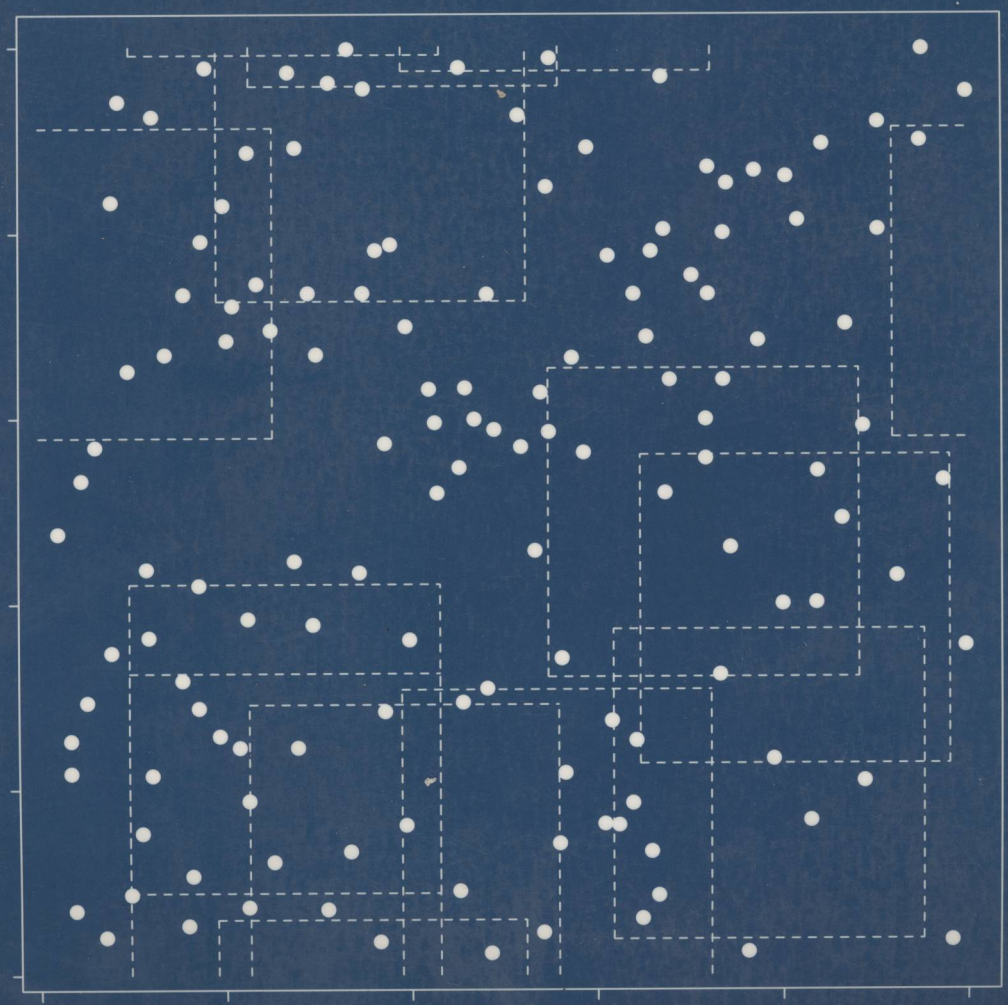
Cambridge Series in Statistical and Probabilistic Mathematics



Bootstrap Methods and their Application

A.C. Davison D.V. Hinkley

Bootstrap methods and their application

A. C. Davison
Professor of Statistics, Department of Mathematics,
Swiss Federal Institute of Technology, Lausanne

D. V. Hinkley
Professor of Statistics, Department of Statistics and Applied Probability,
University of California, Santa Barbara

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge CB2 1RP, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, United Kingdom
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1997

Printed in the United States of America

Typeset in TEX Monotype Times

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data

Davison, A. C. (Anthony Christopher)
Bootstrap methods and their application / A.C. Davison,
D.V. Hinkley.
p. cm.
Includes bibliographical references and index.
ISBN 0 521 57391 2 (hb). ISBN 0 521 57471 4 (pb)
1. Bootstrap (Statistics) I. Hinkley, D. V. II. Title.
QA276.8.D38 1997
519.5'44–dc21 96-30064 CIP

ISBN 0 521 57391 2 hardback
ISBN 0 521 57471 4 paperback

Bootstrap methods and their application

Cambridge Series on Statistical and Probabilistic Mathematics

This series of high quality upper-division textbooks and expository monographs covers all areas of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, mathematical programming, and optimzation. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books contain important applications and discussions of new techniques made possible be advances in computational methods.

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 to 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 to 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 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, M. P. Jones, 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://dmawww.epfl.ch/davison.mosaic/BMA/`

The book could not have been completed without grants from the UK 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 US 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. Particular thanks are due to Claire Cullen Davison for keeping the Davison family going during the writing of this book.

A. C. Davison and D. V. Hinkley
Lausanne and Santa Barbara
May 1997

Contents

1

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 was 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 the 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 in

Diagnosis period		Reporting delay interval (quarters):									Total reports to end
Year	Quarter	0†	1	2	3	4	5	6	···	≥14	of 1992
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.

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 datasets, from

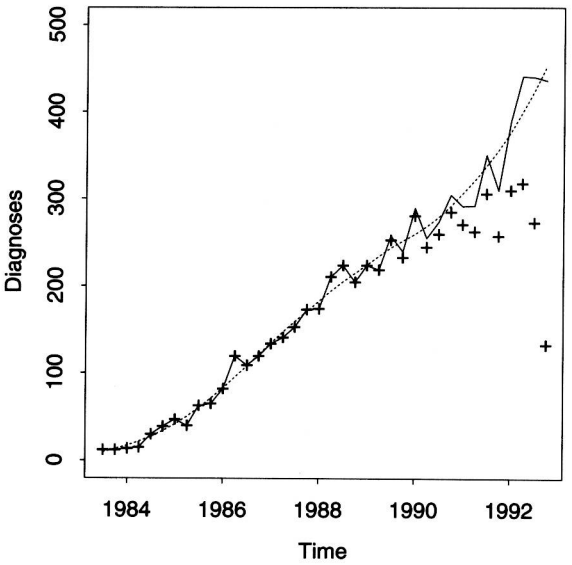


Figure 1.1 Predicted quarterly diagnoses from a parametric model (solid) and a semiparametric model (dots) fitted to the AIDS data, together with the actual totals to the end of 1992 (+).

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 Munchausen, 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 oversimplify 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.

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).

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.

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

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

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 $n^{-1} = 0.08\dot{3}$ 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

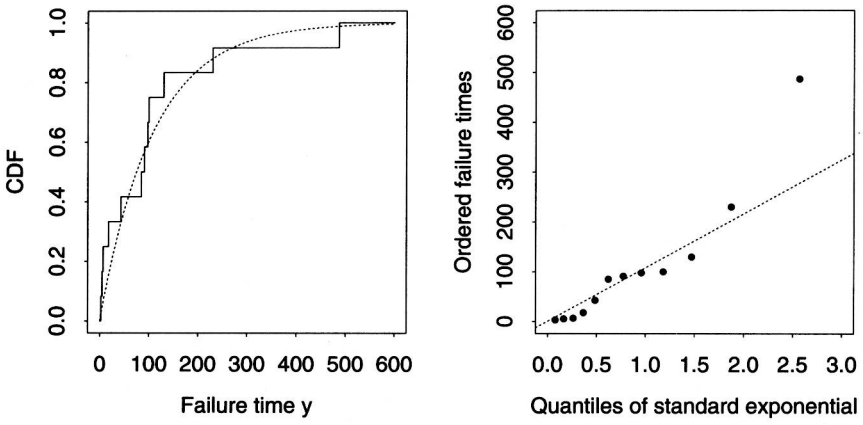


Figure 1.2 Summary displays for the air-conditioning data. The left panel shows the EDF for the data, $\hat{F}$ (solid), and the CDF of a fitted exponential distribution (dots). The right panel shows a plot of the ordered failure times against exponential quantiles, with the fitted exponential model shown as the dotted line.

Q-Q plot — a plot of ordered data values $y_{(j)}$ against the standard exponential quantiles

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

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

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

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: we can tell this because $\bar{Y}/\mu$ 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 city in the United States of America, 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 USA 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 $\theta = \mathrm{E}(X)/\mathrm{E}(U)$. 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 disribution. 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.

Layout of the Book

Chapter 2 describes the properties of resampling methods for use with single samples from parametric and nonparametric models, discusses practical matters such as the numbers of replicate datasets required, and outlines delta methods for variance approximation based on different forms of jackknife. It