

Bootstrap Methods and Their Application

©A.C. Davison and D.V. Hinkley

Contents

<i>Preface</i>	i
1 Introduction	1
2 The Basic Bootstraps	11
2.1 Introduction	11
2.2 Parametric Simulation	15
2.3 Nonparametric Simulation	22
2.4 Simple Confidence Intervals	29
2.5 Reducing Error	31
2.6 Statistical Issues	38
2.7 Nonparametric Approximations for Variance and Bias	45
2.8 Subsampling Methods	55
2.9 Bibliographic Notes	59
2.10 Problems	61
2.11 Practicals	67
3 Further Ideas	71
3.1 Introduction	71
3.2 Several Samples	72
3.3 Semiparametric Models	78
3.4 Smooth Estimates of F	80
3.5 Censoring	83
3.6 Missing Data	89
3.7 Finite Population Sampling	93
3.8 Hierarchical Data	101

3.9	Bootstrapping the Bootstrap	104
3.10	Bootstrap Diagnostics	115
3.11	Choice of Estimator from the Data	120
3.12	Bibliographic Notes	125
3.13	Problems	127
3.14	Practicals	132
4	Tests	137
4.1	Introduction	137
4.2	Resampling for Parametric Tests	141
4.3	Nonparametric Permutation Tests	157
4.4	Nonparametric Bootstrap Tests	162
4.5	Adjusted P-values	177
4.6	Estimating Properties of Tests	182
4.7	Bibliographic Notes	185
4.8	Problems	186
4.9	Practicals	190
5	Confidence Intervals	193
5.1	Introduction	193
5.2	Basic Confidence Limit Methods	195
5.3	Percentile Methods	204
5.4	Theoretical Comparison of Methods	213
5.5	Inversion of Significance Tests	223
5.6	Double Bootstrap Methods	226
5.7	Empirical Comparison of Bootstrap Methods	233
5.8	Multiparameter Methods	235
5.9	Conditional Confidence Regions	241
5.10	Prediction	246
5.11	Bibliographic Notes	249
5.12	Problems	250
5.13	Practicals	255
6	Linear Regression	259
6.1	Introduction	259
6.2	Least Squares Linear Regression	260
6.3	Multiple Linear Regression	277
6.4	Aggregate Prediction Error and Variable Selection	294
6.5	Robust Regression	311
6.6	Bibliographic Notes	320
6.7	Problems	321

6.8	Practicals	326
7	Further Topics in Regression	331
7.1	Introduction	331
7.2	Generalized Linear Models	332
7.3	Survival Data	351
7.4	Other Nonlinear Models	359
7.5	Misclassification Error	364
7.6	Nonparametric Regression	368
7.7	Bibliographic Notes	380
7.8	Problems	382
7.9	Practicals	384
8	Complex Dependence	391
8.1	Introduction	391
8.2	Time Series	391
8.3	Point Processes	422
8.4	Bibliographic Notes	433
8.5	Problems	435
8.6	Practicals	439
9	Improved Calculation	443
9.1	Introduction	443
9.2	Balanced Bootstraps	444
9.3	Control Methods	452
9.4	Importance Resampling	457
9.5	Saddlepoint Approximation	473
9.6	Bibliographic Notes	492
9.7	Problems	494
9.8	Practicals	501
10	Semiparametric Likelihood Inference	506
10.1	Likelihood	506
10.2	Multinomial-Based Likelihoods	507
10.3	Bootstrap Likelihood	514
10.4	Likelihood Based on Confidence Sets	517
10.5	Bayesian Bootstraps	519
10.6	Bibliographic Notes	522
10.7	Problems	523
10.8	Practicals	526

11 Computer Implementation	529
11.1 Introduction	529
11.2 Basic Bootstraps	532
11.3 Further Ideas	538
11.4 Tests	541
11.5 Confidence Intervals	543
11.6 Linear Regression	544
11.7 Further Topics in Regression	547
11.8 Time Series	550
11.9 Improved Simulation	552
11.10 Semiparametric Likelihoods	556
Appendix 1 Cumulant Calculations	558

Preface

The publication in 1979 of Bradley Efron's first article on bootstrap methods was a major event in Statistics, at once synthesizing some of the earlier resampling ideas and establishing a new framework for simulation-based statistical analysis. The idea of replacing complicated and often inaccurate approximations to biases, variances, and other measures of uncertainty by computer simulations caught the imagination of both theoretical researchers and users of statistical methods. Theoreticians sharpened their pencils and set about establishing mathematical conditions under which the idea could work. Once they had overcome their initial skepticism, applied workers sat down at their terminals and began to amass empirical evidence that the bootstrap often did work better than traditional methods. The early trickle of papers quickly became a torrent, with new additions to the literature appearing every month, and it was hard to see when would be a good moment to try and chart the waters. Then the organizers of COMPSTAT'92 invited us to present a course on the topic, and shortly afterwards we began to write this book.

We decided to try and write a balanced account of resampling methods, to include basic aspects of the theory which underpinned the methods, and to show as many applications as we could in order to illustrate the full potential of the methods — warts and all. We quickly realized that in order for us and others to understand and use the bootstrap, we would need suitable software, and producing it led us further towards a practically-oriented treatment. Our view was cemented by two further developments: the appearance of the two excellent books, one by Peter Hall on the asymptotic theory and the other on basic methods by Bradley Efron and Robert Tibshirani; and the chance to give further courses that included practicals. Our experience has been that hands-on computing is essential in coming to grips with resampling ideas, so we have included practicals in this book, as well as more theoretical problems.

As the book expanded, we realized that a fully comprehensive treatment was

beyond us, and that certain topics could be given only a cursory treatment because too little is known about them. So it is that the reader will find only brief accounts of bootstrap methods for hierarchical data, missing data problems, model selection, robust estimation, nonparametric regression, and complex data. But we do try to point the more ambitious reader in the right direction.

No project of this size is produced in a vacuum. The majority of work on the book was completed while we were at the University of Oxford, and we are very grateful to colleagues and students there, who have helped shape our work in various ways. The experience of trying to teach these methods in Oxford and elsewhere — at the Université de Toulouse I, Université de Neuchâtel, Università degli Studi di Padova, Queensland University of Technology, Universidade de São Paulo, and University of Umeå— has been vital, and we are grateful to participants in these courses for prompting us to think more deeply about the material. Readers will be grateful to these people also, for unwittingly debugging some of the problems and practicals. We are also grateful to the organizers of COMPSTAT'92 and CLAPEM V for inviting us to give short courses on our work.

While writing this book we have asked many people for access to data, copies of their programs, papers or reprints; some have then been rewarded by our bombarding them with questions, to which the answers have invariably been courteous and informative. We cannot name all those who have helped in this way, but D. R. Brillinger, P. Hall, B. D. Ripley, H. O'R. Sternberg and G. A. Young have been especially generous. S. Hutchinson and B. D. Ripley have helped considerably with computing matters.

We are grateful to the mostly anonymous reviewers who commented on an early draft of the book, and to R. Gatto and G. A. Young, who later read various parts in detail. At Cambridge University Press, A. Woollatt and D. Tranah have helped greatly in producing the final version, and their patience has been commendable.

We are particularly indebted to two people. V. Ventura read large portions of the book, and helped with various aspects of the computation. A. J. Canty has turned our version of the bootstrap library functions into reliable working code, checked the book for mistakes, and has made numerous suggestions that have improved it enormously. Both of them have contributed greatly — though of course we take responsibility for any errors that remain in the book. We hope that readers will tell us about them, and we will do our best to correct any future versions of the book; see its WWW page, at URL <http://www.cup.cam.ac.uk> --- is this right?.

The book could not have been completed without grants from the U. K. Engineering and Physical Sciences Research Council, which in addition to providing funding for equipment and research assistantships, supported the

work of A. C. Davison through the award of an Advanced Research Fellowship. We also acknowledge support from the U. S. National Science Foundation.

We must also mention the Friday evening sustenance provided at the Eagle and Child, the Lamb and Flag, and the Royal Oak. The projects of many authors have flourished in these amiable establishments.

Finally, we thank our families, friends and colleagues for their patience while this project absorbed our time and energy.

A. C. Davison and D. V. Hinkley
Lausanne and Santa Barbara
October 1996

Introduction

The explicit recognition of uncertainty is central to the statistical sciences. Notions such as prior information, probability models, likelihood, standard errors and confidence limits are all intended to formalize uncertainty and thereby make allowance for it. In simple situations, the uncertainty of an estimate may be gauged by analytical calculation based on an assumed probability model for the available data. But in more complicated problems this approach can be tedious and difficult, and its results are potentially misleading if inappropriate assumptions or simplifications have been made.

For illustration, consider Table 1.1, which is taken from a larger tabulation (Table 7.4) of the numbers of AIDS reports in England and Wales from mid-1983 to the end of 1992. Reports are cross-classified by diagnosis period and length of reporting-delay, in three-month intervals. A blank in the table corresponds to an unknown (as-yet unreported) entry. The problem is to predict the states of the epidemic in 1991 and 1992, which depend heavily on the values missing at the bottom right of the table.

The data support the assumption that the reporting delay does not depend on the diagnosis period. In this case a simple model is that the number of reports in row j and column k of the table has a Poisson distribution with mean $\mu_{jk} = \exp(\alpha_j + \beta_k)$. If all the cells of the table are regarded as independent, then total number of unreported diagnoses in period j has a Poisson distribution with mean

$$\sum_k \mu_{jk} = \exp(\alpha_j) \sum_k \exp(\beta_k),$$

where the sum is over columns with blanks in row j . The eventual total of as-yet unreported diagnoses from period j can be estimated by replacing α_j and β_k by estimates derived from the incomplete table, and thence we obtain the predicted total for period j . Such predictions are shown by the solid line

Diagnosis period		Reporting-delay interval (quarters):									Total reports to end of 1992
Year	Quarter	0 [†]	1	2	3	4	5	6	...	≥14	
1988	1	31	80	16	9	3	2	8	...	6	174
	2	26	99	27	9	8	11	3	...	3	211
	3	31	95	35	13	18	4	6	...	3	224
	4	36	77	20	26	11	3	8	...	2	205
1989	1	32	92	32	10	12	19	12	...	2	224
	2	15	92	14	27	22	21	12	...	1	219
	3	34	104	29	31	18	8	6	...		253
	4	38	101	34	18	9	15	6	...		233
1990	1	31	124	47	24	11	15	8	...		281
	2	32	132	36	10	9	7	6	...		245
	3	49	107	51	17	15	8	9	...		260
	4	44	153	41	16	11	6	5	...		285
1991	1	41	137	29	33	7	11	6	...		271
	2	56	124	39	14	12	7	10			263
	3	53	175	35	17	13	11				306
	4	63	135	24	23	12					258
1992	1	71	161	48	25						310
	2	95	178	39							318
	3	76	181								273
	4	67									133

Table 1.1 Numbers of AIDS reports in England and Wales to the end of 1992 (De Angelis and Gilks, 1994), extracted from Table 7.4. A † indicates a reporting-delay less than one month.

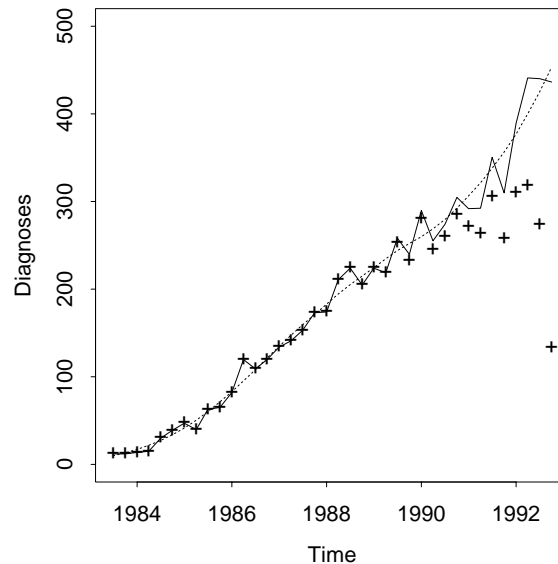
in Figure 1.1, together with the observed total reports to the end of 1992. How good are these predictions?

It would be tedious but possible to put pen to paper and estimate the prediction uncertainty through calculations based on the Poisson model. But in fact the data are much more variable than that model would suggest, and by failing to take this into account we would believe that the predictions are more accurate than they really are. Furthermore, a better approach would be to use a semiparametric model to smooth out the evident variability of the increase in diagnoses from quarter to quarter; the corresponding prediction is the dotted line in Figure 1.1. Analytical calculations for this model would be very unpleasant, and a more flexible line of attack is needed. While more than one approach is possible, the one that we shall develop based on computer simulation is both flexible and straightforward.

Purpose of the Book

Our central goal is to describe how the computer can be harnessed to obtain reliable standard errors, confidence intervals, and other measures of uncertainty for a wide range of problems. The key idea is to resample from the original data — either directly or via a fitted model — to create replicate data sets,

Figure 1.1
 Predicted diagnoses from a parametric model (solid) and a semiparametric model (dots) fitted to the AIDS data, together with the actual totals to the end of 1992 (+).



from which the variability of the quantities of interest can be assessed without long-winded and error-prone analytical calculation. Because this approach involves repeating the original data analysis procedure with many replicate sets of data, these are sometimes called *computer-intensive methods*. Another name for them is *bootstrap methods*, because to use the data to generate more data seems analogous to a trick used by the fictional Baron Münchhausen, who when he found himself at the bottom of a lake got out by pulling himself up by his bootstraps. In the simplest nonparametric problems we do literally sample from the data, and a common initial reaction is that this is a fraud. In fact it is not. It turns out that a wide range of statistical problems can be tackled this way, liberating the investigator from the need to over-simplify complex problems. The approach can also be applied in simple problems, to check the adequacy of standard measures of uncertainty, to relax assumptions, and to give quick approximate solutions. An example of this is random sampling to estimate the permutation distribution of a nonparametric test statistic.

It is of course true that in many applications we can be fairly confident in a particular parametric model and the standard analysis based on that model. Even so, it can still be helpful to see what can be inferred without particular parametric model assumptions. This is in the spirit of *robustness of validity* of the statistical analysis performed. Nonparametric bootstrap analysis allows us to do this.

Despite its scope and usefulness, resampling must be carefully applied. Unless certain basic ideas are understood, it is all too easy to produce a solution to the wrong problem, or a bad solution to the right one. Bootstrap methods are intended to help avoid tedious calculations based on questionable assumptions, and this they do. But they cannot replace clear critical thought about the problem, appropriate design of the investigation and data analysis, and incisive presentation of conclusions.

In this book we describe how resampling methods can be used, and evaluate their performance, in a wide range of contexts. Our focus is on the methods and their practical application rather than on the underlying theory, accounts of which are available elsewhere. This book is intended to be useful to the many investigators who want to know how and when the methods can safely be applied, and how to tell when things have gone wrong. The mathematical level of the book reflects this: we have aimed for a clear account of the key ideas without an overload of technical detail.

Examples

Bootstrap methods can be applied both when there is a well-defined probability model for data and when there is not. In our initial development of the methods we shall make frequent use of two simple examples, one of each type, to illustrate the main points.

Example 1.1 (Air-conditioning data) Table 1.2 gives $n = 12$ times between failures of air-conditioning equipment, for which we wish to estimate the underlying mean or its reciprocal, the failure rate. A simple model for this problem is that the times are sampled from an exponential distribution.

3	5	7	18	43	85	91	98	100	130	230	487
---	---	---	----	----	----	----	----	-----	-----	-----	-----

Table 1.2
Service-hours
between failures of
the air-conditioning
equipment in a
Boeing 720 jet
aircraft (Proschan,
1963).

The dotted line in the left panel of Figure 1.2 is the cumulative distribution function (CDF)

$$F_{\mu}(y) = \begin{cases} 0, & y \leq 0, \\ 1 - \exp(-y/\mu), & y > 0. \end{cases}$$

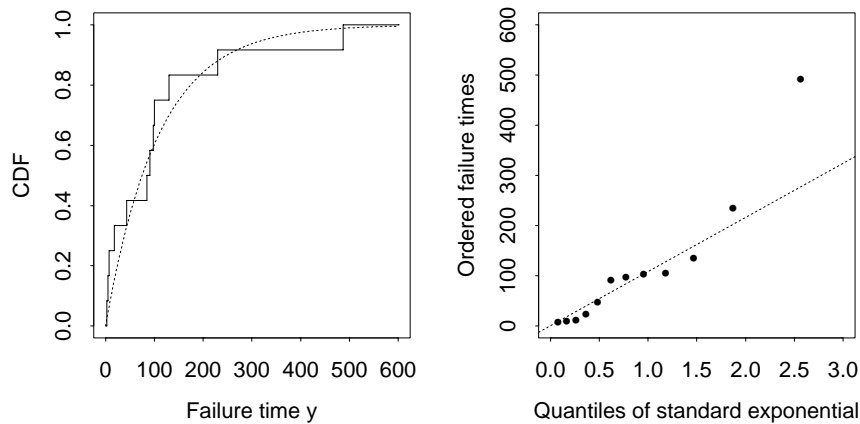
for the fitted exponential distribution with mean μ set equal to the sample average, $\bar{y} = 108.083$. The solid line on the same plot is the nonparametric equivalent, the empirical distribution function (EDF) for the data, which

places equal probabilities $n^{-1} = 0.083 \dots$ at each sample value. Comparison of the two curves suggests that the exponential model fits reasonably well. An alternative view of this is shown in the right panel of the figure, which is an exponential Q-Q plot — a plot of ordered data values $y_{(j)}$ against the standard exponential quantiles

$$F_{\mu}^{-1} \left(\frac{j}{n+1} \right) \Bigg|_{\mu=1} = -\log \left(1 - \frac{j}{n+1} \right).$$

Figure 1.2

Summary displays for the air-conditioning data. The left panel shows the EDF for the data, F (solid), and the CDF of a fitted exponential distribution (dots). The right panel shows a plot of the ordered failure times against exponential quantiles, with the fitted exponential model shown as the dotted line.



Although these plots suggest reasonable agreement with the exponential model, the sample is rather too small to have much confidence in this. In the data source the more general gamma model with mean μ and index κ is used; its density is

$$f_{\mu, \kappa}(y) = \frac{1}{\Gamma(\kappa)} \left(\frac{\kappa}{\mu} \right)^{\kappa} y^{\kappa-1} \exp(-\kappa y/\mu), \quad y > 0, \quad \mu, \kappa > 0. \quad (1.1)$$

For our sample the estimated index is $\hat{\kappa} = 0.71$, which does not differ significantly ($P = 0.29$) from the value $\kappa = 1$ that corresponds to the exponential model. Our reason for mentioning this will become apparent in Chapter 2.

Basic properties of the estimator $T = \bar{Y}$ for μ are easy to obtain theoretically under the exponential model. For example, it is easy to show that T is unbiased and has variance μ^2/n . Approximate confidence intervals for μ can be calculated using these properties in conjunction with a normal approximation for the distribution of T , although this does not work very well: see

can tell this because \bar{Y}/μ has an exact gamma distribution, which leads to exact confidence limits. Things are more complicated under the more general gamma model, because the index κ is only estimated, and so in a traditional approach we would use approximations — such as a normal approximation for the distribution of T , or a chi-squared approximation for the log likelihood ratio statistic. The parametric simulation methods of Section 2.2 can be used alongside these approximations, to diagnose problems with them, or to replace them entirely. ■

Example 1.2 (City population data) Table 1.3 reports $n = 49$ data pairs, each corresponding to a US city, the pair being the 1920 and 1930 populations of the city, which we denote by u and x . The data are plotted in Figure 1.3. Interest here is in the ratio of means, because this would enable us to estimate the total population of the US in 1930 from the 1920 figure. If the cities form a random sample with (U, X) denoting the pair of population values for a randomly selected city, then the total 1930 population is the product of the total 1920 population and the ratio of expectations $\theta = E(X)/E(U)$. This ratio is the parameter of interest.

In this case there is no obvious parametric model for the joint distribution of (U, X) , so it is natural to estimate θ by its empirical analog, $T = \bar{X}/\bar{U}$, the ratio of sample averages. We are then concerned with the uncertainty in T . If we had a plausible parametric model — for example, that the pair (U, X) has a bivariate lognormal distribution — then theoretical calculations like those in Example 1.1 would lead to bias and variance estimates for use in a normal approximation, which in turn would provide approximate confidence intervals for θ . Without such a model we must use nonparametric analysis. It is still possible to estimate the bias and variance of T , as we shall see, and this makes normal approximation still feasible, as well as more complex approaches to setting confidence intervals. ■

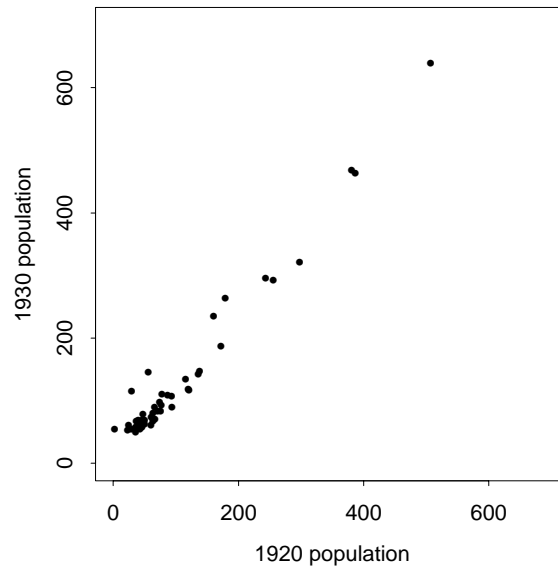
Example 1.1 is special in that an exact distribution is available for the statistic of interest and can be used to calculate confidence limits, at least under the exponential model. But for parametric models in general this will not be true. In Section 2.2 we shall show how to use parametric simulation to obtain approximate distributions, either by approximating moments for use in normal approximations, or — when these are inaccurate — directly.

In Example 1.2 we make no assumptions about the form of the data distribution. But still, as we shall show in Section 2.3, simulation can be used to obtain properties of T , even to approximate its distribution. Much of Chapter 2 is devoted to this.

Table 1.3
Populations in
thousands of $n = 49$
large United States
cities in 1920 (u)
and in 1930 (x)
(Cochran, 1977,
p. 152).

u	x	u	x	u	x
138	143	76	80	67	67
93	104	381	464	120	115
61	69	387	459	172	183
179	260	78	106	66	86
48	75	60	57	46	65
37	63	507	634	121	113
29	50	50	64	44	58
23	48	77	89	64	63
30	111	64	77	56	142
2	50	40	60	40	64
38	52	136	139	116	130
46	53	243	291	87	105
71	79	256	288	43	61
25	57	94	85	43	50
298	317	36	46	161	232
74	93	45	53	36	54
50	58				

Figure 1.3
Populations of 49
large United States
cities (in 1000s) in
1920 and 1930.



Layout of the Book

Chapter 2 describes the properties of resampling methods for use with single samples from parametric and nonparametric models, discusses practical mat-

ters such as the numbers of replicate data sets required, and outlines delta methods for variance approximation based on different forms of jackknife. It also contains a basic discussion of confidence intervals and of the ideas that underlie bootstrap methods.

Chapter 3 outlines how the basic ideas are extended to several samples, semiparametric and smooth models, simple cases where data have hierarchical structure or are sampled from a finite population, and to situations where data are incomplete because censored or missing. It goes on to discuss how the simulation output itself may be used to detect problems — so-called bootstrap diagnostics — and how it may be useful to bootstrap the bootstrap.

In Chapter 4 we review the basic principles of significance testing, and then describe Monte Carlo tests, including those using Markov Chain simulation, and parametric bootstrap tests. This is followed by discussion of nonparametric permutation tests, and the more general methods of semi- and nonparametric bootstrap tests. A double bootstrap method is detailed for improved approximation of P-values.

Confidence intervals are the subject of Chapter 5. After outlining basic ideas, we describe how to construct simple confidence intervals based on simulations, and then go on to more complex methods, such as the studentized bootstrap, percentile methods, the double bootstrap and test inversion. The main methods are compared empirically in Section 5.7. Then there are brief accounts of confidence regions for multivariate parameters, and of prediction intervals.

The three subsequent chapters deal with more complex problems. Chapter 6 describes how the basic resampling methods may be applied in linear regression problems, including tests for coefficients, prediction analysis, and variable selection. Chapter 7 deals with more complex regression situations: generalized linear models, other nonlinear models, semi- and nonparametric regression, survival analysis, and classification error. Chapter 8 details methods appropriate for time series, spatial data, and point processes.

Chapter 9 describes how variance reduction techniques such as balanced simulation, control variates, and importance sampling can be adapted to yield improved simulations, with the aim of reducing the amount of simulation needed for an answer of given accuracy. It also shows how saddlepoint methods can sometimes be used to avoid simulation entirely.

Chapter 10 describes various semiparametric versions of the likelihood function, the ideas underlying which are closely related to resampling methods. It also briefly outlines a Bayesian version of the bootstrap.

Chapters 2–10 contain problems intended to reinforce the reader's understanding of both methods and theory, and in some cases problems develop topics that could not be included in the text. Some of these demand a knowl-

edge of moments and cumulants, basic facts about which are sketched in the Appendix.

The book also contains practicals that apply resampling routines written in the S language to sets of data. The practicals are intended to reinforce the ideas in each chapter, to supplement the more theoretical problems, and to give examples on which readers can base analyses of their own data.

It would be possible to give different sorts of course based on this book. One would be a ‘theoretical’ course based on the problems and another an ‘applied’ course based on the practicals; we prefer to blend the two.

Although a library of routines for use with the statistical package **SPlus** is bundled with it, most of the book can be read without reference to particular software packages. Apart from the practicals, the exception to this is Chapter 11, which is a short introduction to the main resampling routines, arranged roughly in the order with which the corresponding ideas appear in earlier chapters. Readers intending to use the bundled routines will find it useful to work through the relevant sections of Chapter 11 before attempting the practicals.

Notation

Although we believe that our notation is largely standard, there are not enough letters in the English and Greek alphabets for us to be entirely consistent. Greek letters such as θ , β and ν generally denote parameters or other unknowns, while α is used for error rates in connexion with significance tests and confidence sets. English letters X , Y , Z , and so forth are used for random variables, which take values x , y , z . Thus the estimator T has observed value t , which may be an estimate of the unknown parameter θ . The letter V is used for a variance estimate, and the letter p for a probability, except for regression models, where p is the number of covariates.

Probability, expectation, variance and covariance are denoted $\Pr(\cdot)$, $E(\cdot)$, $\text{var}(\cdot)$ and $\text{cov}(\cdot, \cdot)$, while the joint cumulant of Y_1 , Y_1Y_2 and Y_3 is denoted $\text{cum}(Y_1, Y_1Y_2, Y_3)$. We use $I\{A\}$ to denote the indicator random variable, which takes values one if the event A is true and zero otherwise. A related function is the Heaviside function

$$H(u) = \begin{cases} 0, & u < 0, \\ 1, & u \geq 0. \end{cases}$$

We use $\#\{A\}$ to denote the number of elements in the set A , and $\#\{A_r\}$ for the number of events A_r that occur in a sequence A_1, A_2, \dots

The data values in a sample of size n are typically denoted by y_1, \dots, y_n , the observed values of the random variables Y_1, \dots, Y_n ; their average is $\bar{y} = n^{-1} \sum y_j$.

We mostly reserve Z for random variables that are standard normal, at least approximately, and use Q for random variables with other (approximately) known distributions. As usual $N(\mu, \sigma^2)$ represents the normal distribution with mean μ and variance σ^2 , while z_α is often the α quantile of the standard normal distribution.

The letter R is reserved for the number of replicate simulations. Simulated copies of a statistic T are denoted T_r^* , $r = 1, \dots, R$, whose ordered values are $T_{(1)}^* \leq \dots \leq T_{(R)}^*$. Expectation, variance and probability calculated with respect to the simulation distribution are written $\Pr^*(\cdot)$, $E^*(\cdot)$ and $\text{var}^*(\cdot)$.

Where possible we avoid boldface type, and rely on the context to make it plain when we are dealing with vectors or matrices; a^T denotes the matrix transpose of a vector or matrix a .

We use PDF, CDF, and EDF as shorthand for probability density function, cumulative distribution function, and empirical distribution function. The letters F and G are used for CDFs, and f and g are generally used for the corresponding PDFs. An exception to this is that f_{rj}^* denotes the frequency with which y_j appears in the r th resample.

The end of each example is marked \blacksquare , and the end of each algorithm is marked \bullet .

The Basic Bootstraps

2.1 Introduction

In this chapter we discuss techniques which are applicable to a single, homogeneous sample of data, denoted by y_1, \dots, y_n . The sample values are thought of as the outcomes of independent and identically distributed random variables Y_1, \dots, Y_n whose *probability density function* (PDF) and *cumulative distribution function* (CDF) we shall denote by f and F , respectively. The sample is to be used to make inferences about a population characteristic, generically denoted by θ , using a statistic T whose value in the sample is t . We assume for the moment that the choice of T has been made and that it is an estimate for θ , which we take to be a scalar.

Our attention is focussed on questions concerning the probability distribution of T . For example, what are its bias, its standard error, or its quantiles? What are likely values under a certain null hypothesis of interest? How do we calculate confidence limits for θ using T ?

There are two situations to distinguish, the parametric and the nonparametric. When there is a particular mathematical model, with adjustable constants or parameters ψ that fully determine f , such a model is called *parametric* and statistical methods based on this model are parametric methods. In this case the parameter of interest θ is a component of or function of ψ . When no such mathematical model is used, the statistical analysis is *nonparametric*, and uses only the fact that the random variables Y_j are independent and identically distributed. Even if there is a plausible parametric model, a nonparametric analysis can still be useful to assess the robustness of conclusions drawn from a parametric analysis.

An important role is played in nonparametric analysis by the *empirical distribution* which puts equal probabilities n^{-1} at each sample value y_j . The corresponding estimate of F is the *empirical distribution function* (EDF) \hat{F}

which is defined as the sample proportion

$$\hat{F}(y) = \frac{\#\{y_j \leq y\}}{n}.$$

$\#\{A\}$ means the number of times the event A occurs.

More formally

$$\hat{F}(y) = \frac{1}{n} \sum_{j=1}^n H(y - y_j), \quad (2.1)$$

where $H(u)$ is the unit step function which jumps from 0 to 1 at $u = 0$. Notice that the values of the EDF are fixed $(0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n}{n})$, so the EDF is equivalent to its points of increase, the ordered values $y_{(1)} \leq \dots \leq y_{(n)}$ of the data. An example of the EDF was shown in the left panel of Figure 1.2.

When there are repeat values in the sample, as would often occur with discrete data, the EDF assigns probabilities proportional to the sample frequencies at each distinct observed value y . The formal definition (2.1) still applies.

The EDF plays the role of fitted model when no mathematical form is assumed for F , analogous to a parametric CDF with parameters replaced by their estimates.

2.1.1 Statistical functions

Many simple statistics can be thought of in terms of properties of the EDF. For example, the sample average $\bar{y} = n^{-1} \sum y_j$ is the mean of the EDF; see Example 2.1 below. More generally the statistic of interest t will be a symmetric function of y_1, \dots, y_n , meaning that t is unaffected by reordering the data. This implies that t depends only on the ordered values $y_{(1)} \leq \dots \leq y_{(n)}$, or equivalently on the EDF \hat{F} . Often this can be expressed simply as $t = t(\hat{F})$, where $t(\cdot)$ is a *statistical function* — essentially just a mathematical expression of the algorithm for computing t from \hat{F} . Such a statistical function is of central importance in the nonparametric case because it also defines the parameter of interest θ through the “algorithm” $\theta = t(F)$. This corresponds to the qualitative idea that θ is a characteristic of the population described by F . Simple examples of such functions are the mean and variance of Y , which are respectively defined as

$$t(F) = \int y dF(y), \quad t(F) = \int y^2 dF(y) - \left\{ \int y dF(y) \right\}^2. \quad (2.2)$$

The same definition of θ applies in parametric problems, although then θ is more usually defined explicitly as one of the model parameters ψ .

The relationship between the estimate t and \hat{F} can usually be expressed as $t = t(\hat{F})$, corresponding to the relation $\theta = t(F)$ between the characteristic of interest and the underlying distribution. The statistical function $t(\cdot)$ defines

both the parameter and its estimate, but we shall use $t(\cdot)$ to represent the function, and t to represent the estimate of θ based on the observed data y_1, \dots, y_n .

Example 2.1 (Average) The sample average, \bar{y} , estimates the population mean

$$\mu = \int y dF(y).$$

To show that $\bar{y} = t(\hat{F})$, we substitute for \hat{F} in the defining function at (2.2) to obtain

$$\begin{aligned} t(\hat{F}) &= \int y d\hat{F}(y) = \int y d\left(\frac{1}{n} \sum_{j=1}^n H(y - y_j)\right) = \frac{1}{n} \sum_{j=1}^n \int y dH(y - y_j) \\ &= \frac{1}{n} \sum_{j=1}^n y_j = \bar{y}, \end{aligned}$$

because $\int a(y) dH(y - x) = a(x)$ for any function $a(\cdot)$. ■

Example 2.2 (City population data) For the problem outlined in Example 1.2, the parameter of interest is the ratio of means $\theta = E(X)/E(U)$. In this case F is the bivariate CDF of $Y = (U, X)$, and the bivariate EDF \hat{F} puts probability n^{-1} at each of the data pairs (u_j, x_j) . The statistical function version of θ simply uses the definition of mean for both numerator and denominator, so that

$$t(F) = \frac{\int x dF(u, x)}{\int u dF(u, x)}.$$

The corresponding estimate of θ is

$$t = t(\hat{F}) = \frac{\int x d\hat{F}(u, x)}{\int u d\hat{F}(u, x)} = \frac{\bar{x}}{\bar{u}},$$

with $\bar{x} = n^{-1} \sum x_j$ and $\bar{u} = n^{-1} \sum u_j$. ■

It is quite straightforward to show that (2.1) implies convergence of \hat{F} to F as $n \rightarrow \infty$ (Problem 2.1). Then if $t(\cdot)$ is continuous in an appropriate sense, the definition $T = t(\cdot)$ implies that T converges to θ as $n \rightarrow \infty$, which is the property of consistency.

Not all estimates are exactly of the form $t(\hat{F})$. For example, if $t(F) = \text{var}(Y)$ then the usual unbiased sample variance is $nt(\hat{F})/(n-1)$. Also the sample median is not exactly $\hat{F}^{-1}(\frac{1}{2})$. Such small discrepancies are fairly unimportant as far as applying the bootstrap techniques discussed in this book. In a very formal development we could write $T = t_n(\hat{F})$ and require that $t_n \rightarrow t$ as $n \rightarrow \infty$, possibly even that $t_n - t = O(n^{-1})$. But such formality would be excessive

A quantity A_n is said to be $O(n^d)$ if $\lim_{n \rightarrow \infty} n^{-d} A_n = a$ for some finite a , and $o(n^d)$ if $\lim_{n \rightarrow \infty} n^{-d} A_n = 0$.

here, and we shall assume in general discussion that $T = t(\hat{F})$. (One case that does require special treatment is nonparametric density estimation, which we discuss in Example 5.13.)

The representation $\theta = t(F)$ defines the parameter and its estimator T in a robust way, without any assumption about F , other than that θ exists. This guarantees that T estimates the right thing, no matter what F is. Thus the sample average \bar{y} is the only statistic that is generally valid as an estimate of the population mean μ : only if Y is symmetrically distributed about μ will statistics such as trimmed averages also estimate μ . This property, which guarantees that the correct characteristic of the underlying distribution is estimated, whatever that distribution is, is sometimes called *robustness of specification*.

2.1.2 Objectives

Much of statistical theory is devoted to calculating approximate distributions for particular statistics T , on which to base inferences about their estimands θ . Suppose, for example, that we want to calculate a $(1 - 2\alpha)$ confidence interval for θ . It may be possible to show that T is approximately normal with mean $\theta + \beta$ and variance ν ; here β is the bias of T . If β and ν are both known, then we can write

$$\Pr(T \leq t \mid F) \doteq \Phi\left(\frac{t - (\theta + \beta)}{\nu^{1/2}}\right), \quad (2.3)$$

where $\Phi(\cdot)$ is the standard normal integral. If the α quantile of the standard normal distribution is $z_\alpha = \Phi^{-1}(\alpha)$, then an approximate $(1 - 2\alpha)$ confidence interval for θ has limits

$$t - \beta - \nu^{1/2}z_{1-\alpha}, \quad t - \beta - \nu^{1/2}z_\alpha, \quad (2.4)$$

as follows from

$$\Pr(\beta + \nu^{1/2}z_\alpha \leq T - \theta \leq \beta + \nu^{1/2}z_{1-\alpha}) \doteq 1 - 2\alpha.$$

There is a catch, however, which is that in practice the bias β and variance ν will not be known. So to use the normal approximation we must replace β and ν with estimates. To see how to do this, note that we can express β and ν as

$$\beta = b(F) = E(T \mid F) - t(F), \quad \nu = v(F) = \text{var}(T \mid F), \quad (2.5)$$

thereby stressing their dependence on the underlying distribution. We use expressions such as $E(T \mid F)$ to mean that the random variables from which T is calculated have distribution F ; here a pedantic equivalent would be $E\{t(\hat{F}) \mid Y_1, \dots, Y_n \stackrel{iid}{\sim} F\}$. Suppose that F is estimated by \hat{F} , which might be the empirical distribution function, or a fitted parametric distribution. Then

estimates of bias and variance are obtained simply by substituting \hat{F} for F in (2.5), that is

$$B = b(\hat{F}) = E(T | \hat{F}) - t(\hat{F}), \quad V = v(\hat{F}) = \text{var}(T | \hat{F}). \quad (2.6)$$

These estimates B and V are used in place of β and ν in equations such as (2.4).

Example 2.3 (Air conditioning data) Under the exponential model for the data in Example 1.1, the mean failure time μ is estimated by the average $T = \bar{Y}$, which has a gamma distribution with mean μ and shape parameter $\kappa = n$. Therefore the bias and variance of T are $b(F) = 0$ and $v(F) = \mu^2/n$, and these are estimated by 0 and \bar{y}^2/n . Since $n = 12$, $\bar{y} = 108.083$, and $z_\alpha = -1.96$, a 95% confidence interval for μ based on the normal approximation (2.3) is $\bar{y} \pm 1.96n^{-1/2}\bar{y} = (46.93, 169.24)$. ■

Estimates such as those in (2.6) are bootstrap estimates. Here they have been used in conjunction with a normal approximation, which sometimes will be adequate. However, the bootstrap approach of substituting estimates can be applied more ambitiously to improve upon the normal approximation and other first-order theoretical approximations. The elaboration of the bootstrap approach is the purpose of this book.

2.2 Parametric Simulation

In the previous section we pointed out that theoretical properties of T might be hard to determine with sufficient accuracy. We now describe the sound practical alternative of repeated simulation of data sets from a fitted parametric model, and empirical calculation of relevant properties of T .

Suppose that we have a particular parametric model for the distribution of the data y_1, \dots, y_n . We shall use $F_\psi(y)$ and $f_\psi(y)$ to denote the CDF and PDF respectively. When ψ is estimated by $\hat{\psi}$ — often but not invariably its maximum likelihood estimate — its substitution in the model gives the *fitted model*, with CDF $\hat{F}(y) = F_{\hat{\psi}}(y)$, which can be used to calculate properties of T , sometimes exactly. We shall use Y^* to denote the random variable distributed according to the fitted model \hat{F} , and the superscript $*$ will be used with E , var and so forth when these moments are calculated according to the fitted distribution. Occasionally it will also be useful to write $\hat{\psi} = \psi^*$ to emphasise that this is the parameter value for the simulation model.

Example 2.4 (Air-conditioning data) We have already calculated the mean and variance under the fitted exponential model for the estimator $T = \bar{Y}$ of Example 1.1. Our sample estimate for the mean μ is $t = \bar{y}$. So here

Y^* is exponential with mean \bar{y} . In the notation just introduced, we have by theoretical calculation with this exponential distribution that

$$E^*(\bar{Y}^*) = \bar{y}, \quad \text{var}^*(\bar{Y}^*) = \bar{y}^2/n.$$

Note that the estimated bias of \bar{Y} is zero, being the difference between $E^*(\bar{Y}^*)$ and the value $\mu^* = \bar{y}$ for the mean of the fitted distribution. These moments were used to calculate an approximate normal confidence interval in Example 2.3.

If, however, we wished to calculate the bias and variance of $T = \log \bar{Y}$ under the fitted model, i.e. $E^*(\log \bar{Y}^*) - \log \bar{y}$ and $\text{var}^*(\log \bar{Y}^*)$, exact calculation is more difficult. The delta method of Section 2.7.1 would give approximate values $-(2n)^{-1}$ and n^{-1} . But more accurate approximations can be obtained using simulated samples of Y^* 's.

Similar results and comments would apply if instead we chose to use the more general gamma model for this example. Then Y^* would be a gamma random variable with mean \bar{y} and index $\hat{\kappa}$. ■

2.2.1 Moment estimates

So now suppose that theoretical calculation with the fitted model is too complex. Approximations may not be available, or they may be untrustworthy, perhaps because the sample size is small. The alternative is to estimate the properties we require from simulated data sets. We write such a data set as Y_1^*, \dots, Y_n^* where the Y_j^* are independently sampled from the fitted distribution \hat{F} . When the statistic of interest is calculated from a simulated data set, we denote it by T^* . From R repetitions of the data simulation we obtain T_1^*, \dots, T_R^* . Properties of $T - \theta$ are then estimated from T_1^*, \dots, T_R^* . For example, the estimator of the bias $b(F) = E(T | F) - \theta$ of T is

$$B = b(\hat{F}) = E(T | \hat{F}) - t = E^*(T^*) - t,$$

and this in turn is estimated by

$$B_R = R^{-1} \sum_{r=1}^R T_r^* - t = \bar{T}^* - t. \quad (2.7)$$

Note that in the simulation t is the parameter value for the model, so that $T^* - t$ is the simulation analogue of $T - \theta$. The corresponding estimator of the variance of T is

$$V_R = \frac{1}{R-1} \sum_{r=1}^R (T_r^* - \bar{T}^*)^2, \quad (2.8)$$

with similar estimators for other moments.

These empirical approximations are justified by the law of large numbers.

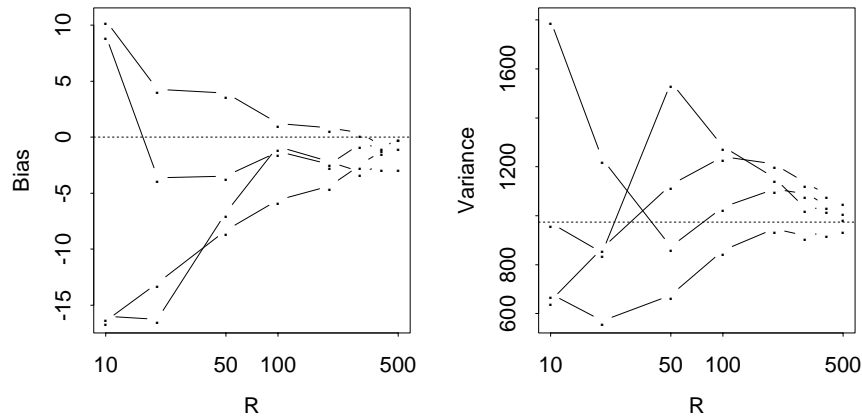
For example, B_R converges to B , the exact value under the fitted model, as R increases. We usually drop the subscript R from B_R , V_R , and so forth unless we are explicitly discussing the effect of R . How to choose R will be illustrated in the examples that follow, and discussed in Section 2.5.2.

It is important to recognize that we are not estimating absolute properties of T , but rather of T relative to θ . Usually this involves the estimation error $T - \theta$, but we should not ignore the possibility that T/θ (equivalently $\log T - \log \theta$) or some other relevant measure of estimation error might be more appropriate, depending upon the context. Bootstrap simulation methods will apply to any such measure.

Example 2.5 (Air-conditioning data) Consider Example 1.1 again. As we have seen, simulation is unnecessary in practice for this problem because the moments are easy to calculate theoretically, but the example is useful for illustration. Here the fitted model is an exponential distribution for the failure times, with mean estimated by the sample average $\bar{y} = 108.083$. All simulated failure times Y^* are generated from this distribution.

Figure 2.1 shows the results from several simulations, four for each of eight values of R , in each of which the empirical biases and variances of $T^* = \bar{Y}^*$ have been calculated according to (2.7) and (2.8). On both panels the “correct” values, namely zero and $\bar{y}^2/n = (108.083)^2/12 = 973.5$, are indicated by horizontal dotted lines.

Figure 2.1
Empirical biases and variances of \bar{Y}^* for the air-conditioning data from four repetitions of parametric simulation. Each line shows how the estimated bias and variance for $R = 10$ initial simulations change when further simulations are successively added. Note how the variability decreases as the simulation size increases, and how the simulated values converge to the exact values under the fitted exponential model, given by the horizontal dotted lines.



Evidently the larger is R the closer is the simulation calculation to the right answer. How large a value of R is needed? Figure 2.1 suggests that for some

purposes $R = 100$ or 200 will be adequate, but that $R = 10$ will not be large enough. In this problem the accuracy of the empirical approximations is quite easy to determine from the fact that $n\bar{Y}/\mu$ has a gamma distribution with index n . The simulation variances of B_R and V_R are

$$\frac{t^2}{nR}, \quad \frac{t^4}{n^2} \left(\frac{2}{R-1} + \frac{6}{nR} \right),$$

and we can use these to say how large R should be in order that the simulated values have a specified accuracy. For example, the coefficients of variation of V_R at $R = 100$ and 1000 are respectively 0.16 and 0.05 . However for a complicated problem where simulation was really necessary, such calculations could not be done, and general rules are needed to suggest how large R should be. These are discussed in Section 2.5.2. ■

2.2.2 Distribution and quantile estimates

The simulation estimates of bias and variance will sometimes be of interest in their own right, but more usually would be used with normal approximations for T , particularly for large samples. For situations like those in Examples 1.1 and 1.2, however, the normal approximation is intrinsically inaccurate. This can be seen from a normal Q-Q plot of the simulated values t_1^*, \dots, t_R^* , that is, a plot of the ordered values $t_{(1)}^* < \dots < t_{(R)}^*$ against expected normal order statistics. It is the empirical distribution of these simulated values which can provide a more accurate distributional approximation, as we shall now see.

If as is often the case we are approximating the distribution of $T - \theta$ by that of $T^* - t$, then cumulative probabilities are estimated simply by the empirical distribution function of the simulated values $t^* - t$. More formally, if $G(u) = \Pr(T - \theta \leq u)$, then the simulation estimate of $G(u)$ is

$$\hat{G}_R(u) = \frac{\#\{t_r^* - t \leq u\}}{R} = \frac{1}{R} \sum_{r=1}^R I\{t_r^* - t \leq u\},$$

$\#\{A\}$ means the number of times event A occurs.

where $I\{A\}$ is the indicator of the event A , equal to 1 if A is true and 0 otherwise. As R increases, so this estimate will converge to $\hat{G}(u)$, the exact CDF of $T^* - t$ under sampling from the fitted model. Just as with the moment approximations discussed earlier, so the approximation \hat{G}_R to G contains two sources of error, i.e. that between \hat{G} and G due to data variability and that between \hat{G}_R and \hat{G} due to finite simulation.

We are often interested in quantiles of the distribution of $T - \theta$, and these are approximated using ordered values of $t^* - t$. The underlying result used here is that if X_1, \dots, X_N are independently distributed with CDF K and if

$X_{(j)}$ denotes the j th ordered value, then

$$E(X_{(j)}) \doteq K^{-1} \left(\frac{j}{N+1} \right).$$

This implies that a sensible estimate of $K^{-1}(p)$ is $X_{((N+1)p)}$, assuming that $(N+1)p$ is an integer. So we estimate the p quantile of $T - \theta$ by the $(R+1)p$ th ordered value of $t^* - t$, that is $t_{((R+1)p)}^* - t$. We assume that R is chosen so that $(R+1)p$ is an integer.

The simulation approximation \hat{G}_R and the corresponding quantiles are in principle better than results obtained by normal approximation, provided that R is large enough, because they avoid the supposition that the distribution of $T^* - t$ has a particular form.

Example 2.6 (Air-conditioning data) The simulation experiments described in Example 2.5 can be used to study the simulation approximations to the distribution and quantiles of $\bar{Y} - \mu$. First, Figure 2.2 shows normal Q-Q plots of t^* values for $R = 99$ (top left panel) and $R = 999$ (top right panel). Clearly a normal approximation would not be accurate in the tails, and this is already fairly clear with $R = 99$. For reference, the lower half of Figure 2.2 shows corresponding Q-Q plots with exact gamma quantiles.

The nonnormality of T^* is also reasonably clear on histograms of t^* values, shown in Figure 2.3, at least at the larger value $R = 999$. Corresponding density estimate plots provide smoother displays of the same information.

We look next at the estimated quantiles of $\bar{Y} - \mu$. The p quantile is approximated by $\bar{y}_{((R+1)p)}^* - \bar{y}$ for $p = 0.05$ and 0.95 . The values of R are $19, 39, 99, 199, \dots, 999$, chosen to ensure that $(R+1)p$ is an integer throughout. Thus at $R = 19$ the 0.05 quantile is approximated by $\bar{y}_{(1)}^* - \bar{y}$ and so forth. In order to assess the magnitude of simulation error, we ran four independent simulations at $R = 19, 39, 99, \dots, 999$. The results are plotted in Figure 2.4. Also shown by dotted lines are the exact quantiles under the model, which the simulations approach as R increases. There is large variability in the approximate quantiles for R less than 100 and it appears that 500 or more simulations are required to get accurate results.

The same simulations can be used in other ways. For example, we might want to know about $\log \bar{Y} - \log \mu$, in which case the empirical properties of $\log \bar{y}^* - \log \bar{y}$ are relevant. ■

The illustration used here is very simple, but essentially the same methods can be used in arbitrarily complicated parametric problems. For example, distributions of likelihood ratio statistics can be approximated when large-sample approximations are inaccurate (Example 2.11 in Section 2.4) or fail entirely. In Chapters 4 and 5 respectively we show how parametric bootstrap methods can be used to calculate significance tests and confidence sets.

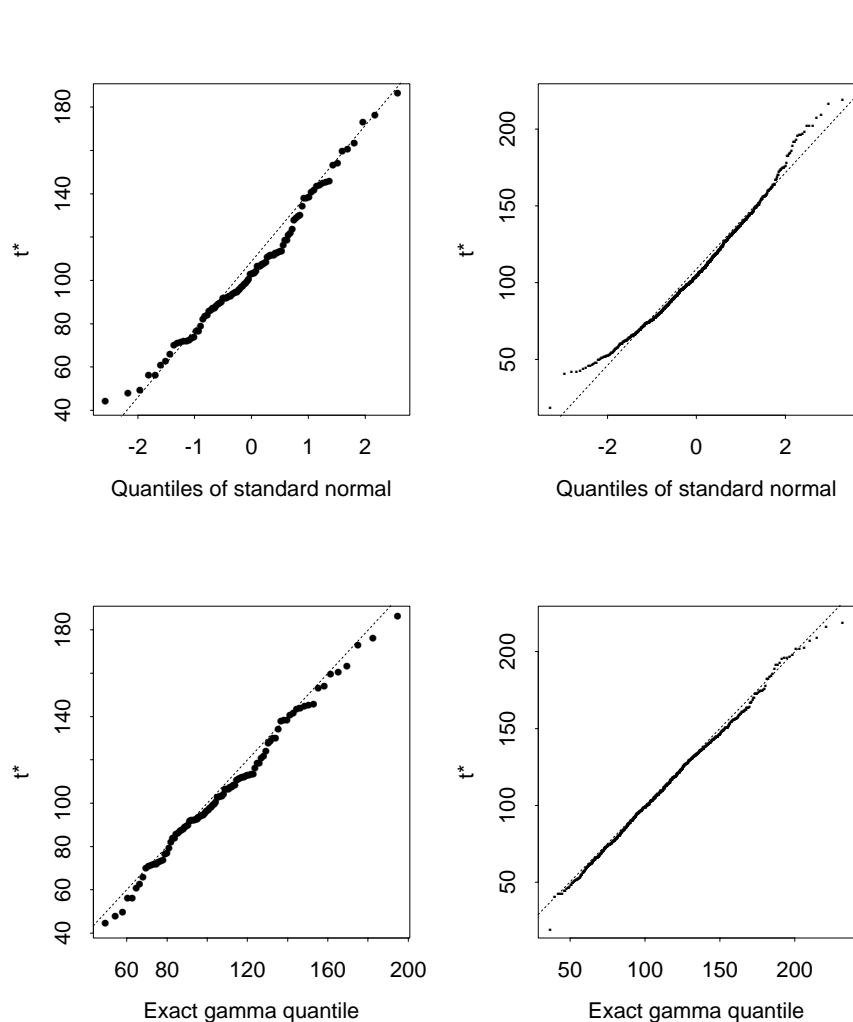


Figure 2.2 Normal (upper) and gamma (lower) Q-Q plots of t^* values based on $R = 99$ (left) and $R = 999$ (right) simulations from the fitted exponential model for the air-conditioning data.

It is sometimes useful to be able to look at the density of T , for example to see if it is multimodal, skewed, or otherwise differs appreciably from normality. A rough idea of the density $g(u)$ of $U = T - \theta$, say, can be had from a histogram of the values of $t^* - t$. A somewhat better picture is offered by a kernel density estimate, defined by

$$\hat{g}_h(u) = \frac{1}{Rh} \sum_{r=1}^R w \left\{ \frac{u - (t_r^* - t)}{h} \right\}, \quad (2.9)$$

where w is a symmetric PDF with zero mean and h is a positive bandwidth that determines the smoothness of \hat{g}_h . The estimate \hat{g}_h is non-negative and

Figure 2.3
Histograms of t^* values based on $R = 99$ (left) and $R = 999$ (right) simulations from the fitted exponential model for the air-conditioning data.

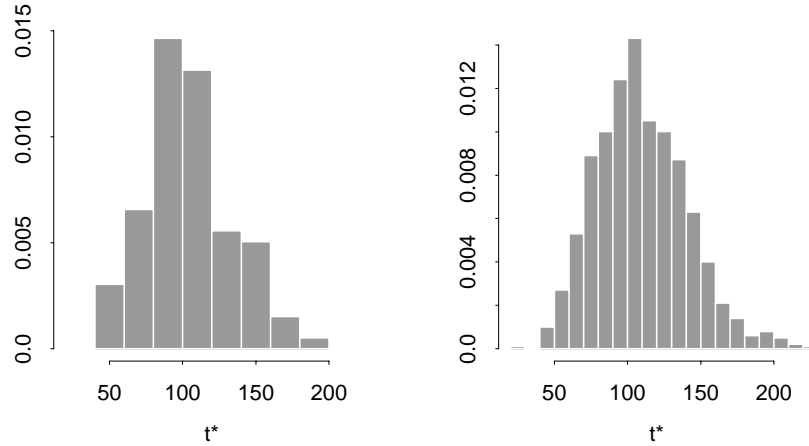
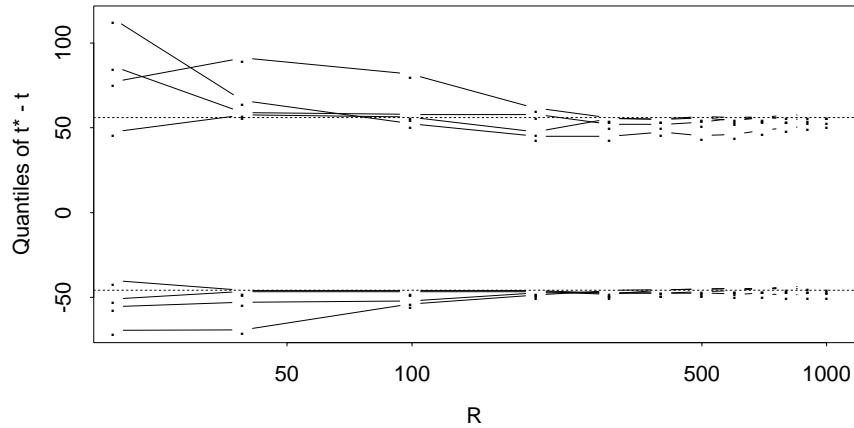


Figure 2.4
Empirical quantiles ($p = 0.05, 0.95$) of $T^* - t$ under resampling from the fitted exponential model for the air-conditioning data. The horizontal dotted lines are the exact quantiles under the model.



has unit integral. It is insensitive to the choice of $w(\cdot)$, for which we use the standard normal density. The choice of h is more important. The key is to produce a smooth result, while not flattening out significant modes. If the choice of h is quite large, as may be if $R \leq 100$, then one should rescale the density estimate to make its mean and variance agree with the estimated mean b_R and variance v_R of $T - \theta$; see Problem 3.8.

As a general rule, good estimates of density require at least $R = 1000$: density estimation is usually harder than probability or quantile estimation.

Note that the same methods of estimating density, distribution function and quantiles can be applied to any transformation of T . We shall discuss this further in Section 2.5.

2.3 Nonparametric Simulation

Suppose that we have no parametric model, but that it is sensible to assume that Y_1, \dots, Y_n are independent and identically distributed according to an unknown distribution function F . We use the EDF \hat{F} to estimate the unknown CDF F . We shall use \hat{F} just as we would a parametric model: theoretical calculation if possible, otherwise simulation of data sets and empirical calculation of required properties. In only very simple cases are exact theoretical calculations possible, but we shall see later that good theoretical approximations can be obtained in many problems involving sample moments.

Example 2.7 (Average) In the case of the average, exact moments under sampling from the EDF are easily found. For example,

$$E^*(\bar{Y}^*) = E^*(Y^*) = \sum_{j=1}^n \frac{1}{n} y_j = \bar{y}$$

and similarly

$$\begin{aligned} \text{var}^*(\bar{Y}^*) &= \frac{1}{n} \text{var}^*(Y^*) = \frac{1}{n} E^* \{Y^* - E^*(Y^*)\}^2 \\ &= \frac{1}{n} \times \sum_{j=1}^n \frac{1}{n} (y_j - \bar{y})^2 = \frac{(n-1)}{n} \times \frac{1}{n(n-1)} \sum_{j=1}^n (y_j - \bar{y})^2. \end{aligned}$$

Apart from the factor $(n-1)/n$, this is the usual result for the estimated variance of \bar{Y} . ■

Other simple statistics such as the sample variance and sample median are also easy to handle (Problems 2.3, 2.4).

To apply simulation with the EDF is very straightforward. Because the EDF puts equal probabilities on the original data values y_1, \dots, y_n , each Y^* is independently sampled at random from those data values. Therefore the simulated sample Y_1^*, \dots, Y_n^* is a random sample taken with replacement from the data. This simplicity is special to the case of a homogeneous sample, but many extensions are straightforward. This resampling procedure is called the *nonparametric bootstrap*.

Example 2.8 (City population data) Here we look at the ratio estimate for the problem described in Example 1.2. For convenience we consider a subset of the data in Table 1.3, comprising the first ten pairs. This is an application with no obvious parametric model, so nonparametric simulation makes

good sense. Table 2.1 shows the data and the first simulated sample, which has been drawn by randomly selecting subscript j^* from the set $\{1, \dots, n\}$ with equal probability and taking $(u^*, x^*) = (u_{j^*}, x_{j^*})$. In this sample $j^* = 1$ never occurs and $j^* = 2$ occurs three times, so that the first data pair is never selected, the second is selected three times, and so forth.

Table 2.1 The data set for ratio estimation, and one synthetic sample. The values j^* are chosen randomly with equal probability from $\{1, \dots, n\}$ with replacement; the simulated pairs are (u_{j^*}, x_{j^*}) .

j	1	2	3	4	5	6	7	8	9	10
u	138	93	61	179	48	37	29	23	30	2
x	143	104	69	260	75	63	50	48	111	50
j^*	6	7	2	2	3	3	10	7	2	9
u^*	37	29	93	93	61	61	2	29	93	30
x^*	63	50	104	104	69	69	50	50	104	111

Table 2.2 shows the same simulated sample, plus eight more, expressed in terms of the frequencies of original data pairs. The ratio t^* for each simulated sample is recorded in the last column of the table. After the R sets of calculations, the bias and variance estimates are calculated according to (2.7) and (2.8). The results are, for the $R = 9$ replicates shown,

$$b = 1.582 - 1.520 = 0.062, \quad v = 0.03907.$$

Table 2.2 Frequencies with which each original data pair appears in each of $R = 9$ nonparametric bootstrap samples for the data on US cities.

j	1	2	3	4	5	6	7	8	9	10	Statistic
	u	138	93	61	179	48	37	29	23	30	
x	143	104	69	260	75	63	50	48	111	50	
Data	Numbers of times each pair sampled										$t = 1.520$
	1	1	1	1	1	1	1	1	1	1	
Replicate r											
1		3	2			1	2		1	1	$t_1^* = 1.466$
2	1		1		2	2	1		2	1	$t_2^* = 1.761$
3	1	1		1		1			4	2	$t_3^* = 1.951$
4		1	2		1	1	2	2		1	$t_4^* = 1.542$
5	3			1	3		1	1	1		$t_5^* = 1.371$
6	1	1	2			1		1	1	3	$t_6^* = 1.686$
7	1	1	2	2	2		1			1	$t_7^* = 1.378$
8	2		1		3	1	1	1	1		$t_8^* = 1.420$
9		1	1	1	2	1		2	1	1	$t_9^* = 1.660$

A simple approximate distribution for $T - \theta$ is $N(b, v)$. With the results so far, this is $N(0.062, 0.0391)$, but this is unlikely to be accurate enough and a larger value of R should be used. In a simulation with $R = 999$ we obtained $b = 1.5755 - 1.5203 = 0.0552$ and $v = 0.0601$. The latter is appreciably bigger than the nonparametric delta method variance estimate for t , which in Example 2.12 is shown to equal 0.0325. Below we shall usually use the more compact notation $v_L(\hat{F}) = v_L$. This variance estimate is based on an expansion of t in terms of the y_j , analogous to Taylor series expansion and described in Section 2.7.2. The discrepancy between v and v_L is due partly to a few extreme values of t^* , an issue we discuss in Section 2.3.2.

The left panel of Figure 2.5 shows a histogram of t^* , whose skewness is evident: use of a normal approximation here would be very inaccurate.

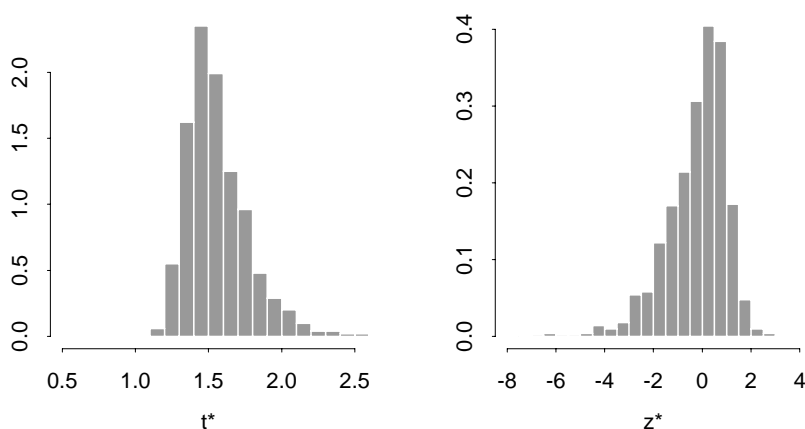


Figure 2.5 City population data. Histograms of t^* and z^* under nonparametric resampling for sample of size $n = 10$, $R = 999$ simulations. Note the skewness of both t^* and z^* .

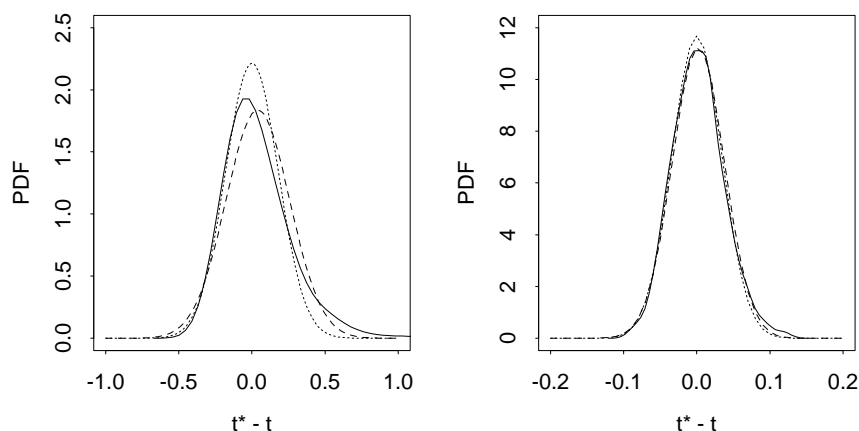
We can use the same simulations to estimate distributions of related statistics, such as transformed estimates or studentized estimates. The right panel of Figure 2.5 shows a corresponding histogram of studentized values $z^* = (t^* - t)/v_L^{*1/2}$, where v_L^* is the nonparametric delta method variance estimate based on a simulated sample. That is,

$$v_L^* = n^{-2} \sum_{j=1}^n \frac{(x_j^* - t^* u_j^*)^2}{\bar{u}^{*2}}.$$

The corresponding theoretical approximation for Z is the $N(0, 1)$ distribution, which we would judge also inaccurate in view of the strong skewness in the histogram. We shall discuss the rationale for the use of z^* in Section 2.4.

One natural question to ask here is what effect the small sample size has on the accuracy of normal approximations. This can be answered in part by plotting density estimates. The left panel of Figure 2.6 shows three estimated densities for $T^* - t$ with our sample of $n = 10$, a kernel density estimate based on our simulations, the $N(b, v)$ approximation with moments computed from the same simulations, and the $N(0, v_L)$ approximation. The right panel shows corresponding density approximations for the full data with $n = 49$; the empirical bias and variance of T are $b = 0.00118$ and $v = 0.001290$, and the delta method variance approximation is $v_L = 0.001166$. At the larger sample size the normal approximations seem very accurate. ■

Figure 2.6 Density estimates for $T^* - t$ based on 999 nonparametric simulations for the city population data. The left panel is for the sample of size $n = 10$ in Table 2.1, and the right panel shows the corresponding estimates for the entire dataset of size $n = 49$. Each plot shows a kernel density estimate (solid), the $N(b, v)$ approximation (dashes), with these moments computed from the same simulations, and the $N(0, v_L)$ approximation (dots).



2.3.1 Comparison with parametric methods

A natural question to ask is how well the nonparametric resampling methods might compare to parametric methods, when the latter are appropriate. Equally important is the question as to which parametric model would produce results like those for nonparametric resampling; this is another way of asking just what the nonparametric bootstrap does. Some insight into these questions can be gained by revisiting Example 1.1.

Example 2.9 (Air-conditioning data) We now look at the results of applying nonparametric resampling to the air-conditioning data. One might expect to obtain results similar to those in Example 2.5, where exponential

resampling was used, since we found in Example 1.1 that the data appear compatible with an exponential model.

Figure 2.7 is the nonparametric analogue of Figure 2.4, and shows quantiles of $T^* - t$. It appears that $R = 500$ or so is needed to get reliable quantile estimates; $R = 100$ is enough for the corresponding plot for bias and variance. Under nonparametric resampling there is no reason why the quantiles should approach the theoretical quantiles under the exponential model, and it seems that they do not do so. This suggestion is confirmed by the Q-Q plots in Figure 2.8. The first panel compares the ordered values of t^* from $R = 999$ nonparametric simulations with theoretical quantiles under the fitted exponential model, and the second panel compares the t^* with theoretical quantiles under the best-fitting gamma model with index $\hat{\kappa} = 0.71$. The agreement in the second panel is strikingly good. On reflection this is natural, because the EDF is closer to the larger gamma model than to the exponential model. ■

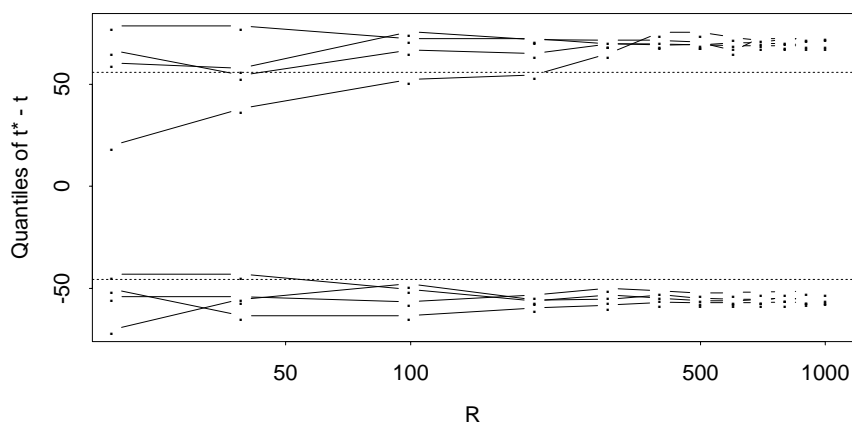
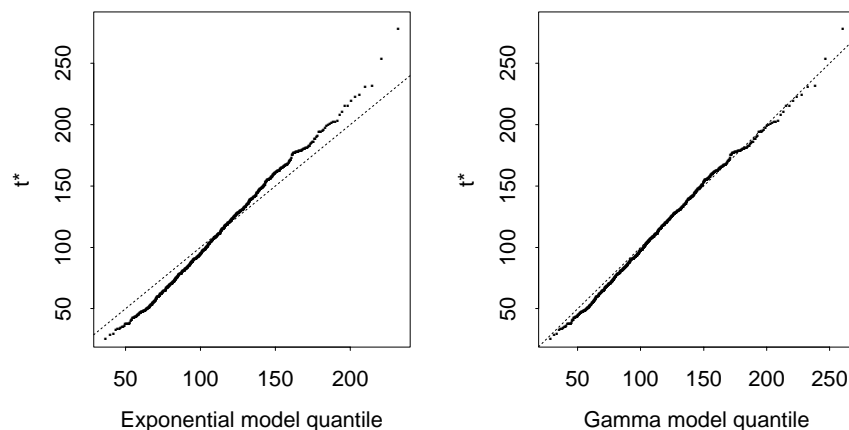


Figure 2.7
Empirical quantiles ($p = 0.05, 0.95$) of $T^* - t$ under nonparametric resampling from the air-conditioning data. The horizontal lines are the exact quantiles based on the fitted exponential model.

2.3.2 Effects of discreteness

For intrinsically continuous data, a major difference between parametric and nonparametric resampling lies in the discreteness of the latter. Under nonparametric resampling, T^* and related quantities will have discrete distributions, even though they may be approximating continuous distributions. This makes results somewhat “fuzzy” compared to their parametric counterparts.

Figure 2.8 Q-Q plots of \bar{y}^* under nonparametric resampling from the air-conditioning data, first against theoretical quantiles under fitted exponential model (left panel) and then against theoretical quantiles under fitted gamma model (right panel).



Example 2.10 (Air-conditioning data) For the nonparametric simulation discussed in the previous example, the right panels of Figure 2.9 show the scatter-plots of sample standard deviation versus sample average for $R = 99$ and $R = 999$ simulated data sets. Corresponding plots for the exponential simulation are shown in the left panels. The qualitative feature to be read from any one of these plots is that data standard deviation is proportional to data average. The discreteness of the nonparametric model (the EDF) adds noise whose peculiar banded structure is evident at $R = 999$, although the qualitative structure is still apparent. ■

For a statistic that is symmetric in the data values, there are up to

$$m_n = \binom{2n-1}{n-1} = \frac{(2n-1)!}{n!(n-1)!}$$

possible values of T^* , depending upon the smoothness of the statistical function $t(\cdot)$. Even for moderately small samples the support of the distribution of T^* will often be fairly dense: values of m_n for $n = 7$ and 11 are $1,716$ and $352,716$ (Problem 2.5). It would therefore usually be harmless to think of there being a PDF for T^* , and to approximate it, either using simulation results as in Figure 2.6 or theoretically (Section 9.5). There are exceptions, however, most notably when T is a sample quantile. The case of the sample median is discussed in Section 2.6.2; see also Problem 2.4 and Example 2.15.

For many practical applications of the simulation results, the effects of discreteness are likely to be fairly minimal. However, one possible problem is that outliers are more likely to occur in the simulation output. For example, in Example 2.8 there were three outliers in the simulation, and these inflated

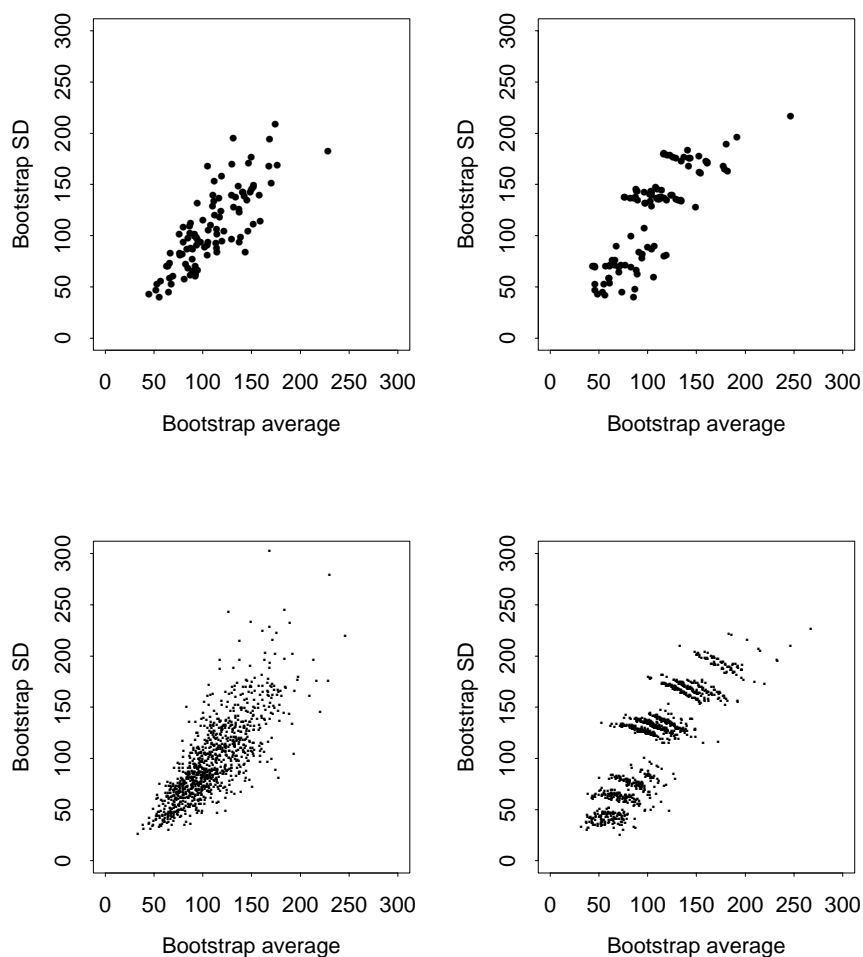


Figure 2.9 Scatter plots of sample standard deviation versus sample average for samples generated by parametric simulation from the fitted exponential model (left panels) and by nonparametric resampling (right panels). Top line is for $R = 99$ and bottom line is for $R = 999$.

the estimate v^* of the variance of T^* . Such outliers should be evident on a normal Q-Q plot (or comparable relevant plot), and when found they should be omitted. More generally, a statistic that depends heavily on a few quantiles can be sensitive to the repeated values that occur under nonparametric sampling, and it can be useful to smooth the original data when dealing with such statistics; see Section 3.4.