Design and Analysis of Cross-Over Trials *B. Jones and M.G. Kenward* (1989)
35 Empirical Bayes Method, 2nd edition *J.S. Maritz and T. Lwin* (1989)
Symmetric Multivariate and Related Distributions *K.-T. Fang, S. Kotz and K. Ng* (1989)
Generalized Linear Models, 2nd edition *P. McCullagh and J.A. Nelder* (1989)
38 Cyclic Designs *J.A. John* (1987)
39 Analog Estimation Methods in Econometrics *C.F. Manski* (1988)
40 Subset Selection in Regression *A.J. Miller* (1990)
41 Analysis of Repeated Measures *M. Crowder and D.J. Hand* (1990)
42 Statistical Reasoning with Imprecise Probabilities *P. Walley* (1990)
43 Generalized Additive Models *T.J. Hastie and R.J. Tibshirani* (1990)
Inspection Errors for Attributes in Quality Control *N.L. Johnson, S. Kotz and X. Wu* (1991)
45 The Analysis of Contingency Tables, 2nd edition *B.S. Everitt* (1992)
46 The Analysis of Quantal Response Data *B.J.T. Morgan* (1992)
47 Longitudinal Data with Serial Correlation: A State-Space Approach *R.H. Jones* (1993)
Differential Geometry and Statistics *M.K. Murray and J.W. Rice* (1993)
49 Markov Models and Optimization *M.H.A. Davies* (1993)
50 Chaos and Networks: Statistical and Probabilistic Aspects *Edited by O. Barndorff-Nielsen et al.* (1993)
Number Theoretic Methods in Statistics *K.-T. Fang and W. Yuan* (1993)
52 Inference and Asymptotics *O. Barndorff-Nielsen and D.R. Cox* (1993)
53 Practical Risk Theory for Actuaries *C.D. Daykin, T. Pentikainen and M. Pesonen* (1993)
54 Statistical Concepts and Applications in Medicine *J. Aitchison and I.J. Lauder* (1994)
55 Predictive Inference *S. Geisser* (1993)
56 Model-Free Curve Estimation *M. Tarter and M. Lock* (1993)
57 An Introduction to the Bootstrap *B. Efron and R. Tibshirani* (1993)
(Full details concerning this series are available from the Publishers.)

An Introduction to the Bootstrap

Bradley Efron
Department of Statistics
Stanford University
and
Robert J. Tibshirani
Department of Preventative Medicine and Biostatistics
and Department of Statistics, University of Toronto

CHAPMAN & HALL/CRC
Boca Raton London New York Washington, D.C.

Library of Congress Cataloging-in-Publication Data

Efron, Bradley.
An introduction to the bootstrap/Brad Efron, Rob Tibshirani.
p. cm.
Includes bibliographical references.
ISBN 0-412-04231-2
1. Bootstrap (Statistics).I. Tibshirani, Robert. II. Title.
QA276.8.E3745 1993
519.5'44—dc20 93-4489
CIP

First CRC Press reprint 1998
Originally published by Chapman & Hall

International Standard Book Number 0-412-04231-2
Library of Congress Card Number 93-4489
Printed in the United States of America 1 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

TO

CHERYL, CHARLIE, RYAN AND JULIE

AND TO THE MEMORY OF

RUPERT G. MILLER, JR.

supported the development of statistical theory at Stanford, including much of the theory behind this book. The second author would like to thank his wife Cheryl for her understanding and support during this entire project, and his parents for a lifetime of encouragement. He gratefully acknowledges the support of the Natural Sciences and Engineering Research Council of Canada.

Palo Alto and Toronto Bradley Efron
June 1993 Robert Tibshirani

CHAPTER 1

Introduction

Statistics is the science of learning from experience, especially experience that arrives a little bit at a time. The earliest information science was statistics, originating in about 1650. This century has seen statistical techniques become the analytic methods of choice in biomedical science, psychology, education, economics, communications theory, sociology, genetic studies, epidemiology, and other areas. Recently, traditional sciences like geology, physics, and astronomy have begun to make increasing use of statistical methods as they focus on areas that demand informational efficiency, such as the study of rare and exotic particles or extremely distant galaxies.

Most people are not natural-born statisticians. Left to our own devices we are not very good at picking out patterns from a sea of noisy data. To put it another way, we are all too good at picking out non-existent patterns that happen to suit our purposes. Statistical theory attacks the problem from both ends. It provides optimal methods for finding a real signal in a noisy background, and also provides strict checks against the overinterpretation of random patterns.

Statistical theory attempts to answer three basic questions:

(1) How should I collect my data?

(2) How should I analyze and summarize the data that I've collected?

(3) How accurate are my data summaries?

Question 3 constitutes part of the process known as statistical inference. The bootstrap is a recently developed technique for making certain kinds of statistical inferences. It is only recently developed because it requires modern computer power to simplify the often intricate calculations of traditional statistical theory.

The explanations that we will give for the bootstrap, and other

computer-based methods, involve explanations of traditional ideas in statistical inference. The basic ideas of statistics haven't changed, but their implementation has. The modern computer lets us apply these ideas flexibly, quickly, easily, and with a minimum of mathematical assumptions. Our primary purpose in the book is to explain when and why bootstrap methods work, and how they can be applied in a wide variety of real data-analytic situations.

All three basic statistical concepts, data collection, summary and inference, are illustrated in the New York Times excerpt of Figure 1.1. A study was done to see if small aspirin doses would prevent heart attacks in healthy middle-aged men. The data for the aspirin study were collected in a particularly efficient way: by a controlled, randomized, double-blind study. One half of the subjects received aspirin and the other half received a control substance, or placebo, with no active ingredients. The subjects were randomly assigned to the aspirin or placebo groups. Both the subjects and the supervising physicians were blinded to the assignments, with the statisticians keeping a secret code of who received which substance. Scientists, like everyone else, want the project they are working on to succeed. The elaborate precautions of a controlled, randomized, blinded experiment guard against seeing benefits that don't exist, while maximizing the chance of detecting a genuine positive effect.

The summary statistics in the newspaper article are very simple:

	heart attacks (fatal plus non-fatal)	subjects
aspirin group:	104	11037
placebo group:	189	11034

We will see examples of much more complicated summaries in later chapters. One advantage of using a good experimental design is a simplification of its results. What strikes the eye here is the lower rate of heart attacks in the aspirin group. The ratio of the two rates is

$$\widehat{\theta} = \frac{104/11037}{189/11034} = .55. \tag{1.1}$$

If this study can be believed, and its solid design makes it very believable, the aspirin-takers only have 55% as many heart attacks as placebo-takers.

Of course we are not really interested in $\widehat{\theta}$, the estimated ratio. What we would like to know is θ, the true ratio, that is the ratio

HEART ATTACK RISK FOUND TO BE CUT BY TAKING ASPIRIN

LIFESAVING EFFECTS SEEN

Study Finds Benefit of Tablet Every Other Day Is Much Greater Than Expected

By HAROLD M. SCHMECK Jr.

A major nationwide study shows that a single aspirin tablet every other day can sharply reduce a man's risk of heart attack and death from heart attack.

The lifesaving effects were so dramatic that the study was halted in mid-December so that the results could be reported as soon as possible to the participants and to the medical profession in general.

The magnitude of the beneficial effect was far greater than expected, Dr. Charles H. Hennekens of Harvard, principal investigator in the research, said in a telephone interview. The risk of myocardial infarction, the technical name for heart attack, was cut almost in half.

'Extreme Beneficial Effect'

A special report said the results showed "a statistically extreme beneficial effect" from the use of aspirin. The report is to be published Thursday in The New England Journal of Medicine.

In recent years smaller studies have demonstrated that a person who has had one heart attack can reduce the risk of a second by taking aspirin, but there had been no proof that the beneficial effect would extend to the general male population.

Dr. Claude Lenfant, the director of the National Heart Lung and Blood Institute, said the findings were "extremely important," but he said the general public should not take the report as an indication that everyone should start taking aspirin.

Figure 1.1. *Front-page news from the New York Times of January 27, 1987. Reproduced by permission of the New York Times.*

we would see if we could treat all subjects, and not just a sample of them. The value $\widehat{\theta} = .55$ is only an estimate of θ. The sample seems large here, 22071 subjects in all, but the conclusion that aspirin works is really based on a smaller number, the 293 observed heart attacks. How do we know that $\widehat{\theta}$ might not come out much less favorably if the experiment were run again?

This is where statistical inference comes in. Statistical theory allows us to make the following inference: the true value of θ lies in the interval

$$.43 < \theta < .70 \tag{1.2}$$

with 95% confidence. Statement (1.2) is a classical confidence interval, of the type discussed in Chapters 12–14, and 22. It says that if we ran a much bigger experiment, with millions of subjects, the ratio of rates probably wouldn't be too much different than (1.1). We almost certainly wouldn't decide that θ exceeded 1, that is that aspirin was actually harmful. It is really rather amazing that the same data that give us an estimated value, $\widehat{\theta} = .55$ in this case, also can give us a good idea of the estimate's accuracy.

Statistical inference is serious business. A lot can ride on the decision of whether or not an observed effect is real. The aspirin study tracked strokes as well as heart attacks, with the following results:

	strokes	subjects
aspirin group:	119	11037
placebo group:	98	11034

(1.3)

For strokes, the ratio of rates is

$$\widehat{\theta} = \frac{119/11037}{98/11034} = 1.21. \tag{1.4}$$

It now looks like taking aspirin is actually harmful. However the interval for the true stroke ratio θ turns out to be

$$.93 < \theta < 1.59 \tag{1.5}$$

with 95% confidence. This includes the neutral value $\theta = 1$, at which aspirin would be no better or worse than placebo vis-à-vis strokes. In the language of statistical hypothesis testing, aspirin was found to be significantly beneficial for preventing heart attacks, but not significantly harmful for causing strokes. The opposite conclusion had been reached in an older, smaller study concerning men

see bs_ex1.m

who had experienced previous heart attacks. The aspirin treatment remains mildly controversial for such patients.

The bootstrap is a data-based simulation method for statistical inference, which can be used to produce inferences like (1.2) and (1.5). The use of the term bootstrap derives from the phrase *to pull oneself up by one's bootstrap*, widely thought to be based on one of the eighteenth century Adventures of Baron Munchausen, by Rudolph Erich Raspe. (The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps.) It is not the same as the term "bootstrap" used in computer science meaning to "boot" a computer from a set of core instructions, though the derivation is similar.

Here is how the bootstrap works in the stroke example. We create two populations: the first consisting of 119 ones and 11037-119=10918 zeroes, and the second consisting of 98 ones and 11034-98=10936 zeroes. We draw with replacement a sample of 11037 items from the first population, and a sample of 11034 items from the second population. Each of these is called a *bootstrap sample*. From these we derive the bootstrap replicate of $\hat{\theta}$:

$$\hat{\theta}^* = \frac{\text{Proportion of ones in bootstrap sample \#1}}{\text{Proportion of ones in bootstrap sample \#2}}. \tag{1.6}$$

We repeat this process a large number of times, say 1000 times, and obtain 1000 *bootstrap replicates* $\hat{\theta}^*$. This process is easy to implement on a computer, as we will see later. These 1000 replicates contain information that can be used to make inferences from our data. For example, the standard deviation turned out to be 0.17 in a batch of 1000 replicates that we generated. The value 0.17 is an estimate of the standard error of the ratio of rates $\hat{\theta}$. This indicates that the observed ratio $\hat{\theta} = 1.21$ is only a little more than one standard error larger than 1, and so the neutral value $\theta = 1$ cannot be ruled out. A rough 95% confidence interval like (1.5) can be derived by taking the 25th and 975th largest of the 1000 replicates, which in this case turned out to be (.93, 1.60).

In this simple example, the confidence interval derived from the bootstrap agrees very closely with the one derived from statistical theory. Bootstrap methods are intended to simplify the calculation of inferences like (1.2) and (1.5), producing them in an automatic way even in situations much more complicated than the aspirin study.

The terminology of statistical summaries and inferences, like regression, correlation, analysis of variance, discriminant analysis, standard error, significance level and confidence interval, has become the lingua franca of all disciplines that deal with noisy data. We will be examining what this language means and how it works in practice. The particular goal of bootstrap theory is a computer-based implementation of basic statistical concepts. In some ways it is easier to understand these concepts in computer-based contexts than through traditional mathematical exposition.

1.1 An overview of this book

This book describes the bootstrap and other methods for assessing statistical accuracy. The bootstrap does not work in isolation but rather is applied to a wide variety of statistical procedures. Part of the objective of this book is expose the reader to many exciting and useful statistical techniques through real-data examples. Some of the techniques described include nonparametric regression, density estimation, classification trees, and least median of squares regression.

Here is a chapter-by-chapter synopsis of the book. **Chapter 2** introduces the bootstrap estimate of standard error for a simple mean. **Chapters 3–5** contain some basic background material, and may be skimmed by readers eager to get to the details of the bootstrap in **Chapter 6**. Random samples, populations, and basic probability theory are reviewed in **Chapter 3**. **Chapter 4** defines the empirical distribution function estimate of the population, which simply estimates the probability of each of n data items to be $1/n$. **Chapter 4** also shows that many familiar statistics can be viewed as "plug-in" estimates, that is, estimates obtained by plugging in the empirical distribution function for the unknown distribution of the population. **Chapter 5** reviews standard error estimation for a mean, and shows how the usual textbook formula can be derived as a simple plug-in estimate.

The bootstrap is defined in **Chapter 6**, for estimating the standard error of a statistic from a single sample. The bootstrap standard error estimate is a plug-in estimate that rarely can be computed exactly; instead a simulation ("resampling") method is used for approximating it.

Chapter 7 describes the application of bootstrap standard errors in two complicated examples: a principal components analysis and a curve fitting problem.

Up to this point, only one-sample data problems have been discussed. The application of the bootstrap to more complicated data structures is discussed in **Chapter 8**. A two-sample problem and a time-series analysis are described.

Regression analysis and the bootstrap are discussed and illustrated in **Chapter 9**. The bootstrap estimate of standard error is applied in a number of different ways and the results are discussed in two examples.

The use of the bootstrap for estimation of bias is the topic of **Chapter 10**, and the pros and cons of bias correction are discussed. **Chapter 11** describes the jackknife method in some detail. We see that the jackknife is a simple closed-form approximation to the bootstrap, in the context of standard error and bias estimation.

The use of the bootstrap for construction of confidence intervals is described in **Chapters 12, 13** and **14**. There are a number of different approaches to this important topic and we devote quite a bit of space to them. In **Chapter 12** we discuss the bootstrap-t approach, which generalizes the usual Student's t method for constructing confidence intervals. The percentile method (**Chapter 13**) uses instead the percentiles of the bootstrap distribution to define confidence limits. The BC_a (bias-corrected accelerated interval) makes important corrections to the percentile interval and is described in **Chapter 14**.

Chapter 15 covers permutation tests, a time-honored and useful set of tools for hypothesis testing. Their close relationship with the bootstrap is discussed; **Chapter 16** shows how the bootstrap can be used in more general hypothesis testing problems.

Prediction error estimation arises in regression and classification problems, and we describe some approaches for it in **Chapter 17**. Cross-validation and bootstrap methods are described and illustrated. Extending this idea, **Chapter 18** shows how the bootstrap and cross-validation can be used to adapt estimators to a set of data.

Like any statistic, bootstrap estimates are random variables and so have inherent error associated with them. When using the bootstrap for making inferences, it is important to get an idea of the magnitude of this error. In **Chapter 19** we discuss the jackknife-after-bootstrap method for estimating the standard error of a bootstrap quantity.

Chapters 20–25 contain more advanced material on selected

topics, and delve more deeply into some of the material introduced in the previous chapters. The relationship between the bootstrap and jackknife is studied via the "resampling picture" in **Chapter 20. Chapter 21** gives an overview of non-parametric and parametric inference, and relates the bootstrap to a number of other techniques for estimating standard errors. These include the delta method, Fisher information, infinitesimal jackknife, and the sandwich estimator.

Some advanced topics in bootstrap confidence intervals are discussed in **Chapter 22**, providing some of the underlying basis for the techniques introduced in Chapters 12–14. **Chapter 23** describes methods for efficient computation of bootstrap estimates including control variates and importance sampling. In **Chapter 24** the construction of approximate likelihoods is discussed. The bootstrap and other related methods are used to construct a "non-parametric" likelihood in situations where a parametric model is not specified.

Chapter 25 describes in detail a bioequivalence study in which the bootstrap is used to estimate power and sample size. In **Chapter 26** we discuss some general issues concerning the bootstrap and its role in statistical inference.

Finally, the **Appendix** contains a description of a number of different computer programs for the methods discussed in this book.

1.2 Information for instructors

We envision that this book can provide the basis for (at least) two different one semester courses. An upper-year undergraduate or first-year graduate course could be taught from some or all of the first 19 chapters, possibly covering Chapter 25 as well (both authors have done this). In addition, a more advanced graduate course could be taught from a selection of Chapters 6–19, and a selection of Chapters 20–26. For an advanced course, supplementary material might be used, such as Peter Hall's book *The Bootstrap and Edgeworth Expansion* or journal papers on selected technical topics. The Bibliographic notes in the book contain many suggestions for background reading.

We have provided numerous exercises at the end of each chapter. Some of these involve computing, since it is important for the student to get hands-on experience for learning the material. The bootstrap is most effectively used in a high-level language for data

analysis and graphics. Our language of choice (at present) is "S" (or "S-PLUS"), and a number of S programs appear in the Appendix. Most of these programs could be easily translated into other languages such as Gauss, Lisp-Stat, or Matlab. Details on the availability of S and S-PLUS are given in the Appendix.

1.3 Some of the notation used in the book

Lower case bold letters such as $\mathbf{x}$ refer to vectors, that is, $\mathbf{x} = (x_1, x_2, \ldots x_n)$. Matrices are denoted by upper case bold letters such as $\mathbf{X}$, while a plain uppercase letter like X refers to a random variable. The transpose of a vector is written as $\mathbf{x}^T$. A superscript "*" indicates a bootstrap random variable: for example, $\mathbf{x}^*$ indicates a bootstrap data set generated from a data set $\mathbf{x}$. Parameters are denoted by Greek letters such as θ. A hat on a letter indicates an estimate, such as $\hat{\theta}$. The letters F and G refer to populations. In Chapter 21 the same symbols are used for the cumulative distribution function of a population. I_C is the indicator function equal to 1 if condition C is true and 0 otherwise. For example, $I_{\{x<2\}} = 1$ if $x < 2$ and 0 otherwise. The notation $\mathrm{tr}(A)$ refers to the trace of the matrix A, that is, the sum of the diagonal elements. The derivatives of a function $g(x)$ are denoted by $g'(x), g''(x)$ and so on.

The notation

$$F \rightarrow (x_1, x_2, \ldots x_n)$$

indicates an independent and identically distributed sample drawn from F. Equivalently, we also write $x_i \overset{\text{i.i.d.}}{\sim} F$ for $i = 1, 2, \ldots n$.

Notation such as $\#\{x_i > 3\}$ means the number of x_is greater than 3. $\log x$ refers to the natural logarithm of x.

CHAPTER 2

The accuracy of a sample mean

The bootstrap is a computer-based method for assigning measures of accuracy to statistical estimates. The basic idea behind the bootstrap is very simple, and goes back at least two centuries. After reviewing some background material, this book describes the bootstrap method, its implementation on the computer, and its application to some real data analysis problems. First though, this chapter focuses on the one example of a statistical estimator where we really don't need a computer to assess accuracy: the sample mean. In addition to previewing the bootstrap, this gives us a chance to review some fundamental ideas from elementary statistics. We begin with a simple example concerning means and their estimated accuracies.

Table 2.1 shows the results of a small experiment, in which 7 out of 16 mice were randomly selected to receive a new medical treatment, while the remaining 9 were assigned to the non-treatment (control) group. The treatment was intended to prolong survival after a test surgery. The table shows the survival time following surgery, in days, for all 16 mice.

Did the treatment prolong survival? A comparison of the means for the two groups offers preliminary grounds for optimism. Let $x_1, x_2, \cdots, x_7$ indicate the lifetimes in the treatment group, so $x_1 = 94, x_2 = 197, \cdots, x_7 = 23$, and likewise let $y_1, y_2, \cdots, y_9$ indicate the control group lifetimes. The group means are

$$\bar{x} = \sum_{i=1}^{7} x_i/7 = 86.86 \quad \text{and} \quad \bar{y} = \sum_{i=1}^{9} y_i/9 = 56.22, \tag{2.1}$$

so the difference $\bar{x} - \bar{y}$ equals 30.63, suggesting a considerable life-prolonging effect for the treatment.

But how accurate are these estimates? After all, the means (2.1) are based on small samples, only 7 and 9 mice, respectively. In

Table 2.1. *The mouse data. Sixteen mice were randomly assigned to a treatment group or a control group. Shown are their survival times, in days, following a test surgery. Did the treatment prolong survival?*

Group	Data			(Sample Size)	Mean	Estimated Standard Error
Treatment:	94	197	16			
	38	99	141			
	23			(7)	86.86	25.24
Control:	52	104	146			
	10	51	30			
	40	27	46	(9)	56.22	14.14
				Difference:	30.63	28.93

order to answer this question, we need an estimate of the accuracy of the sample means $\bar{x}$ and $\bar{y}$. For sample means, and essentially *only* for sample means, an accuracy formula is easy to obtain.

The *estimated standard error* of a mean $\bar{x}$ based on n independent data points $x_1, x_2, \cdots, x_n$, $\bar{x} = \sum_{i=1}^{n} x_i/n$, is given by the formula

$$\sqrt{\frac{s^2}{n}} \tag{2.2}$$

where $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$. (This formula, and standard errors in general, are discussed more carefully in Chapter 5.) The standard error of any estimator is defined to be the square root of its variance, that is, the estimator's root mean square variability around its expectation. This is the most common measure of an estimator's accuracy. Roughly speaking, an estimator will be less than one standard error away from its expectation about 68% of the time, and less than two standard errors away about 95% of the time.

If the estimated standard errors in the mouse experiment were very small, say less than 1, then we would know that $\bar{x}$ and $\bar{y}$ were close to their expected values, and that the observed difference of 30.63 was probably a good estimate of the true survival-prolonging

capability of the treatment. On the other hand, if formula (2.2) gave big estimated standard errors, say 50, then the difference estimate would be too inaccurate to depend on.

The actual situation is shown at the right of Table 2.1. The estimated standard errors, calculated from (2.2), are 25.24 for $\bar{x}$ and 14.14 for $\bar{y}$. The standard error for the difference $\bar{x} - \bar{y}$ equals $28.93 = \sqrt{25.24^2 + 14.14^2}$ (since the variance of the difference of two independent quantities is the sum of their variances). We see that the observed difference 30.63 is only $30.63/28.93 = 1.05$ estimated standard errors greater than zero. Readers familiar with hypothesis testing theory will recognize this as an *insignificant* result, one that could easily arise by chance even if the treatment really had no effect at all.

There are more precise ways to verify this disappointing result, (e.g. the permutation test of Chapter 15), but usually, as in this case, estimated standard errors are an excellent first step toward thinking critically about statistical estimates. Unfortunately standard errors have a major disadvantage: for most statistical estimators other than the mean there is no formula like (2.2) to provide estimated standard errors. In other words, it is hard to assess the accuracy of an estimate other than the mean.

Suppose for example, we want to compare the two groups in Table 2.1 by their medians rather than their means. The two medians are 94 for treatment and 46 for control, giving an estimated difference of 48, considerably more than the difference of the means. But how accurate are these medians? Answering such questions is where the bootstrap, and other computer-based techniques, come in. The remainder of this chapter gives a brief preview of the bootstrap estimate of standard error, a method which will be fully discussed in succeeding chapters.

Suppose we observe independent data points $x_1, x_2, \cdots, x_n$, for convenience denoted by the vector $\mathbf{x} = (x_1, x_2, \cdots, x_n)$, from which we compute a statistic of interest $s(\mathbf{x})$. For example the data might be the $n = 9$ control group observations in Table 2.1, and $s(\mathbf{x})$ might be the sample mean.

The bootstrap estimate of standard error, invented by Efron in 1979, looks completely different than (2.2), but in fact it is closely related, as we shall see. A *bootstrap sample* $\mathbf{x}^* = (x_1^*, x_2^*, \cdots, x_n^*)$ is obtained by randomly sampling n times, with replacement, from the original data points $x_1, x_2, \cdots, x_n$. For instance, with $n = 7$ we might obtain $\mathbf{x}^* = (x_5, x_7, x_5, x_4, x_7, x_3, x_1)$.

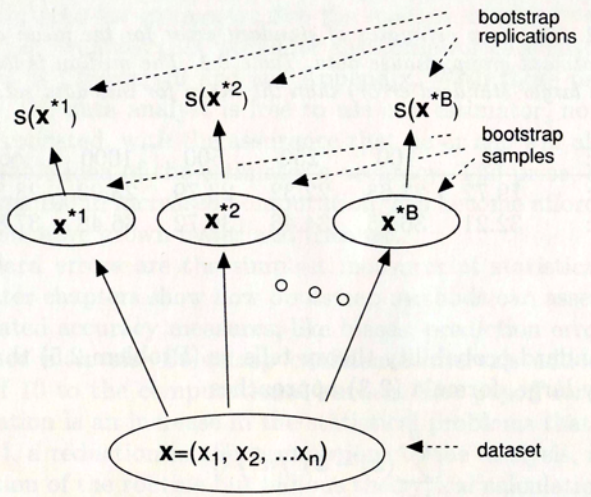

Figure 2.1. *Schematic of the bootstrap process for estimating the standard error of a statistic $s(\mathbf{x})$. B bootstrap samples are generated from the original data set. Each bootstrap sample has n elements, generated by sampling with replacement n times from the original data set. Bootstrap replicates $s(\mathbf{x}^{*1}), s(\mathbf{x}^{*2}), \ldots s(\mathbf{x}^{*B})$ are obtained by calculating the value of the statistic $s(\mathbf{x})$ on each bootstrap sample. Finally, the standard deviation of the values $s(\mathbf{x}^{*1}), s(\mathbf{x}^{*2}), \ldots s(\mathbf{x}^{*B})$ is our estimate of the standard error of $s(\mathbf{x})$.*

Figure 2.1 is a schematic of the bootstrap process. The bootstrap algorithm begins by generating a large number of independent bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \cdots, \mathbf{x}^{*B}$, each of size n. Typical values for B, the number of bootstrap samples, range from 50 to 200 for standard error estimation. Corresponding to each bootstrap sample is a *bootstrap replication* of s, namely $s(\mathbf{x}^{*b})$, the value of the statistic s evaluated for $\mathbf{x}^{*b}$. If $s(\mathbf{x})$ is the sample median, for instance, then $s(\mathbf{x}^*)$ is the median of the bootstrap sample. The bootstrap estimate of standard error is the standard deviation of the bootstrap replications,

$$\widehat{\text{se}}_{\text{boot}} = \left\{ \sum_{b=1}^{B} [s(\mathbf{x}^{*b}) - s(\cdot)]^2 / (B-1) \right\}^{\frac{1}{2}}, \tag{2.3}$$

where $s(\cdot) = \sum_{b=1}^{B} s(\mathbf{x}^{*b})/B$. Suppose $s(\mathbf{x})$ is the mean $\bar{x}$. In this

Table 2.2. *Bootstrap estimates of standard error for the mean and median; treatment group, mouse data, Table 2.1. The median is less accurate (has larger standard error) than the mean for this data set.*

B:	50	100	250	500	1000	∞
mean:	19.72	23.63	22.32	23.79	23.02	23.36
median:	32.21	36.35	34.46	36.72	36.48	37.83

case, standard probability theory tells us (Problem 2.5) that as B gets very large, formula (2.3) approaches

$$\{\sum_{i=1}^{n}(x_i - \bar{x})^2/n^2\}^{\frac{1}{2}}. \tag{2.4}$$

This is almost the same as formula (2.2). We could make it exactly the same by multiplying definition (2.3) by the factor $[n/(n-1)]^{\frac{1}{2}}$, but there is no real advantage in doing so.

Table 2.2 shows bootstrap estimated standard errors for the mean and the median, for the treatment group mouse data of Table 2.1. The estimated standard errors settle down to limiting values as the number of bootstrap samples B increases. The limiting value 23.36 for the mean is obtained from (2.4). The formula for the limiting value 37.83 for the standard error of the median is quite complicated: see Problem 2.4 for a derivation.

We are now in a position to assess the precision of the difference in medians between the two groups. The bootstrap procedure described above was applied to the control group, producing a standard error estimate of 11.54 based on $B = 100$ replications ($B = \infty$ gave 9.73). Therefore, using $B = 100$, the observed difference of 48 has an estimated standard error of $\sqrt{36.35^2 + 11.54^2} = 38.14$, and hence is $48/38.14 = 1.26$ standard errors greater than zero. This is larger than the observed difference in means, but is still insignificant.

For most statistics we don't have a formula for the limiting value of the standard error, but in fact no formula is needed. Instead we use the numerical output of the bootstrap program, for some convenient value of B. We will see in Chapters 6 and 19, that B in the range 50 to 200 usually makes $\widehat{\text{se}}_{\text{boot}}$ a good standard error

estimator, even for estimators like the median. It is easy to write a bootstrap program that works for any computable statistic $s(\mathbf{x})$, as shown in Chapters 6 and the Appendix. With these programs in place, the data analyst is free to use any estimator, no matter how complicated, with the assurance that he or she will also have a reasonable idea of the estimator's accuracy. The price, a factor of perhaps 100 in increased computation, has become affordable as computers have grown faster and cheaper.

Standard errors are the simplest measures of statistical accuracy. Later chapters show how bootstrap methods can assess more complicated accuracy measures, like biases, prediction errors, and confidence intervals. Bootstrap confidence intervals add another factor of 10 to the computational burden. The payoff for all this computation is an increase in the statistical problems that can be analyzed, a reduction in the assumptions of the analysis, and the elimination of the routine but tedious theoretical calculations usually associated with accuracy assessment.

2.1 Problems

2.1† Suppose that the mouse survival times were expressed in weeks instead of days, so that the entries in Table 2.1 were all divided by 7.

(a) What effect would this have on $\bar{x}$ and on its estimated standard error (2.2)? Why does this make sense?

(b) What effect would this have on the ratio of the difference $\bar{x} - \bar{y}$ to its estimated standard error?

2.2 Imagine the treatment group in Table 2.1 consisted of R repetitions of the data actually shown, where R is a positive integer. That is, the treatment data consisted of R 94's, R 197's, etc. What effect would this have on the estimated standard error (2.2)?

2.3 It is usually true that the error of a statistical estimator decreases at a rate of about 1 over the square root of the sample size. Does this agree with the result of Problem 2.2?

2.4 Let $x_{(1)} < x_{(2)} < x_{(3)} < x_{(4)} < x_{(5)} < x_{(6)} < x_{(7)}$ be an ordered sample of size $n = 7$. Let $\mathbf{x}^*$ be a bootstrap sample, and $s(\mathbf{x}^*)$ be the corresponding bootstrap replication of the median. Show that

(a) $s(\mathbf{x}^*)$ equals one of the original data values $x_{(i)}$, $i = 1, 2, \cdots, 7$.

(b) † $s(\mathbf{x}^*)$ equals $x_{(i)}$ with probability

$$p(i) = \sum_{j=0}^{3} \{\text{Bi}(j; n, \frac{i-1}{n}) - \text{Bi}(j; n, \frac{i}{n})\}, \qquad (2.5)$$

where $\text{Bi}(j; n, p)$ is the binomial probability $\binom{n}{j} p^j (1-p)^{n-j}$. [The numerical values of $p(i)$ are .0102, .0981, .2386, .3062, .2386, .0981, .0102. These values were used to compute $\widehat{\text{se}}_{\text{boot}}\{\text{median}\} = 37.83$, for $B = \infty$, Table 2.2.]

2.5 Apply the weak law of large numbers to show that expression (2.3) approaches expression (2.4) as n goes to infinity.

† Indicates a difficult or more advanced problem.

CHAPTER 3

Random samples and probabilities

3.1 Introduction

Statistics is the theory of accumulating information, especially information that arrives a little bit at a time. A typical statistical situation was illustrated by the mouse data of Table 2.1. No one mouse provides much information, since the individual results are so variable, but seven, or nine mice considered together begin to be quite informative. Statistical theory concerns the best ways of extracting this information. Probability theory provides the mathematical framework for statistical inference. This chapter reviews the simplest probabilistic model used to model random data: the case where the observations are a random sample from a single unknown population, whose properties we are trying to learn from the observed data.

3.2 Random samples

It is easiest to visualize random samples in terms of a finite population or "universe" $\mathcal{U}$ of individual units $U_1, U_2, \cdots, U_N$, any one of which is equally likely to be selected in a single random draw. The population of units might be all the registered voters in an area undergoing a political survey, all the men that might conceivably be selected for a medical experiment, all the high schools in the United States, etc. The individual units have properties we would like to learn, like a political opinion, a medical survival time, or a graduation rate. It is too difficult and expensive to examine every unit in $\mathcal{U}$, so we select for observation a random sample of manageable size.

A *random sample of size* n is defined to be a collection of n