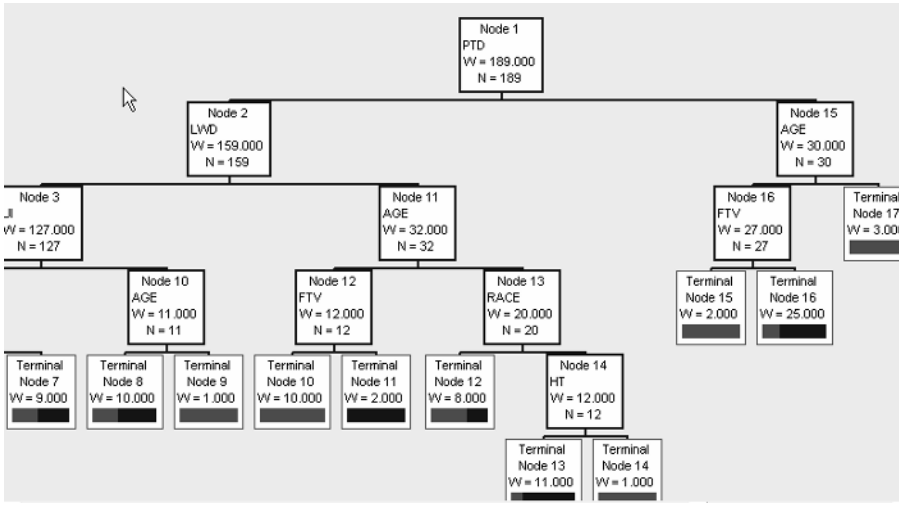


Fig. 9.13. Decision tree for equal costs, unequal frequency of categories in population.

- TAX full-value property-tax rate per $10,000,
- PT pupil-teacher ratio by town,
- LSTAT % lower status of the population.

The data may be downloaded from http://lib.stat.cmu.edu/datasets/boston.

Using *R* to obtain the desired results via stepwise regression

- `summary(step(lm(MV~ CRIM+ZN+INDUS+CHAS+NOX+RM+AGE+DIS+RAD+TAX+PT+B+ LSTAT)))`

and fitting with respect to the AIC criterion yields the model

```
Step:  AIC= 1585.76
MV = 36.3 -0.11CRIM + 0.05ZN + 2.7CHAS + 17.38NOX +
     3.80RM + 1.49DIS + 0.30RAD -0.01 TAX + 0.94PT +
     0.009B - 0.52LSTAT
With Multiple R-Squared: 0.7406,   Adjusted R-squared:
     0.7348
```

Using *R* to develop a tree

- library("tree")
- `bos.tr=tree(MV~ CRIM+ZN+INDUS+CHAS+NOX+RM+AGE+DIS+RAD+TAX+PT+ B+LSTAT,BHP)`
- `summary(bos.tr)`

yields the following output

```
Regression tree:
Variables actually used in tree construction:
[1] ''RM'' ''LSTAT'' ''DIS'' ''CRIM'' ''PT''
Number of terminal nodes: 9
Residual mean deviance: 13.55 = 6734 / 497
```

The command

- `plot(bos.tr); text(bos.tr, srt=90)`

yields the tree displayed in Figure 9.14. Note that the same variables are employed as were obtained via a stepwise regression.

CART's output (Figure 9.15) for the same problem, even after pruning, would suggest the inclusion of additional variables. For example, nitric oxide concentration is found to be of importance.

CART also offers the option of minimizing the least absolute deviation rather than the least squares as shown in Figure 9.16.

9.3.2 Which Variables?

The cost of collecting data may vary from predictor to predictor. As our potential predictors are interdependent, we may wish to consider what would happen if we were to substitute variables. CART output, Figure 9.17, orders the predictors in their order of importance should they be used in the absence of other information. Though crime rates figure early in the decision tree of Figure 9.15, information concerning crime rates may not be readily available. Figure 9.18 depicts the resultant decision tree if we eliminate crime rate from our potential predictors.

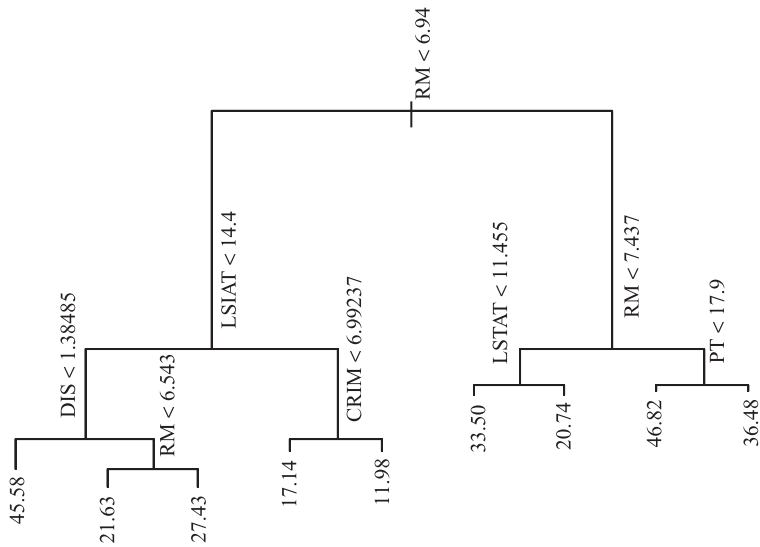


Fig. 9.14. Regression tree for predicting median home value.

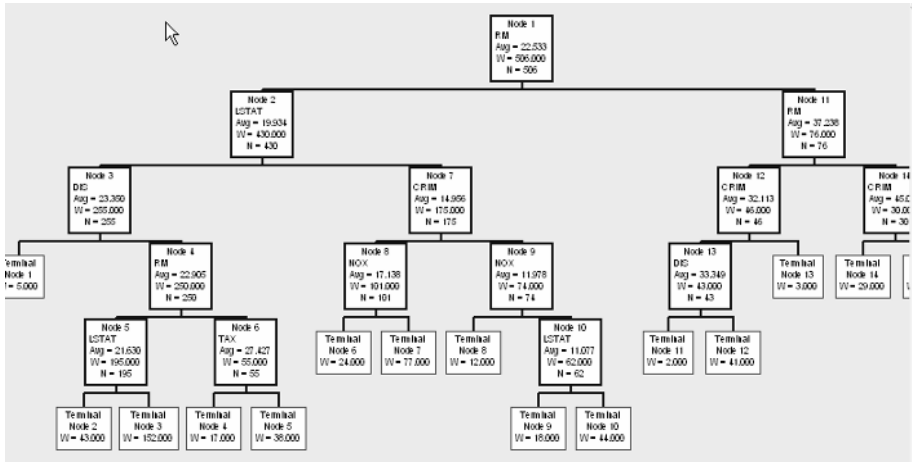


Fig. 9.15. CART decision tree for predicting median house price (OLS).

9.4 To Learn More

Automated construction of a decision tree by repeated resampling dates back to Morgan and Sonquist [1963]. Breiman et al. [1984] pioneered the majority of the algorithms in use today. Clark and Pregibon [1992] provide a probability basis for the approach. Comparisons of the regression and tree approaches were made by Nurminen [2003] and Perlich, Provost, and Simonoff [2003].

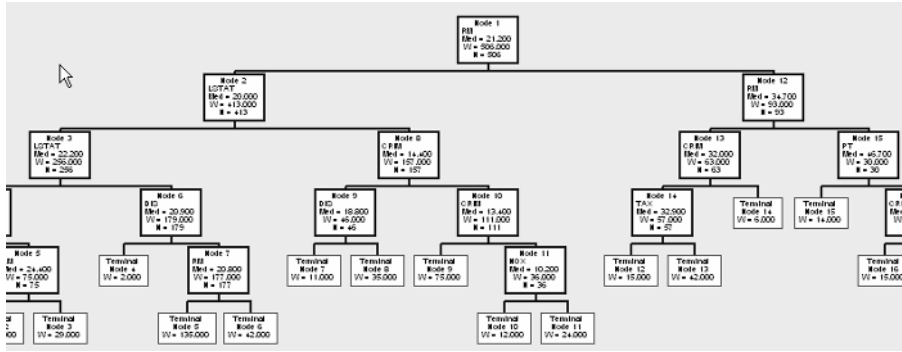


Fig. 9.16. CART decision tree for predicting median house price (LAD).

Variable	Score
LSTAT	100.00
RM	89.56
DIS	28.55
NOX	26.21
PT	25.15
INDUS	22.59
TAX	19.88
AGE	15.69
CRIM	13.18
ZN	11.46
RAD	4.93

Fig. 9.17. Variable selection in order of importance (OLS).

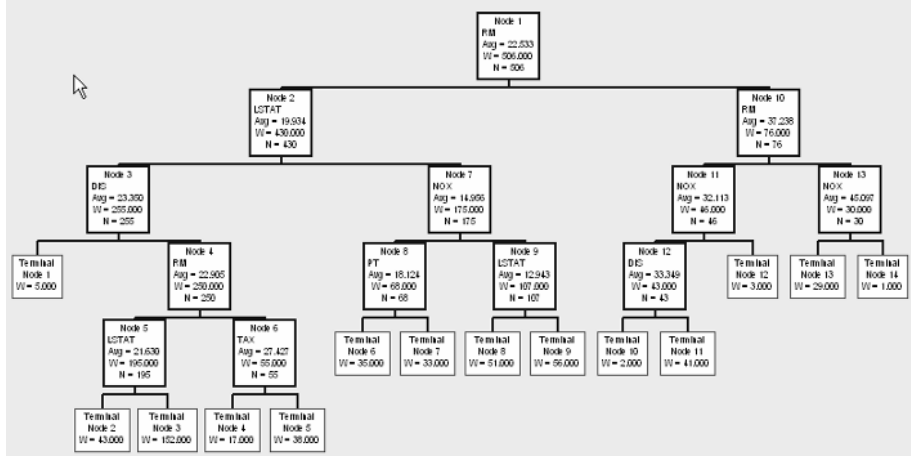


Fig. 9.18. Revised CART decision tree for predicting median house price (OLS).

9.5 Exercises

1. Using the Hosmer and Lemeshow data of the previous chapter (Exercise 10) and discarding "race" as a predictor, fit a decision tree. Does it use the same variables as a stepwise regression? Investigate the effects of pruning.
2. Using the data from http://www.sgi.com/tech/mlc/db/breast.test, build a decision tree to forecast malignancy. A desciption of this database is located at http://www.sgi.com/tech/mlc/db/breastLoss.names
3. Analyze the data located at http://www.cba.ua.edu/∼mhardin/hepatitisdatasets/ by both stepwise regression and decision tree methods. Validate your findings by one or more methods.

Answers to Selected Exercises

Do you always turn to the last page of a detective mystery first? Can't wait to learn whether Little Nell will die? Even so, read this chapter only AFTER you've tried to do the exercises. (The proof will be if you've a sheet of paper in front of you on which you've written out the answers.)

The object of the exercises at the end of each of the preceding chapters is to prepare you for real-world applications. These exercises force you to think through the concepts while giving you the opportunity to work through, test, and modify the program listings before you apply them to your own data. No pain, no gain. Working through problems is the one sure way to gain reusable experience. Or to put it another way: Consult this chapter only AFTER you've tried to do the exercises.

Chapter 2

1, 3, 6. If you repeat your calculations you may get quite different results, since random samples are involved. In any event, the confidence interval must include the original estimate. Note in question 6 that the larger the original sample, the less variation from one set of bootstrap samples to another.

2. Adding $5''$ to every observation is equivalent to shifting the entire frequency distribution $5''$ to the right. Thus, the mean, the median, and every other percentile will increase by $5''$ while the variance will be left unchanged.
   If you convert from feet to inches, you will need to multiply every observation by 12. The mean and standard deviation will be in inches also, so you will need to multiply their original values by 12. As the variance is the square of the standard deviation, you will need to multiply its former value by $12 \times 12 = 144$.
3. Combine the data from the various PPO's to form a single sample.
6. Although it was suggested that you use as many different methods as possible, you should use and report only one in practice. The percentile bootstrap should never be used in practice; strictly a classroom tool, its sole virtue is that it is easy to explain. Which of the remaining bootstrap estimates you use will depend upon the circumstances.

7. As you are attempting to determine the precision of the original estimate and as precision is a function of sample size, your bootstrap samples should be the same size as the original sample.

Chapter 3

1. There are 8 choose 4 possible relabelings of which only one is more extreme than the one we observed.
2. Use the fact that the sum of all the observations in both samples and the sum of the squares of the observations in both samples are the same for all relabelings.
4. p-value close to 0.
5. These differences are paired and you would not use the same statistic as in exercise 4.
6. Presumably you are to test the hypothesis that the fuel additive described in Section 3.3.5 increases mileage by *at least* 10%. Add 10% to each of the control values. Then test the hypothesis that there is no difference against the one-sided alternative that the control values are greater.

Chapter 4

1.a. When the difference is zero, the power is identical with the significance level. When the difference is very large, the power is 1.
3. *d* is always true.
4. Both *a* and *b* are true.
5. You should always decide which statistic you will use before you look at the results of a statistical analysis. Otherwise, the true significance level will far exceed the hypothetical one. See *Common Errors in Statistics* (Wiley, 2003).
8. If you put a $2 bill in your pocket originally, you will always accept the $2 hypothesis so the significance level is zero. If you put a $20 bill in your pocket originally, you will reject the $2 hypothesis 50% of the time so the power is 50%.
9. a. par, b. par or perm, c: par, d: perm, e: par, f: par, g: perm rank, h: perm.
13. For $n = 100$, allowing one defective would result in a significance level of 1-pbinom(1,100,.01) or 26% and a power of 1-pbinom(1,100,.02) or 60%.
14. The paired test as it eliminates one major source of variation.
16. The width of confidence intervals is a decreasing function of sample size. The width of a confidence interval for the difference in means is a function of the size of the combined samples.

Chapter 5

2a.
```
data= c( 221.1,224.2,217.8,208.8,206.9,205.9,211.1,
        198.4,213.0,208.3,214.1,209.1,221.1,208.8,
        211.1,224.2,206.9,198.4)
```

```
size=c(3,3,3,3,3,3)
> f1=F1(size,data)
> #Number MC of simulations determines precision of
     p-value
> MC = 1600
> cnt = 0
> for (i in 1:MC){
+ pdata = sample (data)
+ f1p=F1(size,pdata)
+ # counting number of rearrangements for which F1
     greater than or equal to original
+ if (f1 <= f1p) cnt=cnt+1
+ }
> cnt/N
[1] 0.448
```

4a. If the order in which plants (subjects, experimental units) are selected makes a difference, then the two methods are not equivalent.

4b. Among the methods are using two coins to assign one of four treatments (HH, HT, TH, and TT) and three coins to assign eight treatments. The six-sided die could be used to assign three rows and two columns. Let 1, 2, and 3 assign to the first column.

6. Subtract the initial weight from the final weight in each instance; then make an unordered k-sample comparison of these differences.

8. 1

10. Weight your mean estimate in accordance with the relative sizes of the two populations. To create a confidence interval by bootstrap means, take separate bootstrap samples from the Burb sample and from the City sample; form an estimate of the mean again weighting the individual bootstrap means in accordance with the relative sizes of the two populations.

12. Analyze as an ordered k-sample comparison. Would yours be a one-sided or a two-sided test?

14. Analyze as an unordered k-sample comparison. Would yours be a one-sided or a two-sided test?

15. $9!/3!^3$; $6!/2!^3 \times 3! = 540$.

Chapter 6

1. As we do not know the relative numbers of black and white births during the same period, we cannot answer this question.
2. As we do not know the sizes of the two samples, we cannot answer this question.
4. Analyzing as a singly ordered table, we find a p-value close to 0.
6. Maybe. If both teams are in the same league, the entries would not be independent and Fisher's Exact Test would not be applicable.
9. Two rows and three ordered columns. A Monte Carlo is necessary due to the large sample sizes. Association with the virus is significant at the 1% level.

11. Using either the Jonckheere–Terpstra test or the linear-by-linear association test for doubly ordered tables, differences among the rows are significant at the 1% level.

Chapter 7

4. As with most real-world data, we first need to reformat the observations so as to satisfy the requirements imposed by the software. Thus to use Blossom to obtain Hotelling's-T^2 we have

Id	hmpg	time	hrqi	smpg	srpi
1	20	1	0	19	0
2	23	1	0	22	0
3	21	1	0	20	0
4	25	1	0	24	0
5	18	1	0	17	0
6	17	1	0	16	0
7	18	1	0	17	0
8	24	1	0	23	0
9	20	1	0	19	0
10	24	1	0	22	0
11	23	1	0	22	0
12	19	1	0	18	0
1	24	2	0	23.5	1
2	25	2	0	24.5	1
3	21	2	1	20.5	0
4	22	2	0	20.5	−1
5	23	2	1	22.5	1
6	18	2	−1	16.5	−1
7	17	2	0	16.5	0
8	28	2	1	27.5	0
9	24	2	0	23.5	1
10	27	2	0	25.5	0
11	21	2	0	20.5	0
12	23	2	1	22.5	1

```
MRPP HMPG HRQI SMPG SRPI * TIME * ID
```

Probability (Pearson Type III) of a smaller or equal delta = 0.08

Chapter 8

2. a. positive linear
   b. + or − linear
   c. nonlinear
   d. negative log linear initially

e. nonlinear
f. negative log linear
g. exponential initially

8. Reserve the latest and most current observations for use in validation.
10. Compare with the answer you got for exercise 1 of Chapter 9.

Bibliography

1. Adams D C, Anthony C D. Using randomization techniques to analyse behavioural data. *Animal Behav*. 1996; 51: 733–738.
2. Adderley E E. Nonparametric methods of analysis applied to large-scale seeding experiments. *J. Meteorology*. 1961; 18: 692–694.
3. Agresti A. *Categorical Data Analysis*. New York: John Wiley & Sons; 1990.
4. Agresti A. A survey of exact inference for contingency tables. *Stat. Sci*. 1992; 7: 131–177.
5. Alderson M R, Nayak R A. Study of space-time clustering in Hodgkin's disease in the Manchester Region. *Brit. J. Preventive and Social Medicine*. 1971; 25: 168–173.
6. Aly E-E A A. Simple tests for dispersive ordering. *Stat. Prob. Ltr*. 1990; 9: 323–325.
7. Andersen P K, Borgan O, Gill R D and Keiding N. *Statistical Models Based on Counting Processes*. New York: Springer. 1993.
8. Anderson E. The irises of the Gaspe peninsula. *Bull. Amer. Iris Soc.* 1935; 59: 2–5.
9. Anderson M J, Legendre P. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *J. Staist. Comp. Simul.* 1999; 62: 271–303.
10. Andrews D W K, Buchinsky M. A three-step method for choosing the number of bootstrap repetitions. *Econometrica* 2000; 68(1): 23–51.
11. Antretter E, Dunkel D & Haring C. The WHO/EURO multi-centre study of suicidal behaviour: Findings of the Austrian research centre in Europe-wide comparison. *Wien Klin Wochenschr*. 2000; 112: 955–964.
12. Arndt S, Cizadlo T, Andreasen N C, Heckel D, Gold S & Oleary D S. Tests for comparing images based on randomization and permutation methods. *J. Cerebral Blood Flow and Metabolism*. 1996; 16: 1271–9.
13. Baglivo J, Olivier D & Pagano M. Methods for the analysis of contingency tables with large and small cell counts. *JASA*. 1988; 83: 1006–1013.
14. Barbe P, Bertail P. *The Weighted Bootstrap*. New York: Springer-Verlag; 1995.
15. Barbella P, Denby L & Glandwehr J M. Beyond exploratory data analysis: The randomization test. *Math Teacher*. 1990; 83: 144–149.
16. Barton D E, David F N. Randomization basis for multivariate tests. *Bull. Int. Statist. Inst*. 1961; 39(2): 455–467.
17. Basu D. Discussion of Joseph Berkson's paper "In dispraise of the exact test". *J. Statist. Plan. Infer*. 1979; 3: 189–192.
18. Beran R. Prepivoting to reduce error rate of confidence sets. *Biometrika,* 1987; 74: 151–173.
19. Berger V W. Pros and cons of permutation tests. *Statistics in Medicine* 2000; 19: 1319–1328.
20. Berkson J. In dispraise of the exact test. *J. Statist. Plan. Inf.* 1978; 2: 27–42.

21. Berry K J, Kvamme K L & Mielke P W jr. Permutation techniques for the spatial analysis of the distribution of artifacts into classes. *Amer. Antiquity*. 1980; 45: 55–59.
22. Berry K J, Kvamme K L & Mielke P W jr. Improvements in the permutation test for the spatial analysis of the distribution of artifacts into classes. *Amer. Antiquity*. 1983; 48: 547–553.
23. Bishop Y M M, Fienberg S E & Holland P W. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge MA: MIT Press, 1975.
24. Blair C, Higgins J J, Karinsky W, Krom R & Rey J D. A study of multivariate permutation tests which may replace Hotelling's T test in prescribed circumstances. *Multivariate Beh. Res*. 1994; 29: 141–163.
25. Blair R C, Troendle J F & Beck R W. Control of familywise errors in multiple endpoint assessments via stepwise permutation tests. *Statistics in Medicine*. 1996; 15: 1107–1121.
26. Boess F G, Balasuvramanian R, Brammer M J & Campbell I C. Stimulation of muscarinic acetylcholine receptors increases synaptosomal free calcium concentration by protein kinase-dependent opening of L-type calcium channels. *J. Neurochem*. 1990; 55: 230–236.
27. Boschloo R D. Raised conditional level of significance for the 2×2 table when testing the equality of two probabilities. *Statist. Neer*. 1970; 24: 1–35.
28. Boyett J M, Shuster J J. Nonparametric one-sided tests in multivariate analysis with medical applications. *JASA*. 1977; 72: 665–668.
29. Bradley J V. *Distribution Free Statistical Tests*. New Jersey: Prentice Hall; 1968.
30. Breiman L, Friedman J H, Olshen R A & Stone C J. *Classification and Regression Trees.* Monterey CA: Wadsworth and Brooks; 1984.
31. Bross I D J. Taking a covariable into account. *JASA*. 1964; 59: 725–736.
32. Bryant E H. Morphometric adaptation of the housefly, Musa domestica L; in the United States. *Evolution*. 1977; 31: 580–956.
33. Bullmore E, Brammer M, et al. Statistical methods for estimation and inference for functional MR image analysis. *Magn. Res M*. 1996; 35: 261–327.
34. Burgess A P. Gruzelier J H. Short duration power changes in the EEG during recognition memory for words and faces. *Psychophysiology*. 2000; 37: 596–606.
35. Busby D G. Effects of aerial spraying of fenithrothion on breeding white-throated sparrows. *J. Appl. Ecol*. 1990; 27: 745–755.
36. Cade B. Comparison of tree basal area and canopy cover in habitat models: Subalpine forest. *J. Alpine Mgmt*. 1997; 61: 326–335.
37. Cade B, Hoffman H. Differential migration of blue grouse in Colorado. *Auk*. 1993; 110: 70–77.
38. Cade B S, Noon B R. A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*. 2003; 1: 412–420.
39. Cade B S, Richards J D. Permutation tests for least absolute deviation regression. *Biometrics*. 1996; 52: 886–902.
40. Canty A J, Davison A C, Hinkley D V & Ventura V. Bootstrap diagnostics. http://www.stat.cmu.edu/www/cmu-stats/tr/tr726/tr726.html
41. Carter C, Catlett J. Assessing credit card applications using machine learning. *IEEE Expert.* 1987; 2: 71–79.
42. Chernick M R. *Bootstrap Methods: A Practitioner's Guide*. New York: Wiley; 1999.
43. Chhikara R K. State of the art in credit evaluation. *Amer. J. Agric. Econ*. 1989; 71: 1138–1144.
44. Clark L A, Pregibon D. Tree-based models. In *Statistical Models in S,* eds. Chambers J M and Hastie T J. New York: Chapman and Hall; 1992.

45. Cliff A D, Ord J K. Evaluating the percentage points of a spatial autocorrelation coefficient. *Geog. Anal*. 1971; 3: 51–62.
46. Cliff A D, Ord J K. *Spatial Processes: Models and Applications*. London: Pion Ltd; 1981.
47. Cohen A, Sackrowitz H B. Methods of reducing loss of efficiency due to discreteness of distributions. *Statist. Methods Med. Res*. 2003; 12: 23–36.
48. Davis A W. On the effects of moderate nonnormality on Roy's largest root test. *JASA*. 1982; 77: 896–900.
49. Daw N C, Arnold J T, Abushullaih B A, Stenberg P E, White M M, Jayawardene D, Srivastava D K & Jackson C W A. Single intravenous dose murine megakaryocyte growth and development factor potently stimulates platelet production challenging the necessity for daily administration. *Blood.* 1998; 91: 466–474.
50. Diaconis P, Efron B. Computer intensive methods in statistics. *Sci. Amer*. 1983; 48: 116–130.
51. DiCiccio T J, Hall P & Romano J P. On smoothing the bootstrap. *Ann. Statist*. 1989; 17: 692–704.
52. DiCiccio T J, Romano J. A review of bootstrap confidence intervals (with discussions). *JRSS B*. 1988; 50: 163–170.
53. Dietz E J. Permutation tests for the association between two distance matrices. *Systemic Zoology*. 1983; 32: 21–26.
54. Diggle P J, Lange N & Benes F M. Analysis of variance for replicated spatial point paterns in Clinical Neuroanatomy. *JASA*. 1991; 86: 618–625.
55. Do K A, Hall P. On importance resampling for the bootstrap. *Biometrika*. 1991; 78: 161–167.
56. Doolittle R F. Similar amino acid sequences: chance or common ancestry. *Science*. 1981; 214: 149–159.
57. Douglas M E, Endler J A. Quantitative matrix comparisons in ecological and evolutionary investigations. *J. Theoret. Biol*. 1982; 99: 777–795.
58. Draper D, Hodges J S, Mallows C L & Pregibon D. Exchangeability and data analysis (with discussion). *J. Roy. Statist. Soc. A*. 1993; 156: 9–28.
59. Dubuisson B, Lavison P. Surveillance of a nuclear reactor by use of a pattern recognition methodology. *IEEE Trans. Systems, Man, & Cybernetics*. 1980; 10: 603–609.
60. Dupont W D. Sensitivity of Fisher's exact test to minor perturbations in 2×2 contingency tables. *Statist. Med*. 1986; 5: 629–35.
61. Eden T, Yates F. On the validity of Fisher's z test when applied to an actual sample of nonnormal data. *J. Agricultural Sci.* 1933; 23: 6–16.
62. Edgington E S. *Randomization Tests*. 3rd ed. New York: Marcel Dekker; 1995.
63. Efron B. Bootstrap methods: Another look at the jackknife. *Annals Statist*. 1979; 7: 1–26.
64. Efron B. *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: SIAM; 1982.
65. Efron B. Estimating the error rate of a prediction rule: improvements on cross-validation. *JASA*. 1983; 78: 316–331.
66. Efron B. Better bootstrap confidence intervals. (with disc.) *JASA*. 1987; 82: 171–200.
67. Efron B. Bootstrap confidence intervals: good or bad? (with discussion). *Psychol Bull*. 1988; 104: 293–296.
68. Efron B. Six questions raised by the bootstrap. R. LePage and L. Billard, eds. *Exploring the Limits of the Bootstrap*. New York: Wiley; 1992.
69. Efron B, DiCiccio T. More accurate confidence intervals in exponential families. *Biometrika*. 1992; 79: 231–245.
70. Efron B, Tibshirani R. Bootstrap measures for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci*. 1986; 1: 54–77.

71. Efron B, Tibshirani R. Statistical data analysis in the computer age. *Science*. 1991; 253: 390–395.
72. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. New York: Chapman & Hall; 1993.
73. Falk M, Reiss R D. Weak convergence of smoothed and nonsmoothed bootstrap quantiles estimates. *Ann. Prob*. 1989; 17: 362–371.
74. Faris P D, Sainsbury R S. The role of the Pontis Oralis in the generation of RSA activity in the hippocampus of the guinea pig. *Psych. & Beh*. 1990; 47: 1193–1199.
75. Farrar D A, Crump K S. Exact statistical tests for any cocarcinogenic effect in animal assays. *Fund. Appl. Toxicol*. 1988; 11: 652–663.
76. Farrar D A, Crump K S. Exact statistical tests for any cocarcinogenic effect in animal assays. II age adjusted tests. *Fund. Appl. Toxicol*. 1991; 15: 710–721.
77. Feinstein A R. Clinical biostatistics XXIII. The role of randomization in sampling, testing, allocation, and credulous idolatry (part 2). *Clinical Pharm*. 1973; 14: 989–1019.
78. Festinger L C, Carlsmith J M. Cognitive consequences of forced compliance. *J. Abnorm. Soc. Psych.* 1959; 58: 203–210.
79. Fisher R A. *Design of Experiments*. New York: Hafner; 1935.
80. Fisher R A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 1936; 7, 179–188.
81. Ford R D, Colom L V & Bland B H. The classification of medial septum-diagonal band cells as theta-on or theta-off in relation to hippo campal EEG states. *Brain Res*. 1989; 493: 269–282.
82. Foutz R N, Jensen D R, & Anderson G W. Multiple comparisons in the randomization analysis of designed experiments with growth curve responses. *Biometrics*. 1985; 41: 29–37.
83. Frank D, Trzos R J and Good P. Evaluating drug-induced chromosome alterations. *Mutation Res*. 1978; 56: 311–17.
84. Fraumeni J F, Li F P. Hodgkin's disease in childhood: an epidemiological study. *J. Nat. Cancer Inst.* 1969; 42: 681–691.
85. Freedman D A. A note on screening regression equations. *Amer. Statist.* 1983; 37: 152–155.
86. Freeman G H, Halton J H. Note on an exact treatment of contingency, goodness of fit, and other problems of significance. *Biometrika*. 1951; 38: 141–149.
87. Gabriel K R. Some statistical issues in weather experimentation. *Commun Statist. A.* 1979; 8: 975–1015. Gart J. Pont and interval estimation of the common odds ratio in the ocmbination of 2×2 tables with fixed margins. *Biometrika*. 1970; 57: 471–475.
88. Gabriel K R, Feder P. On the distribution of statistics suitable for evaluating rainfall simulation experiments. *Technometrics*. 1969; 11: 149–160.
89. Gabriel K R, Sokal R R. A new statistical approach to geographical variation analysis. *Systematic Zoology*. 1969; 18: 259–270.
90. Gail M, Mantel N. Counting the number of rxc contingency tables with fixed marginals. *JASA*. 1977; 72: 859–862.
91. Gail M H, Tan W Y & Piantadosi S. Tests for no treatment effect in randomized clinical trials. *Biometrika*. 1988; 75: 57–64.
92. Gart J J. *Statistical Methods in Cancer Res., Vol III - The design and analysis of long term animal experiments*. Lyon: IARC Scientific Publications; 1986.
93. Gastwirht J L. Statistical reasoning in the legal setting. *Amer. Statist*. 1992; 46: 55–69.
94. Geisser S. The predictive sample reuse method with applications. *JASA*. 1975; 70: 320–328.
95. Gine E, Zinn J. Necessary conditions for a bootstrap of the mean. *Ann. Statist*. 1989; 17: 684–691.
96. Glass A G, Mantel N. Lack of time-space clustering of childhood leukemia, Los Angeles County 1960–64. *Cancer Res*. 1969; 29: 1995–2001.

97. Glass A G, Mantel N, Gunz F W & Spears G F S. Time-space clustering of childhood leukemia in New Zealand. *J. Nat. Cancer Inst.* 1971; 47: 329–336.
98. Gleason J R. Algorithms for balanced bootstrap simulations. *Amer. Statist*. 1988; 42: 263–266.
99. Gliddentracey C, Greenwood A K. A validation study of the Spanish self directed search using back translation procedures. *J. Career Assess*. 1997; 5: 105–113.
100. Gliddentracey C E, Parraga M I. Assessing the structure of vocational interests among Bolivian university students. *J. Vocational Beh*. 1996; 48: 96–106.
101. Gong G. Cross-validation, the jackknife and the bootstrap: Excess error in forward logistic regression. *JASA*. 1986: 81: 108–113.
102. Gonzalez Jose R, Garcia-Moro C, Dahinten S & Hernandez M. Origin of Fueguian-Patagonians: an approach to population history and structure using R matrix and matrix permutation methods. *Amer. J. Human Biol*. 2002; 14: 308–20.
103. Good P I. *Introduction to Statistics via Resampling Methods and R.* New York: Wiley; 2005.
104. Good P I. *Permutation Tests*. New York: Springer Verlag; 3rd ed, 2004.
105. Good P I. Extensions of the concept of exchangeability and their applications, *J. Modern Appl. Statist. Methods*. 2002; 1: 243–247.
106. Good P I. *Applying Statistics in the Courtroom*. London: Chapman and Hall; 2001.
107. Good P I. *Permutation Tests*. New York: Springer Verlag; 2nd ed, 2000.
108. Good P I. Detection of a treatment effect when not all experimental subjects respond to treatment. *Biometrics*. 1979; 35: 483–489.
109. Good P I, Hardin J. *Common Errors in Statistics.* New York: Wiley; 2003.
110. Goodman L, Kruskal W. Measures of association for cross-classification. *JASA*. 1954; 49: 732–764.
111. Graubard B I, Korn E L. Choice of column scores for testing independence in ordered 2 by K contingency tables. *Biometrics*. 1987; 43: 471–476.
112. Graves G W, Whinston A B. An algorithm for the quadratic assignment probability. *Mgmt. Science.* 1970; 17: 453–471.
113. Grossman D C, Cummings P, Koepsell T D. et al. Firearm safety counseling in primary care pediatrics: A randomized controlled trial. *Pediatrics* 2000; 106: 22–26.
114. Gupta et al. *Community Dentistry and Oral Epidemiology*. 1980; 8: 287–333.
115. Haber M. A comparison of some conditional and unconditional exact tests for 2×2 contingency tables. *Comm. Statist. A*. 1987; 18: 147–156.
116. Hall P. On the bootstrap and confidence intervals. *Ann. Statist*. 1986; 14: 1431–1452.
117. Hall P. Theoretical comparison of bootstrap confidence intervals (with discussion). *Ann. Statist*. 1988; 16: 927–985.
118. Hall P, Hart J D. Bootstrap test for difference between means in nonparametric regression. *JASA*. 1990; 85: 1039–1049.
119. Hall P, Martin M A. On bootstrap resampling and iteration. *Biometrika*. 1988; 75: 661–671.
120. Hall P, Titterington M. The effect of simulation order on level accuracy and power of Monte Carlo tests. *JRSS B*. 1989; 51: 459–467.
121. Hall P, Wilson S R. Two guidelines for bootstrap hypothesis testing. *Biometrics*. 1991; 47: 757–62.
122. Halter J H. A rigorous derivation of the exact contingency formula. *Proc. Cambridge Phil. Soc.* 1969; 65: 527–530.
123. Hårdle W. *Smoothing Techniques with Implementation in S*. New York: Springer-Verlag; 1991.
124. Hartigan J A. Using subsample values as typical values. *JASA*. 1969; 64: 1303–1317.

125. Hartigan J A. Error analysis by replaced samples. *J. Roy. Statist. Soc. B.* 1971; 33: 98–110.
126. Hasegawa M, Kishino H & Yano T. Phylogenetic inference from DNA sequence data. K. Matusita, editor. *Statistical Theory and Data Analysis*. Amsterdam: North Holland; 1988.
127. Hayasaka S, Nichols T E. Combining voxel intensity and cluster extent with permutation test framework. *Neuroimage*. 2004; 23: 54–63.
128. Hettmansperger T P. *Statistical Inference Based on Ranks*. New York: Wiley; 1984.
129. Higgins J J, Noble W. A permutation test for a repeated measures design. *Applied Statistics* Agriculture. 1993; 5: 240–55.
130. Highton R. Comparison of microgeographic variation in morphological and electrophoretic traits. In Hecht MK, Steer WC, & B Wallace eds. *Evolutionary Biology* New York: Plenum; 1977; 10: 397–436.
131. Hinkley D V, Shi S. Importance sampling and the nested bootstrap. *Biometrika*. 1989; 76: 435–446.
132. Hisdal H, Stahl K, Tallaksen L M et al. Have streamflow droughts in Europe become more severe or frequent? *Int J Climatol*. 2001; 21: 317–321.
133. Hjorth J S U. *Computer Intensive Statistical Methods: Validation, Model Selection and Bootstrap*. New York: Chapman & Hall; 1994.
134. Hollander M, Pena E. Nonparametric tests under restricted treatment assignment rules. *JASA*. 1988; 83: 1144–1151.
135. Hollander M, Sethuraman J. Testing for agreement between two groups of judges. *Biometrika*. 1978; 65: 403–412.
136. Holmes M C, Williams R E O. The distribution of carriers of streptococcus pyrogenes among 2413 healthy children. *J. Hyg. Camd*. 1954; 52: 165–179.
137. Hosmer D W, Lemeshow S. Best subsets logistic regression. *Biometrics*. 1989.
138. Hossein-Zadeh G A, Ardekani B A & Soltanian-Zadeh H. Activation detection in fMRI using a maximum energy ratio statistic obtained by adaptive spatial filtering. *IEEE Trans. Med. Imaging*. 2003; 22: 795–805.
139. Howard M (pseud for Good P). Randomization in the analysis of experiments and clinical trials. *American Laboratory*. 1981; 13: 98–102.
140. Hubert L J. Combinatorial data analysis: Association and partial association. *Psychometrika*. 1985; 50: 449–467.
141. Hubert L J, Baker F B. Analyzing distinctive features confusion matrix. *J. Educ. Statist*. 1977; 2: 79–98.
142. Hubert L J, Baker F B. Evaluating the conformity of sociometric measurements. *Psychometrika*. 1978; 43: 31–42.
143. Hubert L J, Golledge R G & Costanzo C M. Analysis of variance procedures based on a proximity measure between subjects. *Psych. Bull*. 1982; 91: 424–30.
144. Hubert L J, Golledge R G, Costanzo C M, Gale N & Halperin W C. Nonparametric tests for directional data. In Bahrenberg G, Fischer M & P. Nijkamp, eds. *Recent Developments in Spatial Analysis: Methodology , Measurement, Models.* Aldershot UK: Gower; 1984: 171–190.
145. Hubert L J, Schultz J. Quadratic assignment as a general data analysis strategy. *Brit. J. Math. Stat. Psych*. 1976; 29: 190–241.
146. Ingenbleek J F. Tests simultanes de permutation des rangs pour bruit-blanc multivarie. *Statist. Anal Donnees*. 1981; 6: 60–65.
147. Jackson D A. Ratios in acquatic sciences: Statistical shortcomings with mean depth and the morphoedaphic index. *Canadian J Fisheries and Acquatic Sciences*. 1990; 47: 1788–1795.
148. Jin M Z. On the multisample pemutation test when the experimental units are nonuniform and random experimental errors exist. *J System Sci Math Sci.* 1984; 4: 117–127, 236–243.

149. Johns M V jr. Importance sampling for bootstrap confidence intervals. *JASA*. 1988; 83: 709–714.
150. Jones H L. Investigating the properties of a sample mean by employing random subsample means. *JASA*. 1956; 51: 54–83.
151. Karlin S, Ghandour G, Ost F, Tauare S & Korph K. New approaches for computer analysis of DNA sequences. *Proc. Nat. Acad. Sci*., USA. 1983; 80: 5660–5664.
152. Karlin S, Williams P T. Permutation methods for the structured exploratory data analysis (SEDA) of familial trait values. *Amer. J. Human Genetics*. 1984; 36: 873–898.
153. Kazdin A E. Statistical analysis for single-case experimental designs. In *Strategies for Studying Behavioral Change*. M Hersen and DH Barlow, editors. New York: Pergamon; 1976.
154. Kazdin A E. Obstacles in using randomization tests in single-case experiments. *J. Educ. Statist.* 1980; 5: 253–260.
155. Keller-McNulty S, Higgens J J. Effect of tail weight and outliers on power and type I error of robust permuatation tests for location. *Commun. Stat.—Theory and Methods*. 1987; 16: 17–35.
156. Kempthorne O. *Design and Analysis of Experiments*. New York: Wiley; 1952.
157. Kempthorne O. The randomization theory of experimental inference. *JASA*. 1955; 50: 946–967.
158. Kempthorne O. Some aspects of experimental inference. *JASA*. 1966; 61: 11–34.
159. Kempthorne O. Why randomize? *J. Statist. Prob. Infer*. 1977; 1: 1–26.
160. Klauber M R. Two-sample randomization tests for space-time clustering. *Biometrics*. 1971; 27: 129–142.
161. Klauber M R. Space-time clustering tests for more than two samples. *Biometrics*. 1975; 31: 719–726.
162. Klauber M R, Mustacchi A. Space-time clustering of childhood leukemia in San Francisco. *Cancer Res*. 1970; 30: 1969–1973.
163. Knight K. On the bootstrap of the sample mean in the infinite variance case. *Annal Statist.* 1989; 17: 1168–1173.
164. Koch G (ed). *Exchangeability in Probability and Statistics*. Amsterdam: North Holland; 1982.
165. Koziol J A, Maxwell D A, Fukushima M, Colmer A & Pilch Y H A. Distribution-free test for tumor-growth curve analyses with applications to an animal tumor immunotherapy experiment. *Biometrics*. 1981; 37: 383–90.
166. Krewski D, Brennan J & M Bickis. The power of the Fisher permutation test in 2 by k tables. *Commun. Stat. B*. 1984; 13: 433–448.
167. Kryscio R J, Meyers M H, Prusiner S I, Heise H W & B W Christine. The space-time distribution of Hodgkin's disease in Connecticut, 1940–1969. *J. Nat. Cancer Inst*. 1973; 50: 1107–1110.
168. Lachin J M. Properties of sample randomization in clinical trials. *Contr. Clin. Trials.* 1988; 9: 312–326.
169. Lachin J N. Statistical properties of randomization in clinical trials. *Contr. Clin. Trials.* 1988; 9: 289–311.
170. Lahiri S N. Bootstrapping the studentized sample mean of lattice variables. *J. Multiv. Anal.* 1993; 45: 247–256.
171. Laitenberger O, Atkinson C, Schlich M. et al. An experimental comparison of reading techniques for defect detection in UML design documents. *J. Syst. Software.* 2000; 53: 183–204.
172. Lehmann E L. *Testing Statistical Hypotheses*. 2nd ed. New York: John Wiley and Sons; 1986.

173. Levin D A. The organization of genetic variability in Phlox drummondi. *Evolution*. 1977; 31: 477–494.
174. Loh W-Y. Bootstrap calibration for confidence interval construction and selection. *Statist. Sinica.* 1991; 1: 479–495.
175. Loh W-Y. Calibrating confidence coefficients. *JASA*. 1987; 82: 155–162.
176. Loughin TM; Noble W. A permutation test for effects in an unreplicated factorial design. *Technometrics*. 1997; 39: 180–190.
177. Lunneborg C E. Estimating the correlation coefficient: The bootstrap approach. *Psychol. Bull*. 1985; 98: 209–215.
178. Makinodan T, Albright J W, Peter C P, Good P I & Hedrick M L. Reduced humoral activity in long-lived mice. *Immunology*. 1976; 31: 400–408.
179. Manly B F J. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. (2nd ed.). London: Chapman & Hall; 1997.
180. Mantel N. The detection of disease clustering and a generalized regression approach. *Cancer Res*. 1967; 27: 209–220.
181. Mantel N, Bailar J C. A class of permutational and multinomial tests arising in epidemiological research. *Biometrics*. 1970; 26: 687–700.
182. Mapleson W W. The use of GLIM and the bootstrap in assessing a clinical trial of two drugs. *Statist. Med*. 1986; 5: 363–374.
183. Marcus L F. Measurement of selection using distance statistics in prehistoric orangutan pongo pygamous palaeosumativens. *Evolution*. 1969; 23: 301.
184. Maritz J S. *Distribution Free Statistical Methods*. (2nd ed.) London: Chapman & Hall; 1996.
185. Marron J S. A comparison of cross-validation techniques in density estimation. *Ann. Statist*. 1987; 15: 152–162.
186. Martin M A. On bootstrap iteration for converge correction in confidence intervals. *JASA*. 1990; 85: 1105–1108.
187. Maxwell SE; Cole DA. A comparison of methods for increasing power in randomized between-subjects designs. *Psych Bull*. 1991; 110: 328–337.
188. McCarthy P J. Pseudo-replication: Half samples. *Review Int. Statist. Inst*. 1969; 37: 239–264.
189. McDonald L L, Davis B M & Miliken G A. A nonrandomized unconditional test for comparing two proportions in 2×2 contingency tables. *Technometrics*. 1977; 19: 145–158.
190. McKinney P W, Young M J, Hartz A & Bi-Fong Lee M. The inexact use of Fisher's exact test in six major medical journals. *J. American Medical Association*. 1989; 261: 3430–3433.
191. Mehta C R, Patel N R. A network algorithm for the exact treatment of the $2 \times K$ contingency table. *Commun. Statist. B.* 1980; 9: 649–664.
192. Mehta C R, Patel N R. A network algorithm for performing Fisher's exact test in rxc contingency tables. *JASA*. 1983; 78: 427–434.
193. Mehta C R, Patel N R & Gray R. On computing an exact confidence interval for the common odds ratio in several 2×2 contingency tables. *JASA*. 1985; 80: 969–973.
194. Mehta C R, Patel N R & Senchaudhuri P. Importance sampling for estimating exact probabilities in permutational inference. *JASA*. 1988; 83: 999–1005.
195. Merrington M, Spicer C C. Acute leukemia in New England. *Brit J Preventive and Social Medicine*. 1969; 23: 124–127.
196. Micceri T. The unicorn, the normal curve, and other improbable creatures. *Psychol. Bull*. 1989; 105: 156–166.
197. Mielke P W. Non-metric statistical analysis: Some metric alternatives. *J. Statist. Plan. Infer*. 1986; 13: 377–387.

198. Mielke P W Jr. The application of multivariate permutation methods based on distance functions in the earth sciences. *Earth-Science Rev.* 1991; 31: 55–71.
199. Mielke P W Jr, Berry K J. Fisher's exact probability test for cross-classification tables. *Educational and Psychological Measurement*. 1992; 52: 97–101.
200. Mielke P W Jr, Berry K J. Multivariate tests for correlated data in completely randomized designs. *J. Educ. Behav. Statist*. 1999; 24: 109–131.
201. Milano F, Maggi E, & del Turco M R. Evaluation of the effect of a quality control programme in mammography on technical and exposure parameters. *Radiat Prot Dosim*. 2000; 90: 263–266.
202. Mitchell-Olds T. Analysis of local variation in plant size. *Ecology*. 1987; 68: 82–7.
203. Mooney C Z, Duval R D. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Newbury Park CA: Sage Publications; 1993.
204. Morgan J N, Sonquist J A. Problems in the analysis of surveuy data and a proposal. *JASA*. 1963; 58: 415–434.
205. Mosteller F, Tukey J W. *Data Analysis and Regression: a second course in statistics*. Menlo Park: Addison-Wesley; 1977.
206. Mueller L D, Altenberg L. Statistical inference on measures of niche overlap. *Ecology*. 1985; 66: 1204–1210.
207. Nguyen T T. A generalization of Fisher's exact test in pxq contingency tables using more concordant relations. *Commun. Statist. B*. 1985; 14: 633–645.
208. Noreen E. *Computer Intensive Methods for Testing Hypotheses*. New York: John Wiley & Sons; 1989.
209. North B V, Curtis D, Cassell P G, Hitman G A & Sham P C. Assessing optimal neural network architecture for identifying disease-associated multi-marker genotypes using a permutation test, and application to calpain 10 polymorphisms associated with diabetes. *Ann. Hum. Genet*. 2003; 67: 348–56.
210. Nurminen M. Prognostic models for predicting delayed onset of renal allograft function. *Internet Journal of Epidemiology*. 2003; 1: 1.
211. Oden N L. Allocation of effort in Monte Carlo simulations for power of permutation tests. *JASA*. 1991; 86: 1074–76.
212. Passing H. Exact simultaneous comparisons with controls in an rxc contingency table. *Biometrical J*. 1984; 26: 643–654.
213. Patefield W M. Exact tests for trends in ordered contingency tables. *Appl. Statist*. 1982; 31: 32–43.
214. Patil C H K. Cochran's Q test: exact distribution. *JASA*. 1975; 70: 186–189.
215. Pearson E S. Some aspects of the problem of randomization. *Biometrika*. 1937; 29: 53–64.
216. Penninckx W, Hartmann C, Massart D L & Smeyersverbeke J. Validation of the calibration procedure in atomic absorption spectrometric methods. *J Analytical Atomic Spectrometry*. 1996; 11: 237–246.
217. Peritz E. Exact tests for matched pairs: studies with covariates. *Commun. Statist. A*. 1982; 11: 2157–2167 (errata 12: 1209–1210).
218. Peritz E. Modified Mantel-Haenszel procedures for matched pairs. *Commun. Statist. A*. 1985; 14: 2263–2285.
219. Perlich C, Provost F & Simonoff J S. Tree induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research.* 2003; 4: 211–255.
220. Pesarin F. On a nonparametric combination method for dependent permutation tests with applications. *Psychotherapy and Psychosomatics*. 1990; 54: 172–179.
221. Pesarin F. A nonparametric combination method for dependent permutation tests with application to some problems with repeated measures. In *Industrial Statistics*, Kitsos C P and Edler L, eds. Physics-Verlag: Heidelberg, 1997; 259–68.

222. Pesarin F. *Multivariate Permutation Tests*. NewYork: Wiley; 2001.
223. Petrondas D A, Gabriel R K. Multiple comparisons by rerandomization tests. *JASA*. 1983; 78: 949–957.
224. Pitman E J G. Significance tests which may be applied to samples from any population. *Roy. Statist. Soc. Suppl*. 1937; 4: 119–130, 225–232.
225. Pitman E J G. Significance tests which may be applied to samples from any population. Part III. The analysis of variance test. *Biometrika*. 1938; 29: 322–335.
226. Phipps M C. Small samples and the tilted bootstrap. *Theory of Stochastic Processes*. 1997; 19: 355–362.
227. Plackett R L, Hewlett P S. A unified theory of quantal responses to mixtures of drugs. The fitting to data of certain models for two non-interactive drugs with complete positive correlation of tolerances. *Biometrics*. 1963; 19: 517–531.
228. Poi B P. 2004. From the help desk: Some bootstrapping techniques. *Stata Journal* 4(3): 312–328.
229. Pollard E, Lackland K H & Rothrey P. The detection of density dependence from a series of annual censuses. *Ecology*. 1987; 68: 2046–2055.
230. Ponton D, Copp G H. Early dry-season community structure and habitat use of young fish in tributaries of the river Sinnamary (French-Guiana South-America) before and after hydrodam operation. *Environ. Biol. Fishes.* 1997; 50: 235–56.
231. Prager M H, Hoenig J M. Superposed epoch analysis: A randomization test of environmental effects on recruitment with application to chub mackerel. *Trans. Amer. Fisheries Soc*. 1989; 18: 608–619.
232. Praska Rao B L S. *Nonparametric Functional Estimation*. New York: Academic Press; 1983.
233. Priesendorfer R W, Barnett T P. Numerical model/reality intercomparison tests using small-sample statistics. *J. Atmospheric Sciences.* 1983; 40: 1884–96.
234. Puri M L, Sen P K. On a class of multivariate, multisample rank-order tests. *Sankyha Series A*. 1966; 28: 353–376.
235. Quinn, J F. On the statistical detection of cycles in extinctions in the marine fossil record. *Paleobiology*. 1987; 13: 465–78.
236. Ritland C, Ritland K. Variation of sex allocation among eight taxa of the Minimuls guttatus species complex (Scrophulariaceae). *Amer. J. Botany*. 1989; 76.
237. Rasmussen J. Estimating correlation coefficients: bootstrap and parametric approaches. *Psych. Bull*. 1987; 101: 136–139.
238. Raz J, Zheng H, Ombao H & Turetsky B. Statistical tests for fMRI based on experimental randomization. *Neuroimage*. 2003; 19: 226–32.
239. Romano J P. A bootstrap revival of some nonparametric distance tests. *JASA*. 1988; 83: 698–708.
240. Royaltey H H, Astrachen E & Sokal R R. Tests for patterns in geographic variation. *Geographic Analysis*. 1975; 7: 369–95.
241. Ryan T P. *Modern Regression Methods*. New York: John Wiley & Sons; 1997.
242. Ryan J M, Tracey T J G & Rounds J. Generalizability of Holland's structure of vocational interests across ethnicity, gender, and socioeconomic status. *J. Counseling Psych*. 1996; 43: 330–337.
243. Ryman N, Reuterwall C, Nygren K & Nygren T. Genetic variation and differentiation in Scandiavian moose (Alces Alces): Are large mammals monomorphic? *Evolution*. 1980; 34: 1037–1049.
244. Salapatek P, Kessen W. Visual scanning of triangles by the human newborn. *J Exper Child Psych*. 1966; 3: 155–167.

245. Scheffe H. *Analysis of Variance*. New York: John Wiley & Sons; 1959.
246. Schenker N. Qualms about Bootstrap confidence intervals. *JASA* 1985; 80: 360–361.
247. Schultz J R, Hubert L. A nonparametric test for the correspondence between two proximity matrices. *J. Educ. Statist*. 1976; 1: 59–67.
248. Selander R K, Kaufman D W. Genetic structure of populations of the brown snail (Helix aspersa). I Microgeographic variation. *Evolution*. 1975; 29: 385–401.
249. Shao J, Tu D. *The Jacknife and the Bootstrap*. New York: Springer; 1995.
250. Shen C D, Quade D. A randomization test for a three-period three-treatment crossover experiment. *Commun. Statist*. B. 1986; 12: 183–199.
251. Shimbukaro F I, Lazar S, Dyson H B & Chernick M R. A quasi-optical method for measuring the complex permittivity of materials. *IEEE Trans. Microwave Theor. Techn.* 1984; 32: 659–665.
252. Shuster J J. *Practical Handbook of Sample Size Guidelines for Clinical Trials*. Boca Raton FL: CRC Press; 1993.
253. Shuster J J, Boyett J M. Nonparametric multiple comparison procedures. *JASA*. 1979; 74: 379–82.
254. Siemiatycki J. Mantel's space-time clustering statistic: computing higher moments and a comparison of various data transforms. *J Statist. Comput. Simul*. 1978; 7: 13–31.
255. Siemiatycki J, McDonald A D. Neural tube defects in Quebec: A search for evidence of 'clustering' in time and space. *Brit. J. Prev. Soc. Med*. 1972; 26: 10–14.
256. Silverman B W. Using kernel density estimates to investigate multimodality. *JRSS* B. 1981; 43: 97–99.
257. Silverman B W, Young G A. The bootstrap: to smooth or not to smooth. *Biometrika*. 1987; 74: 469–479.
258. Simon J L. *Basic Research Methods in Social Science*. New York: Random House; 1969.
259. Singh K. On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist*. 1981; 9: 1187–1195.
260. Smith P W F, Forester J J & McDonald J W. Monte Carlo exact tests for square contingency tables. *J. Royal Statist. Soc. A*. 1996; 59: 309–21.
261. Smith P G, Pike M C. Generalization of two tests for the detection of household aggregation of disease. *Biometrics*. 1976; 32: 817–28.
262. Smythe R T. Conditional inference for restricted randomization designs. *Ann. Math. Statist*. 1988; 16: 1155–1161.
263. Sokal R R. Testing statistical significance in geographical variation patterns. *Systematic Zoo*. 1979; 28: 227–232.
264. Solow A R. A randomization test for misclassification problems in discriminatory analysis. *Ecology*. 1990; 71: 2379–2382.
265. Stine R. An introduction to bootstrap methods: examples and ideas. In *Modern Methods of Data Analysis*. J. Fox & J S Long, eds. Newbury Park CA: Sage Publications; 1990: 353–373.
266. Stone M. Cross-validation choice and assessment of statistical predictions. *JASA*. 1974; B36: 111–147.
267. Streitberg B, Roehmel J. On tests that are uniformly more powerful than the Wilcoxon-Mann-Whitney test. *Biometrics* 1990; 46: 481–484.
268. Syrjala S E. A statistical test for a difference between the spatial distributions of two populations. *Ecology*. 1996; 77: 75–80.
269. Thaler-Neto A, Fries R & Thaller G. Risk ratio as parameter for the genetic characterization of complex binary traits in cattle. A simulation study under various genetic models using halfsib families. *J. Anim. Breed.* Genetics. 2000; 117: 153–167.
270. Therneau T M, Grambsch P M. *Modeling Survival Data*. NY: Springer. 2000.

271. Thompson J R, Bridges E & Ensor K. Marketplace competition in the personal computer industry. *Decision Science*s. 1992: 467–477.
272. Tibshirani R J. Variance stabilization and the bootstrap. *Biometrika*. 1988; 75: 433–444.
273. Titterington D M, Murray G D, Spiegelhalter D J, Skene A M, Habbema J D F & Gelke G J. Comparison of discrimination techniques applied to a complex data set of head-injured patients. *JRSS A.* 1981; 144: 145–175.
274. Tracy D S, Khan K A. Comparison of some MRPP and standard rank tests for three equal sized samples. *Commun. Statist. B.* 1990; 19: 315–333.
275. Tracy D S, Tajuddin I H. Empirical power comparisons of two MRPP rank tests. *Commun. Statist. A.* 1986; 15: 551–570.
276. Troendle J F. A stepwise resampling method of multiple hypothesis testing. *JASA*. 1995; 90: 370–378.
277. Tsuji R. The structuring of trust relations in groups and the transition of within-group order. *Sociol Theor Method*. 2000; 15: 197–208.
278. Tsutakawa R K, Yang S L. Permutation tests applied to antibiotic drug resistance. *JASA*. 1974; 69: 87–92.
279. Tu D, Zhang L. Jacknife approximations for some nonparametreic confidence intervals of functional parameters based on normalizing transformations. *Compu. Statist*. 1992; 7: 3–5.
280. Tukey J W. Improving crucial randomized experiments—especially in weather modification—by double randomization and rank combination. In LeCam L & Binckly P, eds. *Proceeding of the Berkeley Conference in Honor of J Neyman and J Kiefer*. Hayward CA: Wadsworth; 1985; 1: 79–108.
281. Tukey J W, Brillinger D R & Jones L V. Management of Weather Resources: Vol II: The role of statistics in weather resources management. Washington DC: Department of Commerce, US Government Printing Office; 1978.
282. Valdes-Perez R E. Some recent human-computer studies in science and what accounts for them. *AI Magazine*. 1995; 16: 37–44.
283. Valdes-Perez R E, Pericliev V. Computer enumeration of significant implicational universals of kinship terminology. *Cross-Cult. Res.* 1999; 33: 162–174.
284. vanKeerberghen P, Vandenbosch C, Smeyers-Verbeke J & Massart D L. Some robust statistical procedures applied to the analysis of chemical data. *Chemometrics and Intelligent Laboratory Systems*. 1991; 12: 3–13.
285. Vanlier J B. Limitations of thermophilic anaerobic waste-water treatment and the consequences for process design. Antonie Van Leeuwenhoek Int. *J. General Molecular Microbiol*. 1996; 69: 1–14.
286. van-Putten B. On the construction of multivariate permutation tests in the multivariate two-sample case. *Statist. Neerlandica*. 1987; 41: 191–201.
287. Varga J, Toth B. Genetic variability and reproductive mode of Aspergillus fumigatus. *Infect. Genet. Evol.* 2003; 3: 3–17.
288. Wåhrendorf J, Brown C C. Bootstrapping a basic inequality in the analysis of the joint action of two drugs. *Biometrics*. 1980; 36: 653–657.
289. Wald A, Wolfowitz J. Statistical tests based on permutations of the observations. *Ann. Math. Statist*. 1944; 15: 358–372.
290. Westfall D H Young S S. *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*. New York: John Wiley; 1993.
291. Wei L J. Exact two-sample permutation tests based on the randomized play-the-winner rule. *Biometrika*. 1988; 75: 603–605.
292. Welch W J. Rerandomizing the median in matched-pairs designs. *Biometrika*. 1987; 74: 609–614.

293. Whaley F S. The equivalence of three individually derived permutation procedures for testing the homogenity of multidimensional samples. *Biometrics*. 1983; 39: 741–745.
294. Williams-Blangero S. Clan-structured migration and phenotypic differentiation in the Jirels of Nepal. *Human Biology*. 1989; 61: 143–157.
295. Wilk M B. The randomization analysis of a generalized randomized block design. *Biometrika*. 1955; 42: 70–79.
296. Wilk M B, Kempthorne O. Some aspects of the analysis of factorial experiments in a completely randomized design. *Ann. Math. Statist*. 1956; 27: 950–984.
297. Wilk M B, Kempthorne O. Nonadditivities in a Latin square design. *JASA*. 1957; 52: 218–236.
298. Witztum D, Rips E & Rosenberg Y. Equidistant letter sequences in the Book of Genesis. *Statist. Science*. 1994; 89: 768–76.
299. Wong M A, Lane T. A *k*th nearest neighbor clustering procedure, *JRSS* B. 1983; 45: 362–368.
300. Wong R K W, Chidambaram N & Mielke P W. Applications of multi-response permutation procedures and median regression for covariate analyses of possible weather modification effects on hail responses. *Atmosphere-Ocean. 1983*; 21: 1–13.
301. Wu C F J. Jackknife, bootstrap, and other resampling methods in regression analysis (with discuss.) *Annals Statist.* 1986; 14: 1261–1350.
302. Xu R Li X. A comparison of parametric versus permutation methods with applications to general and temporal microarray gene expression data. *Bioinformatics*. 2003; 19: 1284–9.
303. Young G A. Bootstrap: More than a stab in the dark. *Statist. Science*. 1994; 9: 382–415.
304. Zempo N, Kayama N, Kenagy R D, Lea H J & Clowes A W. Regulation of vascular smooth-muscle-cell migration and proliferation in vitro and in injured rat arteries by a synthetic matrix metalloprotinase inhibitor. *Art. Throm. V*. 1996; 16: 28–33.
305. Zerbe G O. Randomization analysis of the completely randomized design extended to growth and response curves. *JASA*. 1979; 74: 215–221.
306. Zerbe G O. Randomization analysis of randomized block design extended to growth and response curves. *Commun Statist. A*. 1979; 8: 191–205.
307. Zerbe G O, Murphy J R. On multiple comparisons in the randomization analysis of growth and response curves. *Biometrics*. 1986; 42: 795–804.
308. Zerbe G O, Walker S H. A randomization test for comparison of groups of growth curves with different polynomial design matricies. *Biometrics*. 1977; 33: 653–657.
309. Zucker S, Mazeh T. On the statistical significance of the Hipparcos astronomic orbit of ρ Coronae Borealis. *Mon. Not. R. Astron. Soc.* 2003.

Glossary

Accuracy – An *accurate* estimate is close to the estimated quantity.

Bootstrap Sample – A sample taken at random and with replacement from an existing sample rather than from the population at large.

Contingency Table – The entries in a contingency table are the frequencies with which specific events were observed. The marginals of the contingency table are the frequencies with which various categories of events were observed. These categories may be ordered or unordered.

Deterministic – A phenomenon is *deterministic* when its outcome is totally predictable.

Distribution-Free Methods – Require no assumptions about or knowledge of the distribution of the observations. Permutation tests, the nonparametric bootstrap, and nonparametric decision trees are all distribution-free methods.

Empirical Distribution Function – The distribution function of the observations as distinguished from the distribution function of the population itself.

Exchangeable Observations – Their joint probability distribution remains unchanged by rearrangements of their labels.

Experimental Design – A division of an experiment into blocks based on the values of one or more factors. These values may be ordinal or categorical.

Functional – Any numeric characteristic of a population such as a percentile, a mean, a standard deviation, or a combination thereof.

Nonparametric Tests – A misnomer as many such tests concern parameters. Distribution-free test is what is usually intended by this expression.

Parametric Methods – Take advantage of our knowledge of the distribution of the observations. Rejection regions and confidence intervals of parametric procedures are based on this knowledge. If our assumptions concerning the distribution of the observations are in error then the corresponding parametric methods are in error.

Permutation Distribution – Should be called a rearrangement distribution as it consists of the values taken by a test statistic for each of the possible relabelings of a set of observations.

Permutation Test – The significance levels of a permutation test are established by reference to the permutation distribution of a test statistic.

p-value – A function of the sample and the sample statistic. Thus, it will vary from sample to sample.

Power of a Test – Defined as the probability of rejecting the hypothesis when a specific alternative is true. Thus the power is 1 minus the probability of making a Type II error.

Precision – Precise estimates take almost the same value from sample to sample. A precise interval estimate is a narrow one.

Rank Test – A permutation test in which the ranks of the observations rather than their original values are used.

Resampling Method – Any estimation, hypothesis testing, or modelling method that requires repeated resampling from the data at hand. Bootstrap, decision trees, and permutation tests are all resampling methods.

Significance Level – Defined as the probability of making a Type I error.

Stochastic Outcomes – May take any of a distribution of values. A "random fluctuation" is stochastic.

Type I Error – Made when we accept the alternative hypothesis and the primary hypothesis is true.

Type II Error – Made when we accept the primary hypothesis, yet the null hypothesis is true.

Unbiased Confidence Interval – A confidence interval that has a greater probability of containing the true value of a parameter than of any false value.

Unbiased Test – A hypothesis test that is more likely to reject any false hypothesis than any true one (if, that is, all the assumptions on which the test is based are satisfied).

Author Index

Subject Index