

Jun Shao Dongsheng Tu

The Jackknife and Bootstrap



Springer

Jun Shao
Department of Statistics
University of Wisconsin, Madison
1210 West Dayton Street
Madison, WI 53706-1685 USA

Dongsheng Tu
Institute of System Science
Academia Sinica
Beijing, 100080
People's Republic of China

With 4 figures.

Library of Congress Cataloging-in-Publication Data
Shao, Jun.

The jackknife and bootstrap / Jun Shao, Dongsheng Tu.
p. cm. — (Springer series in statistics)
Includes bibliographical references and index.
ISBN 978-1-4612-6903-8 ISBN 978-1-4612-0795-5 (eBook)
DOI 10.1007/978-1-4612-0795-5
1. Jackknife (Statistics). 2. Bootstrap (Statistics).
3. Resampling (Statistics). 4. Estimation theory. I. Tu,
Dongsheng. II. Title. III. Series.
QA276.6.S46 1995
519.5'44—dc20

95-15074

Printed on acid-free paper.

© 1995 Springer Science+Business Media New York
Originally published by Springer-Verlag New York, Inc in 1995
Softcover reprint of the hardcover 1st edition 1995

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher Springer Science+Business Media, LLC,
except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use of general descriptive names, trade names, trademarks, etc., in this publication, even if the former are not especially identified, is not to be taken as a sign that such names, as understood by the Trade Marks and Merchandise Marks Act, may accordingly be used freely by anyone.

Production managed by Hal Henglein; manufacturing supervised by Joe Quatela.
Photocomposed pages prepared from the authors' LaTeX file.

9 8 7 6 5 4 3 2 (Second corrected printing, 1996)

ISBN 978-1-4612-6903-8

SPIN 10544454

To Guang and Shurong

Preface

The jackknife and bootstrap are the most popular data-resampling methods used in statistical analysis. The resampling methods replace theoretical derivations required in applying traditional methods (such as substitution and linearization) in statistical analysis by repeatedly resampling the original data and making inferences from the resamples. Because of the availability of inexpensive and fast computing, these computer-intensive methods have caught on very rapidly in recent years and are particularly appreciated by applied statisticians.

The primary aims of this book are

- (1) to provide a systematic introduction to the theory of the jackknife, the bootstrap, and other resampling methods developed in the last twenty years;
- (2) to provide a guide for applied statisticians: practitioners often use (or misuse) the resampling methods in situations where no theoretical confirmation has been made; and
- (3) to stimulate the use of the jackknife and bootstrap and further developments of the resampling methods.

The theoretical properties of the jackknife and bootstrap methods are studied in this book in an asymptotic framework. Theorems are illustrated by examples. Finite sample properties of the jackknife and bootstrap are mostly investigated by examples and/or empirical simulation studies. In addition to the theory for the jackknife and bootstrap methods in problems with independent and identically distributed (i.i.d.) data, we try to cover, as much as we can, the applications of the jackknife and bootstrap in various complicated non-i.i.d. data problems.

Chapter 1 introduces some basic ideas and motivations for using the jackknife and bootstrap. It also describes the scope of our studies in this book. Chapters 2 and 3 contain general theory for the jackknife and the bootstrap, respectively, which sets up theoretical fundamentals. Some technical tools are introduced and discussed in these two chapters for readers

interested in theoretical studies. Beginning with Chapter 4, each chapter focuses on an important topic concerning the application of the jackknife, the bootstrap, and other related methods. Chapter 4 studies bootstrap confidence sets in some depth: we consider various bootstrap confidence sets developed in recent years and provide asymptotic and empirical comparisons. Bootstrap hypothesis testing is also studied in Chapter 4. Chapter 5 discusses some computational aspects of the jackknife and bootstrap methods. Chapter 6 considers sample survey problems, one of the non-i.i.d. data problems in which the use of the resampling methods (e.g., the jackknife and balanced repeated replication) has a long history. Chapter 7 focuses on applications of the jackknife and bootstrap to linear models, one of the most useful models in statistical applications. Chapter 8 contains some recent developments of the jackknife and bootstrap in various other important statistical fields such as nonlinear regression, generalized linear models, Cox's regression, nonparametric density estimation, nonparametric regression, and multivariate analysis. Applications of the jackknife and bootstrap for dependent data (time series) are studied in Chapter 9. The last chapter introduces two resampling methods that are generalizations of the bootstrap, namely the Bayesian bootstrap and the random weighting. Except for the first chapter, each chapter ends with conclusions and discussions.

Some useful asymptotic results that are often cited in this book are provided in Appendix A. A list of notation is given in Appendix B.

Some knowledge of mathematical statistics (with a standard textbook such as Bickel and Doksum, 1977) is assumed. The reader should be familiar with concepts such as probability, distribution, expectation, estimators, bias, variance, confidence sets, and hypothesis tests. For reading Chapters 6–9, some knowledge of the fields under consideration is required: sample surveys for Chapter 6; linear models for Chapter 7; nonlinear models, generalized linear models, nonparametric regression, and multivariate analysis for the respective sections of Chapter 8; and time series for Chapter 9. Some knowledge of prior and posterior distributions in Bayesian analysis is needed for reading Chapter 10. The mathematical level of the book is flexible: a practical user with a knowledge of calculus and a notion of vectors and matrices can understand all of the basic ideas, discussions, and recommendations in the book by skipping the derivations and proofs (we actually omitted some difficult proofs); with a knowledge of advanced calculus, matrix algebra, and basic asymptotic tools in mathematical statistics (Appendix A and Chapter 1 of Serfling, 1980), one can fully understand the derivations and most of the proofs. A few places (e.g., Section 2.2 and part of Sections 2.4 and 3.3) involving more advanced mathematics (such as real analysis) can be skipped without affecting the reading of the rest of the book.

The Edgeworth and Cornish-Fisher expansions are very important tools

in studying the accuracy of the bootstrap distribution estimators and bootstrap confidence sets. However, the derivations and rigorous proofs of these expansions involve difficult mathematics, which can be found in a recent book by Hall (1992d) (some special cases can be found in Appendix A). Thus, we only state these expansions (with the required regularity conditions) when they are needed, without providing detailed proofs. This does not affect the understanding of our discussions.

Although conceived primarily as a research monograph, the book is suitable for a second-year graduate level course or a research seminar. The following are outlines for various possible one-semester courses.

(I) AN INTRODUCTION TO JACKKNIFE AND BOOTSTRAP (non-theoretical)

Chapter 1, Sections 2.1 and 2.3, Sections 3.4 and 3.5, Chapter 4 (except Sections 4.2 and 4.3), and Chapter 5. Technical proofs should be skipped. If there is time, include Sections 6.1, 6.2, 6.3, 7.1, 7.2, and 7.3.

(II) JACKKNIFE AND BOOTSTRAP FOR COMPLEX DATA (non-theoretical)

Chapter 1, Chapter 6 (except Section 6.4), Chapter 7 (except Section 7.5), Chapter 8, and Chapter 9. Technical proofs should be skipped. If there is time, include some materials from Chapter 5 or Section 10.1.

(III) THEORY OF JACKKNIFE AND BOOTSTRAP (theoretical)

Chapters 1–5. If there is time, include some materials from Chapters 6 and 7 or Chapter 10.

(IV) JACKKNIFE AND BOOTSTRAP FOR COMPLEX DATA (theoretical)

Chapters 1, 6–9. If there is time, include Chapter 5 or Chapter 10.

Of course, one may combine either (I)–(II) or (III)–(IV) for a two-semester course.

This book is essentially based on the authors' lecture notes for graduate level courses taught at Purdue University in 1988, at the University of Waterloo in 1989, and at the University of Ottawa in 1991 and 1992. We are very grateful to students and colleagues who provided helpful comments. Special thanks are due to C. R. Rao, who provided some critical comments on a preliminary draft of this book; J. N. K. Rao, who read part of the manuscript and provided useful suggestions; and anonymous referees and Springer-Verlag Production and Copy Editors, who helped to improve the presentation. We also would like to express our appreciation to the National Sciences and Engineering Research Council of Canada for support during the writing of the book.

Madison and Ottawa
May, 1995

J. Shao
D. Tu

Contents

Preface	vii
Chapter 1. Introduction	1
1.1 Statistics and Their Sampling Distributions	1
1.2 The Traditional Approach	2
1.3 The Jackknife	4
1.4 The Bootstrap	9
1.5 Extensions to Complex Problems	17
1.6 Scope of Our Studies	19
Chapter 2. Theory for the Jackknife	23
2.1 Variance Estimation for Functions of Means	23
2.1.1 Consistency	24
2.1.2 Other properties	28
2.1.3 Discussions and examples	29
2.2 Variance Estimation for Functionals	32
2.2.1 Differentiability and consistency	33
2.2.2 Examples	37
2.2.3 Convergence rate	42
2.2.4 Other differential approaches	44
2.3 The Delete-d Jackknife	49
2.3.1 Variance estimation	50
2.3.2 Jackknife histograms	55
2.4 Other Applications	60
2.4.1 Bias estimation	61
2.4.2 Bias reduction	64

2.4.3	Miscellaneous results	68
2.5	Conclusions and Discussions	69
Chapter 3. Theory for the Bootstrap		71
3.1	Techniques in Proving Consistency	72
3.1.1	Bootstrap distribution estimators	72
3.1.2	Mallows' distance	73
3.1.3	Berry-Esséen's inequality	74
3.1.4	Imitation	76
3.1.5	Linearization	78
3.1.6	Convergence in moments	79
3.2	Consistency: Some Major Results	80
3.2.1	Distribution estimators	80
3.2.2	Variance estimators	86
3.3	Accuracy and Asymptotic Comparisons	91
3.3.1	Convergence rate	91
3.3.2	Asymptotic minimaxity	97
3.3.3	Asymptotic mean squared error	99
3.3.4	Asymptotic relative error	102
3.3.5	Conclusions	104
3.4	Fixed Sample Performance	104
3.4.1	Moment estimators	105
3.4.2	Distribution estimators	108
3.4.3	Conclusions	112
3.5	Smoothed Bootstrap	113
3.5.1	Empirical evidences and examples	113
3.5.2	Sample quantiles	116
3.5.3	Remarks	117
3.6	Nonregular Cases	118
3.7	Conclusions and Discussions	127
Chapter 4. Bootstrap Confidence Sets and Hypothesis Tests		129
4.1	Bootstrap Confidence Sets	129
4.1.1	The bootstrap-t	131
4.1.2	The bootstrap percentile	132

4.1.3	The bootstrap bias-corrected percentile	133
4.1.4	The bootstrap accelerated bias-corrected percentile .	135
4.1.5	The hybrid bootstrap	140
4.2	Asymptotic Theory	141
4.2.1	Consistency	141
4.2.2	Accuracy	144
4.2.3	Other asymptotic comparisons	152
4.3	The Iterative Bootstrap and Other Methods	155
4.3.1	The iterative bootstrap	155
4.3.2	Bootstrap calibrating	160
4.3.3	The automatic percentile and variance stabilizing .	161
4.3.4	Fixed width bootstrap confidence intervals	164
4.3.5	Likelihood based bootstrap confidence sets	165
4.4	Empirical Comparisons	166
4.4.1	The bootstrap-t, percentile, BC, and BC _a	166
4.4.2	The bootstrap and other asymptotic methods	170
4.4.3	The iterative bootstrap and bootstrap calibration .	173
4.4.4	Summary	176
4.5	Bootstrap Hypothesis Tests	176
4.5.1	General description	177
4.5.2	Two-sided hypotheses with nuisance parameters .	179
4.5.3	Bootstrap distance tests	182
4.5.4	Other results and discussions	184
4.6	Conclusions and Discussions	188
Chapter 5. Computational Methods		190
5.1	The Delete-1 Jackknife	190
5.1.1	The one-step jackknife	191
5.1.2	Grouping and random subsampling	195
5.2	The Delete-d Jackknife	197
5.2.1	Balanced subsampling	197
5.2.2	Random subsampling	198
5.3	Analytic Approaches for the Bootstrap	200
5.3.1	The delta method	201
5.3.2	Jackknife approximations	202

5.3.3	Saddle point approximations	203
5.3.4	Remarks	205
5.4	Simulation Approaches for the Bootstrap	206
5.4.1	The simple Monte Carlo method	207
5.4.2	Balanced bootstrap resampling	211
5.4.3	Centering after Monte Carlo	215
5.4.4	The linear bootstrap	219
5.4.5	Antithetic bootstrap resampling	221
5.4.6	Importance bootstrap resampling	223
5.4.7	The one-step bootstrap	228
5.5	Conclusions and Discussions	230
Chapter 6. Applications to Sample Surveys		232
6.1	Sampling Designs and Estimates	232
6.2	Resampling Methods	238
6.2.1	The jackknife	238
6.2.2	The balanced repeated replication	241
6.2.3	Approximated BRR methods	244
6.2.4	The bootstrap	246
6.3	Comparisons by Simulation	251
6.4	Asymptotic Results	258
6.4.1	Assumptions	258
6.4.2	The jackknife and BRR for functions of averages .	260
6.4.3	The RGBRR and RSBRR for functions of averages .	264
6.4.4	The bootstrap for functions of averages	267
6.4.5	The BRR and bootstrap for sample quantiles	268
6.5	Resampling Under Imputation	270
6.5.1	Hot deck imputation	271
6.5.2	An adjusted jackknife	273
6.5.3	Multiple bootstrap hot deck imputation	277
6.5.4	Bootstrapping under imputation	278
6.6	Conclusions and Discussions	281
Chapter 7. Applications to Linear Models		283
7.1	Linear Models and Regression Estimates	283
7.2	Variance and Bias Estimation	285

7.2.1	Weighted and unweighted jackknives	285
7.2.2	Three types of bootstraps	289
7.2.3	Robustness and efficiency	292
7.3	Inference and Prediction Using the Bootstrap	295
7.3.1	Confidence sets	295
7.3.2	Simultaneous confidence intervals	298
7.3.3	Hypothesis tests	301
7.3.4	Prediction	303
7.4	Model Selection	306
7.4.1	Cross-validation	307
7.4.2	The bootstrap	311
7.5	Asymptotic Theory	313
7.5.1	Variance estimators	313
7.5.2	Bias estimators	318
7.5.3	Bootstrap distribution estimators	320
7.5.4	Inference and prediction	324
7.5.5	Model selection	326
7.6	Conclusions and Discussions	329
Chapter 8. Applications to Nonlinear, Nonparametric, and Multivariate Models		331
8.1	Nonlinear Regression	331
8.1.1	Jackknife variance estimators	333
8.1.2	Bootstrap distributions and confidence sets	335
8.1.3	Cross-validation for model selection	337
8.2	Generalized Linear Models	338
8.2.1	Jackknife variance estimators	340
8.2.2	Bootstrap procedures	341
8.2.3	Model selection by bootstrapping	343
8.3	Cox's Regression Models	345
8.3.1	Jackknife variance estimators	346
8.3.2	Bootstrap procedures	349
8.4	Kernel Density Estimation	350
8.4.1	Bandwidth selection by cross-validation	351
8.4.2	Bandwidth selection by bootstrapping	353
8.4.3	Bootstrap confidence sets	356

8.5	Nonparametric Regression	360
8.5.1	Kernel estimates for fixed design	360
8.5.2	Kernel estimates for random regressor	364
8.5.3	Nearest neighbor estimates	366
8.5.4	Smoothing splines	370
8.6	Multivariate Analysis	373
8.6.1	Analysis of covariance matrix	373
8.6.2	Multivariate linear models	376
8.6.3	Discriminant analysis	379
8.6.4	Factor analysis and clustering	382
8.7	Conclusions and Discussions	384
Chapter 9. Applications to Time Series and Other Dependent Data		386
9.1	m-Dependent Data	387
9.2	Markov Chains	392
9.3	Autoregressive Time Series	394
9.3.1	Bootstrapping residuals	395
9.3.2	Model selection	397
9.4	Other Time Series	400
9.4.1	ARMA(p, q) models	401
9.4.2	Linear regression with time series errors	403
9.4.3	Dynamical linear regression	406
9.5	Stationary Processes	407
9.5.1	Moving block and circular block	407
9.5.2	Consistency of the bootstrap	410
9.5.3	Accuracy of the bootstrap	411
9.5.4	Remarks	413
9.6	Conclusions and Discussions	414
Chapter 10. Bayesian Bootstrap and Random Weighting		416
10.1	Bayesian Bootstrap	416
10.1.1	Bayesian bootstrap with a noninformative prior	417
10.1.2	Bayesian bootstrap using prior information	420
10.1.3	The weighted likelihood bootstrap	422
10.1.4	Some remarks	424

10.2 Random Weighting	425
10.2.1 Motivation	425
10.2.2 Consistency	427
10.2.3 Asymptotic accuracy	429
10.3 Random Weighting for Functionals and Linear Models	434
10.3.1 Statistical functionals	434
10.3.2 Linear models	437
10.4 Empirical Results for Random Weighting	440
10.5 Conclusions and Discussions	445
Appendix A. Asymptotic Results	447
A.1 Modes of Convergence	447
A.2 Convergence of Transformations	448
A.3 $O(\cdot)$, $o(\cdot)$, and Stochastic $O(\cdot)$, $o(\cdot)$	448
A.4 The Borel-Cantelli Lemma	449
A.5 The Law of Large Numbers	449
A.6 The Law of the Iterated Logarithm	450
A.7 Uniform Integrability	450
A.8 The Central Limit Theorem	451
A.9 The Berry-Esséen Theorem	451
A.10 Edgeworth Expansions	452
A.11 Cornish-Fisher Expansions	454
Appendix B. Notation	455
References	457
Author Index	493
Subject Index	499

Chapter 1

Introduction

1.1 Statistics and Their Sampling Distributions

The basic objective of statistical analysis is “extracting all the information from the data” (Rao, 1989) to deduce properties of the population that generated the data. Statistical analyses are generally based on *statistics*, which are functions of data and are selected according to some principle (e.g., the likelihood principle, the substitution principle, sufficiency, and robustness). For example, the sample mean is an estimate of the population center; studentized statistics are used for constructing confidence sets and testing hypotheses in statistical inference.

Prior to data collection, a statistic is a random quantity having a probability distribution, called the *sampling distribution* of the statistic. Most statistical procedures require some knowledge of the sampling distribution of the statistic being used for analysis. The type of knowledge needed depends on the type of analysis. Constructing confidence sets and testing hypotheses, for example, require the knowledge of the sampling distribution itself or of the percentiles of the sampling distribution. On the other hand, in an estimation problem, it is essential to have an indication of the estimator’s accuracy, since any estimator may have an estimation error. The knowledge of accuracy measures such as the variance, bias, and mean squared error of the estimator is therefore required in this case. These accuracy measures are characteristics of the estimator’s sampling distribution. Accuracy measures can also be used to select the best estimator from a class of appropriate estimators.

The sampling distribution of a statistic and its characteristics usually depend on the underlying population and therefore are unknown. They have to be estimated or approximated from the observed data in most

estimation or inference problems. For selecting an estimator from a given class of estimators, sometimes we do not need to estimate or approximate the accuracy measure used to assess estimators; for example, there may exist an estimator that is the most accurate of all the estimators in the given class, regardless of the underlying population [e.g., the uniformly minimum variance unbiased estimator; see Rao (1973, Section 5a) or Lehmann (1983, Section 2.1)]. In most situations, however, the relative accuracy of the estimators does depend on the underlying population, and we have to use the data to estimate the relative accuracy for selecting an estimator.

The *jackknife* and the *bootstrap* are two methods for estimating or approximating the sampling distribution of a statistic and its characteristics.

1.2 The Traditional Approach

As was described in the previous section, a crucial step in statistical analysis is to use the data to approximate or estimate some accuracy measures of a given statistic (estimator), such as the bias, the variance, and the mean squared error. Before we introduce the jackknife and the bootstrap, let us glance over the traditional approach in estimating accuracy measures.

In the traditional approach, an accuracy measure is estimated by an empirical analog of an explicit theoretical formula of the accuracy measure (or its approximation), which is derived from a postulated model. Let us use the variance as an illustration. Let X_1, \dots, X_n denote the data set of n independent and identically distributed (i.i.d.) observations from an unknown distribution F and let $T_n = T_n(X_1, \dots, X_n)$ be a given statistic. Then the variance of T_n is

$$\text{var}(T_n) = \int \left[T_n(x) - \int T_n(y) d \prod_{i=1}^n F(y_i) \right]^2 d \prod_{i=1}^n F(x_i), \quad (1.1)$$

where $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$. When T_n is simple, we can obtain an explicit expression of $\text{var}(T_n)$ as a function of some unknown quantities and then estimate $\text{var}(T_n)$ by substituting the unknown quantities with their estimates.

Example 1.1. Simple functions of the sample mean. When $T_n = \bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ is the sample mean, $\text{var}(\bar{X}_n) = n^{-1} \text{var}(X_1)$. Thus, we can estimate $\text{var}(\bar{X}_n)$ by estimating $\text{var}(X_1)$, a single unknown parameter. If F does not belong to a parametric family, then $\text{var}(X_1)$ is usually estimated by the sample variance $(n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$.

When T_n is a *very simple* function of \bar{X}_n , a not too complicated formula

of $\text{var}(T_n)$ can be obtained. For example, if $T_n = \bar{X}_n^2$, then

$$\text{var}(\bar{X}_n^2) = \frac{4\mu^2\alpha_2}{n} + \frac{4\mu\alpha_3}{n^2} + \frac{\alpha_4}{n^3}, \quad (1.2)$$

where $\mu = EX_1$ and $\alpha_k = E(X_1 - \mu)^k$ is the k th central moment of X_1 . We can then estimate $\text{var}(\bar{X}_n^2)$ by substituting μ and α_k with their estimators \bar{X}_n and

$$\hat{\alpha}_k = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^k, \quad k = 2, 3, 4. \quad (1.3)$$

However, there are not many statistics as simple as the sample mean. For most statistics, the expression in (1.1) is too complicated to be useful in obtaining estimators of $\text{var}(T_n)$, and it is very hard or impossible to obtain an exact and explicit formula for $\text{var}(T_n)$. Thus, in the traditional approach, we try to simplify the problem by considering approximations or asymptotic expansions of $\text{var}(T_n)$. Under some regularity conditions, we can often establish

$$\lim_{n \rightarrow \infty} n[\text{var}(T_n)] = \sigma_F^2, \quad (1.4)$$

where $\sigma_F^2 = \sigma^2(F)$ is a simple function of F or $\sigma_F^2 = \sigma^2(\gamma)$ for a vector γ of unknown parameters. We then estimate $\text{var}(T_n)$ by the empirical analog of the approximation, i.e., $\sigma^2(\hat{F})/n$ or $\sigma^2(\hat{\gamma})/n$, where \hat{F} and $\hat{\gamma}$ are estimators of F and γ , respectively.

Example 1.1 (continued). From (1.2), (1.4) holds in the case of $T_n = \bar{X}_n^2$ with $\sigma_F^2 = 4\mu^2\alpha_2$. This simply says that the expression in (1.2) can be approximated by a simpler form which is the first order term on the right-hand side of (1.2). Using (1.4) to obtain an estimate of $\text{var}(\bar{X}_n^2)$, we do not need to estimate α_3 and α_4 .

Example 1.2. Trimmed sample mean. Consider the α -trimmed sample mean $\bar{X}_n^{(\alpha)}$ in robust estimation of the center of a symmetric F :

$$\bar{X}_n^{(\alpha)} = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{(i)},$$

where $[t]$ is the integer part of t and $X_{(i)}$ is the i th order statistic. An exact and explicit form of $\text{var}(\bar{X}_n^{(\alpha)})$ does not exist, but we can show that (1.4) holds with

$$\sigma_F^2 = 2(1 - 2\alpha)^{-2} \left[\int_0^{F^{-1}(1-\alpha)} x^2 dF(x) + \alpha F^{-1}(1 - \alpha) \right],$$

where

$$F^{-1}(t) = \inf\{x : F(x) \geq t\}. \quad (1.5)$$

Therefore, we can estimate $\text{var}(\bar{X}_n^{(\alpha)})$ by $\sigma_{\hat{F}}^2$, where \hat{F} is an estimator of F , e.g., \hat{F} is the *empirical distribution* F_n defined by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I\{X_i \leq x\}, \quad (1.6)$$

where $I\{A\}$ is the indicator function of the set A .

Here are some of the weaknesses and disadvantages of the traditional approach that have been discovered over the years:

- (1) Sometimes it requires a very large sample size n in order to have accurate variance (or other accuracy measure) estimators, because of the use of an approximate formula such as the σ_F^2 in (1.4).
- (2) The theoretical formula or its approximation is based on the postulated model. When the model is *slightly* wrong, the obtained estimator of accuracy may not be valid any more.
- (3) To apply the traditional approach to many different problems, we have to derive a theoretical formula such as σ_F^2 in (1.4) for *each* problem. These derivations may be difficult or tedious. Furthermore, the derivation of the theoretical formula requires that the data analysts have a good knowledge of mathematics and theoretical statistics.
- (4) Sometimes the derivation of the theoretical formula or its approximation is very difficult or even impossible. This has been experienced since statistical methods based on empirical processes were introduced. The limiting variances of statistics that are functionals of empirical processes are functionals of Gaussian processes. Expressing them as simple functions of model parameters is very hard.
- (5) The theoretical formula may be too complicated to be useful in estimating the accuracy measure.

These issues will be addressed further for particular problems in later chapters.

1.3 The Jackknife

Quenouille (1949) introduced a method, later named the *jackknife*, to estimate the bias of an estimator by deleting one datum each time from the original data set and recalculating the estimator based on the rest of the data. Let $T_n = T_n(X_1, \dots, X_n)$ be an estimator of an unknown parameter

θ . The bias of T_n is defined as

$$\text{bias}(T_n) = E(T_n) - \theta. \quad (1.7)$$

Let $T_{n-1,i} = T_{n-1}(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ be the given statistic but based on $n-1$ observations $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$, $i = 1, \dots, n$. Quenouille's *jackknife bias estimator* is

$$b_{\text{JACK}} = (n-1)(\bar{T}_n - T_n), \quad (1.8)$$

where $\bar{T}_n = n^{-1} \sum_{i=1}^n T_{n-1,i}$. This leads to a bias-reduced *jackknife estimator* of θ ,

$$T_{\text{JACK}} = T_n - b_{\text{JACK}} = nT_n - (n-1)\bar{T}_n. \quad (1.9)$$

The jackknife estimators b_{JACK} and T_{JACK} can be heuristically justified as follows. Suppose that

$$\text{bias}(T_n) = \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right) \quad (1.10)$$

(see Appendix A.3 for the notation O and o), where a and b are unknown but do not depend on n . Since $T_{n-1,i}$, $i = 1, \dots, n$, are identically distributed,

$$\text{bias}(T_{n-1,i}) = \frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{(n-1)^3}\right), \quad (1.11)$$

and $\text{bias}(\bar{T}_n)$ has the same expression as that in (1.11). Therefore,

$$\begin{aligned} E(b_{\text{JACK}}) &= (n-1)[\text{bias}(\bar{T}_n) - \text{bias}(T_n)] \\ &= (n-1)\left[\left(\frac{1}{n-1} - \frac{1}{n}\right)a + \left(\frac{1}{(n-1)^2} - \frac{1}{n^2}\right)b + O\left(\frac{1}{n^3}\right)\right] \\ &= \frac{a}{n} + \frac{(2n-1)b}{n^2(n-1)} + O\left(\frac{1}{n^2}\right), \end{aligned}$$

which means that as an estimator of the bias of T_n , b_{JACK} is correct (unbiased) up to the order n^{-2} . It follows that

$$\text{bias}(T_{\text{JACK}}) = \text{bias}(T_n) - E(b_{\text{JACK}}) = -\frac{b}{n(n-1)} + O\left(\frac{1}{n^2}\right),$$

i.e., the bias of T_{JACK} is of the order n^{-2} . The jackknife produces a bias-reduced estimator by removing the first order term in $\text{bias}(T_n)$. This property of the jackknife will be studied more rigorously in Chapter 2.

The jackknife has become a more valuable tool since Tukey (1958) found that the jackknife can also be used to construct variance estimators. A

heuristic justification for using the jackknife in variance estimation is given by Tukey (1958). Note that T_{JACK} can be written as

$$T_{\text{JACK}} = \frac{1}{n} \sum_{i=1}^n [nT_n - (n-1)T_{n-1,i}].$$

Tukey defined

$$\tilde{T}_{n,i} = nT_n - (n-1)T_{n-1,i}, \quad i = 1, \dots, n, \quad (1.12)$$

as the jackknife pseudovalues and conjectured that

- (A) The pseudovalues $\tilde{T}_{n,i}$, $i = 1, \dots, n$, may be treated as though they were i.i.d.;
- (B) $\tilde{T}_{n,i}$ has approximately the same variance as $\sqrt{n}T_n$.

Under (A) and (B), it is natural to estimate $\text{var}(\sqrt{n}T_n)$ by the sample variance based on $\tilde{T}_{n,1}, \dots, \tilde{T}_{n,n}$, i.e., to estimate $\text{var}(T_n)$ by

$$\begin{aligned} v_{\text{JACK}} &= \frac{1}{n(n-1)} \sum_{i=1}^n \left(\tilde{T}_{n,i} - \frac{1}{n} \sum_{j=1}^n \tilde{T}_{n,j} \right)^2 \\ &= \frac{n-1}{n} \sum_{i=1}^n \left(T_{n-1,i} - \frac{1}{n} \sum_{j=1}^n T_{n-1,j} \right)^2. \end{aligned} \quad (1.13)$$

This estimator is the well-known (delete-1) *jackknife variance estimator* for T_n .

The jackknife is less dependent on model assumptions and does not need the theoretical formula required by the traditional approach. However, the jackknife requires repeatedly computing the statistic n times, which was virtually impossible in the old days. The invention of electronic computers now has made it possible for us to use the jackknife method. Nowadays, the jackknife, as its name indicates, has become a popular and useful tool in statistical analysis. Many agencies (e.g., Statistics Canada) have computer software to implement the computation of the jackknife estimators.

We now study some examples.

Example 1.1 (continued). For the simple case of $T_n = \bar{X}_n$, T_n is unbiased for μ . It is easy to show that

$$T_{n-1,i} = \bar{X}_{n-1,i} = (n\bar{X}_n - X_i)/(n-1) \quad (1.14)$$

and, therefore, $\bar{T}_n = \bar{X}_n$, $b_{\text{JACK}} = 0$ and $T_{\text{JACK}} = T_n = \bar{X}_n$. Also, in this case, $\tilde{T}_{n,i}$ in (1.12) is exactly X_i , whose variance is equal to $\text{var}(\sqrt{n}\bar{X}_n)$.

Hence, Tukey's conjectures (A) and (B) are true. In this case, v_{JACK} reduces to $[n(n-1)]^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$, and the jackknife and the traditional approach produce the same variance estimator.

Next, we consider the case of $T_n = \bar{X}_n^2$ as an estimator of μ^2 . Note that $\text{bias}(\bar{X}_n^2) = \alpha_2/n$. Using (1.14), we obtain the jackknife bias estimator

$$\begin{aligned} b_{\text{JACK}} &= \frac{n-1}{n} \sum_{i=1}^n (\bar{X}_{n-1,i}^2 - \bar{X}_n^2) \\ &= \frac{n-1}{n} \sum_{i=1}^n (\bar{X}_{n-1,i} - \bar{X}_n)(\bar{X}_{n-1,i} + \bar{X}_n) \\ &= \frac{1}{n(n-1)} \sum_{i=1}^n (\bar{X}_n - X_i)[2(n-1)\bar{X}_n + (\bar{X}_n - X_i)] \\ &= \frac{1}{n(n-1)} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{\hat{\alpha}_2}{n}, \end{aligned}$$

where $\hat{\alpha}_2$ is defined in (1.3). In this case, the jackknife again agrees with the traditional substitution method: replacing α_2 in $\text{bias}(\bar{X}_n^2)$ by $\hat{\alpha}_2$. The jackknife estimator of μ^2 is

$$T_{\text{JACK}} = \bar{X}_n^2 - \hat{\alpha}_2/n,$$

which is unbiased. The jackknife removes the bias of T_n completely.

A straightforward calculation shows that

$$\begin{aligned} v_{\text{JACK}} &= \frac{n-1}{n} \sum_{i=1}^n (\bar{X}_{n-1,i}^2 - \bar{X}_n^2)^2 - \frac{\hat{\alpha}_2^2}{n^2(n-1)} \\ &= \frac{4\bar{X}_n^2\hat{\alpha}_2}{n} - \frac{4\bar{X}_n\hat{\alpha}_3}{n(n-1)} + \frac{\hat{\alpha}_4}{n(n-1)^2} - \frac{\hat{\alpha}_2^2}{n^2(n-1)}, \quad (1.15) \end{aligned}$$

where $\hat{\alpha}_k$ are given in (1.3). Comparing (1.15) with (1.2), we conclude that the jackknife and the traditional approach (exact or approximate) provide different variance estimators, but these variance estimators have the same first order term $4\bar{X}_n^2\hat{\alpha}_2/n$. This is actually true for many general statistics and will be studied further in Chapter 2.

Example 1.2 (continued). We now study the jackknife estimators for the α -trimmed sample mean $\bar{X}_n^{(\alpha)}$. For simplicity, we assume that $m = (n-1)\alpha$ is an integer. Then,

$$\bar{X}_{n-1,i}^{(\alpha)} = \frac{1}{(n-1)(1-2\alpha)} \sum_{m+1 \leq j \leq n-m, j \neq i} X_{(j)}$$

when $m+1 \leq i \leq n-m$; $\bar{X}_{n-1,i}^{(\alpha)} = \bar{X}_{n-1,m}^{(\alpha)}$ when $i < m+1$; and $\bar{X}_{n-1,i}^{(\alpha)} = \bar{X}_{n-1,n-m+1}^{(\alpha)}$ when $i > n-m$. Consequently,

$$\bar{X}_{n-1,i}^{(\alpha)} - \bar{X}_n^{(\alpha)} = \frac{1}{(n-1)(1-2\alpha)} (\bar{X}_n^{(\alpha)} - X_{(w_i)}),$$

where $w_i = i$ when $m+1 \leq i \leq n-m$, $w_i = m$ when $i < m+1$ and $w_i = n-m+1$ when $i > n-m$. Hence

$$b_{\text{JACK}} = (n-1)(\bar{T}_n - T_n) = \frac{1}{1-2\alpha} (\bar{X}_n^{(\alpha)} - \bar{X}_n^{(w)}),$$

where $\bar{X}_n^{(w)} = n^{-1} \sum_{i=1}^n X_{(w_i)}$ is the Winsorized sample mean, and

$$T_{\text{JACK}} = \frac{1}{1-2\alpha} \bar{X}_n^{(w)} - \frac{2\alpha}{1-2\alpha} \bar{X}_n^{(\alpha)}.$$

The jackknife variance estimator is

$$v_{\text{JACK}} = \frac{1}{n(n-1)(1-2\alpha)^2} \sum_{i=1}^n (X_{(w_i)} - \bar{X}_n^{(w)})^2.$$

Example 1.3. V-statistics of order 2. Consider a V-statistic of order 2,

$$T_n = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n h(X_i, X_j),$$

where h is a function on \mathbb{R}^2 satisfying $h(x, y) = h(y, x)$. A straightforward calculation shows that

$$\begin{aligned} (n-1)\bar{T}_n &= \frac{n-1}{n} \sum_{i=1}^n \frac{1}{(n-1)^2} \sum_{j \neq i} \sum_{k \neq i} h(X_j, X_k) \\ &= \frac{1}{n} \sum_{i=1}^n h(X_i, X_i) + \frac{2(n-2)}{n(n-1)} \sum_{k < j} h(X_j, X_k). \end{aligned}$$

Then

$$b_{\text{JACK}} = \frac{1}{n^2} \sum_{i=1}^n h(X_i, X_i) - \frac{2}{n^2(n-1)} \sum_{k < j} h(X_j, X_k)$$

and

$$T_{\text{JACK}} = \frac{2}{n(n-1)} \sum_{k < j} h(X_j, X_k),$$

which is a U-statistic of order 2. This result will be used later in Chapter 2. It is not easy, however, to obtain an explicit formula for v_{JACK} with a general function h .

Consider a more specific case where $h(x, y) = (x - y)^2/2$. Then

$$T_n = \frac{1}{2n^2} \sum_{i=1}^n \sum_{j=1}^n (X_i - X_j)^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

and

$$T_{\text{JACK}} = \frac{1}{n(n-1)} \sum_{i < j} (X_i - X_j)^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

Hence, the difference between T_n and T_{JACK} is that the former is the sample variance using n in the denominator, whereas the latter is the sample variance using $n-1$ in the denominator. It is well-known that T_{JACK} is unbiased for $\text{var}(X_1)$ and T_n has a bias of order n^{-1} . The jackknife variance estimator v_{JACK} in this case is

$$v_{\text{JACK}} = \frac{n}{(n-1)^2} \hat{\alpha}_4 - \frac{1}{n-1} \hat{\alpha}_2^2,$$

where $\hat{\alpha}_k$ are defined in (1.3).

We have seen in Examples 1.1-1.3 that in some cases the jackknife estimator is an explicit function of X_1, \dots, X_n , and hence the repeated calculation of the given statistic is unnecessary. However, usually this is not the case. The attempt to express jackknife estimators as explicit functions of X_1, \dots, X_n is usually of theoretical interest only. (We did it here for illustration.) It is the property of replacing complicated derivations by repeated computations that makes the jackknife very useful in practice.

1.4 The Bootstrap

A data set of size n has $2^n - 1$ nonempty subsets; the jackknife only utilizes n of them. The jackknife may be improved by using statistics based on more than n , or even all $2^n - 1$, subsets. This idea was discussed by Hartigan (1969), but it requires more computer power than the jackknife.

The development in computer technology in the last two decades was very rapid. The high speed and power of the new generation of computers stimulated the development of new statistical methods that are computer-intensive and more reliable, and have broader applications. The *bootstrap*, introduced by Efron (1979), is one of these methods.

Variance estimation

Suppose that the data X_1, \dots, X_n are i.i.d. from F and that F is estimated by \hat{F} . Substituting \hat{F} for F in (1.1), we obtain the *bootstrap variance*

estimator

$$\begin{aligned} v_{\text{BOOT}} &= \int \left[T_n(x) - \int T_n(y) d \prod_{i=1}^n \hat{F}(y_i) \right]^2 d \prod_{i=1}^n \hat{F}(x_i) \\ &= \text{var}_*[T_n(X_1^*, \dots, X_n^*) | X_1, \dots, X_n], \end{aligned} \quad (1.16)$$

where $\{X_1^*, \dots, X_n^*\}$ is an i.i.d. sample from \hat{F} and is called a *bootstrap sample*, and $\text{var}_*[\cdot | X_1, \dots, X_n]$ denotes the conditional variance for given X_1, \dots, X_n . Equation (1.16) is the *theoretical form* of the bootstrap variance estimator for T_n . It may not be used directly for practical applications when v_{BOOT} is not an explicit function of X_1, \dots, X_n . If v_{BOOT} is an explicit function of X_1, \dots, X_n , then it is actually a substitution estimator of $\text{var}(T_n)$.

Example 1.1 (continued). When $T_n = \bar{X}_n$ and $\hat{F} = F_n$, the empirical distribution defined in (1.6), plugging F_n into the expression

$$\text{var}(\bar{X}_n) = \frac{1}{n} \text{var}(X_1) = \frac{1}{n} \int [x - \int y dF(y)]^2 dF(x)$$

we obtain $v_{\text{BOOT}} = n^{-2} \sum_{i=1}^n (X_i - \bar{X}_n)^2$.

When $T_n = \bar{X}_n^2$, using (1.2), $\mu = \int x dF(x)$, and $\alpha_k = \int (x - \mu)^k dF(x)$, we obtain

$$v_{\text{BOOT}} = \frac{4\bar{X}_n^2 \tilde{\alpha}_2}{n} + \frac{4\bar{X}_n \tilde{\alpha}_3}{n^2} + \frac{\tilde{\alpha}_4}{n^3},$$

where $\tilde{\alpha}_k = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^k$.

Example 1.4. The sample median. Let $T_n = F_n^{-1}(\frac{1}{2})$ be the sample median, where F_n^{-1} is defined by (1.5) with F replaced by F_n in (1.6). Assume that $n = 2m - 1$ for an integer m . Then, $T_n = X_{(m)}$. Using $\hat{F} = F_n$, we can obtain v_{BOOT} as follows. Let X_1^*, \dots, X_n^* be an i.i.d. sample from F_n . Then

$$\begin{aligned} p_k &= P_*\{X_{(m)}^* = X_{(k)} | X_1, \dots, X_n\} \\ &= \sum_{j=0}^{m-1} \binom{n}{j} \frac{(k-1)^j (n-k+1)^{n-j} - k^j (n-k)^{n-j}}{n^n}. \end{aligned} \quad (1.17)$$

Let $F_{X_{(m)}}$ be the distribution of $X_{(m)}$. Since

$$\text{var}(X_{(m)}) = \int [x - \int y dF_{X_{(m)}}(y)]^2 dF_{X_{(m)}}(x),$$

replacing $F_{X_{(m)}}$ by the bootstrap distribution of $X_{(m)}^*$ given by (1.17), we obtain

$$v_{\text{BOOT}} = \sum_{k=1}^n p_k \left(X_{(k)} - \sum_{j=1}^n p_j X_{(j)} \right)^2.$$

This leads to the same estimator given by Maritz and Jarrett (1978).

However, as we discussed earlier, the expression in (1.1) or (1.16) is usually complicated and v_{BOOT} is not an explicit function of X_1, \dots, X_n .

When the right-hand side of (1.1) is not simple, we cannot evaluate $\text{var}(T_n)$ exactly, even if F is known. In statistics, there is an old technique called the Monte Carlo method that can be used to approximate $\text{var}(T_n)$ numerically when F is known. That is, we repeatedly draw new data sets from F and then use the sample variance of the values of T_n computed from new data sets as a numerical approximation to $\text{var}(T_n)$. This idea can be used to approximate v_{BOOT} since \hat{F} is a known distribution. That is, we can draw $\{X_{1b}^*, \dots, X_{nb}^*\}$, $b = 1, \dots, B$, independently from \hat{F} , conditioned on X_1, \dots, X_n ; compute $T_{n,b}^* = T_n(X_{1b}^*, \dots, X_{nb}^*)$ and approximate v_{BOOT} by the following Monte Carlo approximation:

$$v_{\text{BOOT}}^{(B)} = \frac{1}{B} \sum_{b=1}^B \left(T_{n,b}^* - \frac{1}{B} \sum_{l=1}^B T_{n,l}^* \right)^2. \quad (1.18)$$

From the law of large numbers (see Appendix A.5), $v_{\text{BOOT}} = \lim_{B \rightarrow \infty} v_{\text{BOOT}}^{(B)}$ a.s. Both v_{BOOT} and its Monte Carlo approximation $v_{\text{BOOT}}^{(B)}$ may be called bootstrap estimators. In fact, $v_{\text{BOOT}}^{(B)}$ is more useful in practical applications, whereas in theoretical studies we usually focus on v_{BOOT} .

Some other methods for approximating the theoretical formula of the bootstrap estimator are introduced in Chapter 5.

Thus, the bootstrap method is a mixture of two techniques: the substitution principle and the numerical approximation. When v_{BOOT} in (1.16) is an explicit function of X_1, \dots, X_n , the bootstrap coincides with the traditional substitution approach; otherwise, the bootstrap approximates v_{BOOT} numerically [e.g., using $v_{\text{BOOT}}^{(B)}$ in (1.18)], whereas the traditional approach analytically approximates $\text{var}(T_n)$ first and then estimates the unknown quantities in the formula of the approximation to $\text{var}(T_n)$. This understanding of the bootstrap helps us to generalize the use of the bootstrap to more complicated problems.

The distribution \hat{F} used to generate the bootstrap data sets can be any estimator (parametric or nonparametric) of F based on X_1, \dots, X_n . A simple nonparametric estimator of F is the empirical distribution F_n defined in (1.6). Choice of \hat{F} will be discussed in Examples 1.5–1.6 and in later chapters.

Relationship between the bootstrap and the jackknife

In (1.18), $v_{\text{BOOT}}^{(B)}$ can be viewed as a generalization of the jackknife variance estimator in the sense that it is a measure of the variation among the values

of the given statistic based on many different data sets. In the jackknife, the given statistic is recalculated for n fixed data sets that are subsets of the original data set, whereas in the bootstrap the recomputations are based on many *bootstrap data sets* or *resamples* $\{X_{ib}^*, i = 1, \dots, n\}$, $b = 1, \dots, B$, that are randomly generated from the original data set. This is why $v_{\text{BOOT}}^{(B)}$ and v_{JACK} are also called *data-resampling* or simply *resampling estimators*. For the same reason, the jackknife and the bootstrap are also referred to as *resampling methods*.

The bootstrap often has a very close relationship with other data reuse methods such as the jackknife (Efron, 1982; Rao and Wu, 1988; Sitter, 1992a). We will show in Chapter 5 that the jackknife variance estimator v_{JACK} is an approximation to the bootstrap variance estimator v_{BOOT} when the statistic T_n is sufficiently smooth. But this does not imply that v_{BOOT} or $v_{\text{BOOT}}^{(B)}$ is always better than v_{JACK} . Since the jackknife requires fewer computations than the bootstrap, bootstrapping may not be necessary in variance estimation when the jackknife can be applied. On the other hand, the bootstrap can be conveniently applied to other problems, which will be discussed next. In some cases, the jackknife can be used to estimate the accuracy of bootstrap estimates (Efron, 1992a).

Estimation of sampling distribution

While it is not obvious how to estimate the entire sampling distribution of T_n by jackknifing, the bootstrap can be readily used to obtain a distribution estimator for T_n .

More generally, we consider the problem of estimating the sampling distribution of a random variable (or a root) $\mathfrak{R}_n(X_1, \dots, X_n, F)$:

$$H_F(x) = P\{\mathfrak{R}_n(X_1, \dots, X_n, F) \leq x\}, \quad (1.19)$$

where X_1, \dots, X_n are i.i.d. from F . Note that H_F depends on n , but the subscript n is omitted for simplicity. For estimating the sampling distribution of T_n , we simply set $\mathfrak{R}_n(X_1, \dots, X_n, F) = T_n$. When T_n is used to construct a confidence set for an unknown parameter θ related to F , we often use $T_n - \theta$ or a studentized version of $T_n - \theta$: $(T_n - \theta)/S_n$, where S_n is an estimator of the standard deviation of T_n . We then need the distribution of $T_n - \theta$ or $(T_n - \theta)/S_n$, and in this case $\mathfrak{R}_n(X_1, \dots, X_n, F) = T_n - \theta$ or $= (T_n - \theta)/S_n$.

In the traditional approach, we look again for a simple theoretical formula for $H_F(x)$, exact or approximate, and substitute the unknown quantities in the theoretical formula by their estimators. For example, when $\mathfrak{R}_n(X_1, \dots, X_n, F) = T_n - \theta$, in many cases $H_F(x)$ can be approximated by $\Phi(x\sqrt{n}/\sigma)$, where Φ is the standard normal distribution and σ is an unknown parameter related to F . If $\hat{\sigma}$ is an estimator of σ , then we esti-

mate $H_F(x)$ by $\Phi(x\sqrt{n}/\hat{\sigma})$. For $\mathfrak{R}_n(X_1, \dots, X_n, F) = (T_n - \theta)/S_n$, in many cases $H_F(x)$ can be approximated by $\Phi(x)$. This approach, however, has the same weaknesses and disadvantages as those described in Section 1.2.

Following the recipe for obtaining bootstrap variance estimators, we can immediately obtain the bootstrap estimator of $H_F(x)$. First, we substitute F by \hat{F} , an estimator of F , and obtain

$$H_{\text{BOOT}}(x) = H_{\hat{F}}(x) = P_*\{\mathfrak{R}_n(X_1^*, \dots, X_n^*, \hat{F}) \leq x \mid X_1, \dots, X_n\}, \quad (1.20)$$

where X_1^*, \dots, X_n^* are i.i.d. from \hat{F} and $P_*\{\cdot \mid X_1, \dots, X_n\}$ denotes the conditional probability for given X_1, \dots, X_n . If $H_{\text{BOOT}}(x)$ is an explicit function of X_1, \dots, X_n , then it is the *bootstrap estimator* of $H_F(x)$; otherwise, we may use a Monte Carlo approximation to $H_{\text{BOOT}}(x)$:

$$H_{\text{BOOT}}^{(B)}(x) = \frac{1}{B} \sum_{b=1}^B I\{\mathfrak{R}_n(X_{1b}^*, \dots, X_{nb}^*, \hat{F}) \leq x\}, \quad (1.21)$$

where $\{X_{1b}^*, \dots, X_{nb}^*\}, b = 1, \dots, B$, are independent bootstrap data from \hat{F} .

Example 1.5. Location-scale family. Suppose that X_1, \dots, X_n are i.i.d. from a distribution with density $\frac{1}{\tau}f_0(\frac{x-\mu}{\tau})$, where μ and $\tau > 0$ are unknown parameters and f_0 is a known density satisfying $\int x f_0(x)dx = 0$ and $\int x^2 f_0(x)dx = 1$. Suppose that μ and τ^2 are estimated by \bar{X}_n and $\hat{\tau}^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$, respectively. Let $S_n^2 = \hat{\tau}^2/n$. Consider the studentized variable

$$\mathfrak{R}_n(X_1, \dots, X_n, F) = (\bar{X}_n - \mu)/S_n.$$

Let $t_i = (X_i - \mu)/\tau$. Then t_i has density f_0 . Since

$$(\bar{X}_n - \mu)/S_n = \bar{t}_n / \sqrt{[\tilde{t}_n - (\bar{t}_n)^2]/(n-1)},$$

where $\bar{t}_n = n^{-1} \sum_{i=1}^n t_i$ and $\tilde{t}_n = n^{-1} \sum_{i=1}^n t_i^2$, the distribution H_F is independent of F , but its form may not be known explicitly. The bootstrap estimator H_{BOOT} of H_F is the conditional (given X_1, \dots, X_n) distribution of $(\bar{X}_n^* - \bar{X}_n)/S_n^*$, where \bar{X}_n^* is the sample mean of X_1^*, \dots, X_n^* , which are i.i.d. from $\frac{1}{\hat{\tau}}f_0(\frac{x-\bar{X}_n}{\hat{\tau}})$ and $S_n^{*2} = [n(n-1)]^{-1} \sum_{i=1}^n (X_i^* - \bar{X}_n^*)^2$. The bootstrap estimator H_{BOOT} in this case is exactly equal to H_F , since

$$(\bar{X}_n^* - \bar{X}_n)/S_n^* = \bar{t}_n^* / \sqrt{[\tilde{t}_n^* - (\bar{t}_n^*)^2]/(n-1)}$$

with $\bar{t}_n^* = n^{-1} \sum_{i=1}^n t_i^*$, $t_i^* = (X_i^* - \bar{X}_n)/\hat{\tau}$, $\tilde{t}_n^* = n^{-1} \sum_{i=1}^n t_i^{*2}$, and the conditional density of t_i^* is exactly f_0 . If H_{BOOT} is approximated by $H_{\text{BOOT}}^{(B)}$ in (1.21), then the solution is the same as that produced by the Monte Carlo method.

Other applications

Bootstrap confidence sets for an unknown parameter θ can be obtained using the percentiles of H_{BOOT} or $H_{\text{BOOT}}^{(B)}$ (Efron, 1982). This will be discussed further in later chapters.

Sometimes, an accuracy measure of a given statistic T_n is a characteristic of the sampling distribution of T_n . If an estimator of the sampling distribution of T_n is available, a natural estimator of the accuracy measure is the corresponding characteristic of the estimated sampling distribution of T_n . The bootstrap estimators of the accuracy measure can be obtained by using the characteristic of H_{BOOT} or $H_{\text{BOOT}}^{(B)}$. For example, the bootstrap variance estimator v_{BOOT} in (1.16) [or $v_{\text{BOOT}}^{(B)}$ in (1.18)] is actually the variance of the distribution H_{BOOT} [or $H_{\text{BOOT}}^{(B)}$] with $\mathfrak{R}_n = T_n$. Another example is the *interquartile range* of the sampling distribution of T_n . Its bootstrap estimator is the interquartile range of H_{BOOT} or $H_{\text{BOOT}}^{(B)}$.

Bootstrap estimators of other accuracy measures can be obtained in a similar manner. Consider the bias of T_n as an estimator of θ . From (1.7),

$$\text{bias}(T_n) = \int xdH_F(x) - \theta,$$

where $H_F(x)$ is given by (1.19) with $\mathfrak{R}_n = T_n$. We can substitute the unknown F and θ in $\text{bias}(T_n)$ by their estimators \hat{F} and T_n , respectively, and obtain the *bootstrap bias estimator*

$$b_{\text{BOOT}} = \int xdH_{\text{BOOT}}(x) - T_n. \quad (1.22)$$

When the integral in (1.22) has no explicit form, we use the Monte Carlo approximation

$$\begin{aligned} b_{\text{BOOT}}^{(B)} &= \int xdH_{\text{BOOT}}^{(B)}(x) - T_n \\ &= \frac{1}{B} \sum_{b=1}^B T_n(X_{1b}^*, \dots, X_{nb}^*) - T_n. \end{aligned} \quad (1.23)$$

Similar to variance estimators, $b_{\text{BOOT}}^{(B)}$ is also related to the jackknife bias estimator b_{JACK} in (1.8).

Example 1.1 (continued). When $T_n = \bar{X}_n$ and $\hat{F} = F_n$,

$$\int xdH_{\text{BOOT}}(x) = E_*(\bar{X}_n^* | X_1, \dots, X_n) = \bar{X}_n.$$

Hence, $b_{\text{BOOT}} = 0$. When $T_n = \bar{X}_n^2$,

$$\begin{aligned}
\int x dH_{\text{BOOT}}(x) &= E_*(\bar{X}_n^{*2} | X_1, \dots, X_n) \\
&= \text{var}_*(\bar{X}_n^* | X_1, \dots, X_n) + [E_*(\bar{X}_n^* | X_1, \dots, X_n)]^2 \\
&= \frac{1}{n^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2 + \bar{X}_n^2.
\end{aligned}$$

Hence, $b_{\text{BOOT}} = n^{-2} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. Note that in this problem $b_{\text{BOOT}} = [(n-1)/n]b_{\text{JACK}}$.

The mean squared error of T_n as an estimator of θ is

$$E(T_n - \theta)^2 = \text{var}(T_n) + [\text{bias}(T_n)]^2.$$

The bootstrap estimator of $E(T_n - \theta)^2$ is then $v_{\text{BOOT}} + (b_{\text{BOOT}})^2$, and its Monte Carlo approximation is

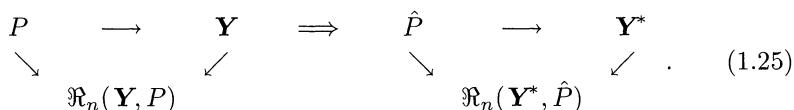
$$v_{\text{BOOT}}^{(B)} + (b_{\text{BOOT}}^{(B)})^2 = \frac{1}{B} \sum_{b=1}^B [T_n(X_{1b}^*, \dots, X_{nb}^*) - T_n]^2. \quad (1.24)$$

Some other uses of the bootstrap procedure, in areas such as testing hypotheses, prediction, and model selection, will be discussed later. It is these broad applications that make the bootstrap very appealing.

A variety of numerical examples of applications of the bootstrap method, together with many useful algorithms, can be found in Efron and Tibshirani (1993).

Summary and discussion

Now we can summarize the bootstrap method in a general situation. Suppose that we have data $\mathbf{Y} = (Y_1, \dots, Y_n)$ (not necessarily i.i.d.) and P is a statistical model under which the data are obtained. Usually, P can be described by the joint distribution of \mathbf{Y} or by some quantities that can uniquely determine this joint distribution. Let $\mathfrak{R}_n(\mathbf{Y}, P)$ be a random variable, and suppose that we want to estimate its distribution. The first step in the bootstrap is to estimate the model P with data \mathbf{Y} . Let \hat{P} be a bootstrap data set generated from the estimated model \hat{P} . The conditional distribution of $\mathfrak{R}_n(\mathbf{Y}^*, \hat{P})$ given \mathbf{Y} is then the bootstrap estimator of the distribution of $\mathfrak{R}_n(\mathbf{Y}, P)$. Using a diagram, Efron and Tibshirani (1986) summarized this process as follows:



The spirit of the bootstrap is to use the sampling behavior of the triplex $(\hat{P}, \mathbf{Y}^*, \mathfrak{R}_n(\mathbf{Y}^*, \hat{P}))$ to mimic that of $(P, \mathbf{Y}, \mathfrak{R}_n(\mathbf{Y}, P))$, where the relationship among \hat{P} , \mathbf{Y}^* and $\mathfrak{R}_n(\mathbf{Y}^*, \hat{P})$ is the same as that among P , \mathbf{Y} and $\mathfrak{R}_n(\mathbf{Y}, P)$. If $\hat{P} = P$ exactly, then the distribution of $\mathfrak{R}_n(\mathbf{Y}^*, \hat{P})$ is exactly the same as that of $\mathfrak{R}_n(\mathbf{Y}, P)$. Even if $\hat{P} \neq P$, the distributions of $\mathfrak{R}_n(\mathbf{Y}^*, \hat{P})$ and $\mathfrak{R}_n(\mathbf{Y}, P)$ sometimes can still be the same (Example 1.5).

While the bootstrap is based on the principle of substitution and mimicking sampling behavior, its application is usually carried out with data-resampling, i.e., when the conditional distribution of $\mathfrak{R}_n(\mathbf{Y}^*, \hat{P})$ is not an explicit function of \mathbf{Y} , Monte Carlo or some other method is required for computing the bootstrap estimators. This is why the bootstrap is usually classified as a data-resampling method, although data-resampling is not absolutely necessary in applying the bootstrap. For instance, bootstrap estimates can be computed by enumerating all possible bootstrap samples when the sample size n is small (Fisher and Hall, 1991).

The bootstrap can be applied to all situations where a model P can be established and estimated by \hat{P} . However, it is crucial how the model P is *postulated* and *estimated*. We illustrate this through some examples.

Example 1.6. One sample problem. The data X_1, \dots, X_n are i.i.d. from a distribution F . The joint distribution of X_1, \dots, X_n is determined by F . Hence, $P = F$. If F belongs to a parametric family, then $P = F_\theta$, where θ is a vector of unknown parameters. In the parametric case, θ is first estimated by an estimator $\hat{\theta}$ and P is then estimated by $\hat{P} = F_{\hat{\theta}}$. The bootstrap data set X_1^*, \dots, X_n^* is now generated from $F_{\hat{\theta}}$. This bootstrap method is often called the *parametric bootstrap*. In the nonparametric case, P is estimated by $\hat{P} = F_n$, the empirical distribution given by (1.6). The bootstrap data set X_1^*, \dots, X_n^* is then generated from F_n . This bootstrap method is often called the *nonparametric bootstrap*. Note that the nonparametric bootstrap can be used for both parametric and nonparametric models.

The parametric bootstrap depends on the parametric model assumption. The nonparametric bootstrap is “model assumption free”, but it is not as efficient as the parametric bootstrap when the parametric model is correct. In general, the performance of the bootstrap relies on how well we can identify and estimate the model. Even in the nonparametric case, we can replace F_n by a smoothed estimator of F when we know F is smooth. This produces better nonparametric bootstrap estimators. We will return to this subject later.

Example 1.7. Linear regression model. Suppose that the data Y_1, \dots, Y_n are independent and $Y_i = (y_i, x'_i)$, $i = 1, \dots, n$, where the x_i are p -vectors and x' is the transpose of x . Depending on whether or not the x_i are random, we have two different models in this problem.

Case 1. The x_i are nonrandom, $y_i = x'_i \beta + \varepsilon_i$, β is a p -vector of unknown parameters, and $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. from an unknown distribution F_ε with 0 mean. In this case, P can be identified as (β, F_ε) . Suppose that β is estimated by $\hat{\beta}$ (e.g., the least squares estimator). Then F_ε can be estimated by the empirical distribution \hat{F}_ε putting mass n^{-1} to $\hat{\varepsilon}_i - n^{-1} \sum_{j=1}^n \hat{\varepsilon}_j, i = 1, \dots, n$, where $\hat{\varepsilon}_i = y_i - x'_i \hat{\beta}$ is the i th residual. P is now estimated by $\hat{P} = (\hat{\beta}, \hat{F}_\varepsilon)$. To generate bootstrap data Y_1^*, \dots, Y_n^* , we first generate i.i.d. data $\varepsilon_1^*, \dots, \varepsilon_n^*$ from \hat{F}_ε and then define $y_i^* = x'_i \hat{\beta} + \varepsilon_i^*$, $Y_i^* = (y_i^*, x'_i)$, $i = 1, \dots, n$. This bootstrap method is often called *bootstrapping residuals* or *the bootstrap based on residuals*.

Case 2. The x_i are random, $(y_i, x'_i), i = 1, \dots, n$, are i.i.d. from an unknown $(p+1)$ -variate distribution F , and $E(y_i | x_i) = x'_i \beta$. In this case, $P = F$ and can be estimated by the empirical distribution function F_n putting mass n^{-1} to the pair $(y_i, x'_i), i = 1, \dots, n$. The bootstrap data are generated from F_n . This bootstrap method is often called *bootstrapping pairs* or *the paired bootstrap*.

Therefore, the bootstrap is a “model-dependent” method in terms of its implementation (how the bootstrap data are generated) and performance, although the bootstrap requires no theoretical formula for the quantity to be estimated and is less model-dependent than the traditional approach.

1.5 Extensions to Complex Problems

Although the jackknife and the bootstrap are first used in i.i.d. data problems, their extensions to non-i.i.d. data problems are sometimes straightforward. In fact, we have shown in Section 1.4 how the bootstrap is used for non-i.i.d. data \mathbf{Y} . With non-i.i.d. complex data, it is very difficult, or even impossible in some cases, to derive the theoretical formula required by the traditional approach. This is an important reason why the jackknife and the bootstrap are so attractive and have become so popular.

Naturally, some properties that the jackknife or the bootstrap estimator have in the i.i.d. case may be lost in the non-i.i.d. case. Blindly applying the jackknife or the bootstrap may lead to incorrect results. To get correct results and/or retain the properties that the jackknife or the bootstrap has in the i.i.d. case, we have to modify the jackknife or the bootstrap to take into account the non-i.i.d. nature of the data. More detailed discussions will be provided later.

We now give an outline of the applications of the jackknife and the bootstrap in non-i.i.d. complex data problems that will be studied in later chapters.

Two non-i.i.d. data problems in which the jackknife, the bootstrap, and their modifications are widely used are sample surveys and regression analyses.

A recent trend in sample survey problems is the use of one or a combination of several of the following modern sampling techniques: stratified sampling, cluster sampling, unequal probability sampling, and multistage sampling (Kish and Frankel, 1974). Because of the complexity of the sampling design, the resulting data are heavily non-i.i.d. With slight modifications, the jackknife can still be used to obtain variance estimators. Nowadays, the jackknife is one of the two most popular methods for variance estimation used in survey problems (the other one is the traditional linearization and substitution method). Many data analysts prefer the jackknife simply because it does not require complicated and/or difficult derivation of the theoretical formula for the variance (or its approximation) of the given statistic T_n . The application of the bootstrap to survey data is, however, a more delicate problem. Firstly, without some substantial modifications, the bootstrap may provide an incorrect solution (Rao and Wu, 1988). Secondly, for variance estimation, the bootstrap requires considerably more computations than the jackknife, and the gain in doing these extra computations is inappreciable or unknown. Finally, the theory of the bootstrap distribution estimators and bootstrap confidence sets in complex survey problems has not been well developed, although some attempts have been made (Rao and Wu, 1988). We shall return to these topics in Chapter 6.

Miller (1974) studied the jackknife in linear models. Since then, a considerable amount of work has been done in this area. The main focus has been on the validity of the jackknife and bootstrap variance estimators; the accuracy of the bootstrap confidence sets, bootstrap simultaneous confidence intervals, and bootstrap prediction intervals; the use of the *weighted* jackknife and bootstrap to take into account the imbalance of the regression data; the use of various sampling schemes to generate bootstrap data (see Example 1.7); and the robustness of the jackknife and bootstrap estimators against heteroscedasticity. Chapter 7 provides a complete study of applications of the jackknife and bootstrap in linear models.

There are a number of parametric, semiparametric, and nonparametric nonlinear models that are extensions of the linear regression model from different directions, for example, the nonlinear regression models, the generalized linear models, Cox's regression model, and nonparametric regression models. Many of the results for the jackknife and bootstrap in linear models can be extended to these nonlinear models. This will be studied in Chapter 8, which also includes applications of the jackknife and bootstrap in other complex problems such as nonparametric density estimation and multivariate analysis.

Although the jackknife and the bootstrap procedures described in this chapter cannot be applied to a problem involving dependent data such as a time series, some of their modifications, as described in Chapter 9, work well for dependent data from a stationary process.

In addition to their applications to variance estimation, distribution estimation, and other inference and prediction problems, the jackknife and the bootstrap can also be used in a very important and complex problem: the model selection. The jackknife is closely related to a method called *cross-validation* for selecting the explanatory variables that can be used to predict future responses in linear, nonlinear, generalized linear, and Cox's regression models. However, some modifications have to be made in order to obtain accurate jackknife (cross-validation) and bootstrap variable selection procedures in these problems. The cross-validation method and the bootstrap are widely used for selecting bandwidth in nonparametric density estimation and nonparametric regression. A modified bootstrap can also be used to select the order of an autoregressive time series model. These topics are all covered in Chapters 7–9.

The idea of data-resampling can be applied to the problem of computing posterior distributions in Bayesian analysis. This leads to the Bayesian bootstrap (Rubin, 1981) and a closely related method called random weighting (Zheng, 1987a). These methods can also be used as alternatives to the bootstrap; the details are provided in Chapter 10.

1.6 Scope of Our Studies

Ideally, solid theory should be behind practice. Despite the fact that the jackknife and bootstrap are practically and intuitively appealing as described in the previous sections, theoretical confirmation should be made for using these methods, especially when the data are from a complex statistical model. Theoretical studies are, in most cases, limited to asymptotics, i.e., the limiting (as the sample size tends to infinity) behaviors of the jackknife and bootstrap estimators. Although in real applications the sample size cannot increase infinitely, the asymptotic theory does provide a guide to the proper use of the methodology. Asymptotically incorrect methods should not be used, and asymptotically accurate and efficient methods should be recommended to replace inaccurate methods.

Asymptotic studies are often supplemented by empirical (finite sample) studies. An empirical study, however, may not allow us to see the whole picture, since we are not able to do an exhaustive empirical study for all different possible populations generating the data.

We now outline the scope of our studies of the jackknife and bootstrap. The following issues will be addressed in the rest of this book.

Consistency

For a sequence of estimators $\{\delta_n, n = 1, 2, \dots\}$ of an unknown parameter δ , $\{\delta_n\}$ is weakly consistent (or, simply, δ_n is consistent) if, as $n \rightarrow \infty$,

$$\delta_n \rightarrow_p \delta \quad (1.26)$$

for all possible values of δ , where \rightarrow_p denotes convergence in probability (see Appendix A.1). δ_n is strongly consistent if

$$\delta_n \rightarrow_{a.s.} \delta \quad (1.27)$$

for all possible values of δ , where $\rightarrow_{a.s.}$ denotes convergence almost surely (see Appendix A.1). Strong consistency implies weak consistency. Although weak consistency is sufficient for most statistical problems, results of strong consistency will be presented in this book whenever they are available.

The definition of consistency in (1.26)–(1.27) has to be modified slightly when estimating the variance of T_n , since $\text{var}(T_n)$ depends on n and usually $\text{var}(T_n) \rightarrow 0$ as $n \rightarrow \infty$. Note that when $\delta \neq 0$, (1.26) or (1.27) is equivalent to $\delta_n/\delta \rightarrow_p 1$ or $\delta_n/\delta \rightarrow_{a.s.} 1$. Hence, a variance estimator v_n of $\text{var}(T_n)$ (e.g., v_{JACK} or v_{BOOT}) is said to be weakly consistent if

$$v_n/\text{var}(T_n) \rightarrow_p 1 \quad (1.28)$$

or strongly consistent if

$$v_n/\text{var}(T_n) \rightarrow_{a.s.} 1. \quad (1.29)$$

The consistency of bias estimators can be defined similarly.

When the bootstrap is applied to estimate the sampling distribution H_F , H_F and its estimator H_{BOOT} are functions, and, therefore, the consistency of H_{BOOT} has to be redefined. This will be discussed in Chapter 3.

As an estimator of δ , δ_n is consistent means that δ_n is asymptotically correct and, therefore, for sufficiently large sample sizes, it is expected that δ_n is very close to δ or δ_n will be close to δ with high probability. Thus, consistency is a minimum requirement for any estimator. Inconsistent estimators should not be used.

The problem of variance estimation needs to be rephrased in the cases where $\text{var}(T_n)$ does not exist for each n (e.g., $T_n = \bar{X}_n^2$ and $EX_i^4 = \infty$), but

$$(T_n - \theta)/\sigma_n \rightarrow_d Z, \quad (1.30)$$

where Z is a random variable with 0 mean and unit variance, $\{\sigma_n\}$ is a sequence of nonrandom positive numbers, and \rightarrow_d denotes convergence in distribution (see Appendix A.1). In general, the existence of σ_n and

(1.30) does not imply the existence of $\text{var}(T_n)$; and the existence of $\text{var}(T_n)$ does not imply (1.30). However, when both $\text{var}(T_n)$ and σ_n^2 exist and (1.30) holds, very often $\text{var}(T_n)/\sigma_n^2 \rightarrow 1$. For example, if (1.4) holds, then $\sigma_n^2 = \sigma_F^2/n$.

In (1.30), σ_n^2 is called the *asymptotic* or *approximate* variance of T_n . When σ_n^2 exists but $\text{var}(T_n)$ does not, v_{JACK} and v_{BOOT} are estimators of σ_n^2 , and their consistency is defined by (1.28)–(1.29) with $\text{var}(T_n)$ replaced by σ_n^2 . When both $\text{var}(T_n)$ and σ_n^2 exist, v_{JACK} and v_{BOOT} can be used to estimate both of them, which is justified if $\text{var}(T_n)/\sigma_n^2 \rightarrow 1$.

Asymptotic accuracy

The consistency of an estimator tells us that, for large n , the error in estimation is likely to be small but not whether the order of the error is $n^{-1}, n^{-1/2}, n^{-1/4}$, and so on. It is desirable to measure the accuracy of the jackknife and bootstrap estimators in terms of their rates of convergence and other quantities such as their asymptotic mean squared errors. The convergence rate and asymptotic mean squared errors are usually used for theoretical comparisons among various estimators. For example, we may want to know whether the bootstrap provides better estimators than the traditional approach. Sometimes we also would like to compare estimators obtained by using different bootstrap sampling schemes (see Example 1.7).

Bootstrap confidence sets

One of the main tasks in statistical inference is the construction of confidence sets for unknown parameters of interest. Theory for bootstrap confidence sets has rapidly developed since the invention of the bootstrap. It is important to know when the bootstrap provides a better confidence set than the traditional method, in terms of the coverage probability of the confidence set. Since there are many different versions of bootstrap confidence sets, it is also important to compare their performances. We will also study many interesting techniques, such as the bootstrap accelerated bias-corrected, the bootstrap prepivoting, the bootstrap calibration, and the iterative bootstrap methods, that have been introduced recently to increase the accuracy of the bootstrap confidence sets.

Extensions to complex problems

Although the formulas of the jackknife and bootstrap estimators for the i.i.d. data can still be used for complex data, they may not provide correct answers. In addition, when the model is complex, there may exist more than one way to remove data in the case of the jackknife or to obtain resamples in the case of the bootstrap. Therefore, we need to study how to take into account the special features of the model or the data set to

produce correct and efficient jackknife and bootstrap estimators.

Model selection

As we discussed in Section 1.5, an important application of the jackknife and bootstrap methods is to use them for model selection in regression analysis and bandwidth selection in nonparametric curve estimation. An essential theoretical justification for a model (bandwidth) selection procedure is its consistency, i.e., whether the model (bandwidth) selected by the particular procedure is close to the optimal model (bandwidth). A precise definition of the consistency of a model (bandwidth) selection procedure will be given in later chapters.

Robustness

As we pointed out earlier, unlike the traditional approach, the jackknife and the bootstrap do not rest on a theoretical formula that is derived under some model assumptions. Hence, their performances are less susceptible to violation of these model assumptions. The robustness of the jackknife and the bootstrap was recognized by Tukey and subsequent researchers. For example, Hinkley (1977) and Wu (1986) found that the jackknife variance estimator for the least squares estimator in linear models is robust against the presence of unequal error variances. The robustness property of the jackknife and the bootstrap provides another strong motivation for the use of these data-resampling methods.

Computation

Despite the power of modern computers, it is still very important to study how to efficiently compute the jackknife and the bootstrap estimators, especially when we compute bootstrap confidence sets. Even for variance estimation, the computations of the jackknife and the bootstrap estimators can also be cumbersome if the sample size is very large (e.g., a large scale survey problem). We will discuss various analytic and simulation-type approximations proposed recently to replace the Monte Carlo approximation, an old and popular but perhaps inefficient method.

Empirical simulation

Results from empirical simulation studies often will be presented to examine finite sample properties of various methods. These empirical results provide some numerical examples as complementary studies of the asymptotic theory. They allow us to compare the performances of some methods that are asymptotically equivalent and to assess the actual difference between two methods in the case where one method is shown to be asymptotically better than the other.

Chapter 2

Theory for the Jackknife

This chapter presents theory for the jackknife in the case where the data are i.i.d. Many results can be extended in a straightforward manner to more complicated cases, which will be studied in later chapters. We begin this chapter by first focusing on jackknife variance estimators. The basic theoretical consideration in using jackknife variance estimators is their consistency, which is especially crucial when the jackknife variance estimators are used in large sample statistical inference problems such as constructing confidence sets for some unknown parameters. A complete theory for the consistency of the jackknife variance estimators is given in Sections 2.1 and 2.2. We will show that the success of the jackknife variance estimator for a given statistic T_n relies on the smoothness of T_n , which can be characterized by the differentiability of the function that generates T_n . The jackknife variance estimator may be inconsistent for a statistic that is not very smooth; however, its inconsistency can be rectified by using the *delete-d jackknife*, an extended version of the jackknife that removes more than one datum at a time (Section 2.3). The delete-d jackknife also provides a jackknife estimator of the sampling distribution of T_n , known as the *jackknife histogram*. Other applications of the jackknife, such as bias estimation and bias reduction, are discussed in Section 2.4. Some empirical results are given as examples.

2.1 Variance Estimation for Functions of Means

Throughout this chapter, we assume that X_1, \dots, X_n are i.i.d. nondegenerate random p -vectors sampled from an unknown p -dimensional distribution F . Let $T_n = T_n(X_1, \dots, X_n)$ be a given statistic.

For the simplest case where $p = 1$ and $T_n = \bar{X}_n = \sum_{i=1}^n X_i/n$, the

sample mean, we showed in Example 1.1 that the jackknife estimate v_{JACK} in (1.13) reduces to the usual estimator $[n(n-1)]^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ and its properties are well known (consistency, unbiasedness, etc.).

Many useful statistics are functions of the sample mean \bar{X}_n , i.e., $T_n = g(\bar{X}_n)$ with a function g from \mathbb{R}^p to \mathbb{R} . The following is an example.

Example 2.1. Ratio and correlation coefficient. Suppose that $(Y_i, Z_i)' \in \mathbb{R}^2$ are i.i.d. data and we wish to make an inference about the ratio $\gamma = EY_1/EZ_1$ based on the sample ratio $\hat{\gamma}_n = \bar{Y}_n/\bar{Z}_n$ or a function of $\hat{\gamma}_n$ such as $\log \hat{\gamma}_n$, where $\bar{Y}_n = \sum_{i=1}^n Y_i/n$ and $\bar{Z}_n = \sum_{i=1}^n Z_i/n$. In this case, the statistic of interest is a function of $\bar{X}_n = (\bar{Y}_n, \bar{Z}_n)'$.

Another important parameter is the correlation coefficient ρ between Y_1 and Z_1 . The sample correlation coefficient

$$\hat{\rho}_n = \sum_{i=1}^n (Y_i - \bar{Y}_n)(Z_i - \bar{Z}_n) / \left[\sum_{i=1}^n (Y_i - \bar{Y}_n)^2 \sum_{i=1}^n (Z_i - \bar{Z}_n)^2 \right]^{1/2}$$

can be written as $g(\bar{X}_n)$ for some function g and $X'_i = (Y_i, Z_i, Y_i^2, Z_i^2, Y_i Z_i)$, $i = 1, \dots, n$. Inference about ρ is often based on a transformed statistic $\hat{\phi}_n = \frac{1}{2} \log[(1 + \hat{\rho}_n)/(1 - \hat{\rho}_n)]$, which is also a function of \bar{X}_n .

More examples are given in Section 2.1.3.

2.1.1 Consistency

The first theorem concerning the consistency of the jackknife variance estimator was given by Miller (1964) for the case of $T_n = g(\bar{X}_n)$. We also begin with this relatively simple case, not only because functions of \bar{X}_n are important statistics but also because the proof of consistency in the case of $T_n = g(\bar{X}_n)$ highlights the method of proofs in general cases.

Let $\mu = E(X_1)$ and $\Sigma = \text{var}(X_1)$. Under the conditions that Σ exists and g is differentiable at μ with $\nabla g(\mu) \neq 0$, we have

$$(T_n - \theta)/\sigma_n \rightarrow_d N(0, 1), \quad (2.1)$$

where $N(0, 1)$ denotes a standard normal random variable, $\theta = g(\mu)$,

$$\sigma_n^2 = n^{-1} \nabla g(\mu)' \Sigma \nabla g(\mu), \quad (2.2)$$

and ∇g is the gradient of g (see Appendices A.2 and A.8). σ_n^2 is then the asymptotic variance of T_n (see the discussion in Section 1.6). Note that for $T_n = g(\bar{X}_n)$, $\text{var}(T_n)$ does not exist in many cases. Hence, the jackknife estimator v_{JACK} is applied to estimate σ_n^2 . If v_{JACK} is consistent according to (1.28) or (1.29), then by (2.1),

$$(T_n - \theta)/\sqrt{v_{\text{JACK}}} \rightarrow_d N(0, 1),$$

which is a useful result for large sample statistical inferences.

Theorem 2.1. Suppose that Σ exists, $T_n = g(\bar{X}_n)$, ∇g is defined in a neighborhood of μ , $\nabla g(\mu) \neq 0$, and ∇g is continuous at μ . Then the jackknife variance estimator for T_n is strongly consistent, i.e.,

$$v_{\text{JACK}}/\sigma_n^2 \rightarrow_{a.s.} 1, \quad (2.3)$$

where σ_n^2 is given by (2.2).

Proof. Let $\bar{X}_{n-1,i} = \sum_{j \neq i} X_j / (n - 1)$ be the sample mean based on $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$. From the mean-value theorem, we have

$$\begin{aligned} T_{n-1,i} - T_n &= g(\bar{X}_{n-1,i}) - g(\bar{X}_n) \\ &= \nabla g(\xi_{n,i})'(\bar{X}_{n-1,i} - \bar{X}_n) \\ &= \nabla g(\bar{X}_n)'(\bar{X}_{n-1,i} - \bar{X}_n) + R_{n,i}, \end{aligned}$$

where $R_{n,i} = [\nabla g(\xi_{n,i}) - \nabla g(\bar{X}_n)]'(\bar{X}_{n-1,i} - \bar{X}_n)$ and $\xi_{n,i}$ is a point on the line segment between $\bar{X}_{n-1,i}$ and \bar{X}_n . From

$$\bar{X}_{n-1,i} - \bar{X}_n = (n - 1)^{-1}(\bar{X}_n - X_i), \quad (2.4)$$

it follows that $\sum_{i=1}^n (\bar{X}_{n-1,i} - \bar{X}_n) = 0$ and

$$\frac{1}{n} \sum_{i=1}^n (T_{n-1,i} - T_n) = \frac{1}{n} \sum_{i=1}^n R_{n,i} = \bar{R}_n.$$

From the definition of the jackknife estimator v_{JACK} in (1.13),

$$\begin{aligned} v_{\text{JACK}} &= \frac{n-1}{n} \sum_{i=1}^n \left(T_{n-1,i} - \frac{1}{n} \sum_{j=1}^n T_{n-1,j} \right)^2 \\ &= A_n + B_n + 2C_n, \end{aligned}$$

where

$$A_n = \frac{n-1}{n} \nabla g(\bar{X}_n)' \sum_{i=1}^n (\bar{X}_{n-1,i} - \bar{X}_n)(\bar{X}_{n-1,i} - \bar{X}_n)' \nabla g(\bar{X}_n),$$

$$B_n = \frac{n-1}{n} \sum_{i=1}^n (R_{n,i} - \bar{R}_n)^2,$$

and

$$C_n = \frac{n-1}{n} \sum_{i=1}^n (R_{n,i} - \bar{R}_n)(\bar{X}_{n-1,i} - \bar{X}_n)' \nabla g(\bar{X}_n).$$

By the Cauchy-Schwarz inequality, $C_n^2 \leq A_n B_n$. Hence, (2.3) follows from

$$A_n/\sigma_n^2 \rightarrow_{a.s.} 1 \quad (2.5)$$

and

$$B_n/\sigma_n^2 \rightarrow_{a.s.} 0. \quad (2.6)$$

We first show (2.5). From (2.4),

$$A_n = \frac{1}{n(n-1)} \nabla g(\bar{X}_n)' \sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)' \nabla g(\bar{X}_n).$$

Hence, (2.5) follows from $n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)' \rightarrow_{a.s.} \Sigma$, by the strong law of large numbers, and $\nabla g(\bar{X}_n) \rightarrow_{a.s.} \nabla g(\mu)$, by the continuity of ∇g at μ and $\bar{X}_n \rightarrow_{a.s.} \mu$ (see Appendices A.2 and A.5).

We now show (2.6). From (2.4),

$$(n-1) \sum_{i=1}^n \|\bar{X}_{n-1,i} - \bar{X}_n\|^2 = \frac{1}{n-1} \sum_{i=1}^n \|X_i - \bar{X}_n\|^2 \rightarrow_{a.s.} \text{tr}(\Sigma), \quad (2.7)$$

where $\text{tr}(A)$ denotes the trace of the matrix A . Hence,

$$\max_{i \leq n} \|\bar{X}_{n-1,i} - \bar{X}_n\|^2 \rightarrow_{a.s.} 0, \quad (2.8)$$

where $\|\cdot\|$ denotes the Euclidean norm in \mathbb{R}^p . It follows from (2.8), the continuity of ∇g at μ , and $\|\xi_{n,i} - \bar{X}_n\| \leq \|\bar{X}_{n-1,i} - \bar{X}_n\|$ that

$$u_n = \max_{i \leq n} \|\nabla g(\xi_{n,i}) - \nabla g(\bar{X}_n)\| \rightarrow_{a.s.} 0.$$

Hence,

$$\frac{B_n}{\sigma_n^2} \leq \frac{n-1}{\sigma_n^2 n} \sum_{i=1}^n R_{n,i}^2 \leq \frac{u_n}{\sigma_n^2} \sum_{i=1}^n \|\bar{X}_{n-1,i} - \bar{X}_n\|^2 \rightarrow_{a.s.} 0,$$

since, from (2.2) and (2.7), $\sum_{i=1}^n \|\bar{X}_{n-1,i} - \bar{X}_n\|^2 / \sigma_n^2 = O(1)$ a.s. This proves (2.6) and completes the proof of (2.3). \square

The argument used in the proof of Theorem 2.1 is very typical in proving the consistency of jackknife estimators and will be used frequently in the rest of the book. Examining the proof, it can be seen that the following two key facts ensure the consistency of v_{JACK} :

- (1) For the linear statistic $T_n = l' \bar{X}_n$, where l is a p -vector, the jackknife variance estimator for T_n is

$$\frac{n-1}{n} \sum_{i=1}^n [l' (\bar{X}_{n-1,i} - \bar{X}_n)]^2 = \frac{1}{n(n-1)} \sum_{i=1}^n [l' (X_i - \bar{X}_n)]^2,$$

which is consistent for $n^{-1} l' \Sigma l$ according to the law of large numbers.

- (2) $T_{n-1,i} - T_n$ can be approximated by $l'(\bar{X}_{n-1,i} - \bar{X}_n)$ with $l = \nabla g(\bar{X}_n)$ and the contributions of the remainders $R_{n,i}$ are sufficiently small [result (2.6)].

Fact (2) indicates that the consistency of v_{JACK} requires a certain degree of smoothness from $T_n = g(\bar{X}_n)$, i.e., how well T_n is approximated by some linear statistic. In Theorem 2.1, the smoothness of T_n is characterized by the differentiability of g . In fact, the consistency of v_{JACK} requires the continuity of the derivative ∇g , which is more than what is required for result (2.1). Recall that in establishing (2.1) we only require the differentiability of g (see Appendix A.2).

Thus, in theoretical studies of v_{JACK} for general statistics T_n , we should look for some kind of linear approximation for T_n , perhaps by using some notion of the differentiability of the function that generates T_n . This will be studied in Section 2.2.

Let m be a fixed positive integer and $h(x_1, \dots, x_m)$ be a vector-valued function from \mathbb{R}^{pm} to \mathbb{R}^p satisfying $h(x_1, \dots, x_m) = h(x_{i_1}, \dots, x_{i_m})$ for any permutation $\{i_1, \dots, i_m\}$ of $\{1, \dots, m\}$. For $n > m$, define

$$U_n = \binom{n}{m}^{-1} \sum_c h(X_{i_1}, \dots, X_{i_m}), \quad (2.9)$$

where \sum_c denotes the summation over the $\binom{n}{m}$ combinations of m distinct elements $\{i_1, \dots, i_m\}$ from $\{1, \dots, n\}$. The statistic U_n , called a U-statistic, is an unbiased estimator of $\mu_m = E[h(X_1, \dots, X_m)]$ and a generalization of the simple sample mean \bar{X}_n . Examples of U-statistics can be found in Serfling (1980, Chapter 5). An estimator of $\theta = g(\mu_m)$ is then $T_n = g(U_n)$.

The results in Theorem 2.1 can be extended to the case of $T_n = g(U_n)$. Using the theory for U-statistics (e.g., Chapter 5 in Serfling, 1980), we can show that $T_{n-1,i} - T_n$ can be approximated by $\nabla g(U_n)'(\bar{Z}_{n-1,i} - \bar{Z}_n)$, where $\bar{Z}_n = n^{-1} \sum_{i=1}^n Z_i$, $\bar{Z}_{n-1,i} = (n-1)^{-1} \sum_{j \neq i} Z_j$, $Z_i = \tilde{h}(X_i)$, and $\tilde{h}(x) = mE[h(x, X_2, \dots, X_m) - \mu_m]$. Thus, the asymptotic variance of T_n is $\sigma_n^2 = \nabla g(\mu_m)'E[(\bar{Z}_n - E\bar{Z}_n)(\bar{Z}_n - E\bar{Z}_n)']\nabla g(\mu_m)$, and the consistency of the jackknife estimator v_{JACK} for σ_n^2 can be established along the line of the proof of Theorem 2.1 (see also Arvesen, 1969). We state the following result without giving a proof. Note that Theorem 2.1 is a special case of Theorem 2.2 with $m = 1$ and $h(X_i) = X_i$.

Theorem 2.2. Suppose that $E\|h(X_1, \dots, X_m)\|^2 < \infty$, $T_n = g(U_n)$, and ∇g is continuous at μ_m with $\nabla g(\mu_m) \neq 0$. Then (2.3) holds.

A statistic closely related to the U-statistic in (2.9) is the V-statistic

defined by

$$V_n = \frac{1}{n^m} \sum_{i_1=1}^n \cdots \sum_{i_m=1}^n h(X_{i_1}, \dots, X_{i_m}).$$

Under the conditions in Theorem 2.2 and an extra condition that for any $1 \leq i_1 \leq \dots \leq i_m \leq m$, $E\|h(X_{i_1}, \dots, X_{i_m})\|^2 < \infty$, it can be shown that the result in Theorem 2.2 holds with U_n replaced by V_n (see, e.g., Sen, 1977).

2.1.2 Other properties

A variance estimator can also be assessed by its bias or variance. A simple and exact form of the bias of v_{JACK} is hard to obtain. Using an ANOVA decomposition, Efron and Stein (1981) showed that for a statistic T_n having a finite second order moment,

$$\text{var}(T_n) = \sum_{j=1}^n \binom{n-1}{j-1} \frac{\varrho_j}{jn^{2j-1}} \quad (2.10)$$

and

$$Ev_{\text{JACK}} = \frac{n-1}{n} \sum_{j=1}^{n-1} \binom{n-2}{j-1} \frac{\varrho_j}{(n-1)^{2j-1}}, \quad (2.11)$$

where ϱ_j are unknown positive constants. Let $\text{var}_{n-1} = \text{var}(T_{n-1,n})$, i.e., the variance of the given statistic based on $n-1$ observations. Then, (2.10) and (2.11) imply

$$Ev_{\text{JACK}} - \frac{(n-1)\text{var}_{n-1}}{n} = \frac{n-1}{n} \sum_{j=2}^{n-1} \binom{n-2}{j-1} \frac{(j-1)\varrho_j}{j(n-1)^{2j-1}}.$$

This indicates that if $(n-1)\text{var}_{n-1}/n \geq \text{var}(T_n)$, then v_{JACK} is conservative in the sense that $Ev_{\text{JACK}} \geq \text{var}(T_n)$.

Results (2.10) and (2.11), however, do not imply that the bias of v_{JACK} is of the order $O(n^{-2})$, since the number of terms in the sum in (2.11) increases as n increases. In some simple cases, we may obtain the order of the bias of v_{JACK} directly.

Example 1.3 (continued). Consider $T_n = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. Its jackknife estimator is $v_{\text{JACK}} = [(1 - n^{-1})\hat{\alpha}_4 - \hat{\alpha}_2^2]/(n-1)$ (see Section 1.3). Applying Theorem 2.2.3A in Serfling (1980), we obtain that

$$\text{var}(T_n) = \frac{\alpha_4 - \alpha_2^2}{n} + O\left(\frac{1}{n^2}\right),$$

$$E\hat{\alpha}_4 = \alpha_4 + \frac{\alpha_4 - 2\alpha_2}{n} + O\left(\frac{1}{n^2}\right),$$

$$E\hat{\alpha}_2^2 = \alpha_2^2 + \frac{\alpha_4 - \alpha_2^2}{n} + O\left(\frac{1}{n^2}\right),$$

and therefore $\text{bias}(v_{\text{JACK}}) = O(n^{-2})$.

For the case of $T_n = g(\bar{X}_n)$ with a smooth g , it can be shown that v_{JACK} is asymptotically unbiased in the sense that $\text{bias}(v_{\text{JACK}}) = o(n^{-1})$ (Shao, 1988a).

The variance of v_{JACK} is even more complicated and is not discussed here. Some empirical results are shown in Sections 2.1.3 and 2.3.1. We can also assess v_{JACK} by its convergence rate or asymptotic efficiency (see Section 2.2.3).

2.1.3 Discussions and examples

The σ_n^2 in (2.2) can also be estimated via the traditional approach. Since (2.2) is a theoretical formula for σ_n^2 that involves unknown μ and Σ , an estimator of σ_n^2 can be obtained by substituting μ with \bar{X}_n and Σ with the sample covariance matrix

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)'.$$

The resulting estimator is

$$v_L = n^{-1} \nabla g(\bar{X}_n)' \hat{\Sigma} \nabla g(\bar{X}_n),$$

which is exactly the same as A_n in (2.5). Since the theoretical formula (2.2) is obtained by linearization, v_L is also called the linearization estimator. Clearly, $v_{\text{JACK}}/v_L \rightarrow_{a.s.} 1$ (there is, in fact, a higher order asymptotic equivalence between v_{JACK} and v_L ; see Theorems 2.7 and 6.3). Therefore, the jackknife estimator does not have any apparent superiority in terms of asymptotic performance. Then, why would we want to use a jackknife variance estimator in this problem?

The computation of v_L requires the derivation of the explicit form of ∇g and the evaluation of ∇g at \bar{X}_n , whereas the computation of v_{JACK} requires n evaluations of g at $\bar{X}_{n-1,i}$, $i = 1, \dots, n$, which is routine and easy to program. The derivation of ∇g can be very complicated and/or tedious, or sometimes is impossible (e.g., the function g is implicitly defined). In such cases the jackknife is preferred. We now study two examples.

Example 2.2. Estimation of poverty line. In the study of income shares or wealth distributions, sometimes we need to estimate the poverty line (or

low income cutoff) of the population. For the i th sampled family, let z_i be expenditure on “necessities”, y_i be total income, and x_{it} , $t = 1, \dots, m$, be variables such as urbanization category and family size. Then,

$$\log z_i = \gamma_1 + \gamma_2 \log y_i + \sum_{t=1}^m \beta_t x_{it} + \text{error}, \quad (2.12)$$

where γ_1 , γ_2 , β_t , $t = 1, \dots, m$, are unknown parameters. Let γ_0 be the overall proportion of income spent on “necessities”. Then the poverty line θ is defined to be the solution of

$$\log[(\gamma_0 + 0.2)\theta] = \gamma_1 + \gamma_2 \log \theta + \sum_{t=1}^m \beta_t x_{0t} \quad (2.13)$$

for a particular set of x_{01}, \dots, x_{0m} (Mantel and Singh, 1991).

Suppose that γ_0 is estimated by $\hat{\gamma}_0 = \bar{z}/\bar{y}$, $\bar{z} = n^{-1} \sum_{i=1}^n z_i$, $\bar{y} = n^{-1} \sum_{i=1}^n y_i$, γ_j and β_t are estimated by the least squares estimators under (2.12), say $\hat{\gamma}_j$ and $\hat{\beta}_t$, $j = 1, 2$, $t = 1, \dots, m$. Then an estimator $\hat{\theta}$ of θ is the solution of (2.13) with γ_j and β_t replaced by $\hat{\gamma}_j$ and $\hat{\beta}_t$. Since $\hat{\theta}$ is a function of $\hat{\gamma}_j$ and $\hat{\beta}_t$, which are functions of sample means, $\hat{\theta}$ can be written as $g(\bar{X}_n)$ for a function g and $X_i = (z_i, y_i, \log z_i, \log y_i, x_{i1}, \dots, x_{im})'$.

Since the function g is not explicit, the derivation of ∇g is difficult. In fact, even if θ is an explicit function of γ_j and β_t , the derivation of the derivatives will be very tedious. Hence, it is not easy to use the traditional variance estimator v_L in this problem. The jackknife variance estimator v_{JACK} , however, can be readily used.

Example 2.3. Degradation models. In the study of the reliability of a system component, very often we have the following type of degradation measurements (data):

$$y_{ij} = z(t_j)' \Theta_i + \varepsilon_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, m.$$

Here y_{ij} is the measurement of the i th sample component at time t_j ; $z(t)$ is a q -vector whose components are known functions of the time t ; Θ_i are unobservable random q -vectors that are i.i.d. with a q -variate normal distribution $N_q(\theta, \Sigma_\Theta)$, where θ and Σ_Θ are unknown; the ε_{ij} are i.i.d. measurement errors with mean 0 and variance σ_ε^2 ; and the Θ_i and ε_{ij} are mutually independent. As a function of t , $z(t)'\Theta$ is the degradation curve for a particular component and $z(t)'\theta$ is the *mean* degradation curve. Suppose that a component will fail to work if $z(t)'\Theta < \eta$, a given critical value. Assume that $z(t)'\Theta$ is always a decreasing function of t . Then the reliability function of a component is

$$R(t) = P\{z(t)'\Theta > \eta\} = \Phi\left(\frac{z(t)'\theta - \eta}{s(t)}\right),$$

where $s(t) = \sqrt{z(t)' \Sigma_\Theta z(t)}$ and Φ is the standard normal distribution function. For a fixed t , estimators of $R(t)$ can be obtained by estimating θ and Σ_Θ , since Φ is a known function. It can be shown that the best linear estimator of θ is the least squares estimator

$$\hat{\theta}_n = (Z'Z)^{-1} Z' \bar{Y},$$

where $Z = (z(t_1), \dots, z(t_m))'$, $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$, and $Y_i = (y_{i1}, \dots, y_{im})'$. The estimation of Σ_Θ is more difficult. An almost unbiased and consistent (as $n \rightarrow \infty$) estimator of Σ_Θ is

$$\frac{1}{n} \sum_{i=1}^n (Z'Z)^{-1} Z' (Y_i - \bar{Y}) (Y_i - \bar{Y})' Z (Z'Z)^{-1} - \hat{\sigma}_\varepsilon^2 (Z'Z)^{-1},$$

where

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n(m-q)} \sum_{i=1}^n [Y_i' Y_i - Y_i' Z (Z'Z)^{-1} Z' Y_i].$$

Hence, an estimator of $R(t)$ is

$$\hat{R}(t) = \Phi \left(\frac{z(t)' \hat{\theta}_n - \eta}{\hat{s}(t)} \right),$$

where

$$\hat{s}(t) = \left\{ \frac{1}{n} \sum_{i=1}^n [z(t)' (Z'Z)^{-1} Z' (Y_i - \bar{Y})]^2 - \hat{\sigma}_\varepsilon^2 z(t)' (Z'Z)^{-1} z(t) \right\}^{1/2}.$$

If we define $x_{i1} = z(t)' (Z'Z)^{-1} Z' Y_i$, $x_{i2} = [z(t)' (Z'Z)^{-1} Z' Y_i]^2$, $x_{i3} = [Y_i' Y_i - Y_i' Z (Z'Z)^{-1} Z' Y_i]/(m-q)$, and $X_i = (x_{i1}, x_{i2}, x_{i3})'$, then it is apparent that $\hat{R}(t)$ can be written as $g(\bar{X}_n)$ for a function

$$g(x_1, x_2, x_3) = \Phi \left((x_1 - \eta) / \sqrt{x_2 - x_1^2 - x_3 z(t)' (Z'Z)^{-1} z(t)} \right).$$

Suppose that ε_{ij} has a finite fourth moment, which implies the existence of $\Sigma = \text{var}(X_i)$. The asymptotic variance of $\hat{R}(t)$ can be immediately estimated by the jackknife estimator v_{JACK} . To use the traditional variance estimator v_L , the explicit formula of ∇g is needed. In this case, the derivation of ∇g is not complicated.

In the case where both v_L and v_{JACK} can be easily computed, an empirical simulation study may help us to decide which estimator should be adopted. As an example, we consider some empirical simulation results for the problem described in Example 2.1.

Table 2.1. Simulation results in Example 2.1. [Adapted from Efron (1982), by permission of Society for Industrial and Applied Mathematics]

Case	True SD	Jackknife				Linearization			
		ME	SD	RB	RM	ME	SD	RB	RM
1	.37	.37	.11	0%	.11	.35	.09	-5%	.09
2	.67	.70	.33	4%	.33	.53	.14	-21%	.20
3	.22	.22	.09	2%	.09	.18	.06	-20%	.07
4	.30	.31	.09	5%	.09	.24	.05	-18%	.08

Example 2.1 (continued). Simulation estimates of ME (mean of \sqrt{v} , $v = v_{\text{JACK}}$ or v_L) and SD (standard deviation of \sqrt{v}) obtained from Efron (1982, pp. 17-18), are given in Table 2.1. Table 2.1 also contains RB (relative bias of \sqrt{v} as an estimator of the true standard deviation) and RM (root mean squared error of \sqrt{v}). Four cases are considered. In cases 1 and 2, $T_n = \log \hat{\gamma}_n$; Y_i and Z_i are independent, Y_i is distributed as the uniform on $[0, 1]$, and Z_i is distributed as W in case 1 and as $W^2/2$ in case 2, where W is distributed as the exponential with mean 1; $n = 10$ and the simulation size is 100. In case 3, $T_n = \hat{\rho}_n$; (Y_i, Z_i) is distributed as the bivariate normal with $EY_i = EZ_i = 0$, $EY_i^2 = EZ_i^2 = 1$, and $\rho = EY_i Z_i = 0.5$; $n = 14$ and the simulation size is 200. The setting in case 4 is the same as that in case 3 except that $T_n = \hat{\phi}_n$.

A common feature for all four cases is that the linearization estimator is less variable but more biased than the jackknife estimator. In fact, the linearization estimator is badly downward-biased in cases 2–4. If a badly downward-biased variance estimator v is used in constructing the confidence interval $[T_n - z_{1-\alpha}\sqrt{v}, T_n + z_{1-\alpha}\sqrt{v}]$, where $z_{1-\alpha}$ is the $(1 - \alpha)$ th quantile of the standard normal distribution, then the coverage probability of the confidence interval may be seriously below its nominal level $1 - 2\alpha$. Therefore, in terms of stability we prefer v_L , whereas to have a conservative variance estimator or confidence interval we should choose v_{JACK} .

2.2 Variance Estimation for Functionals¹

In this section, we study jackknife variance estimation in a more general situation where $T_n = T(F_n)$. Here, F_n is the empirical distribution putting mass n^{-1} to each X_i and T is a real-valued functional defined on \mathcal{F} , a convex class of p -dimensional distributions containing F and all degenerate distributions. For the functions of the sample mean studied in Section 2.1, $T_n = g(\bar{X}_n)$ can be written as $T(F_n)$ with $T(G) = g(\int x dG(x))$, $G \in \mathcal{F}$.

¹This section may be omitted because it involves high level mathematics.

Other examples of $T_n = T(F_n)$ are M-estimators, L-statistics, R-estimators, and rank statistics (see Section 2.2.2). In fact, most (not all) commonly used statistics can be written as $T(F_n)$ for some T . Note that the consistency of v_{JACK} was established in Section 2.1 for *continuously differentiable* g when $T_n = g(\bar{X}_n)$. We now study the asymptotic behavior of v_{JACK} for $T_n = T(F_n)$ using a suitably defined differential of T .

2.2.1 Differentiability and consistency

There are several different versions of differentials for a given functional T on \mathcal{F} . The weakest differential of T is the Gâteaux differential. Let \mathcal{D} be the linear space generated by members of \mathcal{F} .

Definition 2.1. A functional T on \mathcal{F} is said to be Gâteaux differentiable at $G \in \mathcal{F}$ if there is a linear functional L_G on \mathcal{D} such that $H \in \mathcal{F}$ and $G + t(H - G) \in \mathcal{F}$ imply

$$\lim_{t \rightarrow 0} \left[\frac{T(G + t(H - G)) - T(G)}{t} - L_G(H - G) \right] = 0.$$

$L_G(H - G)$ is called the Gâteaux differential of T at G in the direction of $H - G$. If we define $h(t) = T(G + t(H - G))$, then $L_G(H - G)$ is simply the usual derivative of $h(t)$ at $t = 0$. Let δ_x denote the p -dimensional distribution degenerated at the point x and $\phi_G(x) = L_G(\delta_x - G)$. Then $\phi_F(x)$ is the influence function of T at F (see Hampel, 1974). It is assumed throughout that $E[\phi_F(X_1)] = 0$ and $E[\phi_F(X_1)]^2 < \infty$.

Gâteaux differentiability is too weak to be useful in asymptotic analysis. We consider the following stronger differentiability. Let ρ be a metric (distance) on \mathcal{D} .

Definition 2.2. A functional T on \mathcal{F} is said to be ρ -Hadamard differentiable at $G \in \mathcal{F}$ if there is a linear functional L_G on \mathcal{D} such that for any sequence of numbers $t_k \rightarrow 0$ and $\{D, D_k, k = 1, 2, \dots\} \subset \mathcal{D}$ satisfying $\rho(D_k, D) \rightarrow 0$ and $G + t_k D_k \in \mathcal{F}$,

$$\lim_{k \rightarrow \infty} \left[\frac{T(G + t_k D_k) - T(G)}{t_k} - L_G(D_k) \right] = 0.$$

Hadamard differentiability is also referred to as compact differentiability and is clearly stronger than Gâteaux differentiability. In some cases, we need an even stronger differentiability.

Definition 2.3. A functional T on \mathcal{F} is said to be ρ -Fréchet differentiable at $G \in \mathcal{F}$ if there is a linear functional L_G on \mathcal{D} such that for any sequence

$\{G_k\}$ satisfying $G_k \in \mathcal{F}$ and $\rho(G_k, G) \rightarrow 0$,

$$\lim_{k \rightarrow \infty} \frac{T(G_k) - T(G) - L_G(G_k - G)}{\rho(G_k, G)} = 0.$$

The most commonly used metric on \mathcal{D} is the one generated by the sup-norm, i.e., for any D_1 and $D_2 \in \mathcal{D}$,

$$\rho_\infty(D_1, D_2) = \|D_1 - D_2\|_\infty = \sup_x |D_1(x) - D_2(x)|.$$

However, it is necessary to consider other metrics. For a function of the mean $T(G) = g(\int x dG(x))$ with a differentiable function g , T is not necessarily ρ_∞ -Hadamard differentiable but is ρ_1 -Fréchet differentiable, where for any positive integer r , ρ_r is defined by

$$\rho_r(D_1, D_2) = \|D_1 - D_2\|_r = \left[\int |D_1(x) - D_2(x)|^r dx \right]^{1/r}.$$

Note that if ρ and $\tilde{\rho}$ are two metrics on \mathcal{D} satisfying $\tilde{\rho}(D_1, D_2) \leq c\rho(D_1, D_2)$ for a constant c and all $D_1, D_2 \in \mathcal{D}$, then $\tilde{\rho}$ -Hadamard (Fréchet) differentiability implies ρ -Hadamard (Fréchet) differentiability. This suggests the use of the metric $\rho_{\infty+r}$ defined by

$$\rho_{\infty+r}(D_1, D_2) = \rho_\infty(D_1, D_2) + \rho_r(D_1, D_2). \quad (2.14)$$

There are functionals that are neither ρ_∞ -Hadamard differentiable nor ρ_r -Hadamard differentiable, but are $\rho_{\infty+r}$ -Hadamard differentiable.

It is known that the ρ_∞ -Hadamard differentiability of T at F implies

$$\sqrt{n}[T(F_n) - T(F)]/\sigma \rightarrow_d N(0, 1), \quad (2.15)$$

where $\sigma^2 = E[\phi_F(X_1)]^2$ (see, e.g., Fernholz, 1983). When T is ρ -Fréchet differentiable at F with a metric ρ satisfying

$$\rho(F_n, F) = O_p(n^{-1/2}), \quad (2.16)$$

result (2.15) also holds (see Appendix A.3 for the notation O_p and o_p). Note that (2.16) holds for ρ_∞ and for ρ_r if we assume

$$\int \{F(x)[1 - F(x)]\}^{r/2} dx < \infty. \quad (2.17)$$

Hence, (2.17) and $\rho_{\infty+r}$ -Fréchet differentiability at F imply (2.15).

However, even the Fréchet differentiability of T does not ensure the consistency of the jackknife estimator v_{JACK} . For example, $T(G) = g(\int x dG)$ is

ρ_1 -Fréchet differentiable at F if g is differentiable at $\mu = \int x dF$. But if the derivative of g is not continuous at μ , v_{JACK} is not necessarily consistent. The proof of Theorem 2.1 indicates that the consistency of v_{JACK} requires that T be differentiable *continuously* in some sense.

Definition 2.4. A functional T is continuously Gâteaux differentiable at $G \in \mathcal{F}$ if T is Gâteaux differentiable at G and for any sequences of numbers $t_k \rightarrow 0$ and $G_k \in \mathcal{F}$ satisfying $\rho_\infty(G_k, G) \rightarrow 0$,

$$\lim_{k \rightarrow \infty} \left[\frac{T(G_k + t_k(\delta_x - G_k)) - T(G_k)}{t_k} - L_G(\delta_x - G_k) \right] = 0$$

uniformly in x , where δ_x is the distribution degenerated at the point x .

Example 2.4. Convolution functionals. Suppose that $p = 1$ and for a fixed $z \in \mathbb{R}$,

$$T(G) = \int G(z - y) dG(y), \quad G \in \mathcal{F}.$$

Note that $T(G)$ is the convolution of G evaluated at z . It can be shown that T is ρ_∞ -Hadamard differentiable at any $G \in \mathcal{F}$. Hence, $T(F_n)$ satisfies (2.1) with some σ_n^2 . For sequences $t_k \rightarrow 0$ and $G_k \in \mathcal{F}$ with $\rho_\infty(G_k, G) \rightarrow 0$, let $D_k = \delta_x - G_k$ and $H_k = G_k + t_k(\delta_x - G_k)$. Then

$$T(H_k) - T(G_k) = 2t_k \int D_k(z - y) dG_k(y) + t_k^2 \int D_k(z - y) dD_k(y),$$

where the last term is $O(t_k^2)$. Let $L_G(D) = 2 \int D(z - y) dG(y)$. Then

$$T(H_k) - T(G_k) = t_k L_G(D_k) + 2t_k \int D_k(z - y) d(G_k - G)(y) + O(t_k^2).$$

Hence, T is continuously Gâteaux differentiable at G , since

$$\left| \int D_k(z - y) d(G_k - G)(y) \right| \leq \rho_\infty(G_k, G) \| \delta_x - G_k \|_v \leq \rho_\infty(G_k, G),$$

where $\|\cdot\|_v$ is the total variation norm (Serfling, 1980, p. 254).

When the metric ρ_∞ is used, continuous Gâteaux differentiability is just enough for establishing the consistency of v_{JACK} . If a metric other than ρ_∞ is considered, we may need a stronger differentiability.

Definition 2.5. A functional T is continuously ρ -Fréchet differentiable at $G \in \mathcal{F}$ if T is ρ -Fréchet differentiable at G and $\rho(G_k, G) \rightarrow 0$ and $\rho(H_k, G) \rightarrow 0$ imply

$$\lim_{k \rightarrow \infty} \frac{T(H_k) - T(G_k) - L_G(H_k - G_k)}{\rho(H_k, G_k)} = 0.$$

It can be verified that continuous ρ_∞ -Fréchet differentiability implies continuous Gâteaux differentiability. Similarly, we can define continuous ρ -Hadamard differentiability of T at G by replacing $T(G + t_k D_k) - T(G)$ in Definition 2.2 by $T(G_k + t_k D_k) - T(G_k)$ with G_k satisfying $\rho(G_k, G) \rightarrow 0$.

Theorem 2.3. *Assume that T is continuously Gâteaux differentiable at F with $\phi_F \not\equiv 0$. Then the jackknife estimator v_{JACK} is strongly consistent, i.e.,*

$$n v_{\text{JACK}} / \sigma^2 \rightarrow_{a.s.} 1, \quad (2.18)$$

where $\sigma^2 = E[\phi_F(X_1)]^2$ and $\phi_F(x)$ is the influence function of T .

Proof. Let $F_{n-1,i}$ be the empirical distribution based on $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$. Note that $F_{n-1,i} = F_n + t_n(\delta_{X_i} - F_n)$ with $t_n = -1/(n-1)$. Then the continuous differentiability of T and the fact that $\rho_\infty(F_n, F) \rightarrow_{a.s.} 0$ imply

$$\frac{T(F_{n-1,i}) - T(F_n)}{t_n} - L_F(\delta_{X_i} - F_n) \rightarrow_{a.s.} 0 \quad \text{uniformly in } i.$$

Hence,

$$\max_{i \leq n} |(n-1)[T(F_n) - T(F_{n-1,i})] - (Z_i - \bar{Z}_n)| \rightarrow_{a.s.} 0,$$

where $Z_i = \phi_F(X_i)$ and $\bar{Z}_n = n^{-1} \sum_{i=1}^n Z_i$. Then,

$$nv_{\text{JACK}} = \frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z}_n)^2 + o(1) \quad a.s.,$$

and the result (2.18) follows from the strong law of large numbers. \square

Theorem 2.4. *Let ρ be a metric on \mathcal{D} . Assume that ρ satisfies*

$$\rho(F_n, F) \rightarrow_{a.s.} 0 \quad \text{and} \quad \sum_{i=1}^n [\rho(F_{ni}, F_n)]^2 = O(n^{-1}) \quad a.s. \quad (2.19)$$

If T is continuously ρ -Fréchet differentiable at F with $\phi_F \not\equiv 0$, then (2.18) holds.

Proof. Let $Z_i = \phi_F(X_i)$, $\bar{Z}_{n-1,i} = (n-1)^{-1} \sum_{j \neq i} Z_j$, $R_{ni} = T(F_{n-1,i}) - T(F_n) - \bar{Z}_{n-1,i} + \bar{Z}_n$, and $\bar{R} = n^{-1} \sum_{i=1}^n R_{ni}$. Then

$$\begin{aligned} nv_{\text{JACK}} &= (n-1)^{-1} \sum_{i=1}^n (Z_i - \bar{Z}_n)^2 + (n-1) \sum_{i=1}^n (R_{ni} - \bar{R})^2 \\ &\quad + 2(n-1) \sum_{i=1}^n R_{ni} (\bar{Z}_{n-1,i} - \bar{Z}_n). \end{aligned}$$

From the strong law of large numbers, $(n-1)^{-1} \sum_{i=1}^n (Z_i - \bar{Z}_n)^2 \rightarrow_{a.s.} \sigma^2$. It remains to be shown that $(n-1) \sum_{i=1}^n R_{ni}^2 \rightarrow_{a.s.} 0$. From (2.19),

$$\max_{i \leq n} \rho(F_{ni}, F) \leq \rho(F_n, F) + \max_{i \leq n} \rho(F_{ni}, F_n) \rightarrow_{a.s.} 0.$$

If T is continuously ρ -Fréchet differentiable at F , then for any $\epsilon > 0$,

$$R_{ni}^2 \leq \epsilon^2 [\rho(F_{ni}, F_n)]^2 \quad \text{for all } i \leq n \text{ and sufficiently large } n.$$

Thus, $(n-1) \sum_{i=1}^n R_{ni}^2 \leq \epsilon^2 (n-1) \sum_{i=1}^n [\rho(F_{ni}, F_n)]^2$ and $(n-1) \sum_{i=1}^n R_{ni}^2 \rightarrow_{a.s.} 0$ follows from (2.19). This proves (2.18). \square

Note that Theorem 2.4 is mainly for the situations where Theorem 2.3 is not applicable, i.e., T is not continuously Gâteaux differentiable and a metric ρ other than ρ_∞ has to be considered. The metric ρ , however, has to have property (2.19). Shao (1993a) shows that (2.19) holds for ρ_r if the second moment of F exists. Since $\rho_\infty(F_{ni}, F_n) \leq n^{-1}$, (2.19) also holds for the metric $\rho_{\infty+r}$ in (2.14). Thus, from Theorems 2.3 and 2.4, the jackknife estimator v_{JACK} is a consistent estimator of σ^2 when T is either continuously Gâteaux differentiable or continuously $\rho_{\infty+r}$ -Fréchet differentiable at F .

2.2.2 Examples

We now present some examples of functionals $T(F_n)$, and applications of Theorems 2.1–2.4 and their extensions.

Example 2.5. M-estimators. An M-functional $T(G)$ is defined to be a solution of

$$\int r(x, T(G)) dG(x) = \min_t \int r(x, t) dG(x),$$

where $r(x, t)$ is a real-valued function on \mathbb{R}^{p+1} . Let $\theta = T(F)$. $T(F_n)$ is called the M-estimator of θ . M-estimators are extensions of the maximum likelihood estimators. Examples can be found in Serfling (1980) and Huber (1981). Assume that $\psi(x, t) = \partial r(x, t)/\partial t$ exists and $\lambda_G(t) = \int \psi(x, t) dG(x)$ is well defined. Consequently, $\lambda_G(T(G)) = 0$. Assume further that $\lambda'_G = d\lambda_G/dt$ exists at $T(G)$ and $\lambda'_G(T(G)) \neq 0$. Then T is Gâteaux differentiable at G with $\phi_G(x) = -\psi(x, \theta)/\lambda'_G(\theta)$.

Shao (1993a) showed that T is continuously ρ_∞ -Fréchet differentiable at F under the following conditions:

- (a) $T(G_k) \rightarrow \theta$ for $\rho_\infty(G_k, F) \rightarrow 0$;
- (b) $\lambda'_{G_k}(\xi_k) \rightarrow \lambda'_F(\theta)$ for $\rho_\infty(G_k, F) \rightarrow 0$ and $\xi_k \rightarrow \theta$;
- (c) $\|\psi(\cdot, t) - \psi(\cdot, \theta)\|_v \rightarrow 0$ as $t \rightarrow \theta$; and

(d) $\lambda_H(T(G)) = O(\rho_\infty(H, G))$ (T is continuously $\rho_{\infty+1}$ -Fréchet differentiable at F if ρ_∞ is replaced by $\rho_{\infty+1}$).

Clarke (1983, 1986) established some results for the continuity of T that imply condition (a). Condition (b) is implied by A_4 in Clarke (1983) or A'_4 in Clarke (1986). In particular, condition (b) is satisfied if both $\psi(x, t)$ and $\partial\psi(x, t)/\partial t$ are bounded and continuous. A sufficient condition for (d) is $\|\psi(\cdot, \theta)\|_v < \infty$, since $|\lambda_H(T(G))| \leq \|\psi(\cdot, T(G))\|_v \rho_\infty(H, G)$.

Thus, by Theorems 2.3 and 2.4, the jackknife variance estimator v_{JACK} is strongly consistent for M-estimators. The consistency of v_{JACK} can also be established if condition (a) is replaced by a weaker condition,

$$\max_{i \leq n} |T(F_{ni}) - \theta| \rightarrow_{a.s.} 0, \quad (2.20)$$

which is a necessary condition for the consistency of v_{JACK} if $T(F_n) \rightarrow_{a.s.} \theta$. For example, (2.20) holds if ψ is nondecreasing in t and there is a neighborhood N_θ of θ such that for each fixed x , $\psi(x, t)$ is continuous on N_θ , $|\psi(x, t)| \leq M(x)$ for $t \in N_\theta$, and $E[M(X_1)] < \infty$.

Reeds (1978) also proved the consistency of v_{JACK} , assuming that there is a function $\tilde{M}(x)$ satisfying $E[\tilde{M}(X_1)] < \infty$ and

$$|\partial\psi(x, t)/\partial t - \partial\psi(x, s)/\partial s| \leq \tilde{M}(x)|t - s|^a$$

for all t, s and some constant $a > 0$.

Example 2.6. Smooth L-statistics. An L-functional is defined as

$$T(G) = \int x J(G(x)) dG(x), \quad G \in \mathcal{F}, \quad (2.21)$$

where $J(t)$ is a function on $[0, 1]$. $T(F_n)$ is called an L-statistic or L-estimator of $T(F)$. When $J \equiv 1$, $T(F_n)$ is simply the sample mean. When $J(t) = (\beta - \alpha)^{-1}$ for $\alpha \leq t \leq \beta$ and $J(t) = 0$ otherwise, $T(F_n)$ is the trimmed mean. When $J(t) = 4t - 2$, $T(F_n)$ is Gini's mean difference. Other examples can be found in Serfling (1980). We consider smooth J functions (the corresponding L-statistics are smooth L-statistics).

An L-statistic is trimmed if the function $J(t) = 0$ for $t < \alpha$ or $t > \beta$, $0 < \alpha < \beta < 1$. Parr (1985) proved that for a trimmed L-statistic, if J is bounded and continuous a.e. Lebesgue and a.e. F^{-1} , then the corresponding functional T is continuously ρ_∞ -Fréchet differentiable at F with

$$\phi_F(x) = \int [F(y) - I\{y \geq x\}] J(F(y)) dy.$$

Since continuous ρ_∞ -Fréchet differentiability implies continuous Gâteaux differentiability, the jackknife estimator v_{JACK} is strongly consistent by Theorem 2.3.

For an untrimmed L-statistic, T may not be ρ_∞ -Hadamard differentiable. In general, T is also not necessarily ρ_1 -Hadamard differentiable. Using the metric $\rho_{\infty+1}$ in (2.14), Shao (1993a) showed that if J is bounded, continuous a.e. Lebesgue and a.e. F^{-1} , and continuous on $[0, \alpha] \cup (1 - \alpha, 1]$ for a constant $\alpha > 0$, then T is continuously $\rho_{\infty+1}$ -Fréchet differentiable at F . Hence, by Theorem 2.4, v_{JACK} is strongly consistent for an L-statistic $T(F_n)$ when J satisfies the conditions previously stated and $E\|X_1\|^2 < \infty$ [which ensures (2.19) for $\rho_{\infty+1}$].

The consistency of v_{JACK} for L-statistics was also established by Parr and Schucany (1982) using a different approach.

Example 2.7. Linear rank statistics. Let $\mathcal{F} = \{\text{all distributions on } \mathbb{R}\}$ and, for $G \in \mathcal{F}$,

$$T(G) = \int_0^\infty J(\tilde{G}(x))dG(x), \quad (2.22)$$

where J is a differentiable function on $[0, 1]$ and satisfies $J(1 - t) = -J(t)$, and

$$\tilde{G}(x) = G(x) - G((-x)-), \quad x \geq 0.$$

$T(F_n)$ is then a linear rank statistic. Note that the Wilcoxon signed rank statistic and the Winsorized signed rank statistic are special cases of $T(F_n)$. For any $G \in \mathcal{F}$, T in (2.22) is ρ_∞ -Hadamard differentiable at G and

$$L_G(D) = \int_0^\infty J'(\tilde{G}(x))\tilde{D}(x)dG(x) + \int_0^\infty J(\tilde{G}(x))dD(x), \quad D \in \mathcal{D}.$$

In general, T is not ρ_∞ -Fréchet differentiable. Shao (1993a) showed that T is continuously Gâteaux differentiable at F if J' is continuous on $[0, 1]$ and $\|J'\|_v < \infty$. Therefore, Theorem 2.3 is applicable to linear rank statistics.

Example 2.8. Cramér-von Mises test statistic. Let F_0 be a specified hypothetical distribution and

$$T(G) = \int [G(x) - F_0(x)]^2 dF_0(x).$$

$T(F_n)$ is then the Cramér-von Mises test statistic for the test problem:

$$H_0: F = F_0 \quad \text{versus} \quad H_1: F \neq F_0.$$

Let $D = H - G$ and $L_F(D) = 2 \int D(x)[F(x) - F_0(x)]dF_0(x)$. Then,

$$\begin{aligned} |T(H) - T(G) - L_F(D)| &= \left| \int D(x)(H + G - 2F)(x)dF_0(x) \right| \\ &\leq \rho_\infty(H, G)[\rho_\infty(H, F) + \rho_\infty(G, F)]. \end{aligned}$$

Hence, T is continuously ρ_∞ -Fréchet differentiable at F , and, by Theorem 2.3, v_{JACK} is strongly consistent for $T(F_n)$.

Example 2.9. Two-sample linear rank statistics. Consider a test problem concerning two unknown q -variate distributions F_X and F_Y :

$$H_0 : F_X = F_Y \quad \text{versus} \quad H_1 : F_X \neq F_Y. \quad (2.23)$$

Let $\{X_1, \dots, X_n\}$ be an i.i.d. sample from F_X and $\{Y_1, \dots, Y_m\}$ be an i.i.d. sample from F_Y . We assume that $n = m$, but we do not assume that $\{X_1, \dots, X_n\}$ and $\{Y_1, \dots, Y_m\}$ are independent. However, the following discussion can be modified and applied to the case where $n \neq m$, $n/m \rightarrow \lambda \in (0, 1)$, and $\{X_1, \dots, X_n\}$ and $\{Y_1, \dots, Y_m\}$ are independent.

Let \mathcal{F} be the collection of all distributions on \mathbb{R}^{2q} , and, for any $G \in \mathcal{F}$,

$$T(G) = \int J(\bar{G}(x)) dG(x, \infty), \quad (2.24)$$

where $\bar{G}(x) = [G(x, \infty) + G(\infty, x)]/2$ and J is the same as that in (2.22). Let F be the joint distribution of (X_1, Y_1) . Then, $\bar{F}(x) = [F_X(x) + F_Y(x)]/2$. Note that $T(F)$ can be used as a measure of the closeness of F_X and F_Y . Let F_n be the empirical distribution putting mass n^{-1} to (X_i, Y_i) , $i = 1, \dots, n$. $T(F_n)$ is called a two-sample linear rank statistic. The test procedure based on $T(F_n)$ rejects H_0 in (2.23) if $|T(F_n)|$ is large. The well-known two-sample Wilcoxon statistic is a special case of $T(F_n)$ with $J(t) = t - \frac{1}{2}$.

Similar to the one-sample linear rank statistics in (2.22), it can be shown that, if J' exists, T in (2.24) is ρ_∞ -Hadamard differentiable at any $G \in \mathcal{F}$ with the influence function

$$\phi_G(x, y) = \int J'(\bar{G}(u)) (\bar{\delta}_{x,y} - \bar{G})(u) dG(u, \infty) + J(\bar{G}(x)) - T(G);$$

and, furthermore, if J' is continuous on $[0, 1]$ and $\|J'\|_v < \infty$, T is also continuously Gâteaux differentiable. Hence, Theorem 2.3 is applicable to the two-sample linear rank statistics.

Example 2.10. R-estimators. R-estimators are closely related to rank statistics. Let $\mathcal{F} = \{\text{all distributions on } \mathbb{R}\}$ and $T(G)$ be a solution of

$$\int J([G(x) + 1 - G(2t - x)]/2) dG(x) = 0, \quad G \in \mathcal{F},$$

where J is the same as that in (2.22). $T(F_n)$ is called an R-estimator of $T(F)$. Assume that F is symmetric about $T(F)$ and F has a density f . Then, T is ρ_∞ -Hadamard differentiable at F with the influence function

$$\phi_F(x) = J(F(x)) \left/ \int J'(F(x)) f(x) dF(x) \right..$$

However, it is unknown whether T is continuously Gâteaux differentiable. Hence, Theorem 2.3 is not applicable to R-estimators. We will return to the discussion of the consistency of v_{JACK} for R-estimators later.

Example 2.11. The statistics based on generalized empirical distributions. A generalized empirical distribution is defined by the U-statistic [see (2.9)]

$$\Xi_n(x) = \binom{n}{m}^{-1} \sum_c I\{h(X_{i_1}, \dots, X_{i_m}) \leq x\}, \quad (2.25)$$

where m is a fixed integer. Apparently, the empirical distribution F_n in (1.6) is a special case of (2.25) with $m = 1$ and $h(x) = x$. Let T be a functional on \mathcal{F} and $T_n = T(\Xi_n)$. The simplest example of this kind of statistic is given by the functional $T(G) = \int x dG(x)$; $T(\Xi_n)$ is then a U-statistic with kernel $h(x_1, \dots, x_m)$ given by (2.9). Another example is the Hodges–Lehmann estimator, $T_n = \Xi_n^{-1}(\frac{1}{2})$, with $h(x_1, x_2) = (x_1 + x_2)/2$. When T is an L-functional defined by (2.21), $T(\Xi_n)$ is a generalized L-statistic studied in Serfling (1984). The class of generalized L-statistics includes trimmed sample variance, trimmed U-statistics, and Winsorized U-statistics.

Serfling (1984) and Helmers, Janssen and Serfling (1988) established some asymptotic results for generalized empirical distributions and the statistics based on them. In particular,

$$\sqrt{n}[T(\Xi_n) - T(\Xi)] \rightarrow_d N(0, \sigma^2) \quad (2.26)$$

for some σ^2 , where Ξ is the distribution of $h(X_{i_1}, \dots, X_{i_m})$.

Let $T_{n-1,i} = T(\Xi_{n-1,i})$, $i = 1, \dots, n$, where $\Xi_{n-1,i}$ is defined by (2.25) with the i th observation X_i removed. The jackknife estimator v_{JACK} in (1.13) can be used to estimate σ^2/n , the asymptotic variance of T_n . Although Theorem 2.4 cannot be applied to this case directly, the proof of the following result is very similar to that of Theorem 2.4 (using the result in Theorem 2.2) and, therefore, is omitted here.

Theorem 2.5. Suppose that T is continuously ρ_∞ -Fréchet differentiable at Ξ and that the kernel of the “U-statistic” $L_\Xi(\Xi_n - \Xi)$ has a nonzero finite second moment. Then the jackknife estimator v_{JACK} is strongly consistent for σ^2/n , where σ^2 is given in (2.26). The result still holds if T is continuously $\rho_{\infty+1}$ -Fréchet differentiable at Ξ and $E[h(X_1, \dots, X_m)]^2 < \infty$.

Finally, we consider a very large class of statistics formed by transformations of statistics given in the previous examples. That is, $T_n = g(\mathbf{T}_n)$ with $\mathbf{T}_n = \mathbf{T}(F_n)$, where \mathbf{T} is a q -vector of functionals on \mathcal{F} and g is a function from \mathbb{R}^q to \mathbb{R} . Suppose that

$$\sqrt{n}[\mathbf{T}_n - \mathbf{T}(F)] \rightarrow_d N_q(0, V), \quad (2.27)$$

where $N_q(0, V)$ is a q -variate normal random vector with mean 0 and covariance matrix V . Then (2.1) holds for T_n with $\sigma_n^2 = n^{-1} \nabla g(\theta)' V \nabla g(\theta)$, where $\theta = \mathbf{T}(F)$. The following result is an extension of Theorem 2.1.

Theorem 2.6. *Suppose that (2.27) holds with $V > 0$, ∇g is continuous at θ , $\nabla g(\theta) \neq 0$, and the components of \mathbf{T} satisfy the conditions in either Theorem 2.3 or Theorem 2.4. Then the jackknife estimator v_{JACK} is strongly consistent, i.e., (2.3) holds with $\sigma_n^2 = n^{-1} \nabla g(\theta)' V \nabla g(\theta)$.*

Proof. Note that the proofs of Theorems 2.3 and 2.4 can be extended in a straightforward manner to the case where \mathbf{T} is a vector of functionals satisfying the conditions of the theorems. That is,

$$(n-1) \sum_{i=1}^n [\mathbf{T}(F_{n-1,i}) - \bar{\mathbf{T}}_n][\mathbf{T}(F_{n-1,i}) - \bar{\mathbf{T}}_n]' \rightarrow_{a.s.} V, \quad (2.28)$$

where $\bar{\mathbf{T}}_n = n^{-1} \sum_{i=1}^n \mathbf{T}(F_{n-1,i})$, and

$$(n-1) \sum_{i=1}^n \|\mathbf{T}(F_{n-1,i}) - \bar{\mathbf{T}}_n\|^2 \rightarrow_{a.s.} \text{tr}(V).$$

It follows from (2.28) and the continuity of ∇g that

$$(n-1) \sum_{i=1}^n \{[\mathbf{T}(F_{n-1,i}) - \bar{\mathbf{T}}_n]' \nabla g(\bar{\mathbf{T}}_n)\}^2 \rightarrow_{a.s.} \nabla g(\theta)' V \nabla g(\theta).$$

The rest of the proof is similar to that of Theorem 2.1 and is omitted. \square

2.2.3 Convergence rate

The accuracy of a variance estimator can be described by the convergence rate and some measure of the asymptotic efficiency of the variance estimator. Consider the simple case where $T_n = \bar{X}_n$ and $p = 1$. Then, $v_{\text{JACK}} = [n(n-1)]^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ and

$$n^{3/2} [v_{\text{JACK}} - \text{var}(\bar{X}_n)] \rightarrow_d N(0, \tau)$$

or, equivalently,

$$\sqrt{n} (nv_{\text{JACK}} - \sigma^2) \rightarrow_d N(0, \tau),$$

where $\tau = \text{var}(X_1 - EX_1)^2$ and $\sigma^2 = \text{var}(X_1)$. Hence, the convergence rate of v_{JACK} is $n^{-1/2}$ [note that both v_{JACK} and $\text{var}(\bar{X}_n)$ are of the order n^{-1}] and τ is a measure of the asymptotic efficiency of v_{JACK} . For the general differentiable functional T , τ should be defined by

$$\tau = \text{var}[\phi_F(X_1)]^2, \quad (2.29)$$

since $\phi_F(X_i)$ plays the same role as $X_i - EX_i$. We need to define a concept similar to the Lipschitz continuity of the derivative of a function on \mathbb{R}^p .

Definition 2.6. A functional T on \mathcal{F} is said to be ρ -Lipschitz differentiable of order $\delta > 0$ at $G \in \mathcal{F}$ if there is a linear functional L_G on \mathcal{D} such that, for any H_k and $G_k \in \mathcal{F}$ satisfying $\rho(H_k, G) \rightarrow 0$ and $\rho(G_k, G) \rightarrow 0$,

$$T(H_k) - T(G_k) - L_G(H_k - G_k) = O([\rho(H_k, G_k)]^{1+\delta}).$$

Note that ρ -Lipschitz differentiability implies continuous ρ -Fréchet differentiability since $\delta > 0$.

We now provide two examples of functionals that are Lipschitz differentiable. The first example is functions of the mean: $T(G) = g(\int x dG(x))$ with a differentiable function g from \mathbb{R}^p to \mathbb{R} . If ∇g is Lipschitz continuous of order δ in the sense that $\|\nabla g(t) - \nabla g(s)\| \leq c\|t - s\|^\delta$ for a constant $c > 0$ and for all t and s in a neighborhood of $\mu = EX_1$, then T is ρ_1 -Lipschitz differentiable of order $\delta > 0$ at F . The second example is the L-functional defined in (2.21). If the L-functional T is trimmed and the function J in (2.21) is Lipschitz continuous of order $\delta > 0$ on $[\alpha, \beta] \subset [0, 1]$, then T is ρ_∞ -Lipschitz differentiable of order δ at F . Furthermore, τ in (2.29) is finite since $\phi_F(x)$ is bounded. For an untrimmed L-functional T , if J is Lipschitz continuous of order $\delta > 0$ on $[0, 1]$, then T is $\rho_{\infty+1}$ -Lipschitz differentiable of order δ at F . Furthermore, τ in (2.29) is finite if $E\|X_i\|^4 < \infty$.

For the convergence rate and asymptotic efficiency of v_{JACK} , we have the following result.

Theorem 2.7. Suppose that T is ρ_∞ -Lipschitz differentiable of order $\delta > 0$ at F and τ in (2.29) is finite. Then

$$nv_{\text{JACK}} - \sigma^2 = O_p(c_n),$$

where $\sigma^2 = E[\phi_F(X_1)]^2$ and $c_n = \max(n^{-\delta}, n^{-1/2})$; and, if $\delta > \frac{1}{2}$,

$$\sqrt{n}(nv_{\text{JACK}} - \sigma^2) \rightarrow_d N(0, \tau).$$

The same results hold if T is $\rho_{\infty+1}$ -Lipschitz differentiable of order $\delta > 0$ at F and $E\|X_1\|^{2(1+\delta)} < \infty$.

Proof. Let $Z_i = \phi_F(X_i)$, $\bar{Z} = n^{-1} \sum_{i=1}^n Z_i$, $s_n^2 = (n-1)^{-1} \sum_{i=1}^n (Z_i - \bar{Z})^2$, and $R_{ni} = (n-1)(T_n - T_{n-1,i}) - (Z_i - \bar{Z})$. From the Lipschitz differentiability of T ,

$$R_{ni} = O(n[\rho(F_{n-1,i}, F_n)]^{1+\delta}) \quad a.s. \quad \text{uniformly in } i,$$

where ρ is either ρ_∞ or $\rho_{\infty+1}$. Then,

$$\frac{1}{n-1} \sum_{i=1}^n R_{ni}^2 = O(n) \sum_{i=1}^n [\rho(F_{n-1,i}, F_n)]^{2(1+\delta)} \quad a.s.$$

For ρ_∞ ,

$$\sum_{i=1}^n [\rho_\infty(F_{n-1,i}, F_n)]^{2(1+\delta)} \leq \frac{n}{(n-1)^{2(1+\delta)}}.$$

For ρ_1 , it can be shown (Shao, 1992a) that

$$\sum_{i=1}^n [\rho_1(F_{n-1,i}, F_n)]^{2(1+\delta)} = O(n^{-2\delta-1}) \text{ a.s.}$$

Therefore, in all cases,

$$\frac{1}{n-1} \sum_{i=1}^n R_{ni}^2 = O(n^{-2\delta}) \text{ a.s.}$$

Note that

$$\sqrt{n}(s_n^2 - \sigma^2) \rightarrow_d N(0, \tau)$$

and

$$\left| \frac{1}{n-1} \sum_{i=1}^n R_{ni}(Z_i - \bar{Z}) \right| \leq s_n \left(\frac{1}{n-1} \sum_{i=1}^n R_{ni}^2 \right)^{1/2} = O(n^{-\delta}) \text{ a.s.}$$

Therefore,

$$nv_{\text{JACK}} - \sigma^2 = s_n^2 - \sigma^2 + O_p(n^{-\delta}) = O_p(n^{-1/2}) + O_p(n^{-\delta}).$$

If $\delta > \frac{1}{2}$, $O_p(n^{-\delta}) = o_p(n^{-1/2})$. This completes the proof. \square

2.2.4 Other differential approaches

There are two other differential approaches that are useful in establishing the consistency of jackknife variance estimators.

Second order differentiable functionals

The key role of the differentiability of a functional T is that it ensures that $T(G) - T(F)$ can be approximated by a linear functional up to a certain accuracy. We may also approximate $T(G) - T(F)$ by a quadratic functional, which leads to the “second order differential” approach introduced by Beran (1984a) and furthered in Sen (1988), Shao and Wu (1989), and Shao (1991a).

Definition 2.7. A functional T on \mathcal{F} is said to be second order ρ -Hadamard differentiable at $G \in \mathcal{F}$ if there is a quadratic functional Q_G on \mathcal{D} such that for any $t_k \rightarrow 0$, $D_k \in \mathcal{D}$, $D \in \mathcal{D}$, and $\rho(D_k, D) \rightarrow 0$,

$$\lim_{k \rightarrow \infty} \frac{T(G + t_k D_k) - T(G) - Q_G(t_k D_k)}{t_k^2} = 0.$$

T is said to be second order ρ -Fréchet differentiable at G if for any $G_k \in \mathcal{F}$ with $\rho(G_k, G) \rightarrow 0$,

$$\lim_{k \rightarrow \infty} \frac{T(G_k) - T(G) - Q_G(G_k - G)}{[\rho(G_k, G)]^2} = 0.$$

In most cases, $Q_G(D)$ can be expressed as

$$Q_G(D) = \iint \psi_G(x, Y) d(G+D)(x) d(G+D)(y) \quad (2.30)$$

for a function ψ_G on \mathbb{R}^2 satisfying $\iint \psi_G(x, y) dG(x) dG(y) = 0$ and $\psi_G(x, y) = \psi_G(y, x)$.

The reader should not have the impression that second order differentiability is stronger than the (first order) differentiability that we studied in the previous sections. In fact, there is no definite relation between these two kinds of differentiability. The best example of a second order differentiable T which is not first order differentiable is the variance functional

$$T(G) = \frac{1}{2} \iint (x - y)^2 dG(x) dG(y), \quad G \in \mathcal{F}.$$

T is not ρ_∞ -Hadamard differentiable according to Definition 2.2, but it is second order ρ_∞ -Fréchet differentiable at any $G \in \mathcal{F}$, since T is a quadratic functional.

If T is second order ρ_∞ -Hadamard differentiable at F , (2.1) holds with

$$\sigma_n^2 = \frac{1}{n} E \left[2 \int \psi_F(X_1, y) dF(y) \right]^2, \quad (2.31)$$

provided that

$$E[\psi_F(X_1, X_2)]^2 < \infty, \quad E|\psi_F(X_1, X_1)| < \infty. \quad (2.32)$$

Result (2.1) also holds if T is second order $\rho_{\infty+r}$ -Fréchet differentiable at F and (2.17) holds (see, e.g., Shao, 1991a).

We have the following result concerning the consistency of v_{JACK} for a second order differentiable T :

Theorem 2.8. *Suppose that T is second order ρ_∞ -Fréchet differentiable at F with $\psi_F(x, y)$ satisfying (2.32) and $\int \psi_F(x, y) dF(y) \not\equiv 0$. Then the jackknife estimator v_{JACK} is weakly consistent, i.e.,*

$$v_{\text{JACK}} / \sigma_n^2 \rightarrow_p 1 \quad (2.33)$$

with σ_n^2 given by (2.31). Result (2.33) also holds if ρ_∞ is replaced by $\rho_{\infty+r}$ and $E\|X_1\|^{2/r} < \infty$.

Proof. We prove the first assertion for illustration. The proof of the second assertion is in Shao (1991a). From the second order differentiability of T ,

$$T(F_n) - T(F) = V_n + R_n,$$

where

$$V_n = \iint \psi_F(x, y) dF_n(x) dF_n(y) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \psi_F(X_i, X_j)$$

and R_n is the remainder satisfying $R_n / [\rho(F_n, F)]^2 \rightarrow_p 0$. Similarly,

$$T(F_{n-1,i}) - T(F) = V_{n-1,i} + R_{n-1,i}, \quad i = 1, \dots, n.$$

Since V_n is a “V-statistic” (see Section 2.1.1), its jackknife variance estimator is consistent. Also, σ_n^2 in (2.31) is the asymptotic variance of V_n since $V_n / \sigma_n \rightarrow_d N(0, 1)$. Therefore, result (2.33) follows if we can show that

$$n \sum_{i=1}^n R_{n-1,i}^2 \rightarrow_p 0. \quad (2.34)$$

From the second order ρ_∞ -Fréchet differentiability of T , for any $\epsilon > 0$ there is a $\delta_\epsilon > 0$ such that, if $\rho_\infty(F_{n-1,i}, F) < \delta_\epsilon$, $|R_{n-1,i}| \leq \epsilon [\rho_\infty(F_{n-1,i}, F)]^2$. Hence, for any $\eta > 0$,

$$\begin{aligned} P\left\{ n \sum_{i=1}^n R_{n-1,i}^2 > \eta \right\} &\leq P\left\{ \max_{i \leq n} \rho_\infty(F_{n-1,i}, F) \geq \delta_\epsilon \right\} \\ &+ P\left\{ n \sum_{i=1}^n [\rho_\infty(F_{n-1,i}, F)]^4 \geq \eta/\epsilon \right\}. \end{aligned} \quad (2.35)$$

Note that

$$\begin{aligned} \max_{i \leq n} \rho_\infty(F_{n-1,i}, F) &\leq \rho_\infty(F_n, F) + \max_{i \leq n} \rho_\infty(F_{n-1,i}, F_n) \\ &\leq \rho_\infty(F_n, F) + n^{-1} \rightarrow_{a.s.} 0 \end{aligned}$$

and

$$\begin{aligned} n \sum_{i=1}^n [\rho_\infty(F_{n-1,i}, F)]^4 &\leq 4n^2 [\rho_\infty(F_n, F)]^4 + 4n \sum_{i=1}^n [\rho_\infty(F_{n-1,i}, F_n)]^4 \\ &\leq 4n^2 [\rho_\infty(F_n, F)]^4 + 4n^{-2} = O_p(1). \end{aligned}$$

Hence, (2.34) follows from (2.35) since η and ϵ are arbitrary. \square

Actually, Shao (1991a) proved (2.33) for a second order ρ_∞ -Hadamard differentiable T at F (the proof requires higher level mathematics and therefore we do not discuss it here).

Unlike the situation where we use first order differentials (Theorems 2.3–2.4), we *do not* require that T be *continuously* second order differentiable in establishing result (2.33).

Second order differentiability will be especially useful for studying jackknife bias estimators (see Section 2.4).

Uniformly differentiable functionals

For a function g on \mathbb{R} differentiable at x_0 , the continuous differentiability described in Section 2.2.1 is equivalent to

$$\lim_{x \rightarrow x_0, y \rightarrow x_0} \frac{g(x) - g(y) - \nabla g(x_0)(x - y)}{x - y} = 0,$$

which is true if ∇g is defined in a neighborhood of x_0 and is continuous at x_0 . Under these conditions on ∇g we also have

$$\lim_{\epsilon \rightarrow 0} \sup_{|x - x_0| < \epsilon, |y - x_0| < \epsilon} \left| \frac{g(x) - g(y) - \nabla g(y)(x - y)}{x - y} \right| = 0, \quad (2.36)$$

i.e., g is *uniformly* differentiable in neighborhoods of x_0 that shrink to $\{x_0\}$. Note that (2.36) is equivalent to

$$\lim_{k \rightarrow \infty} \frac{g(x_k) - g(y_k) - \nabla g(y_k)(x_k - y_k)}{x_k - y_k} = 0,$$

where $\{x_k\}$ and $\{y_k\}$ are any sequences satisfying $x_k \rightarrow x_0$ and $y_k \rightarrow x_0$. This leads to the following definitions.

Definition 2.8. A functional T on \mathcal{F} is uniformly Gâteaux differentiable at $G \in \mathcal{F}$ if T is Gâteaux differentiable in a neighborhood of G and, for any sequences of numbers $t_k \rightarrow 0$ and $G_k \in \mathcal{F}$ satisfying $\rho_\infty(G_k, G) \rightarrow 0$,

$$\lim_{k \rightarrow \infty} \left[\frac{T(G_k + t_k(\delta_x - G_k)) - T(G_k)}{t_k} - \phi_{G_k}(x) \right] = 0 \quad (2.37)$$

uniformly in x .

Definition 2.9. A functional T on \mathcal{F} is uniformly ρ -Fréchet differentiable at $G \in \mathcal{F}$ if T is ρ -Gâteaux differentiable in a neighborhood of G , and $\rho(G_k, G) \rightarrow 0$ and $\rho(H_k, G) \rightarrow 0$ imply

$$\lim_{k \rightarrow \infty} \frac{T(H_k) - T(G_k) - L_{G_k}(H_k - G_k)}{\rho(H_k, G_k)} = 0.$$

Neither uniform differentiability nor continuous differentiability implies the other. Uniform ρ_∞ -Fréchet differentiability implies uniform Gâteaux differentiability.

Since $F_{n-1,i} = F_n + t_n(\delta_{X_i} - F_n)$ with $t_n = -1/(n-1)$, from (2.37) we obtain (assuming that T is either uniformly Gâteaux differentiable at F or is uniformly $\rho_{\infty+1}$ -Fréchet differentiable at F and $E\|X_1\|^2 < \infty$) that

$$v_{\text{JACK}} = v_{\text{IJACK}} + o(n^{-1}) \quad a.s., \quad (2.38)$$

where

$$v_{\text{IJACK}} = \frac{1}{n^2} \sum_{i=1}^n [\phi_{F_n}(X_i)]^2. \quad (2.39)$$

The estimator v_{IJACK} in (2.39) is called the infinitesimal jackknife estimator (Jaeckel, 1972). From (2.38), v_{JACK} is consistent if and only if v_{IJACK} is consistent. Let

$$\sigma^2(G) = \int [\phi_G(x)]^2 dG(x), \quad G \in \mathcal{F}. \quad (2.40)$$

$\sigma^2(G)$ is well defined in a neighborhood of F if T is either uniformly Gâteaux differentiable or uniformly $\rho_{\infty+1}$ -Fréchet differentiable at F . Then the asymptotic variance of $T(F_n)$ is $\sigma_n^2 = \sigma^2(F)/n$ and $v_{\text{IJACK}} = \sigma^2(F_n)/n$. That is, the infinitesimal jackknife estimator v_{IJACK} is actually a substitution estimator.

Despite its name, the infinitesimal jackknife is closer to the traditional approach. This is because, for calculating v_{IJACK} , the theoretical forms of $\phi_G(x)$ and $\sigma^2(G)$ in (2.40) are required, and no repeated computation of $T_{n-1,i}$ is needed.

If the functional $\sigma^2(G)$ is continuous in the sense that

$$\lim_{\rho(G,F) \rightarrow 0} \sigma^2(G) = \sigma^2(F) \quad (2.41)$$

for a metric ρ with $\rho(F_n, F) \rightarrow_{a.s.} 0$, then v_{IJACK} is strongly consistent (Sen, 1988). However, (2.41) is too strong for the consistency of v_{IJACK} . The following result is established in Shao (1990a).

Theorem 2.9. *Assume that for any $c > 0$, $\sup_{\|x\| \leq c} |\phi_{F_n}(x) - \phi_F(x)| \rightarrow_{a.s.} 0$, and that there is a constant $c_0 > 0$ and a function $h(x) > 0$ such that $E[h(X_1)] < \infty$ and almost surely $[\phi_{F_n}(x)]^2 \leq h(x)$ for $\|x\| \geq c_0$ and all sufficiently large n . Then*

$$v_{\text{IJACK}} / \sigma_n^2 \rightarrow_{a.s.} 1. \quad (2.42)$$

Consequently, both (2.38) and (2.42) imply the strong consistency of the jackknife estimator v_{JACK} .

The reader may verify that the functions ϕ_G in many of the examples in Section 2.2.2 satisfy the conditions in Theorem 2.9. However, v_{JACK} is not well defined for R-estimators since ϕ_G does not exist when G is not differentiable.

2.3 The Delete-d Jackknife

Although we have shown the consistency of the jackknife variance estimator v_{JACK} for many statistics, there are cases where v_{JACK} is inconsistent. The best known examples of inconsistency are the sample quantiles. For the sample median,

$$v_{\text{JACK}}/\sigma_n^2 \xrightarrow{d} (\chi_2^2/2)^2,$$

where σ_n^2 is the asymptotic variance of the sample median and χ_2^2 is a chi-square random variable with 2 degrees of freedom (see Efron, 1982, Chapter 3). The main reason for this inconsistency is that the functionals that generate sample quantiles are not smooth enough, whereas, as we have seen in the previous sections, the consistency of v_{JACK} requires that the functional be smooth.

On the other hand, bootstrapping the sample quantiles does lead to consistent variance estimators under reasonable conditions on F (see Section 3.2.2). This is a major triumph of the bootstrap over the jackknife. Another advantage of the bootstrap over the jackknife is that the bootstrap provides distribution estimators (see Section 1.4).

To remove these two deficiencies of the jackknife, a more general version of the jackknife, the *delete- d jackknife*, was proposed and studied in Wu (1986, 1990), Shao (1988b), and Shao and Wu (1989). Using the delete- d jackknife, we repeatedly compute and use the statistics of the form

$$T_{r,s} = T_r(X_i, i \in s^c), \quad (2.43)$$

where s is a subset of $\{1, \dots, n\}$ with size d , s^c is the complement of s , d is an integer depending on n , $1 \leq d \leq n$, and $r = n - d$. Note that for the given statistic T_n , $T_{r,s}$ is the same statistic but is based on r observations that are obtained by removing $\{X_i, i \in s\}$ from the original data set. Clearly, this is an extension of the jackknife previously studied with $d \equiv 1$.

The delete- d jackknife for variance estimation is studied in Section 2.3.1. In general, the consistency of the delete- d jackknife variance estimator requires less stringent smoothness conditions on T_n than that of the delete-1 jackknife estimator v_{JACK} in (1.13). In particular, the delete- d jackknife variance estimators for sample quantiles are consistent when $d \rightarrow \infty$ as

$n \rightarrow \infty$ with a certain rate. It can be shown that the required number of observations deleted depends on some measures of smoothness of T_n . The less smooth T_n is, the larger d needs to be.

In Section 2.3.2, we study jackknife histograms constructed by using $T_{r,s}$ for all s with size d . The jackknife histogram provides an estimator of the sampling distribution of T_n . This estimator is consistent if and only if both r and d diverge to infinity. The convergence rate of the jackknife distribution estimator is also studied in some simple cases.

2.3.1 Variance estimation

To derive and motivate the delete-d jackknife variance estimator for a given T_n , we start with the simplest situation where jackknifing is not necessary: $T_n = \bar{X}_n$ and $p = 1$. For a subset $s_* \subset \{1, \dots, n\}$ of size d , $\{X_i, i \in s_*^c\}$ can be viewed as a simple random sample of size r without replacement from $\{X_i, i = 1, \dots, n\}$. Note that \bar{X}_{r,s_*} defined by (2.43) for $T_n = \bar{X}_n$ is the sample mean of $\{X_i, i \in s_*^c\}$. Hence, from the sampling theory, the variance of \bar{X}_{r,s_*} as an estimator of the “population” mean \bar{X}_n is

$$\text{var}_*(\bar{X}_{r,s_*}) = \frac{1 - f_n}{r(n-1)} \sum_{i=1}^n (X_i - \bar{X}_n)^2, \quad (2.44)$$

where $f_n = r/n$ is the “sampling fraction” and var_* is the variance taken under the sampling distribution of $\{X_i, i \in s_*\}$ as a simple random sample from $\{X_i, i = 1, \dots, n\}$. On the other hand, a direct computation gives that

$$\text{var}_*(\bar{X}_{r,s_*}) = \frac{1}{N} \sum_{s \in \mathcal{S}} \left(\bar{X}_{r,s} - \frac{1}{N} \sum_{s \in \mathcal{S}} \bar{X}_{r,s} \right)^2, \quad (2.45)$$

where \mathcal{S} is the collection of all the subsets of $\{1, \dots, n\}$ that have size d and $N = \binom{n}{d}$ is the total number of subsets in \mathcal{S} . Noting that $1 - f_n = d/n$, we obtain from (2.44) and (2.45) that

$$\frac{r}{dN} \sum_{s \in \mathcal{S}} \left(\bar{X}_{r,s} - \frac{1}{N} \sum_{s \in \mathcal{S}} \bar{X}_{r,s} \right)^2 = \frac{1}{n(n-1)} \sum_{i=1}^n (X_i - \bar{X}_n)^2. \quad (2.46)$$

Result (2.46) not only motivates the definition of delete-d jackknife variance estimators for more general T_n but also is useful in studying properties of the delete-d jackknife variance estimators.

For a given statistic $T_n = T_n(X_1, \dots, X_n)$, the delete-d jackknife variance estimator is defined to be

$$v_{\text{JACK-d}} = \frac{r}{dN} \sum_{s \in \mathcal{S}} \left(T_{r,s} - \frac{1}{N} \sum_{s \in \mathcal{S}} T_{r,s} \right)^2, \quad (2.47)$$

where $T_{r,s}$ is given by (2.43) and \mathcal{S} is the same as in (2.45). Clearly, when $d \equiv 1$, $v_{\text{JACK-}d}$ reduces to v_{JACK} in (1.13). Also, from (2.46), $v_{\text{JACK-}d} = v_{\text{JACK}} = [n(n-1)]^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ for any d if $T_n = \bar{X}_n$.

The computation of $v_{\text{JACK-}d}$ with a large d may be impractical since N is usually huge. This problem is similar to that of computing the bootstrap estimator v_{BOOT} in (1.16), and it can be resolved in much the same way, by using approximations similar to $v_{\text{BOOT}}^{(B)}$ in (1.18). We will study the computation of $v_{\text{JACK-}d}$ in Chapter 5.

To study the consistency of $v_{\text{JACK-}d}$, we assume that T_n admits the following expansion:

$$T_n = \theta + \frac{1}{n} \sum_{i=1}^n \phi_F(X_i) + R_n, \quad R_n = o_p(n^{-1/2}), \quad (2.48)$$

where θ is an unknown parameter, $\phi_F(X_1)$ has mean 0 and variance $\sigma_F^2 > 0$, and R_n is the remainder term. Since $n^{-1} \sum_{i=1}^n \phi_F(X_i)$ in (2.48) is $O_p(n^{-1/2})$, $R_n = o_p(n^{-1/2})$ is a reasonable assumption and is true for most statistics T_n satisfying (2.1). If $T_n = T(F_n)$ and $\theta = T(F)$ with a differentiable T (see Section 2.2.1), then $\phi_F(x)$ is simply the influence function and usually $R_n = o_p(n^{-1/2})$. Having expansion (2.48) is, however, weaker than having a differential for general statistics T_n . For example, when T_n is the U-statistic in (2.9), T_n cannot be written as $T(F_n)$ for some T ; but T_n admits the expansion (2.48) (see, e.g., Serfling, 1980, Chapter 5).

The order of magnitude of the remainder R_n can be used to measure the smoothness of T_n , i.e., how well T_n is approximated by the linear function $n^{-1} \sum_{i=1}^n \phi_F(X_i)$. However, it is not convenient to use R_n directly since it is random. Instead, we use

$$\tau_n = ER_n^2 \quad (2.49)$$

as a smoothness measure of T_n . The smaller τ_n is, the smoother T_n is.

For an $s \in \mathcal{S}$, let $R_{r,s}$ be the remainder in the expansion (2.48) for $T_{r,s}$:

$$T_{r,s} = \theta + \frac{1}{r} \sum_{i \in s^c} \phi_F(X_i) + R_{r,s}. \quad (2.50)$$

Define $Z_i = \phi_F(X_i)$, $\bar{Z}_n = n^{-1} \sum_{i=1}^n Z_i$, $\bar{Z}_{r,s} = r^{-1} \sum_{i \in s^c} Z_i$, and $W_{r,s} = R_{r,s} - N^{-1} \sum_{s \in \mathcal{S}} R_{r,s}$. From (2.46) and $\sum_{s \in \mathcal{S}} W_{n,s} = 0$, it follows that

$$\begin{aligned} v_{\text{JACK-}d} &= \frac{r}{dN} \sum_{s \in \mathcal{S}} (\bar{Z}_{r,s} - \bar{Z}_n + W_{n,s})^2 \\ &= \frac{s_n^2}{n} + \frac{r}{dN} \sum_{s \in \mathcal{S}} W_{r,s}^2 + \frac{2r}{dN} \sum_{s \in \mathcal{S}} \bar{Z}_{r,s} W_{n,s}, \end{aligned}$$

where $s_n^2 = (n-1)^{-1} \sum_{i=1}^n (Z_i - \bar{Z}_n)^2$. Since $EZ_i = 0$, $EZ_i^2 = \sigma_F^2$, $s_n^2 \rightarrow_{a.s.}$ σ_F^2 , and σ_F^2/n is the asymptotic variance of T_n , $v_{\text{JACK-}d}$ is weakly consistent if

$$\frac{nr}{dN} \sum_{\mathbf{s} \in S} W_{n,\mathbf{s}}^2 \rightarrow_p 0 \quad (2.51)$$

[$v_{\text{JACK-}d}$ is strongly consistent if \rightarrow_p in (2.51) is replaced by $\rightarrow_{a.s.}$]. Note that the above discussion is valid for any d , including $d \equiv 1$. Apparently, a larger d provides higher chance for (2.51) to hold than a smaller d , since d is in the denominator.

Since $\{X_i, i = 1, \dots, n\}$ are exchangeable, $EW_{n,\mathbf{s}}^2$ does not depend on the choice of \mathbf{s} . Let $w_n = EW_{n,\mathbf{s}}^2$. Then (2.51) holds if we choose d so that

$$nrw_n/d \rightarrow 0. \quad (2.52)$$

It is, however, not easy to choose d according to (2.52), since the order of $w_n = EW_{n,\mathbf{s}}^2$ is not easy to obtain. A sufficient condition of (2.52), from which the choice of d can be made directly, is

$$nr\tau_r/d \rightarrow 0, \quad (2.53)$$

where τ_r is given by (2.49). Hence, the smaller τ_r is (the smoother T_n is), the smaller d is required to be.

We have actually proved the following result.

Theorem 2.10. *Assume that τ_n in (2.49) is well defined and that d is chosen according to either (2.52) or (2.53). Then, the delete- d jackknife estimator $v_{\text{JACK-}d}$ is weakly consistent, i.e.,*

$$nv_{\text{JACK-}d}/\sigma_F^2 \rightarrow_p 1. \quad (2.54)$$

An important special case is when d is chosen according to

$$d/n \geq \epsilon_0 \quad \text{for some } \epsilon_0 > 0 \quad \text{and} \quad r \rightarrow \infty. \quad (2.55)$$

In this case, we have the following result.

Theorem 2.11. *Assume that d is chosen according to (2.55). Then, the delete- d jackknife estimator $v_{\text{JACK-}d}$ is weakly consistent in the sense of (2.54) and*

$$v_{\text{JACK-}d}/\text{var}(T_n) \rightarrow_p 1, \quad (2.56)$$

provided that one of the following three conditions holds:

- (a) $n\tau_n = nER_n^2 \rightarrow 0$;
- (b) Condition (2.48) holds and condition (1.4) holds, i.e.,

$$n[\text{var}(T_n)] \rightarrow \sigma_F^2; \quad (2.57)$$

- (c) Condition (2.48) holds and $\{n(T_n - \theta)^2\}$ is uniformly integrable.

Proof. Clearly, condition (a) and (2.55) imply (2.53), which is sufficient for the weak consistency of $v_{\text{JACK-}d}$ (Theorem 2.10). The conditions (a), (b), and (c) are equivalent (see Lemma 1 in Shao and Wu, 1989). Finally, (2.56) follows from (2.54) and (2.57). \square

Example 2.12. Sample quantiles. Suppose that $p = 1$. Let $T(G) = G^{-1}(t)$ be the t -quantile of G , $0 < t < 1$. $T(F_n)$ is then the sample t -quantile used to estimate $\theta = T(F)$. Suppose that F has a positive first order derivative at θ : $f(\theta) > 0$. Then, by Ghosh (1971), (2.48) holds with $\phi_F(x) = (t - I\{x \leq \theta\})/f(\theta)$. Hence, (2.1) holds for $T(F_n)$ with $\sigma_n^2 = \sigma_F^2/n$ and $\sigma_F^2 = t(1-t)/[f(\theta)]^2$.

The delete-1 jackknife estimator v_{JACK} in (1.13) is known to be inconsistent in this case. It can be shown that $v_{\text{JACK-}d}$ is also inconsistent if d does not diverge to infinity as $n \rightarrow \infty$. For d satisfying (2.55), by Theorem 2.11, $v_{\text{JACK-}d}$ is consistent if (2.57) holds, which is satisfied if $E|X_1|^\epsilon < \infty$ for some $\epsilon > 0$ and df/dx exists in a neighborhood of θ . In fact, under these extra conditions, the consistency of $v_{\text{JACK-}d}$ can be achieved with a smaller d , since Duttweiler (1973) showed that for the τ_n in (2.49),

$$\tau_n = \left[\frac{2t(1-t)}{\pi f(\theta)} \right]^2 n^{-3/2} + o(n^{-7/4+\delta})$$

for some $\delta > 0$. Hence, if d satisfies $\sqrt{n}/d \rightarrow 0$ and $r \rightarrow \infty$, then (2.53) holds and $v_{\text{JACK-}d}$ is consistent by Theorem 2.10.

Example 2.13. Poverty proportion. Suppose that $p = 1$. Consider the statistic $T_n = F_n(\gamma_n)$ with $\gamma_n = \frac{1}{2}F_n^{-1}(\frac{1}{2})$, half of the sample median. This statistic is used to estimate the proportion of people whose income is lower than half of the median income of a population F , i.e., $F(\gamma)$ with $\gamma = \frac{1}{2}F^{-1}(\frac{1}{2})$. Assume that F is twice differentiable, $f = dF/dx$ is bounded, and $f(2\gamma) > 0$. Then, by Lemma 2.5.4.E in Serfling (1980),

$$F_n(\gamma_n) - F_n(\gamma) - F(\gamma_n) + F(\gamma) = o_p(n^{-1/2}).$$

From Example 2.12 and $\gamma = \theta/2$,

$$\begin{aligned} F_n(\gamma_n) - F(\gamma) &= F_n(\gamma) - F(\gamma) + F(\gamma_n) - F(\gamma) + o_p(n^{-1/2}) \\ &= \frac{1}{n} \sum_{i=1}^n [I\{X_i \leq \gamma\} - F(\gamma)] + f(\gamma)(\gamma_n - \gamma) + o_p(n^{-1/2}) \\ &= \frac{1}{n} \sum_{i=1}^n \phi_F(X_i) + o_p(n^{-1/2}), \end{aligned} \tag{2.58}$$

where $\phi_F(x) = I\{x \leq \gamma\} - F(\gamma) + f(\gamma)(1/2 - I\{x \leq 2\gamma\})/[2f(2\gamma)]$. This shows that (2.48) holds for $T_n = F_n(\gamma_n)$. Since the delete-1 jackknife

variance estimator is inconsistent for the sample median, we consider the delete-d jackknife variance estimator $v_{\text{JACK}-d}$ for estimating the asymptotic variance of T_n .

Assume that d is a fraction of n and $E|X_1|^\epsilon < \infty$ for some $\epsilon > 0$. To apply Theorem 2.11, we verify condition (c). The first part of condition (c) in Theorem 2.11 has already been verified [see (2.58)]. From Duttweiler (1973), $\{n(\gamma_n - \gamma)^2\}$ is uniformly integrable (Appendix A.7). Since f is bounded,

$$n[F(\gamma_n) - F(\gamma)]^2 \leq \|f\|_\infty^2 n(\gamma_n - \gamma)^2$$

is uniformly integrable. From

$$\begin{aligned} n[T_n - F(\gamma)]^2 &\leq 2n[T_n - F(\gamma_n)]^2 + 2n[F(\gamma_n) - F(\gamma)]^2 \\ &\leq 2n\|F_n - F\|_\infty^2 + 2n[F(\gamma_n) - F(\gamma)]^2 \end{aligned}$$

and the fact that $\{n\|F_n - F\|_\infty^2\}$ is uniformly integrable, we conclude that $\{n[T_n - F(\gamma)]^2\}$ is uniformly integrable. This verifies the second part of condition (c). Hence, $v_{\text{JACK}-d}$ is consistent in this case.

Note that (2.48) is very weak. Hence, for the consistency of $v_{\text{JACK}-d}$ with d being a fraction of n , Theorem 2.11 essentially requires condition (2.57). This condition is not trivial, and it does not hold if the variance of T_n does not exist. However, this condition can be removed or replaced by other conditions. Two examples are: (1) Shao (1988b) showed the consistency of $v_{\text{JACK}-d}$ for sample quantiles, assuming only (2.48); (2) Shao (1991b) showed that condition (2.57) can be replaced by $T_n = T(F_n)$ and T is ρ_∞ -Fréchet differentiable at F . Note that the consistency of the delete-1 jackknife estimator v_{JACK} requires that T be *continuously* differentiable in some sense.

For the R-estimators defined in Example 2.10, it is unknown whether or not the delete-1 jackknife estimator v_{JACK} is consistent, although the simulation results in Schucany and Sheather (1989) showed that v_{JACK} has a good performance. Condition (2.48) is satisfied for R-estimators, since T is ρ_∞ -Hadamard differentiable at F . Condition (2.57) has to be verified for given R-estimators in order to apply the result in Theorem 2.11.

It is interesting to compare the performance of v_{JACK} and $v_{\text{JACK}-d}$ for fixed n and either smooth or nonsmooth statistics. The following example is an empirical simulation comparison.

Example 2.14. Simulation comparison of v_{JACK} and $v_{\text{JACK}-d}$. We consider the case where $n = 40$ and F is one of the following three distributions: $N(\frac{5}{4}, 1)$, the exponential distribution with mean $\frac{1}{2}$, and the Cauchy distribution with median $\frac{5}{2}$ and scale parameter 2. Three jackknife estimators, v_{JACK} , $v_{\text{JACK}-d}$ with $d = 10$ and $d = 20$, of the asymptotic variance σ_F^2 of

Table 2.2. Simulation results in Example 2.14

$T_n = \bar{X}_n^2$				
d	$F = \text{normal}, \sigma_F^2 = 6.25$		$F = \text{exponential}, \sigma_F^2 = 6.256$	
	RB	RM	RB	RM
1	1.2%	2.125	7.2%	4.992
10	1.2%	2.198	9.8%	5.152
20	1.6%	2.250	13.0%	5.520
$T_n = F_n^{-1}(\frac{1}{2})$				
d	$F = \text{normal}, \sigma_F^2 = 6.28$		$F = \text{Cauchy}, \sigma_F^2 = 4.93$	
	RB	RM	RB	RM
1	92.2%	25.56	106.8%	24.02
10	21.8%	6.140	42.1%	6.327
20	8.9%	3.911	29.0%	4.030

$\sqrt{n}T_n$ are considered, where T_n is either \bar{X}_n^2 (a smooth estimator) or $F_n^{-1}(\frac{1}{2})$ (a nonsmooth estimator). To assess the performances of the variance estimators, we compute their RB (relative biases, i.e., biases over the true value of σ_F^2) and RM (root mean squared errors) based on 2000 simulation replications. Monte Carlo approximations are used for the computation of $v_{\text{JACK-}d}$ with $d > 1$ (see Section 5.2). The results are given in Table 2.2.

All three jackknife variance estimators perform well when $T_n = \bar{X}_n^2$. Note that for this T_n , the performance of v_{JACK} is slightly better than those of the other two. On the other hand, when T_n is the sample median, v_{JACK} performs very poorly: its relative bias is over 90%, which shows that it is inconsistent. The performance of $v_{\text{JACK-}d}$ with $d = 10$ and $d = 20$ are much better than v_{JACK} here. Also, when the underlying population is Cauchy, none of these estimators performs well, which illustrates how the convergence speed of a consistent estimator depends on the underlying population.

2.3.2 Jackknife histograms

Using the delete-d jackknife, we can obtain a consistent estimator of the sampling distribution of T_n . Again, we start with the simplest case where $T_n = \bar{X}_n$ and $p = 1$. Let $\mu = EX_1$,

$$H_n(x) = H_{n,F}(x) = P\{\sqrt{n}(\bar{X}_n - \mu) \leq x\}, \quad (2.59)$$

and

$$G_n(x) = G_{n,F}(x) = P\{(\bar{X}_n - \mu)/S_n \leq x\}, \quad (2.60)$$

where $S_n^2 = [n(n-1)]^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. Assume $\text{var}(X_1) = \sigma^2 < \infty$. Then, by the central limit theorem (Appendix A.8),

$$\lim_{n \rightarrow \infty} H_n(x) = \Phi(x/\sigma) \quad \text{and} \quad \lim_{n \rightarrow \infty} G_n(x) = \Phi(x). \quad (2.61)$$

The traditional estimators of $H_n(x)$ and $G_n(x)$ are $\Phi(x/\hat{\sigma})$ and $\Phi(x)$, respectively, where $\hat{\sigma} = \sqrt{n}S_n$.

We follow the notation in Section 2.3.1. For $\mathbf{s}_* \in \mathcal{S}$, \bar{X}_{r,\mathbf{s}_*} is the “sample mean” of $\{X_i, i \in \mathbf{s}_*^c\}$, a simple random sample without replacement from $\{X_1, \dots, X_n\}$; $P_*\{\cdot | X_1, \dots, X_n\} = P_*\{\cdot\}$ is the conditional probability corresponding to the selection of $\{X_i, i \in \mathbf{s}_*^c\}$; and E_* and var_* are the mean and variance, respectively, taken under P_* . Note that $E_*(\bar{X}_{r,\mathbf{s}_*}) = \bar{X}_n$ and

$$\text{var}_*(\bar{X}_{r,\mathbf{s}_*}) = \frac{1-f_n}{r} \hat{\sigma}^2 = \frac{d}{nr} \hat{\sigma}^2,$$

where $f_n = r/n$ is the “finite sample correction”. Hence, an analog of the standardized variable $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ is $\sqrt{nr/d}(\bar{X}_{r,\mathbf{s}_*} - \bar{X}_n)/\hat{\sigma}$. This leads to a jackknife estimator of H_n in (2.59):

$$H_{\text{JACK}}(x) = P_* \left\{ \sqrt{\frac{nr}{d}} (\bar{X}_{r,\mathbf{s}_*} - \bar{X}_n) \leq x \right\}.$$

Since the selection of a simple random sample without replacement from $\{X_1, \dots, X_n\}$ is equivalent to the selection of a subset \mathbf{s}_* from \mathcal{S} with equal probability N^{-1} , we have

$$H_{\text{JACK}}(x) = \frac{1}{N} \sum_{\mathbf{s} \in \mathcal{S}} I \left\{ \sqrt{\frac{nr}{d}} (\bar{X}_{r,\mathbf{s}} - \bar{X}_n) \leq x \right\}. \quad (2.62)$$

Note that H_{JACK} corresponds to the histogram constructed by the N points

$$\sqrt{\frac{nr}{d}} (\bar{X}_{r,\mathbf{s}} - \bar{X}_n), \quad \mathbf{s} \in \mathcal{S}.$$

For this reason, H_{JACK} can be called a *cumulative jackknife histogram*.

Similarly, a jackknife estimator of G_n in (2.60) can be obtained. An analog of the studentized variable $(\bar{X}_n - \mu)/S_n$ is $(\bar{X}_{r,\mathbf{s}_*} - \bar{X}_n)/S_{r,\mathbf{s}_*}$, where

$$S_{r,\mathbf{s}_*}^2 = \frac{1-f_n}{r(r-1)} \sum_{i \in \mathbf{s}_*^c} (X_i - \bar{X}_{r,\mathbf{s}_*})^2.$$

Then, a jackknife estimator of G_n is

$$\begin{aligned} G_{\text{JACK}}(x) &= P_* \{ (\bar{X}_{r,\mathbf{s}_*} - \bar{X}_n)/S_{r,\mathbf{s}_*} \leq x \} \\ &= \frac{1}{N} \sum_{\mathbf{s} \in \mathcal{S}} I \{ (\bar{X}_{r,\mathbf{s}} - \bar{X}_n)/S_{r,\mathbf{s}} \leq x \}. \end{aligned} \quad (2.63)$$

Since N is usually very large, Monte Carlo approximations may be used to numerically compute H_{JACK} and G_{JACK} . See more discussions in Chapter 5.

Wu (1990) established some fundamental results for the asymptotic behavior of jackknife histograms. In view of (2.61), the first part of the following theorem indicates the *inconsistency* of H_{JACK} when *either r or d* is bounded.

Theorem 2.12. *Assume that $\text{var}(X_1) < \infty$.*

- (i) *If either d or r is bounded, then for any x , $H_{\text{JACK}}(x) \rightarrow_{a.s.} K_F(x)$ and $K_F(x) \neq \Phi(x/\sigma)$ except for normal F .*
- (ii) *When both $r \rightarrow \infty$ and $d \rightarrow \infty$,*

$$\|H_{\text{JACK}} - H_n\|_\infty \rightarrow_{a.s.} 0.$$

Proof. (i) It suffices to show the result for the case where d is fixed. The relationship between \mathbf{s} and \mathbf{s}^c yields that $r(\bar{X}_{r,\mathbf{s}} - \bar{X}_n) = -d(\bar{X}_{d,\mathbf{s}^c} - \bar{X}_n)$, and, by (2.62),

$$H_{\text{JACK}}(x) = \frac{1}{N} \sum_{\mathbf{s} \in \mathcal{S}} I\left\{-\sqrt{\frac{nd}{r}}(\bar{X}_{d,\mathbf{s}^c} - \bar{X}_n) \leq x\right\}.$$

Since $n/r \rightarrow 1$ when d is fixed and $\bar{X}_n \rightarrow_{a.s.} \mu$,

$$H_{\text{JACK}}(x) - K_n(x) \rightarrow_{a.s.} 0,$$

where

$$K_n(x) = \frac{1}{N} \sum_{\mathbf{s} \in \mathcal{S}} I\{-\sqrt{d}(\bar{X}_{d,\mathbf{s}^c} - \mu) \leq x\}$$

is a “U-statistic”. From the theory of U-statistics (e.g., Serfling, 1980, Chapter 5), $K_n(x) \rightarrow_{a.s.} K_F(x) = P\{-\sqrt{d}(\bar{X}_d - \mu) \leq x\}$. Since d is fixed, $K_F(x) \neq \Phi(x/\sigma)$ except for normal F . This proves (i).

(ii) From (2.61), Pólya’s theorem (Appendix A.1), and the continuity of Φ , it suffices to show that, for any fixed x ,

$$H_{\text{JACK}}(x) - \Phi(x/\sigma) \rightarrow_{a.s.} 0. \quad (2.64)$$

Define

$$\tilde{H}_{\text{JACK}}(x) = H_{\text{JACK}}(\hat{\sigma}x) = P_* \left\{ \frac{\sqrt{r}(\bar{X}_{r,\mathbf{s}_*} - \bar{X}_n)}{\sqrt{1 - f_n}\hat{\sigma}} \leq x \right\}. \quad (2.65)$$

Then Hájek’s (1960) result on the sufficient condition for the asymptotic normality of \bar{X}_{r,\mathbf{s}_*} based on a simple random sample $\{X_i, i \in \mathbf{s}_*^c\}$ (without

replacement) from the “population” $\{X_1, \dots, X_n\}$ can be applied. That is, $\tilde{H}_{\text{JACK}}(x) - \Phi(x) \rightarrow_{a.s.} 0$ holds if

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 I\{|X_i - \bar{X}_n| \geq \epsilon \sqrt{rd/n} \hat{\sigma}\} = 0 \quad a.s. \quad (2.66)$$

for any $\epsilon > 0$. Since $\text{var}(X_1) < \infty$, $\bar{X}_n \rightarrow_{a.s.} \mu$ and $\hat{\sigma} \rightarrow_{a.s.} \sigma$, (2.66) follows from $\text{var}(X_1) < \infty$ and $rd/n \rightarrow \infty$ when $r \rightarrow \infty$ and $d \rightarrow \infty$. This proves $\|\tilde{H}_{\text{JACK}} - \Phi\|_\infty \rightarrow_{a.s.} 0$ and therefore

$$\tilde{H}_{\text{JACK}}(x/\hat{\sigma}) - \Phi(x/\hat{\sigma}) \rightarrow_{a.s.} 0.$$

From (2.65), $\tilde{H}_{\text{JACK}}(x/\hat{\sigma}) = H_{\text{JACK}}(x)$. Since Φ is continuous, $\Phi(x/\hat{\sigma}) - \Phi(x/\sigma) \rightarrow_{a.s.} 0$. Hence (2.64) holds. \square

Theorem 2.13. *Assume that $\text{var}(X_1) < \infty$.*

- (i) *If either d or r is bounded, then for any x , $G_{\text{JACK}}(x) \rightarrow_{a.s.} Q_F(x)$ and $Q_F(x) \neq \Phi(x)$ except for normal F .*
- (ii) *When both $r \rightarrow \infty$ and $d \rightarrow \infty$, $\|G_{\text{JACK}} - G_n\|_\infty \rightarrow_{a.s.} 0$.*
- (iii) *If $E|X_1|^3 < \infty$, then with $r \rightarrow \infty$ and $d \rightarrow \infty$,*

$$\|G_{\text{JACK}} - G_n\|_\infty = O\left(\sqrt{\frac{n}{rd}}\right) \quad a.s. \quad (2.67)$$

Proof. The proofs of (i) and (ii) are similar to those of Theorem 2.12 and are omitted. For (iii), an application of the result in Robinson (1978) gives

$$\|G_{\text{JACK}} - \Phi\|_\infty \leq C \sqrt{\frac{n}{rd}} \frac{1}{n\hat{\sigma}^3} \sum_{i=1}^n |X_i - \bar{X}_n|^3, \quad (2.68)$$

which is $O(\sqrt{n}/rd)$ a.s. since $E|X_1|^3 < \infty$. Result (2.67) follows from (2.68) and $\|G_n - \Phi\|_\infty = O(n^{-1/2})$ by the Berry-Esséen theorem (Appendix A.9). \square

Result (2.67) provides an upper bound for the rate of convergence of G_{JACK} . The best rate is $O(n^{-1/2})$, which can be reached by choosing d such that $d/n \rightarrow \lambda \in (0, 1)$. Wu (1990) also proved that, under $E|X_1|^3 < \infty$,

$$\|\tilde{H}_{\text{JACK}} - \Phi\|_\infty = O\left(\sqrt{\frac{n}{rd}}\right) \quad a.s., \quad (2.69)$$

where \tilde{H}_{JACK} is given in (2.65). An upper bound for the convergence rate of H_{JACK} can be obtained as follows. Assuming $EX_1^4 < \infty$, we have that $\hat{\sigma} - \sigma = O_p(n^{-1/2})$ and there is a constant $c > 0$ such that

$$\|\Phi_{\hat{\sigma}} - \Phi_\sigma\|_\infty \leq c|\hat{\sigma}^{-1} - \sigma^{-1}| = O_p(n^{-1/2}),$$

where $\Phi_{\hat{\sigma}}(x) = \Phi(x/\hat{\sigma})$ and $\Phi_{\sigma}(x) = \Phi(x/\sigma)$. Then, by (2.65) and (2.69),

$$\begin{aligned}\|H_{\text{JACK}} - \Phi_{\sigma}\|_{\infty} &\leq \|H_{\text{JACK}} - \Phi_{\hat{\sigma}}\|_{\infty} + \|\Phi_{\hat{\sigma}} - \Phi_{\sigma}\|_{\infty} \\ &= \|\tilde{H}_{\text{JACK}} - \Phi\|_{\infty} + O_p(n^{-1/2}) = O_p\left(\sqrt{\frac{n}{rd}}\right).\end{aligned}$$

Since $\|H_n - \Phi_{\sigma}\|_{\infty} = O(n^{-1/2})$, we have

$$\|H_{\text{JACK}} - H_n\|_{\infty} = O\left(\sqrt{\frac{n}{rd}}\right).$$

We now consider the general statistic T_n . Define

$$H_n(x) = H_{n,F}(x) = P\{\sqrt{n}(T_n - \theta) \leq x\} \quad (2.70)$$

and

$$\begin{aligned}H_{\text{JACK}}(x) &= P_*\left\{\sqrt{\frac{nr}{d}}(T_{r,\mathbf{s}_*} - T_n) \leq x\right\} \\ &= \frac{1}{N} \sum_{\mathbf{s} \in \mathcal{S}} I\left\{\sqrt{\frac{nr}{d}}(T_{r,\mathbf{s}} - T_n) \leq x\right\},\end{aligned} \quad (2.71)$$

where $T_{r,\mathbf{s}}$ is given by (2.43). Equations (2.70) and (2.71) are straightforward extensions of (2.59) and (2.62), respectively. If a variance estimator of T_n is available, we can also define a jackknife estimator of the form (2.63) to estimate the sampling distribution (2.60) of the studentized variable. Note that the jackknife variance estimator $v_{\text{JACK-}d}$ in (2.47) is exactly the variance of the jackknife histogram, i.e., $v_{\text{JACK-}d} = \text{var}_*[\sqrt{r/d}(T_{r,\mathbf{s}_*} - T_n)]$.

Wu (1990) proved the following result that covers a very broad class of statistics including many smooth and not-so-smooth statistics.

Theorem 2.14. *Assume that T_n admits the expansion in (2.48). If d is chosen according to (2.55), then*

$$\|H_{\text{JACK}} - H_n\|_{\infty} \rightarrow_p 0, \quad (2.72)$$

where H_n and H_{JACK} are given in (2.70) and (2.71), respectively.

Proof. From (2.48), $\lim_{n \rightarrow \infty} H_n(x) = \Phi(x/\sigma_F)$ with $\sigma_F^2 = E[\phi_F(X_1)]^2$. Without loss of generality, we assume $\sigma_F = 1$. Then it remains to be shown that $\|H_{\text{JACK}} - \Phi\|_{\infty} \rightarrow_p 0$. Let

$$\xi_{n,\mathbf{s}} = \sqrt{\frac{nr}{d}} \left(\frac{1}{r} \sum_{i \in \mathbf{s}^c} Z_i - \frac{1}{n} \sum_{i=1}^n Z_i \right) \quad \text{and} \quad \zeta_{n,\mathbf{s}} = \sqrt{\frac{nr}{d}} (R_{r,\mathbf{s}} - R_n),$$

where $Z_i = \phi_F(X_i)$ and $R_{r,s}$ is given in (2.50). Then

$$\sqrt{\frac{nr}{d}} (T_{r,s} - T_n) = \xi_{n,s} + \zeta_{n,s}.$$

From Theorem 2.12, $\sup_x |P_*\{\xi_{n,s_*} \leq x\} - \Phi(x)| \rightarrow_{a.s.} 0$. Since for any $\epsilon > 0$, $\|H_{\text{JACK}} - \Phi\|_\infty = \sup_x |P_*\{\xi_{n,s_*} + \zeta_{n,s_*} \leq x\} - \Phi(x)|$ is bounded by

$$(2\pi)^{-1/2}\epsilon + 2P_*\{|\zeta_{n,s_*}| \geq \epsilon\} + \sup_x |P_*\{\xi_{n,s_*} \leq x\} - \Phi(x)|,$$

result (2.72) follows if

$$P_*\{|\zeta_{n,s_*}| \geq \epsilon\} \rightarrow_p 0. \quad (2.73)$$

Since X_1, \dots, X_n are i.i.d.,

$$E(P_*\{|\zeta_{n,s_*}| \geq \epsilon\}) = E\left(\frac{1}{N} \sum_{s \in S} I\{\zeta_{n,s} \geq \epsilon\}\right) = P\left\{\sqrt{\frac{nr}{d}} |R_r - R_n| \geq \epsilon\right\}.$$

This equality, together with $\epsilon_0 n \leq d$, $R_r = o_p(r^{-1/2})$ and $R_n = o_p(n^{-1/2})$, yields (2.73). \square

Some more results about jackknife histograms can be found in Shi (1991), Booth and Hall (1993a), and Politis and Romano (1995).

It is interesting to compare the jackknife estimator H_{JACK} in (2.71) with the bootstrap estimator H_{BOOT} in (1.20) with $\mathfrak{R}_n(X_1^*, \dots, X_n^*, \hat{F}) = \sqrt{n}(T_n^* - T_n)$. If $\hat{F} = F_n$ (the empirical distribution) and the finite sample fraction f_n is ignored, then H_{JACK} and H_{BOOT} are the same except that in the bootstrap T_n^* is based on a simple random sample of size n *with replacement* from $\{X_1, \dots, X_n\}$, whereas in the jackknife T_{r,s_*} is based on a simple random sample of size r *without replacement* from $\{X_1, \dots, X_n\}$. The bootstrap may produce more accurate results in terms of the convergence rate in approximation (see Section 3.3) and is more flexible, since \hat{F} is not necessarily F_n . The jackknife histogram, however, is a valuable tool in some situations where the bootstrap does not work (Politis and Romano, 1995).

2.4 Other Applications

In this section we study other applications of the jackknife. Similar to the problem of variance estimation, we show in Section 2.4.1 that the success of the jackknife bias estimator b_{JACK} defined in (1.8) relies on the smoothness of the given statistic T_n . In Section 2.4.2 we show that the jackknife estimator T_{JACK} in (1.9) has the same asymptotic distribution as T_n , but is less biased. In both sections we start with the simple case, functions of the sample mean, and then extend the results to general differentiable functionals.

2.4.1 Bias estimation

Let T_n be an estimator of an unknown parameter θ . We first consider the case where $T_n = g(\bar{X}_n)$, used to estimate $\theta = g(\mu)$. Here, $\mu = EX_1$ and g is a function from \mathbb{R}^p to \mathbb{R} . Suppose that g is second order differentiable and the Hessian $\nabla^2 g(x) = \partial g(x)/\partial x \partial x'$ is continuous at μ . Then

$$\begin{aligned} T_n - \theta &= g(\bar{X}_n) - g(\mu) = \nabla g(\mu)'(\bar{X}_n - \mu) \\ &\quad + \frac{1}{2}(\bar{X}_n - \mu)' \nabla^2 g(\mu)(\bar{X}_n - \mu) + R_n, \end{aligned} \quad (2.74)$$

where $R_n = o_p(n^{-1})$. Let $\nabla^3 g$ denote all of the third order partial derivatives of g . If $\nabla^3 g$ exists and is bounded in a neighborhood of μ and $E\|X_1\|^3 < \infty$, then $R_n = O_p(n^{-2})$. Note that

$$E[\nabla g(\mu)'(\bar{X}_n - \mu)] = 0,$$

$$E\left[\frac{1}{2}(\bar{X}_n - \mu)' \nabla^2 g(\mu)(\bar{X}_n - \mu)\right] = \frac{a}{n},$$

where

$$a = \frac{1}{2}\text{tr}[\nabla^2 g(\mu)\text{var}(X_1)], \quad (2.75)$$

$\text{tr}(A)$ is the trace of the matrix A , and R_n is of a lower stochastic order than $\frac{1}{2}(\bar{X}_n - \mu)' \nabla^2 g(\mu)(\bar{X}_n - \mu)$. For sufficiently smooth g ,

$$ER_n = o(n^{-1}) \quad \text{or} \quad ER_n = O(n^{-2}) \quad (2.76)$$

(see Lehmann, 1983, and Shao, 1988c). If (2.76) holds, then $\text{bias}(T_n) = a/n + o(n^{-1})$ or $\text{bias}(T_n) = a/n + O(n^{-2})$. Even if (2.76) does not hold, we can still define a/n as the *asymptotic bias* of T_n . Note that the asymptotic bias exists even if ET_n does not exist. From the mean-value theorem,

$$\begin{aligned} T_{n-1,i} - T_n &= g(\bar{X}_{n-1,i}) - g(\bar{X}_n) = \nabla g(\bar{X}_n)'(\bar{X}_{n-1,i} - \bar{X}_n) \\ &\quad + \frac{1}{2}(\bar{X}_{n-1,i} - \bar{X}_n)' \nabla^2 g(\xi_{n,i})(\bar{X}_{n-1,i} - \bar{X}_n), \end{aligned}$$

where $\xi_{n,i}$ is a point on the line segment between $\bar{X}_{n-1,i}$ and \bar{X}_n . From (2.4), $\sum_{i=1}^n (\bar{X}_{n-1,i} - \bar{X}_n) = 0$ and

$$\begin{aligned} b_{\text{JACK}} &= \frac{1}{2n(n-1)} \sum_{i=1}^n (X_i - \bar{X}_n)' \nabla^2 g(\xi_{n,i})(X_i - \bar{X}_n) \\ &= \frac{1}{2}\text{tr}\left[\frac{1}{n(n-1)} \sum_{i=1}^n \nabla^2 g(\xi_{n,i})(X_i - \bar{X}_n)(X_i - \bar{X}_n)'\right]. \end{aligned} \quad (2.77)$$

The following result shows that b_{JACK} is consistent.

Theorem 2.15. Assume that $\text{var}(X_1) < \infty$ and $\nabla^2 g$ is continuous at μ . Then,

$$n b_{\text{JACK}} \rightarrow_{a.s.} a, \quad (2.78)$$

where a is defined in (2.75). If, in addition, $\nabla^3 g$ is bounded in a neighborhood of μ and $E\|X_1\|^3 < \infty$, then

$$b_{\text{JACK}} = a/n + O_p(n^{-2}). \quad (2.79)$$

Proof. From (2.75) and (2.77), result (2.78) follows from the continuity of $\nabla^2 g$ and

$$\|\xi_{n,i} - \mu\| \leq \|\xi_{n,i} - \bar{X}_n\| + \|\bar{X}_n - \mu\| \leq \|\bar{X}_{n-1,i} - \bar{X}_n\| + \|\bar{X}_n - \mu\|.$$

When $\nabla^3 g$ is bounded in a neighborhood of μ , there is a constant $c > 0$ such that the absolute values of the elements of $\nabla^2 g(\xi_{n,i}) - \nabla^2 g(\mu)$ are bounded by $c\|\xi_{n,i} - \mu\| \leq c(\|\bar{X}_{n-1,i} - \bar{X}_n\| + \|\bar{X}_n - \mu\|)$ for large n . Hence, result (2.79) follows. \square

If (2.76) holds, then a/n in (2.78) or (2.79) may be replaced by $\text{bias}(T_n)$.

We now consider the general statistical functional $T_n = T(F_n)$.² Note that the bias of $T_n = g(\bar{X}_n)$ is directly related to the second order derivative $\nabla^2 g$. Hence, in the case of $T_n = T(F_n)$, we need to assume that T is second order differentiable (see Section 2.2.4) in order to study the bias of T_n . If T is second order ρ -Fréchet differentiable (see Definition 2.7), where $\rho = \rho_\infty$ or $\rho = \rho_{\infty+r}$ [assuming (2.17)], then $T_n - \theta$ is equal to

$$T(F_n) - T(F) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \psi_F(X_i, X_j) + o_p(n^{-1}),$$

where $\psi_F(x, y) = \psi_F(y, x)$ and $\int \int \psi_F(x, y) dF(x) dF(y) = 0$. Assume that

$$a = E\psi_F(X_1, X_1) = \int \psi_F(x, x) dF(x) \quad (2.80)$$

is well defined. Then the asymptotic bias of T_n is

$$E\left[\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \psi_F(X_i, X_j)\right] = \frac{a}{n},$$

since $E[\psi(X_i, X_j)] = 0$ for $i \neq j$.

To study the consistency of b_{JACK} , we need to assume that T has a stronger second order differential (similar to Definition 2.8). Note that for

²The reader may omit the rest of this section if Section 2.2 was omitted.

the simple case $T_n = g(\bar{X}_n)$ we need to assume the continuity of the second order derivative $\nabla^2 g$.

Definition 2.9. A functional T on \mathcal{F} is uniformly second order ρ -Fréchet differentiable at $G \in \mathcal{F}$ if T is second order Gâteaux differentiable in a neighborhood of G , and $\rho(G_k, G) \rightarrow 0$ and $\rho(H_k, G) \rightarrow 0$ imply

$$\lim_{k \rightarrow \infty} \frac{T(H_k) - T(G_k) - Q_{G_k}(H_k - G_k)}{[\rho(H_k, G_k)]^2} = 0,$$

where $Q_G(D)$ is given by (2.30).

Theorem 2.16. Suppose that T is uniformly second order ρ_∞ -Fréchet differentiable at F with $\psi_F(x, y)$ satisfying (2.32) and that the conditions in Theorem 2.9 hold with $\phi_G(x)$ replaced by $\psi_G(x, x)$. Then (2.78) holds with a given by (2.80). The same result holds if ρ_∞ is replaced by $\rho_{\infty+r}$ (assuming $E\|X_1\|^{2/r} < \infty$, $r = 1$ or 2).

Proof. Let $V_n = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \psi_{F_n}(X_i, X_j)$ and

$$r_n = b_{\text{JACK}} - \frac{n-1}{n} \sum_{i=1}^n (V_{n-1,i} - V_n). \quad (2.81)$$

From the uniform second order ρ -Fréchet differentiability of T , it follows that, for any $\epsilon > 0$, there is a $\delta_\epsilon > 0$ such that for $\rho(F_n, F) < \delta_\epsilon$ and $\max_{i \leq n} \rho(F_{n-1,i}, F) < \delta_\epsilon$,

$$n|r_n| \leq \epsilon n \sum_{i=1}^n [\rho(F_{n-1,i}, F_n)]^2, \quad (2.82)$$

where $\rho = \rho_\infty$ or $\rho_{\infty+r}$. Since ρ_∞ and $\rho_{\infty+r}$ satisfy (2.19), (2.82) implies $nr_n \rightarrow_{a.s.} 0$. Note that V_n is a “V-statistic” of order 2 and the second term on the right-hand side of (2.81) is the jackknife bias estimator for V_n . Hence, by Example 1.3 and (2.81),

$$b_{\text{JACK}} = n^{-1}(W_n - U_n) + r_n,$$

where $W_n = n^{-1} \sum_{i=1}^n \psi_{F_n}(X_i, X_i)$ and $U_n = (nV_n - W_n)/(n-1)$. From the properties of ψ_{F_n} , we know that $V_n = 0$. Therefore,

$$b_{\text{JACK}} = \frac{1}{n} \left(W_n + \frac{1}{n-1} W_n \right) + r_n = \frac{W_n}{n-1} + o(n^{-1}) \quad a.s.$$

Using the same argument as that in the proof of Theorem 2.9, we can show that $W_n \rightarrow_{a.s.} a$. Hence, the result follows. \square

Example 2.6 (continued). Consider the L-functional defined in (2.32). If J' exists and is continuous on $[0, 1]$, then T is uniformly second order $\rho_{\infty+2}$ -Fréchet differentiable at F with

$$\psi_G(x, y) = a_G(x) + a_G(y) + b_G(x, y),$$

where $a_G(x) = \int [G(u) - I\{u \leq x\}]J(G(u))du$ and

$$b_G(x, y) = - \int [G(u) - I\{x \leq u\}][G(u) - I\{y \leq u\}]J'(G(u))du.$$

This can be verified as follows. By Serfling (1980, p. 289),

$$\begin{aligned} R_G(H) &= T(H) - T(G) - \iint \psi_G(x, y)dG(x)dG(y). \\ &= \int W(H(x), G(x))[H(x) - G(x)]^2 dx, \end{aligned}$$

where $W(s, t) = [\int_t^s J(u)du - J(t)(s-t) - 2^{-1}J'(t)(s-t)^2]/(s-t)^2$ if $s \neq t$ and $W(s, t) = 0$ if $s = t$. Then,

$$|R_G(H)| \leq \sup_x |W(H(x), G(x))|[\rho_2(H, G)]^2.$$

From the continuity of J' , $\sup_x |W(H(x), G(x))| \rightarrow 0$ as $\rho_\infty(G, H) \rightarrow 0$. Hence, $R_G(H) = o([\rho_{\infty+2}(H, G)]^2)$ as $\rho_\infty(H, F) \rightarrow 0$ and $\rho_\infty(G, F) \rightarrow 0$.

If $E\|X_1\|^2 < \infty$, it can be shown that the conditions on $\psi_F(x, y)$ in Theorem 2.16 are satisfied. If J is trimmed and J' is continuous, then T is uniformly second order ρ_∞ -Fréchet differentiable at F and the condition $E\|X_1\|^2 < \infty$ is not necessary.

2.4.2 Bias reduction

Subtracting the estimated bias b_{JACK} from the original estimator T_n , we obtain the jackknife estimator $T_{\text{JACK}} = T_n - b_{\text{JACK}}$. However, T_{JACK} does not necessarily have a smaller bias than T_n since b_{JACK} is only an estimated bias. Naturally, bias reduction is closely related to bias estimation: a good bias estimator leads to a bias-reduced T_{JACK} .

The jackknife estimator T_{JACK} is motivated in Section 1.3 under assumption (1.10), which is somewhat too strong. With weaker assumptions, we have shown in Section 2.4.1 that the jackknife bias estimator b_{JACK} is asymptotically correct as an estimator of a/n , the asymptotic bias of T_n . Note that a/n is defined even if $E(T_n)$ does not exist. We now use the results in Section 2.4.1 to study the properties of T_{JACK} .

First, consider the case of $T_n = g(\bar{X}_n)$. From expansion (2.74),

$$T_n - \theta = L_n + Q_n + o_p(n^{-1}), \quad (2.83)$$

where $L_n = \nabla g(\mu)'(\bar{X}_n - \mu)$ is a linear functional of \bar{X}_n , $L_n = O_p(n^{-1/2})$, $E(L_n) = 0$, $Q_n = 2^{-1}(\bar{X}_n - \mu)' \nabla^2 g(\mu)(\bar{X}_n - \mu)$ is a quadratic functional of \bar{X}_n , $Q_n = O_p(n^{-1})$, $E(Q_n) = a/n$, and a is given by (2.75). From Theorem 2.15,

$$T_{\text{JACK}} - \theta = L_n + (Q_n - a/n) + o_p(n^{-1}), \quad (2.84)$$

where $Q_n - a/n$ is still of order $O_p(n^{-1})$ but has 0 mean. Comparing (2.83) with (2.84), we conclude that T_{JACK} is less biased than T_n , since the highest order term on the right-hand side of (2.84) that has possibly a nonzero mean is $o_p(n^{-1})$, whereas the highest order term on the right-hand side of (2.83) that has possibly a nonzero mean is $Q_n = O_p(n^{-1})$. In terms of the asymptotic bias, T_n has bias a/n , whereas T_{JACK} is unbiased. Hence, the jackknife does remove the first order bias.

Similar conclusions can be drawn in the general case where $T_n = T(F_n)$, $\theta = T(F)$, and T satisfies the conditions in Theorem 2.16. For a second order Fréchet differentiable T , (2.83) holds with

$$L_n = \frac{2}{n} \sum_{i=1}^n \int \psi_F(X_i, u) dF(u)$$

and

$$Q_n = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \left[\psi_F(X_i, X_j) - \int [\psi_F(X_i, u) + \psi_F(X_j, u)] dF(u) \right].$$

Note that $L_n = O_p(n^{-1/2})$, $Q_n = O_p(n^{-1})$, $E(L_n) = 0$, and $E(Q_n) = a/n$, where a is given in (2.80). Then Theorem 2.16 shows that (2.84) holds.

From (2.83) and (2.84), the asymptotic distributions of T_n and T_{JACK} are the same as that of L_n if $\sigma_n^2 = \text{var}(L_n) > 0$. That is,

$$(T_n - \theta)/\sigma_n \rightarrow_d N(0, 1) \quad \text{and} \quad (T_{\text{JACK}} - \theta)/\sigma_n \rightarrow_d N(0, 1). \quad (2.85)$$

We can also remove the second order bias by jackknifing twice. Gray and Schucany (1972) discusses this further.

A drawback of the jackknife estimator T_{JACK} is that it may have a larger mean squared error than T_n , since bias reduction may be accompanied by variance increase. Suppose that (2.85) holds. Then usually

$$\text{var}(T_n) = \sigma_n^2 + \eta_n \quad \text{and} \quad \text{var}(T_{\text{JACK}}) = \sigma_n^2 + \tilde{\eta}_n,$$

where $\eta_n = o(n^{-1})$ and $\tilde{\eta}_n = o(n^{-1})$. The first order terms in $\text{var}(T_n)$ and $\text{var}(T_{\text{JACK}})$ are the same, but the second order terms may be different.

Let $\Delta_n = \tilde{\eta}_n - \eta_n$ be the increase in variance caused by the jackknife. Then $\Delta_n = o(n^{-1})$, a lower order term than σ_n^2 . Suppose that $\text{bias}(T_n) = a/n + o(n^{-1})$ and $\text{bias}(T_{\text{JACK}}) = o(n^{-1})$. Then the mean squared errors

$$\begin{aligned} E(T_n - \theta)^2 &= \sigma_n^2 + \eta_n + [a/n + o(n^{-1})]^2 \\ &= \sigma_n^2 + \eta_n + a^2/n^2 + o(n^{-2}), \end{aligned}$$

and

$$E(T_{\text{JACK}} - \theta)^2 = \sigma_n^2 + \tilde{\eta}_n + o(n^{-2}).$$

The jackknife estimator T_{JACK} has a larger mean squared error (up to the order n^{-2}) if and only if $a^2/n^2 \leq \Delta_n$. Let us study some examples.

Example 1.1 (continued). For $T_n = \bar{X}_n^2$, $T_{\text{JACK}} = \bar{X}_n^2 - \hat{\alpha}_2/n$. For simplicity we assume that X_1 is distributed as $N(\mu, \sigma^2)$. Then,

$$\begin{aligned} \text{var}(T_n) &= \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n^2}, \quad \text{bias}(T_n) = \frac{\sigma^2}{n}, \\ \text{var}(T_{\text{JACK}}) &= \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n(n-1)}, \quad \text{bias}(T_{\text{JACK}}) = 0, \end{aligned}$$

and $\Delta_n = 2\sigma^4/[n^2(n-1)]$ is of a lower order than $a^2/n^2 = \sigma^4/n^2$. In fact, $\Delta_n < a^2/n^2$ if $n > 3$. Hence, in this example the jackknife estimator T_{JACK} has a smaller mean squared error than T_n .

Example 1.3 (continued). For a V-statistic,

$$T_n = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n h(X_i, X_j),$$

the jackknife estimator is a U-statistic,

$$T_{\text{JACK}} = \frac{2}{n(n-1)} \sum_{i < j} h(X_i, X_j).$$

It is well known that a U-statistic is unbiased (for $E[h(X_1, X_2)]$) but that it may have a larger mean squared error than the corresponding V-statistic. Hence, in this example the choice between T_n and T_{JACK} is a question of preference between a V- and a U-statistic. In Section 1.3, we showed that, if $h(x, y) = (x - y)^2/2$, then

$$T_n = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad \text{and} \quad T_{\text{JACK}} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad (2.86)$$

are estimators of $\text{var}(X_1)$. Assuming that X_1 is distributed as $N(\mu, \sigma^2)$, we obtain that

$$\begin{aligned}\text{var}(T_n) &= \frac{2\sigma^4}{n} - \frac{\sigma^4}{n^2}, \quad \text{bias}(T_n) = -\frac{\sigma^2}{n}, \\ \text{var}(T_{\text{JACK}}) &= \frac{2\sigma^4}{n} + \frac{2\sigma^4}{n(n-1)}, \quad \text{bias}(T_{\text{JACK}}) = 0,\end{aligned}$$

and $\Delta_n = (3n-1)\sigma^4/[n^2(n-1)]$. Note that Δ_n is of the same order as $a^2/n^2 = \sigma^4/n^2$, but $\Delta_n > a^2/n^2$ for all n . Hence, the jackknife estimator always has a larger mean squared error than T_n in this example.

In these examples, the difference between T_n and T_{JACK} is, in the first case, the difference between a V-statistic and a U-statistic, and in the second case, the difference between dividing the sum of squared deviations by n or $(n-1)$ (in estimating the population variance). In terms of the mean squared errors, the jackknife estimator T_{JACK} is always better than the original estimator T_n in Example 1.1 and always worse in Example 1.3. In general, the relative performance between T_{JACK} and T_n is indefinite and depends on the unknown population F .

Example 2.15. Simulation comparison of T_n and T_{JACK} . Parr and Schucany (1982) provided an empirical study of the L-statistic T_n (see Example 2.6) and T_{JACK} with several different functions J symmetric about 0.5 [$J_1(t) = 23.53(t-0.05)$, $0.05 \leq t < 0.1$; $J_1(t) = 1.1765$, $0.1 \leq t \leq 0.5$; $J_2(t) = 4t$, $t \leq 0.5$; $J_3(t) = 6t(1-t)$] and populations F [$N(5, 1)$; the logistic distribution $L(5, 1)$ with $F(x) = (1 + e^{5-x})^{-1}$; the uniform distribution $U(4, 6)$; the exponential distribution $E(1)$ with mean 1; the inverse Gaussian $IG(1, \frac{1}{2})$, and $IG(1, 2)$ (Folks and Chhikara, 1978)]. The variances of T_n and T_{JACK} based on 500 simulation replications are given in Table 2.3.

For the first three populations, T_{JACK} has a larger variance than T_n in most of the cases, although the difference is small. For the last three populations, however, T_{JACK} has a smaller variance in most of the cases and therefore has a smaller mean squared error, since T_{JACK} has a smaller bias than T_n . It is interesting to observe that, in this example, T_n is better when F is symmetric, whereas T_{JACK} is better when F is asymmetric, in which case T_{JACK} might be useful since the bias of T_n is not negligible.

The mean squared error (rather than the bias) is commonly adopted to assess different estimators. But sometimes we also need to consider the bias. For example, although $T_n = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ has a smaller mean squared error than $T_{\text{JACK}} = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ as estimators of the population variance, many people still like the unbiased estimator $(n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ for variance estimation. In conclusion, we should

Table 2.3. Simulation results in Example 2.15 [Adapted from Parr and Schucany (1982), by permission of American Statistical Association]

F	n	J_1		J_2		J_3	
		T_n	T_{JACK}	T_n	T_{JACK}	T_n	T_{JACK}
$N(5, 1)$	5	.203	.203	.209	.238	.205	.213
	10	.093	.111	.100	.106	.097	.103
	20	.049	.050	.051	.052	.049	.051
$L(5, 1)$	5	.621	.621	.594	.625	.607	.597
	10	.337	.334	.318	.320	.320	.319
	20	.163	.159	.159	.159	.159	.158
$U(4, 6)$	5	.016	.016	.019	.026	.018	.021
	10	.009	.013	.012	.013	.011	.012
	20	.005	.006	.006	.007	.006	.006
$E(1)$	5	.174	.174	.153	.146	.160	.151
	10	.098	.090	.085	.083	.087	.084
	20	.043	.041	.040	.040	.041	.040
$IG(1, \frac{1}{2})$	5	.396	.396	.248	.165	.296	.216
	10	.164	.090	.092	.075	.108	.082
	20	.056	.045	.060	.035	.046	.039
$IG(1, 2)$	5	.094	.094	.081	.080	.085	.079
	10	.048	.040	.039	.038	.041	.038
	20	.020	.019	.018	.018	.019	.018

keep in mind that the jackknife estimator T_{JACK} is designed to eliminate bias and, therefore, can be used when the bias is an important issue. We need to balance the advantage of unbiasedness against the drawback of a large mean squared error.

2.4.3 Miscellaneous results

There are other studies and applications of the jackknife. Thorburn (1977) and Shi (1984) showed that Tukey's conjecture (A) (Section 1.3) is true in some sense. Miller (1964), Arvesen (1969), and Thorburn (1977) proved that for the case where T_n is a function of sample means or U-statistics,

$$(T_{\text{JACK}} - \theta)/\sqrt{v_{\text{JACK}}} \rightarrow_d N(0, 1),$$

which is a consequence of result (2.84). Ghosh (1985) established a Berry-Esséen bound for $(T_{\text{JACK}} - \theta)/\sqrt{v_{\text{JACK}}}$ when T_n is a function of U-statistics.

To improve the accuracy of the normal approximation to the distribution of the studentized statistic $(T_n - \theta)/\sqrt{v_{\text{JACK}}}$, Hinkley and Wei (1984)

and Tu and Zhang (1992a) suggested a method based on the estimation of the skewness of T_n by using a weighted sum of delete-1 and delete-2 jackknife pseudovalues. This estimator of skewness is also useful in the construction of the bootstrap confidence sets. These applications will be discussed further in Chapters 3-5.

In Chapters 7 and 8 we will discuss another important application of the jackknife: the use of cross-validation (a method closely related to the jackknife) for model selection and bandwidth selection.

2.5 Conclusions and Discussions

- (1) The use of the jackknife only requires a method of computing a given statistic repeatedly, whereas the traditional method requires some theoretical derivations (e.g., obtaining derivatives of g or the influence function ϕ_F) but fewer computations. The jackknife replaces the theoretical derivations by computations. In terms of their asymptotic behaviors, the jackknife and the traditional (linearization) estimators are usually first order asymptotically equivalent. Some limited empirical results show that the jackknife variance estimator is less biased but more variable than the linearization estimator.
- (2) The jackknife variance estimator v_{JACK} in (1.13) is consistent for many statistics including functions of the sample mean, U-, V-, L-statistics, M-estimators, linear rank statistics, and some statistics based on generalized empirical distributions. v_{JACK} is also consistent for statistics that are smooth transformations of several statistics. However, for R-estimators, the consistency of v_{JACK} is still an unsolved problem.
- (3) The consistency of v_{JACK} usually requires more stringent smoothness conditions on the given statistic T_n than the asymptotic normality of T_n . For example, for $T_n = g(\bar{X}_n)$, the asymptotic normality of T_n requires that g be differentiable at $\mu = EX_1$, whereas the consistency of v_{JACK} requires that ∇g be continuous at μ . Another example is the sample median, which is asymptotically normal but its jackknife variance estimator v_{JACK} is inconsistent.
- (4) For its consistency, the delete-d jackknife estimator $v_{\text{JACK-d}}$ in (2.47) requires less stringent smoothness conditions on the given statistic T_n than v_{JACK} when $d > 1$. The sample median is one example. In fact, the less smooth T_n is, the larger d needs to be. When d is a fraction of n , the consistency of $v_{\text{JACK-d}}$ requires just slightly more stringent smoothness conditions on T_n than the asymptotic normality of T_n . The delete-1 jackknife estimator is, however, computationally simpler than the delete-d jackknife estimator. The computation of the delete-d jackknife estimators will be discussed further in Chapter 5.

- (5) The jackknife provides a bias estimator and a bias-reduced estimator (compared with the original estimator T_n). Their asymptotic validity is justified for a smooth T_n . However, to use the jackknife for bias reduction we should balance the advantage of a small bias against the loss of efficiency.
- (6) The delete-d jackknife provides an additional function of the jackknife methodology: the estimation of the distribution of the given T_n . The jackknife distribution estimator is consistent under very weak conditions. However, the convergence rate of the jackknife distribution estimator is found to be not as good as the bootstrap distribution estimator defined in Section 1.4 and studied in Chapter 3. In spite of this, there are situations in which the jackknife is preferred (Wu, 1990, Section 5; Politis and Romano, 1995).
- (7) The jackknife is a nonparametric method; it works with subsets of the data set. Hence, the jackknife may not be as efficient as the bootstrap estimator, which takes into account some special features of the postulated model (e.g., the parametric bootstrap). But the jackknife is more robust against model assumptions. The jackknife employs a more systematic method of taking resamples than the bootstrap. Hence, there may be more efficient methods of computation for the jackknife.
- (8) For variance estimation, the bootstrap estimator v_{BOOT} in (1.6) may not be consistent even for very smooth T_n , when the underlying distribution F has very heavy tails (see Section 3.2.2). Also, the existing result on the consistency of v_{BOOT} does not cover a broad class of statistics. The computation of v_{BOOT} is usually more difficult than that of v_{JACK} . Thus, for variance estimation with a smooth T_n , the jackknife is preferred to the bootstrap because of both theoretical and computational reasons.
- (9) The jackknife can be easily extended to the multivariate case. For vector statistic T_n , v_{JACK} in (1.13) and $v_{\text{JACK-d}}$ in (2.47) have obvious extensions with $T_{r,s}$ representing vectors and the square being replaced by a vector product. Asymptotic results for each variance and covariance component of v_{JACK} or $v_{\text{JACK-d}}$ can be established in a similar manner.

Chapter 3

Theory for the Bootstrap

The bootstrap is a very convenient and appealing tool for statistical analysis, however, theoretical and/or empirical confirmation should be made of its suitability for the problem at hand. Also, it is important to know the relative performance of the bootstrap versus other existing methods. A general theory for the bootstrap distribution and variance estimation for a given statistic is presented in this chapter. The more delicate problem of constructing bootstrap confidence sets and hypothesis tests will be treated in the next chapter.

In Section 3.1, we discuss some basic techniques used in proving the consistency of bootstrap estimators. Section 3.2 summarizes some major results for the consistency of bootstrap estimators. The consistency of bootstrap estimators is a desirable theoretical property but may not be enough in some cases, especially when other consistent estimators are available. The use of bootstrap estimators can be justified by other more thorough asymptotic properties. In Section 3.3, the accuracy of bootstrap estimators is studied, and asymptotic comparisons of the bootstrap with other existing asymptotic methods, namely, the normal approximation and the Edgeworth expansion, are also provided. Some empirical simulation results are presented in Section 3.4 to examine and compare the performance of bootstrap estimators with other estimators. The issue of whether we should use a smoothed bootstrap is discussed in Section 3.5. Finally, some examples of inconsistent bootstrap estimators are given in Section 3.6, and we show that in some cases the inconsistency can be rectified by changing the bootstrap sample size.

In Section 3.1, our discussion is based on a general model, but we often use the i.i.d. case and the simple nonparametric bootstrap described in Example 1.6 as illustration. Results in Sections 3.2–3.6 are for the i.i.d.

case. Properties of the bootstrap under complex models will be discussed in later chapters.

3.1 Techniques in Proving Consistency

As we discussed in Section 1.6, consistency is an essential requirement for any estimator. In this section, we define the concept of consistency for bootstrap estimators, and describe some basic ideas and techniques in proving their consistency. This not only lays a foundation for the study of the consistency of bootstrap estimators, but also provides a guide for studying the bootstrap in a problem where no theoretical result has been given.

3.1.1 Bootstrap distribution estimators

Let X_1, \dots, X_n be p -dimensional observations from a stochastic model P_n and $\mathfrak{R}_n = \mathfrak{R}_n(X_1, \dots, X_n, P_n)$ be a random s -vector. We are interested in estimating H_{P_n} , the sampling distribution of \mathfrak{R}_n under model P_n . For two vectors $x = (x_1, \dots, x_k)'$ and $y = (y_1, \dots, y_k)'$, $y \leq x$ is defined as $y_j \leq x_j$ for all $j = 1, \dots, k$. Thus,

$$H_{P_n}(x) = P\{\mathfrak{R}_n \leq x \mid P_n\}. \quad (3.1)$$

When X_1, \dots, X_n are i.i.d. from a distribution F , P_n is determined by F and we denote H_{P_n} by $H_{n,F}$ or simply H_n . Let \hat{P}_n be an estimator of P_n based on X_1, \dots, X_n . Then, according to the discussion in Section 1.4, the bootstrap estimator of H_{P_n} is given by

$$H_{\text{BOOT}}(x) = H_{\hat{P}_n}(x) = P_*\{\mathfrak{R}_n^* \leq x \mid \hat{P}_n\}, \quad (3.2)$$

where $\mathfrak{R}_n^* = \mathfrak{R}_n(X_1^*, \dots, X_n^*, \hat{P}_n)$, $\{X_1^*, \dots, X_n^*\}$ is a bootstrap sample from the estimated model \hat{P}_n , and $P_*\{\cdot \mid \hat{P}_n\}$ ($= P_*\{\cdot\}$ for simplicity) is the conditional probability given \hat{P}_n .

For a fixed x , the consistency of $H_{\text{BOOT}}(x)$ as a point estimator of $H_{P_n}(x)$ can be defined using (1.26)–(1.27) or (1.28)–(1.29). But we also need to consider H_{BOOT} as a function estimator of H_{P_n} . The following is a definition of the consistency of H_{BOOT} as a function estimator.

Definition 3.1. Let ρ be a metric on $\mathcal{F}_{\mathbb{R}^s} = \{\text{all distributions on } \mathbb{R}^s\}$. H_{BOOT} is ρ -consistent (or weakly ρ -consistent) if $\rho(H_{\text{BOOT}}, H_{P_n}) \rightarrow_p 0$ as $n \rightarrow \infty$, and H_{BOOT} is strongly ρ -consistent if $\rho(H_{\text{BOOT}}, H_{P_n}) \rightarrow_{a.s.} 0$ as $n \rightarrow \infty$.

A metric ρ is involved in this definition. In the following discussion, we will specify the metric if necessary. The most commonly used metric is

ρ_∞ , the distance generated by the sup-norm (see Appendix B). When ρ_∞ is used, we abbreviate ρ_∞ -consistency to consistency.

Definition 3.1 can be generalized to the cases where \mathfrak{R}_n is an element on a general metric space. For example, $\mathfrak{R}_n = \{\sqrt{n}[F_n(x) - F(x)], x \in \mathbb{R}\}$ is a random element on the space of all functions on \mathbb{R} that are right continuous and have left limits. We will not provide a discussion on this topic since it requires some knowledge of stochastic processes and measure theory.

We now describe some basic ideas and techniques for proving the consistency of the bootstrap estimator H_{BOOT} .

3.1.2 Mallows' distance

Mallows (1972) introduced a metric on $\mathcal{F}_{r,s} = \{G \in \mathcal{F}_{\mathbb{R}^s} : \int \|x\|^r dG(x) < \infty\}$. For two distributions H and G in $\mathcal{F}_{r,s}$, their Mallows' distance is

$$\tilde{\rho}_r(H, G) = \inf_{T_{X,Y}} (E\|X - Y\|^r)^{1/r},$$

where $T_{X,Y}$ is the collection of all possible joint distributions of the pairs (X, Y) whose marginal distributions are H and G , respectively. For random U and V having distributions $H \in \mathcal{F}_{r,s}$ and $G \in \mathcal{F}_{r,s}$, respectively, we define $\tilde{\rho}_r(U, V) = \tilde{\rho}_r(H, G)$.

Some useful results for $\tilde{\rho}_r$ are proved by Bickel and Freedman (1981). The first result is that, if $G_k \in \mathcal{F}_{r,s}$ and ξ_k is the random vector having distribution G_k , $k = 0, 1, 2, \dots$, then $\tilde{\rho}_r(G_k, G_0) \rightarrow 0$ if and only if $\xi_k \rightarrow_d \xi_0$ and $\int \|x\|^r dG_k(x) \rightarrow \int \|x\|^r dG_0(x)$. Hence, $\tilde{\rho}_r$ -convergence is stronger than convergence in distribution.

Let F_n be the empirical distribution putting mass n^{-1} on each X_i , $i = 1, \dots, n$. If X_1, \dots, X_n are i.i.d. with a distribution $F \in \mathcal{F}_{r,p}$, then

$$\tilde{\rho}_r(F_n, F) \rightarrow_{a.s.} 0. \quad (3.3)$$

Let U and V be random vectors and a be a constant. Then,

$$\tilde{\rho}_r(aU, aV) = |a| \tilde{\rho}_r(U, V), \quad (3.4)$$

and, if $E\|U\|^2 < \infty$ and $E\|V\|^2 < \infty$,

$$[\tilde{\rho}_2(U, V)]^2 = [\tilde{\rho}_2(U - EU, V - EV)]^2 + \|EU - EV\|^2. \quad (3.5)$$

Let $\{U_j\}$ and $\{V_j\}$ be two sequences of independent random vectors whose distributions are in $\mathcal{F}_{r,s}$ and $EU_j = EV_j$ for all j . Then,

$$\left[\tilde{\rho}_2 \left(\sum_{j=1}^m U_j, \sum_{j=1}^m V_j \right) \right]^2 \leq \sum_{j=1}^m [\tilde{\rho}_2(U_j, V_j)]^2. \quad (3.6)$$

Bickel and Freedman (1981) used Mallows' distance and its properties in establishing the consistency of the bootstrap estimator H_{BOOT} . We now illustrate this technique by an example.

Example 3.1. The sample mean. Suppose that X_1, \dots, X_n are i.i.d. random p -vectors with distribution $F \in \mathcal{F}_{2,p}$ and $\mathfrak{R}_n = \sqrt{n}(\bar{X}_n - \mu)$, where $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ and $\mu = EX_1$. In this case, $s = p$. We consider the nonparametric bootstrap (see Example 1.6), i.e., the bootstrap data X_1^*, \dots, X_n^* are i.i.d. from F_n , the empirical distribution. Let $H_n = H_{P_n}$ and H_{BOOT} be given by (3.1) and (3.2), respectively, and E_* be the expectation taken under P_* . Then,

$$\begin{aligned}\tilde{\rho}_2(H_{\text{BOOT}}, H_n) &= \tilde{\rho}_2(\sqrt{n}(\bar{X}_n^* - \bar{X}_n), \sqrt{n}(\bar{X}_n - \mu)) \\ &= \frac{1}{\sqrt{n}} \tilde{\rho}_2\left(\sum_{i=1}^n (X_i^* - \bar{X}_n), \sum_{i=1}^n (X_i - \mu)\right) \quad [\text{by (3.4)}] \\ &\leq \sqrt{\frac{1}{n} \sum_{i=1}^n [\tilde{\rho}_2(X_i^* - \bar{X}_n, X_i - \mu)]^2} \quad [\text{by (3.6)}] \\ &= \tilde{\rho}_2(X_1^* - \bar{X}_n, X_1 - \mu) \quad [\text{by exchangeability}] \\ &= \sqrt{[\tilde{\rho}_2(X_1^*, X_1)]^2 - \|EX_1 - E_* X_1^*\|^2} \quad [\text{by (3.5)}] \\ &= \sqrt{[\tilde{\rho}_2(F_n, F)]^2 - \|\mu - \bar{X}_n\|^2} \quad [\text{by definition}] \\ &= o(1) \text{ a.s.} \quad [\text{by (3.3) and } \bar{X}_n \rightarrow_{a.s.} \mu].\end{aligned}$$

This shows the strong $\tilde{\rho}_2$ -consistency of H_{BOOT} for $T_n = \bar{X}_n$.

Because of its properties, Mallows' distance is very suitable for studying linear statistics such as the sample mean and least squares estimators in linear models. One advantage of this approach is that we do not need to find the limit distributions of \mathfrak{R}_n and \mathfrak{R}_n^* . This brings a possibility that the bootstrap can be used in some cases where the limit distribution of \mathfrak{R}_n is difficult to find or does not even exist. Bickel and Freedman (1983) provided such an example for linear models (see Chapter 7).

3.1.3 Berry-Esséen's inequality

When the limit distribution of \mathfrak{R}_n is available, we can establish the consistency of H_{BOOT} by showing that \mathfrak{R}_n^* has the same limit distribution as \mathfrak{R}_n . One way to do this is to use a Berry-Esséen inequality. Suppose that H is a distribution on \mathbb{R} which may depend on some unknown quantities. Berry-Esséen inequalities are of the form

$$\rho_\infty(H_{P_n}, H) \leq c B_n(P_n), \quad (3.7)$$

where c is a constant that does not depend on n and P_n , B_n is a known function of P_n , and $B_n(P_n) \rightarrow 0$. If \hat{P}_n satisfies the conditions under which (3.7) holds, then

$$\rho_\infty(H_{\text{BOOT}}, \hat{H}) \leq cB_n(\hat{P}_n) \quad (3.8)$$

for some \hat{H} . Note that

$$\rho_\infty(H_{\text{BOOT}}, H_{P_n}) \leq \rho_\infty(H_{\text{BOOT}}, \hat{H}) + \rho_\infty(\hat{H}, H) + \rho_\infty(H, H_{P_n}). \quad (3.9)$$

Hence, if we can show that $\rho_\infty(\hat{H}, H) \rightarrow_{a.s.} 0$ and $B_n(\hat{P}_n) \rightarrow_{a.s.} 0$, then the consistency of H_{BOOT} follows from (3.7)–(3.9). This technique was first used by Singh (1981). We use Example 3.1 again for illustration.

Example 3.1 (continued). Consider the case of $p = 1$, $\mathfrak{R}_n = \sqrt{n}(\bar{X}_n - \mu)$ and $\mathfrak{R}_n^* = \sqrt{n}(\bar{X}_n^* - \bar{X}_n)$. From the Berry-Esséen theorem for the sample mean \bar{Z}_n of i.i.d. random variables Z_1, \dots, Z_n having distribution F_Z (Appendix A.9), we obtain

$$\sup_x |P\{\sqrt{n}(\bar{Z}_n - EZ_1)/\sqrt{\text{var}(Z_1)} \leq x\} - \Phi(x)| \leq cB_n(F_Z) \quad (3.10)$$

for any F_Z with a finite third order moment, where Φ is the standard normal distribution and $B_n(F_Z) = n^{-1/2}E|Z_1 - EZ_1|^3/[\text{var}(Z_1)]^{3/2}$. Replacing F_Z and \bar{Z}_n in (3.10) by the empirical distribution F_n of X_1, \dots, X_n and \bar{X}_n^* , respectively, we obtain that

$$\sup_x |P_*\{\sqrt{n}(\bar{X}_n^* - \bar{X}_n)/\hat{\sigma} \leq x\} - \Phi(x)| \leq cB_n(F_n), \quad (3.11)$$

where $B_n(F_n) = \hat{\sigma}^{-3}n^{-3/2}\sum_{i=1}^n |X_i - \bar{X}_n|^3$ and $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. Define

$$\Phi_a(x) = \Phi(x/a). \quad (3.12)$$

Then (3.11) becomes $\rho_\infty(H_{\text{BOOT}}, \Phi_{\hat{\sigma}}) \leq cB_n(F_n)$, i.e., (3.8) holds with $\hat{H} = \Phi_{\hat{\sigma}}$. Since $\hat{\sigma}^2 \rightarrow_{a.s.} \sigma^2 = \text{var}(X_1)$, $\rho_\infty(\Phi_{\hat{\sigma}}, \Phi_\sigma) \rightarrow_{a.s.} 0$. By the central limit theorem, $\rho_\infty(H_n, \Phi_\sigma) \rightarrow 0$. From (3.9) with $H = \Phi_\sigma$ and $\hat{H} = \Phi_{\hat{\sigma}}$, $\rho_\infty(H_{\text{BOOT}}, H_n) \rightarrow_{a.s.} 0$ if $B_n(F_n) \rightarrow_{a.s.} 0$, which follows from $n^{-3/2}\sum_{i=1}^n |X_i - \bar{X}_n|^3 \rightarrow_{a.s.} 0$. But this is a direct consequence of the Marcinkiewicz strong law of large numbers under $EX_1^2 < \infty$ (see Appendix A.5).

In principle, this technique can be applied to any situation in which a Berry-Esséen type inequality is available. However, it is usually difficult to establish such inequalities. An advantage of using this technique is that we can find the convergence rate of bootstrap estimators (see Section 3.3).

3.1.4 Imitation

When we can derive the limit of H_{P_n} , i.e., $\rho(H_{P_n}, H) \rightarrow 0$ for some H , another way of proving the consistency of H_{BOOT} is to show $\rho(H_{\text{BOOT}}, H) \rightarrow 0$ by imitating the proof of $\rho(H_{P_n}, H) \rightarrow 0$. This is actually the most popular method in proving the consistency of bootstrap estimators. We now show two examples and will provide more examples in Section 3.6. The first example is given by Yang (1988).

Example 3.1 (continued). Assume $p = 1$. It is well known that the limit distribution of $H_n(x) = P\{\sqrt{n}(\bar{X}_n - \mu) \leq x\}$ is $\Phi_\sigma(x) = \Phi(x/\sigma)$. A method of proving this fact is by showing that the characteristic function of H_n tends to that of Φ_σ . Let $\psi(t)$ be the characteristic function of a random variable W with mean 0 and variance τ^2 . Then we have (e.g., Billingsley, 1979)

$$\psi(t) = 1 - \frac{1}{2}\tau^2 t^2 + \lambda(t),$$

where $|\lambda(t)| \leq |t|^2 E[\min(|t||W|^3, W^2)]$. Since X_1, \dots, X_n are i.i.d. and $(X_i - \mu)/\sqrt{n}$ has mean 0 and variance σ^2/n , the characteristic function of H_n is

$$\psi_n(t) = [1 - \frac{1}{2n}\sigma^2 t^2 + \lambda_n(t)]^n,$$

where $|\lambda_n(t)| \leq |t|^2 E[\min(|t|n^{-3/2}|X_1 - \mu|^3, n^{-1}|X_1 - \mu|^2)]$. As $n \rightarrow \infty$, $\psi_n(t) \rightarrow \exp(-\sigma^2 t^2/2)$, which is the characteristic function of Φ_σ .

We next derive the limit of $H_{\text{BOOT}}(x) = P_*\{\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x\}$ by imitating the above arguments. Since $(\bar{X}_i^* - \bar{X}_n)/\sqrt{n}$, $i = 1, \dots, n$, are i.i.d. and have mean 0 and variance $\hat{\sigma}^2/n = n^{-2} \sum_{i=1}^n (X_i - \bar{X}_n)^2$, the characteristic function of H_{BOOT} is

$$\psi_n^*(t) = [1 - \frac{1}{2n}\hat{\sigma}^2 t^2 + \lambda_n^*(t)]^n,$$

where

$$\begin{aligned} |\lambda_n^*(t)| &\leq |t|^2 E_*[\min(|t|n^{-3/2}|X_1^* - \bar{X}_n|^3, n^{-1}|X_1^* - \bar{X}_n|^2)] \\ &= \frac{|t|^2}{n} \sum_{i=1}^n [\min(|t|n^{-3/2}|X_i - \bar{X}_n|^3, n^{-1}|X_i - \bar{X}_n|^2)]. \end{aligned}$$

By the Marcinkiewicz strong law of large numbers (Appendix A.5), we obtain that $n|\lambda_n^*(t)| \rightarrow_{a.s.} 0$. Since $\hat{\sigma}^2 \rightarrow_{a.s.} \sigma^2$, we have $\psi_n^*(t) \rightarrow_{a.s.} \exp(-\sigma^2 t^2/2)$. This implies that the limit of H_{BOOT} is Φ_σ almost surely. Hence, H_{BOOT} is strongly consistent.

Since $H_n \rightarrow \Phi_\sigma$ can be established by verifying Lindeberg's condition (see, e.g., Chung, 1974, p. 205), we may also prove $H_{\text{BOOT}} \rightarrow_{a.s.} \Phi_\sigma$ by checking Lindeberg's condition.

Example 3.2. Sample quantiles. Let q be a fixed constant, $0 < q < 1$. The q -quantile of a univariate distribution F is defined to be $\theta = F^{-1}(q)$ [see (1.12)]. The sample q -quantile is then $T_n = F_n^{-1}(q)$, where F_n is the empirical distribution based on i.i.d. data X_1, \dots, X_n . Suppose that F is continuous at θ and that $f(\theta-)$ (the left-hand derivative of F at θ) and $f(\theta+)$ (the right-hand derivative of F at θ) exist and are positive. Using the method given in Serfling (1980, pp. 77-79), we now derive the limit of $H_n(x) = P\{\sqrt{n}(T_n - \theta) \leq x\}$. Let $Z_n(\Delta)$ be a binomial (n, Δ) random variable and $\Lambda_{nt} = \sqrt{n}(\Delta_{nt} - q)/\sqrt{\Delta_{nt}(1 - \Delta_{nt})}$, where $\Delta_{nt} = F(\theta + ta_t n^{-1/2})$, $a_t = \sqrt{q(1-q)}/f(\theta-)$ if $t < 0$, $a_t = \sqrt{q(1-q)}/f(\theta+)$ if $t > 0$, and $a_t = 1$ if $t = 0$. Then,

$$\begin{aligned} H_n(ta_t) &= P\{T_n \leq \theta + ta_t n^{-1/2}\} \\ &= P\{np \leq Z_n(\Delta_{nt})\} \\ &= P\left\{\frac{Z_n(\Delta_{nt}) - n\Delta_{nt}}{\sqrt{n\Delta_{nt}(1 - \Delta_{nt})}} \geq -\Lambda_{nt}\right\}. \end{aligned} \quad (3.13)$$

Using the Berry-Esséen theorem in Appendix A.9 [note that $Z_n(\Delta_{nt})$ can be expressed as a sum of i.i.d. random variables], we obtain that [see (3.10)]

$$\sup_t |H_n(ta_t) - \Phi(\Lambda_{nt})| \leq c \frac{(1 - \Delta_{nt})^2 + \Delta_{nt}^2}{\sqrt{n\Delta_{nt}(1 - \Delta_{nt})}} = O(n^{-1/2}), \quad (3.14)$$

where the last equality follows from the fact that $\Delta_{nt} \rightarrow q$. From the existence of $f(\theta-)$ and $f(\theta+)$ and the definition of a_t , we can show that, for any $t \neq 0$,

$$\frac{F(\theta + ta_t n^{-1/2}) - q}{tn^{-1/2}} - \sqrt{q(1-q)} \rightarrow 0, \quad (3.15)$$

which, with $\Delta_{nt} \rightarrow q$, implies that $\Lambda_{nt} \rightarrow t$. Thus, $\sup_t |H_n(ta_t) - \Phi(t)| \rightarrow 0$ and the limit distribution of H_n is given by $\Phi(x/a_x)$.

We next derive the limit of $H_{\text{BOOT}}(x) = P_*\{\sqrt{n}(T_n^* - T_n) \leq x\}$ using the same arguments, where $T_n^* = F_n^{*-1}(q)$ and F_n^* is the empirical distribution based on $\{X_1^*, \dots, X_n^*\}$, an i.i.d. sample from F_n . We only consider the case where $f(\theta-) = f(\theta+)$, i.e., F is continuously differentiable at θ . First, we define analogs of Δ_{nt} and Λ_{nt} : $\delta_{nt} = F_n(T_n + ta_t n^{-1/2})$ and $\lambda_{nt} = \sqrt{n}(\delta_{nt} - q)/\sqrt{\delta_{nt}(1 - \delta_{nt})}$. Then, similar to (3.13),

$$\begin{aligned} H_{\text{BOOT}}(ta_t) &= P_*\{T_n^* \leq T_n + ta_t n^{-1/2}\} \\ &= P_*\left\{\frac{Z_n(\delta_{nt}) - n\delta_{nt}}{\sqrt{n\delta_{nt}(1 - \delta_{nt})}} \geq -\lambda_{nt}\right\}. \end{aligned} \quad (3.16)$$

Similar to (3.14),

$$\sup_t |H_{\text{BOOT}}(ta_t) - \Phi(\lambda_{nt})| \leq c \frac{(1 - \delta_{nt})^2 + \delta_{nt}^2}{\sqrt{n\delta_{nt}(1 - \delta_{nt})}} = O(n^{-1/2}) \text{ a.s.} \quad (3.17)$$

since $\delta_{nt} \rightarrow_{a.s.} q$. Finally, we need to show a result similar to (3.15): $\lambda_{nt} \rightarrow_{a.s.} t$ for any t . By the fact that $f(\theta-) = f(\theta+)$,

$$\frac{F_n(T_n + ta_t n^{-1/2}) - q}{tn^{-1/2}} - \sqrt{q(1-q)} \rightarrow_{a.s.} 0 \quad (3.18)$$

for $t \neq 0$. This and $\delta_{nt} \rightarrow_{a.s.} q$ imply that $\lambda_{nt} \rightarrow_{a.s.} t$. Therefore, the limit of H_{BOOT} is $\Phi(x/a_x)$ almost surely, and H_{BOOT} is strongly consistent.

The reader may try to see if this result (the consistency of H_{BOOT}) still holds in the case where $f(\theta-) \neq f(\theta+)$. We will return to this example in Section 3.6.

3.1.5 Linearization

Linearization is another important technique in proving the consistency of bootstrap estimators, since results for linear statistics are often available or may be established using the techniques previously introduced. Suppose that a given statistic T_n can be approximated by a linear random variable $\bar{Z}_n = n^{-1} \sum_{i=1}^n \phi(X_i)$, i.e.,

$$T_n = \theta + \bar{Z}_n + o_p(n^{-1/2}). \quad (3.19)$$

Let T_n^* and \bar{Z}_n^* be the bootstrap analogs of T_n and \bar{Z}_n , respectively, based on the bootstrap sample $\{X_1^*, \dots, X_n^*\}$. If we can establish a result for T_n^* similar to (3.19), i.e.,

$$T_n^* = \theta + \bar{Z}_n^* + o_p(n^{-1/2}), \quad (3.20)$$

then the limit of $H_{\text{BOOT}}(x) = P_*\{\sqrt{n}(T_n^* - T_n) \leq x\}$ is the same as that of $P_*\{\sqrt{n}(\bar{Z}_n^* - \bar{Z}_n) \leq x\}$. We have thus reduced the problem to a problem involving a “sample mean” \bar{Z}_n , whose bootstrap distribution estimator can be shown to be consistent using the methods in Sections 3.1.2–3.1.4. An example is given as follows.

Example 3.3. Functions of the sample mean. Let g be a function from \mathbb{R}^p to \mathbb{R} and X_1, \dots, X_n be i.i.d. random vectors having finite second moments. Suppose that $\mathfrak{N}_n = \sqrt{n}(T_n - \theta)$, where $T_n = g(\bar{X}_n)$, \bar{X}_n is the sample mean, and $\theta = g(\mu)$, $\mu = EX_1$. If g is differentiable at μ , then

$$T_n - \theta = \nabla g(\mu)'(\bar{X}_n - \mu) + o_p(n^{-1/2}).$$

Let \bar{X}_n^* be the sample mean of a bootstrap sample and $T_n^* = g(\bar{X}_n^*)$. Using the results in Example 3.1, we obtain that, almost surely, the conditional distribution of $\sqrt{n}(\bar{X}_n^* - \bar{X}_n)$ converges to a p -dimensional normal distribution with mean 0. Then, by the differentiability of g ,

$$T_n^* - \theta = \nabla g(\mu)'(\bar{X}_n^* - \mu) + r_n^*,$$

where r_n^* satisfies $P_*\{\sqrt{n}|r_n^*| > \epsilon\} \rightarrow_p 0$ for any $\epsilon > 0$. Hence,

$$T_n^* - T_n = \nabla g(\mu)'(\bar{X}_n^* - \bar{X}_n) + r_n^* + o_p(n^{-1/2})$$

and

$$H_{\text{BOOT}}(x) = P_*\{\sqrt{n}(T_n^* - T_n) \leq x\} \rightarrow_p \Phi(x/\sigma_F)$$

with $\sigma_F^2 = \nabla g(\mu)'\text{var}(X_1)\nabla g(\mu)$, which is the limit distribution of \mathfrak{R}_n . This shows that the bootstrap estimator H_{BOOT} is weakly consistent.

If we assume that g is differentiable in a neighborhood of μ and ∇g is continuous at μ , then

$$T_n^* - T_n = \nabla g(\bar{X}_n)'(\bar{X}_n^* - \bar{X}_n) + \tilde{r}_n^*,$$

where \tilde{r}_n^* satisfies $P_*\{\sqrt{n}|\tilde{r}_n^*| > \epsilon\} \rightarrow_{a.s.} 0$ for any $\epsilon > 0$. Since $\nabla g(\bar{X}_n) \rightarrow_{a.s.} \nabla g(\mu)$, $H_{\text{BOOT}}(x) \rightarrow_{a.s.} \Phi(x/\sigma_F)$, i.e., the bootstrap estimator H_{BOOT} is strongly consistent.

As Example 3.3 indicates, the differential of the functional that generates T_n sometimes helps to establish (3.19) and (3.20). This will be discussed further in Section 3.2.1.

3.1.6 Convergence in moments

In addition to the estimation of the distribution of a given statistic T_n , the bootstrap can also be used to estimate the variance of T_n (see Section 1.4). The definition of the consistency of variance estimators is given in Section 1.6. The bootstrap variance estimator v_{BOOT} defined in (1.16) is actually the variance of H_{BOOT} in (3.2) with $\mathfrak{R}_n^* = T_n^* - T_n$. Thus, while the consistency of H_{BOOT} is a kind of convergence in distribution, the consistency of v_{BOOT} is the convergence of the second order moment of H_{BOOT} . If the distance $\tilde{\rho}_2$ is used, then the consistency of H_{BOOT} implies the consistency of v_{BOOT} . When the distance ρ_∞ is used, however, it is well known that convergence in distribution of a random sequence does not imply convergence in moment, unless the random sequence is also uniformly integrable (see Appendix A.7 and Serfling, 1980, Section 1.4). Thus, showing the uniform integrability of $\{n(T_n^* - T_n)^2\}$ is the main technique used in proving the consistency of v_{BOOT} when it does not have an explicit form.

A sufficient condition for the uniform integrability of a sequence of random variables $\{Z_n\}$ is that $\sup_n E|Z_n|^{1+\delta} < \infty$ for a $\delta > 0$ (e.g., Serfling, 1980, Section 1.4), which is easier to verify. Thus, if the bootstrap estimator H_{BOOT} is weakly or strongly consistent, then a sufficient condition for the weak or strong consistency of $v_{\text{BOOT}} = \text{var}_*(T_n^*)$ is that

$$E_*|\sqrt{n}(T_n^* - T_n)|^{2+\delta} = O_p(1) \quad \text{or} \quad = o(1) \text{ a.s.} \quad (3.21)$$

for a $\delta > 0$. This technique will be applied in Section 3.2.2.

3.2 Consistency: Some Major Results

Since Bickel and Freedman (1981) and Singh (1981) presented their results for the consistency of the bootstrap estimators, many results under various model assumptions have been established in the last decade. In this section, we focus on the simple nonparametric bootstrap in the i.i.d. case. Results for other models will be discussed in later chapters.

Throughout this section, we assume that X_1, \dots, X_n are i.i.d. random p -vectors from F . Let $T_n = T_n(X_1, \dots, X_n)$ be an estimator of an unknown parameter θ and $\mathfrak{R}_n = \sqrt{n}(T_n - \theta)$. Let $\{X_1^*, \dots, X_n^*\}$ be a bootstrap sample from the empirical distribution F_n based on X_1, \dots, X_n , $T_n^* = T_n(X_1^*, \dots, X_n^*)$ and $\mathfrak{R}_n^* = \sqrt{n}(T_n^* - T_n)$. In Section 3.2.1, we study the consistency of H_{BOOT} defined in (3.2), the bootstrap estimator of the sampling distribution $H_n = H_{P_n}$ defined in (3.1), for various commonly used statistics T_n . In Section 3.2.2, we study the consistency of the bootstrap variance estimator $v_{\text{BOOT}} = \text{var}_*(T_n^*)$.

3.2.1 Distribution estimators

We summarize some major results scattered in various journals. The proofs of these results are based on the techniques introduced in Section 3.1, but some of them involve heavy technical arguments and are omitted.

Functions of the sample mean

Functions of the sample mean form a large class of useful statistics. Examples can be found in Section 2.1. Combining Examples 3.1 and 3.3, we obtain the following result.

Theorem 3.1. *Suppose that $E\|X_1\|^2 < \infty$, $T_n = g(\bar{X}_n)$, and g is continuously differentiable at $\mu = EX_1$ with $\nabla g(\mu) \neq 0$. Then the bootstrap estimator H_{BOOT} is strongly consistent for H_n , i.e., $\rho_\infty(H_{\text{BOOT}}, H_n) \rightarrow_{a.s.} 0$.*

The result on the consistency of bootstrap estimators for the sample mean provides a foundation for studying the consistency of bootstrap estimators for general statistics that can be approximated by a “sample mean” (see Section 3.1.5).

It may be of interest to ask what will happen if the second moment of X_1 is not finite. Babu (1984), Athreya (1987), and Knight (1989) gave some examples showing that the bootstrap estimator H_{BOOT} is inconsistent if $E\|X_1\|^2 = \infty$. Giné and Zinn (1989) and Hall (1990a) further proved that, for $p = 1$ and $g(x) = x$, $EX_1^2 < \infty$ is a sufficient and necessary condition for the strong consistency of H_{BOOT} .

Theorem 3.2. Let X_1, \dots, X_n be i.i.d. random variables. If there exist a measurable function A_n of X_1, \dots, X_n , a strictly increasing sequence of numbers $a_n \rightarrow \infty$, and a distribution function $G(x)$ such that

$$P_* \left\{ \frac{1}{a_n} \sum_{i=1}^n X_i^* - A_n \leq x \right\} \rightarrow_{a.s.} G(x),$$

then we have $a_n/\sqrt{n} \rightarrow 1$ and $EX_1^2 < \infty$.

A similar result for the weak consistency of the bootstrap estimator H_{BOOT} is also given in Giné and Zinn (1989). Arcones and Giné (1989), Kinateder (1992), and LePage (1992) showed that the inconsistency of the bootstrap estimator in this case can be rectified by varying the bootstrap sample size (see also Section 3.6) or bootstrapping signs of $X_i - \mu$, $i = 1, \dots, n$.

U-statistic

The U-statistic

$$U_n = \binom{n}{m}^{-1} \sum_c h(X_{i_1}, \dots, X_{i_m}) \quad (3.22)$$

is a generalization of the sample mean and is an unbiased estimator of $\theta = \int \cdots \int h(x_1, \dots, x_m) dF(x_1) \cdots dF(x_m)$. Let H_n be given by (3.1) with $\mathfrak{R}_n = \sqrt{n}(U_n - \theta)$ and H_{BOOT} be given by (3.2) with $\mathfrak{R}_n^* = \sqrt{n}(U_n^* - U_n)$, where U_n^* is the bootstrap analog of U_n in (3.22). The consistency of H_{BOOT} in this case was first proved by Bickel and Freedman (1981). Shi (1986a) weakened the conditions required in Bickel and Freedman (1981). We only state the result for the special case of $m = 2$.

Theorem 3.3. Let X_1, \dots, X_n be i.i.d. random variables and U_n be given by (3.22) with $m = 2$. Suppose that $E[h(X_1, X_2)]^2 < \infty$, $E|h(X_1, X_1)| < \infty$, and $\int h(x, y) dF(y) \not\equiv \text{a constant}$. Then $\rho_\infty(H_{\text{BOOT}}, H_n) \rightarrow_{a.s.} 0$.

Comparing this result with the asymptotic normality of U_n (e.g., Serfling, 1980, p. 192), we find that the consistency of the bootstrap estimator requires an extra condition $E|h(X_1, X_1)| < \infty$. The following example, given by Bickel and Freedman (1981), shows that this condition cannot be dropped.

Example 3.4. U-statistics. Suppose that $m = 2$ and $h(x, y)$ satisfies $h(x, x) = e^{1/x}$. If the population distribution $F(x)$ is assumed to be uniform on $(0, 1)$, then we have $E|h(X_1, X_1)| = \int_0^1 e^{1/x} dx = \infty$. Write $h(x, y) = h_1(x, y) + h_2(x, y)$ with $h_1(x, y) = h(x, y)I\{x \neq y\}$. Let U_{n1} and U_{n2} be the

U -statistics corresponding to h_1 and h_2 , respectively. Then, $U_n = U_{n1} + U_{n2}$. Since F is continuous, we have $U_{n2} \equiv 0$. For the bootstrap analogs U_n^* , U_{n1}^* , and U_{n2}^* of U_n , U_{n1} , and U_{n2} , respectively, we also have $U_n^* = U_{n1}^* + U_{n2}^*$. Define $\mathfrak{R}_{n1}^* = \sqrt{n}(U_{n1}^* - U_{n1})$ and $\mathfrak{R}_{n2}^* = \sqrt{n}(U_{n2}^* - U_{n2})$. Bickel and Freedman (1981) showed that the conditional distribution of \mathfrak{R}_{n1}^* tends to $\Phi(x/\sigma)$ for some $\sigma > 0$, but $\mathfrak{R}_{n2}^* \rightarrow_p \infty$. Since $\sqrt{n}(U_n - \theta) \rightarrow_d N(0, \sigma^2)$, we can conclude that the bootstrap estimator H_{BOOT} is inconsistent.

L-statistics

The smooth L-statistics defined in Example 2.6 are special cases of the following general L-statistics:

$$T_n = \int x J(F_n(x)) dF_n(x) + \sum_{j=1}^m a_j F_n^{-1}(p_j), \quad (3.23)$$

where a_j , p_j , $j = 1, \dots, m$, are known constants, $0 < p_1 < \dots < p_m < 1$, and J is a function on $[0, 1]$. There are other forms of L-statistics; however, the difference among the various forms of L-statistics can be neglected in an asymptotic analysis, and hence we focus on T_n in (3.23) only.

A considerable amount of work has been done in proving the consistency of the bootstrap estimator H_{BOOT} , including Bickel and Freedman (1981), Babu and Singh (1984a), Falk (1988), Klenk and Stute (1987), Shao (1992a), Tu (1986a, 1989), and Tu and Cheng (1989). The following is a summary of their results.

When $J(t) \equiv 0$ and all a_j but one are 0, the L-statistic T_n reduces to a sample quantile, which is discussed in Example 3.2.

A straightforward extension of the result in Example 3.2 gives the following result.

Theorem 3.4. *Let X_1, \dots, X_n be i.i.d. random variables from F and $T_n = \sum_{j=1}^m a_j F_n^{-1}(p_j)$. Suppose that F is differentiable at each $F^{-1}(p_j)$ and $f(F^{-1}(p_j)) > 0$ [$f(t) = df/dt$]. Then, the bootstrap estimator H_{BOOT} is strongly consistent.*

For the case where $a_j = 0$ for all j , we have the following results for trimmed and untrimmed L-statistics, respectively.

Theorem 3.5. *Let T_n be given by (3.23) with $a_j = 0$ for all j .*

(i) *Trimmed L-statistics. Suppose that J is bounded, continuous a.e. Lebesgue and a.e. F^{-1} , $J = 0$ outside an interval $[\alpha, \beta]$, $0 < \alpha < \beta < 1$, and $0 < E[\phi_F(X_1)]^2 < \infty$, where $\phi_F(x) = \int [F(y) - I\{y \geq x\}]J(F(y))dy$. Then the bootstrap estimator H_{BOOT} is strongly consistent.*

- (ii) *Untrimmed L-statistics.* If J is continuous on $[0, 1]$ and $E\|X_1\|^2 < \infty$, then the bootstrap estimator H_{BOOT} is strongly consistent.

The proof of this theorem can be found in the references previously cited. Under the conditions in Theorems 3.4 and 3.5, the strong consistency of H_{BOOT} for general L-statistics T_n given by (3.23) can also be established.

Differentiable functionals¹

We have shown in Chapter 2 that many statistics can be expressed as $T_n = T(F_n)$, where F_n is the empirical distribution and T is a functional defined on \mathcal{F} , a convex class of distributions containing the population distribution F and all degenerate distributions. We have also shown in Chapter 2 that the approach of using a suitably defined differential of T is effective for studying the consistency of jackknife estimators. We now establish the consistency of the bootstrap estimator for $T_n = T(F_n)$ with a T that is differentiable in some sense. The basic idea was outlined in Section 3.1.5. For differentiable T , (3.19) usually holds. If we can establish (3.20) for the bootstrap analog T_n^* , then the bootstrap estimator H_{BOOT} is weakly consistent.

Definitions of Hadamard, Fréchet differentiability and continuous Hadamard, Fréchet differentiability used in the following theorem can be found in Section 2.2.

Theorem 3.6. *Let $T_n = T(F_n)$.*

- (i) *Suppose that T is ρ_∞ -Hadamard differentiable at F with the influence function satisfying $0 < E[\phi_F(X_1)]^2 < \infty$. Then the bootstrap estimator H_{BOOT} is weakly consistent. If, in addition, T is continuously ρ_∞ -Hadamard differentiable at F , then H_{BOOT} is strongly consistent.*
- (ii) *Suppose that T is ρ_r -Fréchet differentiable at F with the influence function satisfying $0 < E[\phi_F(X_1)]^2 < \infty$ and that $\int \{F(x)[1-F(x)]\}^{r/2} dx < \infty$. Then H_{BOOT} is weakly consistent. If, in addition, T is continuously ρ_r -Fréchet differentiable at F , then H_{BOOT} is strongly consistent.*

Proof. A complete proof of this theorem is mathematically too complicated and is omitted. For illustration, we only give a proof for the special case where T is continuously ρ_∞ -Fréchet differentiable at F . A complete proof for (i) can be found in Gill (1989) and Liu, Singh and Lo (1989). A complete proof for (ii) can be found in Yang (1985) and Shao (1992a).

Suppose that T is continuously ρ_∞ -Fréchet differentiable at F . Then

$$T_n - T(F) = \bar{Z}_n + o_p(n^{-1/2}), \quad (3.24)$$

¹This part may be omitted if Section 2.2 was omitted.

where $\bar{Z}_n = n^{-1} \sum_{i=1}^n Z_i$ and $Z_i = \phi_F(X_i)$, $i = 1, \dots, n$. Let F_n^* be the empirical distribution putting mass n^{-1} on X_i^* , $i = 1, \dots, n$, that are i.i.d. from F_n , $T_n^* = T(F_n^*)$, $Z_i^* = \phi_F(X_i^*)$, $\bar{Z}_n^* = n^{-1} \sum_{i=1}^n Z_i^*$, and $r_n^* = T_n^* - T_n - (\bar{Z}_n^* - \bar{Z}_n)$. We first show

$$P_*\{\sqrt{n}|r_n^*| > \eta\} \rightarrow_{a.s.} 0 \quad (3.25)$$

for any $\eta > 0$. From the differentiability of T , for any $\epsilon > 0$, there is a $\delta_\epsilon > 0$ such that

$$P_*\{\sqrt{n}|r_n^*| > \eta\} \leq P_*\{\sqrt{n}\rho_\infty(F_n^*, F_n) > \eta/\epsilon\} + P_*\{\rho_\infty(F_n^*, F_n) > \delta_\epsilon\}.$$

From Dvoretzky, Kiefer, and Wolfowitz's inequality (see, e.g., Serfling, 1980, Theorem 2.1.3A), there is a constant c such that

$$P_*\{\sqrt{n}\rho_\infty(F_n^*, F_n) > \eta/\epsilon\} \leq ce^{-2\eta^2/\epsilon^2}$$

and

$$P_*\{\rho_\infty(F_n^*, F_n) > \delta_\epsilon\} \leq ce^{-2n\delta_\epsilon^2}.$$

Then (3.25) follows since ϵ is arbitrary. This proves that, for almost all given sequences X_1, X_2, \dots , the conditional distribution of $\sqrt{n}(T_n^* - T_n)$ has the same limit as the conditional distribution of $\sqrt{n}(\bar{Z}_n^* - \bar{Z}_n)$. Hence, by Theorem 3.1 and (3.24), $\rho_\infty(H_{\text{BOOT}}, H_n) \rightarrow_{a.s.} 0$. \square

Statistics $T_n = T(F_n)$ with a ρ_∞ -Hadamard differentiable T include M-estimators (Example 2.5), trimmed smooth L-statistics (Example 2.6), linear rank statistics (Example 2.7), R-estimators (Example 2.10), etc. An example of $T_n = T(F_n)$ with a ρ_r -Fréchet differentiable T is the L-statistic defined in (3.23) with $a_j = 0$ for all j and a function J that is Lipschitz continuous of order $r - 1$.

Lohse (1987) defined a concept of K-differentiation and applied it to a discussion of the consistency of the bootstrap estimators for $T_n = T(\eta_n(X_1, \dots, X_n))$, where η_n is a map from the sample space to the collection of all distribution functions.

Statistics based on generalized empirical distributions²

In Example 2.11, we discussed statistics of the form $T_n = T(\Xi_n)$, where

$$\Xi_n(x) = \binom{n}{m}^{-1} \sum_c I\{h(X_{i_1}, \dots, X_{i_m}) \leq x\}$$

[see (2.25)]. Let Ξ_n^* be the bootstrap analog of Ξ_n and $T_n^* = T(\Xi_n^*)$. Helmers, Janssen and Serfling (1988) established some results for Ξ_n^* , which lead to the first part of the following result:

²This part may be omitted if Section 2.2 was omitted.

Theorem 3.7. Let $T_n = T(\Xi_n)$.

- (i) Suppose that T is ρ_∞ -Hadamard differentiable at Ξ [the distribution of $h(X_1, \dots, X_m)$]. Then the bootstrap estimator H_{BOOT} is weakly consistent.
- (ii) Suppose that T is $\rho_{\infty+1}$ -Fréchet differentiable at F and that for all i_1, \dots, i_m satisfying $1 \leq i_1 \leq i_2 \leq \dots \leq i_m \leq m$,

$$\int \{\Xi^{i_1, \dots, i_m}(x)[1 - \Xi^{i_1, \dots, i_m}(x)]\}^{1/2} dx < \infty,$$

where Ξ^{i_1, \dots, i_m} is the distribution of $h(X_{i_1}, \dots, X_{i_m})$. Then H_{BOOT} is weakly consistent.

The proof for (i) is similar to that of part (i) of Theorem 3.6. The proof for (ii) is similar to the proof of Theorem 2 in Shao (1989a).

An example of $T(\Xi_n)$ is the generalized L-statistics (Serfling, 1984):

$$T_n = \int x J(\Xi_n(x)) d\Xi_n(x) + \sum_{t=1}^m a_t \Xi_n^{-1}(p_t), \quad (3.26)$$

where a_t, p_t are constants, $0 < p_t < 1$, and J is a function on $[0, 1]$. Since the quantile functional is ρ_∞ -Hadamard differentiable (Fernholz, 1983) and the functional $\int x J(G(x)) dG(x)$ is ρ_∞ -Fréchet (or $\rho_{\infty+1}$ -Fréchet) differentiable (see Example 2.6), Theorem 3.7 can be applied to T_n in (3.26). In fact, Helmers, Janssen and Serfling (1990) proved the strong consistency of H_{BOOT} for T_n in (3.26) with $a_t = 0$ for all t using the Berry-Esséen bound. Helmers, Janssen and Veraverbeke (1992) proved the strong consistency of H_{BOOT} for T_n in (3.26) with $J \equiv 0$.

Empirical and quantile processes

The bootstrap can also be used to approximate the distributions of stochastic processes. Two useful processes in statistical analysis are the empirical distribution process $\{\sqrt{n}[F_n(x) - F(x)], x \in \mathbb{R}\}$ and the quantile process $\{\sqrt{n}[F_n^{-1}(t) - F^{-1}(t)], t_0 \leq t \leq t_1\}$, where $0 < t_0 \leq t_1 < 1$. Consistency of the bootstrap distribution estimators for these processes (in the sense of a generalized Definition 3.1) are established by Bickel and Freedman (1981), Singh (1981), Shorack (1982), Beran, LeCam and Millar (1987), Gaenssler (1987), Csörgő and Mason (1989), and Giné and Zinn (1990). We omit the detailed discussion, which requires higher level mathematics.

Discussions

We have seen that the bootstrap distribution estimator H_{BOOT} is consistent for many commonly used statistics. However, there are situations where H_{BOOT} is inconsistent. Two examples are given in Theorem 3.2 and Example 3.4, respectively. More examples can be found later (see also Mammen,

1992). The inconsistency of bootstrap estimators is mainly caused by one or a combination of the following reasons:

- (1) The bootstrap is sensitive to the tail behavior of the population F . The consistency of H_{BOOT} requires moment conditions that are more stringent than those needed for the existence of the limit of H_n .
- (2) The consistency of bootstrap estimators requires a certain degree of smoothness from the given statistic T_n .
- (3) The behavior of the bootstrap estimator sometimes depends on the method used to obtain bootstrap data.

We will address further reasons (1)–(3) and provide more examples in Section 3.5 and in later chapters.

3.2.2 Variance estimators

As we discussed in Chapter 1, the bootstrap can also be used to estimate some accuracy measures of a given statistic T_n such as the variance and the bias of T_n . Because of the importance of variance estimation and the similarities between variance estimation and the estimation of other accuracy measures, we focus on the consistency of bootstrap variance estimators.

The main technique was discussed in Section 3.1.6; that is, we try to show that $\{n(T_n^* - T_n)^2\}$ is uniformly integrable or that (3.21) holds.

Again, we start with the simple case where $T_n = g(\bar{X}_n)$, a function of the sample mean. It is expected that the consistency of v_{BOOT} requires smoothness of the function g . However, the following example shows that v_{BOOT} may still be inconsistent even if g is very smooth.

Example 3.5. Inconsistency of v_{BOOT} . Let F be a univariate distribution satisfying $F(x) = 1 - x^{-h}$ if $x > 10$ and $F(x) = |x|^{-h}$ if $x < -10$, where h is a constant. Thus, F has a finite s th moment for any $s < h$. In particular, F has a finite second moment if $h > 2$. Let $t > h$ be a constant and $g(x) = e^{x^t}$. Following the proof in Ghosh *et al.* (1984, Example), the bootstrap variance estimator for the case where $T_n = g(\bar{X}_n)$ is inconsistent if

$$n^{-n+1} |g(X_{(n)})|^2 \rightarrow_{a.s.} \infty, \quad (3.27)$$

where $X_{(n)} = \max(X_1, \dots, X_n)$. In fact, under (3.27), $n v_{\text{BOOT}} \rightarrow_{a.s.} \infty$.

To show (3.27), note that, for any $M > 0$,

$$\begin{aligned} P\{|g(X_{(n)})|^2 < Mn^{n+1}\} &\leq P\{X_{(n)} < [\log(M^{1/2}n^{(n-1)/2})]^{1/t}\} \\ &= \{1 - [\log(M^{1/2}n^{(n-1)/2})]^{-h/t}\}^n \\ &\leq \exp\{-n[\log(M^{1/2}n^{(n-1)/2})]^{-h/t}\} \leq n^{-2} \end{aligned}$$

for large n . Thus, (3.27) follows from the Borel-Cantelli lemma (see Appendix A.4).

This example shows that the bootstrap variance estimator may diverge to infinity while the asymptotic variance of T_n is finite. The inconsistency of the bootstrap estimator is caused by the fact that $|T_n^* - T_n|$ may take some exceptionally large values. Hence, some moment condition has to be imposed for the consistency of v_{BOOT} . We consider a sufficient condition:

$$\max_{i_1, \dots, i_n} |T_n(X_{i_1}, \dots, X_{i_n}) - T_n| / \tau_n \rightarrow_{a.s.} 0, \quad (3.28)$$

where the maximum is taken over all integers i_1, \dots, i_n satisfying $1 \leq i_1 \leq \dots \leq i_n \leq n$, and $\{\tau_n\}$ is a sequence of positive numbers satisfying $\liminf_n \tau_n > 0$ and $\tau_n = O(e^{n^q})$ with a $q \in (0, \frac{1}{2})$.

Theorem 3.8. *Let $T_n = g(\bar{X}_n)$. Suppose that (3.28) holds, $E\|X_1\|^2 < \infty$, and g is continuously differentiable in a neighborhood of $\mu = EX_1$ with $\nabla g(\mu) \neq 0$. Then the bootstrap estimator v_{BOOT} is strongly consistent, i.e., $v_{\text{BOOT}}/\sigma_n^2 \rightarrow_{a.s.} 1$, where $\sigma_n^2 = n^{-1}\nabla g(\mu)' \Sigma \nabla g(\mu)$ and $\Sigma = \text{var}(X_1)$.*

Proof. We only need to show (3.21). Let τ_n be given by (3.28) and

$$\Delta_n^* = \begin{cases} \tau_n & \text{if } T_n^* - T_n > \tau_n \\ T_n^* - T_n & \text{if } |T_n^* - T_n| \leq \tau_n \\ -\tau_n & \text{if } T_n^* - T_n < -\tau_n. \end{cases} \quad (3.29)$$

From condition (3.28),

$$P\{E_*|T_n^* - T_n|^{2+\delta} = E_*|\Delta_n^*|^{2+\delta} \text{ for sufficiently large } n\} = 1. \quad (3.30)$$

Hence, it suffices to show that, for some $\delta > 0$,

$$E_*|\sqrt{n} \Delta_n^*|^{2+\delta} = O(1) \quad a.s. \quad (3.31)$$

We show (3.31) with $\delta = 2$. Since ∇g is continuous in a neighborhood of μ , there are positive constants η and M such that $\text{tr}[\nabla g(x)' \nabla g(x)] \leq M$ if $\|x - \mu\| \leq 2\eta$. By the strong law of large numbers, almost surely,

$$\bar{X}_n \rightarrow \mu \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)' \rightarrow \Sigma. \quad (3.32)$$

Let X_{ij} and \bar{X}_{nj} be the j th components of X_i and \bar{X}_n , respectively. By the Marcinkiewicz strong law of large numbers (Appendix A.5), almost surely,

$$\frac{1}{n^2} \sum_{i=1}^n (X_{ij} - \bar{X}_{nj})^4 \leq \frac{16}{n^2} \sum_{i=1}^n (X_{ij} - EX_{ij})^4 \rightarrow 0, \quad 1 \leq j \leq p. \quad (3.33)$$

Let X_1, X_2, \dots be a sequence such that (3.32) and (3.33) hold. Then, for large n , $\|\bar{X}_n - \mu\| \leq \eta$ and

$$\begin{aligned} E_* |\Delta_n^*|^4 &= E_* |\Delta_n^*|^4 I\{\|\bar{X}_n^* - \bar{X}_n\| \leq \eta\} + E_* |\Delta_n^*|^4 I\{\|\bar{X}_n^* - \bar{X}_n\| > \eta\} \\ &\leq E_* |T_n^* - T_n|^4 I\{\|\bar{X}_n^* - \bar{X}_n\| \leq \eta\} + \tau_n^4 E_* I\{\|\bar{X}_n^* - \bar{X}_n\| > \eta\} \\ &\leq M^2 E_* \|\bar{X}_n^* - \bar{X}_n\|^4 + \tau_n^4 P_* \{\|\bar{X}_n^* - \bar{X}_n\| > \eta\}, \end{aligned}$$

where the last inequality follows from $T_n^* - T_n = \nabla g(\xi_n^*)'(\bar{X}_n^* - \bar{X}_n)$ (the mean-value theorem) with ξ_n^* satisfying $\|\xi_n^* - \mu\| \leq \|\bar{X}_n - \mu\| + \|\xi_n^* - \bar{X}_n\| \leq \eta + \|\bar{X}_n^* - \bar{X}_n\|$. From (3.32) and (3.33), $E_* \|\bar{X}_n^* - \bar{X}_n\|^4 = O(n^{-2})$. It remains to be shown that

$$n^2 \tau_n^4 P_* \{\|\bar{X}_n^* - \bar{X}_n\| > \eta\} = O(1) \quad a.s. \quad (3.34)$$

Note that X_1^*, \dots, X_n^* are i.i.d. with mean \bar{X}_n and $|X_{ij}^* - \bar{X}_{nj}| \leq Y_j$, $j = 1, \dots, p$, where $Y_j = |\max_{i \leq n} X_{ij}| + |\min_{i \leq n} X_{ij}|$. Then, by Bernstein's inequality (e.g., Serfling, 1980, p. 95),

$$P_* \{\|\bar{X}_n^* - \bar{X}_n\| > \eta\} \leq c \exp \left[- \frac{n\eta^2}{2(W + \eta Y)} \right],$$

where c is a constant, $Y = \max_{j \leq p} Y_j$, and

$$W = \sum_{j=1}^p \text{var}_*(X_{ij}^*) = \frac{1}{n} \sum_{j=1}^p \sum_{i=1}^n (X_{ij} - \bar{X}_{nj})^2.$$

Since $E\|X_1\|^2 < \infty$, $(W + \eta Y)/\sqrt{n} \rightarrow_{a.s.} 0$, which, together with the condition that $\tau_n = O(e^{n^q})$, implies (3.34). This completes the proof. \square

From the proof of Theorem 3.8, if we define a variance estimator

$$v_{\text{TB}} = \text{var}_*(\Delta_n^*), \quad (3.35)$$

then v_{TB} is strongly consistent, i.e., $v_{\text{TB}}/\sigma_n^2 \rightarrow_{a.s.} 1$. Condition (3.28) actually ensures that

$$P\{v_{\text{BOOT}} = v_{\text{TB}} \text{ for sufficiently large } n\} = 1.$$

But the consistency of v_{TB} does not require (3.28). This is because Δ_n^* in (3.29) is obtained by truncating $T_n^* - T_n$ at τ_n and $-\tau_n$, and this truncation circumvents the problems of the bootstrap when the distribution of T_n has heavy tails. In Example 3.5, v_{BOOT} is inconsistent but v_{TB} is strongly consistent. In fact, the consistency of v_{BOOT} always implies the consistency of v_{TB} (Shao, 1990b), and v_{TB} may have a better performance than v_{BOOT} when both of them are consistent (Shao, 1992b).

If the computation of v_{BOOT} requires a Monte Carlo approximation [see (1.18)], then v_{TB} can be computed by the same Monte Carlo approximation with $T_{n,b}^* - T_n$ in (1.18) truncated at τ_n and $-\tau_n$.

We now consider other statistics T_n .

Theorem 3.9. *Let F be a univariate distribution, $\theta = F^{-1}(t)$ and $T_n = F_n^{-1}(t)$, $0 < t < 1$. Suppose that F is differentiable in a neighborhood of θ and $f = dF/dx$ is positive at θ . Suppose also that*

$$E|X_1|^\epsilon < \infty \quad (3.36)$$

for an $\epsilon > 0$. Then the bootstrap variance estimator is strongly consistent, i.e., $v_{\text{BOOT}}/\sigma_n^2 \rightarrow_{a.s.} 1$, where $\sigma_n^2 = n^{-1}t(1-t)/[f(\theta)]^2$ is the asymptotic variance of T_n .

This result was first established by Ghosh *et al.* (1984). Babu (1986) obtained the same result under a moment condition weaker than (3.36).

Let v_{TB} be given by (3.35) with $\tau_n \equiv 1$. Under (3.36), result (3.30) holds. On the other hand, the consistency of v_{TB} does not require condition (3.36), which is necessary for the consistency of v_{BOOT} .

Result (3.30) actually holds for a class of estimators of the form

$$T_n = \sum_{i=1}^n c_{ni} X_{(i)} \quad \text{with } \sup_n \sum_{i=1}^n |c_{ni}| < \infty, \quad (3.37)$$

where $X_{(i)}$ is the i th order statistic of random variables X_1, \dots, X_n and c_{ni} are constants. Note that the sample mean, the sample quantile, and more generally, the L-statistics given in Example 2.6 are of the form (3.37). If (3.36) holds, then

$$\frac{|T_n^* - T_n|}{n^{1/\epsilon}} \leq 2 \left(\sup_n \sum_{i=1}^n |c_{ni}| \right) \frac{|X_{(n)}| + |X_{(1)}|}{n^{1/\epsilon}} \rightarrow_{a.s.} 0$$

(see Appendix A.5). Hence, (3.28) holds with $\tau_n = n^{1/\epsilon}$, which implies (3.30).

The following result³ shows that v_{TB} is consistent when T_n is generated by ρ_∞ -Fréchet differentiable functionals. The bootstrap estimator v_{BOOT} is also consistent under the extra condition (3.28).

Theorem 3.10. *Let $T_n = T(F_n)$ and v_{TB} be given by (3.35) with $\tau_n \not\rightarrow 0$ and $\tau_n = O(e^{n^q})$, $q \in (0, \frac{1}{2})$.*

³Omit the rest of this section if Section 2.2 was omitted.

- (i) Assume that T is ρ_∞ -Fréchet differentiable at F . Then v_{TB} is weakly consistent, i.e., $v_{\text{TB}}/\sigma_n^2 \rightarrow_p 1$, where $\sigma_n^2 = n^{-1}E[\phi_F(X_1)]^2 > 0$. If, in addition, (3.28) holds, then $v_{\text{BOOT}}/\sigma_n^2 \rightarrow_p 1$.
- (ii) Assume that T is continuously ρ_∞ -Fréchet differentiable at F . Then v_{TB} is strongly consistent, i.e., $v_{\text{TB}}/\sigma_n^2 \rightarrow_{a.s.} 1$. If, in addition, (3.28) holds, then $v_{\text{BOOT}}/\sigma_n^2 \rightarrow_{a.s.} 1$.

Proof. We prove part (ii) only. The proof of part (i) is in Shao (1990b). From Theorem 3.1 and the discussion in the proof of Theorem 3.8, we only need to show that, for almost all given sequences $X_1, X_2, \dots, \{n\Delta_n^{*2}\}$ is uniformly integrable with respect to P_* . By the continuous differentiability of T ,

$$T_n^* - T_n = \frac{1}{n} \sum_{i=1}^n \phi_F(X_i^*) - \frac{1}{n} \sum_{i=1}^n \phi_F(X_i) + r_n^*$$

and $|r_n^*| \leq \rho_\infty(F_n^*, F_n)$ whenever $\rho_\infty(F_n^*, F) \leq \eta$ and $\rho_\infty(F_n, F) \leq \eta$, where $\eta > 0$ is a fixed constant. Let $B_n = \{\rho_\infty(F_n^*, F) \leq \eta\}$ and B_n^c be the complement of B_n . Since $\rho_\infty(F_n, F) \rightarrow_{a.s.} 0$, for almost all X_1, X_2, \dots , $\rho_\infty(F_n, F) \leq \eta/2$ for sufficiently large n . Then $\rho_\infty(F_n^*, F) > \eta$ implies $\rho_\infty(F_n^*, F_n) > \eta/2$ and, for any $\delta > 0$,

$$\begin{aligned} E_* |\sqrt{n} \Delta_n^{*2} I\{B_n^c\}|^{2+\delta} &\leq n^{1+\delta/2} \tau_n^{2+\delta} P_* \{\rho_\infty(F_n^*, F_n) > \eta/2\} \\ &\leq cn^{1+\delta/2} \tau_n^{2+\delta} \exp(-\eta^2 n/2) \rightarrow 0, \end{aligned}$$

where c is a constant and the last inequality follows from Dvoretzky, Kiefer, and Wolfowitz's inequality. It remains to be shown that for almost all given sequences $X_1, X_2, \dots, \{n\Delta_n^{*2} I\{B_n\}\}$ is uniformly integrable. Note that

$$n\Delta_n^{*2} I\{B_n\} \leq 2n \left[\frac{1}{n} \sum_{i=1}^n \phi_F(X_i^*) - \frac{1}{n} \sum_{i=1}^n \phi_F(X_i) \right]^2 + 2nr_n^{*2} I\{B_n\}$$

and $\{n[n^{-1} \sum_{i=1}^n \phi_F(X_i^*) - n^{-1} \sum_{i=1}^n \phi_F(X_i)]^2\}$ is uniformly integrable a.s., since $n^{-1} \sum_{i=1}^n \phi_F(X_i)$ is a linear functional. Hence, it remains to be shown that for almost all $X_1, X_2, \dots, \{nr_n^{*2} I\{B_n\}\}$ is uniformly integrable. For $\rho_\infty(F_n, F) \leq \eta/2$,

$$\begin{aligned} E_* |\sqrt{n} r_n^* I\{B_n\}|^{2+\delta} &= (2+\delta) \int_0^\infty t^{1+\delta} P_* \{|\sqrt{n} r_n^* I\{B_n\}| > t\} dt \\ &\leq (2+\delta) \int_0^\infty t^{1+\delta} P_* \{\sqrt{n} \rho_\infty(F_n^*, F_n) > t\} dt \\ &\leq c(2+\delta) \int_0^\infty t^{1+\delta} \exp(-2t^2) dt, \end{aligned}$$

where the last inequality follows from Dvoretzky, Kiefer, and Wolfowitz's inequality. This proves that $E_* |\sqrt{n} r_n^* I\{B_n\}|^{2+\delta} = O(1)$ a.s. Hence, the proof is completed. \square

3.3 Accuracy and Asymptotic Comparisons

The consistency of the bootstrap estimator H_{BOOT} in (3.2) asserts that H_{BOOT} will be as close as possible to H_{P_n} in (3.1) as $n \rightarrow \infty$. It is important to study further the convergence rate of $\rho_\infty(H_{\text{BOOT}}, H_{P_n})$ to 0 so that we can know at what rate H_{BOOT} gets close to H_{P_n} . Also, it is of interest to study the accuracy of H_{BOOT} in terms of other measures such as the mean squared error or asymptotic mean squared error.

There may exist other estimators of H_{P_n} , especially when \mathfrak{R}_n is simple. For example, the normal approximation supplies an estimator when \mathfrak{R}_n is asymptotically normal. Another example is the estimator obtained by using a one-term Edgeworth expansion. Therefore, in addition to the attractive features of the bootstrap described in Chapter 1, it is also important to see whether the bootstrap provides a better estimator of H_{P_n} than the other methods.

In this section⁴ we study the accuracy of the bootstrap estimator H_{BOOT} and compare it with other estimators in the case where X_1, \dots, X_n are i.i.d. random variables and the bootstrap data are taken from the empirical distribution F_n . Any comparison among several methods has to be made under a given criterion. After we obtain the convergence rate of the bootstrap estimator, some comparisons are made in terms of the convergence rates in Section 3.3.1. However, in many cases the convergence rate of the bootstrap estimator is the same as that of other estimators. Thus, refined analyses and comparisons are desired. We consider the asymptotic minimaxity in Section 3.3.2, the asymptotic mean squared error in Section 3.3.3, and the asymptotic relative error in Section 3.3.4.

The main tools for the studies in this section are the Berry-Esséen inequalities and the Edgeworth expansions of distribution functions. Some of these inequalities and expansions are listed in Appendix A.10 for the special case of the sample mean. We will only give the forms of the Edgeworth expansions when they are needed. The detailed derivations and proofs of the Edgeworth expansions are omitted, which does not affect the understanding our discussion. The reader may find details about the Berry-Esséen inequalities and the Edgeworth expansions in Petrov (1975) and Hall (1992d).

3.3.1 Convergence rate

The convergence rate of the bootstrap distribution estimators has been studied by many researchers. Results for the case where the given statistic T_n is the sample mean or a function of the sample mean are presented first.

⁴For application purposes, the reader may skip this section and read only the conclusions in Section 3.3.5.

Because it is very difficult to establish the Berry-Esséen inequalities and the Edgeworth expansions, there are not many results available when the statistic T_n is not a function of the sample mean. We summarize some available results in this direction.

The sample mean and standardized sample mean

Consider the problem of the sample mean and use the same notation as in Example 3.1. Bickel and Freedman (1980) and Singh (1981) studied the accuracy of

$$H_{\text{BOOT}}(x) = P_*\{\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x\},$$

the bootstrap estimator of

$$H_n(x) = P\{\sqrt{n}(\bar{X}_n - \mu) \leq x\},$$

and the accuracy of $\tilde{H}_{\text{BOOT}}(x) = H_{\text{BOOT}}(\hat{\sigma}x)$ as an estimator of $\tilde{H}_n(x) = H_n(\sigma x)$, the distribution of the standardized sample mean $\sqrt{n}(\bar{X}_n - \mu)/\sigma$, where σ^2 is the variance of X_i and $\hat{\sigma}^2$ is the sample variance. Their results are summarized in the following theorem.

Theorem 3.11. *Let X_1, \dots, X_n be i.i.d. random variables.*

(i) *If $E|X_1|^4 < \infty$, then*

$$\limsup_{n \rightarrow \infty} \frac{\sqrt{n} \rho_\infty(H_{\text{BOOT}}, H_n)}{\sqrt{\log \log n}} = \frac{\sqrt{\text{var}(\bar{X}_1 - \mu)^2}}{2\sigma^2 \sqrt{\pi e}} \quad a.s.$$

(ii) *If $E|X_1|^3 < \infty$ and F is lattice in the sense that there are constants c and h such that $P\{X_1 = c + kh, k = 0, 1, 2, \dots\} = 1$, then*

$$\limsup_{n \rightarrow \infty} \sqrt{n} \rho_\infty(\tilde{H}_{\text{BOOT}}, \tilde{H}_n) = \frac{h}{\sqrt{2\pi}\sigma} \quad a.s.$$

(iii) *If $E|X_1|^3 < \infty$ and F is nonlattice, then*

$$\sqrt{n} \rho_\infty(\tilde{H}_{\text{BOOT}}, \tilde{H}_n) \rightarrow_{a.s.} 0.$$

Proof. It follows from (3.10) and (3.11) that

$$\rho_\infty(\tilde{H}_n, \Phi) \leq cn^{-1/2}\sigma^{-3}E|X_1 - \mu|^3 \tag{3.38}$$

and

$$\rho_\infty(\tilde{H}_{\text{BOOT}}, \Phi) \leq cn^{-1/2}\hat{\sigma}^{-3}E_*|X_1^* - \bar{X}_n|^3. \tag{3.39}$$

From (3.38), (3.39), $E|X_1|^3 < \infty$, and the fact that $H_n(x) = \tilde{H}_n(x/\sigma)$ and $H_{\text{BOOT}}(x) = \tilde{H}_{\text{BOOT}}(x/\hat{\sigma})$, we have

$$\begin{aligned} \rho_\infty(H_{\text{BOOT}}, H_n) &\leq \rho_\infty(H_{\text{BOOT}}, \Phi_{\hat{\sigma}}) + \rho_\infty(\Phi_\sigma, H_n) + \rho_\infty(\Phi_{\hat{\sigma}}, \Phi_\sigma) \\ &= \rho_\infty(\Phi_{\hat{\sigma}}, \Phi_\sigma) + O(n^{-1/2}) \quad a.s., \end{aligned}$$

where Φ_a is defined in (3.12). Then the result in (i) follows from

$$\limsup_{n \rightarrow \infty} \frac{\sqrt{n} \rho_\infty(\Phi_{\hat{\sigma}}, \Phi_\sigma)}{\sqrt{\log \log n}} = \frac{\sqrt{\text{var}(X_1 - \mu)^2}}{2\sigma^2 \sqrt{\pi e}} \quad a.s.,$$

which can be proved using the law of the iterated logarithm (Appendix A.6).

The proof for (ii) and (iii) relies on asymptotic expansions of distribution functions. Since the expansion for a lattice distribution is more involved, we only outline how to prove (iii) using the Edgeworth expansion.

Suppose that F is nonlattice and $E|X_1|^3 < \infty$. Using the Edgeworth expansion for the sum of i.i.d. random variables (Appendix A.10), we obtain that

$$\sup_x \left| \tilde{H}_n(x) - \Phi(x) - \frac{\alpha_3(1-x^2)\varphi(x)}{6\sigma^3\sqrt{n}} \right| = o\left(\frac{1}{\sqrt{n}}\right), \quad (3.40)$$

where $\alpha_3 = E(X_1 - \mu)^3$ and $\varphi(x) = d\Phi/dx$. We also need a similar expansion for the bootstrap distribution \tilde{H}_{BOOT} , but it cannot be obtained by substituting F_n for F in (3.40), since the validity of (3.40) requires that F be nonlattice. It can be directly shown, however, that

$$\sup_x \left| \tilde{H}_{\text{BOOT}}(x) - \Phi(x) - \frac{\tilde{\alpha}_3(1-x^2)\varphi(x)}{6\hat{\sigma}^3\sqrt{n}} \right| = o\left(\frac{1}{\sqrt{n}}\right) \quad a.s., \quad (3.41)$$

where $\tilde{\alpha}_3 = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^3$. The proof of (3.41) is more difficult than that of (3.40), and the details can be found in Singh (1981) or Hall (1992d). It follows from (3.40) and (3.41) that

$$\rho_\infty(\tilde{H}_{\text{BOOT}}, \tilde{H}_n) \leq \left| \frac{\tilde{\alpha}_3}{6\hat{\sigma}^3} - \frac{\alpha_3}{6\sigma^3} \right| \sup_x \left| \frac{(1-x^2)\varphi(x)}{\sqrt{n}} \right| + o\left(\frac{1}{\sqrt{n}}\right) \quad a.s.$$

Since $\tilde{\alpha}_3 \rightarrow_{a.s.} \alpha_3$ and $\hat{\sigma} \rightarrow_{a.s.} \sigma$, the result in (iii) follows. \square

Theorem 3.11 indicates that if $EX_1^4 < \infty$, then the convergence rate of H_{BOOT} is $O(\sqrt{\log \log n/n})$, which is the same as the convergence rate of the normal approximation $\Phi_{\hat{\sigma}}$; if $E|X_1|^3 < \infty$, then the convergence rate of \tilde{H}_{BOOT} is $O(n^{-1/2})$ in general, but $o(n^{-1/2})$ when F is nonlattice. Note that for the normal approximation Φ to \tilde{H}_n , $\rho_\infty(\tilde{H}_n, \Phi) = O(n^{-1/2})$ by the Berry-Esséen theorem. This rate cannot be improved, whether F is lattice or not. Therefore, the last conclusion of Theorem 3.11 implies that the bootstrap approximation is more accurate than the normal approximation for the standardized variable. This property of the bootstrap also has been proved in some other situations. *It is the discovery of this property that makes the bootstrap more attractive, since it not only estimates the sampling distribution of a given statistic, but also provides a better solution*

than the traditional normal approximation. Abramovitch and Singh (1985) provided some more results on the higher order accuracy of the bootstrap distribution estimators. We will discuss other advantages of the bootstrap over the normal approximation in Sections 3.3.2–3.3.4.

The result in Theorem 3.11 can be extended to the case of functions of the sample mean. Let X_1, \dots, X_n be i.i.d. random p -vectors from a distribution F and g be a function from \mathbb{R}^p to \mathbb{R} . Consider the standardized random variable

$$\mathfrak{R}_n = \sqrt{n}[g(\bar{X}_n) - g(\mu)]/[\nabla g(\mu)' \Sigma \nabla g(\mu)]^{1/2},$$

where $\mu = EX_1$ and $\Sigma = \text{var}(X_1)$. The bootstrap analog of \mathfrak{R}_n is

$$\mathfrak{R}_n^* = \sqrt{n}[g(\bar{X}_n^*) - g(\bar{X}_n)]/[\nabla g(\bar{X}_n)' \hat{\Sigma} \nabla g(\bar{X}_n)]^{1/2},$$

where $\hat{\Sigma} = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)'$. Babu and Singh (1984b) showed that, if $E\|X_1\|^3 < \infty$, $|\int e^{it'x} dF(x)| < 1$ for any $t \neq 0$ ($i = \sqrt{-1}$), and g is twice continuously differentiable in a neighborhood of μ with $\nabla g(\mu) \neq 0$, then

$$\sqrt{n} \sup_x |P_*\{\mathfrak{R}_n^* \leq x\} - P\{\mathfrak{R}_n \leq x\}| \rightarrow_{a.s.} 0.$$

Datta (1992) discussed the convergence rate of bootstrap distribution estimators for \bar{X}_n in terms of the $\tilde{\rho}_r$ distance.

Studentized statistics

In the sample mean problem (Example 3.1), we are often interested in the bootstrap estimation of the distribution of a studentized statistic

$$t_n = \sqrt{n}(\bar{X}_n - \mu)/\hat{\sigma}. \quad (3.42)$$

Babu and Singh (1983) considered a more general problem: the bootstrap approximation for studentized linear combinations of sample means.

Let $\{Y_{11}, \dots, Y_{1n_1}\}, \{Y_{21}, \dots, Y_{2n_2}\}, \dots, \{Y_{k1}, \dots, Y_{kn_k}\}$ be k i.i.d. samples taken independently from G_1, G_2, \dots, G_k , respectively. Define $\mu_j = EY_{j1}$ and $\theta = \sum_{j=1}^k l_j \mu_j$, where l_1, \dots, l_k are known. Let $\bar{Y}_j = n_j^{-1} \sum_{i=1}^{n_j} Y_{ji}$ be the sample mean of the j th sample. As an estimator of θ , $\sum_{j=1}^k l_j \bar{Y}_j$ has variance $\sum_{j=1}^k l_j^2 \sigma_j^2/n_j$, where $\sigma_j^2 = \text{var}(Y_{j1})$. Now we want to estimate the distribution of the studentized statistic

$$t_{kn} = \left(\sum_{j=1}^k l_j \bar{Y}_j - \theta \right) / \left(\sum_{j=1}^k l_j^2 \hat{\sigma}_j^2/n_j \right)^{1/2},$$

where $\hat{\sigma}_j^2 = n_j^{-1} \sum_{i=1}^{n_j} (Y_{ji} - \bar{Y}_j)^2$. If $k = 1$ and $l_1 = 1$, then t_{kn} reduces to t_n in (3.42).

Let $\{Y_{j1}^*, \dots, Y_{jn_j}^*\}$ be an i.i.d. sample from the empirical distribution of the j th sample $\{Y_{j1}, \dots, Y_{jn_j}\}$, $j = 1, \dots, k$, and assume that the bootstrap data are generated independently. Define $\bar{Y}_j^* = n_j^{-1} \sum_{i=1}^{n_j} Y_{ji}^*$ and $\hat{\sigma}_j^{*2} = n_j^{-1} \sum_{i=1}^{n_j} (Y_{ji}^* - \bar{Y}_j^*)^2$. Then the bootstrap analog of t_{kn} is

$$t_{kn}^* = \left(\sum_{j=1}^k l_j \bar{Y}_j^* - \sum_{j=1}^k l_j \bar{Y}_j \right) / \left(\sum_{j=1}^k l_j^2 \hat{\sigma}_j^{*2} / n_j \right)^{1/2}.$$

Theorem 3.12. Suppose that G_j has a finite 6th moment for all $j = 1, \dots, k$ and that for at least one j , G_j is continuous. Assume further that $n/n_j \leq c < \infty$ for all j , where $n = \sum_{j=1}^k n_j$. Then

$$\sqrt{n} \sup_x |P_*\{t_{kn}^* \leq x\} - P\{t_{kn} \leq x\}| \rightarrow_{a.s.} 0.$$

The proof of this theorem can be found in Babu and Singh (1983). This result can be extended to the case where $t_{kn} = [g(\bar{X}_n) - g(\mu)]/h(\bar{X}_n)$, and g and h are smooth functions from \mathbb{R}^p to \mathbb{R} . We will return to study this kind of statistics in Section 3.3.3.

Lahiri (1993a) discussed the convergence rate of the bootstrap distribution estimator for t_n when X_1 is lattice.

U-statistics

Consider the U-statistic given by (3.22) with $m = 2$. Shi (1987) showed that under the condition that $E|h(X_1, X_2)|^3 < \infty$ and $E|h(X_1, X_1)| < \infty$, the convergence rate of the bootstrap estimator of the distribution of $\sqrt{n}(U_n - EU_n)/\sigma_F$ is $O(n^{-1/2})$ a.s., where $\sigma_F^2 = E[\tilde{h}(X_1)]^2$ and $\tilde{h}(x) = 2E[h(x, X_1) - EU_n]$. For a studentized U-statistic $\sqrt{n}(U_n - EU_n)/\hat{\sigma}_F$, where

$$\hat{\sigma}_F^2 = \frac{4(n-1)}{(n-2)^2} \sum_{i=1}^n \left[\frac{1}{n-1} \sum_{j=1}^n h(X_i, X_j) - U_n \right]^2,$$

Helmers (1991) showed that the convergence rate of the bootstrap estimator of the distribution of $\sqrt{n}(U_n - EU_n)/\hat{\sigma}_F$ is $o(n^{-1/2})$ a.s., provided that $E|h(X_1, X_2)|^{4+\epsilon} < \infty$ for some $\epsilon > 0$, $E|h(X_1, X_1)|^3 < \infty$, and the distribution of $\tilde{h}(X_1)$ is nonlattice.

Sample quantiles

It is difficult to discuss the accuracy of bootstrap estimators for the distributions of general L-statistics. But there are many results for sample quantiles. Singh (1981) first proved the following result. Let F_n^* be the bootstrap analog of the empirical distribution F_n , H_n and H_{BOOT} be

given by (3.1) and (3.2), respectively, with $\Re_n = \sqrt{n}[F_n^{-1}(t) - F^{-1}(t)]$ and $\Re_n^* = \sqrt{n}[F_n^{*-1}(t) - F_n^{-1}(t)]$, $0 < t < 1$.

Theorem 3.13. *Suppose that F has a bounded second order derivative in a neighborhood of $\theta = F^{-1}(t)$ and $f(\theta) > 0$, where $f(x) = dF/dx$. Then there is a constant c_F such that*

$$\limsup_{n \rightarrow \infty} \frac{n^{1/4} \rho_\infty(H_{\text{BOOT}}, H_n)}{\sqrt{\log \log n}} = c_F \quad a.s.$$

By the Berry-Esséen inequality (e.g., Serfling, 1980, p. 81),

$$\rho_\infty(H_n, \Phi_{\sigma_F}) = O(n^{-1/2}),$$

where $\sigma_F = \sqrt{t(1-t)/f(\theta)}$ and Φ_a is defined in (3.12). If we have an estimator $\hat{\sigma}_F$ of σ_F and use $\Phi_{\hat{\sigma}_F}$ to estimate H_n , then whether the bootstrap estimator H_{BOOT} is better than $\Phi_{\hat{\sigma}_F}$ depends on the convergence rate of $\hat{\sigma}_F$ to σ_F . Also, it may be of interest to study the accuracy of the bootstrap distribution estimator for the studentized statistic $\sqrt{n}[F_n^{-1}(t) - \theta]/\hat{\sigma}_F$. But it involves Edgeworth expansions for sample quantiles and their bootstrap versions, and, unfortunately, there are no results available for this problem.

Falk and Reiss (1989) studied the weak convergence of the process

$$P_*\{\sqrt{n}[F_n^{*-1}(t) - F_n^{-1}(t)] \leq x\} - P\{\sqrt{n}[F_n^{-1}(t) - F^{-1}(t)] \leq x\}$$

and showed that the bootstrap estimator has a convergence rate $O_p(n^{-1/4})$. Falk (1988) used the same technique to study the accuracy of the bootstrap estimator of the sampling distribution of the statistic $F_n^{-1}(1 - k_n/n)$. A survey of these results was given by Falk (1992a).

Discussions

It is expected that the convergence rate of the bootstrap estimator of the distribution of a standardized or studentized statistic is $o(n^{-1/2})$ [or $O(n^{-1})$]. This is why the bootstrap is said to be second order accurate. [An estimator is said to be k th order accurate if its convergence rate is $O(n^{-k/2})$.] The proof required for this type of result is technically difficult since Edgeworth expansions are involved, and, consequently, results are currently available for only simple statistics, some of which have been discussed in this section. The investigation of the accuracy of the bootstrap estimators is still an active field, and a complete theory is expected to be established in the future. For example, Lahiri (1992a) discussed the convergence rate of the bootstrap distribution estimators for M-estimators, and Lai and Wang (1993) studied Edgeworth expansions for symmetric statistics with censored data.

Theorem 3.11 shows that the bootstrap estimator has a convergence rate $o(n^{-1/2})$ if F is nonlattice and $E|X_1|^3 < \infty$. Hall (1988a) proved that $E|X_1|^3 < \infty$ is also a necessary condition. If $E|X_1|^3 = \infty$, the convergence rate of the bootstrap estimator may be slower than $O(n^{-1/2})$, which is the convergence rate of the normal approximation. This is parallel to Theorem 3.2, which states that the bootstrap estimator is consistent if and only if the second moment of X_1 is finite.

3.3.2 Asymptotic minimaxity

We now consider the accuracy of the bootstrap estimator and make some comparisons by using some other accuracy measures.

Let $T_n = T_n(X_1, \dots, X_n)$ be an estimator of an unknown parameter θ , $\mathfrak{R}_n = \sqrt{n}(T_n - \theta)$, H_n be the distribution of \mathfrak{R}_n , and \mathcal{H} be a collection of estimators of H_n . For any $\hat{H} \in \mathcal{H}$, define

$$\Gamma_{n,c}^{(a)}(\hat{H}) = \sup_{G \in B_F(n,c)} E_G L(\sqrt{n}\rho_\infty(\hat{H}^{(a)}, H_n^{(a)})), \quad (3.43)$$

where L is an increasing loss function defined on $[0, \infty)$, E_G is the expectation under the assumption that X_1 is distributed as G ,

$$B_F(n, c) = \{G : \rho_\infty(G, F) \leq cn^{-1/2}\},$$

and, for any function $h(x)$ on \mathbb{R} , $h^{(a)}(x)$ is defined to be

$$h^{(a)}(x) = \frac{1}{a} \int h(x-y) \left(1 - \frac{|y|}{a}\right) I\{|y| \leq a\} dy.$$

Note that $\Gamma_{n,c}^{(a)}(\hat{H})$ is the maximum risk of \hat{H} as an estimator of H_n over a neighborhood of F . The reason why we use $\hat{H}^{(a)}$ and $H_n^{(a)}$ in (3.43), instead of \hat{H} and H_n , will be explained later.

If an estimator $\hat{H}_{\text{MIN}} \in \mathcal{H}$ satisfies

$$\lim_{n \rightarrow \infty} \Gamma_{n,c}^{(a)}(\hat{H}_{\text{MIN}}) = \lim_{c \rightarrow \infty} \liminf_{n \rightarrow \infty} \inf_{\hat{H} \in \mathcal{H}} \Gamma_{n,c}^{(a)}(\hat{H}),$$

then \hat{H}_{MIN} is said to be asymptotically minimax, since it minimizes the maximum risk over a neighborhood. Asymptotic minimaxity was used by many authors to assess parametric and semiparametric procedures. The reader is referred to Ibragimov and Has'minskii (1981) for a full discussion on this topic.

Under some regularity conditions, Beran (1984b) showed that

$$\lim_{c \rightarrow \infty} \liminf_{n \rightarrow \infty} \inf_{\hat{H} \in \mathcal{H}} \Gamma_{n,c}^{(a)}(\hat{H}) \geq E_F L(a_F |Z|),$$

where a_F is a positive constant and Z is the standard normal random variable, and showed that if L is bounded, then the bootstrap estimator $H_{\text{BOOT}}(x) = P_*\{\sqrt{n}(T_n^* - T_n) \leq x\}$ satisfies

$$\lim_{n \rightarrow \infty} \Gamma_{n,c}^{(a)}(H_{\text{BOOT}}) = E_F L(a_F |Z|) \quad (3.44)$$

for every $c > 0$. Hence, H_{BOOT} is asymptotically minimax.

The traditional normal approximation provides an estimator of the form $H_{\text{NOR}}(x) = \Phi(x/\sigma(F_n))$, where $\sigma(G)$ is a functional and $\sigma(F_n)$ estimates the standard deviation of $\sqrt{n}(T_n - \theta)$. This estimator, however, satisfies

$$\lim_{n \rightarrow \infty} \Gamma_{n,c}^{(a)}(H_{\text{NOR}}) = E_F L\left(\sup_x |a_F Z + h_F^{(a)}(x)|\right),$$

where

$$h_F(x) = [\kappa(F)(x^2 - 1)/6 + b(F)/\sigma(F)]\varphi(x), \quad (3.45)$$

$\varphi(x) = d\Phi/dx$, and $\kappa(G)$ and $b(G)$ are some functionals. Hence, H_{NOR} is asymptotically minimax if and only if $\kappa(F) = 0$ and $b(F) = 0$, which essentially requires that T_n has 0 skewness and a bias of the order $o(n^{-1})$. This shows the superiority of the bootstrap to the traditional normal approximation, although their convergence rates are the same (see Section 3.3.1).

Using a one-term Edgeworth expansion, we can obtain the following estimator which improves H_{NOR} :

$$H_{\text{EDG}}(x) = \Phi(x/\sigma(F_n)) - h_{F_n}(x/\sigma(F_n))/\sqrt{n}, \quad (3.46)$$

where h_F is given in (3.45). Beran (1984b) showed that under the same conditions used in establishing (3.44),

$$\lim_{n \rightarrow \infty} \Gamma_{n,c}^{(a)}(H_{\text{EDG}}) = E_F L(a_F |Z|). \quad (3.47)$$

Hence, H_{EDG} is also asymptotically minimax, and, in terms of the asymptotic minimaxity, the bootstrap estimator is essentially equivalent to the one-term Edgeworth expansion estimator. The latter, however, requires the derivations of the functionals $b(F)$, $\sigma(F)$, and $\kappa(F)$, which are not easy. The bootstrap replaces these derivations by computations.

The regularity conditions used to establish (3.44) and (3.47) are somewhat strong: it is required that the studentized variable $\sqrt{n}(T_n - \theta)/\sigma(F_n)$ admits the one-term Edgeworth expansion uniformly over $B_F(n, c)$. Beran (1984b) showed that the sample mean, the U-statistics, and some differentiable functional statistics satisfy these regularity conditions.

The reason for using $\hat{H}^{(a)}$ and $H_F^{(a)}$ in (3.43) is that they ensure that $\sqrt{n}(T_n - \theta)/\sigma(F_n)$ admits the Edgeworth expansion. If we assume that F

is nonlattice and restrict $B_F(n, c)$ to \mathcal{F}_0 , the set of nonlattice distributions, then we may replace $\Gamma_{n,c}^{(a)}(\hat{H})$ in (3.43) by

$$\Gamma_{n,c}(\hat{H}) = \sup_{G \in B_F(n, c) \cap \mathcal{F}_0} E_G L(\sqrt{n}\rho_\infty(\hat{H}, H_n)).$$

Singh and Babu (1990) proved that results (3.44) and (3.47), with $\Gamma_{n,c}^{(a)}$ replaced by $\Gamma_{n,c}$, hold for the sample mean, U-statistics, and some studentized statistics under some regularity conditions.

3.3.3 Asymptotic mean squared error

The results in Section 3.3.2 are established for bounded loss functions. The simplest unbounded loss function is the quadratic function $L(x) = x^2$, and the corresponding risk of δ_n as an estimator of δ is called the mean squared error: $E(\delta_n - \delta)^2$. However, this requires the existence of the finite second moment of δ_n . Without assuming the existence of the second moment of δ_n , we define the *asymptotic mean squared error* (amse) of δ_n to be

$$\text{amse}(\delta_n) = (\sigma_F^2 + b_F^2)/n^{2\ell},$$

provided that

$$n^\ell(\delta_n - \delta) \rightarrow_d W,$$

where W is a random variable having mean b_F and variance σ_F^2 . The amse is an accuracy measure for δ_n .

Liu and Singh (1987) and Bhattacharya and Qumsiyeh (1989) compared the amse of $H_{\text{BOOT}}(x)$, $H_{\text{NOR}}(x)$, and $H_{\text{EDG}}(x)$ defined in Section 3.3.2 for a fixed x . Suppose that the distribution of $\mathfrak{R}_n = \sqrt{n}(T_n - \theta)$ has the Edgeworth expansion of the form

$$H_n(x) = \Phi\left(\frac{x}{\sigma(F)}\right) + \frac{p(x, F)}{\sqrt{n}}\varphi\left(\frac{x}{\sigma(F)}\right) + o\left(\frac{1}{\sqrt{n}}\right), \quad (3.48)$$

where $\sigma(G)$ is a functional and $p(x, F)$ is a polynomial of x . Suppose further that the conditional distribution of $\sqrt{n}(T_n^* - T_n)$ has a similar expansion:

$$H_{\text{BOOT}}(x) = \Phi\left(\frac{x}{\sigma(F_n)}\right) + \frac{p(x, F_n)}{\sqrt{n}}\varphi\left(\frac{x}{\sigma(F_n)}\right) + o\left(\frac{1}{\sqrt{n}}\right) \text{ a.s.} \quad (3.49)$$

Comparing (3.48) with (3.49), we obtain that

$$\begin{aligned} \sqrt{n}[H_{\text{BOOT}}(x) - H_n(x)] &= \sqrt{n}\left[\Phi\left(\frac{x}{\sigma(F_n)}\right) - \Phi\left(\frac{x}{\sigma(F)}\right)\right] \\ &\quad + \left[p(x, F_n)\varphi\left(\frac{x}{\sigma(F_n)}\right) - p(x, F)\varphi\left(\frac{x}{\sigma(F)}\right)\right] + o(1) \text{ a.s.} \end{aligned} \quad (3.50)$$

Suppose that

$$\sqrt{n}[\sigma(F_n) - \sigma(F)] \rightarrow_d N(0, \vartheta) \quad (3.51)$$

for some $\vartheta > 0$. Then, by (3.50) and (3.51),

$$\sqrt{n}[H_{\text{BOOT}}(x) - H_n(x)] \rightarrow_d N(0, a_x^2 \varphi_x^2), \quad (3.52)$$

where $a_x^2 = \vartheta x^2 / [4\sigma^6(F)]$ and $\varphi_x = \varphi(x/\sigma(F))$. For H_{NOR} , by (3.48) and (3.51),

$$\begin{aligned} \sqrt{n}[H_{\text{NOR}}(x) - H_n(x)] &= \sqrt{n} \left[\Phi\left(\frac{x}{\sigma(F_n)}\right) - \Phi\left(\frac{x}{\sigma(F)}\right) \right] \\ &\quad - p(x, F)\varphi_x + o(1) \rightarrow_d N(-p(x, F)\varphi_x, a_x^2 \varphi_x^2). \end{aligned} \quad (3.53)$$

From (3.52) and (3.53), we conclude that

$$\lim_{n \rightarrow \infty} \frac{\text{amse}(H_{\text{BOOT}}(x))}{\text{amse}(H_{\text{NOR}}(x))} = \frac{a_x^2}{a_x^2 + [p(x, F)]^2},$$

which is less than 1 if $p(x, F) \neq 0$. Thus, in terms of the amse, the bootstrap is better than the normal approximation for all x with $p(x, F) \neq 0$, although they have the same convergence rate (Theorem 3.11).

In the special case where $T_n = \bar{X}_n$ and $\theta = EX_1$, (3.48) and (3.49) hold with $p(x, F) = -\kappa(F)(x^2 - 1)/6$, where $\kappa(F)$ is the skewness of X_1 , provided that $E|X_1|^3 < \infty$ and X_1 is nonlattice. Thus, when $\kappa(F) \neq 0$, the bootstrap estimator has a smaller amse than the normal approximation, except at points $x = \pm 1$. This phenomenon was termed by Liu and Singh (1987) as “a partial correction by the bootstrap”, while the “total correction by the bootstrap” means that the bootstrap estimator has the accuracy of $o(n^{-1/2})$, which is the case for the bootstrap estimator of the distribution of a standardized or a studentized statistic, which will be discussed later in this section. Conditions (3.48) and (3.49) also hold when T_n is a smooth function of the sample mean.

The one-term Edgeworth expansion estimator is of the form

$$H_{\text{EDG}}(x) = \Phi(x/\sigma(F_n)) + p(x, F_n)\varphi(x/\sigma(F_n))/\sqrt{n}$$

[see (3.46)]. From (3.49) and (3.52), we conclude that

$$\lim_{n \rightarrow \infty} \frac{\text{amse}(H_{\text{BOOT}}(x))}{\text{amse}(H_{\text{EDG}}(x))} = 1.$$

That is, for estimating the distribution of $\sqrt{n}(T_n - \theta)$, the bootstrap and the one-term Edgeworth expansion are equivalent in terms of the amse.

However, for estimating the distribution of studentized statistics, the bootstrap may be even better than the one-term Edgeworth expansion. Consider the problem of estimating the distribution G_n of the studentized statistic $\sqrt{n}(T_n - \theta)/\sigma(F_n)$, where $T_n = g(\bar{X}_n)$, \bar{X}_n and F_n are the sample mean and the empirical distribution of i.i.d. random p -vectors X_1, \dots, X_n , $\theta = g(\mu)$, $\mu = EX_1$, and $\sigma(F_n) = \{n^{-1} \sum_{i=1}^n [\nabla g(\bar{X}_n)'(X_i - \bar{X}_n)]^2\}^{1/2}$. Assume that g is four times continuously differentiable in a neighborhood of μ , $E\|X_1\|^8 < \infty$, and that X_1 satisfies Cramér's condition, i.e., the characteristic function $\psi(t)$ of X_1 satisfies

$$\limsup_{\|t\| \rightarrow \infty} |\psi(t)| < 1. \quad (3.54)$$

Then the distribution G_n admits a two-term Edgeworth expansion:

$$G_n(x) = \Phi(x) + \left[\frac{q_1(x, F)}{\sqrt{n}} + \frac{q_2(x, F)}{n} \right] \varphi(x) + o\left(\frac{1}{n}\right), \quad (3.55)$$

where $q_k(x, F)$ are polynomials of x (see Appendix A.10 and Bhattacharya and Ghosh, 1978). The one-term Edgeworth expansion estimator is

$$G_{\text{EDG}}(x) = \Phi(x) + q_1(x, F_n)\varphi(x)/\sqrt{n}.$$

Under some moment conditions,

$$\sqrt{n}[q_1(x, F_n) - q_1(x, F)] \rightarrow_d N(0, b_x^2) \quad (3.56)$$

for some $b_x^2 > 0$. From (3.55) and (3.56),

$$n[G_{\text{EDG}}(x) - G_n(x)] \rightarrow_d N(-q_2(x, F)\varphi(x), b_x^2\varphi^2(x)). \quad (3.57)$$

The bootstrap estimator $G_{\text{BOOT}}(x)$ is the conditional distribution of the bootstrap studentized statistic $\sqrt{n}(T_n^* - T_n)/\sigma(F_n^*)$ and satisfies

$$G_{\text{BOOT}}(x) = \Phi(x) + \left[\frac{q_1(x, F_n)}{\sqrt{n}} + \frac{q_2(x, F_n)}{n} \right] \varphi(x) + o\left(\frac{1}{n}\right) \text{ a.s.} \quad (3.58)$$

under the same conditions for the validity of (3.55) (Babu and Singh, 1984b). From (3.55), (3.56), (3.58), and $q_2(x, F_n) - q_2(x, F) \rightarrow_{a.s.} 0$, we obtain that

$$n[G_{\text{BOOT}}(x) - G_n(x)] \rightarrow_d N(0, b_x^2\varphi^2(x)),$$

which, together with (3.57), implies that

$$\lim_{n \rightarrow \infty} \frac{\text{amse}(G_{\text{BOOT}}(x))}{\text{amse}(G_{\text{EDG}}(x))} = \frac{b_x^2}{b_x^2 + [q_2(x, F)]^2}.$$

This shows that the bootstrap is better than the one-term Edgeworth expansion for all x with $q_2(x, F) \neq 0$, although it can be shown that they have the same convergence rate.

The results in this section can be extended to the case where the loss function is bowl-shaped and bounded (Bhattacharya and Qumsiyeh, 1989).

3.3.4 Asymptotic relative error

Let δ_n and $\tilde{\delta}_n$ be two estimators of δ . We can compare their performances by considering the ratio of their errors in estimating δ :

$$r = \lim_{n \rightarrow \infty} \frac{|\tilde{\delta}_n - \delta|}{|\delta_n - \delta|}.$$

Using this criterion, Hall (1988a, 1990b) compared the bootstrap, the normal approximation, and the one-term Edgeworth expansion estimators of the distribution \tilde{H}_n of the standardized variable $\sqrt{n}(\bar{X}_n - \mu)/\sigma$, where \bar{X}_n is the sample mean of i.i.d. random variables X_1, \dots, X_n , $\mu = EX_1$, and $\sigma^2 = \text{var}(X_1)$. Results for estimating the distribution of studentized statistics can be similarly established.

If $E|X_1|^3 < \infty$ and the distribution of X_1 is nonlattice, then \tilde{H}_n admits the one-term Edgeworth expansion:

$$\tilde{H}_n(x) = \Phi(x) + \frac{\tilde{q}_1(x, F)}{\sqrt{n}} \varphi(x) + o\left(\frac{1}{\sqrt{n}}\right),$$

where $\tilde{q}_1(x, F) = E(X_1 - EX_1)^3(1 - x^2)/6\sigma^3$, and, under the same conditions, the bootstrap estimator $\tilde{H}_{\text{BOOT}}(x) = P_*\{\sqrt{n}(\bar{X}_n^* - \bar{X}_n)/\hat{\sigma} \leq x\}$ has the same expansion for almost all X_1, X_2, \dots (see the discussion in Section 3.3.1). Note that in this case the normal approximation estimator of \tilde{H}_n is $\tilde{H}_{\text{NOR}} = \Phi$. Hence,

$$\begin{aligned} r(x) &= \lim_{n \rightarrow \infty} \frac{|\tilde{H}_{\text{BOOT}}(x) - \tilde{H}_n(x)|}{|\tilde{H}_{\text{NOR}}(x) - \tilde{H}_n(x)|} \\ &= \lim_{n \rightarrow \infty} \frac{o(1)}{\tilde{q}_1(x, F)\varphi(x) + o(1)} = 0 \quad a.s. \end{aligned} \tag{3.59}$$

for all x such that $\tilde{q}_1(x, F) \neq 0$. Thus, the bootstrap is better than the normal approximation. In fact, we have drawn this conclusion in Section 3.3.1.

When $E|X_1|^3 = \infty$, Hall (1988a) showed that $r(x)$ in (3.59) equals 1 a.s., if $P\{|X_1| > x\} = x^{-\alpha}K(x)$, where $2 \leq \alpha < 3$ and the function K is slowly varying at infinity. That is, the bootstrap is equivalent to the normal approximation. If $P\{|X_1| > x\} = x^{-3}(\log x)^\epsilon$ for some $\epsilon > 0$, then $r(x) = \infty$ a.s., i.e., the bootstrap is worse than the normal approximation.

In summary, whether the bootstrap is better than the normal approximation depends on the tail condition of the distribution of X_1 . The bootstrap is better when $E|X_1|^3 < \infty$.

The one-term Edgeworth expansion estimator of $\tilde{H}_n(x)$ is

$$\tilde{H}_{\text{EDG}}(x) = \Phi(x) + \tilde{q}_1(x, F_n)\varphi(x)/\sqrt{n}.$$

Under some conditions (see the discussion in previous sections),

$$\tilde{H}_n(x) = \Phi(x) + \left[\frac{\tilde{q}_1(x, F)}{\sqrt{n}} + \frac{\tilde{q}_2(x, F)}{n} \right] \varphi(x) + o\left(\frac{1}{n}\right)$$

and

$$\tilde{H}_{\text{BOOT}}(x) = \Phi(x) + \left[\frac{\tilde{q}_1(x, F_n)}{\sqrt{n}} + \frac{\tilde{q}_2(x, F_n)}{n} \right] \varphi(x) + o\left(\frac{1}{n}\right) \text{ a.s.}$$

Hence,

$$r_n(x) = \frac{|\tilde{H}_{\text{BOOT}}(x) - \tilde{H}_n(x)|}{|\tilde{H}_{\text{EDG}}(x) - \tilde{H}_n(x)|} \rightarrow_{a.s.} 1$$

if $\tilde{q}_2(x, F) = 0$ and otherwise

$$r_n(x) \rightarrow_d |1 + c_x Z^{-1}| \text{ a.s.,}$$

where Z is a standard normal random variable and c_x^2 is the asymptotic variance of $\sqrt{n}[\tilde{q}_1(x, F_n) - \tilde{q}_1(x, F)]/\tilde{q}_2(x, F)$ (Hall, 1988a). Since Z is random, it is difficult to say which estimator is better using the criterion of the ratio of estimation errors. In view of $P\{|1 + c_x Z^{-1}| < 1\} > 0.5$, we can say that the bootstrap is better than the one-term Edgeworth expansion in terms of frequency (on the average). This is just a variation of the result established in Section 3.3.3.

A further comparison between the bootstrap and the one-term Edgeworth expansion can be made if we allow the x in $r_n(x)$ to vary with n , i.e., we consider $r_n(x_n)$ for a sequence $\{x_n\}$ satisfying $x_n \rightarrow \infty$. This also means that we want to see which estimator is better in estimating the tail probabilities of the sampling distribution of the given statistic.

Using the large deviation formula for tail probability (see, e.g., Petrov, 1975), Hall (1990b) studied the asymptotic behavior of $r_n(x_n)$ with x_n increasing to infinity in various orders. Suppose that (3.54) holds, $\alpha_3 = E(X_1 - \mu)^3 \neq 0$, and that X_1 has a finite moment generating function in a neighborhood of μ . Let $\{x_n\}$ be a sequence of positive numbers. Then we have the following conclusions:

- (1) If $n^{-1/3}x_n \rightarrow 0$, then $r_n(x_n) \rightarrow_p 0$. In this case, the bootstrap is better than the one-term Edgeworth expansion.
- (2) Suppose that $n^{-1/3}x_n \rightarrow c$ and $0 < c < \infty$. If $\alpha_3 > 0$, then $r_n(x_n) \rightarrow_p 0$ and the bootstrap is better than the one-term Edgeworth expansion. If $\alpha_3 < 0$, then $r_n(x_n)$ has a nondegenerate limiting distribution and the limiting probability that the bootstrap is better is larger than half.

- (3) Suppose that $n^{-1/3}x_n \rightarrow \infty$ and $n^{-1/2}x_n \rightarrow 0$. If $\alpha_3 < 0$, then $r_n(x_n) \rightarrow_p 0$ and the bootstrap is better. If $\alpha_3 > 0$, then $r_n(x_n)$ has a nondegenerate limiting distribution. The limiting probability that the one-term Edgeworth expansion is better is equal to half, and the limiting probability that the bootstrap and the one-term Edgeworth expansion are equivalent is also half.

3.3.5 Conclusions

- (1) For estimating the distribution of $\sqrt{n}(\bar{X}_n - \mu)$, the bootstrap estimator has the same convergence rate as the estimator based on the normal approximation, but the former is asymptotically minimax and has a smaller asymptotic mean squared error than the latter. The bootstrap estimator is equivalent to the one-term Edgeworth expansion estimator in terms of the asymptotic minimax and mean squared error criteria.
- (2) For the standardized variable $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ or the studentized variable $\sqrt{n}(\bar{X}_n - \mu)/\hat{\sigma}$, the bootstrap distribution estimator has a faster convergence rate than the estimator based on the normal approximation. The bootstrap estimator is equivalent to the one-term Edgeworth expansion estimator in terms of the convergence rate and asymptotic minimaxity, but is better than the latter in terms of the asymptotic mean squared error.
- (3) The assertions in (1) and (2) may still hold in the case where \bar{X}_n is replaced by a general statistic T_n , for example, $T_n = g(\bar{X}_n)$ with a smooth function g . The theoretical justification for a general T_n , however, is not available, due to the technical difficulties involved in deriving and proving the Edgeworth expansion for the distribution of T_n .
- (4) Whether the bootstrap is better than the Edgeworth expansion method in terms of estimating a tail probability depends on how far we want to go in the direction of the extreme of the tail. Except for very extreme tail probabilities, the bootstrap estimator is better or equivalent to the Edgeworth expansion estimator.

3.4 Fixed Sample Performance

We have discussed the asymptotic properties of the bootstrap in the previous sections. Fixed sample (especially small sample) properties of the bootstrap are also important. Unfortunately, the bootstrap estimators are usually complicated, so that we can only assess their fixed sample properties by empirical simulations carried out under some special circumstances. In the following, we examine by simulation the performance of the bootstrap estimators and compare them with other estimators. Here we only

consider the i.i.d. case and the simple nonparametric bootstrap. Empirical simulation results for other cases will be given later.

Throughout this section, ME denotes the simulation average of an estimator $\hat{\delta}$ of a quantity δ (which is usually known in the simulation study), SD denotes the simulation standard deviation of $\hat{\delta}$, CV denotes the simulation coefficient of variation defined to be the ratio of SD over the absolute value of ME, RB denotes the simulation relative bias defined to be the ratio of ME– δ over the absolute value of δ , and RM denotes the simulation root mean squared error $[SD^2 + (ME - \delta)^2]^{1/2}$. The SD and CV are simulation estimates of the stability of $\hat{\delta}$. The RB and RM measure the actual loss of using $\hat{\delta}$ as an estimate of δ .

3.4.1 Moment estimators

We first consider estimators of some simple functions of the moments of a given statistic T_n , namely, the variance or the standard deviation of T_n , the bias of T_n , and the skewness of T_n [defined to be $E(T_n - ET_n)^3 / [\text{var}(T_n)]^{3/2}$]. In the simulation study, the bootstrap variance, bias, and skewness estimators are respectively approximated by the following Monte Carlo estimates (see Section 1.4 or Chapter 5):

$$v_{\text{BOOT}}^{(B)} = \frac{1}{B} \sum_{b=1}^B (T_{n,b}^* - \bar{T}_n^*)^2, \quad b_{\text{BOOT}}^{(B)} = \bar{T}_n^* - T_n,$$

and

$$sk_{\text{BOOT}}^{(B)} = \frac{1}{B} \sum_{b=1}^B (T_{n,b}^* - \bar{T}_n^*)^3 / [v_{\text{BOOT}}^{(B)}]^{3/2},$$

where $\bar{T}_n^* = B^{-1} \sum_{b=1}^B T_{n,b}^*$.

Parr (1983) compared two bootstrap standard deviation estimators $s_{\text{BOOT}}^{(B_k)} = (v_{\text{BOOT}}^{(B_k)})^{1/2}$, $k = 1, 2$, with the jackknife estimator $s_{\text{JACK}} = v_{\text{JACK}}^{1/2}$ and the normal theory estimator $s_{\text{NOR}} = (1 - \hat{\rho}_n^2) / \sqrt{n - 3}$, for $T_n = \hat{\rho}_n$, the sample correlation coefficient (Example 2.1), based on i.i.d. random bivariate normal vectors with 0 means, unit variances, and correlation coefficient ρ . The results in Table 3.1 are based on 100 simulation replicates. As is expected, s_{NOR} has the best performance, since it is derived under the assumption that the data are from a normal distribution. Both bootstrap estimators are downward-biased. $s_{\text{BOOT}}^{(B_1)}$ and s_{JACK} require almost the same amount of computation ($B_1 \approx n$), but, overall, the jackknife estimator is better. With a larger Monte Carlo size B_2 , the bootstrap estimator $s_{\text{BOOT}}^{(B_2)}$ is less variable than the jackknife in the cases of $\rho = 0$ and $\rho = 0.5$, but performs the same as the jackknife estimator when $\rho = 0.9$.

Table 3.1. Comparison of the bootstrap, jackknife, and normal theory estimators for the SD of the sample correlation coefficient
[Adapted from Parr (1983), by permission of Biometrika]

n	ρ	True SD	$s_{\text{BOOT}}^{(B_1)}$ †		$s_{\text{BOOT}}^{(B_2)}$ †		s_{JACK}		s_{NOR}	
			ME	CV	ME	CV	ME	CV	ME	CV
8	0	.373	.319	.445	.333	.366	.395	.420	.376	.234
	.5	.313	.244	.582	.269	.520	.316	.608	.285	.407
	.9	.114	.098	1.25	.116	1.07	.095	1.01	.082	.732
14	0	.277	.238	.277	.253	.217	.284	.289	.281	.103
	.5	.222	.189	.392	.206	.340	.225	.400	.217	.244
	.9	.067	.059	.746	.066	.621	.063	.651	.059	.559
20	0	.230	.206	.228	.217	.175	.239	.201	.232	.056
	.5	.182	.157	.293	.160	.250	.172	.267	.175	.211
	.9	.053	.044	.500	.046	.457	.046	.435	.046	.391

† $B_1 = 8$, $B_2 = 20$ when $n = 8$; $B_1 = 20$, $B_2 = 100$ when $n = 14$ or 20.

Tu and Zhang (1992a) investigated the jackknife and bootstrap variance estimators v_{JACK} and $v_{\text{BOOT}}^{(B)}$ with $B = 500$ in the case where T_n is Gini's mean difference

$$T_n = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n |X_i - X_j|$$

and X_1, \dots, X_n are i.i.d. from $N(0, 100)$. The results, based on 1000 simulation replicates, are given in Table 3.2. The bootstrap estimator, requiring much more computations than the jackknife estimator in this case, is less variable but downward-biased. Furthermore, the improvement (in terms of the CV or RM) of the bootstrap over the jackknife becomes inappreciable when n is large.

Bootstrap and jackknife estimators of the bias and skewness of T_n were also studied in Tu and Zhang (1992a) and the results are listed in Table 3.2. For the bias estimation, the jackknife estimator b_{JACK} is better than the bootstrap estimator $b_{\text{BOOT}}^{(B)}$. For the skewness estimation, the jackknife estimator is an adjusted estimator proposed by Tu and Zhang (1992a):

$$\begin{aligned} sk_{\text{JACK}} &= \frac{3(n-1)^2}{n^3 v_{\text{JACK}}^{3/2}} \sum_{i \neq j} (T_{n-1,i} - \bar{T}_n)(T_{n-1,j} - \bar{T}_n) \Delta_{ij} \\ &\quad - \frac{(n-1)^3}{n^3 v_{\text{JACK}}^{3/2}} \sum_{i=1}^n (T_{n-1,i} - \bar{T}_n)^3, \end{aligned} \tag{3.60}$$

where $T_{n-1,i}$ is the statistic T_n calculated after removing X_i , $\Delta_{ij} = nT_n - (n-1)(T_{n-1,i} + T_{n-1,j}) + (n-2)T_{n-2,ij}$, and $T_{n-2,ij}$ is T_n calculated after

removing X_i and X_j . However, both jackknife and bootstrap skewness estimators do not perform well unless n is very large. For sk_{JACK} , its RB decreases as n increases, and its performance is satisfactory when n reaches 60. It is not surprising that sk_{JACK} is better than $sk_{BOOT}^{(B)}$ since sk_{JACK} has been adjusted. The unsatisfactory performance of $sk_{BOOT}^{(B)}$ was also reported by Peters and Freedman (1987) in the case where T_n is the sample variance. A possible reason is that the empirical distribution F_n cannot capture the true skewness of the population well (Efron, 1992b).

Table 3.2. Comparison of the bootstrap and jackknife estimators for the variance, bias, and skewness of Gini's mean difference [Adapted from Tu and Zhang (1992a), by permission of Springer-Verlag]

Estimation of VAR (variance)									
n	True VAR	$v_{BOOT}^{(B)}$				v_{JACK}			
		ME	RB	CV	RM	ME	RB	CV	RM
10	5.96	4.74	-21%	.643	3.28	6.64	11%	.748	5.02
20	3.06	2.82	-8%	.486	1.39	3.28	7%	.512	1.69
30	2.12	1.97	-7%	.397	.797	2.17	2%	.399	.865
40	1.60	1.49	-7%	.346	.526	1.62	2%	.365	.591
60	1.07	1.02	-5%	.286	.297	1.09	1%	.283	.309
Estimation of bias									
n	True Bias	$b_{BOOT}^{(B)}$				b_{JACK}			
		ME	RB	CV	RM	ME	RB	CV	RM
10	-1.14	-1.02	11%	.258	.289	-1.13	1%	.245	.277
20	-.575	-.542	6%	.225	.126	-.563	2%	.164	.093
30	-.387	-.368	5%	.214	.081	-.375	3%	.138	.053
40	-.302	-.271	10%	.227	.069	-.281	7%	.117	.039
60	-.195	-.186	5%	.268	.051	-.188	4%	.094	.019
Estimation of SK (skewness)									
n	True SK	$sk_{BOOT}^{(B)}$				sk_{JACK}			
		ME	RB	CV	RM	ME	RB	CV	RM
10	.296	-2.72	-192%	1.03	.633	.068	-77%	5.69	.448
20	.187	-.049	-126%	4.00	.307	.089	-52%	2.29	.226
30	.163	-.003	-101%	56.0	.237	.091	-44%	1.62	.164
40	.149	.031	-80%	4.48	.182	.085	-43%	1.35	.130
60	.105	.041	-60%	3.29	.148	.077	-19%	1.07	.084

3.4.2 Distribution estimators

The distribution of $T_n - \theta$ can be estimated by using the bootstrap with the Monte Carlo approximation:

$$H_{\text{BOOT}}^{(B)}(x) = \frac{1}{B} \sum_{b=1}^B I\{T_{n,b}^* - T_n \leq x\},$$

where $T_{n,b}^*$ is T_n based on the b th bootstrap sample. The bootstrap histogram $\{T_{n,b}^* - T_n : b = 1, \dots, B\}$ can also be used to estimate the density of $T_n - \theta$. Figure 3.1 (obtained from Diaconis and Efron, 1983) presents the bootstrap histogram ($B = 1000$) and the normal theory density estimator in contrast with the true density of the sample correlation coefficient $\hat{\rho}_n$ (Example 2.1) based on the following $n = 15$ pairs of observations from 15 American law schools:

$$\begin{aligned} & (576, 3.39), (635, 3.30), (558, 2.81), (578, 3.03), (666, 3.44), \\ & (580, 3.07), (555, 3.00), (661, 3.43), (651, 3.36), (605, 3.13), \\ & (653, 3.12), (575, 2.74), (545, 2.76), (572, 2.88), (594, 2.96). \end{aligned}$$

The normal theory density estimator is given by (Tong, 1990)

$$\frac{2^{n-3}(1-\rho^2)^{(n-1)/2}}{\pi(n-3)!} (1-x^2)^{(n-4)/2} \sum_{j=0}^{\infty} \left[\Gamma\left(\frac{n+j-1}{2}\right) \right]^2 \frac{(2\rho x)^j}{j!}$$

with ρ estimated by $\hat{\rho}_n$. The bootstrap histogram nicely approximates the true density as well as the normal theory density estimate. The rather close agreement among the peaks of the densities is an artifact of the sample.

A simulation study was performed by Hinkley and Wei (1984) to compare the bootstrap estimator with the normal approximation and Edgeworth expansion estimators of the distribution of studentized ratio estimator:

$$G_n(x) = P\{(\hat{\gamma}_n - \gamma)/\hat{\sigma}_n \leq x\},$$

where $\hat{\gamma}_n = \bar{Y}_n/\bar{Z}_n$ for a random sample (Y_i, Z_i) , $i = 1, \dots, n = 50$, from a bivariate population, $\hat{\sigma}_n^2 = \sum_{i=1}^n (Y_i - \hat{\gamma}_n Z_i)^2 / (\sum_{i=1}^n Z_i)^2$ is the linearization variance estimator, and $\gamma = EY_1/EZ_1$. The bootstrap estimator is approximated by Monte Carlo

$$G_{\text{BOOT}}^{(B)}(x) = \frac{1}{B} \sum_{b=1}^B I\{(\hat{\gamma}_{n,b}^* - \hat{\gamma}_n)/\hat{\sigma}_{n,b}^* \leq x\}$$

with size $B = 1000$. The one-term and two-term Edgeworth expansion estimators are derived from the Edgeworth expansion (3.55) (see Hinkley and Wei, 1984) and are denoted by $G_{\text{EDG-1}}$ and $G_{\text{EDG-2}}$, respectively.

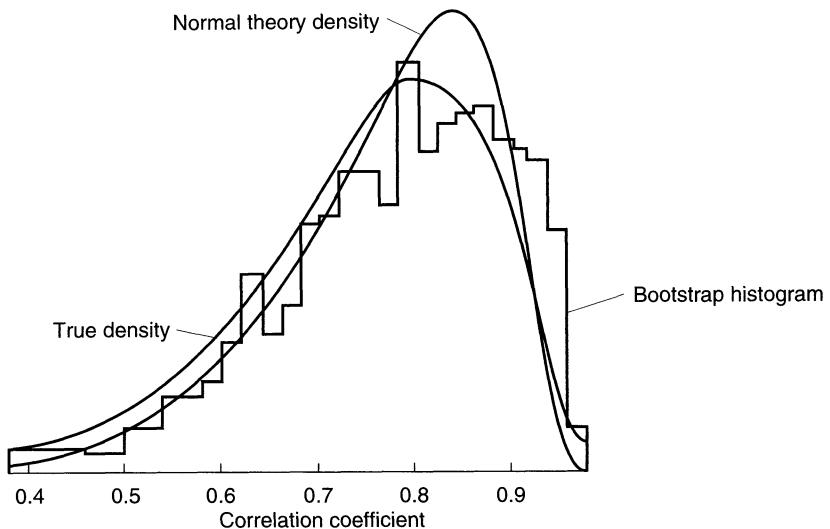


Figure 3.1. Density estimates for the correlation [From Diaconis and Efron (1983), Copyright©(1983) by Scientific American, Inc. All rights reserved]

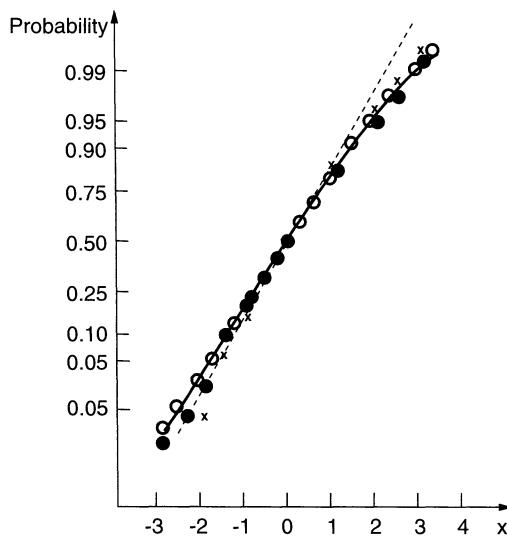


Figure 3.2. Distribution estimates for the studentized ratio [From Hinkley and Wei (1984), Copyright© (1984) by Biometrika, All rights reserved]

Dotted line, standard normal approximation; solid line, the true distribution; crosses, one-term Edgeworth expansion; solid circles, two-term Edgeworth expansion; open circles, bootstrap approximation.

Table 3.3. Comparison of the bootstrap and Edgeworth expansion estimators for the distribution of studentized ratio[†] [Adapted from Hinkley and Wei (1984), by permission of Biometrika]

x	-3	-2	-1.5	-1	0	1	1.5	2	3
True G_n	0.2	2.8	7	16	49	82	90	95	99
$G_{\text{BOOT}}^{(B)}$	ME	0.4	3	7	16	49	82	91	95
	SD	0.3	0.9	1	2	2	2	2	1
$G_{\text{EDG-1}}$	ME	0.1	1.8	6	15	49	83	93	97
	SD	0.1	0.7	1	1	1	1	1	0.1
$G_{\text{EDG-2}}$	ME	0.1	2.5	8	18	49	80	91	96
	SD	0.4	0.9	1	2	1	4	2	0.4

[†]The values of $G_n(x)$, ME, and SD are in %.

The results (based on 1000 simulation replications) in Figure 3.2 show that the bootstrap estimator is clearly better than the normal approximation, slightly better than the one-term Edgeworth expansion estimator, and very close to the two-term Edgeworth expansion estimator. These results confirm the theoretical results given in Section 3.3. The numerical results from Hinkley and Wei (1984) are listed in Table 3.3.

The results in Table 3.3 show that the bootstrap estimator and the two-term Edgeworth expansion estimator are comparable, but this is not generally true since theoretically one can show that the two-term Edgeworth expansion estimator has a faster convergence rate than the bootstrap estimator. We now present another comparison, made by Srivastava and Chan (1989), between the bootstrap and the two-term Edgeworth expansion estimators of

$$H_n(x) = P\{T_n/\theta \leq x\},$$

where $T_n = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ is the sample variance and $\theta = \text{var}(X_1)$. The exact form of H_n can be derived when X_1, \dots, X_n are i.i.d. from a normal distribution or a mixture of two normal distributions (Srivastava and Chan, 1989). Let $\hat{\kappa}_j$ be the j th sample cumulant of X_i . Then, the variance of $\sqrt{n}T_n$ can be estimated by $\hat{\sigma}_F^2 = \hat{\kappa}_4 + 2\hat{\kappa}_2^2$. The two-term Edgeworth expansion estimator of $H_n(x)$ is

$$H_{\text{EDG-2}}(x) = G_{\text{EDG-2}}(\sqrt{n}T_n(x-1)/\hat{\sigma}_F),$$

where

$$\begin{aligned} G_{\text{EDG-2}}(x) &= \Phi(x) - 6^{-1}n^{-1/2}\hat{b}_n\hat{\sigma}_F^{-3}\Phi^{(3)}(x) + n^{-1}\hat{\kappa}_2^2\hat{\sigma}_F^{-2}\Phi^{(2)}(x) \\ &\quad + 24^{-1}n^{-1}\hat{c}_n\hat{\sigma}_F^{-4}\Phi^{(4)}(x) + 2^{-1}n^{-1}(\hat{b}_n/6)^2\hat{\sigma}_F^{-6}\Phi^{(6)}(x) \end{aligned}$$

Table 3.4. Comparison of the bootstrap and Edgeworth expansion estimators for the distribution of sample variance[†] [Adapted from Srivastava and Chan (1989), by permission of Marcel Dekker]

True H_n	2.5	5	10	25	50	75	90	95	97.5	
$n = 20$										
$\varrho = (0, 1, 0)$										
$H_{\text{BOOT}}^{(B)}$ ME	3.2	5.6	10.0	23.0	47.6	75.6	91.7	96.2	98.3	
	SD	3.0	4.2	5.5	5.7	3.2	4.2	4.0	2.7	1.7
$H_{\text{EDG-2}}$ ME	1.1	3.3	8.2	24.0	51.2	77.0	90.8	94.8	96.8	
	SD	0.5	0.4	0.3	0.6	0.7	0.3	0.2	0.1	0.2
$\varrho = (0.2, 0.9, 1.4)$										
$H_{\text{BOOT}}^{(B)}$ ME	2.8	4.9	8.9	21.9	46.6	75.9	92.1	96.7	98.6	
	SD	2.1	2.7	3.7	4.0	2.6	2.9	3.0	2.0	1.2
$H_{\text{EDG-2}}$ ME	1.3	3.6	8.5	24.1	50.9	76.7	90.8	94.9	96.9	
	SD	0.5	0.4	0.3	0.6	0.7	0.3	0.2	0.1	0.2
$\varrho = (0.05, 0.7, 6.7)$										
$H_{\text{BOOT}}^{(B)}$ ME	1.0	1.8	3.5	10.1	29.8	72.7	98.3	99.8	99.9	
	SD	1.1	1.6	2.2	3.5	3.4	2.6	1.1	0.2	0.03
$H_{\text{EDG-2}}$ ME	12.3	15.1	19.1	28.4	43.8	66.7	89.1	95.5	98.2	
	SD	0.7	0.7	0.7	0.7	0.7	0.8	0.5	0.2	0.3
$n = 50$										
$\varrho = (0, 1, 0)$										
$H_{\text{BOOT}}^{(B)}$ ME	2.6	4.9	9.4	23.6	48.9	75.5	90.9	95.7	97.9	
	SD	1.9	2.6	3.5	3.9	2.3	3.3	3.0	2.2	1.5
$H_{\text{EDG-2}}$ ME	1.8	4.2	9.2	24.8	51.0	76.0	90.2	94.8	97.1	
	SD	0.2	0.2	0.1	0.2	0.3	0.3	0.1	0.1	0.1
$\varrho = (0.2, 0.9, 1.4)$										
$H_{\text{BOOT}}^{(B)}$ ME	2.4	4.6	9.1	23.2	48.6	75.6	91.2	95.9	98.1	
	SD	1.1	1.7	2.3	2.8	1.7	2.2	2.1	1.5	0.9
$H_{\text{EDG-2}}$ ME	1.9	4.3	9.3	24.8	50.7	75.8	90.2	94.9	97.2	
	SD	0.2	0.2	0.1	0.2	0.3	0.1	0.1	0.1	0.1
$\varrho = (0.05, 0.7, 6.7)$										
$H_{\text{BOOT}}^{(B)}$ ME	0.3	0.8	1.9	8.3	33.8	81.5	98.9	99.8	99.9	
	SD	0.3	0.5	1.1	2.4	2.6	2.4	0.8	0.2	0.04
$H_{\text{EDG-2}}$ ME	8.2	10.9	15.0	25.4	44.2	70.7	90.3	95.9	98.4	
	SD	0.6	0.5	0.3	0.2	0.6	0.7	0.1	0.2	0.3

[†]All the values are in %.

estimates the distribution of the studentized statistic $\sqrt{n}(T_n - \theta)/\hat{\sigma}_F$,

$$\hat{b}_n = \hat{\kappa}_6 + 12\hat{\kappa}_4\hat{\kappa}_2 + 4\hat{\kappa}_3^2 + 8\hat{\kappa}_2^2,$$

and

$$\hat{c}_n = \hat{\kappa}_8 + 24\hat{\kappa}_6\hat{\kappa}_2 + 32\hat{\kappa}_5\hat{\kappa}_3 + 32\hat{\kappa}_4^2 + 144\hat{\kappa}_4\hat{\kappa}_2^2 + 96\hat{\kappa}_3^2\hat{\kappa}_2 + 48\hat{\kappa}_2^4.$$

Apparently, the derivation of the Edgeworth expansion estimator is difficult and tedious.

The bootstrap estimator of $H_n(x)$ is

$$H_{\text{BOOT}}(x) = P_*\{T_n^*/T_n \leq x\} = G_{\text{BOOT}}(\sqrt{n}T_n(x-1)/\hat{\sigma}_F),$$

where G_{BOOT} is the bootstrap estimator of the distribution of the studentized statistic. In the simulation study, $H_{\text{BOOT}}(x)$ was approximated by Monte Carlo with $B = 1000$; the true population F was $(1-t)N(0, \sigma_1^2) + tN(0, \sigma_2^2)$ with some choices of $\varrho = (t, \sigma_1^2, \sigma_2^2)$ shown in Table 3.4.

The results in Table 3.4, based on 5000 simulation replicates, are extracted from Srivastava and Chan (1989). In the case where F is normal or close to normal, the ME of the two estimators are both close to the true value. The Edgeworth expansion estimator, however, has a smaller SD. When F has a long tail [$\varrho = (0.05, 0.7, 6.7)$], neither estimator performs well, although the Edgeworth expansion estimator still has a smaller SD.

More simulation results can be found in Chapter 4 (for bootstrap confidence sets) and in later chapters.

3.4.3 Conclusions

According to the empirical simulation results previously presented, we have the following conclusions.

- (1) For estimating the variance or standard deviation, the bootstrap estimator is not as efficient as the jackknife estimator when both estimators require relatively the same amount of computation. With more computations (a larger B), the bootstrap estimator may be more efficient than the jackknife estimator, but the difference is negligible when n is large. Furthermore, the bootstrap estimator is downward-biased in many cases.
- (2) The bootstrap estimators of other moments such as the bias and the skewness do not have good performances.
- (3) The bootstrap distribution estimator is better than the normal approximation, and the improvement can be quite substantial. The bootstrap estimator is also slightly better than the one-term Edgeworth expansion estimator, but is slightly worse than the two-term Edgeworth expansion estimator.

3.5 Smoothed Bootstrap

When X_1, \dots, X_n are i.i.d. and H_{P_n} is determined by F , so far we have only considered taking bootstrap data from the empirical distribution F_n . In many problems, it is known that F is smooth (e.g., F is continuous or F has a density f), and, therefore, one might think about drawing bootstrap data from a smoothed estimator \tilde{F}_n of F . Discussions about how to obtain smooth estimators of F can be found in Prakasa Rao (1983) or Silverman (1986).

3.5.1 Empirical evidences and examples

The possible advantages of taking bootstrap data from a smooth estimator \tilde{F}_n were addressed by Efron (1979) when he proposed the bootstrap method. Using empirical simulation, Efron (1982) compared the performances of the smoothed bootstrap resampling from \tilde{F}_n and the nonsmoothed bootstrap resampling from F_n . The statistics he considered are the sample correlation coefficient $\hat{\rho}_n$ based on i.i.d. $X_i = (Y_i, Z_i)'$, $i = 1, \dots, n$, and its variance-stabilizing transformation $\hat{\phi}_n$ (see Example 2.1). The simulation setting is the same as that of case 3 in Example 2.1 (Section 2.1.3). Two different smoothed bootstrap methods were considered. The first one (called the N-smoothed bootstrap) takes bootstrap data from the distribution

$$\tilde{F}_n(y, z) = \frac{1}{n} \sum_{i=1}^n K\left(\frac{y - Y_i}{h_n}, \frac{z - Z_i}{h_n}\right),$$

where K is the bivariate normal distribution function, with mean \bar{X}_n and covariance matrix $\hat{\Sigma} = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)'$, and $h_n = 0.5$ is the parameter controlling the degree of smoothing. The second method (called the U-smoothed bootstrap) is the same as the first, except that K is replaced by the uniform distribution function over a rhombus with mean \bar{X}_n and covariance matrix $\hat{\Sigma}$. In the computation of the nonsmoothed and two smoothed bootstrap standard deviation estimators, Monte Carlo approximations of size 128 were used. Simulation estimates of the ME (mean), CV (coefficient of variation), and RM (root mean squared error) of the bootstrap estimators are listed in Table 3.5.

In terms of the CV and RM, both smoothed bootstrap methods improve the nonsmoothed bootstrap substantially in the case of $T_n = \hat{\phi}_n$ but only slightly in the case of $T_n = \hat{\rho}_n$; the N-smoothed bootstrap is better than the U-smoothed bootstrap. Hence, the effect of smoothing depends on the statistic under consideration and the selection of K (the method of smoothing). The effect of smoothing should also depend on the smoothing parameter h_n , which was not investigated in this simulation study.

Table 3.5. Comparisons of the smoothed and nonsmoothed bootstrap estimators of the standard deviation of T_n [Adapted from Efron (1982), by permission of Society for Industrial and Applied Mathematics]

T_n	σ_n	Nonsmoothed			N-Smoothed			U-Smoothed		
		ME	CV	RM	ME	CV	RM	ME	CV	RM
$\hat{\rho}_n$.218	.206	.320	.067	.200	.300	.063	.205	.298	.062
$\hat{\phi}_n$.299	.301	.216	.065	.296	.139	.041	.298	.195	.058

The simulation result cannot provide a general guide. It is difficult in general to decide whether we should smooth. This can be illustrated by the following simple cases.

Suppose that the quantity to be estimated is

$$A(F) = \int a(x)dF(x).$$

The nonsmoothed bootstrap estimator is the same as the substitution estimator:

$$A(F_n) = \int a(x)dF_n(x) = \frac{1}{n} \sum_{i=1}^n a(X_i).$$

Consider the smooth estimator of F given by

$$\tilde{F}_n(x) = \frac{1}{n} \sum_{i=1}^n K\left(\frac{V^{-1/2}(x - X_i)}{h_n}\right), \quad (3.61)$$

where V is an estimator of the covariance matrix of X_1 , K is a symmetric p -variate distribution with unit covariance matrix, and h_n is a smoothing parameter. The smoothed bootstrap estimator of $A(F)$ is then

$$A(\tilde{F}_n) = \int a(x)d\tilde{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \int a(X_i + h_n V^{1/2}x)dK(x).$$

We now compare the mean squared errors (mse) of $A(F_n)$ and $A(\tilde{F}_n)$. Without loss of generality we assume that $A(F) = 0$. Then,

$$\text{mse}(A(F_n)) = \frac{1}{n} \int [a(x)]^2 dF(x) \quad (3.62)$$

and the convergence rate of $A(F_n)$ is $n^{-1/2}$. Using Taylor's expansion, Silverman and Young (1987) showed that

$$\text{mse}(A(\tilde{F}_n)) = \text{mse}(A(F_n)) + \frac{h_n^2}{n} \int a(x)b(x)dF(x) + O(h_n^4), \quad (3.63)$$

where $b(x) = \text{tr}[V\nabla^2 a(x)]$. Assume that $nh_n^2 \rightarrow 0$ as $n \rightarrow \infty$. Then, for sufficiently large n , $\text{mse}(A(\tilde{F}_n)) < \text{mse}(A(F_n))$ if and only if

$$\int a(x)b(x)dF(x) < 0. \quad (3.64)$$

This shows that, even if F is smooth, the smoothed bootstrap is not better than the nonsmoothed bootstrap when (3.64) does not hold. The statistical meaning of condition (3.64) is not clear.

We next study a more realistic example provided by Wang (1989): the estimation of $p_n(x)$, the density of $\sqrt{n}(\bar{X}_n - \mu)$ at a fixed $x \in \mathbb{R}$, where \bar{X}_n is the sample mean of i.i.d. random variables X_1, \dots, X_n with mean μ . For the nonsmoothed bootstrap, the distribution estimator $H_{\text{BOOT}}(x) = P_*\{\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x\}$ is not differentiable. However, we can define a bootstrap estimator $p_{\text{BOOT}}(x)$ by using finite difference ratios (Wang, 1989). Using \tilde{F}_n in (3.61) with one-dimensional $V = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$, Wang (1989) defined a smoothed bootstrap estimator of $p_n(x)$ by

$$p_{\text{SBOOT}}(x) = \frac{d}{dx} P_*\{\sqrt{n}(\tilde{X}_n^* - \bar{X}_n) \leq x\},$$

where \tilde{X}_n^* is the sample mean of n i.i.d. data from \tilde{F}_n . Under some conditions, Wang (1989) showed that

$$\text{mse}(p_{\text{BOOT}}(x)) = \frac{\nu[(x^2 - \sigma^2)p_n(x)]^2}{4n\sigma^4} + O\left(\frac{1}{n^{3/2}}\right) \quad (3.65)$$

and

$$\begin{aligned} \text{mse}(p_{\text{SBOOT}}(x)) &= \text{mse}(p_{\text{BOOT}}(x)) - \frac{c_1(x/\sigma)h_n^2}{\sigma^2 n} \\ &\quad + \frac{c_2(x/\sigma)h_n^4}{\sigma^2} + o\left(h_n^4 + \frac{h_n^2}{n}\right), \end{aligned} \quad (3.66)$$

where $\sigma^2 = \text{var}(X_1)$, $\nu = \sigma^{-4}E(X_1 - \mu)^4 - 1$,

$$c_1(x) = [\varphi(x)]^2(x^2 - 1)[(x^2 - 1)/2 + \nu(-3x^4 + 12x^2 - 5)/8],$$

$$c_2(x) = [\varphi(x)]^2(x^2 - 1)^2/4,$$

and φ is the standard normal density. Furthermore, for each x such that $c_1(x/\sigma) > 0$, the best choice for h_n is $h_n = \sqrt{c_1(x/\sigma)}/\sqrt{2c_2(x/\sigma)n}$, and, in such a case,

$$\text{mse}(p_{\text{SBOOT}}(x)) = \text{mse}(p_{\text{BOOT}}(x)) - \frac{[c_1(x/\sigma)]^2}{4\sigma^2 c_2(x/\sigma)n^2} + o\left(\frac{1}{n^2}\right),$$

i.e., the smoothed bootstrap estimator has a smaller mean squared error than the nonsmoothed bootstrap estimator when n is large. However, if $c_1(x/\sigma) \leq 0$, then the smoothed bootstrap is not necessarily better. We again have no definite conclusion.

It follows from (3.62)–(3.63) and (3.65)–(3.66) that the ratio of the mean squared errors between the two kinds of bootstrap estimators tends to 1 as $n \rightarrow \infty$. That is, the smoothed and nonsmoothed bootstrap estimators have the same amse and are asymptotically equivalent. This actually is true in many situations where the nonsmoothed bootstrap estimator has convergence rate $n^{-1/2}$. In such cases, we cannot expect a great improvement by smoothing unless n is small. The story is quite different when the statistic T_n is not very smooth, so that the nonsmoothed bootstrap estimator converges slowly. One example is studied next.

3.5.2 Sample quantiles

For a univariate population F , let $\theta = F^{-1}(t)$, $0 < t < 1$. In Section 3.3.1, we stated that, for the sample quantile $F_n^{-1}(t)$, the bootstrap estimator $H_{\text{BOOT}}(x) = P_*\{\sqrt{n}[F_n^{*-1}(t) - F_n^{-1}(t)] \leq x\}$ converges to $H_n(x) = P\{\sqrt{n}[F_n^{-1}(t) - \theta] \leq x\}$ with rate $n^{-1/4}$, where F_n^* is the (nonsmoothed) bootstrap analog of the empirical distribution F_n . Consider the smoothed bootstrap taking i.i.d. $\tilde{X}_1^*, \dots, \tilde{X}_n^*$ from the smoothed estimator

$$\tilde{F}_n(x) = \frac{1}{n} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right),$$

where $K(x)$ is a univariate distribution with 0 mean. Let \tilde{F}_n^* be the bootstrap analog of F_n based on $\tilde{X}_1^*, \dots, \tilde{X}_n^*$, (i.e., \tilde{F}_n^* is the empirical distribution of $\tilde{X}_1^*, \dots, \tilde{X}_n^*$). Then the smoothed bootstrap estimator of H_n is

$$H_{\text{SBOOT}}(x) = P_*\{\sqrt{n}[\tilde{F}_n^{*-1}(t) - \tilde{F}_n^{-1}(t)] \leq x\}.$$

Falk and Reiss (1989) showed that

$$\rho_\infty(H_{\text{SBOOT}}, H_n) = O_p((nh_n)^{-1/2}),$$

provided that: (a) F is three times continuously differentiable near θ and $f(\theta) > 0$, $f(x) = dF/dx$; (b) K has support $[-1, 1]$ and is three times differentiable with bounded derivative; and (c) $nh_n^3 \rightarrow \infty$ and $nh_n^5(\log n)^2 \rightarrow 0$.

If we take $h_n = n^{-1/5-\epsilon}$ for $0 < \epsilon < 2/15$, then $(nh_n)^{-1/2} = n^{-2/5+\epsilon/2} = o(n^{-1/4})$, i.e., H_{SBOOT} has a faster convergence rate than H_{BOOT} .

We next consider variance estimation. As we indicated in Example 1.4, the nonsmoothed bootstrap estimator v_{BOOT} has an explicit form in this case:

$$v_{\text{BOOT}} = v_n(F_n),$$

where

$$v_n(F) = \int x^2 q_n(F(x)) dF(x) - \left[\int x q_n(F(x)) dF(x) \right]^2 \quad (3.67)$$

is the approximate variance of $F_n^{-1}(t)$, $q_n(x) = m(n/m)x^{m-1}(1-x)^{n-m}$, and $m = [nt] + 1$. Similarly, the smoothed bootstrap variance estimator is

$$v_{\text{SBOOT}} = v_n(\tilde{F}_n).$$

Hall and Martin (1988c) showed that if $E|X_1|^\epsilon < \infty$ for some $\epsilon > 0$, $f(\theta) > 0$, and f satisfies $|f(x) - f(y)| \leq c|x - y|^\delta$ for all x and y in a neighborhood of θ and a fixed $\delta \in (\frac{1}{2}, 1]$, then

$$n^{5/4}[v_{\text{BOOT}} - v_n(F)] \rightarrow_d N(0, 2\pi^{-1/2}[t(1-t)]^{3/2}[f(\theta)]^{-4}).$$

Under the same conditions, $v_n(F) = n^{-1}t(1-t)[f(\theta)]^{-2} + o(n^{-1})$. Hence, v_{BOOT} has convergence rate $n^{-1/4}$. This rate is unsatisfactorily slow. Some other estimators of $v_n(F)$ have a faster convergence rate; for example, the kernel estimator proposed by Bloch and Gastwirth (1968). The variance estimator in Csörgő (1983) has the convergence rate $n^{-1/2+\epsilon_0}$ for any $\epsilon_0 > 0$.

Hall, DiCiccio and Romano (1989) showed that

$$n^{1+s/(2s+1)}[v_{\text{SBOOT}} - v_n(F)] \rightarrow_d N(c_1, c_2), \quad (3.68)$$

where $s \geq 2$, c_1 and c_2 are nonzero constants, provided that: (a) $k(x) = dK/dx$ is bounded and dk/dx is an absolutely integrable continuous function of bounded variation; (b) $\int |x^s k(x)| dx < \infty$ and $\int x^j k(x) dx = 0$ for $1 \leq j \leq s-1$ but $\int x^s k(x) dx \neq 0$; (c) f is bounded away from zero in a neighborhood of θ , the j th order derivative of f is bounded for $0 \leq j \leq s$, and $d^s f/dx^s$ is uniformly continuous; (d) $E|X_1|^\epsilon < \infty$ for some $\epsilon > 0$; and (e) $h_n = [c_3/(2nc_4)]^{1/(2s+1)}$ for some positive constants c_3 and c_4 . It follows from (3.68) that the convergence rate of v_{SBOOT} is $n^{-s/(2s+1)}$. Since $s/(2s+1) \geq 2/5 > 1/4$, v_{SBOOT} improves v_{BOOT} in terms of the convergence rate. By selecting an s we may obtain the convergence rate of v_{SBOOT} being $n^{-1/2+\epsilon_0}$ for any $\epsilon_0 > 0$. The smoothing parameter h_n suggested above is selected by minimizing the mean squared error of $v_{\text{SBOOT}} - v_n(F)$. The order of h_n is the same as that of the smoothing parameter minimizing the mean integrated squared error of \tilde{F}_n . Only in the case of $s = 2$ may K be chosen to be a distribution. But $v_{\text{SBOOT}} = v_n(\tilde{F}_n)$ is still well defined even if \tilde{F}_n is not a distribution. Lee and Young (1994) proposed applying a negativity correction to k so that the resulting \tilde{F}_n is a distribution.

3.5.3 Remarks

- (1) Since the smoothed estimator \tilde{F}_n of F is consistent under some conditions on K and h_n , we can expect that the asymptotic results for the nonsmoothed bootstrap presented in the previous sections still hold for

the smoothed bootstrap (Bickel and Freedman, 1981; Singh, 1981).

- (2) The smoothed bootstrap is useful when the sample size n is small. For large n , the smoothed bootstrap estimator improves the nonsmoothed bootstrap estimator when the latter has a slow convergence rate (e.g., a rate much slower than $n^{-1/2}$), which occurs in the case of sample quantiles. More discussions can be found in DeAngelis and Young (1992). However, we should also keep in mind that smoothing usually increases the amount of computational work.
- (3) The results for sample quantiles may be extended to the cases where the statistical procedures are derived based on the L_1 -norm. This includes the L_1 regression estimators and some others (Dodge, 1987).
- (4) One may argue that instead of using the nonsmooth sample quantile $F_n^{-1}(t)$ and taking bootstrap data from a smooth \tilde{F}_n , it might be better to adopt the smooth quantile estimator such as $\tilde{F}_n^{-1}(t)$ and then apply the bootstrap by drawing bootstrap data from the same \tilde{F}_n . However, bootstrapping $\tilde{F}_n^{-1}(t)$ using the same \tilde{F}_n may produce inconsistent bootstrap estimators, and, therefore, a further smoothed bootstrap may be necessary. Thus, for a quantile estimator $\hat{F}^{-1}(t)$, whether \hat{F} is smooth or not, the problem remains the same: smoothing \hat{F} is worthwhile when applying the bootstrap. More detailed discussions will be provided in Chapter 8.
- (5) The problem of determining the smoothing parameter h_n is difficult. When n is large, one may apply methods such as cross-validation, which can be found in the literature on density estimation (Chapter 8).

3.6 Nonregular Cases

Although we have shown that bootstrap estimators are consistent for many commonly used statistics, there are some cases, known as nonregular cases, where the bootstrap yields inconsistent estimators. The causes of inconsistency in bootstrap estimation are given at the end of Section 3.2.1. Examples of inconsistency caused by the lack of moment conditions are given by Theorem 3.2 and Examples 3.4 and 3.5. In this section, we provide some examples of another nonregular case: nonsmooth statistics and/or statistics whose asymptotic distributions are not normal. Recall that we pointed out that the inconsistency of the bootstrap caused by the lack of moment conditions can be rectified by varying the bootstrap sample size (the number of bootstrap data taken from the empirical distribution). We show in this section that the inconsistency caused by nonsmoothness can also be rectified by changing the bootstrap sample size n to m_n with $m_n \rightarrow \infty$ and $m_n/n \rightarrow 0$ as $n \rightarrow \infty$. We will also explain why this change produces a consistent bootstrap estimator.

Example 3.6. Functions of the sample mean with null derivatives. Let X_1, \dots, X_n be i.i.d. p -vectors from F having mean μ and covariance matrix Σ , g be a function from \mathbb{R}^p to \mathbb{R} , $\theta = g(\mu)$, and $T_n = g(\bar{X}_n)$, where \bar{X}_n is the sample mean. Suppose that g is continuously second order differentiable in a neighborhood of μ , $\nabla g(\mu) = 0$, and $\nabla^2 g(\mu) \neq 0$. Using Taylor's expansion and $\nabla g(\mu) = 0$, we obtain that

$$T_n - \theta = \frac{1}{2}(\bar{X}_n - \mu)' \nabla^2 g(\mu)(\bar{X}_n - \mu) + o_p(n^{-1}). \quad (3.69)$$

From (3.69),

$$n(T_n - \theta) \rightarrow_d \frac{1}{2} Z_\Sigma' \nabla^2 g(\mu) Z_\Sigma,$$

where Z_Σ is p -variate normal with mean 0 and covariance matrix Σ . Thus, we should study the bootstrap estimator of the distribution of $n(T_n - \theta)$, not $\sqrt{n}(T_n - \theta)$, in the case of $\nabla g(\mu) = 0$. Let X_1^*, \dots, X_n^* be an i.i.d. bootstrap sample from the empirical distribution F_n , and let \bar{X}_n^* and T_n^* be the bootstrap analogs of \bar{X}_n and T_n , respectively. Then, similar to (3.69),

$$\begin{aligned} T_n^* - T_n &= \nabla g(\bar{X}_n)'(\bar{X}_n^* - \bar{X}_n) + \frac{1}{2}(\bar{X}_n^* - \bar{X}_n)' \nabla^2 g(\bar{X}_n)(\bar{X}_n^* - \bar{X}_n) \\ &\quad + o_p(n^{-1}) \text{ a.s.} \end{aligned} \quad (3.70)$$

By Theorem 3.1, for almost all given sequences X_1, X_2, \dots ,

$$\frac{n}{2}(\bar{X}_n^* - \bar{X}_n)' \nabla^2 g(\bar{X}_n)(\bar{X}_n^* - \bar{X}_n) \rightarrow_d \frac{1}{2} Z_\Sigma' \nabla^2 g(\mu) Z_\Sigma.$$

Since $\nabla g(\mu) = 0$,

$$\sqrt{n} \nabla g(\bar{X}_n) = \sqrt{n} \nabla^2 g(\mu)(\bar{X}_n - \mu) + o_p(1) \rightarrow_d \nabla^2 g(\mu) Z_\Sigma. \quad (3.71)$$

Hence, for almost all given sequences X_1, X_2, \dots , the conditional distribution of $n \nabla g(\bar{X}_n)'(\bar{X}_n^* - \bar{X}_n)$ does not have a limit. It follows from (3.70) that, for almost all given X_1, X_2, \dots , the conditional distribution of $n(T_n^* - T_n)$ does not have a limit. Therefore, the bootstrap estimator of the distribution of $n(T_n - \theta)$ is inconsistent.

Example 3.6 indicates that the bootstrap estimator H_{BOOT} may not have a limit while H_n has a limit. This is caused by an inherent problem of the bootstrap: the bootstrap data are drawn from F_n , which is not exactly F . The effect of this problem is inappreciable in a regular case (i.e., T_n can be well approximated by a linear statistic) but leads to inconsistency in a nonregular case such as Example 3.6. The symptom of this problem in Example 3.6 is that $\nabla g(\bar{X}_n)$ is not necessarily equal to 0 when $\nabla g(\mu) = 0$. As a result, the expansion in (3.70), compared with the expansion in (3.69), has an extra term $\nabla g(\bar{X}_n)'(\bar{X}_n^* - \bar{X}_n)$ that does not converge to 0 fast enough, and, therefore, the conditional distribution of $\mathfrak{R}_n^* = n(T_n^* - T_n)$ cannot mimic the distribution of $\mathfrak{R}_n = n(T_n - \theta)$. When we take bootstrap

data, it is not necessary that we always take n data. Let m be an integer, X_1^*, \dots, X_m^* be i.i.d. from F_n , and $\mathfrak{R}_m^* = \mathfrak{R}_m(X_1^*, \dots, X_m^*, \hat{P}_n)$ be the bootstrap analog of \mathfrak{R}_n based on m bootstrap data. Then we may use the following bootstrap estimator of H_n in (3.1):

$$H_{\text{BOOT-}m}(x) = P_*\{\mathfrak{R}_m^* \leq x\}. \quad (3.72)$$

Bickel and Freedman (1981) actually studied this bootstrap estimator with m as a function of n or with m varying freely. There is no apparent reason why we should always use $m = n$, but $m = n$ ($H_{\text{BOOT-}m} = H_{\text{BOOT}}$) is customarily used, and it works well for regular cases as we have shown in the previous sections. However, we will show that allowing a bootstrap sample size m different from n gives us more freedom for rectifying the inconsistency of the bootstrap estimator in the nonregular cases.

If we could take the sample size $n = \infty$, then $H_{\text{BOOT-}m}$ would be consistent since X_1^*, \dots, X_m^* can almost be viewed as data from F . Of course, taking $n = \infty$ is impossible in practice. However, we may select $m = m_n$ so that F_n converges faster than its bootstrap analog F_m^* , and thus achieve the same effect as taking $n = \infty$ (or generating the bootstrap data from F). Since $F_m^* - F_n = O_p(m^{-1/2})$ and $F_n - F = O_p(n^{-1/2})$ or $O(\sqrt{\log \log n/n})$ a.s., for the weak consistency of $H_{\text{BOOT-}m}$ we may let $m = m_n$ with $m_n \rightarrow \infty$ and $m_n/n \rightarrow 0$ as $n \rightarrow \infty$; for the strong consistency of $H_{\text{BOOT-}m}$ we may let $m = m_n$ with $m_n \rightarrow \infty$ and $m_n \log \log n/n \rightarrow 0$. We now show that this change of bootstrap sample size does provide consistent bootstrap estimators in many nonregular cases where the inconsistency is caused by the lack of smoothness of T_n .

Example 3.6 (continued). Let \bar{X}_m^* and T_m^* be the bootstrap analogs of \bar{X}_n and T_n , respectively, based on a bootstrap sample of size m . Then,

$$\begin{aligned} T_m^* - T_n &= \nabla g(\bar{X}_n)'(\bar{X}_m^* - \bar{X}_n) + \frac{1}{2}(\bar{X}_m^* - \bar{X}_n)'\nabla^2 g(\bar{X}_n)(\bar{X}_m^* - \bar{X}_n) \\ &\quad + o_p(m^{-1}) \text{ a.s.} \end{aligned} \quad (3.73)$$

By Theorem 2.1 in Bickel and Freedman (1981), for almost all given sequences X_1, X_2, \dots ,

$$\frac{m}{2}(\bar{X}_m^* - \bar{X}_n)'\nabla^2 g(\bar{X}_n)(\bar{X}_m^* - \bar{X}_n) \rightarrow_d \frac{1}{2}Z_\Sigma' \nabla^2 g(\mu)Z_\Sigma, \quad (3.74)$$

as $m \rightarrow \infty$ and $n \rightarrow \infty$. The expansion in (3.73) still has the nonzero term $\nabla g(\bar{X}_n)'(\bar{X}_m^* - \bar{X}_n)$, but it is of order $o_p(m^{-1})$ a.s. if $m \log \log n/n \rightarrow 0$ since, by (3.71), $\nabla g(\bar{X}_n) = O(\sqrt{\log \log n/n})$ a.s. This proves that

$$T_m^* - T_n = \frac{1}{2}(\bar{X}_m^* - \bar{X}_n)'\nabla^2 g(\bar{X}_n)(\bar{X}_m^* - \bar{X}_n) + o_p(m^{-1}) \text{ a.s.}, \quad (3.75)$$

as if we had taken $n = \infty$ or $\nabla g(\bar{X}_n)$ had been 0. By (3.74) and (3.75), the bootstrap estimator $H_{\text{BOOT-}m}$ is strongly consistent if $m = m_n \rightarrow \infty$

and $m_n \log \log n/n \rightarrow 0$. We can similarly show that $H_{\text{BOOT},m}$ is weakly consistent if $m_n \rightarrow \infty$ and $m_n/n \rightarrow 0$, since $\nabla g(\bar{X}_n) = O_p(n^{-1/2})$.

The result in Example 3.6 can be extended to the general case where $T_n = T(F_n)$ with a second order differentiable functional T having a null first order differential, i.e., $\phi_F(x) = 0$ for all x . Examples of these types of statistics are some “goodness of fit” test statistics under a null hypothesis. The following is an example.

Example 3.7. Weighted Cramér-von Mises test statistic. Consider the test problem

$$H_0 : F = F_0 \quad \text{versus} \quad H_1 : F \neq F_0,$$

with a given distribution F_0 , and the statistic generated by the functional

$$T(G) = \int w_{F_0}(x)[G(x) - F_0(x)]^2 dF_0(x),$$

where $w_{F_0}(x)$ is a weight function depending on the known distribution F_0 and satisfying $\int w_{F_0}(x)dF_0(x) < \infty$. When $w_{F_0}(x) = 1$ for all x , $T(F_n)$ is the Cramér-von Mises test statistic given in Example 2.8.

It can be shown that T is ρ_∞ -Fréchet differentiable at F with the influence function

$$\phi_F(x) = 2 \int w_{F_0}(y)[I\{x \leq y\} - F(y)][F(y) - F_0(y)]dF_0(y).$$

Under the null hypothesis H_0 , $F = F_0$ and $\phi_F \equiv 0$.

It can also be shown that T is second order ρ_∞ -Fréchet differentiable at F and

$$T(G) - T(F) = \int \phi_F(x)dF(x) + \int \int \eta_F(x, y)d(G - F)(x)d(G - F)(y),$$

where $\eta_F(x, y) = \int w_{F_0}(t)[I\{x \leq t\} - F(t)][I\{y \leq t\} - F(t)]dF_0(t)$. Thus, under H_1 , $\sqrt{n}[T(F_n) - T(F)]$ is asymptotically normal and, under H_0 , $\phi_F \equiv 0$ and $n[T(F_n) - T(F)] \rightarrow_d W$, which is a weighted sum of independent chi-square random variables. Under the null hypothesis H_0 , it can be shown that the bootstrap estimator of the distribution of $n[T(F_n) - T(F)]$ is inconsistent with a bootstrap sample size of n .

We have the following general result that establishes the consistency of the bootstrap estimator $H_{\text{BOOT},m}$ in situations like Example 3.7. Its proof uses the idea described in Example 3.6 and can be found in Shao (1994a).

Theorem 3.14. Let $T_n = T(F_n)$, $H_n(x) = P\{n[T(F_n) - T(F)] \leq x\}$, and $H_{\text{BOOT},m}(x) = P_*\{m[T(F_m^*) - T(F_n)] \leq x\}$, where F_m^* is the empirical distribution based on a bootstrap of size m generated from F_n .

(i) Suppose that T is second order ρ_∞ -Fréchet differentiable at F with the function $\psi_F(x, y)$ in (2.30) satisfying (2.32) and $\int \psi_F(x, y)dF(y) = 0$ for all x . If $m = m_n \rightarrow \infty$ and $m_n/n \rightarrow 0$, then

$$\rho_\infty(H_{\text{BOOT-}m}, H_n) \rightarrow_p 0. \quad (3.76)$$

(ii) Result (3.76) also holds if T is second order $\rho_{\infty+1}$ -Fréchet differentiable at F and $\int \{F(x)[1 - F(x)]\}^{1/2}dx \leq \infty$.

In Examples 3.6 and 3.7, the functionals are differentiable, but their differentials are 0. In the following, we consider the case where the functional is not differentiable.

Example 3.8. Nondifferentiable functions of the sample mean. Consider $T_n = g(\bar{X}_n)$ with a nonsmooth function g . For simplicity, we focus on one-dimensional X_i with $\text{var}(X_i) = \sigma^2 < \infty$. Assume that g is differentiable except at μ and that

$$\lim_{t \rightarrow 0^\pm} \frac{g(\mu + t) - g(\mu)}{t} = g'(\mu \pm)$$

exist but $g'(\mu+) \neq g'(\mu-)$. For example, $g(x) = |x|$ is nondifferentiable at 0, and $g'(0\pm) = \pm 1$. In general, $H_n(x) = P\{\sqrt{n}(T_n - \theta) \leq x\}$ does not converge to a normal distribution and may not have any limit when both $g'(\mu+)$ and $g'(\mu-)$ are nonzero.

To see the inconsistency of $H_{\text{BOOT}}(x) = P_*\{\sqrt{n}(T_n^* - T_n) \leq x\}$, the bootstrap estimator of $H_n(x)$ based on n bootstrap data, we consider the special case of $g(x) = |x|$ and $\mu = 0$. Note that $\sqrt{n}|\bar{X}_n| \rightarrow_d |Z_{\sigma^2}|$ and $\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \rightarrow_d Z_{\sigma^2}$ a.s. But conditional on X_1, X_2, \dots , $\sqrt{n}\bar{X}_n^*$ has no limit, since $\sqrt{n}\bar{X}_n$ does not converge. Thus,

$$\sqrt{n}(|\bar{X}_n^*| - |\bar{X}_n|) = \begin{cases} \sqrt{n}(\bar{X}_n^* - \bar{X}_n) - 2\sqrt{n}\bar{X}_n^*I\{\bar{X}_n^* < 0\} & \bar{X}_n \geq 0 \\ \sqrt{n}(\bar{X}_n - \bar{X}_n^*) + 2\sqrt{n}\bar{X}_n^*I\{\bar{X}_n^* \geq 0\} & \bar{X}_n < 0 \end{cases}$$

has no limit.

The reason for the inconsistency of the bootstrap estimators in this case is that \bar{X}_n is not exactly equal to μ , and both \bar{X}_n and its bootstrap analog \bar{X}_n^* oscillate around the discontinuity point of g' with the same rate $n^{-1/2}$. The inconsistency can be rectified if the bootstrap analog oscillates with a rate slower than $n^{-1/2}$. This again leads us to consider taking a bootstrap sample X_1^*, \dots, X_m^* with $m/n \rightarrow 0$. Define $T_m^* = g(\bar{X}_m^*)$ and $H_{\text{BOOT-}m}(x) = P_*\{\sqrt{m}(T_m^* - T_n) \leq x\}$. If $m \rightarrow \infty$ and $m/n \rightarrow 0$, then $H_{\text{BOOT-}m}$ is weakly consistent; if $m \rightarrow \infty$ and $m \log \log n/n \rightarrow 0$, then $H_{\text{BOOT-}m}$ is strongly consistent. To verify these assertions, we consider the case of $\mu = 0$. For m satisfying $m/n \rightarrow 0$ or $m \log \log n/n \rightarrow 0$,

$$\sqrt{m}(T_n - \theta) \rightarrow_p 0 \quad \text{or} \quad \rightarrow_{a.s.} 0.$$

Using the mean-value theorem, we obtain that

$$\begin{aligned}
H_{\text{BOOT-}m}(x) &= P_*\{\sqrt{m}(T_m^* - \theta) \leq x\} + o(1) \\
&= P_*\{g'(0+)\sqrt{m}\bar{X}_m^* \leq x, \bar{X}_m^* \geq 0\} \\
&\quad + P_*\{g'(0-) \sqrt{m}\bar{X}_m^* \leq x, \bar{X}_m^* < 0\} + o(1) \\
&= P\{g'(0+)\sqrt{n}\bar{X}_n \leq x, \bar{X}_n \geq 0\} \\
&\quad + P\{g'(0-) \sqrt{n}\bar{X}_n \leq x, \bar{X}_n < 0\} + o(1) \\
&= P\{\sqrt{n}(T_n - \theta) \leq x\} + o(1),
\end{aligned}$$

where the third equality follows from the fact that $\sqrt{m}(\bar{X}_m^* - \bar{X}_n)$ has the same limit as $\sqrt{n}\bar{X}_n$ and $o(1)$ is $o_p(1)$ for m satisfying $m/n \rightarrow 0$ or $o(1)$ a.s. for m satisfying $m \log \log n/n \rightarrow 0$.

We have actually shown an example in which H_n may not have any limit, but its bootstrap estimator $H_{\text{BOOT-}m}$ is consistent.

The next example is given by Bickel and Freedman (1981) and Loh (1984).

Example 3.9. Extreme order statistics. Let X_1, \dots, X_n be i.i.d. random variables from a distribution F and let $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ be the order statistics. $X_{(1)}$ and $X_{(n)}$ are called extreme order statistics. We focus on $X_{(n)}$; the result for $X_{(1)}$ can be similarly obtained. Loh (1984) showed that, if $F(\theta) = 1$ for some θ , $F(x) < 1$ for $x < \theta$, and F belongs to the domain of attraction of the type II extreme value law, i.e.,

$$H_n(x) = P\{n^{1/\delta}(\theta - X_{(n)}) \leq x\} \rightarrow 1 - e^{-(x/\theta)^\delta}, \quad (3.77)$$

where $\delta > 0$ is a fixed constant, then for almost all sequences X_1, X_2, \dots , the bootstrap estimator $H_{\text{BOOT}}(x) = P_*\{n^{1/\delta}(X_{(n)} - X_{(n)}^*) \leq x\}$ of $H_n(x)$ is inconsistent, where $X_{(n)}^*$ is the maximum of X_1^*, \dots, X_n^* that are i.i.d. from the empirical distribution F_n . Indeed, it is easy to see that

$$H_n(0) = P\{X_{(n)} = \theta\} \equiv 0, \quad (3.78)$$

whereas

$$H_{\text{BOOT}}(0) = P_*\{X_{(n)}^* = X_{(n)}\} = 1 - (1 - n^{-1})^n \rightarrow 1 - e^{-1}. \quad (3.79)$$

The reasons for the inconsistency of the bootstrap estimators in Examples 3.6-3.9 are the same, although the problems in these examples are entirely different. Although $X_{(n)} \rightarrow \theta$, it never equals θ in view of (3.78). But when the bootstrap sample size is n , $X_{(n)}^*$ has the same convergence rate as $X_{(n)}$, which leads to (3.79) and the inconsistency of the bootstrap estimator. Bickel and Freedman (1981) pointed out that this inconsistency

cannot be mended by smoothing, i.e., taking bootstrap data from a smooth estimator \tilde{F}_n of F . However, with a bootstrap sample size m satisfying $m/n \rightarrow 0$, $X_{(n)}$ converges to θ faster than its bootstrap analog $X_{(m)}^*$, and the inconsistency of the bootstrap can be rectified. Indeed, if $m/n \rightarrow 0$, then, in contrast with (3.79),

$$P_*\{X_{(m)}^* = X_{(n)}\} = 1 - (1 - n^{-1})^m \rightarrow 0.$$

Swanepoel (1986) first showed that, in the case where F is the uniform on $[0, \theta]$, the inconsistency of the bootstrap can be rectified by using the bootstrap sample size $m = o(n^{(\epsilon+1)/2}/\sqrt{\log n})$, with $0 < \epsilon < 1$. His result is a special case of the following result due to Deheuvels, Mason and Shorack (1993). Let H_n be given by (3.77) and

$$H_{\text{BOOT-}m}(x) = P_*\{m^{1/\delta}(X_{(n)} - X_{(m)}^*) \leq x\},$$

where $X_{(m)}^*$ is the maximum of X_1^*, \dots, X_m^* that are i.i.d. from F_n .

Theorem 3.15. *Assume that $F(\theta) = 1$, $F(x) < 1$ for $x < \theta$, and (3.77) holds.*

- (i) *If $m = m_n \rightarrow \infty$ and $m_n/n \rightarrow 0$, then the bootstrap estimator $H_{\text{BOOT-}m}$ is weakly consistent, i.e., $\rho_\infty(H_{\text{BOOT-}m}, H_n) \rightarrow_p 0$.*
- (ii) *If $m_n \rightarrow \infty$ and $m_n \log \log n/n \rightarrow 0$, then the bootstrap estimator $H_{\text{BOOT-}m}$ is strongly consistent, i.e., $\rho_\infty(H_{\text{BOOT-}m}, H_n) \rightarrow_{a.s.} 0$.*

Proof. We show (i) for illustration. Since $X_{(n)} - \theta = O_p(n^{-1/\delta})$ and $m/n \rightarrow 0$, $m^{1/\delta}(X_{(n)} - \theta) = o_p(1)$. Hence, in the following we may assume that for any $\epsilon > 0$, $\theta - \epsilon m^{-1/\delta} \leq X_{(n)} \leq \theta + \epsilon m^{-1/\delta}$ for all sufficiently large n . Consider $x > 0$. Since

$$H_{\text{BOOT-}m}(x) = 1 - [P_*\{X_1^* < X_{(n)} - xm^{-1/\delta}\}]^m,$$

we obtain that for any $\epsilon < x$,

$$[F_n(\theta - x_\epsilon m^{-1/\delta})]^m \leq 1 - H_{\text{BOOT-}m}(x) \leq [F_n(\theta - x_{-\epsilon} m^{-1/\delta})]^m, \quad (3.80)$$

where $x_a = x + a$. From (3.77) and $H_n(x) = 1 - [F(\theta - xn^{-1/\delta})]^n$, we conclude that

$$\frac{1 - F(\theta - xm^{-1/\delta})}{x^\delta/m} \rightarrow 1. \quad (3.81)$$

Note that $E[F_n(\theta - xm^{-1/\delta})] = F(\theta - xm^{-1/\delta})$ and

$$\text{var}[F_n(\theta - xm^{-1/\delta})] \leq n^{-1}[1 - F(\theta - xm^{-1/\delta})] = O(m^{-1}n^{-1})$$

by (3.81). Hence, by Bernstein's inequality (e.g., Serfling, 1980, p. 95),

$$P\{|F_n(\theta - xm^{-1/\delta}) - F(\theta - xm^{-1/\delta})| > \eta/m\} \leq 2e^{-cn/m},$$

where c is a constant depending on η . This implies that for m satisfying $m \rightarrow \infty$ and $m/n \rightarrow 0$,

$$F_n(\theta - xm^{-1/\delta}) - F(\theta - xm^{-1/\delta}) = o_p(m^{-1})$$

for any fixed x and therefore

$$[F_n(\theta - (x \pm \epsilon)m^{-1/\delta})]^m - e^{-[(x \pm \epsilon)/\theta]^\delta} \rightarrow_p 0. \quad (3.82)$$

The results follow from (3.77), (3.80), and (3.82), since ϵ is arbitrary. \square

We now consider a type of nonsmooth estimator based on some tests.

Example 3.10. Estimators based on tests. We consider the simple case where X_1, \dots, X_n are i.i.d. random variables with mean μ and variance σ^2 . Similar but more complicated cases are studied in Shao (1994a). Note that a large sample 2α -level test for the hypothesis $\mu = 0$ has rejection region $|\bar{X}_n| > z_{1-\alpha}S_n$, where \bar{X}_n is the sample mean, $S_n^2 = n^{-2} \sum_{i=1}^n (X_i - \bar{X}_n)^2$, and $z_{1-\alpha} = \Phi^{-1}(1 - \alpha)$. Thus, an estimator of μ based on the test of $\mu = 0$ is defined by

$$T_n = \begin{cases} \bar{X}_n & |\bar{X}_n| > z_{1-\alpha}S_n \\ 0 & |\bar{X}_n| \leq z_{1-\alpha}S_n. \end{cases}$$

Let $H_n(x) = P\{\sqrt{n}(T_n - \mu) \leq x\}$. When $\mu \neq 0$, $P\{|\bar{X}_n| \leq z_{1-\alpha}S_n\} \rightarrow 0$ and

$$H_n(x) = P\{\sqrt{n}(\bar{X}_n - \mu) \leq x\} + o(1) \rightarrow \Phi(x/\sigma).$$

When $\mu = 0$,

$$P\{T_n = 0\} = P\{|\bar{X}_n| \leq z_{1-\alpha}S_n\} \rightarrow 1 - 2\alpha \quad (3.83)$$

and $H_n(x)$ converges weakly to a distribution $H(x)$ that is symmetric about 0 and

$$H(x) = \begin{cases} 1 - 2\alpha & 0 \leq x \leq z_{1-\alpha}\sigma \\ \Phi(x/\sigma) & x > z_{1-\alpha}\sigma. \end{cases}$$

Let X_1^*, \dots, X_m^* be i.i.d. bootstrap data from F_n and T_m^* be the bootstrap analog of T_n . If $m = n$ and $T_n = 0$,

$$P_*\{T_n^* = 0\} = \Phi(z_{1-\alpha} - \sqrt{n}\bar{X}_n/\sigma) - \Phi(z_\alpha - \sqrt{n}\bar{X}_n/\sigma) + o(1) \text{ a.s. } (3.84)$$

(Shao, 1994a). Since $\sqrt{n}\bar{X}_n$ is asymptotically nondegenerate when $\mu = 0$, the bootstrap estimator H_{BOOT} is inconsistent in view of (3.83) and (3.84).

Now suppose that $m = m_n \rightarrow \infty$ and $m_n \log \log n/n \rightarrow 0$. If $\mu \neq 0$, then, by Theorem 3.1, $P_*\{T_m^* = 0\} = P_*\{|\bar{X}_m^*| \leq z_{1-\alpha} S_m^*\} \rightarrow 0$ a.s. and therefore $H_{\text{BOOT-}m}(x) \rightarrow_{a.s.} \Phi(x/\sigma)$. Now assume that $\mu = 0$. Then,

$$\sqrt{m}|T_n| \leq \sqrt{m}|\bar{X}_n| \rightarrow_{a.s.} 0. \quad (3.85)$$

Using (3.85) and Theorem 3.1, we obtain that for $x < 0$,

$$\begin{aligned} P_*\{\sqrt{m}(T_m^* - T_n) \leq x\} &= P_*\{\sqrt{m}T_m^* \leq x\} + o(1) \quad a.s. \\ &= P_*\{\sqrt{m}\bar{X}_m^* \leq x, |\bar{X}_m^*| > z_{1-\alpha} S_m^*\} + o(1) \quad a.s. \\ &= P\{\sqrt{n}\bar{X}_n \leq x, |\bar{X}_n| > z_{1-\alpha} S_n\} + o(1) \quad a.s. \end{aligned}$$

A similar result for $x > 0$ can also be established. This proves that $H_{\text{BOOT-}m}$ is strongly consistent. Similarly, we can show the weak consistency of $H_{\text{BOOT-}m}$ if $m \rightarrow \infty$ and $m/n \rightarrow 0$.

The last example is a continuation of Example 3.2 in Section 3.1.

Example 3.2 (continued). For the sample quantile $T_n = F_n^{-1}(q)$, $0 < q < 1$, we have established the consistency of the bootstrap estimator H_{BOOT} of the distribution H_n of $\sqrt{n}(T_n - \theta)$, under the regularity condition that F is differentiable at θ and the derivative $f(\theta) > 0$. But what happens if F is not differentiable at θ or $f(\theta) = 0$?

Consider the case where $f(\theta-)$ and $f(\theta+)$ exist and are positive but $f(\theta-) \neq f(\theta+)$. In Section 3.1.4, we have shown that (3.13)-(3.15) hold and H_n has a limit given by $\Phi(x/a_x)$, which is not normal (in fact, it is not in the location-scale family). We have also shown that (3.16) and (3.17) [the bootstrap analogs of (3.13) and (3.14)] hold. However, (3.18) [the bootstrap analog of (3.15)] does not hold if $f(\theta-) \neq f(\theta+)$, which leads to the inconsistency of H_{BOOT} . The reason why (3.15) holds but (3.18) does not is that, in (3.15), $\theta + ta_t n^{-1/2}$ is always larger (smaller) than θ if $t > 0$ ($t < 0$), whereas in (3.18), $T_n + ta_t n^{-1/2}$ fluctuate around θ with $0 < \lim_{n \rightarrow \infty} P\{T_n + ta_t n^{-1/2} > \theta\} < 1$, and, therefore, the sequence $\{[F_n(T_n + ta_t n^{-1/2}) - q]/(tn^{-1/2})\}$ has no limit when $f(\theta-) \neq f(\theta+)$.

Once more it is found that the inconsistency of the bootstrap estimator is caused by the nonsmoothness of T_n and the fact that $T_n \neq \theta$. If we draw a bootstrap sample of size m_n with $m_n/n \rightarrow 0$ or $m_n \log \log n/n \rightarrow 0$, then the bootstrap analog of (3.15) holds, i.e.,

$$\frac{F_n(T_n + ta_t m^{-1/2}) - q}{tm^{-1/2}} - \sqrt{q(1-q)} \rightarrow_p 0 \quad \text{or} \quad \rightarrow_{a.s.} 0.$$

Rigorous proofs of the results in this example are given by Huang, Sen and Shao (1995). Results for the case where $f(\theta)$ exists but $f(\theta) = 0$ are also given.

3.7 Conclusions and Discussions

- (1) Like the jackknife, the use of the bootstrap does not require the theoretical derivations such as obtaining the derivatives, the influence function, the form of the asymptotic variance and the Edgeworth expansion, etc. In addition to this advantage, the bootstrap estimator of the sampling distribution of a given statistic may be more accurate than the estimator obtained using the traditional normal approximation. The use of the bootstrap requires more computations than the traditional methods. The computation of the bootstrap estimators will be discussed in Chapter 5.
- (2) For the i.i.d. case and the simple nonparametric bootstrap, the bootstrap estimators of the distributions of many commonly used statistics are consistent. For nonstandardized and nonstudentized statistics, the convergence rate of the bootstrap estimators is the same as the traditional normal approximation; however, a refined comparison of the bootstrap and the normal approximation is in favor of the former: that is, the bootstrap makes a partial correction of the estimation error and is asymptotically minimax. Also, for these types of statistics, the bootstrap is asymptotically equivalent to the one-term Edgeworth expansion method in terms of the asymptotic mean squared error. For a standardized or a studentized statistic, the bootstrap estimator and the one-term Edgeworth expansion estimator usually have the same convergence rate, which is faster than the normal approximation; however, the bootstrap is better than the one-term Edgeworth expansion in terms of the asymptotic mean squared error and the relative error in estimation. These theoretical results have been confirmed by the empirical simulation results in Section 3.4.
- (3) We introduced in Section 3.1 some basic tools used in establishing the consistency of bootstrap estimators. To study the accuracy (the convergence rate) of the bootstrap estimator and to asymptotically compare the bootstrap with other methods, the basic tools are the Berry-Esséen inequalities and the Edgeworth expansions, which are usually hard to establish when the given statistic is not simple. That is why our results for the accuracy of the bootstrap are limited to some simple statistics.
- (4) For the nonparametric bootstrap, the consistency of the bootstrap variance estimator requires more stringent moment conditions than the consistency of the bootstrap distribution estimator. This is shown in Section 3.2.2 and is supported by the simulation results in Schucany and Sheather (1989). Also, the bootstrap variance estimator is often downward-biased and is not as efficient as the jackknife variance estimator when they are consistent and require the same amount of com-

putation. Hence, the bootstrap is not recommended if one only needs a variance estimator. The bootstrap is recommended for more complicated problems such as estimating sampling distributions and constructing confidence sets. The bootstrap distribution estimator may be more accurate than the jackknife histogram (see Theorem 2.13 and Section 3.3), although the latter may be useful in some occasions (Politis and Romano, 1995).

- (5) The nonparametric bootstrap can be improved by other bootstrap methods that take into account the special nature of the model. In the i.i.d. nonparametric model, the smoothed bootstrap improves the simple nonparametric bootstrap for the problem of sample quantiles. However, when the simple nonparametric bootstrap has a good convergence rate, the effect of smoothing is not clear. This will be discussed further in later chapters.
- (6) Although bootstrap estimators are consistent for estimating the distributions of regular statistics, there are some exceptional cases. Usually the consistency of the bootstrap distribution estimator requires some smoothness conditions that are almost the same as those required for the asymptotic normality of the given statistic and certain moment conditions (see the discussion at the end of Section 3.2.1). The inconsistency of the bootstrap estimator in nonregular cases can be rectified by changing the bootstrap sample size from n to m_n , which has a slower rate of diverging to infinity than n . Some examples are given in Section 3.6, and other examples will be given in later chapters. In fact, Politis and Romano (1993a) showed that, if $m_n^2/n \rightarrow 0$, then the bootstrap distribution estimator is consistent as long as the original statistic has a limiting distribution. However, in regular cases, using $m_n < n$ may cause a loss in efficiency.

Chapter 4

Bootstrap Confidence Sets and Hypothesis Tests

In this chapter, we study the applications of the bootstrap in two main components of statistical inference: constructing confidence sets and testing hypotheses. Five different bootstrap confidence sets are introduced in Section 4.1. Asymptotic properties and asymptotic comparisons of these five bootstrap confidence sets and the confidence sets obtained by normal approximation are given in Section 4.2. More sophisticated techniques for bootstrap confidence sets have been developed in recent years, some of which are described in Section 4.3. Fixed sample comparisons of various confidence sets are made in Section 4.4 through empirical simulation studies. Bootstrap hypothesis tests are introduced in Section 4.5.

4.1 Bootstrap Confidence Sets

Let X_1, \dots, X_n be i.i.d. random p -vectors from an unknown F and $\theta = T(F)$ be a parameter of interest. If $\mathcal{C}_n = \mathcal{C}_n(X_1, \dots, X_n)$ is a subset of \mathbb{R} depending only on X_1, \dots, X_n and

$$P\{\theta \in \mathcal{C}_n\} \geq 1 - \alpha, \quad (4.1)$$

where α is a given constant satisfying $0 < \alpha < 1$, then \mathcal{C}_n is said to be a *confidence set* for θ with level $1 - \alpha$. The probability on the left-hand side of (4.1) is called the *coverage probability* of \mathcal{C}_n . If the equality in (4.1) holds, then \mathcal{C}_n is said to be a confidence set with *confidence coefficient* $1 - \alpha$ or, in short, a $1 - \alpha$ confidence set. Let $\underline{\theta} = \underline{\theta}_n(X_1, \dots, X_n)$ and $\bar{\theta} = \bar{\theta}_n(X_1, \dots, X_n)$ be two statistics. Then intervals $(-\infty, \bar{\theta}]$ and $[\underline{\theta}, \infty)$ are called one-sided confidence intervals, and $[\underline{\theta}, \bar{\theta}]$ is called a two-sided

confidence interval. $\underline{\theta}$ (or $\bar{\theta}$) is also called a lower (or upper) confidence bound. If $\underline{\theta}$ and $\bar{\theta}$ are $1 - \alpha$ lower and upper confidence bounds for θ , respectively, then $[\underline{\theta}, \bar{\theta}]$ is a $1 - 2\alpha$ *equal-tail* two-sided confidence interval for θ .

The *desired* level of a confidence set is called the *nominal level*, which is usually given. We will use $1 - \alpha$ and $1 - 2\alpha$ to denote the nominal levels of one-sided and two-sided confidence intervals, respectively. A confidence set is said to be *exact* if its confidence coefficient is exactly equal to its nominal level.

In most cases, confidence sets are constructed by considering a “pivotal quantity” $\mathfrak{R}_n = \mathfrak{R}_n(X_1, \dots, X_n, F)$ whose distribution G_n is known (independent of F). If we can deduce $\underline{\theta} \leq \theta \leq \bar{\theta}$ from the inequality $L \leq \mathfrak{R}_n \leq U$, then $[\underline{\theta}, \bar{\theta}]$ is a confidence interval with a certain level (by properly selecting L and U). For example, consider the case where θ is a location parameter. Then \mathfrak{R}_n is usually of the form $(\hat{\theta}_n - \theta)/\hat{\sigma}_n$, where $\hat{\theta}_n$ is an estimator of θ and $\hat{\sigma}_n^2$ is a variance estimator for $\hat{\theta}_n$, and an exact $1 - 2\alpha$ confidence interval for θ is

$$[\hat{\theta}_n - \hat{\sigma}_n G_n^{-1}(1 - \alpha), \hat{\theta}_n - \hat{\sigma}_n G_n^{-1}(\alpha)]. \quad (4.2)$$

To find such a pivotal quantity in a given problem is usually difficult. That is, it is not easy to find a \mathfrak{R}_n with a known distribution G_n . If G_n is unknown, then the interval in (4.2) cannot be used as a confidence interval and we have to approximate G_n . In the traditional asymptotic approach we replace the unknown G_n by its limit. Consider the interval in (4.2). If G_n has a known limit G (independent of F), then we replace G_n in (4.2) by G . If G_n has a limit G_ϑ depending on some unknown ϑ , then we replace G_n in (4.2) by $G_{\hat{\vartheta}}$, where $\hat{\vartheta}$ is a consistent estimator of ϑ . The bootstrap can be applied to obtain a confidence set simply by replacing G_n with its bootstrap estimator G_{BOOT} .

Because of the use of an approximation, the coverage probabilities of many confidence sets may be lower or higher than their nominal levels. For example, the confidence interval in (4.2) is exact when G_n is known but may not have level $1 - 2\alpha$ for fixed n when G_n is unknown and is estimated.

As we discussed in the earlier chapters, the main disadvantages of the traditional asymptotic approach are its requirement that the limit G be derived analytically and explicitly, and its lack of high order accuracy. On the other hand, the bootstrap can be used to construct easy-to-use confidence sets with higher accuracy.

We now introduce some commonly used bootstrap confidence sets. Properties of these bootstrap confidence sets are studied in Sections 4.2 and 4.4. Computational issues are discussed in Chapter 5. Because of the similarities among confidence intervals, and lower and upper confidence

bounds, we concentrate on lower confidence bounds.

4.1.1 The bootstrap-t

Throughout this chapter, $\{X_1^*, \dots, X_n^*\}$ denotes an i.i.d. sample from \hat{F} , an estimator of F (either parametric or nonparametric).

The *bootstrap-t* method (Efron, 1982) is based on a given studentized “pivot” $\mathfrak{R}_n = (\hat{\theta}_n - \theta)/\hat{\sigma}_n$, where $\hat{\theta}_n$ is an estimator of θ and $\hat{\sigma}_n^2$ is a variance estimator for $\hat{\theta}_n$. If the distribution G_n of \mathfrak{R}_n is unknown, then it is estimated by the bootstrap estimator G_{BOOT} defined by

$$G_{\text{BOOT}}(x) = P_*\{\mathfrak{R}_n^* \leq x\}, \quad (4.3)$$

where $\mathfrak{R}_n^* = (\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}_n^*$, and $\hat{\theta}_n^*$ and $\hat{\sigma}_n^*$ are bootstrap analogs of $\hat{\theta}_n$ and $\hat{\sigma}_n$, respectively. The resulting lower confidence bound for θ is

$$\underline{\theta}_{\text{BT}} = \hat{\theta}_n - \hat{\sigma}_n G_{\text{BOOT}}^{-1}(1 - \alpha), \quad (4.4)$$

which will be called the bootstrap-t lower confidence bound.

Example 4.1. Inference for location parameter. Suppose that X_1, \dots, X_n are i.i.d. random variables. Let $\theta = EX_1$ and $\hat{\theta}_n = \bar{X}_n$. Consider the studentized pivot $\mathfrak{R}_n = (\bar{X}_n - \theta)/S_n$, where $S_n^2 = \hat{\sigma}^2/n$ and $\hat{\sigma}^2$ is the usual sample variance.

(i) Consider the parametric case where $F(x) = F_0((x-\theta)/\sigma)$, $\sigma^2 = \text{var}(X_1)$, and F_0 is a known distribution function. Then the distribution of \mathfrak{R}_n does not depend on F , although its explicit form is not necessarily known. In this case, the bootstrap-t method is the same as the traditional method and produces an exact lower confidence bound for all n , although numerical computation is required if the explicit form of the distribution of \mathfrak{R}_n is unknown.

(ii) For the nonparametric case, the traditional approach uses a normal approximation to approximate the distribution of \mathfrak{R}_n . The bootstrap-t involves generating bootstrap data sets from the empirical distribution F_n and calculating $G_{\text{BOOT}}^{-1}(1 - \alpha)$. The bootstrap-t in this case provides a better solution than the traditional normal approximation (see Section 4.2).

The bootstrap-t method is simple and easy to understand. Although we show in Section 4.2 that the confidence bound $\underline{\theta}_{\text{BT}}$ is of high accuracy, it has the following two disadvantages:

- (1) To use the bootstrap-t method we need a variance estimator $\hat{\sigma}_n^2$. If no other variance estimator is available, we may use the bootstrap variance estimator v_{BOOT} for $\hat{\theta}_n$. However, if both G_{BOOT} and v_{BOOT} have to be approximated by Monte Carlo, we need to do nested boot-

strapping, i.e., we generate B_1 bootstrap data sets to approximate G_{BOOT} ; for each given bootstrap data set, we generate B_2 bootstrap data sets to approximate v_{BOOT} . Hence, the total number of bootstrap data sets used to compute $\underline{\theta}_{\text{BT}}$ is $B_1 B_2$. If $B_1 = 1000$ and $B_2 = 250$, then $B_1 B_2 = 250,000$.

- (2) The bootstrap-t method is not invariant under reparametrization. Invariance is preferred by some statisticians (see the discussion in Lehmann, 1986). However, this is not a fault of the bootstrap. Any confidence set based on a studentized pivot is not invariant.

4.1.2 The bootstrap percentile

Let $\hat{\theta}_n$ be an estimator of θ and $\hat{\theta}_n^*$ be its bootstrap analog based on X_1^*, \dots, X_n^* . Define

$$K_{\text{BOOT}}(x) = P_*\{\hat{\theta}_n^* \leq x\}. \quad (4.5)$$

The *bootstrap percentile* method (Efron, 1981a) gives the following lower confidence bound for θ :

$$\underline{\theta}_{\text{BP}} = K_{\text{BOOT}}^{-1}(\alpha). \quad (4.6)$$

The name percentile comes from the fact that $K_{\text{BOOT}}^{-1}(\alpha)$ is a percentile of the bootstrap distribution K_{BOOT} in (4.5). Apparently, this method is invariant under reparametrization.

Example 4.2. The sample median. For most cases, the computation of $\underline{\theta}_{\text{BP}}$ requires numerical approximations. However, for the sample median (more generally, sample quantiles), we can obtain an exact formula for $\underline{\theta}_{\text{BP}}$. Assume that $p = 1$ and $n = 2m - 1$ is odd. Let θ be the median of F and $X_{(i)}$ be the order statistics. Then $\hat{\theta}_n = X_{(m)}$ and $\hat{\theta}_n^* = X_{(m)}^*$ takes values $X_{(1)}, \dots, X_{(n)}$ with probabilities given by (1.17). Therefore,

$$\underline{\theta}_{\text{BP}} = X_{(\tilde{k})}$$

with \tilde{k} being the smallest k such that

$$\sum_{j=m}^n \binom{n}{j} \frac{k^j (n-k)^{n-j}}{n^n} \geq \alpha.$$

We now provide a justification of the bootstrap percentile method that allows us to see what assumptions are required for a good performance of the bootstrap percentile confidence set.

Suppose that there exists an increasing transformation $\phi_n(x)$ such that

$$P\{\hat{\phi}_n - \phi_n(\theta) \leq x\} = \Psi(x) \quad (4.7)$$

holds for all possible F (including the case $F = \hat{F}$), where $\hat{\phi}_n = \phi_n(\hat{\theta}_n)$, and Ψ is a continuous, increasing, and symmetric [$\Psi(x) = 1 - \Psi(-x)$] distribution. When $\Psi = \Phi$, the standard normal distribution, the function ϕ_n is just the normalizing and variance stabilizing transformation. Under (4.7), if ϕ_n and Ψ are known, we can obtain the following exact lower confidence bound for θ :

$$\underline{\theta}_{\text{EXACT}} = \phi_n^{-1}(\hat{\phi}_n + z_\alpha),$$

where $z_\alpha = \Psi^{-1}(\alpha)$.

We now show that $\underline{\theta}_{\text{BP}} = \underline{\theta}_{\text{EXACT}}$ and therefore we can still use this lower confidence bound even if ϕ_n and/or Ψ are unknown. Let $w_n = \phi_n(\underline{\theta}_{\text{BP}}) - \hat{\phi}_n$. From the fact that assumption (4.7) holds for $F = \hat{F}$,

$$\Psi(w_n) = P_*\{\hat{\phi}_n^* - \hat{\phi}_n \leq w_n\} = P_*\{\hat{\theta}_n^* \leq \underline{\theta}_{\text{BP}}\} = \alpha,$$

where $\hat{\phi}_n^* = \phi_n(\hat{\theta}_n^*)$ and the last equality follows from the definition of $\underline{\theta}_{\text{BP}}$ and the assumption on Ψ . Hence, $w_n = z_\alpha = \Psi^{-1}(\alpha)$ and

$$\underline{\theta}_{\text{BP}} = \phi_n^{-1}(\hat{\phi}_n + z_\alpha) = \underline{\theta}_{\text{EXACT}}.$$

We have shown that the bootstrap percentile lower confidence bound is exact for all n if assumption (4.7) holds exactly. If assumption (4.7) holds approximately for large n , then the bootstrap percentile lower confidence bound is asymptotically valid and its performance depends on how good the approximation is. Assumption (4.7) holds exactly in some situations and approximately in others. A typical example is the case where F is bivariate normal with correlation coefficient θ and $\hat{\theta}_n$ is the sample correlation coefficient (Example 2.1). Then, a normalizing and variance stabilizing transformation is $\phi(x) = \tanh^{-1}(x)$.

However, usually ϕ_n is nonlinear and the bias of $\hat{\phi}_n - \phi_n(\theta)$ does not vanish quickly as $n \rightarrow \infty$. Hence, assumption (4.7) holds approximately, but the approximation is good only when n is very large. Thus, the bootstrap percentile confidence sets are simple but may not be very accurate unless n is very large. Some improvements can be found by relaxing the assumption we have made in (4.7). This leads to the development of the following two methods that also adopt some percentiles of the bootstrap distribution K_{BOOT} as confidence bounds. Both of these two methods produce invariant confidence sets.

4.1.3 The bootstrap bias-corrected percentile

The discussion in the last section suggests that we should accommodate a bias term in assumption (4.7). Efron (1982) considered a more general

assumption:

$$P\{\hat{\phi}_n - \phi_n(\theta) + z_0 \leq x\} = \Psi(x), \quad (4.8)$$

where z_0 is a constant that may depend on F and n , $\hat{\phi}_n$ is still an increasing transformation, and Ψ is still assumed continuous, strictly increasing, and symmetric. When $z_0 = 0$, (4.8) reduces to (4.7). To see that assumption (4.8) is more general, we consider the situation where F is bivariate normal with correlation coefficient θ and $\hat{\theta}_n$ is the sample correlation coefficient. In this case, $\sqrt{n}[\tanh^{-1}(\hat{\theta}_n) - \tanh^{-1}(\theta)] - \theta/(2\sqrt{n})$ is better approximated by $N(0, 1)$ than by $\sqrt{n}[\tanh^{-1}(\hat{\theta}_n) - \tanh^{-1}(\theta)]$ when n is not very large.

If ϕ_n , z_0 , and Ψ are known, we can obtain an exact lower confidence bound

$$\underline{\theta}_{\text{EXACT}} = \phi_n^{-1}(\hat{\phi}_n + z_\alpha + z_0).$$

Applying assumption (4.8) to $F = \hat{F}$, we obtain that

$$K_{\text{BOOT}}(\hat{\theta}_n) = P_*\{\hat{\phi}_n^* - \hat{\phi}_n + z_0 \leq z_0\} = \Psi(z_0),$$

where K_{BOOT} is given in (4.5). This implies

$$z_0 = \Psi^{-1}(K_{\text{BOOT}}(\hat{\theta}_n)). \quad (4.9)$$

Also from (4.8),

$$\begin{aligned} 1 - \alpha &= \Psi(-z_\alpha) \\ &= \Psi(\hat{\phi}_n - \phi_n(\underline{\theta}_{\text{EXACT}}) + z_0) \\ &= P_*\{\hat{\phi}_n^* - \hat{\phi}_n \leq \hat{\phi}_n - \phi_n(\underline{\theta}_{\text{EXACT}})\} \\ &= P_*\{\hat{\theta}_n^* \leq \phi_n^{-1}(\hat{\phi}_n - z_\alpha - z_0)\}, \end{aligned}$$

which implies

$$\phi_n^{-1}(\hat{\phi}_n - z_\alpha - z_0) = K_{\text{BOOT}}^{-1}(1 - \alpha). \quad (4.10)$$

Since (4.10) holds for any α , it implies that for $0 < x < 1$,

$$K_{\text{BOOT}}^{-1}(x) = \phi_n^{-1}(\hat{\phi}_n + \Psi^{-1}(x) - z_0). \quad (4.11)$$

By the definition of $\underline{\theta}_{\text{EXACT}}$ and (4.11),

$$\underline{\theta}_{\text{EXACT}} = K_{\text{BOOT}}^{-1}(\Psi(z_\alpha + 2z_0)).$$

Now, assuming Ψ is known and using (4.9), we obtain the *bootstrap bias-corrected percentile* (BC) lower confidence bound for θ (Efron, 1981a):

$$\underline{\theta}_{\text{BC}} = K_{\text{BOOT}}^{-1}\left(\Psi(z_\alpha + 2\Psi^{-1}(K_{\text{BOOT}}(\hat{\theta}_n)))\right). \quad (4.12)$$

Usually, we can use $\Psi = \Phi$. Since $\Psi^{-1}(\frac{1}{2}) = 0$, $\underline{\theta}_{\text{BC}}$ reduces to $\underline{\theta}_{\text{BP}}$ if $K_{\text{BOOT}}(\hat{\theta}_n) = \frac{1}{2}$, i.e., $\hat{\theta}_n$ is the median of the bootstrap distribution K_{BOOT} . Hence, $\underline{\theta}_{\text{BC}}$ is a bias-adjusted version of the bootstrap percentile method. Again, $\underline{\theta}_{\text{BC}}$ is exact for all n if (4.8) holds exactly and is asymptotically valid if (4.8) holds approximately.

By taking the bias into account, the bootstrap BC does improve the simple bootstrap percentile method. This will be supported by the theoretical and empirical results discussed later. However, there are still many cases where assumption (4.8) cannot be fulfilled nicely and the bootstrap BC does not work well. Schenker (1985) examined the following example.

Example 4.3. Inference for variance. Suppose that F is the distribution of $N(\mu, \sigma^2)$. Let $\theta = \sigma^2$ and $\hat{\theta}_n = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. In this case, a normalizing transformation is Wilson and Hilferty's (1931) transformation

$$\phi_n(x) = \sqrt{\frac{9(n-1)}{2}} \left[\sqrt[3]{\frac{nx}{n-1}} - 1 + \frac{2}{9(n-1)} \right]. \quad (4.13)$$

For this $\phi_n(x)$, we approximately have

$$P\{\sigma^{-2/3}[\hat{\phi}_n - \phi_n(\sigma^2)] + \phi_n(1) \leq x\} = \Phi(x).$$

Although the distribution of $\hat{\theta}_n$ is transformed to a normal distribution, the variance of the transformed variable still depends on F , i.e., we are not able to achieve normalizing and variance stabilizing simultaneously. Hence, assumption (4.8) does not hold. The simulation result in Schenker (1985) shows that the coverage probability of the bootstrap BC confidence interval is substantially below its nominal level for small to moderate n .

4.1.4 The bootstrap accelerated bias-corrected percentile

In view of Schenker's example (Example 4.3), Efron (1987) introduced a more comprehensive assumption that takes into account the skewness:

$$P\left\{ \frac{\hat{\phi}_n - \phi_n(\theta)}{1 + a\phi_n(\theta)} + z_0 \leq x \right\} = \Psi(x), \quad (4.14)$$

where ϕ_n , z_0 , and Ψ are the same as those in (4.8), but a is an extra parameter (depending on F and n). According to Efron (1987), a measures how fast the standard deviation of $\hat{\phi}_n$ is changing with respect to $\phi_n(\theta)$ and therefore is called the *acceleration constant*. Clearly, assumption (4.14) is more general than (4.7) and (4.8). In Example 4.3, the transformation ϕ_n in (4.13) approximately satisfies (4.14) with $a = z_0 = [9(n-1)/2]^{-1/2}$. Some other examples are given in Efron (1987), DiCiccio and Tibshirani (1987), and Konishi (1991).

If ϕ_n , z_0 , a , and Ψ are known, then an exact confidence bound for θ is

$$\underline{\theta}_{\text{EXACT}} = \phi_n^{-1}(\hat{\phi}_n + (z_\alpha + z_0)(1 + a\hat{\phi}_n)/[1 - a(z_\alpha + z_0)]).$$

Now, like the previous discussions, we show that $\underline{\theta}_{\text{EXACT}}$ is equal to a percentile of K_{BOOT} . First, note that (4.9) still holds. Next, using assumption (4.14) and the same proof as that of (4.11), we obtain that for $0 < x < 1$,

$$K_{\text{BOOT}}^{-1}(x) = \phi_n^{-1}(\hat{\phi}_n + [\Psi^{-1}(x) - z_0](1 + a\hat{\phi}_n)). \quad (4.15)$$

Letting $x = \Psi(z_0 + (z_\alpha + z_0)/[1 - a(z_\alpha + z_0)])$ in (4.15), we obtain that

$$\underline{\theta}_{\text{BC}_a}(a) = K_{\text{BOOT}}^{-1}(\Psi(z_0 + (z_\alpha + z_0)/[1 - a(z_\alpha + z_0)])) = \underline{\theta}_{\text{EXACT}}. \quad (4.16)$$

That is, if Ψ and a are known, then $\underline{\theta}_{\text{BC}_a}(a)$ is an exact lower confidence bound for θ for all n . If we can estimate a by \hat{a} and substitute it for a in (4.16), then the resulting lower confidence bound

$$\underline{\theta}_{\text{BC}_a} = \underline{\theta}_{\text{BC}_a}(\hat{a}) \quad (4.17)$$

is the *bootstrap accelerated bias-corrected percentile* (BC_a) lower confidence bound proposed by Efron (1987).

The parameter a , however, is not easy to determine or estimate. Efron (1987) offered some methods to approximate a . We now describe the determination of a based on the summary results in DiCiccio and Romano (1988b) and Konishi (1991).

First, consider the parametric case where $F = F_\theta$, $\theta \in \mathbb{R}$. Suppose that $\text{var}(\hat{\theta}_n) = \sigma^2(\theta)/n$, and that the first three cumulants (see the definitions in Appendix A.10) of $t_n = \sqrt{n}(\hat{\theta}_n - \theta)/\sigma(\theta)$ satisfy

$$\begin{aligned} \kappa_1(t_n) &= \gamma_1(\theta)/[\sqrt{n}\sigma(\theta)] + O(n^{-1}), \\ \kappa_2(t_n) &= 1 + O(n^{-1}), \\ \kappa_3(t_n) &= \gamma_3(\theta)/[\sqrt{n}\sigma^3(\theta)] + O(n^{-1}) \end{aligned}$$

with some functions γ_j . Suppose also that the derivatives ϕ'_n and ϕ''_n are continuous. Then we have the following expansion (Pfanzagel, 1985):

$$\begin{aligned} P\left\{ \frac{\sqrt{n}[\hat{\phi}_n - \phi_n(\theta)]}{\sigma(\theta)\phi'_n(\theta)} - \frac{1}{\sqrt{n}}\left[\frac{\gamma_1(\theta)}{\sigma(\theta)} + \frac{\sigma(\theta)\phi''_n(\theta)}{2\phi'_n(\theta)} \right] \leq x \right\} \\ = \Phi(x) - \frac{(x^2 - 1)\varphi(x)}{\sqrt{n}}\left[\frac{\gamma_3(\theta)}{6\sigma^3(\theta)} + \frac{\sigma(\theta)\phi''_n(\theta)}{2\phi'_n(\theta)} \right] + O\left(\frac{1}{n}\right). \quad (4.18) \end{aligned}$$

Since ϕ_n also satisfies assumption (4.14), we obtain that

$$\sigma(\theta)\phi'_n(\theta)/\sqrt{n} = 1 + a\phi_n(\theta).$$

This means that ϕ_n has the form

$$\phi_n(x) = \frac{1}{a} \left[\exp \left(\int_0^x \frac{a\sqrt{n}}{\sigma(\theta)} d\theta \right) - 1 \right],$$

which is a composition of a variance-stabilizing transformation followed by a skewness-reducing transformation. Now, we can find an a such that

$$P \left\{ \frac{\hat{\phi}_n - \phi_n(\theta)}{1 + a\phi_n(\theta)} + z_0 \leq x \right\} = \Phi(x) + O\left(\frac{1}{n}\right),$$

i.e., (4.14) holds up to the order of $O(n^{-1})$ with $\Psi = \Phi$. This can be done by setting the coefficient of the second term on the right-hand side of (4.18) to 0, which yields that

$$a = a(\theta) = \frac{1}{\sqrt{n}} \left[\frac{d\sigma}{d\theta} - \frac{\gamma_3(\theta)}{3\sigma^3(\theta)} \right]$$

and

$$z_0 = z_0(\theta) = -\frac{1}{\sqrt{n}} \left[\frac{\gamma_1(\theta)}{\sigma(\theta)} + \frac{\sigma(\theta)\phi_n''(\theta)}{2\phi_n'(\theta)} \right] = -\frac{1}{\sqrt{n}} \left[\frac{\gamma_1(\theta)}{\sigma(\theta)} - \frac{\gamma_3(\theta)}{6\sigma^3(\theta)} \right].$$

If $\hat{\theta}_n$ is the maximum likelihood estimator of θ and $l(\theta)$ is the log-likelihood function, then

$$a = z_0 = \left[\frac{1}{6} E \frac{\partial^3 l(\theta)}{\partial \theta^3} \right] / \left[-E \frac{\partial^2 l(\theta)}{\partial \theta^2} \right]^{3/2} = \frac{1}{6} \left[\text{skewness of } \frac{\partial l(\theta)}{\partial \theta} \right],$$

which is the same as that in Efron (1987). We can then use the estimators $\hat{a} = a(\hat{\theta}_n)$ and $\hat{z}_0 = z_0(\hat{\theta}_n)$.

The same method can be used in a multiparametric model where $F = F_\eta$, $\eta = (\eta_1, \dots, \eta_p)' \in \mathbb{R}^p$, and $\theta = g(\eta)$ (DiCiccio and Romano, 1988b). We only give the result for the case where $\hat{\theta}_n = g(\hat{\eta}_n)$ and $\hat{\eta}_n$ is the maximum likelihood estimator of η . Let $l(\eta)$ be the log-likelihood function, $l_i = \partial l(\eta)/\partial \eta_i$, $\kappa_{ij} = E(l_i l_j)$, $\kappa_{ijk} = E(l_i l_j l_k)$, κ^{ij} be the (i, j) th element of the inverse of the matrix whose (i, j) th element is κ_{ij} , $g_i = \partial g(\eta)/\partial \eta_i$, and $t_i = \sum_{j=1}^p \kappa^{ij} g_j$. Then

$$a = a(\eta) = \left(\frac{1}{6} \sum_{i=1}^p \sum_{j=1}^p \sum_{k=1}^p \kappa_{ijk} t_i t_j t_k \right) / \left(\sum_{i=1}^p \sum_{j=1}^p \kappa_{ij} t_i t_j \right)^{3/2}. \quad (4.19)$$

This is also derived by Efron (1987) by reducing F_n to a least favorable family (Stein, 1956). In this case, we can take $\hat{a} = a(\hat{\eta}_n)$. Let us see some examples of the calculation of $a(\eta)$.

Example 4.4. Calculation of the acceleration constant.

(i) Variance. Consider Example 4.3, where $F = N(\mu, \sigma^2)$ and $\theta = \sigma^2$. In this case, $\eta = (\mu, \theta)'$ and $g(\eta) = \theta$. The maximum likelihood estimator of θ is $\hat{\theta}_n = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. A straightforward calculation shows that $l_1 = \partial l(\eta)/\partial \mu = \sum_{i=1}^n (X_i - \mu)/\theta$, $l_2 = \partial l(\eta)/\partial \theta = \sum_{i=1}^n (X_i - \mu)^2/(2\theta^2) - n/(2\theta)$, $\kappa_{11} = n/\theta$, $\kappa_{22} = n/(2\theta^2)$, $\kappa_{12} = 0$, $t_1 = 0$, $t_2 = 2\theta^2/n$, and $\kappa_{222} = n/\theta^3$. Then, applying (4.19), we get $a(\eta) = \sqrt{2}/(3\sqrt{n})$. In this case, a does not depend on F . The same result can be obtained if we treat μ as a known constant.

(ii) Ratio. Suppose that $(Y_i, Z_i)', i = 1, \dots, n$, are i.i.d. from a bivariate distribution with density $f(y, z) = [\Gamma(c)]^{-1} \eta_1 \eta_2^c z^{c-1} \exp[-(\eta_1 y + \eta_2 z)]$ for $y > 0$ and $z > 0$, where $c > 0$ is a constant. Let $\theta = \eta_2/\eta_1$. Then the maximum likelihood estimator of θ is \bar{Y}_n/\bar{Z}_n . It can be shown that $\kappa_{11} = n/\eta_1^2$, $\kappa_{22} = n/\eta_2^2$, $\kappa_{12} = 0$, $t_1 = -n^{-1}$, $t_2 = \eta_2^2/(\eta_1 n)$, $\kappa_{111} = n^3 E(\eta_1^{-1} - \bar{Y}_n)^3$, $\kappa_{222} = n^3 E(\eta_2^{-1} - \bar{Z}_n)^3$, and the rest of $\kappa_{ijk} = 0$. By (4.19), $a(\eta) = (c-1)/[3\sqrt{(c+1)cn}]$.

Even in a simple situation [Example 4.4(ii)], the calculation of $a(\eta)$ can be quite tedious. Some problems for the calculation of a were pointed out by Loh and Wu (1987).

Finally, we consider a nonparametric case where $\theta = T(F)$ for a functional T , $\hat{\theta}_n = T(F_n)$, and F_n is the empirical distribution (Konishi, 1991). Suppose that we have the following Edgeworth expansions:

$$P\left\{ \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(F_n)} - \frac{b(F)}{\sqrt{n}} \leq x \right\} = \Phi(x) - \frac{k(F)(x^2 - 1)\varphi(x)}{6\sqrt{n}} + O\left(\frac{1}{n}\right), \quad (4.20)$$

$$P\left\{ \frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(F)} - \frac{\tilde{b}(F)}{\sqrt{n}} \leq x \right\} = \Phi(x) - \frac{\tilde{k}(F)(x^2 - 1)\varphi(x)}{6\sqrt{n}} + O\left(\frac{1}{n}\right) \quad (4.21)$$

uniformly in x , where $\sigma(F)$, $b(F)$, $k(F)$, $\tilde{b}(F)$, and $\tilde{k}(F)$ are some functionals. Hence, (4.14) holds with $\Psi = \Phi$. Let

$$\underline{\theta} = \hat{\theta}_n - n^{-1/2} \sigma(F_n) \{-z_\alpha + n^{-1/2}[b(F) + k(F)(z_\alpha^2 - 1)/6]\}.$$

Then, by (4.20),

$$P\{\underline{\theta} \leq \theta\} = 1 - \alpha + O(n^{-1}),$$

i.e., if $b(F)$ and $k(F)$ are known, $\underline{\theta}$ is a lower confidence bound with error $O(n^{-1})$. Suppose that $a = O(n^{-1/2})$ and $z_0 = O(n^{-1/2})$. From (4.16) and Taylor's expansion,

$$\begin{aligned} K_{\text{BOOT}}(\underline{\theta}_{\text{BCa}}) &= \Phi(z_0 + (z_\alpha + z_0)/[1 - a(z_\alpha + z_0)]) \\ &= \Phi(z_\alpha) + (2z_0 + az_\alpha^2)\varphi(z_\alpha) + O(n^{-1}). \end{aligned} \quad (4.22)$$

Suppose that we select a so that $\underline{\theta}_{\text{BC}_a} = \underline{\theta} + O(n^{-3/2})$. Let

$$c_n = n^{-1/2}[b(F) + \tilde{b}(F) + k(F)(z_\alpha^2 - 1)/6 + O(n^{-1/2})].$$

Then

$$\begin{aligned} K_{\text{BOOT}}(\underline{\theta}_{\text{BC}_a}) &= P_*\{\hat{\theta}_n^* \leq \underline{\theta} + O(n^{-3/2})\} \\ &= P_*\{\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)/\sigma(F_n) - \tilde{b}(F)/\sqrt{n} \leq z_\alpha - c_n\} \\ &= P\{\sqrt{n}(\hat{\theta}_n - \theta)/\sigma(F) - \tilde{b}(F)/\sqrt{n} \leq z_\alpha - c_n\} \\ &= \Phi(z_\alpha - c_n) - \frac{\tilde{k}(F)[(z_\alpha - c_n)^2 - 1]\varphi(z_\alpha - c_n)}{6\sqrt{n}} + O\left(\frac{1}{n}\right) \\ &= \Phi(z_\alpha) - \left[c_n + \frac{\tilde{k}(F)(z_\alpha^2 - 1)}{6}\right] \frac{\varphi(z_\alpha)}{\sqrt{n}} + O\left(\frac{1}{n}\right), \end{aligned} \quad (4.23)$$

where the third equality follows from assumption (4.14), the fourth equality follows from (4.21), and the last equality follows from Taylor's expansion. Comparing (4.22) with (4.23), we obtain that

$$-\{b(F) + \tilde{b}(F) + [k(F) + \tilde{k}(F)](z_\alpha^2 - 1)/6\}/\sqrt{n} = 2z_0 + az_\alpha^2,$$

which implies

$$a = a(F) = -[k(F) + \tilde{k}(F)]/(6\sqrt{n}) \quad (4.24)$$

and

$$z_0 = z_0(F) = -[b(F) + \tilde{b}(F)]/(2\sqrt{n}) - a(F)/2.$$

In particular, if

$$\hat{\theta}_n = \theta + \frac{1}{n} \sum_{i=1}^n \phi_F(X_i) + \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \eta_F(X_i, X_j) + o_p\left(\frac{1}{n}\right),$$

then $\sigma^2(F) = \int [\phi_F(x)]^2 dF(x)$ and

$$a = a(F) = \frac{1}{6\sqrt{n}\sigma^3(F)} \int [\phi_F(x)]^3 dF(x), \quad (4.25)$$

which is similar to that in Efron (1987).

To obtain \hat{a} , we simply replace F by F_n in $a(F)$. This involves the derivation of ϕ_F .

Example 4.4 (continued). (iii) Suppose that $T(F) = g(\int x dF(x))$ for a differentiable g . Then $\phi_F(x) = \nabla g(\mu)(x - \mu)$, where $\mu = EX_1$. According to (4.25),

$$\hat{a} = a(F_n) = \left(\frac{1}{6} \sum_{i=1}^n W_i^3\right) / \left(\sum_{i=1}^n W_i^2\right)^{3/2},$$

where $W_i = \nabla g(\bar{X}_n)'(X_i - \bar{X}_n)$.

To avoid the derivation of ϕ_F , Efron (1987) suggested the use of

$$\hat{a} = \left(\frac{1}{6} \sum_{i=1}^n U_i^3 \right) / \left(\sum_{i=1}^n U_i^2 \right)^{3/2}, \quad (4.26)$$

where $U_i = [T((1-\epsilon)F_n + \epsilon\delta_{X_i}) - T(F_n)]/\epsilon$ with a small ϵ (e.g., $\epsilon = 0.001$).

Another way to estimate a and z_0 is to use the jackknife. Noting that a and z_0 are functions of the bias, variance, and skewness of $\hat{\theta}_n$, Tu and Zhang (1992b) suggested the following jackknife approximations to a and z_0 :

$$\begin{aligned} \hat{a}_{\text{JACK}} &= \frac{(n-1)^3}{6n^3 v_{\text{JACK}}^{3/2}} \sum_{i=1}^n \left(\hat{\theta}_{n-1,i} - \frac{1}{n} \sum_{j=1}^n \hat{\theta}_{n-1,j} \right)^3, \\ \hat{z}_{\text{JACK}} &= b_{\text{JACK}} / (nv_{\text{JACK}}) - sk_{\text{JACK}} / 6, \end{aligned} \quad (4.27)$$

where $\hat{\theta}_{n-1,i} = T(F_{n-1,i})$ and b_{JACK} , v_{JACK} , and sk_{JACK} are defined in (1.8), (1.13), and (3.60), respectively, with $T_n = \hat{\theta}_n$. A similar formula for a is also given by Frangos and Schucany (1990).

The bootstrap BC_a is invariant under reparametrization. Since the constructions of a and z_0 are based on considerations of achieving high order accuracy, it is expected that the bootstrap BC_a produces accurate confidence sets. Asymptotic and fixed sample properties of the bootstrap BC_a confidence sets are studied in Sections 4.2 and 4.4, respectively.

A disadvantage of the bootstrap BC_a is that the determination of a is not easy, especially when the problem under consideration is complex. Also, the improvement of the bootstrap BC_a over the simple bootstrap percentile relies on the smoothness of $\hat{\theta}_n$. Hall and Martin (1989) showed that the bootstrap percentile confidence intervals for nonsmooth estimators, such as the sample median, cannot be improved.

4.1.5 The hybrid bootstrap

As has been shown in the previous discussions, the bootstrap-t and BC_a methods provide accurate confidence sets. However, these two methods are not always practical: the bootstrap-t requires a good variance estimator and the bootstrap BC_a requires the estimation of the acceleration constant a . In this section we introduce another method that may not be as accurate as the bootstrap-t or BC_a but is convenient to use. Let

$$H_{\text{BOOT}}(x) = P_*\{n^\epsilon (\hat{\theta}_n^* - \hat{\theta}_n) \leq x\},$$

where ι is a fixed constant. From the results in Chapter 3, we know that H_{BOOT} is a good approximation to the distribution of $n^\iota(\hat{\theta}_n - \theta)$. Since

$$1 - \alpha \approx P\{n^\iota(\hat{\theta}_n - \theta) \leq H_{\text{BOOT}}^{-1}(1 - \alpha)\},$$

an approximate lower confidence bound for θ is

$$\underline{\theta}_{\text{HB}} = \hat{\theta}_n - n^{-\iota} H_{\text{BOOT}}^{-1}(1 - \alpha). \quad (4.28)$$

This method treats the percentile of the bootstrap distribution H_{BOOT} as the percentile of H_n , the distribution of $n^\iota(\hat{\theta}_n - \theta)$; hence it is called the *hybrid bootstrap* method. Since in many cases H_{BOOT} is a valid asymptotic approximation to H_n (Chapter 3), this method seems more reasonable than the bootstrap percentile. In fact, this method is used in practice more frequently than any other bootstrap method, especially when the problem under consideration is complex. Theoretical properties of this method and comparisons among the various bootstrap confidence sets are studied next.

4.2 Asymptotic Theory

The main focus of the asymptotic study of the performance of a confidence set is whether the coverage probability of the confidence set converges to the nominal level as $n \rightarrow \infty$, and how fast it converges. Other accuracy measures, such as the length of a confidence interval, are also studied sometimes. In this section, we study and compare the asymptotic properties of the five bootstrap confidence sets introduced in Section 4.1 and the confidence sets obtained by using the traditional normal approximation.

The discussion in much of Sections 4.2.2 and 4.2.3 rests not only on the Edgeworth expansions but also on the (inverse) Cornish-Fisher expansions of the quantiles of statistics' sampling distributions and of the quantiles of bootstrap distributions. Some of these expansions are given in Appendices A.10 and A.11 for some special cases. We will present these expansions without detailed derivation and proof. The reader may refer to Hall (1992d) for more details.

4.2.1 Consistency

If a confidence set is not exact, then the first concern is whether it is asymptotically exact in the following sense.

Definition 4.1. A confidence set \mathcal{C}_n for θ with nominal level $1 - \alpha$ is consistent if, as $n \rightarrow \infty$,

$$P\{\theta \in \mathcal{C}_n\} \rightarrow 1 - \alpha. \quad (4.29)$$

Again, we focus on the lower confidence bound. Let H_n and G_n be the distributions of $n^\iota(\hat{\theta}_n - \theta)$ and the studentized pivot $(\hat{\theta}_n - \theta)/\hat{\sigma}_n$, respectively, where ι is a fixed constant ($\iota = \frac{1}{2}$ in most cases), and let H_{BOOT} and G_{BOOT} be their bootstrap estimators.

- Theorem 4.1.** (i) Suppose that G_{BOOT} is consistent in the sense of Definition 3.1. Then $\underline{\theta}_{\text{BP}}$ in (4.4) is consistent.
(ii) Suppose that H_{BOOT} is consistent in the sense of Definition 3.1. Then $\underline{\theta}_{\text{HB}}$ in (4.28) is consistent.
(iii) Suppose that H_{BOOT} is consistent and

$$\lim_{n \rightarrow \infty} \rho_\infty(H_n, H) = 0 \quad (4.30)$$

for a continuous, strictly increasing, and symmetric H . Then $\underline{\theta}_{\text{BP}}$ in (4.6), $\underline{\theta}_{\text{BC}}$ in (4.12), and $\underline{\theta}_{\text{BC}_a}$ in (4.17) are consistent.

Proof. Results in (i) and (ii) follow from the consistency of G_{BOOT} and H_{BOOT} and the continuity of the functional $T(G) = G^{-1}(1-\alpha)$ with respect to ρ_∞ . The result for $\underline{\theta}_{\text{BP}}$ follows from

$$\begin{aligned} P\{\underline{\theta}_{\text{BP}} \leq \theta\} &= P\{\alpha \leq K_{\text{BOOT}}(\theta)\} \\ &= P\{\alpha \leq H_{\text{BOOT}}(n^\iota(\theta - \hat{\theta}_n))\} \\ &= P\{n^\iota(\hat{\theta}_n - \theta) \leq -H_{\text{BOOT}}^{-1}(\alpha)\} \\ &= P\{n^\iota(\hat{\theta}_n - \theta) \leq -H^{-1}(\alpha)\} + o(1) \\ &= H(-H^{-1}(\alpha)) + o(1) \rightarrow 1 - \alpha, \end{aligned}$$

where the last two equalities follow from the consistency of H_{BOOT} and (4.30). Finally, the results for $\underline{\theta}_{\text{BC}}$ and $\underline{\theta}_{\text{BC}_a}$ can be proved by noting that

$$z_0 = \Psi^{-1}(K_{\text{BOOT}}(\hat{\theta}_n)) = \Psi^{-1}(H_{\text{BOOT}}(0)) \rightarrow_p \Psi^{-1}(H(0)) = 0. \quad \square$$

Note that H in (4.30) is not the same as Ψ in assumption (4.8) or (4.14). Usually $H(x) = \Phi(x/\sigma_F)$ for some $\sigma_F > 0$, whereas $\Psi = \Phi$. Also, condition (4.30) is much weaker than assumption (4.8) or (4.14), since the latter requires variance stabilizing. Hence, the bootstrap BC and BC_a confidence sets can be consistent under a weaker assumption than (4.8) or (4.14).

It is not surprising that all five bootstrap methods produce consistent confidence sets. We will compare other asymptotic properties of these methods in Sections 4.2.2 and 4.2.3.

Theorem 4.1 can be applied to Examples 4.1 and 4.2 to show the consistency of the bootstrap confidence sets. We consider another example.

Example 4.5. Confidence sets for a distribution. Consider the problem of constructing a confidence set for $F \in \mathcal{F} = \{\text{all distributions on } \mathbb{R}^p\}$. We first introduce a metric on \mathcal{F} . Let $B_p = \{x \in \mathbb{R}^p : \|x\| = 1\}$. For any $y \in B_p$ and $t \in \mathbb{R}$, define $A(y, t) = \{x \in \mathbb{R}^p : x'y \leq t\}$, the so-called “half-space”. For any G and H in \mathcal{F} , define their distance by

$$\rho(G, H) = \sup_{y \in B_p, |t| < \infty} |P_G(A(y, t)) - P_H(A(y, t))|,$$

where P_G is the probability on \mathbb{R}^p corresponding to G .

A confidence set for F is

$$\mathcal{C}_n = \{G \in \mathcal{F} : \rho(F_n, G) \leq c\},$$

where F_n is the empirical distribution and c satisfies

$$P\{\rho(F_n, F) \leq c\} = 1 - \alpha. \quad (4.31)$$

However, it is not easy to determine c by (4.31). The exact distribution of $\rho(F_n, F)$ does not have an explicit form. Asymptotically we have

$$\sqrt{n}\rho(F_n, F) \rightarrow_d \|W\|_\infty$$

(Dudley, 1978), where $W = \{W(y, t) : y \in B_p, t \in \mathbb{R}\}$ is a Gaussian process with mean 0 and covariance function

$$E[W(y, t)W(x, s)] = P\{A(y, t) \cap A(x, s)\} - P\{A(y, t)\}P\{A(x, s)\}.$$

But the distribution function of the limit $\|W\|_\infty$ is not tractable, and hence we cannot use it to determine c .

Beran and Millar (1986) suggested the hybrid bootstrap confidence set

$$\mathcal{C}_{HB} = \{G \in \mathcal{F} : \rho(F_n, G) \leq c_n\},$$

where $c_n = n^{-1/2}H_{BOOT}^{-1}(1 - \alpha)$ and $H_{BOOT}(x) = P_*\{\sqrt{n}\rho(F_n^*, F_n) \leq x\}$. The confidence set \mathcal{C}_{HB} can be computed by Monte Carlo or by a stochastic approximation (Chapter 5). Using some results in LeCam (1983), we can show that the bootstrap estimator H_{BOOT} is consistent and, therefore, by Theorem 4.1, \mathcal{C}_{HB} is consistent. This example shows the superiority of the bootstrap over the asymptotic method for some complex problems.

From Theorem 4.1, we know that in general the consistency of the bootstrap distribution estimator implies the consistency of the bootstrap confidence sets. But the reverse may not be true. Loh (1984) gives an example.

Example 3.8 (continued). Let $X_{(n)}$ be the maximum of a sample. In Chapter 3 we showed that the bootstrap estimator of the distribution of

$n^{1/\delta}(\theta - X_{(n)})$ is inconsistent. However, in some cases the hybrid bootstrap confidence sets may still be consistent. Note that

$$P_*\{n^{1/\delta}(X_{(n)} - X_{(n)}^*) < n^{1/\delta}(X_{(n)} - X_{(n-i)})\} \rightarrow 1 - e^{-i}. \quad (4.32)$$

Suppose that $\alpha = e^{-i}$ for some positive integer i . Then, by (4.32), $q_n = n^{1/\delta}(X_{(n)} - X_{(n-i)})$ is approximately the $(1-\alpha)$ th quantile of the bootstrap distribution. Loh (1984) showed that

$$P\{\theta \leq X_{(n)} + n^{-1/\delta}q_n\} \rightarrow 1 - e^{-i}$$

if and only if $\delta = -\log(1 - e^{-1})/\log 2$. That is, the hybrid bootstrap upper confidence bound $X_{(n)} + n^{-1/\delta}q_n$ for θ is consistent if and only if $\delta = -\log(1 - e^{-1})/\log 2$. This also gives an example of an inconsistent bootstrap confidence set.

4.2.2 Accuracy

Like the consistency of a point estimator, the consistency of a confidence set is a very essential requirement. To compare several consistent confidence sets, we need to study the convergence rates of the coverage probabilities of confidence sets to the nominal level.

Definition 4.2. A confidence set \mathcal{C}_n of θ is said to be k th-order (asymptotically) accurate if

$$P\{\theta \in \mathcal{C}_n\} - (1 - \alpha) = O(n^{-k/2}).$$

The order of accuracy of bootstrap-t confidence sets was first discussed by Hall (1986). He continued his study for other types of bootstrap confidence sets in Hall (1988b). We now discuss the results in Hall (1986, 1988b) and sketch the proofs. Consider the case of $\theta = g(\mu)$, $\mu = EX_1$, $\hat{\theta}_n = g(\bar{X}_n)$, and $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. Suppose that g is continuously differentiable on \mathbb{R}^p and $\nabla g(\mu) \neq 0$. Then the asymptotic variance of $\sqrt{n}(\hat{\theta}_n - \theta)$ and its estimator are, respectively,

$$\sigma_n^2 = n^{-1} \nabla g(\mu)' \Sigma \nabla g(\mu) \quad \text{and} \quad \hat{\sigma}_n^2 = n^{-1} \nabla g(\bar{X}_n)' \hat{\Sigma} \nabla g(\bar{X}_n),$$

where $\Sigma = \text{var}(X_1)$ and $\hat{\Sigma} = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)'$. Let G_n be the distribution of the studentized pivot $(\hat{\theta}_n - \theta)/\hat{\sigma}_n$. If G_n is known, then we have an exact lower confidence bound for θ :

$$\underline{\theta}_{\text{EXACT}} = \hat{\theta}_n - \hat{\sigma}_n G_n^{-1}(1 - \alpha), \quad (4.33)$$

which is not useful if G_n is unknown. In the following, we compare the simple nonparametric bootstrap lower confidence bounds with $\underline{\theta}_{\text{EXACT}}$ in

(4.33). As consequences of these comparisons, we obtain the accuracy of related one- and two-sided confidence intervals of θ .

The bootstrap-t

The bootstrap-t lower confidence bound $\underline{\theta}_{\text{BT}}$ is given by (4.4). Suppose that G_n admits a two-term Edgeworth expansion (see Appendix A.10):

$$G_n(x) = \Phi(x) + \left[\frac{q_1(x, F)}{\sqrt{n}} + \frac{q_2(x, F)}{n} \right] \varphi(x) + o\left(\frac{1}{n}\right). \quad (4.34)$$

Then $G_n^{-1}(t)$ has an expansion of (inverse) Cornish-Fisher type on a compact interval (see Appendix A.11):

$$G_n^{-1}(t) = z(t) + \frac{q_{11}(z(t), F)}{\sqrt{n}} + \frac{q_{12}(z(t), F)}{n} + o\left(\frac{1}{n}\right), \quad (4.35)$$

where $z(t) = \Phi^{-1}(t)$ and $q_{11}(x, F) = -q_1(x, F)$.

Example 4.6. The sample mean. Suppose that $p = 1$ and $g(x) = x$. According to Bhattacharya and Ghosh (1978), expansion (4.34) holds under some conditions with $q_1(x, F) = \gamma(2x^2 + 1)/6$ and $q_2(x, F) = x[\kappa(x^2 - 3)/12 - \gamma^2(x^4 + 2x^2 - 3)/18 - (x^2 + 3)/4]$, where $\gamma = E(X_1 - \mu)^3/\sigma^3$ (skewness), $\kappa = E(X_1 - \mu)^4/\sigma^4 - 3$ (kurtosis), and $\sigma^2 = \text{var}(X_1)$. From Hall (1983a), expansion (4.35) holds with $q_{11}(x, F) = -\gamma(2x^2 + 1)/6$ and $q_{12}(x, F) = x[(x^2 + 3)/4 - \kappa(x^2 - 3)/12 + 5\gamma^2(4x^2 - 1)/72]$.

From (4.35), we can write $\underline{\theta}_{\text{EXACT}}$ as

$$\underline{\theta}_{\text{EXACT}} = \hat{\theta}_n - \hat{\sigma}_n \left[z_{1-\alpha} + \sum_{j=1}^2 \frac{q_{1j}(z_{1-\alpha}, F)}{n^{j/2}} + o\left(\frac{1}{n}\right) \right], \quad (4.36)$$

where $z_t = z(t)$. As was discussed in Chapter 3, under some conditions

$$G_{\text{BOOT}}(x) = \Phi(x) + \left[\frac{q_1(x, F_n)}{\sqrt{n}} + \frac{q_2(x, F_n)}{n} \right] \varphi(x) + o\left(\frac{1}{n}\right) \text{ a.s.}$$

[see (3.58)]. Similarly, we can obtain a Cornish-Fisher expansion:

$$G_{\text{BOOT}}^{-1}(t) = z(t) + \frac{q_{11}(z(t), F_n)}{\sqrt{n}} + \frac{q_{12}(z(t), F_n)}{n} + o\left(\frac{1}{n}\right) \text{ a.s.}$$

Hence, the bootstrap-t lower confidence bound can be written as

$$\underline{\theta}_{\text{BT}} = \hat{\theta}_n - \hat{\sigma}_n \left[z_{1-\alpha} + \sum_{j=1}^2 \frac{q_{1j}(z_{1-\alpha}, F_n)}{n^{j/2}} + o\left(\frac{1}{n}\right) \right] \text{ a.s.} \quad (4.37)$$

Under some moment conditions, we have that for each x ,

$$q_{1j}(x, F_n) - q_{1j}(x, F) = O_p(n^{-1/2}), \quad j = 1, 2.$$

Then, comparing (4.36) with (4.37), we obtain that

$$\underline{\theta}_{\text{BT}} - \underline{\theta}_{\text{EXACT}} = O_p(n^{-3/2}), \quad (4.38)$$

since $\hat{\sigma}_n = O_p(n^{-1/2})$. Furthermore,

$$\begin{aligned} P\{\underline{\theta}_{\text{BT}} \leq \theta\} &= P\left\{\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \leq G_{\text{BOOT}}^{-1}(1 - \alpha)\right\} \\ &= P\left\{\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \leq z_{1-\alpha} + \sum_{j=1}^2 \frac{q_{1j}(z_{1-\alpha}, F_n)}{n^{j/2}}\right\} + o\left(\frac{1}{n}\right) \\ &= P\left\{\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \leq z_{1-\alpha} + \sum_{j=1}^2 \frac{q_{1j}(z_{1-\alpha}, F)}{n^{j/2}}\right\} + \frac{\psi(z_{1-\alpha})}{n} + o\left(\frac{1}{n}\right) \\ &= 1 - \alpha + \frac{\psi(z_{1-\alpha})\varphi(z_{1-\alpha})}{n} + o\left(\frac{1}{n}\right), \end{aligned} \quad (4.39)$$

where the last two equalities can be justified by a somewhat complicated argument (Hall, 1983a, 1986) and where $\psi(x)$ is a polynomial whose coefficients are functions of moments of F . In Example 4.6,

$$\psi(x) = x(1 + 2x^2)(\kappa - 3\gamma^2/2)/6.$$

Result (4.39) implies that $[\underline{\theta}_{\text{BT}}, \infty)$ is second order accurate in the sense of Definition 4.2. The same can be concluded for the bootstrap-t upper confidence bound and the equal-tail two-sided bootstrap-t confidence interval for θ .

The hybrid bootstrap

The hybrid bootstrap lower confidence bound $\underline{\theta}_{\text{HB}}$ is given by (4.28). Let \tilde{H}_{BOOT} be the bootstrap estimator of \tilde{H}_n , the distribution of the standardized variable $(\hat{\theta}_n - \theta)/\sigma_n$. Then $n^{-1/2}H_{\text{BOOT}}^{-1}(1 - \alpha) = \hat{\sigma}_n \tilde{H}_{\text{BOOT}}^{-1}(1 - \alpha)$ and

$$\underline{\theta}_{\text{HB}} = \hat{\theta}_n - \hat{\sigma}_n \tilde{H}_{\text{BOOT}}^{-1}(1 - \alpha),$$

which can be viewed as a bootstrap estimator of

$$\underline{\theta}_h = \hat{\theta}_n - \hat{\sigma}_n \tilde{H}_n^{-1}(1 - \alpha).$$

Note that $\underline{\theta}_h$ is not exact even if \tilde{H}_n is known, since it is obtained by muddling up $G_n^{-1}(1 - \alpha)$ and $\tilde{H}_n^{-1}(1 - \alpha)$. Similar to G_n , under some

conditions \tilde{H}_n admits an Edgeworth expansion

$$\tilde{H}_n(x) = \Phi(x) + \left[\frac{\tilde{q}_1(x, F)}{\sqrt{n}} + \frac{\tilde{q}_2(x, F)}{n} \right] \varphi(x) + o\left(\frac{1}{n}\right) \quad (4.40)$$

and a Cornish-Fisher expansion

$$\tilde{H}_n^{-1}(t) = z(t) + \frac{\tilde{q}_{11}(z(t), F)}{\sqrt{n}} + \frac{\tilde{q}_{12}(z(t), F)}{n} + o\left(\frac{1}{n}\right).$$

Example 4.6 (continued). In this case, $\tilde{q}_1(x, F) = \gamma(1-x^2)/6$, $\tilde{q}_2(x, F) = -x[\kappa(x^2-3)/24 + \gamma^2(x^4-10x^2+15)/72]$, $\tilde{q}_{11}(x, F) = -\tilde{q}_1(x, F)$, and $\tilde{q}_{12}(x, F) = x[\kappa(x^2-3)/24 - \gamma^2(2x^2-5)/36]$.

Similar expansions exist for \tilde{H}_{BOOT} :

$$\tilde{H}_{\text{BOOT}}(x) = \Phi(x) + \left[\frac{\tilde{q}_1(x, F_n)}{\sqrt{n}} + \frac{\tilde{q}_2(x, F_n)}{n} \right] \varphi(x) + o\left(\frac{1}{n}\right) \text{ a.s.} \quad (4.41)$$

and

$$\tilde{H}_{\text{BOOT}}^{-1}(t) = z(t) + \frac{\tilde{q}_{11}(z(t), F_n)}{\sqrt{n}} + \frac{\tilde{q}_{12}(z(t), F_n)}{n} + o\left(\frac{1}{n}\right) \text{ a.s.} \quad (4.42)$$

Thus, we obtain that

$$\underline{\theta}_{\text{HB}} = \hat{\theta}_n - \hat{\sigma}_n \left[z_{1-\alpha} + \sum_{j=1}^2 \frac{\tilde{q}_{1j}(z_{1-\alpha}, F_n)}{n^{j/2}} + o\left(\frac{1}{n}\right) \right] \text{ a.s.} \quad (4.43)$$

Comparing (4.36) with (4.43), we find that

$$\underline{\theta}_{\text{HB}} - \underline{\theta}_{\text{EXACT}} = O_p(n^{-1}), \quad (4.44)$$

since $q_{11}(x, F)$ and $\tilde{q}_{11}(x, F)$ are usually different. Results (4.38) and (4.44) imply that $\underline{\theta}_{\text{HB}}$ is not as close to $\underline{\theta}_{\text{EXACT}}$ as $\underline{\theta}_{\text{BT}}$. Similar to (4.39), we can show that

$$\begin{aligned} P\{\underline{\theta}_{\text{HB}} \leq \theta\} &= P\left\{ \frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \leq \tilde{H}_{\text{BOOT}}^{-1}(1-\alpha) \right\} \\ &= P\left\{ \frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \leq z_{1-\alpha} + \sum_{j=1}^2 \frac{\tilde{q}_{1j}(z_{1-\alpha}, F_n)}{n^{j/2}} \right\} + o\left(\frac{1}{n}\right) \\ &= P\left\{ \frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \leq z_{1-\alpha} + \sum_{j=1}^2 \frac{q_{1j}(z_{1-\alpha}, F)}{n^{j/2}} \right\} + O\left(\frac{1}{n}\right) \\ &= 1 - \alpha + \frac{\tilde{\psi}(z_{1-\alpha})\varphi(z_{1-\alpha})}{\sqrt{n}} + O\left(\frac{1}{n}\right), \end{aligned} \quad (4.45)$$

where $\tilde{\psi}(x)$ is an even polynomial [$\tilde{\psi}(x) = \gamma x^2/2$ in Example 4.6]. This implies that when $\tilde{\psi} \neq 0$, $[\underline{\theta}_{\text{HB}}, \infty)$ is only first order accurate.

However, the equal-tail two-sided hybrid bootstrap confidence interval

$$[\underline{\theta}_{\text{HB}}, \bar{\theta}_{\text{HB}}] = [\hat{\theta}_n - \hat{\sigma}_n \tilde{H}_{\text{BOOT}}^{-1}(1 - \alpha), \hat{\theta}_n - \hat{\sigma}_n \tilde{H}_{\text{BOOT}}^{-1}(\alpha)]$$

is second order accurate, as is the equal-tail two-sided bootstrap-t confidence interval, since

$$\begin{aligned} P\{\underline{\theta}_{\text{HB}} \leq \theta \leq \bar{\theta}_{\text{HB}}\} &= P\{\theta \leq \bar{\theta}_{\text{HB}}\} - P\{\theta < \underline{\theta}_{\text{HB}}\} \\ &= 1 - \alpha + n^{-1/2} \tilde{\psi}(z_{1-\alpha}) \varphi(z_{1-\alpha}) \\ &\quad - \alpha - n^{-1/2} \tilde{\psi}(z_\alpha) \varphi(z_\alpha) + O(n^{-1}) \\ &= 1 - 2\alpha + O(n^{-1}) \end{aligned}$$

by the fact that $\tilde{\psi}$ and φ are even functions and $z_{1-\alpha} = -z_\alpha$.

The bootstrap percentile

Let K_{BOOT} be given by (4.5). Then

$$\underline{\theta}_{\text{BP}} = K_{\text{BOOT}}^{-1}(\alpha) = \hat{\theta}_n + \hat{\sigma}_n \tilde{H}_{\text{BOOT}}^{-1}(\alpha).$$

Comparing $\underline{\theta}_{\text{BP}}$ with $\underline{\theta}_{\text{BT}}$ and $\underline{\theta}_{\text{HB}}$, we find that the bootstrap percentile method muddles up not only \tilde{H}_{BOOT} and G_{BOOT} , but also $\tilde{H}_{\text{BOOT}}^{-1}(\alpha)$ and $-\tilde{H}_{\text{BOOT}}^{-1}(1 - \alpha)$. If \tilde{H}_{BOOT} is asymptotically symmetric, then the bootstrap percentile is equivalent to the hybrid bootstrap. Thus, one-sided bootstrap percentile confidence intervals are only first order accurate. However, the equal-tail two-sided bootstrap percentile confidence interval is second order accurate, as is the equal-tail two-sided hybrid bootstrap confidence interval.

The bootstrap bias-corrected percentile

Since $\hat{\theta}_n$ is asymptotically normal, we can use $\Psi = \Phi$ for the bootstrap BC and BC_a methods. Let

$$\tilde{\alpha}_n = \Phi(z_\alpha + 2z_0).$$

Then the bootstrap BC lower confidence bound $\underline{\theta}_{\text{BC}}$ given by (4.12) is just the $\tilde{\alpha}_n$ th quantile of K_{BOOT} . Since, by (4.41),

$$K_{\text{BOOT}}(\hat{\theta}_n) = \tilde{H}_{\text{BOOT}}(0) = \Phi(0) + \frac{\tilde{q}_1(0, F_n)}{\sqrt{n}} \varphi(0) + O_p\left(\frac{1}{n}\right),$$

we have

$$z_0 = \Phi^{-1}(K_{\text{BOOT}}(\hat{\theta}_n)) = \frac{\tilde{q}_1(0, F_n)}{\sqrt{n}} + O_p\left(\frac{1}{n}\right) \quad (4.46)$$

and

$$\tilde{\alpha}_n = \alpha + \frac{2\tilde{q}_1(0, F_n)}{\sqrt{n}} \varphi(z_\alpha) + O_p\left(\frac{1}{n}\right). \quad (4.47)$$

From (4.42) and (4.47),

$$\begin{aligned} \tilde{H}_{\text{BOOT}}^{-1}(\tilde{\alpha}_n) &= z(\tilde{\alpha}_n) + \frac{\tilde{q}_{11}(z(\tilde{\alpha}_n), F_n)}{\sqrt{n}} \varphi(z(\tilde{\alpha}_n)) + O_p\left(\frac{1}{n}\right) \\ &= z_\alpha + \frac{2\tilde{q}_1(0, F_n) + \tilde{q}_{11}(z_\alpha, F_n)}{\sqrt{n}} + O_p\left(\frac{1}{n}\right). \end{aligned}$$

Then, from (4.12) and $K_{\text{BOOT}}^{-1}(\tilde{\alpha}_n) = \hat{\theta}_n + \hat{\sigma}_n \tilde{H}_{\text{BOOT}}^{-1}(\tilde{\alpha}_n)$,

$$\underline{\theta}_{\text{BC}} = \hat{\theta}_n + \hat{\sigma}_n \left[z_\alpha + \frac{2\tilde{q}_1(0, F_n) + \tilde{q}_{11}(z_\alpha, F_n)}{\sqrt{n}} + O_p\left(\frac{1}{n}\right) \right]. \quad (4.48)$$

Comparing (4.36) with (4.48), we conclude that

$$\underline{\theta}_{\text{BC}} - \underline{\theta}_{\text{EXACT}} = O_p(n^{-1}), \quad (4.49)$$

which is similar to (4.44). It also follows from (4.48) that

$$P\{\underline{\theta}_{\text{BC}} \leq \theta\} = 1 - \alpha + \frac{\bar{\psi}(z_{1-\alpha})\varphi(z_{1-\alpha})}{\sqrt{n}} + O\left(\frac{1}{n}\right) \quad (4.50)$$

with an even polynomial $\bar{\psi}(x)$ [$\bar{\psi}(x) = \gamma(x^2 + 2)/6$ in Example 4.6]. Hence, $[\underline{\theta}_{\text{BC}}, \infty)$ is first order accurate in general. In fact,

$$\underline{\theta}_{\text{BC}} - \underline{\theta}_{\text{BP}} = 2\tilde{q}_1(0, F_n)\hat{\sigma}_n n^{-1/2} + O_p(n^{-3/2}),$$

and therefore the bootstrap BC and percentile methods have the same performance in terms of their asymptotic order. The bootstrap BC, however, is a partial improvement over the simple percentile method in the sense that $\bar{\psi}(z_{1-\alpha})$ in (4.50) is smaller than $\bar{\psi}(z_{1-\alpha})$ in (4.45) in absolute value.

The bootstrap accelerated bias-corrected percentile

Note that (4.20) and (4.21) are alternative forms of (4.34) and (4.40), respectively. Therefore,

$$q_1(x, F) = b(F) + k(F)(x^2 - 1)/6$$

and

$$\tilde{q}_1(x, F) = \tilde{b}(F) + \tilde{k}(F)(x^2 - 1)/6.$$

Also, since $q_1(0, F) = \tilde{q}_1(0, F)$ (e.g., both of them are $\gamma/6$ in Example 4.6), we have by (4.24),

$$\hat{a} = a(F_n) = [q_1(z_\alpha, F_n) + \tilde{q}_1(z_\alpha, F_n) - 2q_1(0, F_n)]/(\sqrt{n}z_\alpha^2). \quad (4.51)$$

From (4.16), $\underline{\theta}_{\text{BC}_a}$ is the $\tilde{\alpha}_n(\hat{a})$ th quantile of K_{BOOT} with

$$\begin{aligned}\tilde{\alpha}_n(\hat{a}) &= \Phi(z_0 + (z_\alpha + z_0)/[1 - \hat{a}(z_\alpha + z_0)]) \\ &= \Phi(z_\alpha + 2z_0 + \hat{a}z_\alpha^2 + O_p(n^{-1})).\end{aligned}$$

From (4.42),

$$\tilde{H}_{\text{BOOT}}^{-1}(\tilde{\alpha}_n(\hat{a})) = z(\alpha) + 2z_0 + \hat{a}z_\alpha^2 - n^{-1/2}\tilde{q}_1(z_\alpha, F_n) + O_p(n^{-1}).$$

Hence,

$$\begin{aligned}\underline{\theta}_{\text{BC}_a} &= K_{\text{BOOT}}(\tilde{\alpha}_n(\hat{a})) = \hat{\theta}_n + \hat{\sigma}_n \tilde{H}_{\text{BOOT}}^{-1}(\tilde{\alpha}_n(\hat{a})) \\ &= \hat{\theta}_n + \hat{\sigma}_n [z_\alpha + 2z_0 + \hat{a}z_\alpha^2 - n^{-1/2}\tilde{q}_1(z_\alpha, F_n) + O_p(n^{-1})] \\ &= \hat{\theta}_n + \hat{\sigma}_n [z_\alpha + n^{-1/2}q_1(z_\alpha, F_n) + O_p(n^{-1})],\end{aligned}$$

where the last equality follows from (4.46) and (4.51). Comparing this result with (4.37), we conclude that

$$\underline{\theta}_{\text{BC}_a} - \underline{\theta}_{\text{BT}} = O_p(n^{-3/2}). \quad (4.52)$$

Hence, the bootstrap BC_a and the bootstrap-t confidence sets have the same asymptotic performance, and they are second order accurate.

The normal approximation

The traditional method is to use a normal approximation, which produces the following lower confidence bound for θ :

$$\underline{\theta}_{\text{NOR}} = \hat{\theta}_n - \hat{\sigma}_n z_{1-\alpha}. \quad (4.53)$$

From (4.43) and (4.53), we obtain that

$$\underline{\theta}_{\text{NOR}} - \underline{\theta}_{\text{HB}} = O_p(n^{-1}). \quad (4.54)$$

Hence, $[\underline{\theta}_{\text{NOR}}, \infty)$ has the same order of accuracy as $[\underline{\theta}_{\text{HB}}, \infty)$.

Summary and discussion

For the case of $\hat{\theta}_n = g(\bar{X}_n)$ with a smooth function, we have shown the following results:

- (1) The bootstrap-t and bootstrap BC_a one-sided confidence intervals are second order accurate, whereas the one-sided confidence intervals produced by the bootstrap percentile, bootstrap BC, hybrid bootstrap, and normal approximation are in general first order accurate.
- (2) The equal-tail two-sided confidence intervals produced by all five bootstrap methods and the normal approximation are second order accurate.

The results indicate that, in terms of the accuracy of one-sided confidence intervals, the bootstrap-t and bootstrap BC_a are better than the bootstrap percentile, bootstrap BC, hybrid bootstrap, and normal approximation. However, since the use of the bootstrap-t and bootstrap BC_a is limited (the former requires a good variance estimator whereas the latter requires a good estimator of the acceleration parameter), the other three bootstrap methods are still useful, in spite of their lack of high accuracy.

In addition to the Edgeworth expansions for the sampling distribution of standardized or studentized statistics and their bootstrap estimators, the Cornish-Fisher expansions play a crucial role in the study of the accuracy of bootstrap confidence sets. This limits the result to the cases where these expansions exist. For the bootstrap-t and bootstrap BC_a, the forms of the variance and acceleration parameter estimators also have effects on the properties of the resulting confidence sets, which makes the theoretical study more complicated. Götze (1989) proved the second order accuracy of the bootstrap-t confidence sets based on a class of statistics expressed as a function of $n^{-1/2} \sum_{i=1}^n h(X_i)$, where h is a known function. Helmers (1991) showed that the bootstrap-t confidence sets are second order accurate for the mean of U-statistics.

For nonsmooth statistics such as the sample quantiles, the story is quite different: *both* one-sided and two-sided confidence intervals given by the bootstrap percentile or hybrid bootstrap method are only first order accurate (Falk and Kaufman, 1991). Hall and Martin (1989) proved that the bootstrap BC_a confidence bounds are still first order accurate, and they therefore show no improvement over the bootstrap percentile confidence bounds.

Babu and Bose (1989) introduced another approach to study the accuracy of the bootstrap confidence sets. Let G_n be the distribution of a pivot \mathfrak{R}_n and G_{BOOT} be the bootstrap estimator of G_n . They showed that for a positive sequence of numbers $\{\epsilon_n\}$, if

$$P\{\rho_\infty(G_n, G_{\text{BOOT}}) \geq \epsilon_n\} \leq \epsilon_n,$$

then

$$\sup_{0 \leq \alpha \leq 1} |P\{\mathfrak{R}_n < G_{\text{BOOT}}^{-1}(\alpha)\} - \alpha| \leq 2\epsilon_n + j(G_n),$$

where

$$j(G_n) = \sup_x [G_n(x) - G_n(x-)].$$

That is, the order of accuracy of the bootstrap confidence set depends on the accuracy of the bootstrap distribution estimator and the size of the jumps in G_n . If we apply this result to $\mathfrak{R}_n = \sqrt{n}[g(\bar{X}_n) - g(\theta)]/\hat{\sigma}_n$, then under some conditions we obtain that

$$P\{\mathfrak{R}_n \leq G_{\text{BOOT}}^{-1}(1 - \alpha)\} = 1 - \alpha + O(n^{-1}(\log n)^{3/2}),$$

which is weaker than the result in Hall (1986). However, this approach requires fewer regularity conditions and may be applied to more general situations.

The order of accuracy of the bootstrap confidence sets can be improved further using the iterative bootstrap introduced in Section 4.4.

4.2.3 Other asymptotic comparisons

In terms of the order of accuracy of confidence sets, we have shown that the bootstrap-t and bootstrap BC_a are better than the other bootstrap methods and the normal approximation. Here we make some other asymptotic comparisons.

Asymptotic minimaxity

A confidence set can be viewed as a set-valued estimator. Therefore, we can define a risk function for a confidence set. Suppose that we are interested in constructing a confidence set for a parameter $\theta = \theta(F) \in \Theta$, where Θ can be an abstract space (e.g., $\Theta = \mathcal{F}$ in Example 4.5), although in most applications $\Theta = \mathbb{R}$ or \mathbb{R}^d . Let $\|\cdot\|$ be a norm on Θ ($\|\cdot\|$ is the Euclidean norm if $\Theta = \mathbb{R}^d$). Let

$$C(\hat{z}_n, \hat{r}_n) = \{x \in \Theta : \|x - \hat{z}_n\| \leq \hat{r}_n\}$$

be a confidence set for θ with level $1 - \alpha$, where \hat{z}_n and \hat{r}_n are functions of X_1, \dots, X_n . For example, if $\Theta = \mathbb{R}$, then $C(\hat{z}_n, \hat{r}_n) = [\hat{z}_n - \hat{r}_n, \hat{z}_n + \hat{r}_n]$ is a confidence interval centered at \hat{z}_n with radius (half-length) \hat{r}_n . Let \mathcal{A}_α be the collection of all such confidence sets with level $1 - \alpha$. For each $C \in \mathcal{A}_\alpha$, we define its risk function by

$$R_n(C, \theta) = E[L(\sqrt{n} \sup_{y \in C} \|y - \theta\|)],$$

where L is a bounded and increasing loss function on $[0, \infty)$. When $\Theta = \mathbb{R}$, it can be easily shown that

$$R_n(C, \theta) = E[L(\sqrt{n}(|\hat{z}_n - \theta| + \hat{r}_n))],$$

which takes the miscenterery of C and the length of C into account simultaneously.

Beran and Millar (1985) considered the hybrid bootstrap confidence set

$$C_{\text{BOOT}} = C(\theta(F_n), \tilde{r}_n),$$

where F_n is the empirical distribution and \tilde{r}_n is the $(1 - \alpha)$ th quantile of the bootstrap distribution of $\|\theta(F_n^*) - \theta(F_n)\|$. Beran and Millar (1985)

proved that under some conditions,

$$\lim_{t \uparrow \infty} \lim_{n \rightarrow \infty} \sup_{G \in N(t,n)} R_n(C_{\text{BOOT}}, \theta(G)) = \lim_{t \uparrow \infty} \lim_{n \rightarrow \infty} \inf_{C \in \mathcal{A}_\alpha} \sup_{G \in N(t,n)} R_n(C, \theta(G)),$$

where $N(t, n) = \{G : \rho(G, F) \leq ta_n\}$ for a sequence $\{a_n\}$ satisfying $a_n \geq \sqrt{n}$ and $a_n \rightarrow 0$. This means that the hybrid bootstrap confidence set is asymptotically minimax. In particular, letting $\Theta = \mathbb{R}$, we conclude that the two-sided hybrid bootstrap confidence interval is asymptotically minimax.

Comparison of first order accurate confidence bounds

The one-sided confidence intervals produced by the hybrid bootstrap, bootstrap BC, and normal approximation are first order accurate. We can further compare these confidence sets in terms of their errors in coverage probability in a special case (Liu and Singh, 1987).

Example 4.6 (continued). Under some conditions on F , we have the following expansions (see the discussion in Section 4.2.2):

$$P\{\underline{\theta}_{\text{HB}} \leq \theta\} = 1 - \alpha + \frac{\gamma z_\alpha^2 \varphi(z_\alpha)}{2\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right),$$

$$P\{\underline{\theta}_{\text{BC}} \leq \theta\} = 1 - \alpha + \frac{\gamma(z_\alpha^2 + 2)}{6\sqrt{n}} \varphi(z_\alpha) + o\left(\frac{1}{\sqrt{n}}\right),$$

and

$$P\{\underline{\theta}_{\text{NOR}} \leq \theta\} = 1 - \alpha + \frac{\gamma(2z_\alpha^2 + 1)\varphi(z_\alpha)}{6\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right).$$

Let

$$e(\underline{\theta}) = P\{\underline{\theta} \leq \theta\} - (1 - \alpha)$$

be the error in coverage probability for the lower confidence bound $\underline{\theta}$. Then

$$|e(\underline{\theta}_{\text{HB}})| = |e(\underline{\theta}_{\text{NOR}})| + C_n(z_\alpha, F) + o(n^{-1/2})$$

and

$$|e(\underline{\theta}_{\text{NOR}})| = |e(\underline{\theta}_{\text{BC}})| + C_n(z_\alpha, F) + o(n^{-1/2}),$$

where $C_n(x, F) = |\gamma|(x^2 - 1)\varphi(x)/(6n^{1/2})$. Assume that $\gamma \neq 0$. When $z_\alpha^2 > 1$, which is usually the case in practice, $C_n(z_\alpha, F) > 0$, and therefore the bootstrap BC is better than the normal approximation, which is better than the hybrid bootstrap, in terms of absolute error in coverage probability.

The normal approximation method needs a variance estimator [see (4.53)], which is not required by the hybrid bootstrap and bootstrap BC.

When a variance estimator is available, we can use the bootstrap-t lower confidence bound, which is second order accurate.

Comparison of lengths

For equal-tail two-sided confidence intervals, it has been shown that all of the confidence intervals discussed so far are second order accurate. Another important criterion to assess a two-sided confidence interval is the shortness of the interval. Let $\underline{\theta}$ and $\bar{\theta}$ be $1 - \alpha$ lower and upper confidence bounds, respectively. Then, the $1 - 2\alpha$ equal-tail two-sided confidence interval $[\underline{\theta}, \bar{\theta}]$ has length $l(\alpha) = \bar{\theta} - \underline{\theta}$. Suppose that $\underline{\theta}$ and $\bar{\theta}$ are obtained by using one of the bootstrap methods defined in Section 4.1 or the normal approximation and that $\theta = g(\mu)$ and $\hat{\theta}_n = g(\bar{X}_n)$. Using asymptotic expansions previously discussed, we have

$$\underline{\theta} = \hat{\theta}_n + \hat{\sigma}_n \left[z_\alpha + \sum_{j=1}^3 \frac{h_j(z_\alpha, F_n)}{n^{j/2}} + O\left(\frac{1}{n^2}\right) \right]$$

and

$$\bar{\theta} = \hat{\theta}_n + \hat{\sigma}_n \left[z_{1-\alpha} + \sum_{j=1}^3 \frac{h_j(z_{1-\alpha}, F_n)}{n^{j/2}} + O\left(\frac{1}{n^2}\right) \right],$$

where h_j are some functions. Since h_1 and h_3 are even functions and h_2 is odd,

$$l(\alpha) = 2\hat{\sigma}_n \left[z_{1-\alpha} + \frac{h_2(z_{1-\alpha}, F_n)}{n} + O\left(\frac{1}{n^2}\right) \right].$$

Thus, the length of an equal-tail two-sided confidence interval is determined by the function h_2 . It is, however, very hard to derive the function h_2 in general. Hall (1988b) considered an example.

Example 4.6 (continued). For the six different methods we have been considering, Hall (1988b) tabulated the values of $h_2(z_{1-\alpha}, F)$ in terms of the skewness γ and the kurtosis κ , for $\alpha = 0.025$.

From Table 4.1, we know that the lengths of the confidence intervals depend on the skewness and kurtosis of the underlying distribution. When $\gamma = \kappa = 0$, all intervals, except the bootstrap-t confidence interval, have the same length up to the order of $O(n^{-5/2})$. Although the bootstrap-t confidence interval has high accuracy, it tends to be longer than other intervals. This is confirmed by the empirical results in Section 4.4.

Hall (1988b) suggested a bootstrap-t confidence interval having the shortest length, but not equal-tail. Let G_{BOOT} be the bootstrap distribution defined in (4.3). We choose a pair (\hat{x}, \hat{y}) satisfying

$$G_{\text{BOOT}}(\hat{x}) \geq 1 - \alpha, \quad G_{\text{BOOT}}(\hat{x}) - G_{\text{BOOT}}(-\hat{y}) \geq 1 - 2\alpha$$

Table 4.1. Values of $h_2(z_{0.975}, F)$ [Adapted from Hall (1988b),
by permission of Institute of Mathematical Statistics]

Method	$h_2(z_{0.975}, F)$
Bootstrap-t	$-0.14\kappa + 1.96\gamma^2 + 3.35$
Hybrid bootstrap	$0.069\kappa - 0.15\gamma^2$
Bootstrap percentile	$0.069\kappa - 0.15\gamma^2$
Bootstrap BC	$0.069\kappa + 0.072\gamma^2$
Bootstrap BC _a	$0.069\kappa + 0.81\gamma^2$
Normal approximation	0

and $\hat{x} + \hat{y} = \min(x + y)$. Then the shortest approximate $1 - 2\alpha$ bootstrap-t confidence interval is

$$[\hat{\theta}_n - \hat{\sigma}_n \hat{y}, \hat{\theta}_n + \hat{\sigma}_n \hat{x}]. \quad (4.55)$$

Hall (1988b) showed that the length of this confidence interval is shorter than the equal-tail two-sided bootstrap-t confidence interval by an amount of order $n^{-3/2}$. In Example 4.6, the difference between the lengths of the equal-tail two-sided bootstrap-t confidence interval and interval (4.55) is $\frac{4}{9}\sigma\gamma^2 z_{1-\alpha} n^{-3/2} + O(n^{-5/2})$. Surprisingly, it can also be shown that the coverage probability of interval (4.55) is closer to the nominal level than that of the equal-tail two-sided bootstrap-t confidence interval in this example.

4.3 The Iterative Bootstrap and Other Methods

We have shown that, for one-sided confidence intervals, the bootstrap percentile, bootstrap BC, and hybrid bootstrap are first order accurate and the bootstrap-t and bootstrap BC_a are second order accurate. In view of the existing empirical results showing that second order accurate confidence sets are sometimes still not accurate enough (e.g., see Section 4.4), one might ask whether we can obtain a confidence set with an even higher order of accuracy. The answer is affirmative if one has the resources for more intensive computations. In this section we introduce two important methods, the iterative bootstrap and the bootstrap calibration, by which we can obtain confidence sets of any order of accuracy (at least in theory). Some other modified and special purpose bootstrap confidence sets are also discussed in this section.

4.3.1 The iterative bootstrap

The hybrid bootstrap and the bootstrap-t are based on the bootstrap distribution estimators for $n'(\hat{\theta}_n - \theta)$ and $(\hat{\theta}_n - \theta)/\hat{\sigma}_n$, respectively. Beran

(1987) argued that the reason why the bootstrap-t is better than the hybrid bootstrap is that $(\hat{\theta}_n - \theta)/\hat{\sigma}_n$ would be a better pivot than $n^t(\hat{\theta}_n - \theta)$ in the sense that the distribution of $(\hat{\theta}_n - \theta)/\hat{\sigma}_n$ is less dependent on the unknown F . For example, under the parametric model in Example 4.1, the distribution of $(\bar{X}_n - \theta)/S_n$ is independent of F , whereas the distribution of $\sqrt{n}(\bar{X}_n - \theta)$ usually depends on F .

Hence, to obtain an accurate bootstrap confidence set, it is important to choose a good pivot before bootstrapping. Studentizing is one solution, but it requires a variance estimator $\hat{\sigma}_n^2$. Furthermore, we cannot make a studentized statistic more pivotal by studentizing. Beran (1987) suggested the following method to find a good pivot. Let $\mathfrak{R}_n^{(0)}$ be the pivot to start with, $H_n^{(0)}$ be the distribution of $\mathfrak{R}_n^{(0)}$, and $H_{\text{BOOT}}^{(0)}$ be the bootstrap estimator of $H_n^{(0)}$. Define

$$\mathfrak{R}_n^{(1)} = H_{\text{BOOT}}^{(0)}(\mathfrak{R}_n^{(0)}). \quad (4.56)$$

If we replace $H_{\text{BOOT}}^{(0)}$ in (4.56) by $H_n^{(0)}$, then $\mathfrak{R}_n^{(1)}$ has a uniform distribution $U(0, 1)$ that is independent of F . Hence, it is expected that $\mathfrak{R}_n^{(1)}$ is a better pivot than $\mathfrak{R}_n^{(0)}$. Let $H_{\text{BOOT}}^{(1)}$ be the bootstrap estimator of $H_n^{(1)}$, the distribution of $\mathfrak{R}_n^{(1)}$. Then $\mathfrak{R}_n^{(2)} = H_{\text{BOOT}}^{(1)}(\mathfrak{R}_n^{(1)})$ is a better pivot than $\mathfrak{R}_n^{(1)}$. In general, let $H_n^{(j)}$ be the distribution of $\mathfrak{R}_n^{(j)}$ and $H_{\text{BOOT}}^{(j)}$ be the bootstrap estimator of $H_n^{(j)}$, $j = 0, 1, 2, \dots$. Then we can use the following confidence sets for θ :

$$\mathcal{C}_{\text{PREB}}^{(j)} = \{\theta : \mathfrak{R}_n^{(j)} \leq (H_{\text{BOOT}}^{(j)})^{-1}(1 - \alpha)\}, \quad j = 0, 1, 2, \dots \quad (4.57)$$

Note that for each j , $\mathcal{C}_{\text{PREB}}^{(j)}$ is a hybrid bootstrap confidence set based on the pivot $\mathfrak{R}_n^{(j)}$. Since $\mathfrak{R}_n^{(j+1)}$ is a better pivot than $\mathfrak{R}_n^{(j)}$, we obtain a sequence of confidence sets with increasing accuracies. This method is termed *bootstrap prepivoting*. Beran (1987) showed that, if the distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ has a two-term Edgeworth expansion, then the one-sided confidence interval $\mathcal{C}_{\text{PREB}}^{(1)}$ based on the initial pivot $\mathfrak{R}_n^{(0)} = \sqrt{n}(\hat{\theta}_n - \theta)$ is second order accurate, and the two-sided confidence interval $\mathcal{C}_{\text{PREB}}^{(1)}$ based on the initial pivot $\mathfrak{R}_n^{(0)} = \sqrt{n}|\hat{\theta}_n - \theta|$ is third order accurate. Hence, the bootstrap prepivoting with one iteration improves the hybrid bootstrap. It is expected that the one-sided confidence interval $\mathcal{C}_{\text{PREB}}^{(2)}$ based on the initial pivot $\mathfrak{R}_n^{(0)} = \sqrt{n}(\hat{\theta}_n - \theta)$ is third order accurate, i.e., it is better than the one-sided bootstrap-t or bootstrap BC_a confidence interval. More detailed discussions can be found in Beran (1987) and Beran and Ducharme (1991). Some empirical results are shown in Section 4.4.

It seems that, in using this iterative method, we can start with an arbitrary pivot and obtain a bootstrap confidence set that is as accurate

as we want it to be. However, since more computations are required for higher stage bootstrapping, the practical implementation of this method is very hard, or impossible, with current computational ability. We explain this with the computation of $\mathcal{C}_{\text{PREB}}^{(1)}$. Suppose that we use the Monte Carlo approximation. Let $\{X_{1b}^*, \dots, X_{nb}^*\}$ be i.i.d. samples from F_n , $b = 1, \dots, B_1$. Then we can use $H_{\text{BOOT}}^{(0, B_1)}$, the empirical distribution of $\mathfrak{R}_{nb}^{(0)*}$, $b = 1, \dots, B_1$, to approximate $H_{\text{BOOT}}^{(0)}$. For each b , let F_{nb}^* be the empirical distribution of $X_{1b}^*, \dots, X_{nb}^*$ and $\{X_{1bk}^{**}, \dots, X_{nbk}^{**}\}$ be i.i.d. bootstrap samples from F_{nb}^* , $k = 1, \dots, B_2$. Define

$$z_b^* = \frac{1}{B_2} \sum_{k=1}^{B_2} I\{\mathfrak{R}_n(X_{1bk}^{**}, \dots, X_{nbk}^{**}, F_{nb}^*) \leq \mathfrak{R}_n(X_{1b}^*, \dots, X_{nb}^*, F_n)\}$$

for $b = 1, \dots, B_1$. Then, $H_{\text{BOOT}}^{(1)}$ can be approximated by $H_{\text{BOOT}}^{(1, B_1 B_2)}$, the empirical distribution of z_b^* , $b = 1, \dots, B_1$, and the confidence set $\mathcal{C}_{\text{PREB}}^{(1)}$ can be approximated by

$$\{\theta : \mathfrak{R}_n \leq (H_{\text{BOOT}}^{(0, B_1)})^{-1}((H_{\text{BOOT}}^{(1, B_1 B_2)})^{-1}(1 - \alpha))\}.$$

The second stage bootstrap sampling is nested in the first stage bootstrap sampling. Thus, the total number of bootstrap data sets we need is $B_1 B_2$, which is why this method is also called the *double bootstrap*. If each stage requires 1000 bootstrap replicates, then the total number of bootstrap replicates is $10^6!$. Similarly, to compute $\mathcal{C}_{\text{PREB}}^{(j)}$, we need $(1000)^{j+1}$ bootstrap replicates, $j = 2, 3, \dots$, which is impossible to carry out.

Thus, the requirement of a huge number of computations limits the application of the bootstrap prepivoting method, even though some numerical methods to reduce the computation in iterative bootstrap have been suggested (e.g., Hinkley and Shi, 1989; Newton and Geyer, 1994).

A very similar method, *bootstrap inverting*, is given in Hall (1986). Instead of using (4.57), we define

$$\mathcal{C}_{\text{INV B}}^{(j)} = \{\theta : \mathfrak{R}_n^{(j)} \leq (H_{\text{BOOT}}^{(j)})^{-1}(1 - \alpha)\}, \quad j = 0, 1, 2, \dots, \quad (4.58)$$

where

$$\mathfrak{R}_n^{(j)} = \mathfrak{R}_n^{(j-1)} - (H_{\text{BOOT}}^{(j-1)})^{-1}(1 - \alpha), \quad j = 1, 2, \dots,$$

and $H_{\text{BOOT}}^{(j)}$ is the bootstrap estimator of the distribution of $\mathfrak{R}_n^{(j)}$. Note that the difference between bootstrap prepivoting and bootstrap inverting is in the calculation of $\mathfrak{R}_n^{(j)}$. For each j , the orders of accuracy of $\mathcal{C}_{\text{INV B}}^{(j)}$ and $\mathcal{C}_{\text{PREB}}^{(j)}$ in (4.57) are the same, and so are their required amounts of computation: they both require nested bootstrap sampling.

Suppose that $\mathfrak{R}_n^{(0)}$ admits an Edgeworth expansion

$$P\{\mathfrak{R}_n^{(0)} \leq x\} = \Phi(x) + \sum_{j=1}^k \frac{\eta_j(x)\varphi(x)}{n^{j/2}} + O\left(\frac{1}{n^{(k+1)/2}}\right),$$

where η_j are some polynomials whose coefficients depend on F . Then there exist polynomials ζ_j such that

$$P\left\{\mathfrak{R}_n^{(0)} \leq x - \sum_{j=1}^k \frac{\zeta_j(x)}{n^{j/2}}\right\} = \Phi(x) + O\left(\frac{1}{n^{(k+1)/2}}\right) \quad (4.59)$$

(Hall, 1983a). Let $\hat{\zeta}_j$ be estimators of ζ_j . Then (4.59) still holds with ζ_j replaced by $\hat{\zeta}_j$, and we can obtain a $(k+1)$ th order accurate confidence set

$$\mathcal{C}_{\text{EDG}}^{(k)} = \left\{ \theta : \mathfrak{R}_n^{(0)} \leq z_{1-\alpha} - \sum_{j=1}^k \frac{\hat{\zeta}_j(z_{1-\alpha})}{n^{j/2}} \right\}. \quad (4.60)$$

To use $\mathcal{C}_{\text{EDG}}^{(k)}$, we need to derive the polynomials ζ_j , which is not easy. For each k , the asymptotic performance of $\mathcal{C}_{\text{INV_B}}^{(k)}$ is the same as that of $\mathcal{C}_{\text{EDG}}^{(k)}$ (Hall, 1986). Hence, the analytic Edgeworth expansion inversion can be accomplished by iterative bootstrapping.

Bootstrap prepivoting and inverting are special cases of a general iterative bootstrap introduced by Hall and Martin (1988b) to solve statistical problems (not necessarily constructing confidence sets). Let $f_t(F, F_n)$ be a functional indexed by t . Many statistical problems can be formulated as a problem of choosing t so that

$$E[f_t(F, F_n)] = 0. \quad (4.61)$$

The resulting t can be written as $t(F)$. Two examples are as follows.

- (1) Bias estimation. Let $T_n = T(F_n)$ be an estimator of $\theta = T(F)$ for a functional T . Then the estimation of the bias of T_n is equivalent to finding a $t \in \mathbb{R}$ such that (4.61) holds with

$$f_t(F, F_n) = T(F_n) - T(F) - t. \quad (4.62)$$

- (2) Interval estimation. Suppose that we would like to construct a lower confidence bound for $\theta = T(F)$ based on $\hat{\theta}_n = T(F_n)$ and a variance estimator $\hat{\sigma}_n^2 = \sigma^2(F_n)$. Then the problem is equivalent to finding a $t \in \mathbb{R}$ such that (4.61) holds with

$$f_t(F, F_n) = I\{T(F_n) - t\sigma(F_n) \leq T(F)\} - (1 - \alpha). \quad (4.63)$$

A bootstrap solution to this problem is to find $t_n = t(F_n)$ such that

$$E_*[f_{t_n}(F_n, F_n^*)] = 0,$$

where F_n^* is the empirical distribution based on the bootstrap data and E_* is the conditional expectation with respect to bootstrap sampling. When f_t is given by (4.62), the solution t_n is the bootstrap bias estimator defined in (1.22). When f_t is given by (4.63), what we obtain is just the bootstrap-t lower confidence bound.

In general, t_n is not necessarily a solution of (4.61). There may exist a u and a function $\psi(t, s)$ satisfying $\psi(t_n, 0) = t_n$ such that

$$E[f_{\psi(t_n, u)}(F, F_n)] = 0.$$

If we can find u and ψ , then we have a perfect solution to (4.61). Let $f_u^{(1)}(F, F_n) = f_{\psi(t_n, u)}(F, F_n)$. Then the problem reduces to solving (4.61) with f_t replaced by $f_u^{(1)}$. This process can be applied repeatedly and therefore is called the *iterative bootstrap* method.

Let f_t be given by (4.63) and $Q_F(s)$ be the s th quantile of the distribution of $[T(F_n) - T(F)]/\sigma(F_n)$. If $t_n = Q_{F_n}(1 - \alpha)$ and $\psi(t_n, u) = t_n + u$, then the method reduces to bootstrap inverting. If $t_n = Q_{F_n}(1 - \alpha)$ and $\psi(t_n, u) = Q_{F_n}(1 - \alpha + u)$, then the method reduces to bootstrap prepivoting.

The computational algorithm for this method with one iteration can be described as follows. Draw i.i.d. samples $\{X_{1b}^*, \dots, X_{nb}^*\}$ from F_n , $b = 1, \dots, B_1$. For each b , let F_{nb}^* be the empirical distribution of $X_{1b}^*, \dots, X_{nb}^*$ and draw i.i.d. samples $\{X_{1bk}^{**}, \dots, X_{nbk}^{**}\}$ from F_{nb}^* , $k = 1, \dots, B_2$. Let F_{nbk}^{**} be the empirical distribution of $X_{1bk}^{**}, \dots, X_{nbk}^{**}$. We first solve the equation

$$\frac{1}{B_2} \sum_{k=1}^{B_2} f_t(F_{nb}^*, F_{nbk}^{**}) = 0$$

for each $b = 1, \dots, B_1$. Denote the solution by t_b^* . Then the final solution is obtained by solving

$$\frac{1}{B_1} \sum_{b=1}^{B_1} f_{\psi(t_b^*, u)}(F_n, F_{nb}^*) = 0.$$

Hall and Martin (1988b) showed that if

$$E[f_{t_n}(F, F_n)] = c(F)n^{-j} + O(n^{-(j+1)})$$

for a smooth functional $c(F)$ and a positive j , then

$$E[f_{\psi(t_n, u)}(F, F_n)] = O(n^{-(j+1)}),$$

i.e., the order of the error is reduced from $O(n^{-j})$ to $O(n^{-(j+1)})$ by a one-step bootstrap iteration.

4.3.2 Bootstrap calibrating

This method was first proposed by Loh (1987) and subsequently refined in Loh (1988, 1991). The basic idea is to improve the original confidence set by adjusting its nominal level. Let \mathcal{C}_n be a confidence set for θ with nominal level $1 - \alpha$ and π_n be the actual coverage probability of \mathcal{C}_n . The value of π_n , which may not be $1 - \alpha$, can be estimated by a bootstrap estimator $\hat{\pi}_n$. If we find that $\hat{\pi}_n$ is far from $1 - \alpha$, then we adjust the nominal level $1 - \alpha$ to $1 - \tilde{\alpha}$ and construct a new confidence set for θ with nominal level $1 - \tilde{\alpha}$. This confidence set is called the *bootstrap calibration confidence set*. Bootstrap calibrating can be used iteratively as follows: estimate the true coverage probability of the bootstrap calibration confidence set and find the difference between the estimated coverage probability and $1 - \alpha$. If the difference is still large, we can adjust the nominal level of the calibrated confidence set.

The key for bootstrap calibrating is how to determine the new nominal level $1 - \tilde{\alpha}$ in each step. We now discuss the method suggested by Loh (1988, 1991) in the case where the initial confidence sets are obtained by normal approximation.

Consider first the lower confidence bound $\underline{\theta}_{\text{NOR}}$ defined in (4.53). The coverage probability $\pi_n = P\{\underline{\theta}_{\text{NOR}} \leq \theta\}$ can be estimated by the bootstrap estimator (approximated by Monte Carlo if necessary)

$$\hat{\pi}_n = G_{\text{BOOT}}(z_{1-\alpha}) = P_*\{(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}_n^* \leq z_{1-\alpha}\}.$$

When the bootstrap distribution has an Edgeworth expansion, we have

$$\hat{\pi}_n = 1 - \alpha + \left[\frac{q_1(z_{1-\alpha}, F_n)}{\sqrt{n}} + \frac{q_2(z_{1-\alpha}, F_n)}{n} \right] \varphi(z_{1-\alpha}) + O_p\left(\frac{1}{n^{3/2}}\right).$$

Let h be any increasing, unbounded, and twice differentiable function on $(0, 1)$ and define

$$\delta = h(\hat{\pi}_n) - h(1 - \alpha).$$

Then

$$\begin{aligned} \delta &= \left[\frac{q_1(z_{1-\alpha}, F_n)}{\sqrt{n}} + \frac{q_2(z_{1-\alpha}, F_n)}{n} \right] \varphi(z_{1-\alpha}) h'(1 - \alpha) \\ &\quad + \frac{[q_1(z_{1-\alpha}, F_n) \varphi(z_{1-\alpha})]^2}{2n} h''(1 - \alpha) + O_p\left(\frac{1}{n^{3/2}}\right). \end{aligned} \quad (4.64)$$

Let

$$\tilde{\alpha} = 1 - h^{-1}(h(1 - \alpha) - \delta).$$

The bootstrap calibration lower confidence bound is

$$\underline{\theta}_{\text{CLB}} = \hat{\theta}_n - \hat{\sigma}_n z_{1-\tilde{\alpha}}. \quad (4.65)$$

By (4.64),

$$1 - \tilde{\alpha} = 1 - \alpha + \frac{q_1(z_{1-\alpha}, F_n) \varphi(z_{1-\alpha})}{\sqrt{n}} + O_p\left(\frac{1}{n}\right).$$

Therefore,

$$z_{1-\tilde{\alpha}} = z_{1-\alpha} + \frac{q_1(z_{1-\alpha}, F_n)}{\sqrt{n}} + O_p\left(\frac{1}{n}\right)$$

and

$$\underline{\theta}_{\text{CLB}} = \hat{\theta}_n - \hat{\sigma}_n \left[z_{1-\alpha} + \frac{q_1(z_{1-\alpha}, F_n)}{\sqrt{n}} + O_p\left(\frac{1}{n}\right) \right]. \quad (4.66)$$

Comparing (4.66) with (4.37) and (4.52), we find that

$$\underline{\theta}_{\text{CLB}} - \underline{\theta}_{\text{BT}} = O_p(n^{-3/2}) \quad \text{and} \quad \underline{\theta}_{\text{CLB}} - \underline{\theta}_{\text{BC}_a} = O_p(n^{-3/2}).$$

Thus, $[\underline{\theta}_{\text{CLB}}, \infty)$ is second order accurate. Loh (1988) showed that we can even select the function h to obtain a third order accurate $[\underline{\theta}_{\text{CLB}}, \infty)$.

We can take $[\underline{\theta}_{\text{CLB}}, \bar{\theta}_{\text{CLB}}]$ as a two-sided confidence interval; it is still second order accurate. By calibrating directly the equal-tail two-sided normal approximation confidence interval

$$[\underline{\theta}_{\text{NOR}}, \bar{\theta}_{\text{NOR}}] = [\hat{\theta}_n - \hat{\sigma}_n z_{1-\alpha}, \hat{\theta}_n + \hat{\sigma}_n z_{1-\alpha}], \quad (4.67)$$

we can obtain a higher order accurate confidence interval. Let $\hat{\pi}_n$ be the bootstrap estimator of the coverage probability $P\{\underline{\theta}_{\text{NOR}} \leq \theta \leq \bar{\theta}_{\text{NOR}}\}$, $\delta = h(\hat{\pi}_n) - h(1 - 2\alpha)$, and $\tilde{\alpha} = [1 - h^{-1}(h(1 - 2\alpha) - \delta)]/2$. Then the two-sided bootstrap calibration confidence interval is the interval given by (4.67) with α replaced by $\tilde{\alpha}$. Under some conditions, Loh (1991) showed that this confidence interval is fourth order accurate. The length of this interval excesses the length of the interval in (4.67) by an amount of order $O_p(n^{-3/2})$.

If we begin with a second order accurate confidence set constructed from a one-term Edgeworth expansion, then bootstrap calibrating yields a confidence set which is third order accurate in the one-sided case and fourth order accurate in the two-sided case. The calibration can be iterated to obtain higher order accurate confidence sets. Loh (1991) also introduced other methods for calibration and calibrating other types of confidence sets. If we start with a bootstrap confidence set, then bootstrap calibrating is the same as Beran's bootstrap prepivoting (DiCiccio and Romano, 1988b).

Some empirical results are presented in Section 4.4.

4.3.3 The automatic percentile and variance stabilizing

If assumption (4.14) is true for all n , then the bootstrap BC_a confidence sets are exact for all n . However, using the bootstrap BC_a requires knowing

or estimating the function Ψ and the constants z_0 and a . We have seen that it is not easy to determine the constant a . DiCiccio and Romano (1988a, 1989) proposed another percentile method that avoids the determination of Ψ , a , and z_0 and therefore is called the *automatic percentile* method.

We introduce this method with the parametric model: $F = F_\theta, \theta \in \mathbb{R}$. Define

$$K_\theta(x) = P_\theta\{\hat{\theta}_n \leq x\},$$

where P_θ is the probability law of X_1, \dots, X_n when the true value of the parameter is θ . Therefore, $K_{\hat{\theta}_n}$ is the bootstrap estimator of K_θ . Let θ_0 be any initial guess value of θ ,

$$\theta_1 = K_{\theta_0}^{-1}(1 - \alpha),$$

and

$$\underline{\theta}_{\text{AUB}} = K_{\hat{\theta}_n}^{-1}(K_{\theta_1}(\theta_0)). \quad (4.68)$$

For each fixed n , if assumption (4.14) holds, then $\underline{\theta}_{\text{AUB}}$ is an exact lower confidence bound for θ . That is, $\underline{\theta}_{\text{AUB}}$ is the same as $\underline{\theta}_{\text{BCa}}$ and $\underline{\theta}_{\text{EXACT}}$ in (4.16). This assertion can be proved as follows. Under assumption (4.14) (for simplicity we omit the subscript n),

$$K_{\hat{\theta}}(\theta) = P_{\hat{\theta}}\{\hat{\theta}^* \leq \theta\} = \Psi(z_0 + [\phi(\theta) - \phi(\hat{\theta})]/[1 + a\phi(\hat{\theta})]),$$

and

$$K_{\theta_1}(\theta_0) = P_{\theta_1}\{\hat{\theta} \leq \theta_0\} = \Psi(z_0 + [\phi(\theta_0) - \phi(\theta_1)]/[1 + a\phi(\theta_1)]).$$

Then, by (4.68),

$$\begin{aligned} P_\theta\{\underline{\theta}_{\text{AUB}} \leq \theta\} &= P_\theta\{K_{\theta_1}(\theta_0) \leq K_{\hat{\theta}}(\theta)\} \\ &= P_\theta\left\{\frac{\phi(\theta_0) - \phi(\theta_1)}{1 + a\phi(\theta_1)} \leq \frac{\phi(\theta) - \phi(\hat{\theta})}{1 + a\phi(\hat{\theta})}\right\} \\ &= P_\theta\left\{\phi^{-1}\left(\phi(\hat{\theta}) + [\phi(\theta_0) - \phi(\theta_1)]\frac{1 + a\phi(\hat{\theta})}{1 + a\phi(\theta_1)}\right) \leq \theta\right\} \\ &= P_\theta\left\{\phi^{-1}\left(\phi(\hat{\theta}) + \frac{\xi(\theta_0, \theta_1)[1 + a\phi(\hat{\theta})]}{1 - a\xi(\theta_0, \theta_1)}\right) \leq \theta\right\}, \end{aligned}$$

where $\xi(\theta_0, \theta_1) = [\phi(\theta_0) - \phi(\theta_1)]/[1 + a\phi(\theta_0)]$. The result then follows from

$$\xi(\theta_0, \theta_1) = z_\alpha + z_0,$$

which is a consequence of the definition of θ_1 and (4.14):

$$\begin{aligned} \Psi(-\xi(\theta_0, \theta_1) + z_0) &= P_{\theta_0}\left\{\frac{\phi(\hat{\theta}) - \phi(\theta_0)}{1 + a\phi(\theta_0)} + z_0 \leq \frac{\phi(\theta_1) - \phi(\theta_0)}{1 + a\phi(\theta_0)} + z_0\right\} \\ &= P_{\theta_0}\{\hat{\theta} \leq \theta_1\} = 1 - \alpha. \end{aligned}$$

If the form of K_θ is unknown when θ is known, then using this method requires the simulation of the distribution K_θ at three θ values: $\hat{\theta}_n$, θ_0 , and θ_1 . This involves more computations than those required by the bootstrap BC_a.

If assumption (4.14) holds exactly, then the initial value θ_0 has no effect. However, since assumption (4.14) often holds approximately, the choice of θ_0 does affect the performance of the automatic percentile confidence sets. DiCiccio and Romano (1989) showed that if we start with an initial θ_{0n} satisfying

$$|\underline{\theta}_{\text{EXACT}} - \theta_{0n}| = O_p(n^{-k/2}),$$

then under some conditions, including the existence of the Edgeworth expansions,

$$P\{\underline{\theta}_{\text{AUB}} \leq \theta\} = 1 - \alpha + O(n^{-(k+1)/2}).$$

But in general $k \leq 1$ (Martin, 1990, Remark 3.1). Therefore, this method can at most be used to improve a first order accurate one-sided confidence interval to be second order accurate.

DiCiccio and Romano (1990) extended the automatic percentile method to the nonparametric model by reducing the nonparametric family to a least favorable parametric family and applying the above method to the reduced parametric family.

Although the bootstrap percentile, BC, and BC_a methods are motivated by the existence of a variance-stabilizing transformation, they do not actually use this transformation in constructing confidence sets, i.e., the bootstrap distributions of the untransformed variables are used. In some cases, we can use the bootstrap to estimate the variance-stabilizing transformations and then construct confidence sets by bootstrapping the transformed variables. This is the main idea behind the *variance-stabilizing bootstrap* proposed by Tibshirani (1988).

In some problems, the variance (or the asymptotic variance) of the estimator $\hat{\theta}_n$ depends on F only through the parameter θ , i.e.,

$$\sigma_n^2 = \sigma_n^2(\theta). \quad (4.69)$$

If σ_n^2 is a known function, then assumption (4.14) holds approximately with the transformation

$$\phi_n(\theta) = \int^\theta \frac{ds}{\sigma_n(s)},$$

and we obtain bootstrap confidence sets by bootstrapping the transformed variable. Let $\bar{G}_{\text{BOOT}}(x)$ be the bootstrap estimator of the distribution of the transformed variable $\phi_n(\hat{\theta}_n) - \phi_n(\theta)$. Then the variance-stabilizing bootstrap lower confidence bound for θ is $\phi_n^{-1}(\phi_n(\hat{\theta}_n) - \bar{G}_{\text{BOOT}}^{-1}(1 - \alpha))$, which is

actually the bootstrap-t lower confidence bound based on the transformed variable; hence, it is second order accurate.

In general, the form of σ_n^2 may not be known. Tibshirani (1988) suggested the following method.

- (a) Draw B_1 independent sets of bootstrap data $\{X_{1b}^*, \dots, X_{nb}^*\}$, $b = 1, \dots, B_1$, from \hat{F} . For each b , calculate $\hat{\theta}_{nb}^* = \hat{\theta}_n(X_{1b}^*, \dots, X_{nb}^*)$.
- (b) For each b , generate independent bootstrap data $\{X_{1bk}^{**}, \dots, X_{nbk}^{**}\}$ from the empirical distribution based on $X_{1b}^*, \dots, X_{nb}^*$, $k = 1, \dots, B_2$. Let \hat{w}_b^* be the sample variance of $\hat{\theta}_n(X_{1bk}^{**}, \dots, X_{nbk}^{**})$, $k = 1, \dots, B_2$.
- (c) Obtain an estimate for the variance function σ_n^2 by smoothing \hat{w}_b^* versus $\hat{\theta}_{nb}^*$. For example, we can use the kernel smoothing method to obtain

$$\hat{\sigma}_n^2(\theta) = \left[\sum_{b=1}^{B_1} \hat{w}_{nb}^* \Phi\left(\frac{\theta - \hat{\theta}_{nb}^*}{h_n}\right) \right] / \left[\sum_{b=1}^{B_1} \Phi\left(\frac{\theta - \hat{\theta}_{nb}^*}{h_n}\right) \right].$$

- (d) Generate B_3 independent sets of bootstrap samples $\{X_{1j}^*, \dots, X_{nj}^*\}$, $j = 1, \dots, B_3$, from \hat{F} . For each j , use some numerical method to calculate

$$\phi_n(\hat{\theta}_{nj}^*) = \int_{-\infty}^{\hat{\theta}_{nj}^*} \frac{ds}{\hat{\sigma}_n(s)}.$$

Then calculate the quantile of \bar{G}_{BOOT} using $\phi_n(\hat{\theta}_{nj}^*)$, $j = 1, \dots, B_3$.

Note that the bootstrap sampling in step (b) is nested in step (a). The total number of bootstrap replicates required is $B = B_1 B_2 + B_3$. In Chapter 5, we will see that roughly B_1 , B_2 , and B_3 should be 100, 25, and 1000, respectively, and thus $B = 3500$.

However, if assumption (4.69) does not hold, then we cannot expect this method to work well. Tibshirani (1988) examined several examples by empirical simulation. The variance-stabilizing bootstrap works well for the parameters such as the correlation coefficient for which assumption (4.69) holds but badly for the parameters such as the median for which assumption (4.69) is seriously violated, although there are also some examples showing that this method still works for situations where assumption (4.69) is invalid. Some extensions of this method can be found in DiCiccio and Romano (1990).

4.3.4 Fixed width bootstrap confidence intervals

Although for a given nominal level we may find the shortest confidence interval, it is impossible to obtain a confidence interval with a given nominal level and a given length if the sample size n is fixed. Thus, some sequential

confidence intervals have been suggested to solve this problem. Swanepoel, van Wyk and Venter (1983) proposed a sequential method to construct a bootstrap confidence interval for a parameter θ with level $1 - 2\alpha$ and fixed width $2d$. Let $\hat{\theta}_n$ be an estimator of θ , $\hat{\theta}_n^*$ be its bootstrap analog, and

$$\hat{\pi}_n = P_* \{ \hat{\theta}_n^* - d \leq \hat{\theta}_n \leq \hat{\theta}_n^* + d \}.$$

Then the stopping time is taken as

$$N_{\text{BOOT}} = \inf \{ n \geq n_0 : \hat{\pi}_n \geq 1 - \alpha \},$$

where n_0 is an integer larger than $z_{1-\alpha}/2$. The resulting confidence interval is

$$[\hat{\theta}_{N_{\text{BOOT}}} - d, \hat{\theta}_{N_{\text{BOOT}}} + d]. \quad (4.70)$$

If the bootstrap data are taken from the empirical distribution F_n and θ is the population mean or median, then

$$\lim_{d \rightarrow 0} P \{ \hat{\theta}_{N_{\text{BOOT}}} - d \leq \theta \leq \hat{\theta}_{N_{\text{BOOT}}} + d \} = 1 - \alpha \quad (4.71)$$

and

$$\lim_{d \rightarrow 0} d^2 N_{\text{BOOT}} = \sigma_F^2 z_{1-\alpha}^2 \text{ a.s.}, \quad (4.72)$$

where σ_F^2 is the asymptotic variance of $\sqrt{n}(\hat{\theta}_n - \theta)$.

Result (4.71) indicates that the confidence interval in (4.70) is consistent in a sense different from Definition 4.1. Recall that the normal approximation confidence interval with fixed sample size n is

$$[\hat{\theta}_n - \hat{\sigma}_F z_{1-\alpha}/\sqrt{n}, \hat{\theta}_n + \hat{\sigma}_F z_{1-\alpha}/\sqrt{n}], \quad (4.73)$$

where $\hat{\sigma}_F$ is an estimator of σ_F . If we require the length of this interval to be $2d$, then n should satisfy $n = [\hat{\sigma}_F^2 z_{1-\alpha}^2/d^2]$, where $[x]$ is the integer part of x . From (4.72), we know that as $d \rightarrow 0$, $N_{\text{BOOT}}/\tilde{n} \rightarrow 1$ a.s., where $\tilde{n} = [\sigma_F^2 z_{1-\alpha}^2/d^2]$ can be approximated by $[\hat{\sigma}_F^2 z_{1-\alpha}^2/d^2]$. That is, when d is small, (4.72) tells us that the average sample size of the fixed width bootstrap confidence interval (4.70) is close to the sample size of the nonsequential confidence interval (4.73).

4.3.5 Likelihood based bootstrap confidence sets

So far we have mainly focused on the confidence sets for a real parameter θ , although evidently many procedures can be extended to vector-valued θ or even more complicated θ . Here we consider the case where θ is vector-valued and introduce a method called the *likelihood based bootstrap* (Hall, 1987).

Suppose that $\theta \in \mathbb{R}^q$. Our aim is to construct a confidence set \mathcal{C}_n such that $P\{\theta \in \mathcal{C}_n\} = 1 - \alpha$ and all parameter values inside \mathcal{C}_n have a higher likelihood than those outside \mathcal{C}_n . Let $\hat{\theta}_n$ be an estimator of θ , $\hat{\Sigma}_n$ be a variance estimator for $\hat{\theta}_n$, and $T_n = \hat{\Sigma}_n^{-1/2}(\hat{\theta}_n - \theta)$. Suppose that T_n has a density f_n . If f_n is known, then a $1 - \alpha$ confidence set for θ is

$$\mathcal{C}_n = \{\theta : f_n(\hat{\Sigma}_n^{-1/2}(\hat{\theta}_n - \theta)) \geq c_n\},$$

where c_n is chosen so that $P\{f_n(T_n) \geq c_n\} = 1 - \alpha$. When f_n is unknown, a bootstrap estimate of c_n can be obtained as follows. Generate B independent sets of bootstrap data from the empirical distribution, use each of them to calculate the variable T_n , and obtain T_b^* , $b = 1, \dots, B$. Let

$$\hat{f}_n(t) = \frac{1}{Bh_n^q} \sum_{b=1}^B \kappa\left(\frac{t - T_b^*}{h_n}\right), \quad t \in \mathbb{R}^q$$

be a kernel density estimator of f_n , where $\kappa(t)$ is a given kernel function and $\{h_n\}$ is a sequence of positive numbers. Let \hat{c}_n be a value such that the set $\{t : \hat{f}_n(t) \geq \hat{c}_n\}$ excludes just $100\alpha\%$ of the values T_b^* , $b = 1, \dots, B$. Then \hat{c}_n is the bootstrap estimate of c_n , and the likelihood based bootstrap confidence set for θ with an approximate level $1 - \alpha$ is

$$\mathcal{C}_{\text{BT}} = \{\theta : \hat{f}_n(\hat{\Sigma}_n^{-1/2}(\hat{\theta}_n - \theta)) \geq \hat{c}_n\}.$$

The choices of $\kappa(t)$ and h_n are discussed in Hall (1987). He also showed that \mathcal{C}_{BT} is second order accurate. If we use the variable $\sqrt{n}(\hat{\theta}_n - \theta)$ for bootstrapping, we can obtain the hybrid bootstrap confidence set, but it is only first order accurate.

4.4 Empirical Comparisons

In this section, we examine and compare (by simulation or example) various bootstrap confidence sets and other confidence sets (e.g., confidence sets obtained by normal approximation). Throughout this section, AV denotes the simulation average of a confidence bound, CP denotes the simulation coverage probability of a confidence set \mathcal{C} , and EL denotes the simulation estimates of the expected length of a two-sided confidence interval. All of the bootstrap confidence sets are computed by Monte Carlo with size B .

4.4.1 The bootstrap-t, percentile, BC, and BC_a

We first look at parametric bootstrap confidence sets. DiCiccio and Romano (1988b) considered the ratio problem described in Example 4.4(ii) with $c = 0.4$ and $n = 5$. For the bootstrap BC and BC_a , $z_0 = -0.1217$ and

$a = -0.1195$. The results are summarized in Table 4.2.

From Table 4.2, the bootstrap percentile is improved by the bootstrap BC, which is further improved by the bootstrap BC_a . The bootstrap percentile and BC lower confidence bounds are larger than the exact lower confidence bound, resulting in undercoverage. On the other hand, the bootstrap percentile and BC upper confidence bounds overcover the true θ . As a result, the equal-tail two-sided bootstrap percentile and BC confidence intervals have better performances. This supports the asymptotic theory in Section 4.2: the bootstrap percentile and BC one-sided confidence intervals are first order accurate; the bootstrap BC_a one-sided confidence intervals and all of the two-sided confidence intervals are second order accurate.

Table 4.2. Comparison of the bootstrap percentile (BP), BC, and BC_a confidence sets for $\theta = \eta_2/\eta_1$ in Example 4.4
 $(1 - \alpha = 0.975)$ [Adapted from DiCiccio and Romano
 (1988b), by permission of Royal Statistical Society]

Method	Lower Bound		Upper Bound		Two-Sided	
	CP	AV	CP	AV	CP	EL
Exact	0.975	$0.113\hat{\theta}$	0.975	$4.468\hat{\theta}$	0.950	$4.355\hat{\theta}$
BP	0.919	$0.224\hat{\theta}$	0.998	$8.844\hat{\theta}$	0.917	$8.620\hat{\theta}$
BC	0.941	$0.184\hat{\theta}$	0.992	$6.506\hat{\theta}$	0.934	$6.322\hat{\theta}$
BC_a	0.978	$0.105\hat{\theta}$	0.974	$4.404\hat{\theta}$	0.952	$4.299\hat{\theta}$

Table 4.3. Comparison of the normal approximation (NOR), bootstrap percentile (BP), BC, and BC_a confidence intervals for variance ($1 - 2\alpha = 0.9$)
 [Adapted from DiCiccio and Tibshirani (1987),
 by permission of American Statistical Association]

Method		Left	Right	Two-Sided	
		AV	AV	CP	EL
Parametric	Exact	0.630	1.878	0.900	1.248
	NOR	0.466	1.531	0.890	1.065
	BP	0.520	1.585	0.893	1.065
	BC	0.578	1.670	0.893	1.092
	BC_a	0.628	1.860	0.903	1.232
Nonparametric	BP	0.484	1.363	0.757	0.879
	BC	0.592	1.467	0.807	0.875
	BC_a	0.617	1.524	0.807	0.907

While the performance of the parametric bootstrap BC_a is excellent, it has been found that the nonparametric bootstrap BC_a does not do well in many problems. One obvious reason is that the acceleration constant a can usually be precisely determined in parametric problems (in some cases a does not depend on F) but not in nonparametric problems. Bickel (1992) claimed that in general the parametric and nonparametric BC_a bounds are not second order equivalent.

DiCiccio and Tibshirani (1987) performed an empirical simulation study of the problem of constructing a confidence interval for the population variance θ based on i.i.d. X_1, \dots, X_n from $N(0, 1)$ ($n = 20$) and the sample variance $\hat{\theta}_n = (n - 1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. The parametric and nonparametric bootstraps were done by resampling from $N(0, \hat{\theta}_n)$ and the empirical distribution, respectively. The bootstrap Monte Carlo size is $B = 1000$ and the simulation size is 300. Their results are given in Table 4.3.

The exact confidence interval is obtained using the fact that $(n - 1)\hat{\theta}_n/\theta$ has a chi-square distribution. The parametric normal approximation is done by estimating the asymptotic variance of $\hat{\theta}_n$ by $2\hat{\theta}_n^2/n$. The results for parametric bootstrap confidence intervals in Table 4.3 are very similar to those in Table 4.2: the normal approximation, the bootstrap percentile, and BC confidence intervals undercover on the right but overcover on the left and hence have a nearly correct coverage probability; the bootstrap BC_a confidence interval is the only interval capturing the asymmetry of the exact confidence interval. However, none of the nonparametric bootstrap confidence intervals have a good performance. The bootstrap BC improves the bootstrap percentile but is still not good. The bootstrap BC_a has the same CP as the bootstrap BC, but it can be seen from the right and left limits that the bootstrap BC_a produces better one-sided confidence intervals, which confirms the asymptotic theory in Section 4.2.

The unsatisfactory performance of the bootstrap BC_a was also discovered by Loh and Wu (1987). They examined the bootstrap percentile, BC, BC_a , and the bootstrap-t confidence intervals in the case where $\theta = EX_1$ or $\text{var}(X_1)$. Three population distributions of different tails and shapes were considered: $N(0, 1)$, the t-distribution with 5 degrees of freedom, and the Weibull distribution (the latter two distributions were scaled so that they have unit variances). For the problem of $\theta = EX_1$, the usual t-confidence interval, which is exact when the population distribution is normal, was also included in the above study. The sample size, the bootstrap Monte Carlo size, and the simulation size are 20, 200, and 500, respectively. The results are displayed in Table 4.4. The bootstrap BC_a is only a slight improvement over the bootstrap percentile or BC, and all of them perform poorly when the population is asymmetric. The bootstrap-t is good in terms of the CP, but its length tends to be quite a bit longer than those of other intervals.

Table 4.4. Comparison of the bootstrap percentile, BC, BC_a , and bootstrap-t (BT) confidence intervals for θ
 $(1 - 2\alpha = 0.9)$ [Adapted from Loh and Wu (1987),
by permission of American Statistical Association]

θ	Method	Normal		Student		Weibull	
		CP	EL	CP	EL	CP	EL
Mean	t	0.90	0.76	0.91	1.80	0.75	2.80
	BP	0.91	0.71	0.84	1.70	0.73	2.60
	BC	0.90	0.71	0.81	1.70	0.73	2.70
	BC_a	0.89	0.72	0.79	1.80	0.75	3.00
	BT	0.92	0.77	0.83	2.70	0.83	5.50
Variance	BP	0.79	0.86	0.68	1.10	0.65	1.30
	BC	0.82	0.89	0.75	1.20	0.67	1.30
	BC_a	0.83	0.94	0.77	1.30	0.69	1.50
	BT	0.88	1.50	0.85	3.20	0.83	5.50

Table 4.5. Confidence intervals for correlation coefficient
 $(1 - 2\alpha = 0.9)$ [Adapted from Efron (1987), by
permission of American Statistical Association]

Method		Left Limit	Right Limit
Exact		0.496	0.898
NOR	(based on $\hat{\rho}_n$)	0.587	0.965
NOR	[based on $\tanh^{-1}(\hat{\rho}_n)$]	0.490	0.900
BC_a	($a = 0$)	0.488	0.900
BC_a	($a = -0.0817$)	0.430	0.920
BT	($\hat{\sigma}_n = \sqrt{v_{\text{JACK}}}$)	-0.06	0.940
BT	[$\hat{\sigma}_n = (1 - \hat{\rho}_n^2)/\sqrt{n - 3}$]	-0.01	0.910

Table 4.6. Comparison of the bootstrap percentile, BC_a , and normal approximation confidence intervals for correlation coefficient $\rho = 0.9$
 $(1 - 2\alpha = 0.9)$ [Adapted from DiCiccio and Tibshirani (1987),
by permission of American Statistical Association]

Method	Left	Right	Two-Sided	
	AV	AV	CP	EL
NOR (based on $\hat{\rho}_n$)	0.816	0.954	0.930	0.138
NOR (based on $\tanh^{-1}(\hat{\rho}_n)$)	0.757	0.958	0.927	0.201
BP	0.761	0.930	0.820	0.169
BC_a ($a = 0$)	0.742	0.922	0.767	0.180
BC_a ($a = -0.0817$)	0.701	0.914	0.707	0.213

The bootstrap-t is not necessarily always better than the other bootstrap methods. Efron (1987) applied the bootstrap methods to the law school data given in the beginning of Section 3.4.2. With a bootstrap Monte Carlo size $B = 1000$, he computed various confidence intervals for ρ (Table 4.5). Since $a = 0$ is the true value of the acceleration constant under the bivariate normal model, the bootstrap BC_a with $a = 0$ works very well. It is slightly better than the two parametric normal approximation methods. The bootstrap BC_a with estimated $a = -0.0817\{\hat{a} = -[\hat{\rho}_n(3 + \hat{\rho}_n^2)]/[3n^{1/2}(1 + \hat{\rho}_n^2)^{3/2}]\}$ works reasonably well. However, the bootstrap-t performs poorly, especially for the left limit of the interval. Efron (1987) suggested that the bootstrap-t works better on genuine location-scale problems.

However, for an extreme value of $\rho (= 0.9)$, DiCiccio and Tibshirani (1987) found that all three bootstrap percentile methods were worse than the normal approximation. The bootstrap BC_a pulls the bootstrap percentile confidence interval in the wrong direction. Their results (Table 4.6) are based on a simulation of size 300 under a bivariate normal model with 0 means, unit variances, and $n = 15$.

4.4.2 The bootstrap and other asymptotic methods

We turn to the comparison of the bootstrap and other methods such as the normal approximation and the Edgeworth expansion. Yang (1985) compared the hybrid bootstrap, the bootstrap-t, and the normal approximation confidence intervals for population mean based on the sample trimmed mean

$$\hat{\theta}_n = \frac{1}{[n(1 - 2t)] + 1} \sum_{i=[nt]+1}^{n-[nt]} X_{(i)},$$

where $X_{(i)}$ is the i th order statistic of the sample. The results in Table 4.7 are based on 1000 simulations. The bootstrap Monte Carlo size is 500. Since the confidence intervals under consideration are equal-tail two-sided, they are all second order accurate (Section 4.2). However, the CP of the normal approximation confidence interval is apparently too low, especially when $t = 0.2$ and $n = 10$. Both bootstrap methods are better than the normal approximation. The bootstrap-t confidence interval has the most accurate CP, but the longest EL.

It was shown in Chapter 3 that Edgeworth expansions can be utilized to approximate the distributions of statistics. Thus, one may construct confidence sets based on estimated Edgeworth expansions. More details can be found in Section 4.3 [formula (4.60)]. For the ratio estimator $\hat{\gamma}_n = \bar{Y}_n/\bar{Z}_n$, where $(\bar{Y}_n, \bar{Z}_n)'$ is the sample mean based on a sample of size $n = 50$ from

Table 4.7. Comparison of the hybrid bootstrap (HB), bootstrap-t, and normal approximation confidence intervals based on the trimmed sample mean ($1 - 2\alpha = 0.95$) [Adapted from Yang (1985), by permission of Vereniging voor Statistiek]

Method	$t = 0.1$				$t = 0.2$			
	$n = 10$		$n = 20$		$n = 10$		$n = 20$	
	CP	EL	CP	EL	CP	EL	CP	EL
$F = N(0, 1)$								
NOR	0.899	1.033	0.919	0.806	0.865	0.993	0.906	0.805
HB	0.921	1.150	0.934	0.865	0.896	1.171	0.926	0.878
BT	0.960	1.550	0.958	0.948	0.941	1.866	0.953	1.033
$F = \text{double exponential}$								
NOR	0.892	1.214	0.935	0.979	0.867	1.082	0.927	0.879
HB	0.940	1.497	0.954	1.110	0.934	1.444	0.955	1.028
BT	0.936	1.801	0.943	1.143	0.942	1.966	0.941	1.094
$F = \text{student with 4 df}$								
NOR	0.900	1.239	0.929	0.965	0.876	1.131	0.907	0.916
HB	0.942	1.500	0.958	1.091	0.936	1.467	0.936	1.037
BT	0.945	1.820	0.958	1.128	0.931	2.063	0.948	1.160
$F = 0.9N(0, 1) + 0.1N(0, 9)$								
NOR	0.889	1.151	0.917	0.914	0.845	1.072	0.892	0.883
HB	0.918	1.416	0.948	1.049	0.910	1.382	0.923	0.992
BT	0.943	1.680	0.942	1.065	0.928	1.930	0.942	1.122

Table 4.8. CP of confidence sets for the ratio obtained by the normal approximation, bootstrap-t, estimated Edgeworth expansion using the jackknife (EDGJ), and using substitution (EDGS) [Adapted from Hinkley and Wei (1984), by permission of Biometrika]

Method	Lower	Upper	Two-Sided	Lower	Upper	Two-Sided
Nominal	0.950	0.950	0.900	0.975	0.975	0.950
NOR	0.915	0.945	0.860	0.950	0.975	0.925
BT	0.950	0.960	0.910	0.965	0.985	0.950
EDGJ	0.925	0.950	0.875	0.950	0.975	0.925
EDGS	0.920	0.950	0.870	0.950	0.975	0.925

Table 4.9. Comparison of confidence intervals obtained by transformation (TRAN) and the bootstrap BC_a [Adapted from Tu and Zhang (1992b), by permission of Physica-Verlag]

n	F	Method	Two-Sided		Left Limit		Right Limit	
			CP	EL	AV	SD	AV	SD
20	Φ	Exact	0.90	0.98	0.46		1.44	
		TRAN	0.82	1.45	0.33	0.36	1.78	0.96
		BC _a	0.81	0.91	0.58	0.20	1.49	0.49
	Φ_M	Exact	0.90	1.02	0.43		1.45	
		TRAN	0.80	1.59	0.25	0.49	1.84	1.59
		BC _a	0.80	0.93	0.56	0.20	1.49	0.53
30	Φ	Exact	0.90	0.82	0.55		1.37	
		TRAN	0.86	1.27	0.44	0.27	1.70	0.80
		BC _a	0.84	0.80	0.64	0.18	1.44	0.41
	Φ_M	Exact	0.90	0.85	0.53		1.38	
		TRAN	0.86	1.23	0.45	0.23	1.68	0.73
		BC _a	0.84	0.79	0.64	0.18	1.44	0.40
50	Φ	Exact	0.90	0.65	0.65		1.30	
		TRAN	0.92	0.98	0.59	0.14	1.57	0.46
		BC _a	0.88	0.65	0.72	0.16	1.37	0.30
	Φ_M	Exact	0.90	0.67	0.64		1.31	
		TRAN	0.90	1.01	0.58	0.14	1.58	0.53
		BC _a	0.86	0.65	0.72	0.16	1.37	0.30
20	Φ	Exact	0.95	1.14	0.31		1.45	
		TRAN	0.85	1.73	0.20	0.48	1.93	1.09
		BC _a	0.86	1.07	0.51	0.18	1.59	0.53
	Φ_M	Exact	0.95	1.22	0.29		1.51	
		TRAN	0.83	1.90	0.11	0.62	2.01	1.31
		BC _a	0.85	1.10	0.51	0.18	1.59	0.58
30	Φ	Exact	0.95	0.99	0.44		1.43	
		TRAN	0.89	1.52	0.33	0.33	1.85	0.92
		BC _a	0.89	0.95	0.59	0.17	1.54	0.44
	Φ_M	Exact	0.95	1.02	0.42		1.44	
		TRAN	0.90	1.47	0.35	0.28	1.82	0.83
		BC _a	0.89	0.94	0.59	0.17	1.53	0.43
50	Φ	Exact	0.95	0.77	0.58		1.35	
		TRAN	0.95	1.18	0.51	0.15	1.69	0.53
		BC _a	0.94	0.77	0.68	0.14	1.45	0.29
	Φ_M	Exact	0.95	0.80	0.56		1.36	
		TRAN	0.992	1.21	0.49	0.16	1.70	0.61
		BC _a	0.91	0.78	0.67	0.15	1.45	0.33

some bivariate population, Hinkley and Wei (1984) examined the normal approximation, the bootstrap-t, and two estimated Edgeworth expansion confidence sets (for the true ratio) obtained by estimating the Edgeworth expansion using two different methods: the jackknife and the substitution. The usual linearization variance estimator is adopted for the bootstrap-t. The bootstrap Monte Carlo size is 999. The simulation size is 10,000, except for the bootstrap method, where 1000 simulations were carried out. The results in Table 4.8 show that the normal approximation is worse than all of the other three methods. The bootstrap-t confidence sets are better than the two estimated Edgeworth expansion confidence sets in terms of coverage probability. The improvements from using the bootstrap for lower confidence bounds are greater than those for two-sided confidence intervals.

Konishi (1991) constructed a confidence interval based on an estimated transformation ϕ_n in assumption (4.14). Tu and Zhang (1992b) found that the transformation can be estimated using the jackknife. They compared this method with the bootstrap BC_a with a estimated by (4.27), in the case where θ and $\hat{\theta}_n$ are the population variance and sample variance, respectively. A bootstrap Monte Carlo size of 1000 was used. The simulation was done by taking 1000 samples from $F = \Phi$, the standard normal distribution, or $F = \Phi_M$, a mixture distribution $0.8N(0, 0.9) + 0.2N(0, 1.4)$. It follows from the results in Table 4.9 that, in terms of the coverage probability, the bootstrap BC_a confidence interval is better than the confidence interval based on the estimated transformation ϕ_n when the nominal level is 0.95 and $n = 20$ and 30, whereas the latter becomes more accurate when the nominal level is 0.9. However, a drawback of the interval based on the estimated ϕ_n is that it is much longer than the bootstrap BC_a interval and its right limit is very variable. The bootstrap BC_a provides confidence limits and lengths much closer to those of the exact confidence interval. Tu and Zhang (1992b) also found that the performances of the bootstrap BC_a confidence intervals with a estimated by (4.26) and (4.27), respectively, are almost the same.

4.4.3 The iterative bootstrap and bootstrap calibration

In the case where θ and $\hat{\theta}_n$ are, respectively, the population mean and sample mean, Martin (1990) performed an empirical simulation study to compare the bootstrap percentile, the hybrid bootstrap, and the bootstrap-t confidence intervals with their corresponding one-step iterative bootstrap confidence intervals obtained by using the iteration technique described at the end of Section 4.3.1. For the hybrid bootstrap and bootstrap-t, with $\mathfrak{R}_n^{(0)}$ being $\sqrt{n}(\bar{X}_n - \theta)$ and the studentized $\sqrt{n}(\bar{X}_n - \theta)$, respectively, the iterated bootstrap confidence intervals are the same as the bootstrap prepivoting confidence intervals [see (4.57)]. The bootstrap Monte Carlo

sizes for the two stages of resampling were chosen to be $B_1 = B_2 = 299$. The results in Table 4.10 are based on 1000 simulation replicates (each of size $n = 10$) from a folded normal population.

It can be concluded from Table 4.10 that the iterated bootstrap percentile method improves the coverage accuracy and has the best relative performance among the three iterative bootstrap methods. The iterated hybrid bootstrap does not improve coverage accuracy. The iterated bootstrap-t corrects for the overcoverage of the original bootstrap-t confidence interval and reduces the length as well. In each case, the EL increases or decreases according to whether the original interval undercovers or overcovers. Also note that the iteration increases the standard deviation of the length, which shows that it produces some additional variation. The choice of relatively small $B_1 = B_2 = 299$ seems to be sufficient in this smooth problem.

In the same problem, Loh (1991) conducted a simulation study comparing the confidence sets obtained by normal approximation and Edgeworth expansion with their corresponding bootstrap calibration confidence intervals. The function h in the calibration (see Section 4.3.2) was selected to be the inverse standard normal distribution function. The bootstrap-t, which is the same as a particular type of bootstrap calibration, was also included in the study. The bootstrap Monte Carlo size was 500 and the simulation size was 5000.

The results in Table 4.11 indicate that the one-sided confidence intervals based on an Edgeworth expansion has better coverage accuracy than that based on normal approximation, in the case where F is exponential. Two-sided confidence intervals based on these two methods are almost the same. When F is normal, the results for the bootstrap-t, calibrated normal approximation (CLB-N), and calibrated Edgeworth expansion (CLB-E) are similar; they are indistinguishable from the results for the noncalibrated intervals for $n = 25$ or 50 and are marginally superior to the latter for $n = 10$. When F is exponential, however, the calibrated confidence sets have substantially better coverage accuracy than the normal approximation

Table 4.10. Comparison of the iterated and original bootstrap confidence intervals for the mean ($1 - 2\alpha = 0.9$) [Adapted from Martin (1990), by permission of American Statistical Association]

Method	Iterated		Iterated		Iterated	
	CP	CP	EL	EL	SD	SD
BP	0.85	0.89	0.59	0.69	0.17	0.26
HB	0.84	0.84	0.59	0.70	0.16	0.29
BT	0.92	0.89	0.90	0.86	0.46	0.58

Table 4.11. Comparison of the calibrated and original confidence sets for the mean ($1 - \alpha = 0.95$) [Adapted from Loh (1991), by permission of International Chinese Statistical Association]

<i>F</i>	<i>n</i>	Method	Lower	Upper	Two-Sided	
			CP	CP	CP	EL
Normal	10	NOR	0.937	0.938	0.875	1.02
		EDG	0.933	0.938	0.871	1.02
		BT	0.951	0.956	0.914	1.18
		CLB-N	0.947	0.948	0.904	1.13
		CLB-E	0.940	0.945	0.891	1.10
	25	NOR	0.949	0.943	0.891	0.65
		EDG	0.949	0.942	0.891	0.65
		BT	0.953	0.948	0.906	0.68
		CLB-N	0.953	0.948	0.906	0.68
	50	CLB-E	0.952	0.947	0.899	0.67
		NOR	0.946	0.945	0.891	0.46
		EDG	0.947	0.945	0.892	0.46
		BT	0.952	0.947	0.896	0.47
		CLB-N	0.948	0.945	0.896	0.47
Exponential	10	CLB-E	0.947	0.944	0.895	0.47
		NOR	0.976	0.844	0.820	0.95
		EDG	0.954	0.872	0.826	0.95
		BT	0.992	0.827	0.879	1.44
		CLB-N	0.968	0.889	0.859	1.13
	25	CLB-E	0.944	0.901	0.853	1.09
		NOR	0.973	0.886	0.859	0.64
		EDG	0.947	0.919	0.866	0.64
		BT	0.989	0.859	0.888	0.73
		CLB-N	0.951	0.927	0.882	0.69
	50	CLB-E	0.944	0.936	0.886	0.68
		NOR	0.974	0.911	0.885	0.46
		EDG	0.951	0.933	0.885	0.46
		BT	0.989	0.888	0.896	0.48
		CLB-N	0.949	0.936	0.895	0.48
		CLB-E	0.948	0.944	0.895	0.47

confidence sets, especially for the one-sided upper bound. The bootstrap-t produces the longest confidence intervals and does not perform well in the one-sided and exponential case.

4.4.4 Summary

- (1) The empirical results in this section are in general agreement with the asymptotic theory in Sections 4.2 and 4.3. The bootstrap-t, the bootstrap BC_a , the iterative bootstrap, and the bootstrap calibration (and the hybrid bootstrap in some cases) have better coverage accuracy than the normal approximation.
- (2) The bootstrap-t, BC_a , and calibration are even better than the method based on an Edgeworth expansion or a transformation, which asymptotically has the same order of accuracy as that of the bootstrap methods.
- (3) The performance of the bootstrap BC_a heavily depends on the accuracy of the estimated acceleration constant a . In most parametric problems, a can be determined exactly or precisely, and the resulting confidence sets are very accurate. In nonparametric problems, however, an accurate estimator of a is not easy to obtain, and the BC_a confidence sets may perform poorly.
- (4) The bootstrap-t confidence intervals tend to overcover and have long lengths.
- (5) The iterative bootstrap and the bootstrap calibration work well in the problems considered here.

4.5 Bootstrap Hypothesis Tests

It is well known that testing hypotheses is related to constructing confidence sets. Although a hypothesis test can be obtained by constructing an appropriate confidence set, bootstrap hypothesis testing is still an important topic for the following reasons. Firstly, sometimes finding a test directly is much easier than getting a test through constructing a confidence set (which is impossible in some cases). Secondly, the tests obtained directly may be better since they usually take account of the special nature of the hypothesis. Thirdly, for the bootstrap confidence sets, we always generate bootstrap data from an estimated distribution \hat{F} without any restriction. For hypothesis testing, we may generate bootstrap data from either \hat{F} or from an estimated distribution under the restrictions specified by the hypothesis. For example, let $X_i = (Y_i, Z_i)'$, $i = 1, \dots, n$, be i.i.d. bivariate data, and we want to test whether Y_1 and Z_1 are independent. We may generate bootstrap data from the empirical distribution putting mass n^{-1} to each X_i or generate bootstrap data $(Y_i^*, Z_i^*)'$, $i = 1, \dots, n$, by gener-

ating independently Y_i^* and Z_i^* from the marginal empirical distributions based on $\{Y_i\}$ and $\{Z_i\}$, respectively. The latter bootstrap method is more efficient if the bootstrap is used to approximate some quantities under the hypothesis that Y_i and Z_i are independent. Some practitioners apply the bootstrap blindly, which leads to a very low power of the constructed test (see, e.g., Hall and Wilson, 1991). Finally, in addition to constructing the test itself, hypothesis testing requires an estimate of the power of the test and/or the calculation of the P -values.

However, the methodology and theory for bootstrap hypothesis testing are not well developed, partly because of the technical difficulties that will be discussed later. In what follows, we give a general description of the problem and introduce some existing results in this area.

4.5.1 General description

Statistical hypothesis testing can be generally described as follows. Let X_1, \dots, X_n be random p -vectors (not necessarily i.i.d.) having joint distribution $F^{(n)}$, and let $\mathcal{F}^{(n)}$ be the collection of all possible $F^{(n)}$. Let $\mathcal{F}_0^{(n)}$ and $\mathcal{F}_1^{(n)}$ be two disjoint subsets of $\mathcal{F}^{(n)}$. We would like to determine, by using the data X_1, \dots, X_n , whether the hypothesis that $F^{(n)} \in \mathcal{F}_0^{(n)}$ is true, i.e., to test

$$H_0 : F^{(n)} \in \mathcal{F}_0^{(n)} \quad \text{versus} \quad H_1 : F^{(n)} \in \mathcal{F}_1^{(n)}. \quad (4.74)$$

In (4.74), H_0 is called the null hypothesis and H_1 is called the alternative hypothesis. In the special but important case where X_1, \dots, X_n are i.i.d. from F , $F^{(n)}$ is determined by F and (4.74) reduces to

$$H_0 : F \in \mathcal{F}_0 \quad \text{versus} \quad H_1 : F \in \mathcal{F}_1, \quad (4.75)$$

where \mathcal{F} is the collection of all possible F , and \mathcal{F}_0 and \mathcal{F}_1 are disjoint subsets of \mathcal{F} .

Example 4.7. Statistical hypotheses. Consider i.i.d. X_1, \dots, X_n .

- (i) Let F_0 be a known distribution, $\mathcal{F}_0 = \{F_0\}$, and $\mathcal{F}_1 = \{F : F \neq F_0\}$. In this case, \mathcal{F}_0 contains only one element and the distribution of X_1, \dots, X_n is known under the null hypothesis. We usually call this type of H_0 a simple null hypothesis.
- (ii) Assume that F has a finite mean μ . Let Θ_0 be a subset of \mathbb{R}^p , $\mathcal{F}_0 = \{F : \mu \in \Theta_0\}$, and $\mathcal{F}_1 = \{F : \mu \notin \Theta_0\}$. In this case, \mathcal{F}_0 contains infinitely many elements even if Θ_0 is a single point set, and the distribution of X_1, \dots, X_n under H_0 is unknown. This type of H_0 is called a complex null hypothesis. Also, in this case, (4.75) can be written as:

$$H_0 : \mu \in \Theta_0 \quad \text{versus} \quad H_1 : \mu \notin \Theta_0, \quad (4.76)$$

where μ is the parameter of interest and $F(\cdot - \mu)$ can be treated as a nuisance parameter.

(iii) When $p = 1$, sometimes we are interested in whether F is symmetric about an unknown center θ . Then, $\mathcal{F}_0 = \{F : F(x - \theta) = 1 - F(\theta - x)\}$. In this case, the null hypothesis H_0 (or \mathcal{F}_0) is even more complicated.

Constructing a test for (4.74) or (4.75) is equivalent to finding a rejection region \mathcal{R}_n such that we reject the null hypothesis H_0 if and only if $(X_1, \dots, X_n) \in \mathcal{R}_n$. A simple and effective method, called the test statistic approach, is to use a test statistic $T_n = T_n(X_1, \dots, X_n)$ and

$$\mathcal{R}_n = \{x : T_n(x) \geq c_n\}, \quad (4.77)$$

where c_n is called the critical value. The rejection region \mathcal{R}_n [or the critical value c_n if \mathcal{R}_n is given by (4.77)] is determined by controlling the probability of rejecting H_0 when H_0 is in fact true (type I error), i.e.,

$$\sup_{F^{(n)} \in \mathcal{F}_0^{(n)}} P\{(X_1, \dots, X_n) \in \mathcal{R}_n | F^{(n)}\} = \alpha, \quad (4.78)$$

where α is a given small number and $P\{\cdot | F^{(n)}\}$ is the probability corresponding to $F^{(n)}$. For a fixed n , if (4.78) holds, then the test with rejection region \mathcal{R}_n is an exact level α test. Unless the problem under consideration is simple, an exact level α test is difficult or impossible to obtain. We then consider large n approximation, i.e., replace (4.78) by

$$\lim_{n \rightarrow \infty} \sup_{F^{(n)} \in \mathcal{F}_0^{(n)}} P\{(X_1, \dots, X_n) \in \mathcal{R}_n | F^{(n)}\} = \alpha. \quad (4.79)$$

However, when $\mathcal{F}_0^{(n)}$ is complex, there may not exist any test satisfying (4.79) (Bahadur and Savage, 1956). This leads to the following definition.

Definition 4.3. Let α be a given nominal level. A test with rejection region \mathcal{R}_n is asymptotically correct if

$$\lim_{n \rightarrow \infty} P\{(X_1, \dots, X_n) \in \mathcal{R}_n | F^{(n)}\} = \alpha$$

for any sequence $F^{(n)} \in \mathcal{F}_0^{(n)}$, $n = 1, 2, \dots$ The test is consistent if

$$\lim_{n \rightarrow \infty} P\{(X_1, \dots, X_n) \in \mathcal{R}_n | F^{(n)}\} = 1$$

for any sequence $F^{(n)} \in \mathcal{F}_1^{(n)}$, $n = 1, 2, \dots$

In the traditional asymptotic approach, we find an asymptotically correct test by deriving an asymptotic distribution under the null hypothesis

H_0 , e.g., for the test of the form (4.77), a derivation of the asymptotic distribution of T_n is required. In view of the superiority of the bootstrap over the traditional asymptotic approach previously discussed, we can use the bootstrap to approximate the required distribution under H_0 , e.g., the distribution of T_n . The resulting test is then called a *bootstrap hypothesis test*.

Alternative to the construction of bootstrap confidence sets, we need to generate bootstrap data from a distribution under the restrictions specified by H_0 , which excludes the empirical distribution of the original data in most cases. If H_0 is simple, we may easily identify the distribution from which the bootstrap data should be generated. For example, in Example 4.7(i), obviously we should generate bootstrap data from F_0 , in which case the bootstrap is the same as the famous Monte Carlo method for hypothesis testing that has been used for many years now. If H_0 is complex, however, it may be difficult to decide from which distribution we should take the bootstrap data. In Example 4.7(ii), there are many distributions having mean μ_0 , and it is not clear which distribution should be used. Some existing solutions to this problem will be introduced later.

We will also discuss the other two applications of the bootstrap in hypothesis testing problems: the estimation of the power and the P -value of a given test.

4.5.2 Two-sided hypotheses with nuisance parameters

Let X_1, \dots, X_n be i.i.d. from F which can be indexed by $\theta \in \Theta$ and $\vartheta \in \Omega$, i.e., $F = F_{\theta, \vartheta}$. If both Θ and Ω are subsets of Euclidean spaces, then F belongs to a parametric family. But the problem treated here is nonparametric since Θ and Ω can be any metric spaces [see Example 4.7(ii)]. The parameter θ is the quantity of interest and ϑ is a nuisance parameter. The hypothesis under consideration is of the form given by (4.75) with

$$\mathcal{F}_0 = \{F_{\theta, \vartheta} : \theta = \theta_0, \vartheta \in \Omega\},$$

where θ_0 is a specified value in Θ , and

$$\mathcal{F}_1 = \{F_{\theta, \vartheta} : \theta \in \Theta, \theta \neq \theta_0, \vartheta \in \Omega\}.$$

We adopt the test statistic approach. Let T_n be a given test statistic for this problem, and let its distribution be

$$K_{n, \theta, \vartheta}(x) = P_{\theta, \vartheta}\{T_n \leq x\},$$

where $P_{\theta, \vartheta}$ is the probability corresponding to the joint distribution $F_{\theta, \vartheta}$. Suppose that we have an estimator $\hat{\vartheta}_n$ of the nuisance parameter ϑ and

that $\hat{\vartheta}_n$ is consistent under H_0 in the sense that

$$P_{\theta_0, \vartheta} \{ \rho(\hat{\vartheta}_n, \vartheta) > \epsilon \} \rightarrow 0$$

for any $\epsilon > 0$, where ρ is a metric on Ω . Assuming that $K_{n, \theta_0, \vartheta}$ is continuous, we obtain the following bootstrap estimator of c_n in (4.77):

$$\hat{c}_n = K_{n, \theta_0, \hat{\vartheta}_n}^{-1} (1 - \alpha).$$

The resulting bootstrap test is then the test with rejection region

$$\mathcal{R}_n = \{ T_n(X_1, \dots, X_n) \geq \hat{c}_n \}. \quad (4.80)$$

The computation of \hat{c}_n can be done by Monte Carlo: generate B independent sets of bootstrap samples $\{X_{1b}^*, \dots, X_{nb}^*\}$, $b = 1, \dots, B$, from the distribution $F_{\theta_0, \hat{\vartheta}_n}$. Then \hat{c}_n is approximated by the $\{[(1 - \alpha)B] + 1\}$ th order statistic of $T_n(X_{1b}^*, \dots, X_{nb}^*)$, $b = 1, \dots, B$.

The power function of the bootstrap test (4.80) is

$$\beta_n(\theta, \vartheta) = P_{\theta, \vartheta} \{ T_n \geq \hat{c}_n \} \quad \theta \in \Theta, \theta \neq \theta_0.$$

For any given $\theta \neq \theta_0$, $\beta_n(\theta, \vartheta)$ can be estimated by the bootstrap estimator

$$\hat{\beta}_n(\theta) = \beta_n(\theta, \hat{\vartheta}_n).$$

The computation of $\hat{\beta}_n(\theta)$ can be done by using a nested bootstrap algorithm. For a fixed $\theta \neq \theta_0$, generate B_1 independent sets of bootstrap samples $\{X_{1b}^*, \dots, X_{nb}^*\}$, $b = 1, \dots, B_1$, from $F_{\theta, \hat{\vartheta}_n}$. For each b , generate B_2 independent sets of bootstrap samples from $F_{\theta_0, \hat{\vartheta}_{nb}^*}$, where $\hat{\vartheta}_{nb}^* = \hat{\vartheta}_n(X_{1b}^*, \dots, X_{nb}^*)$, and use the method described above to compute \hat{c}_{nb} . Then $\hat{\beta}_n(\theta)$ is approximated by $B_1^{-1} \sum_{b=1}^{B_1} I\{T_n(X_{1b}^*, \dots, X_{nb}^*) \geq \hat{c}_{nb}\}$.

Under some assumptions on the convergence of $\hat{\vartheta}_n$ and the distribution of T_n under H_0 and H_1 , Beran (1986) showed that the bootstrap test (4.80) is asymptotically correct and

$$\sup_{\theta} |\hat{\beta}_n(\theta) - \beta_n(\theta, \vartheta)| \rightarrow_{a.s.} 0,$$

i.e., the bootstrap power estimator is consistent uniformly in θ .

This method can be applied to the situations where an estimator of the nuisance parameter is available. The following are some examples.

Example 4.8. Test for the population mean. Consider i.i.d. X_1, \dots, X_n from $F(x) = G(x - \mu)$, where $G(x)$ has mean 0 and covariance matrix $\Sigma_G > 0$. We take $\theta = \mu$ and $\vartheta = G$ and consider the test problem

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0,$$

which is a special case of (4.76). We take the usual t-test statistic

$$T_n = \|\sqrt{n}\hat{\Sigma}^{-1/2}(\bar{X}_n - \mu_0)\|,$$

where \bar{X}_n and $\hat{\Sigma}$ are the usual sample mean and sample covariance matrix, respectively. To apply the bootstrap method, we need an estimator of G . Beran (1986) showed that the empirical distribution G_n based on $\{X_i - \bar{X}_n, i = 1, \dots, n\}$ is a consistent estimator of G with the Lévy metric under both H_0 and H_1 . Therefore, the bootstrap sampling can be done from $G_n(x - \mu_0)$ for the computation of \hat{c}_n , and from $G_n(x - \mu)$ for the computation of the power estimator $\hat{\beta}_n$. Beran (1986) verified the conditions for the asymptotic correctness of the bootstrap test and the consistency of the bootstrap power estimator in this example.

Example 4.9. Test of multivariate normality. Suppose that X_1, \dots, X_n are i.i.d. from $G(\Sigma^{-1/2}(x - \mu))$, where $\mu \in \mathbb{R}^p$, Σ is a $p \times p$ positive definite matrix, and G is a continuous distribution on \mathbb{R}^p with mean 0, covariance matrix I_p , and finite fourth moments. We take $\theta = G$ and treat $\vartheta = (\mu, \Sigma)$ as the nuisance parameter. Consider the test

$$H_0 : G = \Phi_p \quad \text{versus} \quad H_1 : G \neq \Phi_p,$$

where Φ_p is the distribution function of $N_p(0, I_p)$. The test statistic is

$$T_n = \sqrt{n}\|F_n(\cdot) - \Phi_p(\hat{\Sigma}^{-1/2}(\cdot - \bar{X}_n))\|,$$

where F_n is the empirical distribution, $\|\cdot\|$ is a norm on the space of distribution functions on \mathbb{R}^p , and the other notations are the same as those in Example 4.8. The nuisance parameter in this case can be estimated by $\hat{\vartheta}_n = (\bar{X}_n, \hat{\Sigma})$. The consistency of $\hat{\vartheta}_n$ with a suitable metric under H_0 is discussed by Beran (1986). We can then use the bootstrap test (4.80) by generating bootstrap data from $\Phi_p(\hat{\Sigma}^{-1/2}(\cdot - \bar{X}_n))$ and the bootstrap power estimator by generating bootstrap data from $G(\hat{\Sigma}^{-1/2}(\cdot - \bar{X}_n))$ for any fixed G .

Example 4.10. Test of symmetry about the origin. Let X_1, \dots, X_n be i.i.d. random variables from a continuous F . We want to test whether F is symmetric about the origin. Define

$$\theta(x) = [F(x) + F(-x) - 1]/2 \quad \text{and} \quad \vartheta(x) = [F(x) - F(-x) + 1]/2.$$

Then this problem is equivalent to testing

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta \neq 0.$$

In this example, both parameters θ and ϑ are functions. Let F_n be the empirical distribution. We use the test statistic $T_n = \sqrt{n}\|\hat{\theta}_n\|$, where $\hat{\theta}_n(x)$

is obtained by replacing F in $\theta(x)$ with $F_n(x)$ and $\|\cdot\|$ is the same as that in Example 4.9, and the consistent estimator of ϑ : $\hat{\vartheta}_n(x) = [F_n(x) - F_n(-x) + 1]/2$. Since $F(x) = \theta(x) + \vartheta(x)$ and $\theta(x) \equiv 0$ under H_0 , we take bootstrap data from $\hat{\vartheta}_n(x)$ to compute \hat{c}_n for the bootstrap test.

Some more results can be found in Beran (1986) and Stute, Manteiga and Quindimil (1993).

4.5.3 Bootstrap distance tests

Romano (1988a, 1989) considered the application of the bootstrap to problems with hypotheses tested using some distance test statistic. Suppose that the hypothesis under consideration is given by (4.75) and \mathcal{F}_0 can be characterized by a mapping τ from \mathcal{F} to \mathcal{F}_0 as follows:

$$\mathcal{F}_0 = \{F \in \mathcal{F} : \tau(F) = F\}.$$

Some examples are given later. Let ρ be a metric on \mathcal{F} . Then \mathcal{F}_0 can be specified by $\rho(F, \tau(F)) = 0$. Let F_n be the empirical distribution. Then the departure of $\rho(F_n, \tau(F_n))$ from 0 can be used to test the validity of the null hypothesis. This leads to a nonparametric distance test statistic

$$T_n = \sqrt{n}\rho(F_n, \tau(F_n)).$$

The computation of T_n may not be easy. A stochastic approximation has been suggested by Beran and Millar (1987) to overcome this difficulty. We now need to find the critical value c_n . Let $H_{n,F}$ be the distribution of T_n . Then the critical value for the bootstrap test (4.80) is

$$\hat{c}_n = H_{n,\tau(F_n)}^{-1}(1 - \alpha).$$

The computation of \hat{c}_n can be done by Monte Carlo. Under some conditions Romano (1988a, 1989) showed the asymptotic correctness and consistency of this bootstrap distance test.

Example 4.7(i) (continued). For testing

$$H_0 : F = F_0 \quad \text{versus} \quad H_1 : F \neq F_0$$

with a known F_0 , we can define $\tau(F) \equiv F_0$. To compute the bootstrap critical value, we generate bootstrap data from F_0 . This is equivalent to the test obtained by Monte Carlo.

Example 4.11. Test of independence. Let X_1, \dots, X_n be i.i.d. random p -vectors from F . We want to test whether the components of X_i are

independent. Let F^j be the j th marginal distribution of F . Then the hypothesis is

$$H_0 : F = \prod_{j=1}^p F^j \quad \text{versus} \quad H_1 : F \neq \prod_{j=1}^p F^j.$$

For any distribution G with marginals G^j , $j = 1, \dots, p$, we define $\tau(G) = \prod_{j=1}^p G^j$. Then $F \in \mathcal{F}_0$ if and only if $\tau(F) = F$. The nonparametric test statistic is then $T_n = \sqrt{n}\rho(F_n, \prod_{j=1}^p F_n^j)$. The critical value can be obtained by bootstrap resampling from $\prod_{j=1}^p F_n^j$.

In the above examples there exists an obvious choice of τ . In general, we can apply the following method to determine τ . For any F , choose $\tau(F)$ such that

$$\rho(F, \tau(F)) = \inf_{G \in \mathcal{F}_0} \rho(F, G), \quad (4.81)$$

where ρ is a metric on \mathcal{F} . If this metric is the same as that used in the test statistic T_n , then

$$T_n = \sqrt{n} \inf_{G \in \mathcal{F}_0} \rho(F_n, G),$$

which is called the nonparametric minimum distance test statistic (Beran and Millar, 1987, 1989).

Example 4.12. Goodness of fit test. Let X_1, \dots, X_n be i.i.d. We want to test whether F is in a parametric family $\mathcal{F}_0 = \{F_\theta : \theta \in \Theta\}$. Let $\hat{\theta}_n$ be the minimum distance estimator of θ defined by

$$\rho(F_n, F_{\hat{\theta}_n}) = \inf_{G \in \mathcal{F}_0} \rho(F_n, G) = \inf_{\theta \in \Theta} \rho(F_n, F_\theta).$$

Then we can take $\tau(F_n) = F_{\hat{\theta}_n}$ and the test statistic becomes $T_n = \sqrt{n}\rho(F_n, F_{\hat{\theta}_n})$. The bootstrap critical value can be computed by bootstrap resampling from $F_{\hat{\theta}_n}$.

An empirical simulation study for this type of problem is in Romano (1988a). He considered $\mathcal{F}_0 = \{F_\theta \text{ has density } f_\theta : \theta \in (-1, \infty)\}$, where $f_\theta = (\theta + 1)(\theta + 2)x^\theta(1 - x)$. The test statistic is the T_n described above with $\rho = \rho_\infty$, and $\hat{\theta}_n$ = the maximum likelihood estimator of θ under H_0 . The bootstrap critical value was computed by drawing 200 bootstrap samples from $f_{\hat{\theta}_n}$. For some choices of α , n , and θ , the simulation estimated levels of the bootstrap test (the actual number of times that the bootstrap test rejected H_0 in 1000 simulation trials) are listed in Table 4.12. The estimated levels are reasonably close to the nominal level α .

More complicated examples and discussions about the properties of the bootstrap distance tests can be found in Romano (1988a, 1989).

Table 4.12. Estimated level of bootstrap testing a Beta subfamily [Adapted from Romano (1988a), by permission of American Statistical Association]

n	θ	α				
		0.01	0.05	0.10	0.20	0.30
15	0	0.008	0.52	0.103	0.204	0.321
	1	0.011	0.048	0.101	0.210	0.318
	2	0.015	0.053	0.109	0.221	0.333
30	0	0.008	0.050	0.102	0.204	0.309
	2	0.013	0.054	0.105	0.211	0.322

4.5.4 Other results and discussions

An optimization approach

As we discussed previously, a difficult problem in bootstrap hypothesis testing is the selection of a distribution in \mathcal{F}_0 for generating bootstrap data. Hinkley (1987, 1988) proposed an optimization approach. He suggested that we take bootstrap data from F_{n0} satisfying

$$\rho(F_n, F_{n0}) = \inf_{G \in \mathcal{F}_0} \rho(F_n, G), \quad (4.82)$$

where ρ is a metric on \mathcal{F} and F_n is the empirical distribution. Although this idea is appealing, it is usually difficult to find the solution F_{n0} in (4.82) when \mathcal{F}_0 is complex. When $\tau(F)$ is given by (4.81), we have $F_{n0} = \tau(F_n)$. Hence, Example 4.12 gives a case where we can use this optimization approach. We now provide two more examples.

Example 4.8 (continued). Assume that $p = 1$. To find a distribution F_{n0} satisfying (4.82) for some metric ρ , we may restrict F_{n0} to a class of distributions whose support is $\{X_1, \dots, X_n\}$ (Efron, 1981a); that is, we assume that

$$\mathcal{F}_0 = \left\{ F : F(x) = \sum_{i=1}^n w_i I\{X_i \leq x\}, \sum_{i=1}^n w_i = 1 \text{ and } \sum_{i=1}^n w_i X_i = \mu_0 \right\}.$$

For $F \in \mathcal{F}_0$, we may define the following two distances between F and F_n : (1) $\rho_a(F_n, F) = \sum_{i=1}^n w_i \log(nw_i)$ (Kullback-Leibler's metric); (2) $\rho_b = -n^{-1} \sum_{i=1}^n \log(nw_i)$.

Efron (1981a) found that the solution of (4.82) when $\rho = \rho_a$ is

$$F_{n0}^{(a)}(x) = \sum_{i=1}^n \exp(aX_i) I\{X_i \leq x\} \Bigg/ \sum_{i=1}^n \exp(aX_i),$$

Table 4.13. Comparison of the power of bootstrap tests
 [Adapted from Young (1988), by permission of Physica-Verlag]

Resampling distribution	μ_0 ($F = \text{Normal}$)			μ_0 ($F = \text{Exponential}$)		
	0.0	-0.2	-0.4	0.0	-0.2	-0.4
$F_{n0}^{(a)}$	0.053	0.219	0.536	0.038	0.230	0.686
$F_{n0}^{(b)}$	0.053	0.221	0.538	0.039	0.234	0.673
$G_n(\cdot - \mu_0)$	0.048	0.204	0.513	0.032	0.206	0.664

where a is uniquely defined by $\int x dF_{n0}^{(a)}(x) = \mu_0$. According to Owen (1988), the solution of (4.82) when $\rho = \rho_b$ is

$$F_{n0}^{(b)}(x) = \sum_{i=1}^n \frac{1}{1 + b(X_i - \mu_0)} I\{X_i \leq x\} \Bigg/ \sum_{i=1}^n \frac{1}{1 + b(X_i - \mu_0)},$$

where b is uniquely defined by $\int x dF_{n0}^{(b)}(x) = \mu_0$.

Young (1988) compared the power of the bootstrap tests with bootstrap sampling from $F_{n0}^{(a)}$, $F_{n0}^{(b)}$, or $G_n(x - \mu_0)$ (given in Section 4.5.2). The test statistic is the usual t-statistic. In the simulation, F is either standard normal or exponential with variance 1 centered at 0. The sample size is $n = 20$ and the nominal level is $\alpha = 0.05$. The results based on 20,000 simulations are shown in Table 4.13. The bootstrap tests taking bootstrap data from $F_{n0}^{(a)}$ or $F_{n0}^{(b)}$ are more powerful than the test taking bootstrap data from $G_n(\cdot - \mu_0)$.

Example 4.7(iii) (continued). Consider the problem of testing whether F is symmetric about an unknown center θ . Define

$$h_n(\theta) = \sup_x |F_n(x) - 1 + F_n((2\theta - x)-)|,$$

where $G(x-)$ is the left limit of G at x . An estimator $\hat{\theta}_n$ of θ satisfying

$$h_n(\hat{\theta}_n) = \min_{\theta} h_n(\theta)$$

is

$$\hat{\theta}_n = [m(l) + M(l)]/2,$$

where $m(k) = \max\{(X_i + X_j)/2 : 1 \leq i \leq [(n-k+1)/2], j = n-k+1-i\}$, $M(k) = \min\{(X_i + X_j)/2 : k+1 \leq i \leq [(n+k+1)/2], j = n+k+1-i\}$, $k = 0, 1, 2, \dots, n-1$, and $l = \min\{k : m(k) \leq M(k)\}$. The test statistic for this problem suggested by Schuster and Barker (1987) is

$$T_n = nh_n(\hat{\theta}_n).$$

To estimate the critical value for the bootstrap test (4.80), Schuster (1987) found that the distribution minimizing $\rho_\infty(F_n, G)$ over all symmetric G is

$$F_{n0}(x) = [F_n(x) + 1 - F_n((2\hat{\theta}_n - x)-)]/2,$$

which is the empirical distribution of the $2n$ points $X_i, 2\hat{\theta}_n - X_i, i = 1, \dots, n$. Schuster and Barker (1987) presented some simulation results of the bootstrap test (4.80) using F_{n0} to generate bootstrap data. Arcones and Giné (1991) showed the asymptotic correctness and consistency of this bootstrap test.

P-value

Let $X_1 = x_1, \dots, X_n = x_n$ be the observed values. For the test of the form (4.77), its P -value is defined to be

$$p_n = P\{T_n(X_1, \dots, X_n) \geq T_n(x_1, \dots, x_n) | F^{(n)}\}$$

for any $F^{(n)} \in \mathcal{F}_0^{(n)}$. Here we assume that p_n is uniquely defined. A bootstrap estimator of p_n is

$$\hat{p}_n = P_*\{T_n(X_1^*, \dots, X_n^*) \geq T_n(x_1, \dots, x_n)\},$$

where X_1^*, \dots, X_n^* are the bootstrap data i.i.d. from the empirical distribution based on x_1, \dots, x_n . Some examples can be found in Noreen (1989) and Krewski *et al.* (1991a,b).

Assume that $t_0 = E[T_n(X_1, \dots, X_n)]$ is known when $F \in \mathcal{F}_0$. Let $\hat{t}_0 = E_*[T_n(X_1^*, \dots, X_n^*)]$. Noreen (1989) suggested that under the assumption

$$P_*\{T_n(X_1^*, \dots, X_n^*) - \hat{t}_0 \geq x\} \approx P\{T_n(X_1, \dots, X_n) - t_0 \geq x\}, \quad (4.83)$$

we should use the following bootstrap estimator of the P -value obtained by shifting the center of the bootstrap distribution:

$$\tilde{p}_n = P_*\{T_n(X_1^*, \dots, X_n^*) \geq T_n(x_1, \dots, x_n) - t_0 + \hat{t}_0\}.$$

This estimator is valid if (4.83) holds. Some numerical examples are given in Noreen (1989). If F_{n0} in (4.82) is available, then another estimate of p_n is

$$P_*\{T_n(\tilde{X}_1^*, \dots, \tilde{X}_n^*) \geq T_n(x_1, \dots, x_n)\},$$

where $\{\tilde{X}_1^*, \dots, \tilde{X}_n^*\}$ is a bootstrap sample from F_{n0} .

k-sample problems

It is not difficult to generalize the bootstrap confidence procedures to k -sample problems (Hall and Martin, 1988a). For bootstrap hypothesis testing, however, there are many choices of distributions for generating bootstrap data. Boos, Janssen and Veraverbeke (1989) provided an example.

Example 4.13. Test for scales. Suppose that we have two independent samples $\{X_1, \dots, X_n\}$ and $\{Y_1, \dots, Y_m\}$ from F (with a scale parameter θ_F) and G (with a scale parameter θ_G), respectively. We want to test

$$H_0 : \theta_F = \theta_G \quad \text{versus} \quad H_1 : \theta_F > \theta_G.$$

Let $\hat{\theta}_F$ and $\hat{\theta}_G$ be estimators of θ_F and θ_G , respectively, and let

$$T_{m,n} = [mn/(m+n)]^{1/2}(\log \hat{\theta}_F - \log \hat{\theta}_G)$$

be the test statistic. To compute the critical value of the bootstrap test, we have at least four different resampling plans available:

- (1) Take bootstrap data from the empirical distribution of the combined sample $\{X_1, \dots, X_n, Y_1, \dots, Y_m\}$.
- (2) Take bootstrap data from the empirical distributions of X -sample and Y -sample separately.
- (3) Take bootstrap data from the empirical distribution based on the location aligned and then combined sample $\{X_1 - \hat{\mu}_F, \dots, X_n - \hat{\mu}_F, Y_1 - \hat{\mu}_G, \dots, Y_m - \hat{\mu}_G\}$, where $\hat{\mu}_F$ and $\hat{\mu}_G$ are estimators of the location parameters of F and G , respectively.
- (4) Take bootstrap data from the empirical distributions of $\{X_1/\hat{\theta}_F, \dots, X_n/\hat{\theta}_F\}$ and of $\{Y_1/\hat{\theta}_G, \dots, Y_m/\hat{\theta}_G\}$ separately.

Asymptotic analysis shows that the bootstrap test derived from the first resampling plan is not asymptotically correct; the second plan produces an inconsistent bootstrap test; both the third and fourth plans produce asymptotically correct and consistent bootstrap tests. However, some small sample simulation results showed that the third plan is better.

A similar conclusion was obtained by Boos and Brownie (1989) for the test of homogeneity of k -sample variances. Other examples can be found in Young (1988), Hinkley (1988), and Chen and Loh (1991).

Discussions

We have introduced some bootstrap methods in testing some hypotheses. It should be noted that there are only a limited number of problems that fit quite well into the framework assumed by these approaches. A simple example in which the above methods cannot be used is the test of

$$H_0 : F \in \{F_\theta : \theta \leq \theta_0\} \quad \text{versus} \quad H_1 : F \in \{F_\theta : \theta > \theta_0\}$$

for a fixed θ_0 .

The proper selection of test statistics before bootstrapping is a key factor in improving the order of correctness of the bootstrap tests. Although

the bootstrap tests may be better than the tests based on normal approximation in terms of mean squared error (Liu and Singh, 1987), a test based on a less pivotal test statistic may result in a large difference between the actual level of the test and the nominal level. Bunke and Riemer (1983) demonstrated this for the problem of testing for the mean. Ducharme and Jhun (1986) and Quenneville (1986) showed that the performance of the bootstrap tests can be improved by using a studentized test statistic. Sutton (1993) suggested using Johnson's modified-t statistic to achieve accurate bootstrap tests about the mean of an asymmetric distribution. More advantages of using a pivot as a test statistic are given in Hall and Wilson (1991). Beran (1988a) suggested a bootstrap prepivoting method to increase the order of correctness of the bootstrap tests. The procedure is similar to that discussed for the confidence sets except that the bootstrap sampling should be done from a distribution in $\mathcal{F}_0^{(n)}$ (the collection of distributions under the null hypothesis).

Once the distribution from which we shall generate bootstrap data is determined, the computation of the bootstrap test involves the computation of quantiles of some bootstrap distributions. Some efficient algorithms will be introduced in the next chapter.

4.6 Conclusions and Discussions

- (1) We have introduced some bootstrap confidence sets and studied their asymptotic properties (consistency and order of accuracy) and fixed sample performances. Compared with the confidence sets obtained by using the traditional normal approximation, some bootstrap confidence sets are more accurate (e.g., the bootstrap-t and BC_a), whereas some bootstrap confidence sets have the same accuracy but require neither theoretical derivation nor variance estimation (e.g., the bootstrap percentile, BC, and hybrid bootstrap). In terms of improving the coverage accuracy, the bootstrap methods are better than those based on Edgeworth expansions or transformations.
- (2) Although some bootstrap methods are better than the others (second order accuracy versus first order accuracy), they require either a variance estimator (for the bootstrap-t), an estimator of the acceleration constant (for the bootstrap BC_a), or additional heavy computations (for the iterative bootstrap). Thus, less accurate but easy-to-use confidence sets are still useful in applications.
- (3) The iterative bootstrap and the bootstrap calibration seem to be two promising methods to obtain accurate confidence sets. Further research in computational issues for the former and in the choice of calibration methods for the latter is needed.

- (4) In this chapter, the asymptotic theory for the bootstrap confidence sets is established on the basis of heuristic arguments. Rigorous treatments in some situations (e.g., smooth functions of the sample mean) can be found in the literature (e.g., Hall, 1988b, 1992d). Rigorous proofs of the asymptotic results for general situations are difficult and are still being developed. A comparative and empirical study of bootstrap confidence sets is necessary in the situations where theoretical results are not available.
- (5) There are some other techniques for constructing confidence sets (e.g., conditioning, empirical Bayesian, etc.). In principle, the bootstrap can be applied to any situation where we need an approximation of the distribution of some variable. But the implementation of the bootstrap requires some specific considerations (see, e.g., Hinkley and Schechtman (1987), Laird and Louis (1987), and Davison, Hinkley and Worton (1992)).
- (6) In this chapter, we have treated mostly the construction of confidence sets for a real-valued parameter. There are some investigations into the problem of constructing confidence sets for some other types of parameters. For example, Hall and Martin (1988a) considered bootstrap confidence intervals for two sample parameters; Ducharme *et al.* (1985) and Fisher and Hall (1989) proposed some bootstrap confidence sets for directional data; Bickel and Freedman (1981), Bickel and Krieger (1989), and Csörgö and Mason (1989) considered some bootstrap confidence bands for distribution and quantile functions.
- (7) We have presented some results in bootstrap hypothesis testing, which is not a well-developed topic. The main difficulty is the selection of a distribution (under the null hypothesis) for generating bootstrap data. Theoretical and empirical studies in this area are still called for. Some applications in multivariate analysis will be discussed in Chapter 8.
- (8) The bootstrap methods introduced in this chapter can be applied to both parametric and nonparametric models, although “the bulk of published work concerns the nonparametric case, for this is where the fun lies and the most immediate practical gains, in terms of quick error estimates and p-values, might be expected” (Young, 1994). There are also many applications of the bootstrap in parametric models (e.g., Efron, 1985; Loh, 1985; Bickel and Ghosh, 1990; Martin, 1990; DiCiccio and Efron, 1992; Hinkley, 1994). More examples can be found in Efron and Tibshirani (1993). Even under a parametric model, the parametric bootstrap (Example 1.6) may not always produce the best results (Doss and Chiang, 1994). An adaptive procedure for choosing the parametric and nonparametric bootstrap methods was proposed in Lee (1994).

Chapter 5

Computational Methods

The jackknife and the bootstrap are computer-intensive methods. Modern computers can meet many computing needs in applying these methods. However, research on the computation of the jackknife and bootstrap estimators is still important, because of the following reasons:

- (1) There are cases (e.g., the delete-d jackknife and the iterative bootstrap) where the computation is cumbersome or even impractical.
- (2) More efficient methods for computation are always welcome for saving time and computing expenses.
- (3) Questions such as how many bootstrap data sets should be taken in bootstrap Monte Carlo approximations are often asked by practitioners before they apply the bootstrap method.

In this chapter, we introduce some existing results in efficient numerical computation (approximation) for the jackknife and bootstrap. For simplicity in presentation, we concentrate on the simple case where the data X_1, \dots, X_n are i.i.d. from an unknown distribution F . However, most of the conclusions obtained in this chapter are applicable to other situations, as will be discussed in later chapters.

5.1 The Delete-1 Jackknife

We focus on the jackknife variance estimator v_{JACK} for a given statistic T_n . Only in limited cases can we obtain an explicit formula for v_{JACK} . Two examples are shown in Chapter 1 (Examples 1.1 and 1.2). Another example is the L-statistic of the form

$$T_n = \sum_{i=1}^n c_{ni} X_{(i)}, \quad (5.1)$$

where $X_{(i)}$ is the i th order statistic of the sample X_1, \dots, X_n and c_{ni} are some constants. The trimmed sample mean in Example 1.2 is a special case of (5.1). Huang (1991) showed that $v_{\text{JACK}} = \mathbf{Y}' \mathbf{W} \mathbf{Y}$, where $\mathbf{Y} = (X_{(1)}, \dots, X_{(n)})'$ and \mathbf{W} is an $n \times n$ matrix with the i th diagonal element $(n-1)n^{-1}[(n-i)c_{ni}^2 + (i-1)c_{n(i-1)}^2 - n^{-1}a_{ni}]$, $a_{ni} = (n-1)c_{ni} + (i-1)c_{n(i-1)}$ (c_{nt} is defined to be 0 if $t \leq 1$ or $> n$), and the (i, j) th off diagonal element $(n-1)n^{-1}[(n-j)c_{ni}c_{nj} + (j-i-1)c_{ni}c_{n(j-1)} + (i-1)c_{n(i-1)}c_{n(j-1)} - n^{-1}a_{ni}a_{nj}]$, $i < j$. Huang (1991) also provides an explicit formula of the delete-d jackknife variance estimator $v_{\text{JACK-d}}$ for T_n given by (5.1).

In general, the computation of v_{JACK} involves the calculation of the given statistic n times. This may be cumbersome in either one of the following situations:

- (1) The evaluation of T_n is difficult, e.g., T_n is a fixed point of an iterative process.
- (2) The sample size n is very large.

Section 5.1.1 introduces a method, the one-step jackknife, for situation (1). Some approximations to v_{JACK} in situation (2) are discussed in Section 5.1.2.

5.1.1 The one-step jackknife

Some efforts have been made in modifying the jackknife estimator to simplify its computation when the computation of T_n requires some iterations. For the least squares estimator in a nonlinear model, Fox, Hinkley and Larntz (1980) proposed a linear jackknife estimator that uses the linear term in the Taylor expansion to approximate the jackknife estimator. In the situation where T_n is a fixed point of an iterative process

$$T_n = \lim_{k \rightarrow \infty} T_n^{[k]}, \quad T_n^{[k+1]} = g(T_n^{[k]}, F_n), \quad k = 0, 1, 2, \dots, \quad (5.2)$$

where F_n is the empirical distribution of X_1, \dots, X_n and $g(t, G)$ is an explicitly known functional of t and G , Jorgensen (1987) considered approximations based on the outcome of a single step of iteration. Define $\tilde{F}_{n,i,\epsilon} = (F_n - \epsilon \delta_{X_i})/(1-\epsilon)$. Let $T_{n,i,\epsilon}^{[1]} = g(T_n, \tilde{F}_{n,i,\epsilon})$ be the first iteration with initial point T_n , $i = 1, \dots, n$, and let $g'(t, G) = \partial g(t, G)/\partial t$. Jorgensen (1987) approximates the jackknife variance estimator v_{JACK} for T_n by

$$v_{n,\epsilon}^{[1]} = \frac{1-\epsilon}{n^2 \epsilon^2 [1 - g'(T_n, F_n)]^2} \sum_{i=1}^n \left(T_{n,i,\epsilon}^{[1]} - \frac{1}{n} \sum_{j=1}^n T_{n,j,\epsilon}^{[1]} \right)^2, \quad (5.3)$$

which requires no iteration in its computation. For fixed n , Jorgensen showed that $v_{n,\epsilon}^{[1]}$ converges to the infinitesimal jackknife estimator given in

(2.39) as $\epsilon \rightarrow 0$. In practice, ϵ is usually taken to be n^{-1} and therefore $\hat{F}_{n,i,\epsilon} = F_{n-1,i}$, the empirical distribution based on the $n - 1$ observations $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$.

The most popular estimators requiring iterations in their computation are the M-estimators defined in Example 2.5, i.e., T_n is a solution of

$$\int r(x, T_n) dF_n(x) = \min_t \int r(x, t) dF_n(x).$$

We now confine our attention to the situation where T_n is an M-estimator of a parameter θ and Newton's method is used in evaluating T_n . That is, T_n is the fixed point of the iterative process (5.2) with

$$g(t, G) = t - \lambda_G(t)/\lambda'_G(t), \quad (5.4)$$

where $\lambda_G(t) = \int \psi(x, t) dG(x)$ and $\psi(x, t) = \partial r(x, t)/\partial t$. With an initial point $T_n^{[0]}$, $T_n^{[1]} = g(T_n^{[0]}, F_n)$ is called the one-step M-estimator in the literature and its asymptotic property is well known: T_n and $T_n^{[1]}$ are asymptotically normal with the same mean and variance, provided that $T_n^{[0]}$ satisfies $\sqrt{n}(T_n^{[0]} - \theta) = O_p(1)$. Let $v_{\text{JACK}}^{[1]}(S_0)$ be the jackknife variance estimator for the one-step M-estimator $T_n^{[1]}$, i.e.,

$$v_{\text{JACK}}^{[1]}(S_0) = \frac{n-1}{n} \sum_{i=1}^n \left[T_{n-1,i}^{[1]}(S_0) - \frac{1}{n} \sum_{j=1}^n T_{n-1,j}^{[1]}(S_0) \right]^2,$$

where $T_{n-1,i}^{[1]}(S_0) = g(S_0, F_{n-1,i})$ and S_0 is the initial point used in computing $T_{n-1,i}^{[1]}$. Since T_n and $T_n^{[1]}$ have the same asymptotic variance, $v_{\text{JACK}}^{[1]}(S_0)$ can also be used as a variance estimator for T_n . The computation of $v_{\text{JACK}}^{[1]}(S_0)$ needs no iteration and is much simpler than the computation of v_{JACK} . The estimator $v_{\text{JACK}}^{[1]}(S_0)$ is called a *one-step jackknife estimator*.

We can take any one of T_n , $T_n^{[1]}$, or $T_n^{[0]}$ as the initial point S_0 . If $S_0 = T_n$, then $v_{\text{JACK}}^{[1]}(S_0) = v_{n,\epsilon}^{[1]}$ in (5.3) with $\epsilon = n^{-1}$. This is true since, for the functional g in (5.4), $g'(T_n, F_n) = 0$ when T_n is an M-estimator.

Note that T_n may not be available due to its computational complexity. If $T_n^{[1]}$, instead of T_n , is used as the point estimator of θ , it is more natural to use $T_n^{[1]}$ or $T_n^{[0]}$ as the initial point in computing $T_{n-1,i}^{[1]}$, $i = 1, \dots, n$, for the purpose of estimating the dispersion of $T_n^{[1]}$.

The following theorem shows that if the initial point S_0 is properly chosen, $v_{\text{JACK}}^{[1]}(S_0)$ with different S_0 are asymptotically equivalent, and they are all consistent estimators of the asymptotic variance of T_n . This gives us some flexibility: we may use the one-step jackknife estimator whose computation is the simplest and the most inexpensive.

Theorem 5.1. Assume that for almost all x , $\psi(x, t)$ and $\partial\psi(x, t)/\partial t$ are continuous on a neighborhood \mathcal{N} of θ and there are functions $h_j(x)$, $j = 1, 2$, such that

$$\sup_{t \in \mathcal{N}} [\psi(x, t)]^2 \leq h_1(x), \quad \sup_{t \in \mathcal{N}} [\partial\psi(x, t)/\partial t]^2 \leq h_2(x)$$

and

$$\int h_j(x) dF(x) \leq \infty, \quad j = 1, 2.$$

Then

$$v_{\text{JACK}}^{[1]}(T_n)/\sigma_n^2 \rightarrow_p 1,$$

where σ_n^2 is the asymptotic variance of T_n , and

$$v_{\text{JACK}}^{[1]}(T_n)/v_{\text{JACK}}^{[1]}(S_0) - 1 = O_p(n^{-1/2})$$

for any initial point S_0 satisfying $\sqrt{n}(S_0 - \theta) = O_p(1)$.

The proof of this theorem is omitted and can be found in Shao (1992c). It can also be shown that v_{JACK} and $v_{\text{JACK}}^{[1]}(S_0)$ are asymptotically equivalent. Furthermore, Shao (1992c) extends the result in Theorem 5.1 to nonlinear models.

A simulation study of the finite sample performance of the one-step jackknife estimators was conducted by Shao (1992c). Let X_1, \dots, X_n be i.i.d. from $F(x - \theta)$, where F is symmetric about 0 but unknown and θ is an unknown parameter to be estimated. In the simulation,

$$F(x - \theta) = 0.9\Phi(x - \theta) + 0.1\Phi((x - \theta)/\tau)$$

with $\theta = 1$ and $\tau = 4$, where $\Phi(x)$ is the standard normal distribution. The following two M-estimators were considered:

(1) Huber's estimator with

$$r(x, t) = \begin{cases} (x - t)^2/2 & \text{if } |x - t| \leq K \\ K|x - t| - K^2/2 & \text{if } |x - t| > K, \end{cases}$$

where $K = 1.5$ as suggested in Lehmann (1983, p. 376).

(2) The least 1.5th power estimator with

$$r(x, t) = |x - t|^{1.5}.$$

To compute the M-estimators T_n , Newton's method with $T_n^{[0]} = \bar{X}_n$ was used. The true asymptotic variances of these two M-estimators are $1.1845/n$ and $1.0914/n$, respectively. Hence, the M-estimators are much

more efficient than the customary estimators sample mean and sample median, which have asymptotic variances of $2.5/n$ and $1.8382/n$, respectively.

The following four jackknife estimators were examined: the jackknife estimator v_{JACK} and the one-step jackknife estimators $v_{\text{JACK}}^{[1]}(T_n)$, $v_{\text{JACK}}^{[1]}(T_n^{[1]})$, and $v_{\text{JACK}}^{[1]}(\bar{X}_n)$.

Table 5.1. Simulation ME and VAR of M-estimators and jackknife estimators [Adapted from Shao (1992c), by permission of Institute of Statistical Mathematics]

Huber's estimator						
	T_n	$T_n^{[1]}$	v_{JACK}	$v_{\text{JACK}}^{[1]}(T_n)$	$v_{\text{JACK}}^{[1]}(T_n^{[1]})$	$v_{\text{JACK}}^{[1]}(\bar{X}_n)$
ME	1.02745	1.03045	0.09977	0.09977	0.09944	0.10133
VAR	0.08663	0.08703	0.00419	0.00419	0.00408	0.00440
	$n = 12, \sigma_n^2 = 0.09871$					
ME	1.02768	1.03029	0.05874	0.05874	0.05868	0.05949
VAR	0.05135	0.05156	0.00073	0.00073	0.00073	0.00080
	$n = 20, \sigma_n^2 = 0.05922$					
ME	1.03093	1.03348	0.03273	0.03273	0.03268	0.03313
VAR	0.02873	0.02885	0.00012	0.00012	0.00012	0.00013
	$n = 36, \sigma_n^2 = 0.03290$					
The least 1.5th power estimator						
	T_n	$T_n^{[1]}$	v_{JACK}	$v_{\text{JACK}}^{[1]}(T_n)$	$v_{\text{JACK}}^{[1]}(T_n^{[1]})$	$v_{\text{JACK}}^{[1]}(\bar{X}_n)$
ME	1.00137	1.00092	0.09957	0.09957	0.10217	0.11384
VAR	0.09200	0.09361	0.00545	0.00545	0.00522	0.00565
	$n = 12, \sigma_n^2 = 0.09095$					
ME	1.00002	1.00026	0.05682	0.05682	0.05826	0.06315
VAR	0.05461	0.05533	0.00113	0.00113	0.00106	0.00115
	$n = 20, \sigma_n^2 = 0.05457$					
ME	1.00118	1.00127	0.03131	0.03131	0.03185	0.03365
VAR	0.03037	0.03062	0.00020	0.00020	0.00019	0.00020
	$n = 36, \sigma_n^2 = 0.03032$					

Table 5.1 shows the simulation means (ME) and variances (VAR) of these four jackknife estimators under 5000 replications for $n = 12, 20$, and 36 . For comparison, the simulation means and variances and the asymptotic variances σ_n^2 of T_n and the corresponding one-step estimator $T_n^{[1]}$ are also given in Table 5.1.

The following is a summary of the results in Table 5.1.

- (1) The ME and VAR of the one-step jackknife estimator $v_{\text{JACK}}^{[1]}(T_n)$ are almost identical to those of the jackknife estimator v_{JACK} . The ME and VAR of the other two one-step jackknife estimators are not the same but close to those of v_{JACK} .
- (2) The jackknife and one-step jackknife estimators are good estimators of the asymptotic variance σ_n^2 . For small n , the finite sample variance of T_n or $T_n^{[1]}$ may not be close to σ_n^2 , especially for the case of Huber's estimator. In such cases, the jackknife and one-step jackknife estimators are not very good as estimators of the finite sample variances of T_n and $T_n^{[1]}$.
- (3) Overall, the one-step estimator $v_{\text{JACK}}^{[1]}(T_n^{[1]})$ is preferred in terms of its performance and computational simplicity.

5.1.2 Grouping and random subsampling

When n is very large, it may be worthwhile to find an approximation to v_{JACK} that requires fewer computations. An old idea for computational saving is to use the *grouped jackknife*. This method works by dividing n data points into m groups of size g ($n = mg$) and computing $T_{n-g,i}$, the given statistic based on the data with the i th group of points removed. The grouped jackknife variance estimator is

$$v_{\text{GJACK}} = \frac{m-1}{m} \sum_{i=1}^m \left(T_{n-g,i} - \frac{1}{m} \sum_{j=1}^m T_{n-g,j} \right)^2.$$

Note that the computation of v_{GJACK} essentially requires m evaluations of the given statistic. If there is no clear way to divide the data set, the groups have to be formed randomly. When the groups are formed randomly,

$$\begin{aligned} E_G(v_{\text{GJACK}}) &= (1 - m^{-1}) v_{\text{JACK}_g} \\ &\quad - \frac{(m-1)^2}{mN\tilde{N}} \sum_{\mathbf{s}, \tilde{\mathbf{s}} \in \mathcal{S}, \mathbf{s} \cap \tilde{\mathbf{s}} = \emptyset} T_{n-g,\mathbf{s}} (T_{n-g,\tilde{\mathbf{s}}} - \bar{T}_n), \end{aligned}$$

where E_G is the expectation with respect to the random grouping, v_{JACK_g} is the delete-d jackknife variance estimator (2.47) with $d = g$, $N = \binom{n}{g}$,

$\tilde{N} = \binom{n-g}{g}$, \mathcal{S} is the collection of all subsets of $\{1, \dots, n\}$ with size g , and $\bar{T}_n = N^{-1} \sum_{s \in \mathcal{S}} T_{n-g,s}$.

For the same m evaluations of the given statistic, we may use *random subsampling*: draw a simple random sample $\{i_1, \dots, i_m\}$ without replacement from $\{1, \dots, n\}$ and approximate v_{JACK} by

$$v_{\text{SJACK}} = \frac{n-1}{m} \sum_{t=1}^m \left(T_{n-1,i_t} - \frac{1}{m} \sum_{k=1}^m T_{n-1,i_k} \right)^2.$$

Let E_S be the expectation with respect to the selection of $\{i_1, \dots, i_m\}$. Then

$$E_S(v_{\text{SJACK}}) = [1 - m^{-1}(n-1)^{-1}(n-m)]v_{\text{JACK}}.$$

Shao (1989b) studied the consistency of v_{SJACK} with $m < n$ and $m \rightarrow \infty$.

It is of interest to know which of the two approximations, v_{GJACK} and v_{SJACK} , is better. Since many statistics can be approximated by a “sample mean”, we consider the following example.

Example 5.1. Comparison of v_{GJACK} and v_{SJACK} . Suppose that $T_n = n^{-1} \sum_{i=1}^n Z_i$, where Z_i is a function of X_i , $i = 1, \dots, n$, with $EZ_1^4 < \infty$. Without loss of generality, assume that $EZ_1 = 0$ and $n = mg$. Since v_{GJACK} and v_{SJACK} have similar biases, we only need to consider their variances. It can be shown that

$$n^2 \text{var}(v_{\text{GJACK}}) = \frac{\kappa\sigma^4}{n} + \frac{2\sigma^4}{m-1} + O\left(\frac{1}{mn}\right)$$

and

$$n^2 \text{var}(v_{\text{SJACK}}) = \frac{\kappa\sigma^4}{n} + \frac{2\sigma^4}{n-1} + \frac{(n-m)(n-2)(\kappa+2)\sigma^4}{m(n-1)^2} + O\left(\frac{1}{mn}\right),$$

where $\sigma^2 = \text{var}(Z_1)$ and $\kappa = \sigma^{-4}EZ_1^4 - 3$. Ignoring the terms of order $O(\frac{1}{mn})$, we obtain that

$$\frac{\text{var}(v_{\text{GJACK}})}{\text{var}(v_{\text{JACK}})} = \frac{n-1}{m-1} \frac{(m-1)\kappa + 2n}{(n-1)\kappa + 2n}$$

and

$$\frac{\text{var}(v_{\text{SJACK}})}{\text{var}(v_{\text{JACK}})} = 1 + \frac{(n-m)(n-2)n(\kappa+2)}{m(n-1)^2\kappa + 2mn(n-1)}.$$

Consequently, we have the following interesting fact: if $\kappa \leq 0$, then

$$\text{var}(v_{\text{SJACK}}) < \text{var}(v_{\text{GJACK}});$$

if $\kappa > 0$, then

$$\text{var}(v_{\text{SJACK}}) > \text{var}(v_{\text{GJACK}})$$

for sufficiently large n and m .

5.2 The Delete-d Jackknife

The computation of the delete-d jackknife variance estimator $v_{\text{JACK-}d}$ for a given statistic T_n requires $N = \binom{n}{d}$ evaluations of the statistics $T_{r,s}$, where $r = n - d$, s is a subset of $\{1, \dots, n\}$ with size d , and $T_{r,s}$ is the given statistic based on $X_i, i \notin s$. The computational complexity increases rapidly as d increases. To circumvent this, $T_{r,s}$ may be evaluated only for a subcollection of subsets suitably chosen from \mathcal{S} , the collection of all subsets of $\{1, \dots, n\}$ of size d . Two such methods, *balanced subsampling* and *random subsampling*, are introduced in this section.

5.2.1 Balanced subsampling

The idea behind balanced subsampling is to choose subsets from \mathcal{S} in a systematic manner. Let $\mathcal{B} = \{s_1, \dots, s_m\}$ be a collection of m subsets in \mathcal{S} satisfying the following two properties:

- (1) every $i \in \{1, \dots, n\}$ appears in the same number of subsets in \mathcal{B} ;
- (2) every pair (i, j) , $1 \leq i < j \leq n$, appears together in the same number of subsets in \mathcal{B} .

If each subset is treated as a “block” and each i as a “treatment”, \mathcal{B} is a balanced incomplete block design (BIBD) (John, 1971). An approximation to $v_{\text{JACK-}d}$ is then obtained by using $T_{r,s}$ with $s \in \mathcal{B}$:

$$v_{\text{BJACK-}d} = \frac{r}{dm} \sum_{s \in \mathcal{B}} \left(T_{r,s} - \frac{1}{m} \sum_{s \in \mathcal{B}} T_{r,s} \right)^2.$$

In general, $n \leq m \leq N$ (\mathcal{S} is the BIBD with largest size N). A BIBD with a size much smaller than N can often be found.

When $T_n = \bar{Z}_n = n^{-1} \sum_{i=1}^n Z_i$, where Z_i is a function of X_i , the properties of the BIBD imply that

$$v_{\text{BJACK-}d} = v_{\text{JACK-}d} = \frac{1}{n(n-1)} \sum_{i=1}^n (Z_i - \bar{Z}_n)^2. \quad (5.5)$$

This and the fact that the asymptotic behavior of $v_{\text{JACK-}d}$ is determined by the linear approximations to $T_{r,s}$ ensure that $v_{\text{BJACK-}d}$ preserves the asymptotic properties that $v_{\text{JACK-}d}$ has.

Theorem 5.2. *Assume the conditions in Theorem 2.10 (or Theorem 2.11). Then $v_{\text{BJACK-}d}$ is weakly consistent.*

Proof. Consider the decomposition $T_{r,s} - T_n = L_{r,s} + W_{r,s}$, where $L_{r,s} = r^{-1} \sum_{i \in s^c} \phi_F(X_i) - n^{-1} \sum_{i=1}^n \phi_F(X_i)$, $\phi_F(X_i)$ is given in (2.48), $W_{r,s} =$

$R_{r,s} - N^{-1} \sum_{s \in S} R_{r,s}$, and $R_{r,s}$ is given in (2.50). It follows from (5.5) that

$$\frac{r}{dm} \sum_{s \in \mathcal{B}} L_{r,s}^2 = \frac{1}{n(n-1)} \sum_{i=1}^n \left[\phi_F(X_i) - \frac{1}{n} \sum_{j=1}^n \phi_F(X_j) \right]^2.$$

Since $W_{r,s}$, $s \in \mathcal{B}$, are identically distributed,

$$E\left(\frac{r}{dm} \sum_{s \in \mathcal{B}} W_{r,s}\right) = \frac{rw_n}{d},$$

where w_n is given in (2.52). The rest of the proof is the same as that of Theorem 2.10. \square

5.2.2 Random subsampling

The balanced subsampling method requires the enumeration of balanced subsets (the construction of \mathcal{B}), which may not be easy. To reduce the number of computations, we can also apply random subsampling; that is, draw a simple random sample $\{\mathbf{s}_1, \dots, \mathbf{s}_m\}$ with replacement from S and approximate $v_{\text{JACK-d}}$ by

$$v_{\text{SJACK-d}} = \frac{r}{dm} \sum_{t=1}^m \left(T_{r,\mathbf{s}_t} - \frac{1}{m} \sum_{k=1}^m T_{r,\mathbf{s}_k} \right)^2.$$

We may also consider sampling $\{\mathbf{s}_1, \dots, \mathbf{s}_m\}$ without replacement from S . However, since m is usually much smaller than N , sampling without replacement produces almost the same result as sampling with replacement. Therefore, we focus on the simpler plan: sampling with replacement.

If $d = n/2$ ($r = d$), $v_{\text{SJACK-d}}$ looks very similar to the Monte Carlo approximation to the bootstrap variance estimator $v_{\text{BOOT}}^{(B)}$ given in (1.18). The only difference is that each $T_{n,b}^*$ is based on a simple random sample with replacement from $\{X_1, \dots, X_n\}$, whereas each $T_{r,s}$ is a function of $\{X_i, i \in \mathbf{s}^c\}$, which can be viewed as a simple random sample without replacement from $\{X_1, \dots, X_n\}$.

From sampling theory,

$$E_S(v_{\text{SJACK-d}}) = (1 - m^{-1})v_{\text{JACK-d}},$$

where E_S is the expectation with respect to the selection of $\{\mathbf{s}_1, \dots, \mathbf{s}_m\}$. The following result shows the asymptotic validity of the approximation $v_{\text{SJACK-d}}$. Let

$$\rho_n = \text{var}(T_n^2). \quad (5.6)$$

Theorem 5.3. Assume that T_n has a finite fourth moment. Then

$$v_{\text{SJACK-d}} = v_{\text{JACK-d}} + O_p\left(\frac{r\sqrt{\rho_r}}{d\sqrt{m}}\right).$$

Proof. Let $\tilde{T}_n = N^{-1} \sum_{\mathbf{s} \in S} T_{r,\mathbf{s}}$. From sampling theory, we obtain that

$$\frac{r}{d} E_S \left(\frac{1}{m} \sum_{t=1}^m T_{r,\mathbf{s}_t} - \tilde{T}_n \right)^2 = \frac{1}{m} v_{\text{JACK-d}} = O_p\left(\frac{r\sqrt{\rho_r}}{dm}\right),$$

since

$$E(v_{\text{JACK-d}}) = \frac{r}{d} E(T_{r,\mathbf{s}} - \tilde{T}_n)^2 \leq O\left(\frac{r\sqrt{\rho_r}}{dm}\right).$$

Similarly,

$$\begin{aligned} E_S \left| \frac{r}{dm} \sum_{t=1}^m (T_{r,\mathbf{s}_t} - \tilde{T}_n)^2 - v_{\text{JACK-d}} \right| &\leq \left[\frac{r^2}{d^2 m N} \sum_{\mathbf{s} \in S} (T_{r,\mathbf{s}} - \tilde{T}_n)^4 \right]^{1/2} \\ &= O_p\left(\frac{r\sqrt{\rho_r}}{d\sqrt{m}}\right). \end{aligned}$$

Then the result follows from the fact that $E_S |v_{\text{SJACK-d}} - v_{\text{JACK-d}}|$ is bounded by

$$E_S \left| \frac{r}{dm} \sum_{t=1}^m (T_{r,\mathbf{s}_t} - \tilde{T}_n)^2 - v_{\text{JACK-d}} \right| + \frac{r}{d} E_S \left(\frac{1}{m} \sum_{t=1}^m T_{r,\mathbf{s}_t} - \tilde{T}_n \right)^2. \quad \square$$

Let σ_n^2 be the asymptotic variance of T_n . Then usually $\sigma_n^2 = O(n^{-1})$. If $v_{\text{JACK-d}}$ is consistent, i.e., $v_{\text{JACK-d}}/\sigma_n^2 \rightarrow_p 1$, then it follows from Theorem 5.3 that

$$v_{\text{SJACK-d}}/\sigma_n^2 \rightarrow_p 1,$$

provided that

$$nr\sqrt{\rho_r}/d\sqrt{m} \rightarrow 0. \quad (5.7)$$

That is, $v_{\text{SJACK-d}}$ is consistent if we choose m according to (5.7). Typically, the ρ_n in (5.6) is of the order n^{-2} . Then, when n/d is bounded, $v_{\text{SJACK-d}}$ is consistent as long as $m \rightarrow \infty$.

Although one desires a small m to save computations, m should not be too small, otherwise efficiency will be lost. When the order of $v_{\text{JACK-d}} - \sigma_n^2$ is $n^{-3/2}$ (e.g., when $T_n = \bar{X}_n$), the loss in efficiency caused by random subsampling is asymptotically negligible if

$$nr\sqrt{n\rho_r}/d\sqrt{m} = o(1),$$

which reduces to $n/m \rightarrow 0$ when $\rho_n = O(n^{-2})$ and n/d is bounded. Hence, to retain the efficiency, a suitable m should satisfy $n/m \rightarrow 0$. For example, we may take $m = n^\delta$ with $\delta > 1$ to get an efficient approximation to $v_{\text{SJACK-d}}$. The number of computations is considerably reduced (n^δ is much smaller than N). In fact, when δ is slightly larger than 1, the number of computations required by $v_{\text{SJACK-d}}$ is almost the same as that required by the delete-1 jackknife variance estimator.

Example 5.2. Efficiency of random subsampling. To examine the efficiency of random subsampling, we consider the simulation results in Shao (1987). The statistic T_n under consideration was actually based on non-i.i.d. data: the least squares estimator of the intersection of a polynomial regression model. The sample size $n = 15$ and $v_{\text{SJACK-d}}$ with $d = 3$ ($N = 455$) were considered. Table 5.2 lists the simulation estimates (3000 replications) of the relative increase in root mean squared error caused by $v_{\text{SJACK-d}}$, which is defined as

$$\frac{\sqrt{\text{mse}(v_{\text{SJACK-d}})} - \sqrt{\text{mse}(v_{\text{JACK-d}})}}{\sqrt{\text{mse}(v_{\text{JACK-d}})}}.$$

It can be seen that the efficiency of $v_{\text{SJACK-d}}$ increases very quickly when m increases from 15 to 65 (nearly equals $n^{3/2}$); after $m = 65$, the efficiency increase slows down and is not appreciable in comparison with the large computation reduction by using $m = 65$.

Table 5.2. Relative increase in root mean squared error (RIRM)

m	15	25	35	50	65	80	100
RIRM	23.8%	17.5%	10.4%	8.6%	5.8%	5.7%	2.9%
m/N	0.033	0.055	0.077	0.110	0.143	0.176	0.220

5.3 Analytic Approaches for the Bootstrap

We now turn to the computation of bootstrap estimators. Like the jackknife estimators, the bootstrap estimators have explicit formulas only in a few cases (Examples 1.1 and 1.4). Huang (1991) derived explicit formulas of bootstrap variance estimators for L-statistics (5.1), but the computations using these formulas still involve a large number of calculations of combinations.

In this section, we study analytic solutions or approximations to the bootstrap estimators. Simulation (Monte Carlo) techniques are studied in Section 5.4.

5.3.1 The delta method

Efron (1979) applied the delta method (Taylor's expansion) to approximate the bootstrap bias and variance estimators. Let $\{X_1^*, \dots, X_n^*\}$ be a bootstrap sample from F_n , the empirical distribution based on X_1, \dots, X_n . Define

$$P_i^* = (\text{the number of } X_j^* = X_i, j = 1, \dots, n)/n$$

and

$$\mathbf{P}^* = (P_1^*, \dots, P_n^*)'.$$

Given X_1, \dots, X_n , $n\mathbf{P}^*$ is distributed as the multinomial with parameters n and $\mathbf{P}^0 = \mathbf{1}/n$, where $\mathbf{1} = (1, \dots, 1)'$. Then

$$E_* \mathbf{P}^* = \mathbf{P}^0 \quad \text{and} \quad \text{var}_*(\mathbf{P}^*) = n^{-2} \mathbf{I} - n^{-3} \mathbf{1}\mathbf{1}',$$

where \mathbf{I} is the identity matrix, and E_* and var_* are the bootstrap expectation and variance (conditional on X_1, \dots, X_n), respectively.

Recall that the bootstrap estimator of a moment of a random variable $\mathfrak{R}_n(X_1, \dots, X_n, F)$ is the same moment of the bootstrap analog $\mathfrak{R}_n(X_1^*, \dots, X_n^*, F_n)$. In many cases, $\mathfrak{R}_n(X_1^*, \dots, X_n^*, F_n) = \mathfrak{R}_n(\mathbf{P}^*)$, a functional of \mathbf{P}^* . For example, if $\mathfrak{R}_n(X_1, \dots, X_n, F) = \bar{X}_n - \mu$, where \bar{X}_n and μ are the sample mean and population mean, respectively, then

$$\mathfrak{R}_n(X_1^*, \dots, X_n^*, F_n) = \bar{X}_n^* - \bar{X}_n = (\mathbf{X} - \bar{X}_n \mathbf{1})' \mathbf{P}^*,$$

where $\mathbf{X} = (X_1, \dots, X_n)'$.

Since $E_* \mathbf{P}^* = \mathbf{P}^0$, we can expand $\mathfrak{R}_n(\mathbf{P}^*)$ at \mathbf{P}^0 . Omitting the terms involving derivatives of order three or higher in a multivariate Taylor expansion, we obtain that

$$\mathfrak{R}_n(\mathbf{P}^*) \approx \mathfrak{R}_n(\mathbf{P}^0) + \mathbf{U}'(\mathbf{P}^* - \mathbf{P}^0) + \frac{1}{2}(\mathbf{P}^* - \mathbf{P}^0)' \mathbf{V}(\mathbf{P}^* - \mathbf{P}^0), \quad (5.8)$$

where $\mathbf{U} = \nabla \mathfrak{R}_n(\mathbf{P}^0)$ and $\mathbf{V} = \nabla^2 \mathfrak{R}_n(\mathbf{P}^0)$. From (5.8) we immediately get the following approximations to the bootstrap bias and variance estimators:

$$b_{\text{BOOT}} = E_* \mathfrak{R}_n(\mathbf{P}^*) \approx \mathfrak{R}_n(\mathbf{P}^0) + \frac{1}{2n^2} \text{tr}(\mathbf{V}) \quad (5.9)$$

and

$$v_{\text{BOOT}} = \text{var}_* \mathfrak{R}_n(\mathbf{P}^*) \approx \frac{1}{n^2} \mathbf{U}' \mathbf{U} = \frac{1}{n^2} \sum_{i=1}^n U_i^2, \quad (5.10)$$

where U_i is the i th component of \mathbf{U} .

Example 5.3. Suppose that $X_i = (Y_i, Z_i)'$, $i = 1, \dots, n$, are i.i.d. from an unknown bivariate distribution F . The parameter $\theta = EY_1/EZ_1$ is

estimated by $\hat{\theta}_n = \bar{Y}_n/\bar{Z}_n$, where \bar{Y}_n and \bar{Z}_n are the sample means of Y_1, \dots, Y_n and Z_1, \dots, Z_n , respectively. Define

$$\mathfrak{R}_n(X_1, \dots, X_n, F) = \hat{\theta}_n/\theta.$$

Then

$$\mathfrak{R}_n(\mathbf{P}^*) = \left(\frac{\sum_{i=1}^n P_i^* Y_i}{\sum_{i=1}^n P_i^* Z_i} \right) / \hat{\theta}_n.$$

Let U_i be as given in (5.10) and V_{ij} be the (i, j) th element of \mathbf{V} . Then

$$U_i = \frac{Y_i}{\bar{Y}_n} - \frac{Z_i}{\bar{Z}_n} \quad \text{and} \quad V_{ij} = \frac{2Z_i Z_j}{\bar{Z}_n^2} - \left(\frac{Y_i Z_j}{\bar{Y}_n \bar{Z}_n} - \frac{Y_j Z_i}{\bar{Y}_n \bar{Z}_n} \right).$$

Therefore, by (5.9) and (5.10),

$$b_{\text{BOOT}} \approx 1 - \frac{1}{n^2} \sum_{i=1}^n \left[\left(\frac{Y_i}{\bar{Y}_n} - 1 \right) \left(\frac{Z_i}{\bar{Z}_n} - 1 \right) - \left(\frac{Z_i}{\bar{Z}_n} - 1 \right)^2 \right]$$

and

$$v_{\text{BOOT}} \approx \frac{1}{n^2} \sum_{i=1}^n \left(\frac{Y_i}{\bar{Y}_n} - \frac{Z_i}{\bar{Z}_n} \right)^2.$$

5.3.2 Jackknife approximations

A closer look at (5.9) and (5.10) reveals a relationship between the delete-1 jackknife and the bootstrap. The right-hand sides of (5.9) and (5.10) are exactly the same as the infinitesimal jackknife bias and variance estimators proposed by Jaeckel (1972). If we replace the derivative U_i in (5.10) by the finite difference approximation

$$\tilde{U}_i = \frac{\mathfrak{R}_n(\mathbf{P}^0 - \frac{1}{n-1}(\mathbf{e}_i - \mathbf{1}/n)) - \mathfrak{R}_n(\mathbf{P}^0)}{-\frac{1}{n-1}},$$

where \mathbf{e}_i is the i th coordinate vector in \mathbb{R}^n , then

$$\tilde{U}_i = (n-1)(T_n - T_{n-1,i}),$$

when $\mathfrak{R}_n(X_1, \dots, X_n, F) = T_n - \theta$ for a given statistic T_n . Note that $n^{-1} \sum_{i=1}^n \tilde{U}_i^2$ is almost the same as the delete-1 jackknife variance estimate v_{JACK} in (1.13) except that $n^{-1} \sum_{i=1}^n T_{n-1,i}$ is replaced by T_n . This indicates that the delete-1 jackknife estimators can be viewed as approximations to the bootstrap estimators. Therefore, for the bias and variance estimation, the delete-1 jackknife is simple (relative to the bootstrap) and is also good when the sample size is reasonably large and $\mathfrak{R}_n(\mathbf{P}^*)$ is sufficiently regular in the sense that the approximation (5.8) is valid. However,

for nonregular statistics such as the sample quantiles, this relationship between the delete-1 jackknife and the bootstrap may not exist, since we know that the delete-1 jackknife variance estimators for the sample quantiles are inconsistent.

In the following, we show that the bootstrap distribution estimators can also be approximated by using the jackknife together with the Edgeworth expansions. For a given statistic T_n , let H_{BOOT} be the bootstrap estimator given in (3.2) with $\Re_n = \sqrt{n}(T_n - \theta)$, where θ is a parameter related to F . Let $b_n(F) = ET_n - \theta$, $\sigma_n^2(F)$ be the asymptotic variance of T_n , and $\kappa_n(F) = \sigma_n^{-3}(F)E(T_n - ET_n)^3$ be the skewness of T_n . Under some conditions, the bootstrap estimator has the following Edgeworth expansion:

$$H_{\text{BOOT}}(x) \approx \Phi\left(\frac{x - \sqrt{n}b_n(F_n)}{\sqrt{n}\sigma_n(F_n)}\right) - \kappa_n(F_n)\psi\left(\frac{x - \sqrt{n}b_n(F_n)}{\sqrt{n}\sigma_n(F_n)}\right), \quad (5.11)$$

where $\psi(x) = \frac{1}{6}(x^2 - 1)\varphi(x)$, and Φ and φ are the standard normal distribution and density, respectively. Since the computations of $b_n(F_n)$, $\sigma_n^2(F_n)$, and $\kappa_n(F_n)$ in (5.11) may involve some difficult derivations, we can replace $b_n(F_n)$, $\sigma_n^2(F_n)$, and $\kappa_n(F_n)$ by the jackknife estimators b_{JACK} , v_{JACK} , and sk_{JACK} defined in (1.8), (1.13), and (3.60), respectively. This results in a jackknife and Edgeworth approximation to the bootstrap estimator H_{BOOT} . One can further apply this method to obtain jackknife approximations to the hybrid bootstrap and bootstrap-t confidence sets.

Beran (1984a) provided some variations of b_{JACK} , v_{JACK} , and sk_{JACK} and derived a different jackknife approximation to H_{BOOT} . Since b_{JACK} , v_{JACK} , and sk_{JACK} are consistent estimators, the jackknife approximation error is of the order $o_p(n^{-1/2})$. Beran (1984a) and Zhang and Tu (1990) showed some simulation results for the performance of the jackknife approximations.

The estimation of the skewness of a statistic is a very delicate problem. Although the jackknife estimator sk_{JACK} is consistent, simulation results (Beran, 1984a; Tu and Zhang, 1992b) indicate that it is very biased for a moderate n , which affects the accuracy of the jackknife approximation to H_{BOOT} . Tu and Gross (1994) modified the jackknife skewness estimator to reduce the bias, but the impact of this modification on the accuracy of the jackknife approximation to H_{BOOT} has not yet been studied.

5.3.3 Saddle point approximations

The *saddle point* method is an alternative to the Edgeworth expansion for approximating distribution functions and is often more accurate than the Edgeworth expansion (Reid, 1988). We discuss here the saddle point approximation to the bootstrap estimator H_{BOOT} given in (3.2) with $\Re_n = T_n - \theta$, for some statistic T_n and parameter θ .

We first assume that $T_n = \bar{X}_n$, the sample mean of X_1, \dots, X_n , and $\theta = EX_1$. From the general theory of saddle point methods, the distribution of $T_n - \theta$, $H_n(x)$, can be approximated by (Reid, 1988)

$$H_n(x) \approx \Phi(\omega_x) + \varphi(\omega_x)(\omega_x^{-1} - z_x^{-1}), \quad (5.12)$$

where $\omega_x = \sqrt{2n[\lambda_x x - h(\lambda_x)]}\text{sign}(\lambda_x)$, $h(t) = \log E(e^{tX_1})$ is assumed to be well defined for some $t \neq 0$, $z_x = \lambda_x \sqrt{nh''(\lambda_x)}$, and λ_x is the solution of $h'(\lambda_x) = x$.

Replacing F by the empirical distribution F_n in the right-hand side of (5.12), we immediately get an approximation to $H_{\text{BOOT}}(x)$:

$$H_{\text{BOOT}}(x) \approx H_{\text{SADD}}(x) = \Phi(\tilde{\omega}_x) + \varphi(\tilde{\omega}_x)(\tilde{\omega}_x^{-1} - \tilde{z}_x^{-1}), \quad (5.13)$$

where $\tilde{\omega}_x = \sqrt{2n[\xi_x x - h_n(\xi_x)]}\text{sign}(\xi_x)$, $h_n(t) = \log(n^{-1} \sum_{i=1}^n e^{tX_i})$, $\tilde{z}_x = \xi_x \sqrt{nh''_n(\xi_x)}$, and ξ_x is the unique solution of $h'_n(\xi_x) = x$ (Davison and Hinkley, 1988). The approximation H_{SADD} in (5.13) is just the empirical saddle point approximation discussed in Feuerverger (1989).

One of the conditions required for the saddle point approximation is the continuity of the distribution to be approximated. Since H_{BOOT} is a discrete distribution, we have to slightly modify the approximation in (5.13). Suppose that X_i are given to m decimal places so that \bar{X}_n^* is a multiple of $10^{-m}/n$. When x is a multiple of $10^{-m}/n$, Hinkley and Wang (1988) provided the following modified approximation to $H_{\text{BOOT}}(x)$:

$$H_{\text{SBOOT}}(x) = \begin{cases} \Phi(\hat{\omega}_x) + \varphi(\hat{\omega}_x)(\hat{\omega}_x^{-1} - \hat{z}_x^{-1}) & \text{if } d \neq 0 \\ \frac{1}{2} + \frac{1}{\sqrt{8\pi nh''_n(0)}} \left[\frac{h'''_n(0)}{3h''_n(0)} - \frac{1}{10^m} \right] & \text{if } d = 0, \end{cases}$$

where $d = x + 10^{-m}/n$, $\hat{\omega}_x = \sqrt{2n[\zeta_x d - h_n(d) - \bar{X}_n]}\text{sign}(\zeta_x)$, $\hat{z}_x = 10^m(1 - e^{10^{-m}\zeta_x})\sqrt{nh''_n(\zeta_x)}$, and ζ_x is the solution of $h'_n(\zeta_x) = x$. They proved that

$$H_{\text{BOOT}}(x) = H_{\text{SBOOT}}(x)[1 + O_p(n^{-1})],$$

although the relative error is not strictly uniform in the tails for fixed n .

A numerical example is given in Hinkley and Wang (1988) to assess the accuracy of saddle point approximation H_{SBOOT} with $m = 1$ for a sample of $n = 10$:

$$9.6, 10.4, 13.0, 15.0, 16.6, 17.2, 17.3, 21.8, 24.0, 23.8. \quad (5.14)$$

The true bootstrap distribution was approximated by using Monte Carlo with $B = 50,000$ replicates. Table 5.3 gives the approximations to the p -quantiles of H_{BOOT} for different p 's, using H_{SADD} and H_{SBOOT} .

From Table 5.3, the saddle point approximation to the bootstrap distribution is very accurate, even when n is as small as 10. Also, H_{SADD} and H_{SBOOT} almost provide the same results.

Table 5.3. Values of $z_B = H_{\text{BOOT}}^{-1}(p)$, $z_S = H_{\text{SADD}}^{-1}(p)$, and $z_{SB} = H_{\text{SBOOT}}^{-1}(p)$ [Adapted from Hinkley and Wang (1988), by permission of Institute of Mathematical Statistics]

p	0.0001	0.0005	0.001	0.05	0.01	0.05	0.10	0.20
z_B	-6.34	-5.79	-5.55	-4.81	-4.42	-3.34	-2.69	-1.86
z_S	-6.31	-5.78	-5.52	-4.81	-4.43	-3.33	-2.69	-1.86
z_{SB}	-6.32	-5.79	-5.53	-4.81	-4.44	-3.33	-2.69	-1.86
p	0.80	0.90	0.95	0.99	0.995	0.999	0.9995	0.9999
z_B	1.80	2.87	3.73	5.47	6.12	7.52	8.19	9.33
z_S	1.80	2.85	3.75	5.48	6.12	7.46	7.99	9.12
z_{SB}	1.79	2.85	3.74	5.47	6.12	7.46	7.98	9.11

However, it is not clear how to generalize the saddle point method to complex statistics. For T_n , a solution of the equation

$$\frac{1}{n} \sum_{i=1}^n \psi(X_i, T_n) = 0,$$

Davison and Hinkley (1988) gave a saddle point formula to approximate the bootstrap estimator H_{BOOT} . For a general statistic T_n , we may first approximate it by a quadratic statistic, i.e.,

$$T_n \approx \theta + \frac{1}{n} \sum_{i=1}^n L(X_i, F_n) + \frac{1}{2n^2} \sum_{i=1}^n \sum_{j=1}^n Q(X_i, X_j, F_n),$$

and then apply the saddle point approximation to the quadratic statistic. But even for a quadratic statistic that looks very simple, its cumulant generating function is often nonconvex and thus the saddle point is difficult to find. Daniels and Young (1991) and Wang (1992) discussed the saddle point approximations to the bootstrap distributions for studentized means and some general statistics, respectively. Roche (1993) considered saddle point approximations to parametric bootstrap distributions in an exponential family.

5.3.4 Remarks

- (1) We have introduced some analytic methods to approximate the bootstrap moment and distribution estimators. These methods often lead to some other well-known estimators. For example, the delta methods yield the infinitesimal jackknife bias and variance estimators, and the saddle point approximations produce the empirical saddle point estima-

tors. This phenomenon reveals the connection between the bootstrap and other well-known methods, and offers a better understanding of the bootstrap.

- (2) The methods introduced in this section are often termed as bootstrap methods without resampling since the Monte Carlo sampling is replaced by some analytic work. The computations with these methods are substantially less intensive than the methods based on the Monte Carlo method. As was pointed out by Efron (1987), “the form of any useful improvement over” normal approximation “will be shaped by similar compromises, based on how fast, how cheap and how generally available electronic computation becomes for the majority of applied statisticians.” Hence, attention should be paid to the reduction of bootstrap computations. This supports the application of some less computationally intensive methods such as the jackknife.
- (3) Some analytic methods for approximating the bootstrap BC_a confidence sets were suggested in DiCiccio and Tibshirani (1987), Konishi (1991), and DiCiccio and Efron (1992).
- (4) The methods suggested above, like all other analytic methods, have the common disadvantage that they may only be applicable to some special classes of statistics. Furthermore, some of them may require a difficult theoretical derivation, and others may not provide an accurate approximation.
- (5) The accuracy of various approximations to the bootstrap estimators is important. However, since bootstrap estimators have their own errors in estimating the sampling distribution of a given statistic, it may not be worthwhile to reduce the error of the approximation to the bootstrap estimator much below the error of the bootstrap estimator. In the case of $T_n = \bar{X}_n$, for example, although the saddle point approximation to the bootstrap distribution has a relative error of order $O_p(n^{-1})$, it can only approximate the distribution of $T_n - \theta$ with the accuracy of order $O_p(n^{-1/2})$ (Feuerverger, 1989). Note that the bootstrap only has an accuracy of order $O_p(n^{-1/2})$ for estimating the distribution of an unstudentized statistic.

5.4 Simulation Approaches for the Bootstrap

Since analytic methods may not be easy to apply (especially in complex problems), in many applications the Monte Carlo approximation is adopted to compute bootstrap estimators. The Monte Carlo method is simple and easy to apply, but may be inefficient. More efficient simulation-type methods for bootstrap computations are introduced in this section. They can be applied when computational time or cost is a concern, although none of them always works in all practical problems.

5.4.1 The simple Monte Carlo method

We start our discussion with the simple Monte Carlo method described in Chapter 1. Let \hat{P}_n be an estimated model using data X_1, \dots, X_n . The bootstrap bias, variance, and distribution estimators are, respectively,

$$b_{\text{BOOT}} = E_*[\mathfrak{R}_n(X_1^*, \dots, X_m^*, \hat{P}_n)], \quad (5.15)$$

$$v_{\text{BOOT}} = \text{var}_*[\mathfrak{R}_n(X_1^*, \dots, X_m^*, \hat{P}_n)], \quad (5.16)$$

and

$$H_{\text{BOOT}}(x) = P_*\{\mathfrak{R}_n(X_1^*, \dots, X_m^*, \hat{P}_n) \leq x\}, \quad (5.17)$$

where $\{X_1^*, \dots, X_m^*\}$ is a sample from \hat{P}_n , $\mathfrak{R}_n(\cdot, \cdot)$ is an appropriately defined functional, and E_* , var_* , and P_* are the conditional expectation, variance, and probability, respectively, for given X_1, \dots, X_n . In general, m is not necessarily the same as n . We have shown the advantages of using $m \neq n$ in some cases (Chapter 3). But we shall assume $m = n$ in the following discussion for simplicity.

To apply the simple Monte Carlo method for computing the bootstrap estimators in (5.15)–(5.17), we begin with a generation of B independent samples $\{X_1^{*b}, \dots, X_n^{*b}\}$, $b = 1, \dots, B$, from the estimated model \hat{P}_n . Then we calculate $\mathfrak{R}_n^{*b} = \mathfrak{R}_n(X_1^{*b}, \dots, X_n^{*b}, \hat{P}_n)$ for $b = 1, \dots, B$, and approximate b_{BOOT} , v_{BOOT} , and H_{BOOT} by, respectively,

$$b_{\text{BOOT}}^{(B)} = \frac{1}{B} \sum_{b=1}^B \mathfrak{R}_n^{*b}, \quad (5.18)$$

$$v_{\text{BOOT}}^{(B)} = \frac{1}{B} \sum_{b=1}^B \left(\mathfrak{R}_n^{*b} - \frac{1}{B} \sum_{b=1}^B \mathfrak{R}_n^{*b} \right)^2, \quad (5.19)$$

and

$$H_{\text{BOOT}}^{(B)}(x) = \frac{1}{B} \sum_{b=1}^B I\{\mathfrak{R}_n^{*b} \leq x\}. \quad (5.20)$$

In the case where X_1, \dots, X_n are i.i.d. from a distribution F , $\hat{P}_n = \hat{F}$, an estimator of F . If \hat{F} is the empirical distribution, then $\{X_1^{*b}, \dots, X_n^{*b}\}$ is a simple random sample from $\{X_1, \dots, X_n\}$. The simple Monte Carlo approximation is easy to program and implement. We only need to repeatedly take samples from \hat{P}_n or \hat{F} and calculate the random variable \mathfrak{R}_n .

One has to decide how large the Monte Carlo sample size B should be. This problem, which is often of concern to applied statisticians, is a little bit like the problem of determining the size of the original sample X_1, \dots, X_n . But there are two important differences between them. First, in most cases

we can afford a B much larger than n . Second, we waste our time and resources if we go too far, since more computations do not help to reduce the error of the original bootstrap estimator. Thus, we should select a B so that the error of the Monte Carlo approximation is negligible with respect to the error of the original bootstrap estimator.

In the rest of this chapter, we assume that, in the simple Monte Carlo approximation, the bootstrap data are generated from the empirical distribution, which is a popular choice when X_1, \dots, X_n are i.i.d.

Efron (1987) considered the coefficient of variation (CV) as a measure of the variability of the Monte Carlo approximation using a fixed B . Since the estimation of standard deviation and the construction of confidence intervals are the two major applications of the bootstrap, Efron (1987) only studied the determination of B for these two problems.

The Monte Carlo approximation to the bootstrap standard deviation estimator for an estimator $\hat{\theta}_n$ is

$$s_{\text{BOOT}}^{(B)} = \sqrt{v_{\text{BOOT}}^{(B)}},$$

where $v_{\text{BOOT}}^{(B)}$ is given in (5.19) with $\mathfrak{R}_n = \hat{\theta}_n$. Standard formulas for moment estimators (Serfling, 1980) lead to

$$E_* s_{\text{BOOT}}^{(B)} \approx \sqrt{v_{\text{BOOT}}}$$

and

$$\text{var}_*(s_{\text{BOOT}}^{(B)}) \approx \frac{\rho_{\text{BOOT}} - v_{\text{BOOT}}^2}{4Bv_{\text{BOOT}}},$$

where v_{BOOT} is given in (5.16) and

$$\rho_{\text{BOOT}} = E_*(\hat{\theta}_n^* - E_* \hat{\theta}_n^*)^4.$$

Therefore, the coefficient of variation of $s_{\text{BOOT}}^{(B)}$, conditional on X_1, \dots, X_n , is equal to

$$\text{cv}_*(s_{\text{BOOT}}^{(B)}) = \frac{\sqrt{\text{var}_*(s_{\text{BOOT}}^{(B)})}}{E_* s_{\text{BOOT}}^{(B)}} \approx \sqrt{\frac{\hat{\delta}_n + 2}{4B}}, \quad (5.21)$$

where $\hat{\delta}_n = \rho_{\text{BOOT}} / v_{\text{BOOT}}^2 - 3$ is the conditional kurtosis of $\hat{\theta}_n^*$.

There are at least two ways to determine the Monte Carlo sample size B by making use of (5.21).

The first method, suggested by Efron (1987), is to determine B by setting

$$\text{cv}_*(s_{\text{BOOT}}^{(B)}) = \epsilon_0, \quad (5.22)$$

where ϵ_0 is a given desired level. In many cases, $\hat{\delta}_n \approx 0$ ($\hat{\delta}_n \rightarrow_p 0$ as $n \rightarrow \infty$). Then (5.22) reduces to $B = \frac{1}{2}\epsilon_0^{-2}$. For example, if $\epsilon_0 = 0.05$, then $B = 200$; if $\epsilon_0 = 0.1$, then $B = 50$.

Let $s_{\text{BOOT}} = \sqrt{v_{\text{BOOT}}}$ be the bootstrap standard deviation estimator and $\text{cv}(s_{\text{BOOT}})$ be its coefficient of variation. As we mentioned before, it is not worthwhile to try to make $\text{cv}_*(s_{\text{BOOT}}^{(B)})$ too small with respect to $\text{cv}(s_{\text{BOOT}})$. In most cases, it is sufficient to have $\text{cv}_*(s_{\text{BOOT}}^{(B)})/\text{cv}(s_{\text{BOOT}}) \rightarrow 0$. In some cases, even $\text{cv}_*(s_{\text{BOOT}}^{(B)}) = \text{cv}(s_{\text{BOOT}})$ is enough. The second method of determining B is to set

$$\text{cv}_*(s_{\text{BOOT}}^{(B)}) = \text{cv}(s_{\text{BOOT}}) \text{ or } o(\text{cv}(s_{\text{BOOT}})),$$

which leads to

$$B = \frac{a_n(\hat{\delta}_n + 2)}{4[\text{cv}(s_{\text{BOOT}})]^2}, \quad (5.23)$$

where $a_n \equiv 1$ or $\{a_n\}$ is any sequence of positive numbers diverging to infinity (e.g., $a_n = \log \log n$). The Monte Carlo sample size selected according to (5.23) depends on n and the accuracy of the original bootstrap estimator. However, some bounds for $\text{cv}(s_{\text{BOOT}})$ and $\hat{\delta}_n$ are needed for using (5.23).

A similar idea can be used for determining the Monte Carlo sample size in approximating the bootstrap confidence sets. Let $\theta_{\text{BOOT}}^{(B)}$ be the Monte Carlo approximation to the bootstrap BC_a lower (or upper) confidence bound with level $1-\alpha$ [see (4.16) or (4.17)] calculated by replacing $K_{\text{BOOT}}(x)$ [defined in (4.5)] by its Monte Carlo approximation

$$K_{\text{BOOT}}^{(B)}(x) = \frac{1}{B} \sum_{b=1}^B I\{\hat{\theta}_n^{*b} \leq x\}.$$

If we assume that the function Ψ in (4.16) is Φ and that z_0 and a are known (e.g., $z_0 = a = 0$ and the BC_a reduces to the percentile), then applying the formula for the CV of a sample quantile given in Kendall and Staurt (1979, Chapter 10), we obtain that

$$\text{cv}_*(\theta_{\text{BOOT}}^{(B)} - \hat{\theta}_n) \approx \frac{1}{|z_\alpha|\varphi(z_\alpha)} \sqrt{\frac{\alpha(1-\alpha)}{B}}, \quad (5.24)$$

where $z_\alpha = \Phi^{-1}(\alpha)$. We can determine the Monte Carlo sample size B by using (5.22) with $\text{cv}_*(s_{\text{BOOT}}^{(B)})$ replaced by $\text{cv}_*(\theta_{\text{BOOT}}^{(B)} - \hat{\theta}_n)$. If $\alpha = 0.025$ and $\epsilon_0 = 0.05$, then, using (5.24), we get $B = 740$.

However, if z_0 is approximated by Monte Carlo using (4.9), then the right-hand side of (5.24) should be replaced by

$$\frac{1}{|z_\alpha|} \sqrt{\frac{1}{B} \left[\frac{1}{[\varphi(0)]^2} - \frac{2(1-\alpha)}{\varphi(0)\varphi(z_\alpha)} + \frac{\alpha(1-\alpha)}{[\varphi(z_\alpha)]^2} \right]}$$

and $B = 1170$ when $\alpha = 0.025$ and $\epsilon_0 = 0.05$. One may have to use an even larger B when a Monte Carlo approximation for a is also needed.

Babu and Singh (1983) proved that if $\mathfrak{R}_n = (\bar{X}_n - EX_1)/\hat{\sigma}_n$ is the studentized sample mean and B is a function of n satisfying $B/(n \log n) \rightarrow \infty$, then, as $n \rightarrow \infty$,

$$\sqrt{n} \sup_x |H_{\text{BOOT}}^{(B)}(x) - H_{\text{BOOT}}(x)| \rightarrow_{a.s.} 0.$$

That is, the Monte Carlo approximation $H_{\text{BOOT}}^{(B)}$ is still second order accurate as an estimator of the sampling distribution of \mathfrak{R}_n .

Shi, Wu and Chen (1990) presented a method for the determination of B based on the convergence rates of $H_{\text{BOOT}}^{(B)}$ and $(H_{\text{BOOT}}^{(B)})^{-1}$. Let H_n be the distribution of a random variable \mathfrak{R}_n , H_{BOOT} be the bootstrap estimator of H_n defined in (5.17), and $H_{\text{BOOT}}^{(B)}$ be the Monte Carlo approximation to H_{BOOT} defined in (5.20). Using the Bahadur representations of quantile processes, Shi, Wu and Chen (1990) established that

$$\sup_x |H_{\text{BOOT}}^{(B)}(x) - H_{\text{BOOT}}(x)| = \varepsilon_n + \sqrt{B^{-1} \log \log B}$$

and

$$\sup_{0 < t < 1} |(H_{\text{BOOT}}^{(B)})^{-1}(t) - H_{\text{BOOT}}^{-1}(t)| = O(\varepsilon_n + \sqrt{B^{-1} \log \log B}),$$

where $\varepsilon_n = \sup_x |H_{\text{BOOT}}(x) - H_n(x)|$. The Monte Carlo approximation error is negligible with respect to the error ε_n if we select B by solving

$$B^{-1} \log \log B = o(\varepsilon_n^2). \quad (5.25)$$

However, if n itself is very large, then we may select B by setting

$$B^{-1} \log \log B = C\varepsilon_n^2, \quad (5.26)$$

where C is a constant. If $\varepsilon_n = O_p(n^{-1/2})$ or $o_p(n^{-1/2})$, then we may take $B = na_n$ with a_n satisfying $\log \log n/a_n \rightarrow 0$, which is almost in agreement with Babu and Singh's result. If $\varepsilon_n = O_p(n^{-1})$, then we can take $B = n^2 \log \log n$. If n is less than 30, B can be roughly 1000. One disadvantage of this asymptotic approach is that (5.25) and (5.26) only tell us the order of B as a function of n .

Like the determination of the sample size n , the selection of B is not an easy problem. Sometimes it depends largely on experience. As a rule of thumb, Efron and Tibshirani (1986) suggested, based on rule (5.22), that for bootstrap moment estimators, B should be between 50 and 200; and for bootstrap distribution and quantile estimators, B should be at least 1000.

Booth and Hall (1994) proposed a method to select Monte Carlo sample sizes in the iterated bootstrap method introduced in Section 4.3.1.

5.4.2 Balanced bootstrap resampling

Although the simple Monte Carlo method is easy to use, the cost and time of the computations may be a burden, especially when the statistic is complicated. For example, suppose a confidence interval based on an L-statistic is to be constructed using the bootstrap-t method and the bootstrap variance estimator. To approximate the bootstrap-t percentile, we need to use a Monte Carlo approximation of size 1000 or more. For each Monte Carlo sample, suppose that the bootstrap variance estimate is approximated by Monte Carlo with size 100. Then, in total, we have to order at least 10^5 sets of data of size $n!$ Therefore, it is necessary to develop some bootstrap computational methods that are more efficient than the simple Monte Carlo method so that we can reduce the number of computations. We begin with the *balanced bootstrap resampling* method introduced by Davison, Hinkley and Schechtman (1986).

Let $\{X_i^{*b}, i = 1, \dots, n, b = 1, \dots, B\}$ be the bootstrap data generated by some resampling plan. For $i = 1, \dots, n$ and $b = 1, \dots, B$, define

$$P_i^{*b} = (\text{the number of } X_j^{*b} = X_i, j = 1, \dots, n)/n, \quad (5.27)$$

which are resampling probabilities and satisfy $\sum_{i=1}^n P_i^{*b} = 1$ for any b .

For the simple Monte Carlo method, $n(P_1^{*b}, \dots, P_n^{*b})$, $b = 1, \dots, B$, are i.i.d. from a multinomial distribution. In the special case where $\mathfrak{R}_n = \bar{X}_n - EX_1$, the simple Monte Carlo approximation to the bootstrap bias estimator b_{BOOT} in (5.15) is, by (5.18),

$$b_{\text{BOOT}}^{(B)} = \frac{1}{B} \sum_{b=1}^B \left(\frac{1}{n} \sum_{i=1}^n X_i^{*b} \right) - \bar{X}_n = \frac{1}{B} \sum_{b=1}^B \sum_{i=1}^n P_i^{*b} X_i - \bar{X}_n.$$

$b_{\text{BOOT}}^{(B)}$ is usually not 0, whereas b_{BOOT} is always equal to 0 for this special case. Therefore, $b_{\text{BOOT}}^{(B)}$ has a simulation error and we need a large B in order to obtain an accurate approximation. The balanced bootstrap resampling method is designed to eliminate this simulation error by imposing a restriction on the resampling probabilities in (5.27). Suppose that $\{X_1^{*b}, \dots, X_n^{*b}\}$, $b = 1, \dots, B$, are generated according to a resampling plan with the probabilities P_i^{*b} satisfying an additional condition:

$$\sum_{b=1}^B P_i^{*b} = \frac{B}{n}. \quad (5.28)$$

Let $b_{\text{BB}}^{(B)}$, $v_{\text{BB}}^{(B)}$, and $H_{\text{BB}}^{(B)}$ be the approximations to the bootstrap bias, variance, and distribution estimators (5.15)–(5.17), respectively, using this particular resampling method. In the special case of $\mathfrak{R}_n = \bar{X}_n - EX_1$,

(5.28) implies

$$b_{\text{BB}}^{(B)} = \frac{1}{B} \sum_{i=1}^n \sum_{b=1}^B P_i^{*b} X_i - \bar{X}_n = 0,$$

i.e., there is no simulation error.

Any method that generates bootstrap data with resampling probabilities $\{P_i^{*b} : i = 1, \dots, n, b = 1, \dots, B\}$ satisfying (5.28) is called a balanced bootstrap resampling method. This name comes from the fact that the resampling plan satisfying (5.28) is similar to a balanced randomized block design. Note that (5.28) implies that each $X_i, i = 1, \dots, n$, appears exactly B times in the bootstrap data $\{X_i^{*b}, i = 1, \dots, n, b = 1, \dots, B\}$ and $n(P_1^{*b}, \dots, P_n^{*b})$ has a multivariate hypergeometric distribution.

Balanced bootstrap samples can be obtained from the following operation. We first copy each observation X_i precisely B times so that we have an array of length Bn . Then we randomly permute these Bn numbers and take the $[(b-1)n+1]$ th, ..., the (bn) th elements in the permuted array as the b th bootstrap sample, $b = 1, \dots, B$. An algorithm describing this procedure and some improvements over it are given by Gleason (1988).

Since, for approximating the bootstrap bias estimator, balanced bootstrap resampling eliminates the simulation error in the special case of $\mathfrak{R}_n = \bar{X}_n - E X_1$, we expect that it also improves the simple Monte Carlo method in the case where $\mathfrak{R}_n = \hat{\theta}_n - E \hat{\theta}_n$ and $\hat{\theta}_n$ can be approximated by a linear statistic. We now show that this is true for the case where X_i is univariate and $\hat{\theta}_n = g(\bar{X}_n)$ with a smooth function g . Assuming some moment conditions and using Taylor's expansion and the formulas for cross-product moments of multinomial and multivariate hypergeometric distributions, Davison, Hinkley and Schechtman (1986) and Hall (1989a) showed that

$$E_{\text{MC}}(b_{\text{BOOT}}^{(B)}) = \frac{g''(\bar{X}_n)}{2n^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2 + O_p\left(\frac{1}{n^2}\right)$$

and

$$E_{\text{BB}}(b_{\text{BB}}^{(B)}) = \frac{n(B-1)}{nB-1} \frac{g''(\bar{X}_n)}{2n^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2 + O_p\left(\frac{1}{n^2 B}\right),$$

where E_{MC} and E_{BB} denote the conditional expectations (given X_1, \dots, X_n) with respect to simple Monte Carlo resampling and balanced bootstrap resampling, respectively, and $B \rightarrow \infty$ as $n \rightarrow \infty$ is assumed. Under some further conditions on g and the distribution of X_1 , we can similarly show that

$$\text{var}_{\text{MC}}(b_{\text{BOOT}}^{(B)}) = \frac{[g'(\bar{X}_n)]^2}{n^2 B} \sum_{i=1}^n (X_i - \bar{X}_n)^2 + O_p\left(\frac{1}{n^2 B}\right),$$

$$\text{var}_{\text{BB}}(b_{\text{BB}}^{(B)}) = \frac{[g''(\bar{X}_n)]^2}{2n^4 B} \left[\sum_{i=1}^n (X_i - \bar{X}_n)^2 \right]^2 + O_p\left(\frac{1}{n^2 B^2} + \frac{1}{n^3 B}\right),$$

and, therefore,

$$\text{mse}_{\text{MC}}(b_{\text{BOOT}}^{(B)}) = O_p\left(\frac{1}{nB}\right) \quad \text{and} \quad \text{mse}_{\text{BB}}(b_{\text{BB}}^{(B)}) = O_p\left(\frac{1}{n^2 B}\right),$$

where var_{MC} and var_{BB} (mse_{MC} and mse_{BB}) are the variances (mean squared errors) with respect to simple Monte Carlo resampling and balanced bootstrap resampling, respectively. The order of magnitude of the mean square error is reduced from $(nB)^{-1}$ to $(n^2 B)^{-1}$ by using the balanced bootstrap resampling method in approximating the bootstrap bias estimator. This result can be extended to multivariate X_i and general second order differentiable functional statistics.

As a numerical example, Davison, Hinkley and Schechtman (1986) calculated the variances of the simple Monte Carlo and balanced bootstrap resampling approximations to the bootstrap bias estimator of the estimator $\hat{\theta}_n = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ based on the sample displayed in (5.14) with $n = 10$. In this case, for $B = 19$, $\text{var}_{\text{MC}}(b_{\text{BOOT}}^{(B)}) = 14.31$ and $\text{var}_{\text{BB}}(b_{\text{BB}}^{(B)}) = 1.98$.

Davison, Hinkley and Schechtman (1986) and Hall (1990c) also compared the balanced bootstrap resampling method with the simple Monte Carlo method for variance, distribution, and quantile estimators.

Again, we consider the case of $\hat{\theta}_n = g(\bar{X}_n)$. Then

$$E_{\text{MC}}(v_{\text{BOOT}}^{(B)}) = \frac{[g'(\bar{X}_n)]^2}{n^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2 + o_p\left(\frac{1}{n}\right),$$

$$\text{var}_{\text{MC}}(v_{\text{BOOT}}^{(B)}) = \frac{2[g'(\bar{X}_n)]^4}{n^4 B} \left[\sum_{i=1}^n (X_i - \bar{X}_n)^2 \right]^2 + o_p\left(\frac{1}{n^2 B}\right),$$

and, therefore,

$$n^2 B \text{mse}_{\text{MC}}(v_{\text{BOOT}}^{(B)}) \rightarrow_p 2[g'(EX_1)]^4 [\text{var}(X_1)]^2. \quad (5.29)$$

Similarly,

$$E_{\text{BB}}(v_{\text{BB}}^{(B)}) = \frac{n(B-1)}{nB-1} \frac{[g'(\bar{X}_n)]^2}{n^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2 + o_p\left(\frac{1}{n}\right),$$

$$\text{var}_{\text{BB}}(v_{\text{BB}}^{(B)}) = \frac{2B^2 [g'(\bar{X}_n)]^4}{n(nB-1)(nB-2)(nB-3)} \left[\sum_{i=1}^n (X_i - \bar{X}_n)^2 \right]^2 + o_p\left(\frac{1}{n^2 B}\right),$$

and (5.29) holds with $\text{mse}_{\text{MC}}(v_{\text{BOOT}}^{(B)})$ replaced by $\text{mse}_{\text{BB}}(v_{\text{BB}}^{(B)})$. Thus, in terms of the mean squared error, the simple Monte Carlo and the balanced bootstrap resampling methods produce asymptotically equivalent approximations to the bootstrap variance estimator. For the numerical example considered earlier, Davison, Hinkley and Schechtman (1986) found that when $\hat{\theta}_n = \bar{X}_n$, $\text{var}_{\text{BB}}(v_{\text{BB}}^{(B)}) = 2.17$ and $\text{var}_{\text{MC}}(v_{\text{BOOT}}^{(B)}) = 2.46$. Hence, the balanced bootstrap resampling does not greatly improve the simple Monte Carlo approximation.

Let us now consider the bootstrap estimators of distribution functions and quantiles. It is apparent that

$$E_{\text{MC}}[H_{\text{BOOT}}^{(B)}(x)] = H_{\text{BOOT}}(x)$$

and

$$\text{var}_{\text{MC}}[H_{\text{BOOT}}^{(B)}(x)] = B^{-1}[1 - H_{\text{BOOT}}(x)]H_{\text{BOOT}}(x).$$

Consider the special case of

$$\mathfrak{R}_n = [g(\bar{X}_n) - g(EX_1)]/\hat{\sigma}_n,$$

where $\hat{\sigma}_n^2 = n^{-2}[g'(\bar{X}_n)]^2 \sum_{i=1}^n (X_i - \bar{X}_n)^2$. From the strong consistency of the bootstrap estimator,

$$B\text{mse}_{\text{MC}}(H_{\text{BOOT}}^{(B)}(x)) \rightarrow_{a.s.} [1 - \Phi(x)]\Phi(x)$$

for every fixed x . Using the properties of the multivariate hypergeometric distributions and assuming some conditions on the distribution of X_1 , Hall (1990c) showed that

$$B\text{mse}_{\text{BB}}(H_{\text{BB}}^{(B)}(x)) \rightarrow_{a.s.} [1 - \Phi(x)]\Phi(x) - [\varphi(x)]^2.$$

Therefore,

$$\frac{\text{mse}_{\text{MC}}(H_{\text{BOOT}}^{(B)}(x))}{\text{mse}_{\text{BB}}(H_{\text{BB}}^{(B)}(x))} \rightarrow_{a.s.} \frac{[1 - \Phi(x)]\Phi(x)}{[1 - \Phi(x)]\Phi(x) - [\varphi(x)]^2}. \quad (5.30)$$

From (5.30) we know that the balanced bootstrap resampling method is an improvement over the simple Monte Carlo method in terms of the mean squared error when n is large. The maximum improvement is at $x = 0$. The relative decrease in the mean squared error at $x = 0$ is 63.8%.

Using the Bahadur representations of $(H_{\text{BOOT}}^{(B)})^{-1}(p)$ and $(H_{\text{BB}}^{(B)})^{-1}(p)$, Hall (1990c) showed that

$$\frac{\text{mse}_{\text{MC}}((H_{\text{BOOT}}^{(B)})^{-1}(p))}{\text{mse}_{\text{BB}}((H_{\text{BB}}^{(B)})^{-1}(p))} \rightarrow_{a.s.} \frac{p(1-p)}{p(1-p) - [\varphi(z_p)]^2}.$$

Hence, the balanced bootstrap resampling method is also better than the simple Monte Carlo method in this case. The maximum decrease in the mean squared error is by a factor of 2.75 for $p = 0$. When $p = 0.975$, the reduction in mean squared error is 14%. This was also demonstrated by Davison, Hinkley and Schechtman (1986) through some numerical examples.

From the above discussion, we know that the balanced bootstrap resampling method, as it is designed, improves the simple Monte Carlo method in approximating the bootstrap bias estimator. For approximating the bootstrap variance estimator, the tails of the bootstrap distribution estimator, and the bootstrap quantile estimator with p close to 0 or 1, the improvement of the balanced bootstrap resampling method over the simple Monte Carlo method is not great.

Graham *et al.* (1990) introduced a higher order balanced bootstrap resampling method that eliminates the simulation errors for bootstrap bias estimators of the statistics that can be expressed as

$$\begin{aligned} T_n = \theta + \frac{1}{n} \sum_{i=1}^n \alpha(X_i, F) + \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \beta(X_i, X_j, F) + \dots \\ \dots + \frac{1}{n^k} \sum_{i_1=1}^n \dots \sum_{i_k=1}^n \gamma(X_{i_1}, \dots, X_{i_k}, F). \end{aligned}$$

But higher order balanced bootstrap samples are difficult to generate. Recall that in the context of experimental design, higher order balanced randomized block designs are not easy to construct.

5.4.3 Centering after Monte Carlo

In approximating the bootstrap bias estimator, the balanced bootstrap resampling method improves the simple Monte Carlo method by changing the bootstrap resampling plan. Efron (1990) introduced another method that still uses the simple Monte Carlo method but reduces the simulation error by adjusting the resulting Monte Carlo approximation.

First, we consider the approximation to the bootstrap bias estimator for an estimator $\hat{\theta}_n$. In many cases, $\hat{\theta}_n$ can be written as

$$\hat{\theta}_n = \hbar_n(\mathbf{P}^0),$$

where \mathbf{P}^0 is an n -vector whose components are all equal to n^{-1} and \hbar_n is an n -variate function depending on X_1, \dots, X_n . For example,

$$\bar{X}_n = \mathbf{X}' \mathbf{P}^0 = \hbar_n(\mathbf{P}^0)$$

with $\hbar(t_1, \dots, t_n) = \sum_{i=1}^n t_i X_i$ and $\mathbf{X} = (X_1, \dots, X_n)'$. For $X_1^{*b}, \dots, X_n^{*b}$ i.i.d. from F_n , $b = 1, \dots, B$, let P_i^{*b} be given in (5.27) and

$$\mathbf{P}^{*b} = (P_1^{*b}, \dots, P_n^{*b})', \quad b = 1, \dots, B.$$

Then the simple Monte Carlo approximation to the bootstrap bias estimator is, by (5.18),

$$b_{\text{BOOT}}^{(B)} = \frac{1}{B} \sum_{b=1}^B \hbar_n(\mathbf{P}^{*b}) - \hbar_n(\mathbf{P}^0). \quad (5.31)$$

If \hbar_n is a linear function in the sense that

$$\hbar_n(\mathbf{P}) = a_{n0} + (\mathbf{P} - \mathbf{P}^0)' \mathbf{U}, \quad \mathbf{P} \in \mathbb{R}^n, \quad (5.32)$$

where a_{n0} and the components of \mathbf{U} are some functions of X_1, \dots, X_n , then $b_{\text{BOOT}} \equiv 0$; whereas $b_{\text{BOOT}}^{(B)} = (\bar{\mathbf{P}}^* - \mathbf{P}^0)' \mathbf{U} \neq 0$, where $\bar{\mathbf{P}}^* = B^{-1} \sum_{b=1}^B \mathbf{P}^{*b}$ is the center of \mathbf{P}^{*b} . Efron (1990) argued that, for the Monte Carlo approximation $b_{\text{BOOT}}^{(B)}$ in (5.31), \mathbf{P}^0 should be replaced by the center $\bar{\mathbf{P}}^*$. This adjustment is called “centering”. Define the *centered* Monte Carlo approximation of the bootstrap bias estimator by

$$b_{\text{BC}}^{(B)} = \frac{1}{B} \sum_{b=1}^B \hbar_n(\mathbf{P}^{*b}) - \hbar_n(\bar{\mathbf{P}}^*).$$

Then $b_{\text{BC}}^{(B)} \equiv 0$ for a linear statistic defined in (5.32). For functions of the sample mean, Hall (1989a) proved that

$$\text{mse}_{\text{MC}}(b_{\text{BC}}^{(B)}) = O_p\left(\frac{1}{n^2 B}\right).$$

Therefore, the centered Monte Carlo approximation improves the simple Monte Carlo approximation in terms of the mean squared error.

The extension of the centered Monte Carlo method to the case of approximating the bootstrap variance or quantile estimator is not straightforward. It relies on the orthogonal decomposition (or ANOVA decomposition) of $\hbar_n(\mathbf{P}^{*b})$ discussed by Efron (1982):

$$\hbar_n(\mathbf{P}^{*b}) = a_{n0} + \alpha_n(\mathbf{P}^{*b}) + \beta_n(\mathbf{P}^{*b}), \quad (5.33)$$

where $a_{n0} = E_{\text{MC}}[\hbar_n(\mathbf{P}^{*b})]$, $\alpha_n(\mathbf{P}^{*b}) = \boldsymbol{\alpha}' \mathbf{P}^{*b}$, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)'$ with

$$\alpha_i = n\{E_{\text{MC}}[\hbar_n(\mathbf{P}^{*b}) | X_1^{*b} = X_i] - a_{n0}\}, \quad i = 1, \dots, n, \quad (5.34)$$

and $\beta_n(\mathbf{P}^{*b})$ is the residual term. Define

$$\bar{\theta}_n = \frac{1}{B} \sum_{b=1}^B \hbar_n(\mathbf{P}^{*b}),$$

$$\mathbf{T} = (\hbar_n(\mathbf{P}^{*1}) - \bar{\theta}_n, \dots, \hbar_n(\mathbf{P}^{*B}) - \bar{\theta}_n)',$$

and

$$\mathbf{Q} = (\mathbf{P}^{*1} - \bar{\mathbf{P}}^*, \dots, \mathbf{P}^{*B} - \bar{\mathbf{P}}^*).$$

Efron (1990) showed that the simple Monte Carlo approximation to the bootstrap variance estimator can be written as

$$v_{\text{BOOT}}^{(B)} = \hat{\alpha}' \hat{\Sigma} \hat{\alpha} + \hat{V}_{\beta}, \quad (5.35)$$

where $\hat{\Sigma} = \frac{1}{B} \mathbf{Q} \mathbf{Q}'$, $\hat{\alpha} = (\mathbf{Q} \mathbf{Q}' + \mathbf{1}\mathbf{1}')^{-1} \mathbf{Q} \mathbf{T}$, $\mathbf{1} = (1, \dots, 1)'$, and

$$\hat{V}_{\beta} = \frac{1}{B} \sum_{b=1}^B [\hbar_n(\mathbf{P}^{*b}) - \bar{\theta}_n - (\mathbf{P}^{*b} - \bar{\mathbf{P}}^*)' \hat{\alpha}]^2.$$

From (5.33), the bootstrap variance estimator is equal to

$$\begin{aligned} v_{\text{BOOT}} &= \text{var}_{\text{MC}}[\alpha_n(\mathbf{P}^{*1}) + \beta_n(\mathbf{P}^{*1})] = \text{var}_{\text{MC}}(\alpha' \mathbf{P}^{*1}) + V_{\beta} \\ &= \alpha' \text{var}_{\text{MC}}(\mathbf{P}^{*1}) \alpha + V_{\beta} = \alpha' \alpha / n^2 + V_{\beta}, \end{aligned} \quad (5.36)$$

where $V_{\beta} = \text{var}_{\text{MC}}[\beta_n(\mathbf{P}^{*1})]$. Note that $\hat{\Sigma}$ in (5.35) is an estimate of $\text{var}_{\text{MC}}(\mathbf{P}^{*1})$ in (5.36), but this estimation is unnecessary since v_{BOOT} actually does not depend on $\text{var}_{\text{MC}}(\mathbf{P}^{*1})$ according to (5.36). Also, \hat{V}_{β} is the Monte Carlo approximation to V_{β} . Therefore, we can reduce the simulation error by changing $\hat{\alpha}' \hat{\Sigma} \hat{\alpha}$ to $\hat{\alpha}' \hat{\alpha} / n^2$. This leads to the centered Monte Carlo approximation

$$v_{\text{BC}}^{(B)} = \hat{\alpha}' \hat{\alpha} / n^2 + \hat{V}_{\beta}.$$

Efron (1990) suggested a further adjustment for reducing the bias of $v_{\text{BC}}^{(B)}$ and proposed the following approximation to v_{BOOT} :

$$\tilde{v}_{\text{BC}}^{(B)} = \hat{\alpha}' \hat{\alpha} / n^2 + d_{n,B} \hat{V}_{\beta}, \quad (5.37)$$

where

$$d_{n,B} = \max \left\{ \left[1 - \frac{n(n-1)}{(B-1)(B-n-1)} \right], 0 \right\}.$$

The calculation of $\tilde{v}_{\text{BC}}^{(B)}$ involves the inversion of an $n \times n$ matrix. We may replace $\hat{\alpha}$ by a jackknife estimate to avoid this inversion. Let \mathbf{P}_i^0 be the n -vector whose i th component is 0 and other components are $(n-1)^{-1}$, $i = 1, \dots, n$, $\tilde{\theta}_n = n^{-1} \sum_{i=1}^n \hbar_n(\mathbf{P}_i^0)$, and $\tilde{\mathbf{U}}$ be the n -vector whose i th component is $(n-1)[\tilde{\theta}_n - \hbar_n(\mathbf{P}_i^0)]$, $i = 1, \dots, n$. Then $\hat{\alpha}$ in (5.37) can be replaced by

$$\frac{\sum_{b=1}^B \tilde{\mathbf{U}}' \mathbf{P}^{*b} [\hbar_n(\mathbf{P}^{*b}) - \bar{\theta}_n]}{B \tilde{\mathbf{U}}' \hat{\Sigma} \tilde{\mathbf{U}}} \tilde{\mathbf{U}}.$$

Define

$$R^2 = \text{var}_{\text{MC}}[\alpha_n(\mathbf{P}^{*1})]/\text{var}_{\text{MC}}[\beta_n(\mathbf{P}^{*1})]$$

and

$$\rho = \text{var}_{\text{MC}}(v_{\text{BOOT}}^{(B)})/\text{var}_{\text{MC}}(\tilde{v}_{\text{BC}}^{(B)}).$$

Through a series of simulation experiments, Efron (1990) demonstrated that the closer R^2 is to 1, the larger is ρ . For example, when $B = 100$, for the ratio estimate of two independent normal samples of size 10, $R^2 = 0.976$ and $\rho = 8.7$; for the 20% trimmed mean of a negative exponential sample of size 10, $R^2 = 0.906$ and $\rho = 3.5$. The conclusion is that if the function \hbar_n is nearly linear in the sense of (5.32), then centering greatly improves the simple Monte Carlo approximation.

We next consider the bootstrap quantile estimator

$$K_{\text{BOOT}}^{-1}(p), \quad (5.38)$$

where $K_{\text{BOOT}}(x) = P_*\{\hat{\theta}_n^* \leq x\}$, $\hat{\theta}_n^* = \hbar_n(\mathbf{P}^{*1})$ and p is a fixed constant in $(0, 1)$. The simple Monte Carlo approximation to $K_{\text{BOOT}}^{-1}(p)$ is just the p th quantile of the values $\hbar_n(\mathbf{P}^{*1}), \dots, \hbar_n(\mathbf{P}^{*B})$. Using the decomposition (5.33), we may reduce the simulation error by adjusting $\hbar_n(\mathbf{P}^{*b})$ to

$$\tilde{\theta}_n^{*b} = a_{n0} + l_n^{*b} + \beta_n(\mathbf{P}^{*b}),$$

with l_n^{*b} having stochastic behavior very similar to $\alpha_n(\mathbf{P}^{*b})$, the linear component of $\hbar_n(\mathbf{P}^{*b})$, $b = 1, \dots, B$. For example, we may use an l_n^{*b} whose first four cumulants match those of $\alpha_n(\mathbf{P}^{*b})$. The centered Monte Carlo approximation to $K_{\text{BOOT}}^{-1}(p)$, denoted by $(K_{\text{BC}}^{(B)})^{-1}(p)$, is the p th quantile of the adjusted values $\tilde{\theta}_n^{*b}$, $b = 1, \dots, B$.

To implement this method, Efron (1990) suggested an algorithm based on a cumulant adjustment formula. Suppose that y is a random variable with mean 0, variance 1, skewness γ_y , and kurtosis δ_y . Then for any γ_x and δ_x , the random variable

$$\begin{aligned} x &= \left[1 - \frac{\delta_x - \delta_y}{8} + \frac{\gamma_x - \gamma_y}{36}(5\gamma_x + 7\gamma_y)\right]y + \left[\frac{\gamma_x - \gamma_y}{6}\right](y^2 - 1) \\ &\quad + \left[\frac{\delta_x - \delta_y}{24} - \frac{\gamma_x - \gamma_y}{9}\left(\frac{\gamma_x}{2} + \gamma_y\right)\right]y^3 \end{aligned} \quad (5.39)$$

has mean 0, variance 1, skewness γ_x , and kurtosis δ_x . Approximately, $\alpha_n(\mathbf{P}^{*b})$ has mean 0, standard deviation σ_α/\sqrt{n} , skewness γ_α/\sqrt{n} , and kurtosis δ_α/\sqrt{n} , where

$$\sigma_\alpha^2 = \frac{1}{n} \sum_{i=1}^n \alpha_i^2, \quad \gamma_\alpha = \frac{1}{n\sigma_\alpha^3} \sum_{i=1}^n \alpha_i^3, \quad \delta_\alpha = \frac{1}{n\sigma_\alpha^4} \sum_{i=1}^n \alpha_i^4 - 3,$$

and α_b is given in (5.34). Let y be a discrete uniform random variable taking values $\Phi^{-1}((b - \frac{1}{2})/B)$, $b = 1, \dots, B$, and x be the transformation of y using (5.39) with $\gamma_x = \gamma_\alpha/\sqrt{n}$ and $\delta_x = \delta_\alpha/\sqrt{n}$. Then we can take the b th ordered support point of $\sigma_\alpha x/\sqrt{n}$ as the b th ordered value of l_n^{*b} , $b = 1, \dots, B$.

For the ratio estimates of two independent normal samples of size 10, Efron (1990) found that when $B = 100$, the ratio

$$\text{var}_{\text{MC}}[(K_{\text{BOOT}}^{(B)})^{-1}(p)]/\text{var}_{\text{MC}}[(K_{\text{BC}}^{(B)})^{-1}(p)]$$

is 6.7 if $p = 0.025$, 11.14 if $p = 0.05$, 68.6 if $p = 0.50$, 8.4 if $p = 0.95$, and 5.5 if $p = 0.975$. Note that balanced bootstrap resampling reduces the variance by a factor of up to 2.75. Hence, the centering method is better than the balanced bootstrap resampling method for quantile estimates. Efron (1990) even showed that the centering method is better than second order balanced bootstrap resampling (Graham *et al.* 1990). The centering method, however, involves adjustments that require heavy theoretical derivations.

Do and Hall (1992) studied this method for approximating $K_{\text{BOOT}}(x)$. Their simulation results showed that the approximation in the tails of K_{BOOT} is not as good as that in the center of K_{BOOT} .

5.4.4 The linear bootstrap

The centered Monte Carlo method is based on the idea of first decomposing the statistic $\hat{\theta}_n$ into a linear statistic and a remainder, and then improving the approximation to the bootstrap estimator with $\hat{\theta}_n$ replaced by its linear component. In some cases, the bootstrap estimators for the linear component of $\hat{\theta}_n$ can be exactly obtained by substituting F_n for F in Taylor's expansion of $\hat{\theta}_n$ or, more generally, the influence function of $\hat{\theta}_n$ introduced in Chapter 2. In such cases, we can then obtain better approximations to the bootstrap estimators for $\hat{\theta}_n$ by using the Monte Carlo method only for approximating the corresponding bootstrap estimators for the remainder of $\hat{\theta}_n$. This is called the method of resampling after linear approximation, or simply the *linear bootstrap*.

Suppose that the estimator $\hat{\theta}_n$ of a parameter θ has the expansion

$$\hat{\theta}_n = \theta + \frac{1}{n} \sum_{i=1}^n \phi_F(X_i) + R_n, \quad (5.40)$$

where $\phi_F(x)$ is a function satisfying $\int \phi_F(x)dF(x) = 0$ for any F . The function ϕ_F is the influence function when $\hat{\theta}_n$ is generated by a differentiable

functional. Define

$$\hat{\theta}_{nL} = \theta + \frac{1}{n} \sum_{i=1}^n \phi_F(X_i).$$

Then $\hat{\theta}_{nL}$ is the linear component of $\hat{\theta}_n$ and $R_n = \hat{\theta}_n - \hat{\theta}_{nL}$. Suppose that ϕ_F is well defined when F is replaced by the empirical distribution F_n . Let

$$\hat{\theta}_{nL}^* = \hat{\theta}_n + \frac{1}{n} \sum_{i=1}^n \phi_{F_n}(X_i^*)$$

and

$$R_n^* = \hat{\theta}_n^* - \hat{\theta}_{nL}^*$$

be the bootstrap analogs of $\hat{\theta}_{nL}$ and R_n , respectively. Since

$$E_* \left[\frac{1}{n} \sum_{i=1}^n \phi_{F_n}(X_i^*) \right] = \int \phi_{F_n}(x) dF_n(x) = 0,$$

we have

$$b_{\text{BOOT}} = E_*(\hat{\theta}_n^*) - \hat{\theta}_n = E_*(\hat{\theta}_{nL}^* + R_n^*) - \hat{\theta}_n = E_*(R_n^*) \quad (5.41)$$

and

$$\begin{aligned} v_{\text{BOOT}} &= \text{var}_*(\hat{\theta}_n^*) = \text{var}_*(\hat{\theta}_{nL}^* + R_n^*) \\ &= \frac{1}{n^2} \sum_{i=1}^n [\phi_{F_n}(X_i)]^2 + 2\text{cov}_*(R_n^*, \hat{\theta}_{nL}^*) + \text{var}_*(R_n^*). \end{aligned} \quad (5.42)$$

To compute b_{BOOT} and v_{BOOT} , we only need to apply the Monte Carlo method to approximate the terms in (5.41) and (5.42) related to R_n^* . Let $\{X_1^{*b}, \dots, X_n^{*b}\}$ be the b th independent bootstrap sample from F_n , $b = 1, \dots, B$, $\hat{\theta}_{nL}^{*b}$ and R_n^{*b} be $\hat{\theta}_{nL}^*$ and R_n^* , respectively, with X_i^* replaced by X_i^{*b} . The linear bootstrap approximations to b_{BOOT} and v_{BOOT} are, respectively,

$$b_{\text{BL}}^{(B)} = \frac{1}{B} \sum_{b=1}^B R_n^{*b}$$

and

$$v_{\text{BL}}^{(B)} = \frac{1}{n^2} \sum_{i=1}^n [\phi_{F_n}(X_i)]^2 + \frac{1}{B} \sum_{b=1}^B [2(\hat{\theta}_{nL}^{*b} - \hat{\theta}_n)R_n^{*b} + (R_n^{*b} - \bar{R}_n^*)^2],$$

where $\bar{R}_n^* = B^{-1} \sum_{b=1}^B R_n^{*b}$. The leading term of $v_{\text{BL}}^{(B)}$ contains no simulation error.

When $\hat{\theta}_n = \bar{X}_n$, $b_{\text{BL}}^{(B)} \equiv 0$. When $\hat{\theta}_n$ is a function of \bar{X}_n , Hall (1989a) proved that

$$\text{var}_{\text{MC}}(b_{\text{BL}}^{(B)}) = O_p\left(\frac{1}{n^2 B}\right).$$

Hence, the linear bootstrap method improves the simple Monte Carlo approximation in terms of the mean squared error.

The linear bootstrap can be viewed as a mixture of, and perhaps an improvement over, the delta method and the simple Monte Carlo method. The price paid for the improvement over the delta method is the extra resampling after linearization; and the extra work required for the improvement over the simple Monte Carlo method is the theoretical derivation of the influence function of $\hat{\theta}_n$, which has to be well defined at $F = F_n$ and can be easily computed. The jackknife method may be used to approximate the influence function, as we illustrated in Section 5.3.2.

If we know more terms in the stochastic expansion of $\hat{\theta}_n$, then we can get a method to approximate the bootstrap bias and variance estimators with much higher accuracy. However, higher order stochastic expansions are usually very difficult to obtain for complicated statistics.

The performance of the linear bootstrap may also be improved by using balanced bootstrap resampling in approximating the terms related to R_n^* in (5.41) and (5.42). Davison, Hinkley and Schechtman (1986) presented some numerical examples.

5.4.5 Antithetic bootstrap resampling

Antithetic sampling is a traditional technique for reducing errors in Monte Carlo simulations. Its basic idea can be described as follows. Suppose that $\hat{\theta}_{1n}$ is an estimator of an unknown parameter θ . If we can find another estimator $\hat{\theta}_{2n}$ of θ that has the same variance as $\hat{\theta}_{1n}$ but is negatively correlated to $\hat{\theta}_{1n}$, then the variance of the new estimator

$$\hat{\theta}_n = (\hat{\theta}_{1n} + \hat{\theta}_{2n})/2$$

is less than half of that of $\hat{\theta}_{1n}$.

Hall (1989b) applied this technique to improve the simple Monte Carlo method. Let $\{X_1^{*b}, \dots, X_n^{*b}\}$, $b = 1, \dots, B$, be independent bootstrap samples from F_n and $b_{\text{BOOT}}^{(B)}$, $v_{\text{BOOT}}^{(B)}$, and $H_{\text{BOOT}}^{(B)}$ be the simple Monte Carlo approximations defined in (5.18)–(5.20), respectively. To find another Monte Carlo approximation to b_{BOOT} (v_{BOOT} or H_{BOOT}) that has the same variance as, but is negatively correlated with, $b_{\text{BOOT}}^{(B)}$ ($v_{\text{BOOT}}^{(B)}$ or $H_{\text{BOOT}}^{(B)}$), we can make use of the bootstrap samples $\{X_1^{*b}, \dots, X_n^{*b}\}$, $b = 1, \dots, B$. Note that $X_i^{*b} = X_{\zeta(b,i)}$, where $\zeta(b,i)$, $i = 1, \dots, n$, are independently

and uniformly distributed on $\{1, \dots, n\}$ for each $b = 1, \dots, B$. For example, if $X_1^{*b} = X_3$, then $\zeta(b, 1) = 3$. We define new bootstrap samples as $\{\tilde{X}_i^{*b} = X_{\pi(\zeta(b, i))}, i = 1, \dots, n\}$, $b = 1, \dots, B$, where π is a permutation of integers $1, \dots, n$. Let $\tilde{b}_{\text{BOOT}}^{(B)}$, $\tilde{v}_{\text{BOOT}}^{(B)}$, and $\tilde{H}_{\text{BOOT}}^{(B)}$ be given by (5.18)–(5.20), respectively, with X_i^{*b} replaced by the new bootstrap sample \tilde{X}_i^{*b} . Then the antithetic Monte Carlo approximations to the bootstrap bias, variance, and distribution estimators are, respectively,

$$\begin{aligned} b_{\text{BA}}^{(B)} &= (b_{\text{BOOT}}^{(B)} + \tilde{b}_{\text{BOOT}}^{(B)})/2, \\ v_{\text{BA}}^{(B)} &= (v_{\text{BOOT}}^{(B)} + \tilde{v}_{\text{BOOT}}^{(B)})/2, \end{aligned}$$

and

$$H_{\text{BA}}^{(B)} = (H_{\text{BOOT}}^{(B)} + \tilde{H}_{\text{BOOT}}^{(B)})/2.$$

The problem now is how to obtain the permutation π so that $\tilde{b}_{\text{BOOT}}^{(B)}$ ($\tilde{v}_{\text{BOOT}}^{(B)}$ or $\tilde{H}_{\text{BOOT}}^{(B)}$) has the same variance as, but is negatively correlated with, $b_{\text{BOOT}}^{(B)}$ ($v_{\text{BOOT}}^{(B)}$ or $H_{\text{BOOT}}^{(B)}$). For the case where X_i is univariate and $\hat{\theta}_n = g(\bar{X}_n)$, Hall (1989b) proposed to first replace X_1, \dots, X_n by the order statistics $X_{(1)} \leq \dots \leq X_{(n)}$ and then use $\pi(i) = n - i + 1$. For example, if $X_1^{*b} = X_{(5)}$, then $\tilde{X}_1^{*b} = X_{(n-5+1)} = X_{(n-4)}$. Hall (1989b) proved that such a π not only satisfies the required conditions but also minimizes the variances of the resulting antithetic Monte Carlo approximations over all possible π .

In the special case of $\hat{\theta}_n = \bar{X}_n$, Hall (1989b) showed that

$$\frac{\text{var}_{\text{MC}}(b_{\text{BOOT}}^{(2B)})}{\text{var}_{\text{MC}}(b_{\text{BA}}^{(B)})} \xrightarrow{\text{a.s.}} \rho = \frac{\text{var}(X_1)}{\text{var}(X_1) + \int \xi_p \xi_{1-p} dp - EX_1}, \quad (5.43)$$

where ξ_p is the p th quantile of X_1 . Note that $b_{\text{BOOT}}^{(2B)}$ is the simple Monte Carlo approximation based on $2B$ bootstrap samples. If X_1 has an exponential distribution, then $\rho = 2.816$. If X_1 is symmetric, then $\rho = \infty$. Therefore, antithetic resampling improves the simple Monte Carlo method; but, unlike balanced bootstrap resampling, the ratio on the left-hand side of (5.43) may not diverge to infinity.

In general, if $\hat{\theta}_n$ admits expansion (5.40) and ϕ_{F_n} is well defined, then π can be obtained as follows. Let $Y_{(1)} \leq \dots \leq Y_{(n)}$ be the ordered values of $\phi_{F_n}(X_1), \dots, \phi_{F_n}(X_n)$. Relabel the data X_1, \dots, X_n as $X_{\{1\}}, \dots, X_{\{n\}}$ so that $X_{\{i\}}$ satisfies $Y_{(i)} = \phi_{F_n}(X_{\{i\}})$. Then use $\pi(i) = n - i + 1$. For example, if $X_1^{*b} = X_3$ and $\phi_{F_n}(X_3) = Y_{(6)}$, then $\tilde{X}_1^{*b} = X_{\{n-6+1\}} = X_{\{n-5\}}$.

Consider now the approximations to the bootstrap distribution estimator H_{BOOT} of the studentized random variable $\mathfrak{R}_n = [g(\bar{X}_n) - g(EX_1)]/\hat{\sigma}_n$.

Hall (1989b) showed that

$$\frac{\text{var}_{\text{MC}}[H_{\text{BOOT}}^{(2B)}(x)]}{\text{var}_{\text{MC}}[H_{\text{BA}}^{(B)}(x)]} \rightarrow_{a.s.} \rho(x) = \frac{[1 - \Phi(x)]\Phi(x)}{[1 - \Phi(x)]\Phi(x) + D(x)},$$

where $D(x) = P\{Z \leq x, Z' \leq x\} - [\Phi(x)]^2$, (Z, Z') is distributed as a bivariate normal with 0 means, unit variances, and correlation coefficient

$$r = \left(\int_0^1 \eta_p \eta_{1-p} dp \right) / \left(\int_0^1 \eta_p^2 dp \right) < 0,$$

and η_p is the p th quantile of $\nabla g(\bar{X}_n)(X_1 - EX_1)$. It is known that $D(x) < 0$ for all x , since $r < 0$. Thus, the antithetic method performs strictly better than the simple Monte Carlo method. If X_1 is symmetric, then $\rho(x) = \max\{\Phi(x), 1 - \Phi(x)\}/|2\Phi(x) - 1|$. The improvement is great near 0 since $\rho(0) = \infty$.

Similar results can be obtained for approximating bootstrap quantile estimators. Some simulation results are given in Do (1992).

To calculate an antithetic Monte Carlo approximation, we first need to draw B sets of independent bootstrap samples from F_n and then rank the values of transformed data. The computational labor may not be as heavy as that of drawing $2B$ sets of independent bootstrap samples, yet we have seen that the antithetic Monte Carlo approximation with sample size B is more efficient than the simple Monte Carlo approximation with sample size $2B$. However, like balanced bootstrap resampling, the improvement over the simple Monte Carlo approximation is not large at the tails of the bootstrap distribution estimator. Also, like the linear bootstrap, the derivation and calculation of ϕ_{F_n} is required.

5.4.6 Importance bootstrap resampling

For approximating the bootstrap distribution and quantile estimators, the previously introduced methods do not have large improvements over the simple Monte Carlo method at the tails of the bootstrap distribution estimator. We next introduce the *importance bootstrap resampling* method, which is designed to reduce the simulation errors of the simple Monte Carlo approximation of the bootstrap distribution and quantile estimators.

To motivate the idea of importance bootstrap resampling, we consider the problem of estimating the unknown p th quantile ξ_p of a distribution H . Suppose that $\{Z_1, \dots, Z_B\}$ is an independent sample from H . A simple estimator is the sample quantile defined by $\hat{\xi}_p = H_B^{-1}(p)$, where H_B is the empirical distribution based on Z_1, \dots, Z_B . Assume that H has a positive density at ξ_p , $h(\xi_p) > 0$. Then

$$\sqrt{B}(\hat{\xi}_p - \xi_p) \rightarrow_d N(0, \gamma^2),$$

where $\gamma^2 = p(1-p)/[h(\xi_p)]^2$ (Serfling, 1980). Therefore, $\text{var}(\hat{\xi}_p) \approx \gamma^2/B$, and, if $[h(\xi_p)]^2$ is small, we need a large B to obtain an accurate $\hat{\xi}_p$.

We can obtain a more accurate quantile estimator by sampling from a different distribution if we are able to decide which distribution is used to draw the sample. Suppose that $\{Y_1, \dots, Y_B\}$ is an independent sample from a distribution G with density g . Let $Y_{(1)} \leq \dots \leq Y_{(B)}$ be the order statistics of Y_1, \dots, Y_B and

$$S_r = \frac{1}{B} \sum_{i=1}^r h(Y_{(i)})/g(Y_{(i)}), \quad r = 1, \dots, B.$$

Then we can prove that if r/n is bounded from 0 and 1, $S_r \approx H(G^{-1}(r/n))$. Thus, if $S_r \approx p$, then $Y_{(r)} \approx G^{-1}(r/n) \approx H^{-1}(S_r) \approx \xi_p$. That is, we can use the importance sampling quantile estimator of ξ_p defined by

$$\tilde{\xi}_p = Y_{(R)},$$

where R is determined by $S_{(R)} \leq p$ and $S_{(R+1)} > p$. It can be shown that

$$\sqrt{B}(\tilde{\xi}_p - \xi_p) \rightarrow_d N(0, \tilde{\gamma}^2)$$

(Johns, 1988), where

$$\tilde{\gamma}^2 = \frac{1}{[h(\xi_p)]^2} \left[\int_{-\infty}^{\xi_p} \frac{[h(y)]^2}{g(y)} dy - p^2 \right].$$

Hence, if we select $g(y)$ so that it is close to $h(y)I(y \leq \xi_p)/p^2$, then the variance of $\tilde{\xi}_p$ is very small.

Since h and ξ_p are unknown, we have to estimate g in order to carry out the importance sampling.

We now focus on the approximation of the bootstrap distribution and quantile estimators. In the simple Monte Carlo method, we take B sets of independent bootstrap samples from the empirical distribution F_n of X_1, \dots, X_n . To apply importance resampling, we can generate bootstrap samples from

$$G_n(x) = \sum_{i=1}^n g_i I\{X_i \leq x\}, \quad (5.44)$$

where $g_i > 0$ and $\sum_{i=1}^n g_i = 1$. That is, instead of resampling from the observations X_1, \dots, X_n with equal probability, we select the bootstrap samples by taking into account the importance of some observations. Let $\{Y_1^{*b}, \dots, Y_n^{*b}\}$, $b = 1, \dots, B$, be independent importance bootstrap samples from G_n , $\tilde{\theta}_n^{*b}$ be the analog of $\hat{\theta}_n$ based on $Y_1^{*b}, \dots, Y_n^{*b}$, $\tilde{\theta}_n^{*(b)}$ be the b th

ordered value of $\tilde{\theta}_n^{*b}$, $b = 1, \dots, B$, and $\{Y_1^{*(b)}, \dots, Y_n^{*(b)}\}$ be the importance bootstrap sample corresponding to $\tilde{\theta}_n^{*(b)}$. Define

$$S_r = \frac{1}{B} \sum_{b=1}^r n^{-n} \left(\prod_{i=1}^n \prod_{j=1}^n g_{ij}^{*b} \right)^{-1}, \quad r = 1, \dots, B,$$

where $g_{ij}^{*b} = g_j$ if $Y_i^{*(b)} = X_j$ and $g_{ij}^{*b} = 1$ otherwise. Then the importance bootstrap Monte Carlo approximation to $K_{\text{BOOT}}^{-1}(p)$ defined in (5.38) is

$$(K_{\text{BI}}^{(B)})^{-1}(p) = \tilde{\theta}_n^{*(R)},$$

where R satisfies $S_R \leq p$ and $S_{R+1} > p$.

If we can properly select the probabilities $\{g_i, i = 1, \dots, n\}$ in (5.44), which is actually not easy to do, then $(K_{\text{BI}}^{(B)})^{-1}(p)$ is more accurate than the simple Monte Carlo approximation $(K_{\text{BOOT}}^{(B)})^{-1}(p)$.

Consider the case where $\hat{\theta}_n$ satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h(Z_i) + o_p(1), \quad (5.45)$$

$Z_i = (X_i - \theta)/s(F)$, $s(\cdot)$ is a known scale functional, and h is a known function satisfying $E[h(Z_1)] = 0$ and $\text{var}[h(Z_1)] < \infty$. By minimizing $\text{var}_{\text{BI}}[(H_{\text{BI}}^{(B)})^{-1}(p)]$ over all possible $\{g_i, i = 1, \dots, n\}$, where var_{BI} is the conditional variance under importance resampling, Johns (1988) obtained the following solution:

$$g_i = \exp\{a_p[h_i/(\sum_{i=1}^n h_i^2)] + g_0\}, \quad i = 1, \dots, n,$$

where $h_i = h((X_i - \hat{\theta}_n)/s(F_n))$, a_p minimizes

$$\int_{-\infty}^{z_p} \frac{1}{\sqrt{2\pi}} \exp\{\frac{1}{2}(y-a)^2 - y^2\} dy$$

over a , and g_0 is chosen so that $\sum_{i=1}^n g_i = 1$. An approximate solution for a_p when $p \ll 0.5$ is $a_p = -\sqrt{1+z_p^2}$. When $p = 0.025$, $a_{0.025} = -2.18$.

One example of $\hat{\theta}_n$ satisfying (5.45) is the M-estimator of location defined as the solution of

$$\sum_{i=1}^n \psi\left(\frac{X_i - \hat{\theta}_n}{s(F_n)}\right) = 0,$$

where X_1, \dots, X_n are i.i.d. from a distribution $F(x-\theta)$, which is symmetric about θ , and $\psi(x)$ is a known function satisfying $\psi(x) = -\psi(-x)$. In this case, we can take $h = \psi$.

Do and Hall (1991) considered the importance Monte Carlo approximation for $H_{\text{BOOT}}(x)$ and $H_{\text{BOOT}}^{-1}(p)$, where H_{BOOT} is defined by (5.17) with $\mathfrak{R}_n = [g(\bar{X}_n) - g(E\bar{X}_1)]/\hat{\sigma}_n$, a studentized function of the sample mean. Let $\{Y_1^{*b}, \dots, Y_n^{*b}\}$ be an importance bootstrap sample from G_n in (5.44) and \bar{Y}_n^{*b} be the sample mean based on $Y_1^{*b}, \dots, Y_n^{*b}$. Then the importance Monte Carlo approximations to $H_{\text{BOOT}}(x)$ and $H_{\text{BOOT}}^{-1}(p)$ are, respectively,

$$H_{\text{BI}}^{(B)}(x) = \frac{1}{B} \sum_{b=1}^B I \left\{ \frac{g(\bar{Y}_n^{*b}) - g(\bar{X}_n)}{\hat{\sigma}_n^{*b}} \leq x \right\},$$

and $(H_{\text{BI}}^{(B)})^{-1}(p)$, where $(\hat{\sigma}_n^{*b})^2 = n^{-2} \sum_{i=1}^n [\nabla g(\bar{Y}_n^{*b})(Y_i^{*b} - \bar{Y}_n^{*b})]^2$. Do and Hall (1991) showed that for any fixed x , the $\{g_i(x), i = 1, \dots, n\}$ minimizing $\text{var}_{\text{BI}}[(H_{\text{BI}}^{(B)}(x))]$ is given by

$$g_i(x) = \exp(-A_x \varepsilon_i) / \sum_{i=1}^n \exp(-A_x \varepsilon_i), \quad i = 1, \dots, n,$$

where $\varepsilon_i = [\nabla g(\bar{X}_n)(X_i - \bar{X}_n)]/(n\hat{\sigma}_n)$ and A_x is chosen to minimize the function $\Phi(x - A)e^{A^2}$ over A ; for any fixed p , the $\{g_i(p), i = 1, \dots, n\}$ minimizing $\text{var}_{\text{BI}}[(H_{\text{BI}}^{(B)})^{-1}(p)]$ is given by

$$g_i(p) = \exp(-A_p \varepsilon_i) / \sum_{i=1}^n \exp(-A_p \varepsilon_i), \quad i = 1, \dots, n,$$

where A_p minimizes $\Phi(z_p - A)e^{A^2}$ over A .

With these choices of $\{g_i, i = 1, \dots, n\}$, Do and Hall (1991) proved that

$$\frac{\text{var}_{\text{MC}}[H_{\text{BOOT}}^{(B)}(x)]}{\text{var}_{\text{BI}}[H_{\text{BI}}^{(B)}(x)]} \rightarrow_{a.s.} \rho(x) = \frac{[1 - \Phi(x)]\Phi(x)}{\Phi(x - A_x)e^{A_x^2} - [\Phi(x)]^2}.$$

A similar result holds for the importance bootstrap resampling approximation to the bootstrap quantile estimators. Note that $\rho(x)$ is strictly decreasing, with $\rho(-\infty) = \infty$, $\rho(0) = 1.7$ and $\rho(\infty) = 1$. Therefore, the importance bootstrap resampling approximation is considerably better than the simple Monte Carlo approximation if x is negative. Some values of $\rho(x)$ are tabulated by Hall (1992a) (see Table 5.4).

Thus, for approximating the bootstrap distribution estimator at $x > 0$ or the bootstrap quantile estimator with $p > \frac{1}{2}$, we can first apply importance bootstrap resampling on $-\mathfrak{R}_n$ and then transform the results back. A similar result for $(H_{\text{BI}}^{(B)})^{-1}(p)$ was obtained by Hall (1991a).

Table 5.4. The efficiency of importance bootstrap resampling
[Adapted from Hall (1992a), by permission of John Wiley & Sons]

$\Phi(x)$	0.005	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975
$\rho(x)$	69.0	38.0	17.6	10.0	5.8	1.7	1.1	1.1	1.0

In contrast with the previously introduced methods, importance bootstrap resampling is very effective for approximating the tails of the bootstrap distribution estimators. Do and Hall (1991) compared the simple Monte Carlo, balanced, antithetic, and importance bootstrap resampling approximations to the bootstrap distribution estimator of a studentized function of the sample mean. The distribution of X_1 is assumed to be symmetric. The results shown in Figure 5.1 indicate that at the tails (large $|x|$), antithetic and balanced bootstrap resampling have similar performances and both improve the simple Monte Carlo approximation slightly; antithetic bootstrap resampling is better than balanced bootstrap resampling for x near 0, but slightly worse for moderate $|x|$; importance bootstrap resampling is much better than all of the other methods for large or moderate $|x|$, although it is worse than antithetic and balanced bootstrap resampling at x near 0.

Importance bootstrap resampling, however, does not improve the simple Monte Carlo method for approximating the bootstrap bias and variance estimators, since Do and Hall (1991) showed that in these cases the optimal choice of $\{g_i, i = 1, \dots, n\}$ is $g_i = n^{-1}$ for all i . That is, the simple Monte Carlo approximations to the bootstrap bias and variance estimators cannot be improved by only changing the resampling plan from equal probability sampling to unequal probability sampling.

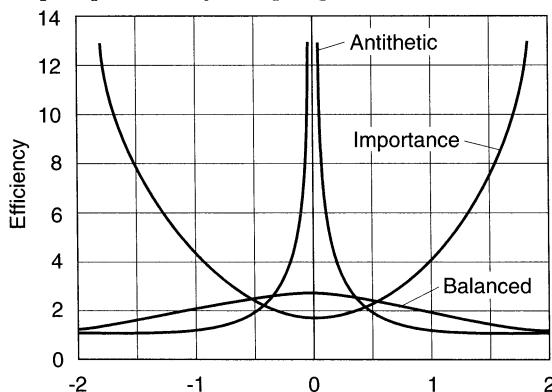


Figure 5.1. Efficiencies for balanced, antithetic, and importance bootstrap resampling, relative to the simple Monte Carlo approximation [From Do and Hall (1991), Copyright© (1984) by Biometrika, All rights reserved]

In conclusion, importance bootstrap resampling is very useful for approximating the tails of the bootstrap distribution estimators and for constructing bootstrap confidence sets. However, in the general case where $\hat{\theta}_n$ is an arbitrary statistic, the probabilities $\{g_i, i = 1, \dots, n\}$ may not be easy to obtain.

5.4.7 The one-step bootstrap

When the computation of the given statistic (estimator) requires some iterations, we can apply the *one-step bootstrap*, an analog of the one-step jackknife studied in Section 5.1.1. The original idea of the one-step bootstrap is due to Schucany and Wang (1991), but in the following we consider a more general situation.

Let P_n be the model that generates the data X_1, \dots, X_n and \hat{P}_n be an estimator of P_n based on X_1, \dots, X_n . Consider a statistic T_n defined by the iterative process

$$T_n = \lim_{k \rightarrow \infty} T_n^{[k]}, \quad T_n^{[k+1]} = g(T_n^{[k]}, \hat{P}_n), \quad k = 0, 1, 2, \dots$$

[see (5.2)], where $g(x, P_n)$ is a function of x for a fixed P_n and is a functional of P_n for a fixed x . An example is given in (5.4).

For simplicity, we focus on a univariate T_n . Under some continuity condition, T_n satisfies

$$T_n = g(T_n, \hat{P}_n),$$

i.e., T_n is a fixed point of $g(x, \hat{P}_n)$.

Let $\{X_1^*, \dots, X_n^*\}$ be a bootstrap sample generated from \hat{P}_n , and let \hat{P}_n^* be the bootstrap analog of \hat{P}_n . Then the bootstrap analog of T_n is the fixed point T_n^* satisfying

$$T_n^* = g(T_n^*, \hat{P}_n^*). \quad (5.46)$$

If the simple Monte Carlo method (or any other bootstrap resampling method) with sample size B is adopted to approximate the bootstrap variance or the distribution estimator for T_n , then the iterative process (5.46) has to be repeated B times.

To save computations, we may replace T_n^* by the one-step iteration

$$T_n^{*[1]} = g(T_n, \hat{P}_n^*), \quad (5.47)$$

where T_n is used as the initial point. The resulting bootstrap estimators are called one-step bootstrap estimators.

Suppose that $g(x, P_n)$ is differentiable for any fixed P_n . Let $g'(x, P_n) = \partial g(x, P_n) / \partial x$. Under weak conditions on $g'(x, P_n)$,

$$T_n^* = g(T_n^*, \hat{P}_n^*) = g(T_n, \hat{P}_n^*) + g'(T_n, \hat{P}_n^*)(T_n^* - T_n) + o_p(|T_n^* - T_n|),$$

and, therefore,

$$T_n^{*[1]} = T_n^* - g'(T_n, \hat{P}_n^*)(T_n^* - T_n) + o_p(|T_n^* - T_n|). \quad (5.48)$$

If

$$g'(T_n, \hat{P}_n^*) \rightarrow_p 0, \quad (5.49)$$

then, by (5.48),

$$T_n^{*[1]} - T_n^* = o_p(|T_n^* - T_n|), \quad (5.50)$$

which implies that the one-step bootstrap is asymptotically equivalent to the original bootstrap.

In many cases (e.g., T_n is an M-estimator),

$$g'(T_n, \hat{P}_n) = 0 \quad (\text{or } \rightarrow_p 0). \quad (5.51)$$

Then (5.49) holds under some continuity condition on g' , e.g.,

$$\lim_{|x-y| \rightarrow 0, \rho(P,Q) \rightarrow 0} |g'(x, P) - g'(y, Q)| = 0, \quad (5.52)$$

where $\rho(P, Q)$ is a suitably defined distance between P and Q .

When (5.51) and (5.52) hold, we can even replace T_n in (5.47) by some other initial point S_0 satisfying $T_n^* - S_0 = O_p(|T_n^* - T_n|)$ (e.g., S_0 equals the initial point used in calculating T_n or the one-step approximation to T_n). Similar to (5.48),

$$\begin{aligned} T_n^{*[1]} &= T_n^* - g'(S_0, \hat{P}_n^*)(T_n^* - S_0) + o_p(|T_n^* - S_0|) \\ &= T_n^* + o_p(|T_n^* - T_n|) \end{aligned}$$

under (5.51) and (5.52).

If (5.49) does not hold, then (5.50) is not true and the one-step approximation $T_n^{*[1]}$ is not close enough to T_n^* . Consequently, we need to make the following adjustment. It follows from (5.48) that

$$[1 - g'(T_n, \hat{P}_n^*)]^{-1}(T_n^{*[1]} - T_n) = T_n^* - T_n + o_p(|T_n^* - T_n|).$$

Hence, if we define an adjusted one-step iteration by

$$\tilde{T}_n^{*[1]} = T_n + [1 - g'(T_n, \hat{P}_n^*)]^{-1}(T_n^{*[1]} - T_n), \quad (5.53)$$

then

$$\tilde{T}_n^{*[1]} - T_n^* = o_p(|T_n^* - T_n|),$$

regardless of whether (5.49) holds or not. If (5.52) holds, then we can save some computations by replacing $g'(T_n, \hat{P}_n^*)$ in (5.53) with $g'(T_n, \hat{P}_n)$.

5.5 Conclusions and Discussions

- (1) We have introduced in this chapter many methods for computing the jackknife and bootstrap estimators. Most of these methods provide approximations to the original jackknife and bootstrap estimators and, therefore, have numerical approximation errors. Some methods are easy to apply but have large approximation errors; others are more sophisticated and have small approximation errors, but may not be easy (or impossible) to implement in some cases. We should note the difference between a numerical approximation error and a statistical error. The former is the error we commit when we approximate the jackknife or bootstrap estimator. The latter, however, refers to the error of the original jackknife or bootstrap estimator. No matter how small we can reduce the numerical approximation error, the statistical error cannot be reduced using any efficient computation method. The statistical error may be reduced by improving the original jackknife or bootstrap estimator, e.g., using the smoothed bootstrap (Chapter 3).
- (2) For the delete-1 jackknife, the computation is a concern only when n is very large or the given statistic is hard to evaluate. Techniques such as grouping, random subsampling, and the one-step jackknife are effective and easy to apply. For the delete-d jackknife, reducing computations is usually necessary. Random subsampling is easy to use, whereas balanced subsampling is more efficient but more difficult to implement.
- (3) There are a variety of methods for computing bootstrap estimators. They are either pure analytic methods or the simple Monte Carlo method with some modifications that require some theoretical work. The simple Monte Carlo method is the easiest to apply and is a general recipe for all problems, but it may require a very large or impractical Monte Carlo sample size B . Other methods work only for some problems under certain circumstances. For example, the delta method and the linear bootstrap can be used to approximate bootstrap moment estimators but not distribution and quantile estimators; the saddle point and the importance bootstrap resampling methods work for approximating bootstrap distribution and quantile estimators but not moment estimators; antithetic and balanced bootstrap resampling are effective for approximating bootstrap moment estimators and the center of the bootstrap distribution estimator, whereas importance bootstrap resampling is more suitable for constructing bootstrap confidence sets; and the one-step bootstrap is used only when the computation of the given statistic requires some iterations. One has to choose a computational method based on the nature of the problem, the computational resources, the accuracy requirement, and the degree of difficulty of the required theoretical derivations.

- (4) We may use a combination of two or several of the methods introduced in this chapter. For example, the linear bootstrap and the one-step bootstrap can be applied in conjunction with any efficient bootstrap resampling method introduced in this chapter. Hall (1990c) and Booth, Hall and Wood (1993) studied the performance of balanced importance bootstrap resampling. But Hall (1989b) showed that we cannot improve antithetic bootstrap resampling by combining it with any of the other bootstrap resampling methods. The performances of other combinations of the computational methods have not been investigated.
- (5) Some bootstrap computational methods introduced in this chapter are useful in cases where we need to apply the smoothed bootstrap (see Chapter 3), the nested bootstrap, bootstrap pre pivoting, and the iterative bootstrap (see Chapter 4). For example, importance bootstrap resampling was applied by Hinkley and Shi (1989) to reduce the computations for the nested or double bootstrap in constructing confidence intervals and by Chen and Do (1992) to compute the smoothed bootstrap estimators. DiCiccio, Martin and Young (1992a,b) used the saddle point approximation to approximate iterated bootstrap confidence intervals.
- (6) Since there are many ways to perform efficient Monte Carlo simulations in numerical analysis, it is of interest to study whether they can be applied to bootstrap computations. One example is the Richardson extrapolation method in Bickel and Yahav (1988). Also, since X_i^{*b} can be mapped to a point in $\{1, \dots, n\}^B$, some methods in experimental design may be useful to obtain the bootstrap samples in a more systematic manner than by simple random sampling.

Chapter 6

Applications to Sample Surveys

A crucial part of sample survey theory is the derivation of a suitable estimator of the variance of a given estimator. The variance estimator can be used in measuring the uncertainty in the estimation, in comparing the efficiencies of sampling designs, and in determining the allocation and stratification under a specific design. It is a common practice to report the estimates in a tabular form along with the variance estimates or estimates of coefficient of variation. Furthermore, consistent variance estimators provide us with large sample confidence sets for the unknown parameters of interest. Confidence sets can also be constructed by directly applying the bootstrap.

This chapter is devoted to studying applications of the jackknife, the bootstrap, and the balanced repeated replication methods in survey problems. After describing sampling designs and some notation in Section 6.1, we introduce in Section 6.2 the basic ideas, formulas, implementations, and properties of these methods, without going through mathematical details. Section 6.3 contains some numerical comparisons using results from empirical studies. Rigorous proofs of asymptotic properties of these methods are given in Section 6.4. In Section 6.5, we study applications of resampling methods to survey data with imputed missing values.

6.1 Sampling Designs and Estimates

An idealistic setup for a sample survey can be described as follows. A population \mathcal{P} consists of N_T distinct units identified through the labels $j = 1, \dots, N_T$. The characteristic of interest Y_j (possibly vector-valued)

associated with unit j can be known by observing unit j . The parameter of interest is a function of Y_j , $j = 1, \dots, N_T$, e.g., the population total $Y_T = Y_1 + \dots + Y_{N_T}$. A sample is a subset \mathbf{s} of \mathcal{P} (and the associated Y -values) selected according to a sampling plan that assigns a known probability $p(\mathbf{s})$ to \mathbf{s} such that $p(\mathbf{s}) \geq 0$ and $\sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) = 1$, where \mathcal{S} is the set of all possible \mathbf{s} . Advantages of taking a sample over a complete enumeration or census of the population include reduced cost and increased speed, scope, and quality of the study. Estimates of the parameters of interest based on the Y -values in the sample \mathbf{s} , however, have estimation errors, because only a fraction (usually small) of the population units are observed. Thus, a statistical analysis is required, which involves three steps: (i) choice of a sampling design [determination of $p(\mathbf{s})$]; (ii) choice of estimates of the parameters of interest; and (iii) construction of variance estimates and confidence sets. (i) and (ii) will be discussed briefly in this section and (iii) will be the main topic in this chapter.

There are essentially three different approaches to carrying out statistical analysis in a sample survey problem. The first is called the design-based approach, also called the probability sampling approach or the randomization approach, which is adopted throughout this chapter. This approach uses probability sampling both for sample selection and for inference from the data. Estimators based on the sample \mathbf{s} are considered to be random with respect to the selection of \mathbf{s} . The sampling distributions and properties of these estimators are with respect to repeated sampling from the population involving all samples $\mathbf{s} \in \mathcal{S}$ and their associated selection probabilities. The second approach is called the model-dependent approach. This approach involves purposive sampling, and the model distribution [generated from hypothetical realizations of (Y_1, \dots, Y_{N_T}) obeying the model] provides valid inferences referring to the particular sample \mathbf{s} that has been drawn. It is best illustrated under a simple regression model:

$$E_m(Y_j) = x'_j \beta, \quad j = 1, \dots, N_T,$$

where E_m is the model expectation. The third approach is a hybrid of the design-based and the model-dependent approaches, called the model-based approach or model-assisted approach. More discussions about these three approaches can be found in Rao and Bellhouse (1990).

The simplest sampling design is the single stage simple random sampling design, which uses equal probability of selecting \mathbf{s} , i.e.,

$$p(\mathbf{s}) = \frac{1}{\binom{N_T}{n_T}} \quad \text{for all } \mathbf{s} \in \mathcal{S},$$

where n_T is the sample size. This is equivalent to taking a simple random sample $\{Y_i, i \in \mathbf{s}\}$ without replacement from the population. Since the

observations in the sample are selected without replacement, they are not i.i.d. but close to i.i.d. if n_T/N_T is small.

Nowadays, one or a combination of several of the following sampling methods is often used in a sample survey: stratified sampling, cluster sampling, unequal probability sampling, and multistage sampling (Kish and Frankel, 1974; Krewski and Rao, 1981). Single stage simple random sampling is rarely used in practice because of both practical and theoretical considerations.

In stratified sampling, the population first is divided into nonoverlapping subpopulations called strata; then a sample is drawn from each stratum, independently across the strata. There are many reasons for stratification: (1) it may produce a gain in precision in the estimates of the parameters of interest when a heterogeneous population is divided into subpopulations, each of which is internally homogeneous; (2) sampling problems may differ markedly in different parts of the population; and (3) administrative considerations may also lead to stratification. More discussions can be found, for example, in Cochran (1977).

In cluster sampling, the sampling unit consists of a group, called a cluster, of smaller subunits. Cluster sampling is used often because of economic considerations. Also, although sometimes the first intention may be to use the subunits as sampling units, it is found that no reliable list of the subunits in the population is available. For example, in many countries, there are no complete lists of the people or houses in a region. From maps of the region, however, it can be divided into units such as cities or blocks in the cities.

Unequal probability sampling is used to improve the efficiency of survey estimates. For example, in cluster sampling, one may greatly increase the efficiency by using sampling with probability proportional to cluster size.

Suppose that a sample of clusters is obtained. If subunits within a selected cluster give similar results, then it may be uneconomical to measure them all. We may select a sample of the subunits in any chosen cluster. This technique is called two stage sampling. One can continue this process to have a multistage sampling (e.g., cities → blocks → houses → people). Of course, at each stage one may use stratified sampling and/or unequal probability sampling.

Many commonly used sampling designs can be unified into the following stratified multistage sampling design that will be considered throughout this chapter. The population under consideration has been stratified into L strata with N_h clusters in the h th stratum. For each h , $n_h \geq 2$ clusters are selected from stratum h , independently across the strata. These first stage clusters are selected by using either unequal probability sampling with replacement or simple random sampling (equal probability sampling)

without replacement. In practice, even when unequal probability sampling is applied, the clusters are often selected without replacement to avoid the selection of the same cluster more than once. Some methods for unequal probability sampling without replacement can be found in Rao, Hartley and Cochran (1962) and Rao (1979). However, at the stage of variance estimation it is a common practice to treat the sample as if the first stage clusters are drawn with replacement, since the calculations are then greatly simplified. This approximation leads to overestimation of variance, but the bias is small if the first stage sampling fractions $f_h = n_h/N_h$ are small. Therefore, for a concise presentation, in first stage sampling we consider sampling without replacement only when simple random sampling is used.

Within the (h, i) th first stage cluster, n_{hi} ultimate units are sampled according to some sampling methods, $i = 1, \dots, n_h$, $h = 1, \dots, L$. Note that we do not specify the number of stages and the sampling methods used after the first stage sampling. The total number of ultimate units in the sample is $n_T = \sum_{h=1}^L \sum_{i=1}^{n_h} n_{hi}$.

Associated with the j th ultimate unit in the i th cluster of stratum h is a vector of characteristics Y_{hij} , $j = 1, \dots, N_{hi}$, $i = 1, \dots, N_h$, $h = 1, \dots, L$, where N_{hi} is the number of ultimate units in the i th cluster of stratum h . The finite population distribution function is then given by

$$F(x) = \frac{1}{N_T} \sum_{h=1}^L \sum_{i=1}^{N_h} \sum_{j=1}^{N_{hi}} \delta_{Y_{hij}}(x), \quad (6.1)$$

where $N_T = \sum_{h=1}^L \sum_{i=1}^{N_h} N_{hi}$ is the total number of ultimate units in the population and δ_y is the distribution function degenerated at y .

Let $\{y_{hij}, j = 1, \dots, n_{hi}, i = 1, \dots, n_h, h = 1, \dots, L\}$ be the observations from the sampled elements and w_{hij} be the survey weight associated with y_{hij} . We assume that the survey weights are constructed so that

$$G(x) = \frac{1}{N_T} \sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} w_{hij} \delta_{y_{hij}}(x) \quad (6.2)$$

is unbiased for the population distribution $F(x)$ in (6.1), i.e.,

$$E[G(x)] = F(x) \quad \text{for any } x,$$

where E is with respect to repeated sampling from the population. The construction of the survey weights in some special cases will be illustrated later.

However, $G(x)$ may not be a distribution function since

$$G(\infty) = \frac{1}{N_T} \sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} w_{hij}$$

is not necessarily equal to 1. Furthermore, in some cases N_T is unknown. Thus, we replace N_T in (6.2) by its unbiased estimator

$$\hat{N}_T = \sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} w_{hij}$$

and obtain

$$\hat{F}(x) = \frac{1}{\hat{N}_T} \sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} w_{hij} \delta_{y_{hij}}(x). \quad (6.3)$$

Note that $\hat{F}(x)$ is a distribution function and is a ratio estimator of $F(x)$.

Let us study some special cases for illustration.

Example 6.1. Stratified one stage simple random sampling. When the sampling design is only one stage and simple random sampling is used within each stratum, $N_T = \sum_{h=1}^L N_h$ and the total number of sampled units is $n_T = \sum_{h=1}^L n_h$. In this case, N_h are known, the survey weights $w_{hi} = N_h/n_h$, and $\hat{F} = G$.

If $L = 1$, i.e., there is no stratification, then the design reduces to simple random sampling.

Example 6.2. Stratified two stage unequal probability sampling. Consider a stratified two stage sampling design. Suppose that in stratum h , n_h first stage clusters are selected from the N_h clusters with probabilities p_{hi} , $i = 1, \dots, n_h$, $\sum_{i=1}^{n_h} p_{hi} = 1$; for each (h, i) , n_{hi} units are selected from the N_{hi} units with probabilities p_{hij} , $j = 1, \dots, N_{hi}$, $\sum_{j=1}^{N_{hi}} p_{hij} = 1$. Then, $w_{hij} = (p_{hi} p_{hij} n_h n_{hi})^{-1}$ and

$$G(x) = \frac{1}{N_T} \sum_{h=1}^L \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{1}{p_{hi} n_{hi}} \sum_{j=1}^{n_{hi}} \frac{1}{p_{hij}} \delta_{y_{hij}}(x).$$

In general, $G(\infty) \neq 1$ and N_T is unknown.

A special case is $p_{hi} \equiv N_h^{-1}$ and $p_{hij} \equiv N_{hi}^{-1}$ (simple random sampling at both stages), where N_h and N_{hi} are known (and so is N_T).

In most survey problems, the parameter of interest can be written as

$$\theta = \theta(N_T, F)$$

with a known functional $\theta(\cdot, \cdot)$. Two important examples are (1) $\theta = g(N_T, Y_T)$, where $Y_T = N_T \int x dF(x)$ is the vector of population totals for Y_{hij} and g is a known differentiable function; and (2) $\theta = F^{-1}(p)$, a quantile of the population (for simplicity, when discussing quantiles we assume

that Y_{hij} is univariate), where p is known and $0 < p < 1$. Other examples can be found in Shao (1994b) and Shao and Rao (1994).

A survey estimator of θ is obtained by replacing N_T and F with \hat{N}_T and \hat{F} , respectively. In most cases, $\hat{\theta}$ can be written as

$$\hat{\theta} = T(\hat{Z}) \quad (6.4)$$

with a known functional T , where

$$\hat{Z} = \sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} w_{hij} z_{hij} \quad (6.5)$$

and z_{hij} is an appropriately defined vector of data from the (h, i, j) th ultimate sample unit. Two important cases are (1) $T = g$, $z_{hij} = (1, y'_{hij})'$, and $\hat{\theta} = g(\hat{Z})$ is the survey estimator of $\theta = g(N_T, Y_T)$ (e.g., ratios, correlation coefficients, and regression coefficients); and (2) $z_{hij} = (1, \delta_{y_{hij}})'$, $\hat{F}(x) = Z_2(x)/Z_1$, where Z_1 and Z_2 are the first and second components of \hat{Z} , and $\hat{\theta} = T(\hat{Z}) = \hat{F}^{-1}(p)$ is the survey estimator of $\theta = F^{-1}(p)$.

When $\hat{\theta} = g(\hat{Z})$, the linearization (Taylor's expansion) method produces the following variance estimator (Rao, 1988; also see Section 2.1.3):

$$v_L = \sum_{h=1}^L \frac{1 - \lambda_h f_h}{n_h} \nabla g(\hat{Z})' s_h^2 \nabla g(\hat{Z}), \quad (6.6)$$

where

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (z_{hi} - \bar{z}_h)(z_{hi} - \bar{z}_h)',$$

$$z_{hi} = \sum_{j=1}^{n_{hi}} n_h w_{hij} z_{hij}, \quad \bar{z}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} z_{hi}, \quad (6.7)$$

$f_h = n_h/N_h$, and λ_h indicates whether the first stage sampling is with or without replacement, i.e.,

$$\lambda_h = \begin{cases} 1 & \text{if the first stage sampling is without replacement} \\ 0 & \text{if the first stage sampling is with replacement.} \end{cases} \quad (6.8)$$

In the special case where the function g is linear, v_L is an unbiased estimator of $\text{var}(\hat{\theta})$. In general, v_L is a consistent estimator of the asymptotic variance of $\hat{\theta} = g(\hat{Z})$ (Krewski and Rao, 1981; Bickel and Freedman, 1984).

The linearization method requires a separate derivation of the derivatives for each g . The data-resampling methods introduced in the next section employ a unified formula for estimating the variances of all nonlinear estimators and require no theoretical derivation.

We now turn to the situation where θ is a population quantile. In this case, linearization is not applicable for deriving a variance estimator for the sample quantile $\hat{\theta}$. Woodruff (1952) provided the following large sample confidence interval for θ with level $1 - 2\alpha$:

$$[\hat{F}^{-1}(p - z_{1-\alpha}\hat{\sigma}(\hat{\theta})), \hat{F}^{-1}(p + z_{1-\alpha}\hat{\sigma}(\hat{\theta}))], \quad (6.9)$$

where $z_{1-\alpha}$ is the $(1-\alpha)$ th quantile of the standard normal distribution and for each fixed x , $\hat{\sigma}^2(x)$ is the variance estimator in (6.6) with $g(x_1, x_2) = x_2/x_1$ and $z_{hij} = (1, \delta_{y_{hij}})'$ that is evaluated at x . By equating Woodruff's interval (6.9) to a normal theory interval, Rao and Wu (1987) obtained the following variance estimator for $\hat{\theta}$:

$$v_w = \left[\frac{\hat{F}^{-1}(p + z_{1-\alpha}\hat{\sigma}(\hat{\theta})) - \hat{F}^{-1}(p - z_{1-\alpha}\hat{\sigma}(\hat{\theta}))}{2z_{1-\alpha}} \right]^2. \quad (6.10)$$

However, the best choice of α is unknown, although $\alpha = 0.05$ is suggested in terms of empirical evidences (Kovar, Rao and Wu, 1988; Rao, Wu and Yue, 1992; Sitter, 1992a,b). The consistency of v_w has been established by Francisco and Fuller (1991) and Shao (1994b).

6.2 Resampling Methods

The most popular data-resampling methods used in sample surveys are the jackknife, the balanced repeated replication, and the bootstrap. Here we present a detailed introduction of how these three methods are applied in survey problems. Theoretical justifications are provided in Section 6.4.

Due to the complexity of the sampling designs, some modifications and adjustments of the jackknife and bootstrap methods are often required in order to apply them to the complex survey data.

6.2.1 The jackknife

To extend the jackknife variance estimator (1.13) to survey problems in which multistage sampling is used, we first need to decide what is the primary unit to delete. In general, it is not a good idea to delete one ultimate unit at a time, since y_{hij} may be dependent for fixed h and i and the dependence structure may not be easily modeled. The jackknife has to be modified in order to apply it to dependent data (see the discussions in Chapter 9).

When the first stage sampling is with replacement or simple random sampling without replacement, the first stage sample clusters are independent or nearly independent. Hence, we may delete one first stage sample

cluster at a time. The total number of first stage sample clusters

$$n = \sum_{h=1}^L n_h \quad (6.11)$$

is large and, therefore, deleting clusters does not lose much efficiency. Furthermore, deleting clusters saves a large number of computations.

For fixed integers $h' \leq L$ and $i' \leq n_{h'}$, let

$$\hat{Z}_{h'i'} = \sum_{h \neq h'} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} w_{hij} z_{hij} + \frac{n_{h'}}{n_{h'} - 1} \sum_{i \neq i'} \sum_{j=1}^{n_{h'i}} w_{h'i} z_{h'i}$$

be the analog of \hat{Z} after the i' th cluster in stratum h' is deleted,

$$\hat{\theta}_{h'i'} = T(\hat{Z}_{h'i'}) \quad \text{and} \quad \hat{\theta}_{h'} = \frac{1}{n_{h'}} \sum_{l=1}^{n_{h'}} \hat{\theta}_{h'l}.$$

A jackknife variance estimator for $\hat{\theta}$ is given by

$$v_{\text{JACK}} = \sum_{h=1}^L \frac{(1 - \lambda_h f_h)(n_h - 1)}{n_h} \sum_{i=1}^{n_h} (\hat{\theta}_{hi} - \hat{\theta}_h)^2. \quad (6.12)$$

In (6.12), $1 - \lambda_h f_h$ [λ_h is defined in (6.8)] is used to account for the effect of sampling without replacement. When $\hat{\theta}$ is a sample quantile, v_{JACK} may be inconsistent (Section 2.3); however, in complex survey problems, v_{JACK} may still perform well (Rao, Wu and Yue, 1992). But in the following discussion about the jackknife, we always focus on the case of $\hat{\theta} = g(\hat{Z})$, a function of weighted averages.

There are at least two different variations of the estimator in (6.12). We may replace $\hat{\theta}_h$ in (6.12) by $n^{-1} \sum_{h=1}^L \sum_{i=1}^{n_h} \hat{\theta}_{hi}$ or by $L^{-1} \sum_{h=1}^L \hat{\theta}_h$. However, these changes have little effect on the performance of the jackknife estimator. Rao and Wu (1985) showed that these jackknife estimators are second order asymptotically equivalent, i.e., they are equal up to the order of $O_p(n^{-2})$, which is supported by the empirical results in Kovar, Rao and Wu (1988).

A special but important case is when $n_h = 2$ and $\lambda_h = 0$ for all h . The jackknife estimator v_{JACK} reduces to the estimator

$$v_{\text{JRR-D}} = \frac{1}{4} \sum_{h=1}^L (\hat{\theta}_{h1} - \hat{\theta}_{h2})^2 \quad (6.13)$$

defined in Kish and Frankel (1974).

When the function g is linear, i.e., $g(x) = c'x$ for a fixed $c \neq 0$, all the jackknife estimators are the same and

$$v_{\text{JACK}} = v_L. \quad (6.14)$$

Example 6.3. Quadratic functions. Rao and Wu (1985) showed that (6.14) holds in another special case where $n_h = 2$ for all h and g is quadratic:

$$g(t) = g(s) + (t - s)' \nabla g(s) + \frac{1}{2}(t - s)' \nabla^2 g(s)(t - s) \quad \text{for all } s, t \quad (6.15)$$

(also, see Efron, 1982, Chapter 8). This can be justified as follows. Note that

$$\hat{Z}_{hi} = \hat{Z} + (n_h - 1)^{-1}(\bar{z}_h - z_{hi}), \quad (6.16)$$

where z_{hi} and \bar{z}_h are as given in (6.7). Since $n_h = 2$, it follows from (6.16) that

$$\hat{Z}_{h1} - \hat{Z} = -(\hat{Z}_{h2} - \hat{Z}).$$

Using (6.15) with $t = \hat{Z}_{hi}$ and $s = \hat{Z}$, we obtain that

$$\frac{1}{2}(\hat{\theta}_{h1} - \hat{\theta}_{h2}) = (\hat{Z}_{h1} - \hat{Z})' \nabla g(\hat{Z}).$$

Then, by (6.12), (6.16), and the fact that $n_h = 2$,

$$v_{\text{JACK}} = \frac{1}{4} \sum_{h=1}^L (1 - \lambda_h f_h)[(z_{h1} - z_{h2})' \nabla g(\hat{Z})]^2,$$

which is exactly the same as v_L in (6.6) with $n_h = 2$.

In general, there is also a close relationship between the linearization variance estimator and the jackknife variance estimator. In fact, Rao and Wu (1985) proved that

$$v_{\text{JACK}}/v_L = 1 + O_p(n^{-1}) \quad (6.17)$$

and

$$v_{\text{JACK}}/v_L = 1 + O_p(n^{-2}) \quad \text{if } n_h = 2 \text{ for all } h. \quad (6.18)$$

Asymptotic properties of v_{JACK} will be studied in Section 6.4.

Recall that the jackknife can also be used for bias estimation and bias reduction. For $\hat{\theta} = g(\hat{Z})$, a correct jackknife bias estimator is

$$b_{\text{JACK}} = \sum_{h=1}^L (n_h - 1)(\hat{\theta}_h - \hat{\theta})$$

(Rao and Wu, 1985). This bias estimator is asymptotically unbiased and consistent. The use of the jackknife estimator $\hat{\theta} - b_{\text{JACK}}$ eliminates the bias of the order n^{-1} . A more detailed discussion can be found in Rao and Wu (1985).

6.2.2 The balanced repeated replication

The balanced repeated replication (BRR) was first proposed by McCarthy (1969) for the case where $n_h = 2$ clusters per stratum are sampled with replacement in the first stage sampling. We start with this special case.

Two clusters per stratum sampled with replacement

A set of R half-samples is formed in a balanced manner, where each half-sample is obtained by selecting one first stage sample cluster from each stratum. This set may be defined by an $R \times L$ matrix with the (r, h) th element $\epsilon_{rh} = +1$ or -1 indicating whether the cluster of the h th stratum in the r th half-sample is the first or the second first stage sample cluster, $r = 1, \dots, R$, $h = 1, \dots, L$, where $\sum_{r=1}^R \epsilon_{rh} = 0$ for all h and $\sum_{r=1}^R \epsilon_{rh} \epsilon_{rh'} = 0$ for all $h \neq h'$, i.e., the columns of the matrix are orthogonal. A minimal set of R balanced half-samples may be constructed from an $R \times R$ Hadamard matrix by choosing any L columns excluding the column of all $+1$'s, where $L \leq R \leq L + 3$. Let $\hat{\theta}^{(r)}$ be the estimator of θ based on the r th half-sample and computed using formulas similar to (6.4)–(6.5), i.e., $\hat{\theta}^{(r)} = T(\hat{Z}^{(r)})$ with

$$\hat{Z}^{(r)} = \sum_{h=1}^L \frac{1}{2} \sum_{i=1}^2 (\epsilon_{rh} + 1) z_{hi},$$

where z_{hi} is as given in (6.7). A BRR variance estimator for $\hat{\theta}$ is given by

$$v_{\text{BRR}} = \frac{1}{R} \sum_{r=1}^R \left(\hat{\theta}^{(r)} - \frac{1}{R} \sum_{r=1}^R \hat{\theta}^{(r)} \right)^2. \quad (6.19)$$

There are several variations of (6.19). For example, $R^{-1} \sum_{r=1}^R \hat{\theta}^{(r)}$ in (6.19) can be replaced by $\hat{\theta}$; and one may use

$$v_{\text{BRR-D}} = \frac{1}{4R} \sum_{r=1}^R (\hat{\theta}^{(r)} - \hat{\theta}_c^{(r)})^2, \quad (6.20)$$

where $\hat{\theta}_c^{(r)}$ is the estimator based on the complement of the r th half-sample used in computing $\hat{\theta}^{(r)}$.

In the case where $\hat{\theta} = g(\hat{Z})$ and g is linear, all of the BRR variance estimators are equal to $v_L = v_{\text{JACK}}$ because of the balanced nature of the half-samples. When g is quadratic in the sense of (6.15),

$$v_{\text{BRR-D}} = v_{\text{JACK}} = v_L$$

(Rao and Wu, 1985). For a general nonlinear function g , Rao and Wu (1985) showed that

$$v_{\text{BRR}}/v_L = 1 + O_p(n^{-1/2})$$

and

$$v_{\text{BRR-D}}/v_L = 1 + O_p(n^{-1}).$$

Comparing these results with (6.18), we find that the BRR variance estimators are not so close to v_L as the jackknife variance estimators are.

General cases

The BRR method has been extended to the general case of $n_h \geq 2$ clusters per stratum. Suppose that for the h th stratum and r th replicate, a set $\mathbf{s}_{rh} \subset \{1, \dots, n_h\}$ of m_h integers is selected. When L is large and all n_h are small, a simple and effective choice of m_h is $m_h = 1$ for all h . The set $\{\mathbf{s}_{rh} : r = 1, \dots, R, h = 1, \dots, L\}$ constitutes a BRR if for fixed h and h' , the number of elements in $\{r : i \in \mathbf{s}_{rh}, i' \in \mathbf{s}_{rh}\}$ does not depend on i and i' ($i \neq i'$); and the number of elements in $\{r : i \in \mathbf{s}_{rh}, i' \in \mathbf{s}_{rh'}\}$ does not depend on i and i' . A trivial BRR is

$$\{\text{all possible subsets of } \{1, \dots, n_h\} \text{ of size } m_h, h = 1, \dots, L\},$$

in which case $R = \prod_{h=1}^L \binom{n_h}{m_h}$.

If we define the BRR variance estimator using (6.19) with $\hat{\theta}^{(r)}$ being the estimator of θ based on the data in the r th replicate, i.e., $\hat{\theta}^{(r)} = T(\hat{Z}^{(r)})$, where

$$\hat{Z}^{(r)} = \sum_{h=1}^L \bar{z}_h^{(r)} \quad \text{and} \quad \bar{z}_h^{(r)} = \frac{1}{m_h} \sum_{i \in \mathbf{s}_{rh}} z_{hi},$$

then, in the case of $\hat{\theta} = g(\hat{Z})$ with linear $g(x) = c'x$ and $\lambda_h = 0$, the BRR variance estimator is

$$\frac{1}{R} \sum_{r=1}^R \left(\hat{\theta}^{(r)} - \frac{1}{R} \sum_{r=1}^R \hat{\theta}^{(r)} \right)^2 = \sum_{h=1}^L \frac{n_h - m_h}{m_h n_h} c' s_h^2 c$$

(by the balanced nature of the replicates), which does not agree with the unbiased and consistent variance estimator v_L in (6.6), unless $m_h = n_h/2$ for all h . Hence, some modification has to be made. Wu (1991) considered the following rescaling adjustment, which was originally introduced in Rao and Wu (1988) for modifying the bootstrap method:

$$\tilde{Z}^{(r)} = \sum_{h=1}^L \left[\sqrt{\frac{(1-\lambda_h f_h)m_h}{n_h - m_h}} \bar{z}_h^{(r)} + \left(1 - \sqrt{\frac{(1-\lambda_h f_h)m_h}{n_h - m_h}} \right) \bar{z}_h \right]. \quad (6.21)$$

An adjusted BRR variance estimator is then defined by

$$v_{\text{BRR}} = \frac{1}{R} \sum_{r=1}^R \left(\tilde{\theta}^{(r)} - \frac{1}{R} \sum_{r=1}^R \tilde{\theta}^{(r)} \right)^2, \quad (6.22)$$

where $\tilde{\theta}^{(r)} = T(\tilde{Z}^{(r)})$, $r = 1, \dots, R$. The adjustment in (6.21) ensures that in the case of $\hat{\theta} = c' \hat{Z}$, v_{BRR} in (6.22) reduces to v_L .

For the case of $\hat{\theta} = \hat{F}^{-1}(p)$, the coefficients in front of $\bar{z}_h^{(r)}$ and \bar{z}_h in formula (6.21) must be nonnegative in order for $\tilde{F}^{(r)}$ (the adjusted estimator of the population distribution function) to be a proper distribution function. This requires that

$$m_h \leq n_h/2 \quad \text{for all } h. \quad (6.23)$$

It is easy to see that the estimator in (6.19) is a special case of that in (6.22). In fact, as long as $m_h = n_h/2$, $\hat{Z}^{(r)}$ and $\hat{F}^{(r)}$ are the same as the adjusted $\tilde{Z}^{(r)}$ and $\tilde{F}^{(r)}$, respectively.

A convenient way of computing $\tilde{\theta}^{(r)}$ is to use the formula for the original estimator $\hat{\theta}$ with the weights w_{hij} changing to

$$w_{hij}^{(r)} = \left[1 + \sqrt{\frac{(1 - \lambda_h f_h)(n_h - m_h)}{m_h}} \right] w_{hij} \quad \text{or} \quad = \left[1 - \sqrt{\frac{(1 - \lambda_h f_h)m_h}{n_h - m_h}} \right] w_{hij}$$

depending on whether or not the (h, i) th sample cluster is selected in the r th replicate. Condition (6.23) ensures that the new weights $w_{hij}^{(r)} \geq 0$. When $n_h = 2$, $\lambda_h = 0$ and $m_h = 1$ for all h , $w_{hij}^{(r)} = 2w_{hij}$ or 0.

Asymptotic properties of the BRR variance estimators will be studied in Section 6.4. For $\hat{\theta} = g(\hat{Z})$, both the BRR and jackknife provide consistent variance estimators. For sample quantiles, however, the jackknife is not applicable, whereas the BRR variance estimators are still consistent under some weak conditions. This is a great advantage of the BRR over the jackknife, especially when one prefers to use a single method for both cases of $\hat{\theta} = g(\hat{Z})$ and $\hat{\theta} = \hat{F}^{-1}(p)$.

To compute the BRR variance estimator, it is desirable to find a BRR with R as small as possible. In the general case of $n_h \geq 2$, however, the construction of a BRR with a feasible R is much more difficult than in the case of $n_h = 2$ per stratum, and the smallest possible R may also be much larger than $L + 4$. In the case of $n_h = p > 2$ clusters per stratum for p prime or power of prime, a BRR can be obtained by using orthogonal arrays of strength two (Gurney and Jewett, 1975), where each balanced sample is obtained by selecting one first stage sample cluster from each stratum. Gupta and Nigam (1987) and Wu (1991) obtained BRR with $m_h = 1$ in the case of unequal n_h using mixed level orthogonal arrays of strength two to construct balanced replicates. More methods for constructing BRR can be found in Sitter (1993).

Example 6.4. Construction of a BRR. Some knowledge of experimental design is helpful in the construction of a BRR. Let $A(R, n_1 \times \dots \times n_L)$ be an $R \times L$ matrix whose h th column has n_h symbols such that for any two

columns each possible combination of symbols appears equally often. Such a matrix is called a mixed orthogonal array of strength two. The following is an example of $A(12, 3 \times 2 \times 2 \times 2 \times 2)$:

$$\left(\begin{array}{ccccc} a & a & a & a & a \\ a & a & b & a & b \\ a & b & a & b & b \\ a & b & b & b & a \\ b & a & a & b & b \\ b & a & b & b & a \\ b & b & a & a & b \\ b & b & b & a & a \\ c & a & a & b & a \\ c & a & b & a & b \\ c & b & a & a & a \\ c & b & b & b & b \end{array} \right) . \quad (6.24)$$

If we let n_h symbols represent the n_h first stage sample clusters in stratum h , then we can form a BRR with $m_h = 1$ whose r th balanced replicate containing one cluster from each stratum corresponds to the (r, h) th entry of the array. The array in (6.24) can be used in the situation where $L = 5$, $n_1 = 3$, $n_h = 2$, $h = 2, 3, 4, 5$, and $R = 12$.

6.2.3 Approximated BRR methods

Despite the existence of several elegant methods of forming balanced replicates, the application of the BRR method is not easy when the n_h are unequal. The enumeration of the balanced replicates may require separate software or involve nontrivial mathematical developments. Furthermore, a feasible BRR may not always exist for arbitrary n_h .

One easy but inefficient alternative is to randomly divide the n_h clusters in stratum h into two groups containing $[n_h/2]$ and $n_h - [n_h/2]$ clusters, respectively, and then apply the BRR method by treating the two groups as two “clusters” (Kish and Frankel, 1970; Wolter, 1985; Valliant, 1987). This method, called the *grouped BRR* (GBRR), requires R repeated computations of the point estimator with $L \leq R \leq L + 3$. When L is large, it provides a consistent variance estimator, although the efficiency of the GBRR variance estimator can be quite low if L is much smaller than n , the total number of first stage sampled clusters (see, e.g., Krewski, 1978). When L is small and n_h are large, however, the GBRR variance estimator is inconsistent (Rao and Shao, 1995).

To increase the efficiency of the GBRR variance estimator, one may independently repeat the grouping R_G times and take the average of the R_G GBRR variance estimators (Rao and Shao, 1995). The resulting variance estimator is called the *repeatedly grouped BRR* (RGBRR). It requires $R_G R$ repeated computations of the point estimator. Rao and Shao (1995) showed that the RGBRR variance estimator is consistent as long as $R_G R \rightarrow \infty$;

the RGBRR variance estimator is quite efficient if R_G and R are chosen so that $R_G R$ is comparable with n . Since R has the same order as L , one must choose a large R_G if L is small. On the other hand, one may simply set $R_G = 1$ when L is large.

Another method is to randomly divide the n_h clusters in stratum h into $n_h/2$ groups and then apply the BRR method by treating the constructed $n/2$ groups as new strata having two clusters, where n_h is assumed even for simplicity and n is the number of first stage sample clusters (Wolter, 1985). The variance estimator obtained using this approximate BRR is consistent as long as $n \rightarrow \infty$ (Rao and Shao, 1995). However, the application of this method may not be easy when n is much larger than L , since it may be difficult to construct a Hadamard matrix with a very large size.

Shao and Wu (1992) introduced a random subsampling method that includes the method in Section 5.2.2 as a special case. Let m_h be fixed, \mathbf{s}_h denote a subset of $\{1, \dots, n_h\}$ of size m_h , and \mathbf{S} be the collection of all possible elements of the form $(\mathbf{s}_1, \dots, \mathbf{s}_L)$. Suppose that

$$\mathbf{S} = \mathbf{S}_1 \cup \mathbf{S}_2 \cup \dots \cup \mathbf{S}_U,$$

where the \mathbf{S}_j are disjoint subsets of \mathbf{S} of ℓ elements. Let $\{\mathbf{S}_j^*, j = 1, \dots, u\}$ be a simple random sample of size u (with or without replacement) from $\{\mathbf{S}_j, j = 1, \dots, U\}$ (u is usually much smaller than U). Let $(\mathbf{s}_{r1}, \dots, \mathbf{s}_{rL})$ be the r th element in the collection

$$\mathbf{S}_1^* \cup \mathbf{S}_2^* \cup \dots \cup \mathbf{S}_u^*,$$

$r = 1, \dots, R = u\ell$. Then the *random subsampling BRR* (RSBRR) variance estimator can still be defined by (6.22) with randomly selected \mathbf{s}_{rh} instead of balanced \mathbf{s}_{rh} . If $\ell = 1$, then this method amounts to taking a simple random sample of size $R = u$ from \mathbf{S} and is the stratified version of the method discussed in Section 5.2.2. Another special case for $n_h = n_0$ and $m_h = 1$ for all h is to define \mathbf{S}_j to be a collection of n_0 mutually exclusive subsamples, each of which contains L first stage sampled clusters with one from each stratum. Then each \mathbf{S}_j amounts to grouping the $n_0 L$ clusters into n_0 exclusive subsamples; there are n_0^{L-1} such groupings to make up \mathbf{S} with $U = n_0^{L-1}$. The RSBRR in this special case is also called the *repeated random group* (RRG) method and is studied in Kovar, Rao and Wu (1988).

In large scale surveys, the total number of first stage sample clusters n can be very large, and, therefore, the computation of the jackknife estimator is cumbersome. Techniques such as grouping and random subsampling introduced in Section 5.1.2 can be applied to circumvent the computational problems. In most cases, grouping or random subsampling should be applied within each stratum.

The grouped jackknife with two groups in each stratum is very similar

to the GBRR method. In fact, the GBRR or RGBRR variance estimator can be used to approximate the jackknife variance estimator.

Asymptotic properties of these approximated BRR and jackknife estimators will be discussed in Section 6.4. Some empirical results on the performance of the RGBRR and the RRG can be found in Section 6.3.

6.2.4 The bootstrap

Recall that the bootstrap can be applied to estimating variance and, more importantly, constructing confidence sets. Here we describe a straightforward extension of the standard bootstrap and some modified bootstrap methods.

The naive bootstrap

A straightforward extension of the bootstrap to survey problems is to apply the standard bootstrap in each stratum. That is, draw a simple random sample $\{\mathbf{y}_{hi}^*, i = 1, \dots, n_h\}$ with replacement from the original sample $\{\mathbf{y}_{hi}, i = 1, \dots, n_h\}$, $h = 1, \dots, L$, independently across the strata, where $\mathbf{y}_{hi} = (y_{h1}, \dots, y_{hn_h})$; calculate $\hat{\theta}^* = T(\hat{Z}^*)$, where

$$\hat{Z}^* = \sum_{h=1}^L \frac{1}{n_h} \sum_{i=1}^{n_h} z_{hi}^*, \quad z_{hi}^* = \sum_{j=1}^{n_{hi}} n_h w_{hij}^* z_{hij}^*, \quad (6.25)$$

z_{hij}^* is the bootstrap analog of z_{hij} based on the bootstrap data y_{hij}^* , the j th component of \mathbf{y}_{hi}^* , and w_{hij}^* is the survey weight associated with y_{hij}^* ; and then obtain bootstrap estimators in the same way as was described in the previous chapters. For example, the bootstrap estimator of the distribution function of $\sqrt{n}(\hat{\theta} - \theta)$ is

$$H_{\text{BOOT}}(x) = P_*\{\sqrt{n}(\hat{\theta}^* - \hat{\theta}) \leq x\}, \quad (6.26)$$

and the bootstrap estimator of the variance of $\hat{\theta}$ is

$$v_{\text{BOOT}} = E_* (\hat{\theta}^* - E_* \hat{\theta}^*)^2 = \text{var}_*(\hat{\theta}^*), \quad (6.27)$$

where P_* , E_* , and var_* refer to the probability, expectation, and variance, respectively, with respect to the bootstrap sampling. If the estimator in (6.26) or (6.27) does not have an explicit form, then a numerical approximation such as the Monte Carlo can be used.

To investigate properties of the bootstrap estimators, we consider the simple case where $\hat{\theta} = c' \hat{Z}$ with a fixed vector c . By (6.25) and (6.27),

$$v_{\text{BOOT}} = \text{var}_*(c' \hat{Z}^*) = \sum_{h=1}^L \frac{1}{n_h} \left(\frac{n_h - 1}{n_h} \right) c' s_h^2 c, \quad (6.28)$$

since

$$\text{var}_*(z_{hi}^*) = \frac{1}{n_h} \sum_{i=1}^{n_h} (z_{hi} - \bar{z}_h)(z_{hi} - \bar{z}_h)' = \left(\frac{n_h - 1}{n_h} \right) s_h^2$$

for any i and h . Comparing result (6.28) with formula (6.6), we find that v_{BOOT} has two potential defects: (1) the lack of the scale factors $1 - \lambda_h f_h$, which may cause some problems when the original first stage clusters are sampled without replacement; and (2) the existence of the superfluous scale factors $\frac{n_h - 1}{n_h}$ in (6.28), which results in a biased and inconsistent variance estimator in the case where n_h are bounded.

Because of result (6.28), the bootstrap distribution estimator H_{BOOT} in (6.26) is inconsistent when some n_h are bounded.

Note that in the i.i.d. case (corresponding to the case of $L = 1$ and simple random sampling with replacement), the scale factor $\frac{n_h - 1}{n_h} = \frac{n - 1}{n}$ does not have an appreciable effect when the sample size n is large. But in survey problems, a large number of n_h may be small. For instance, in the case of two clusters per stratum, $\frac{n_h - 1}{n_h} = \frac{1}{2}$ for all h and, therefore, the relative bias of v_{BOOT} is -50% .

Although v_{BOOT} and H_{BOOT} may be inconsistent, Bickel and Freedman (1984) showed that in the case of $\lambda_h = 0$ for all h , the bootstrap distribution of $c'(\hat{Z}^* - \hat{Z}) / (\sum_{h=1}^L c' s_h^{*2} c/n_h)^{1/2}$, where s_h^{*2} is the bootstrap analog of s_h^2 , is a consistent estimator of the distribution of the studentized statistic $c'(\hat{Z} - E\hat{Z}) / (\sum_{h=1}^L c' s_h^2 c/n_h)^{1/2}$, which provides consistent bootstrap-t confidence sets. However, unlike the i.i.d. case discussed in Chapter 4, these bootstrap-t confidence sets are not second order accurate, because of the scaling problem when some n_h are bounded.

Modified bootstrap methods

If $n_h = n_0$ and $\lambda_h f_h = \lambda_0 f_0$ for all h , then these problems can be avoided by replacing $\hat{\theta}^* - \hat{\theta}$ with $(\frac{n_0}{n_0 - 1})(1 - \lambda_0 f_0)(\hat{\theta}^* - \hat{\theta})$. In general, however, some nontrivial modification has to be made in order to obtain correct bootstrap estimators. Four modified bootstrap methods are introduced next.

(I) The with-replacement bootstrap

McCarthy and Snowden (1985) proposed taking a simple random sample of size m_h , instead of n_h , with replacement from $\{\mathbf{y}_{hi}, i = 1, \dots, n_h\}$, $h = 1, \dots, L$, independently across the strata, where

$$m_h = \frac{n_h - 1}{1 - \lambda_h f_h}. \quad (6.29)$$

The bootstrap estimators are obtained the same way as before, except that \hat{Z}^* is replaced by $\sum_{h=1}^L m_h^{-1} \sum_{i=1}^{m_h} z_{hi}^*$. McCarthy and Snowden (1985) called this method the *with-replacement bootstrap* (BWR).

When $\hat{\theta} = c' \hat{Z}$, the variance estimator obtained from the BWR is, by (6.29),

$$\sum_{h=1}^L \frac{1}{m_h} \left(\frac{n_h - 1}{n_h} \right) c' s_h^2 c = \sum_{h=1}^L \frac{1 - \lambda_h f_h}{n_h} c' s_h^2 c,$$

which is the same as the unbiased and consistent estimator v_L in (6.6). One can also verify that the distribution estimator obtained from the BWR is consistent.

If the original first stage sampling is with replacement, then $m_h = n_h - 1$ for the BWR, which was also suggested by Efron (1982). If m_h in (6.29) is noninteger, then a randomization between the bracketing integers has to be used (and will be discussed later), which results in some efficiency loss.

(II) The rescaling bootstrap

Rao and Wu (1988) proposed a rescaling of the original bootstrap. In this method, the bootstrap sample is selected by using the same scheme as that in the BWR method, except that m_h is not necessarily given by (6.29). After the bootstrap sample is obtained, one calculates $\hat{\theta}^* = T(\tilde{Z}^*)$ with

$$\tilde{Z}^* = \sum_{h=1}^L \left[\sqrt{\frac{(1 - \lambda_h f_h)m_h}{n_h - 1}} \tilde{z}_h^* + \left(1 - \sqrt{\frac{(1 - \lambda_h f_h)m_h}{n_h - 1}} \right) \bar{z}_h \right],$$

where $\tilde{z}_h^* = m_h^{-1} \sum_{i=1}^{m_h} z_{hi}^*$. The bootstrap estimators are obtained the same way as before but with $\hat{\theta}^*$ replaced by $\hat{\theta}^*$. This method is called the *rescaling bootstrap* (BRS).

In the case where $\hat{\theta} = c' \hat{Z}$, for any integers $m_h \geq 1$, the BRS variance estimator

$$\tilde{v}_{\text{BOOT}} = \sum_{h=1}^L \frac{(1 - \lambda_h f_h)m_h}{n_h - 1} c' \text{var}_*(\tilde{z}_h^*) c = v_L.$$

Rao and Wu (1988) also heuristically showed that when $\hat{\theta} = g(\hat{Z})$ with a nonlinear g ,

$$\tilde{v}_{\text{BOOT}}/v_L = 1 + O_p(n^{-1}).$$

Furthermore, the BRS distribution estimators are consistent.

For the choice of m_h , Rao and Wu (1988) suggested

$$m_h \approx \frac{(1 - \lambda_h f_h)(n_h - 2)^2}{(1 - 2\lambda_h f_h)^2(n_h - 1)}$$

(assuming $n_h > 2$), which is obtained by matching the third order moments of $\hat{Z} - E(\hat{Z})$ and $\tilde{Z}^* - E_*(\tilde{Z}^*)$. This choice of m_h has some impact on constructing confidence sets when $\hat{\theta} = g(\hat{Z})$; for example, the bootstrap-t confidence sets are second order accurate (Rao and Wu, 1988).

In the case where $\hat{\theta}$ is a sample quantile, however, $m_h \leq n_h - 1$ is required in order to have a proper distribution function \tilde{F}^* (the BRS analog of \hat{F}).

If $\lambda_h = 0$ (sampling with replacement) and we choose $m_h = n_h - 1$ for all h , then the BRS is the same as the BWR.

When $n_h = 2$ and $m_h = 1$ for all h , the BRS variance estimator is the same as the BRR variance estimator corresponding to the largest BRR scheme; and any BRR variance estimator can be viewed as an approximation to the bootstrap variance estimator.

(III) The mirror-match bootstrap

The BWR is a special case of the *mirror-match bootstrap* (BMM) proposed by Sitter (1992a), which is mainly designed for the case where the original first stage sampling is without replacement and parallels the original sampling scheme more closely. (1) Draw a simple random sample of size $n_h^* < n_h$ without replacement from $\{\mathbf{y}_{hi}, i = 1, \dots, n_h\}$. (2) Repeat step (1) k_h times independently to get $\mathbf{y}_{hi}^*, i = 1, \dots, m_h$, where

$$k_h = \frac{n_h(1 - f_h^*)}{n_h^*(1 - \lambda_h f_h)}, \quad (6.30)$$

$f_h^* = n_h^*/n_h$, and $m_h = k_h n_h^*$ (the case of noninteger k_h will be discussed later). (3) Repeat step (2) independently for each stratum. Note that this method reduces to the BWR if we choose $n_h^* = 1$ for all h . The bootstrap estimators are obtained the same way as before, but using $\tilde{\theta}^*$ in place of $\hat{\theta}^*$, where

$$\tilde{\theta}^* = T(\tilde{Z}^*) \quad \text{and} \quad \tilde{Z}^* = \sum_{h=1}^L \frac{1}{m_h} \sum_{i=1}^{m_h} z_{hi}^*.$$

In the special case of $\hat{\theta} = c' \hat{Z}$, the bootstrap variance estimator obtained from the BMM is equal to

$$\begin{aligned} \text{var}_* \left(\sum_{h=1}^L \frac{1}{m_h} \sum_{i=1}^{m_h} c' z_{hi}^* \right) &= \sum_{h=1}^L \frac{1}{k_h} \text{var}_* \left(\frac{1}{n_h^*} \sum_{i=1}^{n_h^*} c' z_{hi}^* \right) \\ &= \sum_{h=1}^L \frac{1 - f_h^*}{k_h n_h^*} c' s_h^2 c = v_L, \end{aligned}$$

since $\{z_{hi}^*, i = 1, \dots, n_h^*\}$ is a simple random sample without replacement from $\{z_{hi}, i = 1, \dots, n_h\}$ and k_h is given by (6.30). Therefore, the BMM also provides consistent bootstrap estimators for the case of $\hat{\theta} = c' \hat{Z}$. Some asymptotic results for $\hat{\theta} = g(\hat{Z})$ with a nonlinear g are given in Sitter (1992a).

A disadvantage of the BMM (and of the BWR) is that the k_h in (6.30) may not be an integer. A randomization was proposed to handle this problem (Sitter, 1992a). Let \underline{k}_h be the greatest integer less than k_h , \bar{k}_h be the smallest integer greater than k_h , and K_h be a random variable such that

$$P\{K_h = \underline{k}_h\} = p_h \quad \text{and} \quad P\{K_h = \bar{k}_h\} = 1 - p_h,$$

where

$$p_h = (\underline{k}_h^{-1} - \bar{k}_h^{-1}) / (\underline{k}_h^{-1} - \bar{k}_h^{-1}).$$

Assume that $\underline{k}_h \geq 1$, which is implied by $1 \leq n_h^* < n_h / (2 - \lambda_h f_h)$. Then apply the procedure as described, using K_h in place of k_h . The randomization must be applied independently across the strata and repeated at each bootstrap sampling when Monte Carlo approximation is used. With this randomization, one still gets consistent bootstrap estimators.

Sitter (1992a) also suggested the use of $n_h^* = f_h n_h$ to match the third order moments of $\hat{Z} - E(\hat{Z})$ and $\tilde{\hat{Z}}^* - E_*(\tilde{\hat{Z}}^*)$.

(IV) The without-replacement bootstrap

Assume now that the first stage sampling is without replacement so that $\lambda_h = 1$ for all h . The *without-replacement bootstrap* (BWO) was proposed by Gross (1980) and Chao and Lo (1985) in the case of simple random sampling and extended to stratified sampling by Bickel and Freedman (1984). Assuming that $N_h = k_h n_h$ for some integer k_h , the BWO creates a pseudopopulation of size N_h by replicating the sample clusters in stratum h exactly k_h times, and then takes a bootstrap sample of size n_h as a simple random sample without replacement from the pseudopopulation, independently across the strata. The bootstrap estimators are obtained by using the same formulas for the standard bootstrap, e.g., (6.26) and (6.27). Although the BWO is intuitively appealing, it may not provide a consistent variance estimator in the case of $\hat{\theta} = c' \hat{Z}$, since the BWO variance estimator is of the form

$$\sum_{h=1}^L \text{var}_* \left(\frac{1}{n_h} \sum_{i=1}^{n_h} c' z_{hi}^* \right) = \sum_{h=1}^L \frac{1 - f_h}{n_h} \left[\frac{k_h(n_h - 1)}{k_h n_h - 1} \right] c' s_h^2 c.$$

Note that the scaling problem is still there and $\frac{k_h(n_h - 1)}{k_h n_h - 1} \approx \frac{n_h - 1}{n_h}$ for bounded n_h and large N_h . Bickel and Freedman (1984) considered a randomization between two pseudopopulations in each stratum that corrects this problem; however, their method is not always applicable (see McCarthy and Snowden, 1985; Sitter, 1992b).

Sitter (1992b) provided the following modified BWO that circumvents the scaling problem. (1) In stratum h , create a pseudopopulation by replicating $\{y_{hi}, i = 1, \dots, n_h\}$ k_h times and select a simple random sample of

size m_h without replacement from the pseudopopulation, where

$$m_h = n_h - (1 - f_h) \quad \text{and} \quad k_h = \frac{N_h}{n_h} \left(1 - \frac{1 - f_h}{n_h} \right). \quad (6.31)$$

(2) Repeat step (1) independently for each stratum. If the m_h or k_h in (6.31) is not an integer, then we randomize between bracketing integer values (see Sitter, 1992b for details). The m_h and k_h in (6.31) are chosen such that

$$f_h^* = f_h \quad \text{and} \quad \text{var}_*(c' \hat{Z}) = v_L,$$

where $f_h^* = m_h/(k_h n_h)$ is the resampling fraction in stratum h . Hence, this modified BWO provides consistent bootstrap estimators.

6.3 Comparisons by Simulation

Kovar (1985, 1987) and Kovar, Rao and Wu (1988) performed simulation studies to compare various resampling variance estimators and confidence intervals in the case of stratified one stage simple random sampling with replacement, based on several hypothetical populations that resemble real populations encountered in a National Assessment of Educational Progress study. All populations consist of $L = 32$ strata, and each unit in these populations has a two-dimensional vector of variables. The parameters of interest are the ratio of population means, the regression coefficient and correlation coefficient between the two variables, and the population median of the second variable. A detailed description of the population parameters can be found in Kovar (1985, 1987).

We first consider the cases of θ = the ratio, the regression coefficient, and correlation coefficient. The results presented here correspond to normal population 2 in Kovar (1985) with several values of parameters ρ (the true correlation coefficient between the two variables, assumed to be a constant across strata) and cv_1 (the coefficient of variation of the first variable). Table 6.1 lists, based on 100 simulations, the relative bias (RB) and the relative stability (RS) of the linearization (LIN), the jackknife (JACK), the BRR, and the rescaling bootstrap (BRS) variance estimators for the case of $n_h = 2$ per stratum, where

$$\text{RB} = \frac{\text{simulation mean of variance estimator}}{\text{the ture mse}} - 1,$$

and

$$\text{RS} = \frac{(\text{simulation mse of variance estimator})^{1/2}}{\text{the ture mse}}.$$

The bootstrap estimators in all cases are approximated by the simple Monte Carlo approximation with $B = 100$.

Table 6.1. Performances of variance estimators ($n_h = 2$) [Adapted from Kovar, Rao and Wu (1988), by permission of Statistical Society of Canada]

Method	$\rho = 0.8$				$\rho = 0.5$			
	$cv_1 = 0.014$		$cv_1 = 0.142$		$cv_1 = 0.014$		$cv_1 = 0.071$	
	RB	RS	RB	RS	RB	RS	RB	RS
$\theta = \text{ratio}$								
LIN	-0.02	0.27	-0.06	0.54	0.05	0.32	-0.02	0.32
JACK	-0.02	0.27	-0.05	0.54	0.05	0.32	-0.02	0.32
BRR	-0.02	0.27	0.18	0.79	0.05	0.32	-0.01	0.33
BRS ₂ [†]	0.00	0.33	0.22	1.20	0.05	0.37	-0.02	0.35
BRS ₃	-0.04	0.28	0.17	0.87	0.04	0.33	-0.01	0.36
$\theta = \text{regression coefficient}$								
LIN	-0.07	0.43	0.03	0.49	-0.04	0.39	-0.07	0.45
JACK	-0.06	0.44	0.05	0.51	-0.03	0.39	-0.06	0.46
BRR	0.04	0.49	0.20	0.58	0.04	0.42	0.08	0.54
BRS ₂	0.04	0.52	0.23	0.66	0.06	0.48	0.12	0.61
BRS ₃	0.05	0.57	0.26	0.57	0.04	0.45	0.10	0.61
$\theta = \text{correlation coefficient}$								
LIN	0.04	0.56	0.14	0.73	-0.05	0.39	-0.02	0.50
JACK	0.05	0.58	0.17	0.75	-0.04	0.39	-0.01	0.52
BRR	0.19	0.70	0.36	0.91	0.06	0.44	0.12	0.59
BRS ₂	0.22	0.74	0.36	0.93	0.09	0.58	0.16	0.72
BRS ₃	0.22	0.80	0.40	1.00	0.06	0.42	0.10	0.62

† BRS_k is the rescaling bootstrap with $m_h = k$.

The results in Table 6.1 can be summarized as follows:

- (1) The linearization and the jackknife variance estimators perform equivalently in terms of both RB and RS, confirming the second order asymptotic equivalence of these estimators [see (6.18)].
- (2) In the case of $\theta = \text{ratio}$, all of the variance estimators perform well when $cv_1 \leq 0.1$. However, as cv_1 increases, the BRR and BRS variance estimators tend to overestimate. The phenomenon becomes more pronounced as the nonlinearity of the estimators increases (ratio \rightarrow regression coefficient \rightarrow correlation coefficient).
- (3) Overall, the linearization and the jackknife variance estimators have the best performance; the BRR variance estimator is the second best; and the BRS variance estimators are the worst. In addition, the results indicate that increasing m_h in the bootstrap method is not advisable.

Table 6.2 lists, based on 1000 simulations, the coverage probabilities of the lower confidence bound (CPL), the upper confidence bound (CPU), and the two-sided confidence interval (CPI), constructed by using normal approximation with one of v_L , v_{JACK} and v_{BRR} as a variance estimate, or using the bootstrap-t method with rescaling and variance estimator v_L . The nominal levels ($1 - \alpha$ or $1 - \alpha/2$) and the standardized lengths (SEL) of the two-sided confidence intervals (defined to be the mean simulation lengths divided by $2z_{1-\alpha/2}\sqrt{\text{true mse}}$) are also given. The underlying population is the same as that in Table 6.1 with $\rho = 0.8$ and $cv_1 = 0.142$.

The results in Table 6.2 can be summarized as follows:

- (1) In the case of $\theta =$ regression coefficient or correlation coefficient, the confidence sets based on the linearization or jackknife variance estimator tend to be liberal, while the confidence sets based on the BRR variance estimator perform well. The bootstrap-t confidence sets are highly conservative.
- (2) In the case of $\theta =$ ratio, even though the coverage probabilities for all of the two-sided confidence intervals are close to the nominal levels, the left tail is overstated while the right tail is understated for the normal approximation methods. The reverse holds in the case of $\theta =$ correlation coefficient. In either case, the bootstrap-t confidence sets ameliorate the situation, but at the cost of inflated lengths.

Table 6.2. Performances of confidence sets ($n_h = 2$) [Adapted from Kovar, Rao and Wu (1988), by permission of Statistical Society of Canada]

Method	$\alpha = 0.10$				$\alpha = 0.05$			
	CPL	CPU	CPI	SEL	CPL	CPU	CPI	SEL
$\theta =$ ratio								
LIN	0.863	0.938	0.801	0.974	0.918	0.988	0.906	0.974
JACK	0.863	0.942	0.805	0.978	0.918	0.988	0.906	0.978
BRR	0.878	0.974	0.852	1.096	0.924	0.999	0.923	1.096
BRS ₂	0.957	0.963	0.924	1.487	0.980	0.989	0.969	1.565
$\theta =$ regression coefficient								
LIN	0.877	0.887	0.764	0.939	0.922	0.936	0.858	0.939
JACK	0.878	0.888	0.766	0.947	0.923	0.936	0.859	0.947
BRR	0.891	0.910	0.791	1.015	0.935	0.945	0.880	1.015
BRS ₂	0.948	0.961	0.909	1.282	0.991	0.986	0.977	1.622
$\theta =$ correlation coefficient								
LIN	0.916	0.844	0.760	0.946	0.959	0.889	0.848	0.946
JACK	0.916	0.845	0.761	0.954	0.959	0.891	0.850	0.954
BRR	0.928	0.822	0.790	1.034	0.968	0.908	0.876	1.034
BRS ₂	0.955	0.960	0.915	1.575	0.992	0.987	0.979	1.700

Table 6.3. Performances of confidence sets and variance estimators ($n_h = 5$) [Adapted from Kovar, Rao and Wu (1988), by permission of Statistical Society of Canada]

Method	$\alpha = 0.10$			$\alpha = 0.05$			var	
	CPL	CPU	SEL	CPL	CPU	SEL	RB	RS
$\theta = \text{ratio}$								
LIN	0.865	0.944	0.937	0.921	0.978	0.937	-0.09	0.43
JACK	0.865	0.944	0.938	0.921	0.978	0.938	-0.09	0.43
BRS ₂	0.832	0.848	0.692	0.886	0.905	0.708	-0.01	0.51
BRS ₄	0.891	0.924	0.955	0.947	0.958	0.966	-0.00	0.52
$\theta = \text{regression coefficient}$								
LIN	0.902	0.897	0.991	0.948	0.946	0.991	0.00	0.30
JACK	0.906	0.900	1.008	0.948	0.952	1.008	0.04	0.32
BRS ₂	0.844	0.840	0.782	0.910	0.897	0.819	0.02	0.31
BRS ₄	0.908	0.898	1.038	0.949	0.961	1.058	0.02	0.32
$\theta = \text{correlation coefficient}$								
LIN	0.918	0.869	0.949	0.969	0.921	0.949	-0.07	0.34
JACK	0.921	0.875	0.966	0.971	0.928	0.966	-0.04	0.36
BRS ₂	0.844	0.836	0.758	0.901	0.911	0.792	-0.04	0.37
BRS ₄	0.901	0.907	1.007	0.959	0.955	1.033	-0.03	0.38

Similar simulation results for the case of $n_h = 5$ per stratum, based on 1000 simulations, are given in Table 6.3 (the BRR is not included) for the same population with $\rho = 0.8$ and $cv_1 = 0.09$. The results confirm the previous conclusions. In addition, we have the following findings from Table 6.3:

- (1) The one-sided bootstrap-t confidence intervals, with $m_h = n_h - 1 = 4$ for the case of $\theta = \text{ratio}$ and correlation coefficient, are much better than those based on linearization and the jackknife. But for two-sided confidence intervals, the three methods exhibit similar performances in terms of coverage probabilities ($CPI = CPL + CPU - 1$).
- (2) The bootstrap-t confidence sets with $m_h = 2$ [$\approx (n_h - 2)^2 / (n_h - 1)$] give shorter standardized lengths but considerably lower coverage probabilities, suggesting that the choice of m_h based on matching the third moments may not be good.

The results in Table 6.4 are for the case of $\theta = \text{median}$, based on 500 simulations and $B = 500$ for the Monte Carlo approximation to the bootstrap estimators. The population is described in Kovar (1987). The linearization is replaced by its counterpart, Woodruff's method described in (6.9) and

(6.10). The jackknife is omitted because of its inconsistency. The BRR is considered only in the case of $n_h = 2$ for all h . The approximation to the BRR, using random subsampling introduced in Section 6.2.3 (or the RRG; see Kovar, Rao and Wu, 1988), is included for both cases of $n_h = 2$ and $n_h = 5$. The nominal level for the one-sided confidence intervals is 95%. The following is a summary of the results in Table 6.4:

- (1) The performance of the variance estimator based on Woodruff's method depends on the constant α in (6.10), and the choice of $\alpha = 0.05$ seems quite reasonable.
- (2) The BRR variance estimator and its approximation are comparable to Woodruff's estimators. The approximation to the BRR is very good in the case of $n_h = 2$.
- (3) In the case of $n_h = 2$, the bootstrap variance estimator with $m_h = n_h - 1 = 1$ is very similar to the BRR estimator. When $n_h = 5$, however, the bootstrap variance estimators do not perform well.
- (4) All of the confidence sets have approximately the same coverage probabilities and standardized lengths.

Table 6.4. Performances of confidence sets and variance estimators
 $(\hat{\theta} = \text{sample median}, \alpha = 0.05)$ [Adapted from Kovar, Rao and
Wu (1988), by permission of Statistical Society of Canada]

	Method	CPL	CPU	SEL	RB	RS
$n_h = 2$	$W_{0.010}^\dagger$				0.032	0.52
	$W_{0.025}$				0.037	0.57
	$W_{0.050}$	0.948	0.952	0.98	0.052	0.65
	$W_{0.100}$				0.047	0.72
	$W_{0.200}$				0.076	0.90
	BRR	0.952	0.948	1.01	0.115	0.70
	RRG ‡	0.948	0.950	1.01	0.120	0.71
	BRS ₁	0.944	0.936	0.94	0.109	0.68
$n_h = 5$	$W_{0.010}$				0.050	0.40
	$W_{0.025}$				0.047	0.43
	$W_{0.050}$	0.934	0.946	1.00	0.050	0.48
	$W_{0.100}$				0.082	0.57
	$W_{0.200}$				0.110	0.74
	RRG	0.926	0.960	1.01	0.053	0.34
	BRS ₂	0.984	0.944	1.10	0.544	0.84
	BRS ₄	0.930	0.936	1.00	0.147	0.53

\dagger W_t is Woodruff's method with $\alpha = t$.

\ddagger The approximated BRR.

Rao and Shao (1995) studied by simulation the performances of the approximated BRR variance estimators, the GBRR, and the RGBRR variance estimators described in Section 6.2.3. The population under consideration is population 2 in Valliant (1987), which consists of 2000 units and is divided into $L = 5$ strata. More details about the population can be found in Valliant (1987) and Rao and Shao (1995). Single stage simple random sampling with replacement is used within each stratum and $n_h = 48$ for all h . Four different $\hat{\theta}$ are considered: the (stratified) sample mean, the sample median, the separate ratio estimator, and the separate linear regression estimator of the population mean (Cochran, 1977, p. 164, 201), using data from an auxiliary variable. Four variance estimators and the confidence intervals based on these variance estimators and the normal approximation are examined in the simulation: the jackknife variance estimator, the GBRR variance estimator, and the RGBRR variance estimators with $R_G = 15$ and 30. The GBRR variance estimator is obtained by first randomly grouping the 48 units in each stratum into two groups of the same size and then treating the two groups as two units and applying the BRR for the case of $n_h = 2$, using five columns from an 8×8 Hadamard matrix ($R = 8$). The RGBRR variance estimators are calculated by averaging R_G independent GBRR variance estimators. Note that the RGBRR variance estimator with $R_G = 30$ requires $30 \times 8 = 240$ computations of the estimator $\hat{\theta}$, which is the same as that required by the jackknife; the RGBRR variance estimator with $R_G = 15$ reduces the amount of computation by half, and the GBRR variance estimator requires the least amount of computation, only 8 evaluations of $\hat{\theta}$.

Table 6.5 shows the RB and the RS of these variance estimators, and the CPL, the CPU, and the CPI of the lower, upper, and two-sided confidence intervals ($\alpha = 0.05$), respectively, based on 2000 simulations. The following is a summary of the results in Table 6.5:

- (1) The jackknife has the best performance when $\hat{\theta}$ is a smooth estimator but performs poorly when $\hat{\theta}$ is the sample median.
- (2) The GBRR variance estimator has a small relative bias but is very unstable. The coverage probabilities of the confidence intervals based on the GBRR variance estimator are not close to the nominal level.
- (3) The RGBRR variance estimators perform well in all cases; even the choice $R_G = 15$ leads to significant reduction in the relative stability to the GBRR variance estimator.
- (4) There is no substantial difference between the performances of the two RGBRR variance estimators. Thus, the use of $R_G = 15$ may be recommended because it requires only half of the computations needed when using $R_G = 30$.

Table 6.5. Performances of the jackknife, the GBRR, and the RGBRR

Method	RB	RS	CPI	CPL	CPU
The sample mean					
JACK	-0.014	0.112	0.942	0.975	0.967
GBHS	-0.026	0.674	0.880	0.937	0.944
RGBRR, $R_G = 15$	-0.018	0.206	0.939	0.975	0.965
RGBRR, $R_G = 30$	-0.016	0.168	0.943	0.974	0.970
The separate ratio estimator					
JACK	0.006	0.112	0.946	0.974	0.973
GBHS	-0.009	0.654	0.888	0.942	0.947
RGBRR, $R_G = 15$	0.012	0.204	0.947	0.973	0.975
RGBRR, $R_G = 30$	0.013	0.166	0.948	0.974	0.975
The separate regression estimator					
JACK	0.020	0.126	0.951	0.977	0.975
GBHS	0.029	0.685	0.900	0.949	0.952
RGBRR, $R_G = 15$	0.053	0.228	0.954	0.978	0.977
RGBRR, $R_G = 30$	0.050	0.188	0.954	0.979	0.976
The sample median					
JACK	2.362	7.423	0.798	0.917	0.882
GBHS	0.125	1.174	0.830	0.929	0.902
RGBRR, $R_G = 15$	0.088	0.596	0.915	0.975	0.940
RGBRR, $R_G = 30$	0.089	0.568	0.915	0.974	0.941

Sitter (1992b) compared various bootstrap methods (the BWR, the BRS, the BMM with $n_h^* = f_h n_h$, and the modified BWO) based on several populations and 500 simulations under stratified one stage simple random sampling without replacement. Table 6.6 contains the results for the sample median under population 7 in Sitter (1992b). All of the bootstrap estimators are computed by Monte Carlo with $B = 500$. The results indicate that all of the bootstrap methods perform reasonably well, but none of them out-performs the others. Again, it is shown that Woodruff's method depends on α , and $\alpha = 0.05$ is a good choice.

Similar results for smooth functions of weighted averages can be found in Sitter (1992b). Furthermore, Sitter (1992b) observed that the bootstrap methods provide better results than Woodruff's method when the stratification is done using a highly correlated concomitant variable (see also Kovar, Rao and Wu, 1988).

Table 6.6. Performances of confidence sets and variance estimators ($\hat{\theta}$ = sample median) [Adapted from Sitter (1992b), by permission of Statistical Society of Canada]

Method	$\alpha = 0.10$			$\alpha = 0.05$			var	
	CPL	CPU	SEL	CPL	CPU	SEL	RB	RS
$n_h = 5$								
W _{0.010}							2.21	2.46
W _{0.025}							-0.09	0.44
W _{0.050}				0.940	0.978	1.26	0.19	0.59
W _{0.100}	0.942	0.824	0.77				0.69	1.05
W _{0.200}							-0.28	0.70
BWR	0.942	0.864	0.85	0.942	0.978	0.83	0.23	0.95
BRS ₄	0.908	0.938	0.89	0.900	0.938	0.69	0.08	0.87
BMM	0.942	0.846	0.85	0.942	0.978	0.83	0.14	0.84
BWO	0.942	0.824	0.77	0.942	0.930	0.76	0.09	0.86
$n_h = 7$								
W _{0.010}							0.43	0.65
W _{0.025}							-0.29	0.46
W _{0.050}				0.954	0.810	0.69	-0.06	0.47
W _{0.100}	0.998	0.974	1.11				0.33	0.74
W _{0.200}							-0.42	0.67
BWR	0.910	0.842	0.73	0.982	0.978	1.02	0.18	0.79
BRS ₆	0.888	0.882	0.74	0.974	0.898	0.83	0.01	0.68
BMM	0.910	0.872	0.96	0.910	0.980	0.76	0.09	0.81
BWO	0.910	0.980	0.98	0.910	0.980	0.77	0.16	0.87

6.4 Asymptotic Results

This section consists of a rigorous treatment of the consistency of the jackknife and the BRR variance estimators, the asymptotic relationship among various variance estimators, and the consistency of the bootstrap distribution estimators.

6.4.1 Assumptions

The population in a survey problem contains finitely many (N_T) ultimate units, although N_T may be very large. An asymptotic framework is provided by assuming that the finite population under study is a member of a

sequence of finite populations indexed by $k = 1, 2, \dots$. Thus, the population quantities $L, N_T, N_h, N_{hi}, Y_{hij}, F$, and θ ; the sample sizes n_T, n_h, n_{hi} , and n ; the sample values y_{hij} ; the survey weights w_{hij} ; and the survey estimates $\hat{F}, \hat{\theta}, v_L, v_{\text{JACK}}, v_{\text{BRR}}, v_{\text{BOOT}}$, etc., depend on the population index k , but, for simplicity of notation, k will be suppressed in what follows. All limiting processes, however, will be understood to be as $k \rightarrow \infty$. Note that the parameter of interest θ is not fixed as k increases; but we always assume that $\{\theta, k = 1, 2, \dots\}$ is a bounded set.

The total number of first stage sampled clusters is assumed large, i.e., n defined in (6.11) $\rightarrow \infty$ as $k \rightarrow \infty$. Also, without loss of generality, we assume that $\sup_h f_h < 1$ and that, for each k , there is a set $\mathcal{H}_k \subset \{1, \dots, L\}$ (note that L depends on k) such that

$$\sup_{h \in \mathcal{H}_k, k=1,2,\dots} n_h < \infty \quad \text{and} \quad \min_{h \notin \mathcal{H}_k} n_h \rightarrow \infty. \quad (6.32)$$

Note that (6.32) includes the following two common situations in surveys: (1) all of the n_h are small (bounded by a constant), in which case $\mathcal{H}_k = \{1, \dots, L\}$; and (2) all of the n_h are large, in which case $\mathcal{H}_k = \emptyset$.

The following are typical assumptions in asymptotic analysis of survey data.

It is assumed that no survey weight is disproportionately large, i.e.,

$$\max_{h,i,j} \frac{n n_{hi} w_{hij}}{N_T} = O(1). \quad (6.33)$$

Under this assumption, \hat{F} is consistent for F and is asymptotically normal. It is of interest to see what (6.33) reduces to in the special case of Example 6.1 or 6.2. If the sampling design is a stratified one stage simple random sampling (Example 6.1), then $n = n_T$ and (6.33) becomes

$$\max_h \frac{n_T N_h}{n_h N_T} = O(1). \quad (6.34)$$

If the sampling design is a stratified two stage sampling design (Example 6.2) and simple random sampling is used in both stages of sampling, then $w_{hij} = N_h N_{hi} / n_h n_{hi}$ and (6.33) reduces to

$$\max_{i,h} \frac{n N_h N_{hi}}{n_h n_{hi} N_T} = O(1). \quad (6.35)$$

Note that N_{hi} is the number of units in the i th cluster of stratum h . If N_{hi} , $i = 1, \dots, N_h$, are relatively the same or they are bounded, then (6.35) is the same as (6.34).

To make the asymptotic treatment simpler, we redefine \hat{Z} and z_{hi} as follows: \hat{Z} and z_{hi} are still as defined in (6.5) and (6.7), respectively, but

with w_{hij} replaced by w_{hij}/N_T . This change of notation does not have any effect when $\hat{\theta} = T(\hat{F})$ or $\hat{\theta} = g(\hat{Z})$ is proportional to $g(N_T \hat{Z})$ [e.g., $g(\hat{Z}) = c' \hat{Z}$ and $\hat{\theta} = \hat{F}^{-1}(p)$].

With these redefined \hat{Z} and z_{hi} , the following is a Liapunov-type condition for the asymptotic normality of \hat{Z} :

$$\sum_{h=1}^L \sum_{i=1}^{n_h} E \left\| \frac{z_{hi} - Ez_{hi}}{n_h} \right\|^{2+\tau} = O\left(\frac{1}{n^{1+\tau}}\right), \quad (6.36)$$

where $\tau > 0$ is a fixed constant. It is also assumed that

$$\liminf_k [n \text{var}(\hat{Z})] > 0. \quad (6.37)$$

Conditions (6.36) and (6.37) imply that the convergence rate of \hat{Z} to $E\hat{Z}$ is exactly $n^{-1/2}$, which is the case for most problems in sample surveys. Although condition (6.37) can be relaxed, we assume it for simplicity of the presentation.

In the case where $\hat{\theta} = \hat{F}^{-1}(p)$, a sample quantile, some “differentiability” condition on the population distribution function F is required. Although F is not differentiable for each fixed k , we may assume that F is differentiable in the following limiting sense: there exists a sequence of functions $\{f_k(\cdot), k = 1, 2, \dots\}$ such that

$$\lim_{k \rightarrow \infty} \left[\frac{F(\theta + O(n^{-1/2})) - F(\theta)}{O(n^{-1/2})} - f(\theta) \right] = 0, \quad (6.38)$$

$$0 < \inf_k f(\theta) \quad \text{and} \quad \sup_k f(\theta) < \infty. \quad (6.39)$$

That is, f plays the role of the “density” of F . Note that the population index k for F , θ , and $f(\theta)$ is suppressed in (6.38) and (6.39).

6.4.2 The jackknife and BRR for functions of averages

Consider the case of $\hat{\theta} = g(\hat{Z})$. Assuming conditions (6.32)–(6.37) and that g is differentiable, we can show that

$$(\hat{\theta} - \theta)/\sigma \rightarrow_d N(0, 1),$$

where

$$\sigma^2 = \nabla g(\mu)' \text{var}(\hat{Z}) \nabla g(\mu) \quad (6.40)$$

is the asymptotic variance of $\hat{\theta} = g(\hat{Z})$ and $\mu = E\hat{Z}$. The consistency of the jackknife estimator defined in (6.12) can be established in the same way as in the i.i.d. case (see Theorem 2.1).

Theorem 6.1. Suppose that (6.32)–(6.37) hold and that the function g is continuously differentiable with nonzero ∇g in a compact set containing $\{\mu, k = 1, 2, \dots\}$. Then

$$v_{\text{JACK}}/\sigma^2 \rightarrow_p 1, \quad (6.41)$$

where σ^2 is as given in (6.40).

Proof. From (6.16),

$$\sum_{h=1}^L \frac{(1 - \lambda_h f_h)(n_h - 1)}{n_h} \sum_{i=1}^{n_h} [(\hat{Z}_{hi} - \hat{Z})' \nabla g(\hat{Z})]^2 = v_L, \quad (6.42)$$

which is consistent in the sense that

$$v_L/\sigma^2 \rightarrow_p 1. \quad (6.43)$$

This result was established by Krewski and Rao (1981) and Bickel and Freedman (1984); it follows from the continuity of ∇g , condition (6.37), and the law of large numbers (see Appendix A.5). Since (6.42) and (6.43) hold for $g(x) = c'x$ for arbitrary $c \neq 0$,

$$\sum_{h=1}^L \frac{(1 - \lambda_h f_h)(n_h - 1)}{n_h} \sum_{i=1}^{n_h} \|\hat{Z}_{hi} - \hat{Z}\|^2 / \text{tr}[\text{var}(\hat{Z})] \rightarrow_p 1;$$

this and (6.36) imply that

$$\sum_{h=1}^L \sum_{i=1}^{n_h} \|\hat{Z}_{hi} - \hat{Z}\|^2 \rightarrow_p 0.$$

Consequently,

$$\max_{h,i} \|\hat{Z}_{hi} - \hat{Z}\|^2 \rightarrow_p 0. \quad (6.44)$$

The rest of the proof is the same as that of Theorem 2.1 with $\bar{X}_{n-1,i}$ and \bar{X}_n replaced by \hat{Z}_{hi} and \hat{Z} , respectively. \square

A similar result can be obtained for the BRR variance estimator defined in (6.22).

Theorem 6.2. Assume the conditions in Theorem 6.1 and that

$$0 < \epsilon_0 \leq \frac{m_h}{n_p} \leq \frac{1}{2} \quad \text{for all } h \quad (6.45)$$

and

$$R/n^{1+\tau/2} \rightarrow 0, \quad (6.46)$$

where ϵ_0 is a constant and τ is as given in (6.36). Then

$$v_{\text{BRR}}/\sigma^2 \rightarrow_p 1. \quad (6.47)$$

Proof. From the adjustment in (6.21) and the balanced nature of the replicates, $R^{-1} \sum_{r=1}^R \tilde{Z}^{(r)} = \hat{Z}$ and

$$\frac{1}{R} \sum_{r=1}^R [(\tilde{Z}^{(r)} - \hat{Z})' \nabla g(\hat{Z})]^2 = v_L.$$

Following the proofs of Theorems 2.1 and 6.1, result (6.47) holds if

$$\max_{r=1,\dots,R} \|\tilde{Z}^{(r)} - \mu\| \rightarrow_p 0. \quad (6.48)$$

Under condition (6.45), (6.48) follows from

$$\max_{r=1,\dots,R} \|\hat{Z}^{(r)} - \mu\| \rightarrow_p 0. \quad (6.49)$$

Let $\epsilon > 0$ be given. Then

$$\begin{aligned} P\{\max_{r=1,\dots,R} \|\hat{Z}^{(r)} - \mu\| > \epsilon\} &\leq \sum_{r=1}^R P\{\|\hat{Z}^{(r)} - \mu\| > \epsilon\} \\ &\leq \frac{R}{\epsilon^{2+\tau}} E\|\hat{Z}^{(r)} - \mu\|^{2+\tau} = O\left(\frac{R}{n^{1+\tau/2}}\right) \end{aligned}$$

by (6.36). This and (6.46) imply (6.49). \square

Condition (6.45) ensures that the new weights $w_{hij}^{(r)}$ are positive and satisfy (6.33). It holds if n_h are bounded and $m_h = 1$ for all h . Condition (6.46) indicates that choosing a not too large number of replicates in the BRR not only saves computations but also has theoretical advantages.

Results (6.41), (6.43), and (6.47) showed that v_L , v_{JACK} , and v_{BRR} are (first order) asymptotically equivalent. Rao and Wu (1985) studied the second order asymptotic equivalence among v_L , v_{JACK} , and v_{BRR} . Their results are summarized (and extended in the case of the BRR) in the following theorem.

Theorem 6.3. Suppose that (6.33), (6.37), (6.45), and (6.46) hold and that (6.36) holds with $\tau = 2$. Suppose further that the function g is twice continuously differentiable with nonzero ∇g in a compact set containing $\{\mu, k = 1, 2, \dots\}$. Then

- (i) $v_{JACK}/v_L = 1 + O_p(n^{-1})$;
- (ii) $v_{JACK}/v_L = 1 + O_p(n^{-2})$ if $n_h = 2$ for all h ;
- (iii) $v_{BRR}/v_L = 1 + O_p(n^{-1/2})$.

Proof. Under condition (6.37), it suffices to show that $v_{JACK} - v_L = O_p(n^{-2}) [= O_p(n^{-3})]$ when $n_h = 2$ for all h] and $v_{BRR} - v_L = O_p(n^{-3/2})$.

From the second order differentiability of g ,

$$\begin{aligned}\hat{\theta}_{hi} &= \hat{\theta} + (\hat{Z}_{hi} - \hat{Z})' \nabla g(\hat{Z}) + \frac{1}{2} (\hat{Z}_{hi} - \hat{Z})' \nabla^2 g(\xi_{hi}) (\hat{Z}_{hi} - \hat{Z}) \\ &= \hat{\theta} + l_{hi} + q_{hi} + r_{hi},\end{aligned}\quad (6.50)$$

where ξ_{hi} is on the line segment between \hat{Z}_{hi} and \hat{Z} , $l_{hi} = (\hat{Z}_{hi} - \hat{Z})' \nabla g(\hat{Z})$, $q_{hi} = \frac{1}{2} (\hat{Z}_{hi} - \hat{Z})' \nabla^2 g(\hat{Z}) (\hat{Z}_{hi} - \hat{Z})$, and $r_{hi} = \hat{\theta}_{hi} - \hat{\theta} - l_{hi} - q_{hi}$. Under condition (6.36) with $\tau = 2$,

$$\sum_{h=1}^L \frac{(1 - \lambda_h f_h)(n_h - 1)}{n_h} \sum_{i=1}^{n_h} q_{hi}^2 = O_p(n^{-3}), \quad (6.51)$$

which, together with (6.44) and the continuity of $\nabla^2 g$, implies that (6.51) also holds with q_{hi} replaced by r_{hi} . Using these results, expansion (6.50), and result (6.16), we obtain that

$$\begin{aligned}v_{\text{JACK}} &= \sum_{h=1}^L \frac{(1 - \lambda_h f_h)(n_h - 1)}{n_h} \sum_{i=1}^{n_h} (l_{hi} + q_{hi} + r_{hi} - \bar{q}_h - \bar{r}_h)^2 \\ &= v_L + 2 \sum_{h=1}^L \frac{(1 - \lambda_h f_h)(n_h - 1)}{n_h} \sum_{i=1}^{n_h} l_{hi} q_{hi} + O_p(n^{-3}) \\ &= v_L + \sum_{h=1}^L \frac{1 - \lambda_h f_h}{n_h(n_h - 1)^2} \sum_{i=1}^{n_h} u'_{hi} \nabla g(\hat{Z}) u'_{hi} \nabla^2 g(\hat{Z}) u_{hi} + O_p(n^{-3}),\end{aligned}$$

where $\bar{q}_h = n^{-1} \sum_{i=1}^{n_h} q_{hi}$, $\bar{r}_h = n^{-1} \sum_{i=1}^{n_h} r_{hi}$, and $u_{hi} = \bar{z}_h - z_{hi}$. This proves results (i) and (ii), since

$$\sum_{h=1}^L \frac{1 - \lambda_h f_h}{n_h(n_h - 1)^2} \sum_{i=1}^{n_h} u'_{hi} \nabla g(\hat{Z}) u'_{hi} \nabla^2 g(\hat{Z}) u_{hi} = O_p(n^{-2})$$

under condition (6.36) with $\tau = 2$ and

$$\sum_{i=1}^{n_h} u'_{hi} \nabla g(\hat{Z}) u'_{hi} \nabla^2 g(\hat{Z}) u_{hi} = 0$$

whenever $n_h = 2$. Similarly, for $\tilde{\theta}^{(r)}$ in the BRR estimator,

$$\tilde{\theta}^{(r)} = \hat{\theta} + l^{(r)} + q^{(r)}$$

with $l^{(r)} = (\tilde{Z}^{(r)} - \hat{Z})' \nabla g(\hat{Z})$ and $q^{(r)} = \frac{1}{2} (\tilde{Z}^{(r)} - \hat{Z})' \nabla^2 g(\xi^{(r)}) (\tilde{Z}^{(r)} - \hat{Z})$. Using (6.48), condition (6.36) with $\tau = 2$, and the balanced nature of the

replicates, we obtain that

$$\begin{aligned} v_{\text{BRR}} &= \frac{1}{R} \sum_{r=1}^R \left[l^{(r)} + q^{(r)} - \frac{1}{R} \sum_{r=1}^R (l^{(r)} + q^{(r)}) \right]^2 \\ &= v_L + \frac{2}{R} \sum_{r=1}^R l^{(r)} q^{(r)} + O_p(n^{-2}). \end{aligned}$$

Then result (iii) follows from

$$\frac{1}{R} \sum_{r=1}^R l^{(r)} q^{(r)} = O_p(n^{-3/2}). \quad \square$$

The results in Theorem 6.3, together with the simulation results in Section 6.3, indicate that the choice between the jackknife estimator v_{JACK} and the linearization estimator v_L may depend more on nonstatistical considerations, such as the feasibility of their implementation. The order of the magnitude of the difference between the BRR estimator v_{BRR} and v_L is greater than that between v_{JACK} and v_L , although v_{BRR} and v_L are still asymptotically the same up to the order of $n^{-3/2}$. Unlike v_{JACK} and v_L , however, the BRR estimators can be applied to both cases of $\hat{\theta} = g(\hat{Z})$ and $\hat{\theta} = \hat{F}^{-1}(p)$ (studied in Section 6.4.5).

6.4.3 The RGBRR and RSBRR for functions of averages

We next focus on the approximated BRR variance estimators for $\hat{\theta}$. Let v_{RGBRR} and v_{RSBRR} be the RGBRR and the RSBRR variance estimators defined in Section 6.2.3, respectively.

Theorem 6.4. *Assume the conditions in Theorem 6.2 with (6.46) replaced by*

$$\max_{h=1,\dots,L} n_h / n R_G \rightarrow 0 \quad \text{and} \quad R_G R / n^{1+\tau/2} \rightarrow 0. \quad (6.52)$$

Then

$$v_{\text{RGBRR}} / \sigma^2 \rightarrow_p 1, \quad (6.53)$$

where σ^2 is as given in (6.40). Result (6.53) also holds with v_{RGBRR} replaced by v_{RSBRR} if condition (6.52) is changed to

$$R \rightarrow \infty \quad \text{and} \quad R / n^{1+\tau/2} \rightarrow 0.$$

Proof. We prove the case of RGBRR only. A similar proof for the RSBRR can be found in Shao and Wu (1992, Theorem 5.1). Under the conditions

in Theorem 6.2 and condition (6.52),

$$\max_{r=1,\dots,R,t=1,\dots,R_G} \|\tilde{Z}_t^{(r)} - \mu\| \rightarrow_p 0,$$

where $\tilde{Z}_t^{(r)}$ is based on the r th replicate in the t th repeated grouping. Following the proof of Theorem 6.2, we obtain the desired result if we can show (6.53) for the linear case, i.e., $g(x) = c'x$ for a constant vector c .

For simplicity, we consider only the case where n_h is even for all h . Let $v_{GBRR}^{(t)}$ be the t th grouped BRR variance estimator. Then, by $g(x) = c'x$ and the balanced nature of the replicates,

$$v_{RGBRR} = \frac{1}{R_G} \sum_{t=1}^{R_G} v_{GBRR}^{(t)} = \frac{1}{4R_G} \sum_{t=1}^{R_G} \sum_{h=1}^L (1 - \lambda_h f_h) u_h^{(t)}, \quad (6.54)$$

where $u_h^{(t)} = [c'(\bar{z}_{h1}^{(t)} - \bar{z}_{h2}^{(t)})]^2$ and for fixed t and h , $\bar{z}_{h1}^{(t)}$ and $\bar{z}_{h2}^{(t)}$ are obtained by averaging the z_{hi} for the first and the second groups of clusters in stratum h , respectively. Let E_{RG} be the expectation with respect to the random grouping. Then, by (6.54),

$$E_{RG}(v_{RGBRR}) = E_{RG}(v_{GBRR}^{(1)}) = v_L.$$

By the law of large numbers (see Appendix A.5) and condition (6.37), result (6.53) follows from

$$\begin{aligned} \frac{1}{R_G^{1+\tau/2}} \sum_{t=1}^{R_G} \sum_{h=1}^L E(nu_h^{(t)})^{1+\tau/2} &= \frac{1}{R_G^{\tau/2}} \sum_{h=1}^L E(nu_h^{(1)})^{1+\tau/2} \\ &\leq \frac{2n^{1+\tau/2}}{R_G^{\tau/2}} \sum_{h=1}^L E\|\bar{z}_{h1}^{(1)} - E\bar{z}_{h1}^{(1)}\|^{2+\tau} \\ &\leq \frac{c_0 n^{1+\tau/2}}{R_G^{\tau/2}} \sum_{h=1}^L n_h^{\tau/2} \sum_{i=1}^{n_h} E\left\|\frac{z_{hi} - Ez_{hi}}{n_h}\right\|^{2+\tau} \\ &\leq c_0 \max_{h=1,\dots,L} \left(\frac{n_h}{n R_G}\right)^{\tau/2} n^{1+\tau} \sum_{h=1}^L \sum_{i=1}^{n_h} E\left\|\frac{z_{hi} - Ez_{hi}}{n_h}\right\|^{2+\tau} \\ &\rightarrow 0 \end{aligned}$$

[by conditions (6.36) and (6.52)], where c_0 is a constant. \square

Result (6.53) is also true for the grouped jackknife and the random subsampling jackknife variance estimators. The conditions and proofs are very similar and, therefore, are omitted.

Since R has the same order as L and $\max_h n_h/n \rightarrow 0$ holds in most cases where $L \rightarrow \infty$, the condition $\max_h n_h/n R_G \rightarrow 0$ is almost the same as (but slightly stronger than) $R_G R \rightarrow \infty$. As a special case of the RGBRR with $R_G = 1$, the GBRR variance estimator is consistent if $\max_h n_h/n \rightarrow 0$ (which is almost the same as $L \rightarrow \infty$). When L is fixed, however, the GBRR variance estimator (or the RGBRR variance estimator with a fixed R_G) is inconsistent, as the following example shows.

Example 6.5. Inconsistency of the GBRR estimators when L is fixed. Consider the case of stratified one stage simple random sampling with replacement (Example 6.1) and $g(x) = c'x$. In this case,

$$\hat{\theta} = c'\bar{y} = \sum_{h=1}^L \frac{N_h}{N_T} c'\bar{y}_h$$

is an unbiased estimator of $\theta = c'E(\bar{y})$ and

$$\text{var}(\hat{\theta}) = \sum_{h=1}^L \frac{N_h^2 \sigma_h^2}{N_T^2 n_h}, \quad (6.55)$$

where \bar{y}_h is the average of y_{hi} , $i = 1, \dots, n_h$, and $\sigma_h^2 = \text{var}(c'y_{hi})$.

Suppose that in each stratum, the sample size n_h is even and the data are divided into two groups of the same size. Let \bar{y}_{hj} be the average of the data in the j th group of stratum h , $j = 1, 2$, $h = 1, \dots, L$. The GBRR variance estimator is

$$v_{\text{GBRR}} = \frac{1}{4} \sum_{h=1}^L \frac{N_h^2}{N_T^2} [c'(\bar{y}_{h1} - \bar{y}_{h2})]^2 = \sum_{h=1}^L \frac{N_h^2 \sigma_h^2}{N_T^2 n_h} v_h, \quad (6.56)$$

where $v_h = n_h [c'(\bar{y}_{h1} - \bar{y}_{h2})]^2 / 4\sigma_h^2$.

Assume that L is fixed and $\min_h n_h \rightarrow \infty$. Consequently, for $h = 1, \dots, L$,

$$\frac{\sqrt{n_h}}{\sigma_h} \begin{pmatrix} c'[\bar{y}_h - E(\bar{y}_h)] \\ c'(\bar{y}_{h1} - \bar{y}_{h2})/2 \end{pmatrix} \rightarrow_d N(0, I_2) \quad (6.57)$$

and

$$v_h \rightarrow_d \chi_h^2,$$

a chi-square variable with one degree of freedom. Since the data in different strata are independent, it follows from (6.55)-(6.57) that

$$\frac{v_{\text{GBRR}}}{\text{var}(\hat{\theta})} \rightarrow_d \sum_{h=1}^L \omega_h \chi_h^2$$

if

$$\frac{N_h^2 \sigma_h^2}{n_h} \Big/ \sum_{h=1}^L \frac{N_h^2 \sigma_h^2}{n_h} \rightarrow \omega_h \quad \text{for all } h; \quad (6.58)$$

otherwise, $v_{\text{GBRR}}/\text{var}(\hat{\theta})$ does not converge.

In the special case of $L = 1$ (no stratification), $v_{\text{GBRR}}/\text{var}(\hat{\theta}) \rightarrow_d \chi_1^2$ and the studentized statistic

$$(\hat{\theta} - \theta) / \sqrt{v_{\text{GBRR}}} = \sum_{h=1}^L \frac{N_h}{N_T} c' [\bar{y}_h - E(\bar{y}_h)] / \sqrt{v_{\text{GBRR}}} \quad (6.59)$$

is asymptotically distributed as a t-distribution. Hence, one can still use it to make inference on θ , although v_{GBRR} is inconsistent (see Gray and Schucany, 1972). However, this phenomenon no longer exists when $L \geq 2$; by (6.57), the studentized statistic in (6.59) either has no limit distribution or converges in law to $N(0, 1) / (\sum_{h=1}^L \omega_h \chi_h^2)^{1/2}$ when (6.58) holds. Therefore, the studentized statistic in (6.59) cannot be used for making inference unless the limits ω_h in (6.58) are known.

6.4.4 The bootstrap for functions of averages

Under the conditions in Theorem 6.1 and some extra moment conditions, the consistency of the bootstrap variance estimator in (6.27) can also be established, using the same argument in Section 3.2.2. However, since the bootstrap is more useful in estimating distribution functions and constructing confidence sets, we focus on the consistency of the bootstrap estimator

$$H_{\text{BOOT}}(x) = P_* \{ \sqrt{n}(\hat{\theta}^* - \hat{\theta}) \leq x \} \quad (6.60)$$

of the distribution function H of $\sqrt{n}(\hat{\theta} - \theta)$, where $\hat{\theta} = g(\hat{Z})$ and θ^* is the bootstrap analog of $\hat{\theta}$ obtained by using any of the four modified bootstrap procedures introduced in Section 6.2.4. We also consider the bootstrap estimator

$$G_{\text{BOOT}}(x) = P_* \{ (\hat{\theta}^* - \hat{\theta}) / \sqrt{v_L^*} \leq x \}$$

of the distribution function G of the studentized statistic $(\hat{\theta} - \theta) / \sqrt{v_L}$, where v_L^* is a suitably defined bootstrap analog of v_L . For example, if the BWR is applied, then

$$v_L^* = \sum_{h=1}^L \frac{1}{m_h} \nabla g(\hat{Z}^*)' s_h^{*2} \nabla g(\hat{Z}^*),$$

where $s_h^{*2} = (m_h - 1)^{-1} \sum_{i=1}^{m_h} (z_{hi}^* - \bar{z}_h^*)(z_{hi}^* - \bar{z}_h^*)'$; if the BRS is used, then

$$v_L^* = \sum_{h=1}^L \frac{1 - \lambda_h f_h}{n_h - 1} \nabla g(\tilde{Z}^*)' s_h^{*2} \nabla g(\tilde{Z}^*).$$

Theorem 6.5. Assume the conditions in Theorem 6.1. Then,

$$\|H_{\text{BOOT}} - H\|_\infty \rightarrow_p 0 \quad \text{and} \quad \|G_{\text{BOOT}} - G\|_\infty \rightarrow_p 0.$$

Proof. Note that the asymptotic normality of $\sqrt{n}(\hat{\theta} - \theta)$ can be proved by using Liapunov's central limit theorem and Taylor's expansion. The consistency of H_{BOOT} can be derived by imitating this proof. The details are omitted here and can be found in Krewski and Rao (1981) and Bickel and Freedman (1984). The consistency of G_{BOOT} follows from the consistency of H_{BOOT} and

$$v_L^*/v_L \rightarrow_p 1. \quad (6.61)$$

We now prove (6.61) for the BRS; other cases can be treated similarly. Since $\tilde{Z}^* - \hat{Z} \rightarrow_p 0$, (6.61) follows from (6.37) and

$$n \left(\sum_{h=1}^L \frac{1 - \lambda_h f_h}{n_h - 1} s_h^{*2} - \sum_{h=1}^L \frac{1 - \lambda_h f_h}{n_h} s_h^2 \right) \rightarrow_p 0. \quad (6.62)$$

Since the expectation of the left-hand side of (6.62) is 0, the result follows from condition (6.36) and the law of large numbers (Appendix A.5). \square

Unlike the i.i.d. case, results on the accuracy of H_{BOOT} and G_{BOOT} are not available, mainly due to the technical difficulties in establishing Edgeworth expansions for complex survey data, although some attempts have been made (Chen and Sitter, 1993; Liu, 1992; Booth, Butler and Hall, 1994) and the simulation results in Section 6.3 indicated that the one-sided bootstrap confidence intervals are better than the normal approximation based confidence intervals in terms of coverage probability.

6.4.5 The BRR and bootstrap for sample quantiles

Let $p \in (0, 1)$ and $\hat{\theta} = \hat{F}^{-1}(p)$. Francisco and Fuller (1991) and Shao (1994b) obtained the following Bahadur-type representation:

$$\hat{\theta} = \theta + \frac{F(\theta) - \hat{F}(\theta)}{f(\theta)} + o_p\left(\frac{1}{n}\right), \quad (6.63)$$

which holds as long as conditions (6.33), (6.38), and (6.39) are satisfied. Result (6.63) implies that

$$(\hat{\theta} - \theta)/\sigma \rightarrow_d N(0, 1), \quad (6.64)$$

where $\sigma^2 = \sigma_k^2$ is the asymptotic variance of $[F(\theta) - \hat{F}(\theta)]/f(\theta)$.

It is known in this case that the jackknife variance estimator is inconsistent. Shao and Wu (1992) established the consistency of the BRR estimator v_{BRR} in (6.22) for the case of stratified one stage simple random sampling. A general result was obtained in Shao and Rao (1994).

Theorem 6.6. Assume conditions (6.32), (6.33), (6.38), (6.39), (6.45), and

$$R = o(e^{cm}) \quad (6.65)$$

for a constant $c > 0$, where $m = \sum_{h=1}^L m_h$. Then (6.47) and (6.53) hold with σ^2 being as given in (6.64).

Condition (6.65) is almost trivial, since in most applications R is chosen to be of the order $O(n^q)$ with $q \approx 1$. The result for v_{BRR} in Theorem 6.6 can be proved by using result (6.63) and the following two lemmas whose proofs are omitted and can be found in Shao and Rao (1994). The proofs for v_{RGBRR} and v_{RSBRR} are similar to that of Theorem 6.4.

Lemma 6.1. Assume conditions (6.33), (6.38), (6.39), and (6.65). Then,

$$\frac{1}{R} \sum_{r=1}^R (\tilde{\theta}^{(r)} - \hat{\theta})^4 = O_p\left(\frac{1}{m^2}\right).$$

Lemma 6.2. Assume condition (6.33). Let $\{T_k, k = 1, 2, \dots\}$ be a sequence of statistics admitting the expansion $T_k = l_k + r_k$, where $r_k = o_p(n^{-1})$ and l_k is a smooth function of some weighted averages $\bar{\zeta}_k$ satisfying

$$E\|\bar{\zeta}_k - E\bar{\zeta}_k\|^{2(1+\tau)} = O(n^{-(1+\tau)})$$

for some $\tau > 0$, and let σ_k^2 be the asymptotic variance of T_k . Then a sufficient condition for

$$v_{\text{BRR}}/\sigma_k^2 \rightarrow_p 1$$

is

$$\frac{1}{R} \sum_{r=1}^R (T_k^{(r)} - T_k)^{2(1+\tau)} = O_p\left(\frac{1}{m^{1+\tau}}\right).$$

Lemma 6.2 actually provides a result regarding the consistency of the BRR variance estimators for general statistics. Shao and Rao (1994) applied this lemma and established the consistency of v_{BRR} for the case where $T_k = \hat{F}(\hat{\theta}/2)$, the estimated low income proportions.

Let F^* be the bootstrap analog of \hat{F} using any of the four modified bootstrap procedures in Section 6.2.4 and $\theta^* = (F^*)^{-1}(p)$. We now consider the bootstrap estimator H_{BOOT} defined by (6.60).

Theorem 6.7. Assume conditions (6.32), (6.33), (6.38), (6.39), and (6.45). Then,

$$\|H_{\text{BOOT}} - H\|_\infty \rightarrow_p 0.$$

Proof. It suffices to show that a bootstrap analog of representation (6.63) holds, i.e.,

$$\theta^* = \hat{\theta} + \frac{\hat{F}(\theta) - F^*(\theta)}{f(\theta)} + o_p\left(\frac{1}{n}\right),$$

which can be established using a similar argument to that in Francisco and Fuller (1991) or Shao and Wu (1992). The details are omitted. \square

6.5 Resampling Under Imputation

Most surveys have missing observations due to various reasons (Cochran, 1977; Kalton, 1981). There are two types of survey nonresponse: unit nonresponse and item nonresponse. The former occurs when no information is collected from a sampled unit, whereas the latter occurs when the sampled unit cooperates in the survey but fails to provide answers to some of the questions, so that the corresponding vector of characteristics y_{hij} has some missing components.

Compensation procedures for handling missing survey data can be classified into two groups, weighting adjustments and imputation techniques. Weighting adjustments simply change the survey weights of respondents so that the new weights can be used in analyzing the data set that remains to be incomplete. Imputation techniques insert values for missing responses; that is, if a component of y_{hij} is missing, we impute it by a value obtained under a given imputation rule. After the imputation, survey estimates can be computed based on the completed data set. In general, imputation does not increase (in many cases it actually decreases) the statistical precision of survey estimates. The use of imputation is mainly motivated by non-statistical reasons. First, it is not practical to apply weighting adjustments to handle item nonresponse since it would result in different sets of weights for each item in one problem, which would cause serious difficulties for cross-tabulations and other analyses of the relationships between variables (Kalton, 1981). Imputation permits the use of the same survey weights for all items. Second, imputation creates a complete data set so that many standard complete-data algorithms for computing survey estimates can be used. Separated and complicated algorithms for a particular incomplete data set are not needed, and hence it is likely that much processing time is saved. For example, suppose that the sampling design is stratified one stage sampling with an equal number of samples from every stratum, i.e., $n_h = n_0$ for all h , or with proportional allocation (Cochran, 1977, p. 91), i.e., $\frac{n_h}{N_h} = \frac{n_T}{N_T}$. Then an incomplete data set usually does not preserve this balanced nature of the sampling design. Finally, the results obtained from different analyses are bound to be consistent with one another, which is

not necessarily true for the results of analyses of an incomplete data set (Kalton, 1981).

In this section, we consider applications of resampling methods to problems with imputed data. For simplicity, in what follows we assume that y_{hij} are one-dimensional and the first stage sampling is with replacement ($\lambda_h = 0$).

6.5.1 Hot deck imputation

For a given sample, let \mathbf{Y}_r be the collection of respondents and \mathbf{Y}_m be the collection of missing data. A commonly employed imputation rule, called the *hot deck imputation* method (Kalton, 1981; Sedransk, 1985), imputes the missing values in \mathbf{Y}_m with a random sample from \mathbf{Y}_r . Hot deck imputation introduces an additional randomness to the data set. There exist nonrandom imputation methods, such as mean imputation, ratio or regression imputation, and nearest neighbor imputation. The main disadvantage of using nonrandom imputation methods is that they may not preserve the distribution of item values, and hence cannot handle the problem in which the parameter of interest is a functional, such as a quantile, of the population distribution.

We confine our discussion to hot deck imputation. The case of nonrandom imputation is actually simpler and the results obtained here hold for some nonrandom imputation methods. Under the sampling design considered throughout this chapter, each $y_{hij} \in \mathbf{Y}_r$ is selected with (imputation) probability proportional to w_{hij} in the hot deck imputation (Rao and Shao, 1992). Note that the imputation cuts across strata and clusters. A practical reason for imputation cutting across strata and clusters is that some small strata or clusters may not have reasonably large numbers of respondents to allow the imputation to be performed within stratum or cluster. In some cases, one divides the population into several imputation classes and then performs hot deck imputation separately within each imputation class. This may increase the efficiency of imputation, especially when the nonresponse rates in different imputation classes are different.

It is a common practice to treat the imputed values as if they were true values. If the imputation is suitably done, estimators of parameters calculated using standard formulas (for the case of no missing data) and the data with imputed missing values are still asymptotically valid. However, the use of standard formulas for variance estimation, obtained by either linearization or resampling, leads to serious underestimation when the proportion of missing values is appreciable. We illustrate this by the following example.

Example 6.6. The sample mean under simple random sampling. Consider

the case of one stage simple random sampling (without stratification). Let θ be the population mean and $\hat{\theta} = \bar{y} = n^{-1} \sum_{i=1}^n y_i$ if there is no missing data. Suppose that all of the units have the same response probability $p \in (0, 1)$. Let A_r and A_m contain indices for respondents and nonrespondents, respectively. Under hot deck imputation, missing values are imputed by a simple random sample, $\{\eta_i, i \in A_m\}$, with replacement from $\{y_i, i \in A_r\}$. The estimate of θ after imputation is

$$\bar{y}_I = \frac{1}{n} \left(\sum_{i \in A_r} y_i + \sum_{i \in A_m} \eta_i \right).$$

Let n_m be the number of nonrespondents, $n_r = n - n_m$, and let E_I and var_I be the expectation and variance with respect to hot deck imputation, respectively, for given \mathbf{Y}_r . Then

$$E_I(\bar{y}_I) = \frac{1}{n_r} \sum_{i \in A_r} y_i = \bar{y}_r,$$

$$\text{var}_I(\bar{y}_I) = \frac{n_m}{n^2 n_r} \sum_{i \in A_r} (y_i - \bar{y}_r)^2 = \frac{n_m(n_r - 1)}{n^2 n_r} s_r^2,$$

and

$$\text{var}(\bar{y}_I) = \text{var}(\bar{y}_r) + E \left[\frac{n_m(n_r - 1)}{n^2 n_r} s_r^2 \right]. \quad (6.66)$$

If we apply the variance formula for the case of no missing value to the data $\{y_i, i \in A_r, \eta_i, i \in A_m\}$, then in this case the linearization, the jackknife, and the BRR variance estimators are all equal to s_I^2/n with

$$s_I^2 = \frac{1}{n-1} \left[\sum_{i \in A_r} (y_i - \bar{y}_I)^2 + \sum_{i \in A_m} (\eta_i - \bar{y}_I)^2 \right]. \quad (6.67)$$

Since $n_r/n \rightarrow_p p$, $n_r \text{var}(\bar{y}_r)/s_r^2 \rightarrow_p 1$, and

$$\begin{aligned} E[(1 - n^{-1})s_I^2] &= E \left[\frac{1}{n} E_I \left(\sum_{i \in A_r} y_i^2 + \sum_{i \in A_m} \eta_i^2 \right) - E_I(\bar{y}_I^2) \right] \\ &= E \left[\frac{1}{n_r} \sum_{i \in A_r} y_i^2 - \left(\frac{n_m(n_r - 1)}{n^2 n_r} s_r^2 + \bar{y}_r^2 \right) \right] \\ &= E \left[\left(1 - \frac{n_m}{n^2} \right) \frac{n_r - 1}{n_r} s_r^2 \right], \end{aligned}$$

we can show that

$$s_I^2 / [n \text{var}(\bar{y}_I)] \rightarrow_p p/(1 + p - p^2) < 1.$$

Hence, s_I^2/n is inconsistent and has a substantial negative relative bias if p is much smaller than 1.

6.5.2 An adjusted jackknife

In view of Example 6.6, some modifications to the resampling methods are required before applying them to an imputed data set. Rao and Shao (1992) proposed an adjusted jackknife that can be motivated by using Example 6.6. Note that s_{I}^2 in (6.67) divided by n is the same as the jackknife estimator

$$\frac{n-1}{n} \sum_{i=1}^n (\bar{y}_{\text{I},i} - \bar{y}_{\text{I}})^2,$$

where $\bar{y}_{\text{I},i}$ is the average over the data (after imputation) with the i th observation removed. For $i \in A_r$,

$$E_{\text{I}}(\bar{y}_{\text{I},i}) = \frac{1}{n-1} \left(\sum_{j \in A_r, j \neq i} y_j + n_m \bar{y}_r \right) = \frac{n_r - 1}{n-1} \bar{y}_{r,i} + \frac{n_m}{n-1} \bar{y}_r,$$

which is not the same as $\bar{y}_{r,i}$ due to the fact that the missing values are not imputed using $\{y_j, j \in A_r\}$ with y_i removed. Consequently, $\bar{y}_{\text{I},i}$, $i \in A_r$, do not have enough variability, which leads to the underestimation. This suggests the following adjustment: whenever y_i , $i \in A_r$, is deleted, we adjust the imputed value η_j to

$$\tilde{\eta}_{ij} = \eta_j + \bar{y}_{r,i} - \bar{y}_r.$$

The adjusted jackknife variance estimator is then

$$\tilde{v}_{\text{JACK}} = \frac{n-1}{n} \sum_{i=1}^n (\tilde{y}_{\text{I},i} - \bar{y}_{\text{I}})^2,$$

where

$$\tilde{y}_{\text{I},i} = \frac{1}{n-1} \left(\sum_{j \in A_r, j \neq i} y_j + \sum_{j \in A_m} \tilde{\eta}_{ij} \right)$$

when $i \in A_r$, $\tilde{y}_{\text{I},i} = \bar{y}_{\text{I},i}$ when $i \in A_m$, and \bar{y}_{I} is the average of $\tilde{y}_{\text{I},i}$, $i = 1, \dots, n$. A straightforward calculation shows that

$$\tilde{y}_{\text{I},i} = \begin{cases} (n-1)^{-1} [n\bar{y}_{\text{I}} - y_i - n_m(y_i - \bar{y}_r)/(n_r - 1)] & \text{if } i \in A_r \\ (n-1)^{-1} (n\bar{y}_{\text{I}} - \eta_i) & \text{if } i \in A_m \end{cases}$$

and

$$\begin{aligned} \tilde{v}_{\text{JACK}} &= \frac{1}{n(n-1)} \left\{ \sum_{i \in A_r} \left[\bar{y}_{\text{I}} - y_i - \frac{n_m(y_i - \bar{y}_r)}{n_r - 1} \right]^2 + \sum_{i \in A_m} (\bar{y}_{\text{I}} - \eta_i)^2 \right\} \\ &= \frac{1}{n(n-1)} \left[\sum_{i \in A_r} (y_i - \bar{y}_{\text{I}})^2 + \frac{n_m^2}{(n_r - 1)^2} \sum_{i \in A_r} (y_i - \bar{y}_r)^2 \right] \end{aligned}$$

$$\begin{aligned}
& - \frac{2n_m}{n_r - 1} \sum_{i \in A_r} (\bar{y}_1 - y_i)(y_i - \bar{y}_r) + \sum_{i \in A_m} (\eta_i - \bar{y}_1)^2 \Big] \\
& = \frac{1}{n} \left[s_1^2 + \frac{n_m(n+n_r-2)}{(n-1)(n_r-1)} s_r^2 \right].
\end{aligned}$$

Comparing this with (6.66), we conclude that

$$\tilde{v}_{\text{JACK}} / \text{var}(\bar{y}_1) \rightarrow_p 1.$$

We now consider the general case. Suppose that the population of ultimate units is divided into V imputation classes (which may cut across strata and clusters), and within class ν , all of the units have the same probability p_ν to be respondents. Hot deck imputation is then performed independently within each imputation class, using the respondents in the sample and in the imputation class. Let A_r^ν and A_m^ν contain the indices of respondents and nonrespondents, respectively, in imputation class ν . The missing value y_{hij} , $(h, i, j) \in A_m^\nu$, is imputed by η_{hij} selected with replacement from $\{y_{hij}, (h, i, j) \in A_r^\nu\}$ and with probability $p_{hij\nu} = w_{hij} / \sum_{(h,i,j) \in A_r^\nu} w_{hij}$. The imputed estimator of the population total Y_T is given by

$$\hat{Y}_1 = \sum_{\nu=1}^V \left(\sum_{(h,i,j) \in A_r^\nu} w_{hij} y_{hij} + \sum_{(h,i,j) \in A_m^\nu} w_{hij} \eta_{hij} \right).$$

Let

$$\bar{y}_{r\nu} = \sum_{(h,i,j) \in A_r^\nu} w_{hij} y_{hij} / \sum_{(h,i,j) \in A_r^\nu} w_{hij}. \quad (6.68)$$

Then,

$$\begin{aligned}
E_1(\hat{Y}_1) & = \sum_{\nu=1}^V \left[\sum_{(h,i,j) \in A_r^\nu} w_{hij} y_{hij} + \sum_{(h,i,j) \in A_m^\nu} w_{hij} E_1(\eta_{hij}) \right] \\
& = \sum_{\nu=1}^V \left[\sum_{(h,i,j) \in A_r^\nu} w_{hij} y_{hij} + \sum_{(h,i,j) \in A_m^\nu} w_{hij} \bar{y}_{r\nu} \right] \\
& = \sum_{\nu=1}^V \left(\sum_{(h,i,j) \in A_r^\nu \cup A_m^\nu} w_{hij} \right) \bar{y}_{r\nu}
\end{aligned} \quad (6.69)$$

and

$$\begin{aligned}
\text{var}_1(\hat{Y}_1) & = \sum_{\nu=1}^V \left[\sum_{(h,i,j) \in A_r^\nu} w_{hij}^2 \text{var}_1(\eta_{hij}) \right] \\
& = \sum_{\nu=1}^V \left(\sum_{(h,i,j) \in A_m^\nu} w_{hij}^2 \right) \sum_{(h,i,j) \in A_r^\nu} p_{hij\nu} (y_{hij} - \bar{y}_{r\nu})^2.
\end{aligned} \quad (6.70)$$

Let S_1^2 be the asymptotic variance of $E_{\text{I}}(\hat{Y}_{\text{I}})$ in (6.69), S_2^2 be the asymptotic mean of $\text{var}_{\text{I}}(\hat{Y}_{\text{I}})$ in (6.70), i.e.,

$$\begin{aligned} S_2^2 &= \sum_{\nu=1}^V E\left(\sum_{(h,i,j) \in A_m^\nu} w_{hij}^2\right) \left[E\left(\sum_{(h,i,j) \in A_r^\nu} w_{hij} y_{hij}^2\right) \Big/ E\left(\sum_{(h,i,j) \in A_r^\nu} w_{hij}\right) \right. \\ &\quad \left. - E\left(\sum_{(h,i,j) \in A_r^\nu} w_{hij} y_{hij}\right)\right]^2 \Big/ E\left(\sum_{(h,i,j) \in A_r^\nu} w_{hij}\right)^2, \end{aligned} \quad (6.71)$$

and $S^2 = S_1^2 + S_2^2$. It can be shown, by using Lemma 1 of Schenker and Welsh (1988), that

$$(\hat{Y}_{\text{I}} - Y_{\text{T}})/S \rightarrow_d N(0, 1). \quad (6.72)$$

The adjusted jackknife estimator of S^2 is obtained as follows. Let $\bar{y}_{r\nu,hi}$ be as defined in (6.68) but with the (h, i) th cluster removed. For $(h', i', j) \in A_m^\nu$ and $(h', i') \neq (h, i)$, define

$$\tilde{\eta}_{h'i'j}^{(hi)} = \eta_{h'i'j} + \bar{y}_{r\nu,hi} - \bar{y}_{r\nu},$$

$\nu = 1, \dots, V$. The adjusted jackknife estimator is

$$\tilde{v}_{\text{JACK}} = \sum_{h=1}^L \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} (\tilde{Y}_{\text{I},hi} - \tilde{Y}_{\text{I},h})^2, \quad (6.73)$$

where

$$\begin{aligned} \tilde{Y}_{\text{I},hi} &= \sum_{\nu=1}^V \left(\sum_{(h',i',j) \in A_r^\nu, h' \neq h} w_{h'i'j} y_{h'i'j} + \frac{n_h}{n_h - 1} \sum_{(h,i',j) \in A_r^\nu, i' \neq i} w_{hi'j} y_{hi'j} \right. \\ &\quad \left. + \sum_{(h',i',j) \in A_m^\nu, h' \neq h} w_{h'i'j} \tilde{\eta}_{h'i'j}^{(hi)} + \frac{n_h}{n_h - 1} \sum_{(h,i',j) \in A_m^\nu, i' \neq i} w_{hi'j} \tilde{\eta}_{hi'j}^{(hi)} \right) \end{aligned}$$

and

$$\tilde{Y}_{\text{I},h} = \frac{1}{n_h} \sum_{i=1}^{n_h} \tilde{Y}_{\text{I},hi}.$$

Theorem 6.8. Suppose that V is fixed; conditions (6.32)–(6.36) hold; condition (6.36) holds when z_{hij} is replaced by $y_{hij} a_{hij}$, where $a_{hij} = 1$ if y_{hij} is not missing and $a_{hij} = 0$ otherwise; and

$$\frac{1}{N_{\text{T}}} \sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} w_{hij} |y_{hij} - \bar{Y}|^{2+\tau} = O_p(1),$$

where $\bar{Y} = Y_{\text{T}}/N_{\text{T}}$ is the population mean. Then,

$$\tilde{v}_{\text{JACK}}/S^2 \rightarrow_p 1.$$

Proof. We only give a sketched proof for the case of $V = 1$ (one imputation class). The adjusted jackknife estimator can be expressed as

$$\tilde{v}_{\text{JACK}} = A + B + 2C$$

with

$$A = \sum_{h=1}^L \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} \left(\hat{N}_{T,h,i} \bar{y}_{r,h,i} - \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{N}_{T,h,i} \bar{y}_{r,h,i} \right)^2,$$

$$B = \sum_{h=1}^L \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(t_{hi} - \frac{1}{n_h} \sum_{i=1}^{n_h} t_{hi} \right)^2,$$

and

$$C = \sum_{h=1}^L \sum_{i=1}^{n_h} \hat{N}_{T,h,i} \bar{y}_{r,h,i} \left(t_{hi} - \frac{1}{n_h} \sum_{i=1}^{n_h} t_{hi} \right),$$

where $\hat{N}_{T,h,i}$ is \hat{N}_T based on the data with the (h, i) th cluster removed and $t_{hi} = \sum_{j=1}^{n_h} w_{hij}(1 - a_{hij})[\eta_{hij} - E_1(\eta_{hij})]$. From Theorem 6.1,

$$A/S_1^2 \rightarrow_p 1.$$

Note that

$$E_1(B) = \sum_{h=1}^L \sum_{i=1}^{n_h} \text{var}_1(t_{hi}) = \text{var}_1(\hat{Y}_1)$$

and

$$E_1(C) = \sum_{h=1}^L \sum_{i=1}^{n_h} \hat{N}_{T,h,i} \bar{y}_{r,h,i} E_1 \left(t_{hi} - \frac{1}{n_h} \sum_{i=1}^{n_h} t_{hi} \right) = 0.$$

Hence, by the law of large numbers,

$$B/S_2^2 \rightarrow_p 1 \quad \text{and} \quad C/(S_1^2 + S_2^2) \rightarrow_p 0. \quad \square$$

If another characteristic x_{hij} is observed for all ultimate units in the sample and $\hat{\theta}_1 = g(\hat{Y}_1, \hat{X})$, where $\hat{X} = \sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} w_{hij} x_{hij}$, the adjusted jackknife variance estimator for $\hat{\theta}_1$ is given by (6.73) with $\tilde{Y}_{1,h,i}$ replaced by $g(\tilde{Y}_{1,h,i}, \hat{X}_{hi})$. In particular, the choice $x_{hij} \equiv 1$ gives the adjusted jackknife variance estimator for \hat{Y}_1/\hat{N}_T , the imputed estimator of the population mean. If both x - and y -values are missing, and missing x - and y -values are imputed either marginally and independently or jointly, then the adjusted jackknife variance estimator for $\hat{\theta}_1 = g(\hat{Y}_1, \hat{X}_1)$ is given by (6.73) with $\tilde{Y}_{1,h,i}$ replaced by $g(\tilde{Y}_{1,h,i}, \hat{X}_{1,h,i})$. Consistency of the adjusted jackknife variance estimators for smooth g can be established similarly. The result

obviously can be extended to the case of more than two characteristics for each unit.

To implement this adjusted jackknife procedure, the data set must carry identification flags to locate nonrespondents, which is necessary if we want valid variance estimators under a single imputation.

6.5.3 Multiple bootstrap hot deck imputation

There exist non-resampling methods for variance estimation with imputed data. Multiple imputation, proposed by Rubin (1978) and further developed in Rubin (1987) and Rubin and Schenker (1986), can be described as follows. First, $m \geq 2$ independent imputations are performed to form m imputed data sets. Let $\hat{\theta}_{\text{II}}$ and v_{II} be the estimator of θ and the variance estimator for $\hat{\theta}_{\text{II}}$, respectively, computed using standard formulas for the case of no missing value and the l th imputed data set, $l = 1, \dots, m$. Then the imputed estimator of θ and its variance estimator based on multiple imputation are

$$\hat{\theta}_{\text{MI}}^{(m)} = \frac{1}{m} \sum_{l=1}^m \hat{\theta}_{\text{II}} \quad \text{and} \quad v_{\text{MI}}^{(m)} = \hat{W} + \frac{m+1}{m} \hat{B}, \quad (6.74)$$

respectively, where

$$\hat{W} = \frac{1}{m} \sum_{l=1}^m v_{\text{II}} \quad \text{and} \quad \hat{B} = \frac{1}{m-1} \sum_{l=1}^m (\hat{\theta}_{\text{II}} - \hat{\theta}_{\text{MI}}^{(m)})^2.$$

Note that \hat{W} is the average of within-imputation variances and \hat{B} is the between-imputation variance.

The implementation of the multiple imputation procedure does not need identification flags to locate nonrespondents, but requires that multiple complete data sets be maintained.

If the hot deck method described in Section 6.5.1 is used, however, $v_{\text{MI}}^{(m)}$ still leads to an underestimation of $\text{var}(\hat{\theta}_{\text{MI}}^{(m)})$. To adjust this underestimation, a bootstrap hot deck imputation method was suggested (Rubin and Schenker, 1986). First, draw the components of an n_r -dimensional vector \mathbf{Y}^* with replacement from the vector \mathbf{Y}_r of respondents. Then the vector \mathbf{Y}_m of missing values is imputed by a simple random sample from \mathbf{Y}^* . The same process is repeated independently m times and $\hat{\theta}_{\text{MI}}^{(m)}$ and $v_{\text{MI}}^{(m)}$ are calculated by using (6.74) based on the bootstrap hot deck imputed data sets.

Multiple bootstrap hot deck imputation can be justified from a Bayesian perspective (Rubin, 1987). It also has a frequentist (asymptotic) justification when the sampling design is single stage with no stratification.

Example 6.6 (continued). When multiple bootstrap hot deck imputation is used and $\hat{\theta} = \bar{y}$, it can be shown that, for fixed $m \geq 2$,

$$v_{\text{MI}}^{(m)} / \text{var}(\hat{\theta}_{\text{MI}}^{(m)}) \rightarrow_d U^{(m)} \quad (6.75)$$

and

$$(\hat{\theta}_{\text{MI}}^{(m)} - \theta) / \sqrt{v_{\text{MI}}^{(m)}} \rightarrow_d N(0, 1) / \sqrt{U^{(m)}}, \quad (6.76)$$

where

$$U^{(m)} = \frac{mp}{1-p+m} + \frac{(m+1)(1-p)}{1-p+m} \frac{\chi_{m-1}^2}{m-1}$$

and χ_{m-1}^2 is a chi-square random variable with $m-1$ degrees of freedom and independent of $N(0, 1)$. If we also let $m \rightarrow \infty$, then the right-hand sides of (6.75) and (6.76) tend to 1 and $N(0, 1)$, respectively.

From a frequentist perspective, however, there are at least two disadvantages of this method. First, result (6.76) is not very convenient to use in statistical inference. Rubin and Schenker (1986) suggested a further approximation by changing the right-hand side of (6.76) to a t-random variable with degrees of freedom given by

$$(m-1) \left[1 + \left(\frac{m}{m+1} \right) \frac{\hat{W}}{\hat{B}} \right]^2.$$

Second, similar results to (6.75) and (6.76) do not hold when the sampling design is multistage and/or stratified, even for large m , because the imputation cuts across strata and clusters (see Section 6.5.1). Fay (1991) showed the poor performance of $v_{\text{MI}}^{(m)}$ for the stratified one stage sampling case. If the imputation is within each cluster (or stratum), then multiple bootstrap hot deck imputation is expected to perform well when the cluster (or stratum) sizes are large.

Unlike the other methods we studied before, in multiple bootstrap hot deck imputation the bootstrap technique is applied as a device for imputing missing values, rather than a procedure for obtaining variance and distribution estimators. In the following, we show that if the bootstrap is applied to obtain variance and distribution estimators, then it produces asymptotically valid estimators.

6.5.4 Bootstrapping under imputation

Recall that the spirit of the bootstrap, summarized in diagram (1.25), is to mimic the sampling behavior of $(P, \mathbf{Y}, \hat{\theta})$ by using the sampling behavior of $(\hat{P}, \mathbf{Y}^*, \hat{\theta}^*)$, where P is the statistical model and \hat{P} is its estimate based on the data \mathbf{Y} . In the presence of missing values, P consists of two components:

(1) P_s , the population and the sampling design that produce the sample \mathbf{Y}_r ; and (2) P_r , the response mechanism that produces the respondents \mathbf{Y}_r . Thus, we have the following process of obtaining the imputed estimator $\hat{\theta}_I$:

$$\begin{array}{ccccccc} P_r & & & \text{Imputation} & & & \\ \downarrow & & & \downarrow & & & \\ P_s & \longrightarrow & \mathbf{Y}_r & \longrightarrow & \mathbf{Y}_I & \longrightarrow & \hat{\theta}_I, \end{array} \quad (6.77)$$

where \mathbf{Y}_I denotes the imputed data. Let \hat{P}_s and \hat{P}_r be estimates of P_s and P_r , respectively. Then the bootstrap analog $\hat{\theta}_I^*$ of $\hat{\theta}_I$ should be obtained by the process

$$\begin{array}{ccccccc} \hat{P}_r & & & \text{Imputation} & & & \\ \downarrow & & & \downarrow & & & \\ \hat{P}_s & \longrightarrow & \mathbf{Y}_r^* & \longrightarrow & \mathbf{Y}_I^* & \longrightarrow & \hat{\theta}_I^*. \end{array} \quad (6.78)$$

More precisely, $\hat{\theta}_I^*$ should be computed according to the following steps.

- (1) Obtain a bootstrap sample \mathbf{Y}^* from the imputed data set \mathbf{Y}_I as usual.
- (2) Use the identification flag attached to each unit in the bootstrap sample \mathbf{Y}^* to obtain \mathbf{Y}_r^* , the bootstrap analog of \mathbf{Y}_r .
- (3) Apply the same imputation technique used in constructing \mathbf{Y}_I to the units in the bootstrap sample to form \mathbf{Y}_I^* , the bootstrap analog of \mathbf{Y}_I .
- (4) Compute $\hat{\theta}_I^*$ based on the imputed bootstrap data set \mathbf{Y}_I^* by using the same formula used to compute $\hat{\theta}_I$.

Under the sampling design considered throughout this chapter, we can carry out steps (1)-(4) as follows. Let $a_{hij} = 1$ if the (h, i, j) th unit is a respondent and $a_{hij} = 0$ otherwise; and let $u_{hij} = (y_{hij}, a_{hij})$ if the (h, i, j) th unit is a respondent and $u_{hij} = (\eta_{hij}, a_{hij})$ otherwise, where η_{hij} is the imputed value for the missing respondent. First, we draw a simple random sample $\{\mathbf{u}_{hi}^*, i = 1, \dots, n_h - 1\}$ with replacement from $\{\mathbf{u}_{hi}, i = 1, \dots, n_h\}$, $h = 1, \dots, L$, independently across the strata, where $\mathbf{u}_{hi} = (u_{h1i}, \dots, u_{hni})$. Note that in each stratum the bootstrap sample size is $n_h - 1$, i.e., we employ the BWR or the BRS method with $m_h = n_h - 1$ introduced in Section 6.2.4. Let $u_{hij}^* = (y_{hij}^*, a_{hij}^*)$ be the j th component of \mathbf{u}_{hi}^* . Then the bootstrap data set \mathbf{Y}^* contains y_{hij}^* and \mathbf{Y}_r^* contains all of the y_{hij}^* corresponding to $a_{hij}^* = 1$. The imputed bootstrap data set \mathbf{Y}_I^* is constructed by imputing y_{hij}^* corresponding to $a_{hij}^* = 0$ using the same method used to obtain η_{hij} . If η_{hij} are obtained by hot deck imputation described earlier, then the imputed value η_{hij}^* is equal to $y_{hij}^* \in \mathbf{Y}_r^*$ with probability proportional to $w_{hij}^* = \tilde{w}_{hij} n_h / (n_h - 1)$, where \tilde{w}_{hij} is the original survey weight of the unit corresponding to y_{hij}^* . Finally, step (4) can be carried out by using the formula for $\hat{\theta}_I$ with y_{hij} , a_{hij} , and w_{hij}

replaced by y_{hij}^* , a_{hij}^* , and w_{hij}^* , respectively. For example, if $\hat{\theta}_i = \hat{Y}_i$, which can be written as

$$\sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} w_{hij} [a_{hij} y_{hij} + (1 - a_{hij}) \eta_{hij}],$$

then $\hat{\theta}_i^* = \hat{Y}_i^*$ can be computed according to

$$\sum_{h=1}^L \sum_{i=1}^{n_h-1} \sum_{j=1}^{n_{hi}} w_{hij}^* [a_{hij}^* y_{hij}^* + (1 - a_{hij}^*) \eta_{hij}^*].$$

Once the bootstrap analog $\hat{\theta}_i^*$ is obtained, the bootstrap variance and distribution estimators can be computed by using standard formulas, with Monte Carlo approximations if necessary. Under conditions similar to those in Theorem 6.8, these bootstrap estimators are consistent. We consider the special case of $\hat{\theta}_i = \hat{Y}_i$. Other cases can be treated similarly.

Theorem 6.9. *Assume the conditions in Theorem 6.8. Let H be the distribution of $\sqrt{n}(\hat{Y}_i - Y_i)$ and H_{BOOT} be the bootstrap estimator of H defined to be the conditional distribution of $\sqrt{n}(\hat{Y}_i^* - \hat{Y}_i)$ for given \mathbf{Y}_r and $\mathbf{A} = (a_{hij})$. Then*

$$\|H_{\text{BOOT}} - H\|_\infty \rightarrow_p 0.$$

Proof (sketched). Using Lemma 1 in Schenker and Welsh (1988), we can show an analog of result (6.72):

$$(\hat{Y}_i^* - \hat{Y}_i)/S^* \rightarrow_d N(0, 1)$$

(conditional on \mathbf{Y}_r and \mathbf{A}), where S^* is the bootstrap analog of the S in (6.72). Then the result can be proved by showing

$$S^*/S \rightarrow_p 1,$$

using results (6.69)–(6.71) and their bootstrap analogs. \square

Compared with the adjusted jackknife, which also provides a consistent estimator of S^2 (Theorem 6.8), the bootstrap method discussed here does not require any adjustment, but it requires some extra computations since each bootstrap sample has to be imputed. From diagrams (6.77) and (6.78), however, the bootstrap method seems to be more natural and is of higher potential to be extended to other complex problems than the adjusted jackknife (e.g., Efron, 1994). Another advantage of this bootstrap method over the adjusted jackknife is that it is applicable to the case of $\hat{\theta}$ being sample quantiles.

6.6 Conclusions and Discussions

- (1) We studied three resampling methods, the jackknife, the balanced repeated replication (BRR), and the bootstrap, for variance or distribution estimation for smooth functions of weighted averages or sample quantiles based on survey data obtained by stratified multistage sampling. The application of the jackknife, originally used in the i.i.d. case, requires a modification (to account for the sampling fractions) only when the first stage sampling is without replacement [see (6.12)]. The bootstrap method, however, requires some modifications even when the first stage sampling is with replacement. Four modified bootstraps are discussed in Section 6.2.4. The BRR method is specially designed for variance estimation in stratified sampling.
- (2) Throughout this chapter, the primary units for resampling are the first stage sample clusters, rather than the ultimate sample units. This is suitable for survey problems with a large number of first stage clusters. With some minor modifications, we can apply these resampling methods by resampling the k th stage sample clusters for any $k \geq 2$ (e.g., Rao and Wu, 1988; Sitter, 1992a). Applications of the jackknife under two-phase sampling can be found in Rao and Sitter (1994). Applications of the bootstrap under systematic sampling can be found in Kuk (1987, 1989).
- (3) The jackknife method is applicable to the case of variance estimation for functions of weighted averages and is asymptotically equivalent to the linearization method. Both jackknife and linearization variance estimators for functions of weighted averages are consistent and perform well in empirical studies. Hence, the choice between the jackknife and the linearization depends on nonstatistical considerations. Applications of the jackknife for other statistics can be found in Shao (1994b). The BRR variance estimators can be used for both functions of weighted averages and sample quantiles, and they are consistent, although they are not closer to the linearization variance estimator than the jackknife variance estimator. The bootstrap can be applied to both variance and distribution estimation problems. However, the bootstrap variance estimator is not as good as the jackknife or the BRR variance estimator in terms of the empirical results. Furthermore, the bootstrap variance estimator usually requires more computations than the jackknife or the BRR. Thus, the bootstrap is mainly recommended for distribution estimation.
- (4) Unlike in the i.i.d. case, a general theory for the bootstrap confidence sets is not available, although some empirical results showed that the bootstrap one-sided confidence intervals are better than those constructed by normal approximation.

- (5) The construction of the balanced samples for the BRR is easy when $n_h = 2$ for all h (two clusters per stratum). In the general case of unequal n_h , the construction of the balanced samples may be difficult. Some approximated BRR are introduced in Section 6.2.3. In particular, the performances of the repeatedly grouped BRR and the random subsampling BRR are good in terms of both asymptotic and empirical results.
- (6) We also studied problems with imputed missing data. A consistent jackknife variance estimator is derived by adjusting the imputed values for each pseudoreplicate and then applying the standard jackknife formula. An asymptotically valid bootstrap procedure is obtained by imputing the bootstrap data set (in the same way as the original data set is imputed) and then applying the standard bootstrap formula. Both the jackknife and the bootstrap methods require identification flags to the missing values. The bootstrap can be applied in multiple imputation, which does not require identification flags; however, the multiple imputation method may not be asymptotically valid in stratified multistage sampling and requires that multiple complete data sets be maintained.

Chapter 7

Applications to Linear Models

In this chapter, we study the application of the jackknife and the bootstrap to linear models. Section 7.1 describes the forms of linear models and estimation of regression parameters in the models. Variance and bias estimation for the estimators of regression parameters are considered in Section 7.2. Other statistical inferences and analyses based on the bootstrap are discussed in Section 7.3. Model selection using cross-validation (jackknife) or the bootstrap is studied in Section 7.4. Most of the technical details and rigorous proofs of the asymptotic results are deferred to Section 7.5.

7.1 Linear Models and Regression Estimates

Statistical applications are often based on some statistical models. One of the most useful models is the following general linear model:

$$y_i = x_i' \beta + \varepsilon_i, \quad i = 1, \dots, n, \tag{7.1}$$

where y_i is the i th response, x_i is a p -vector of explanatory variables associated with y_i , β is a p -vector of unknown parameters, and ε_i is the random error. Problems like linear regression and one-way and two-way analysis of variance can all be formulated as special cases of model (7.1). Examples of data from model (7.1) can be found in standard text books, for example, Draper and Smith (1981) and Searle (1971).

When the x_i in model (7.1) are deterministic, we assume that the ε_i are independent and have mean 0 and unknown variances σ_i^2 . When the x_i are random, model (7.1) is sometimes referred to as a correlation model. In

a correlation model, the (y_i, x'_i) are assumed to be i.i.d. with finite second order moments and $E(y_i|x_i) = x'_i\beta$; σ_i^2 denotes the conditional variance of y_i given x_i , i.e., $\text{var}(y_i|x_i) = \sigma_i^2$; the conditional expectation $E(\cdot|x_i)$ and the conditional variance $\text{var}(\cdot|x_i)$ will be simplified to E and var , respectively, throughout this chapter.

The parameter of main interest is β , which is referred to as the regression parameter even though the problem under consideration may not be a regression problem. Statistical analysis mainly consists of the estimation of β (model fitting), model testing, model selection, and prediction.

We now describe some commonly used estimators of β . Other statistical inference, prediction, and model selection problems will be described in Sections 7.2–7.4, along with discussions of the jackknife and bootstrap procedures.

It is often assumed that the ε_i have a common variance; that is, $\sigma_i^2 = \sigma^2$ for all i . Under the equal variance assumption, the parameter β in (7.1) is customarily estimated by the least squares estimator (LSE) defined by

$$\hat{\beta}_{\text{LS}} = (X'X)^{-1}X'y, \quad (7.2)$$

where $X' = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)'$. The matrix X is assumed to be of rank p for simplicity; otherwise, $(X'X)^{-1}$ in (7.2) can be replaced by a generalized inverse of $X'X$.

It is easy to see that $E(\hat{\beta}_{\text{LS}}) = \beta$, i.e., $\hat{\beta}_{\text{LS}}$ is unbiased. In general, $\hat{\beta}_{\text{LS}}$ has the minimum variance among the unbiased estimators of β that are linear in y . When $(\varepsilon_1, \dots, \varepsilon_n)' \sim N(0, \sigma^2 I_n)$, $\hat{\beta}_{\text{LS}}$ is the maximum likelihood estimator and the uniformly minimum variance unbiased estimator of β .

There exist other estimators of β . When the equal variance assumption is violated, i.e., σ_i^2 depends on i , a weighted least squares estimator (WLSE)

$$\hat{\beta}_{\text{WLS}} = (X'WX)^{-1}X'Wy \quad (7.3)$$

may be more efficient than the LSE, where W is a diagonal matrix whose i th diagonal element w_i is proportional to an estimate of σ_i^{-2} . For example, if we know that $\sigma_i^2 = \sigma_0^2 \|x_i\|$, $i = 1, \dots, n$, where σ_0 is an unknown scalar, then we may take $w_i = \|x_i\|^{-1}$.

For robustness against outliers in the sample, one may consider an M-estimate of β and σ defined as a solution of

$$\sum_{i=1}^n x_i \psi\left(\frac{y_i - x'_i \beta}{\sigma}\right) = 0 \quad (7.4)$$

for some function ψ (Huber, 1981). Regression L_1 -norm estimates or quantiles can also be used (Koenker and Bassett, 1978; Koenker and Portnoy, 1987; Ruppert and Carroll, 1980).

Assuming that $\sigma_i^2 = \sigma^2$, we obtain that

$$\text{var}(\hat{\beta}_{\text{LS}}) = (X'X)^{-1}X'\text{var}(y)X(X'X)^{-1} = \sigma^2(X'X)^{-1}. \quad (7.5)$$

A customary unbiased estimator of $\text{var}(\hat{\beta}_{\text{LS}})$ is

$$\left(\frac{1}{n-p} \sum_{i=1}^n r_i^2 \right) (X'X)^{-1}, \quad (7.6)$$

where $r_i = y_i - x_i' \hat{\beta}_{\text{LS}}$ is the i th residual.

The variances and biases of other estimators of β do not have exact forms. Hence, the traditional approach requires that these variances and biases be estimated by asymptotic approximations.

7.2 Variance and Bias Estimation

In many situations, one needs to estimate $\theta = g(\beta)$ for a function g from \mathbb{R}^p to \mathbb{R} . If β is estimated by $\hat{\beta}$, then θ is estimated by $\hat{\theta} = g(\hat{\beta})$. In this section, we study applications of the jackknife and bootstrap for estimating the variance and bias of $\hat{\theta}$.

7.2.1 Weighted and unweighted jackknives

Miller (1974) first extended the jackknife to the estimation of the variance and bias of $\hat{\theta}_{\text{LS}} = g(\hat{\beta}_{\text{LS}})$. The extension is quite straightforward: delete the pair (y_i, x_i') and calculate $\hat{\theta}_{\text{LS},i}$, the LSE of θ based on the rest of the data set, $i = 1, \dots, n$, and then estimate the variance and bias of $\hat{\theta}_{\text{LS}}$ by

$$v_{\text{JACK}} = \frac{n-1}{n} \sum_{i=1}^n \left(\hat{\theta}_{\text{LS},i} - \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{\text{LS},i} \right)^2 \quad (7.7)$$

and

$$b_{\text{JACK}} = \frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_{\text{LS},i} - \hat{\theta}_{\text{LS}}), \quad (7.8)$$

respectively. Using the fact that

$$(A + cc')^{-1} = A^{-1} - A^{-1}cc'A^{-1}/(1 + c'A^{-1}c)$$

for a matrix A and a vector c , we conclude that

$$\hat{\beta}_{\text{LS},i} = \hat{\beta}_{\text{LS}} - z_i r_i / (1 - h_i), \quad (7.9)$$

where $z_i = (X'X)^{-1}x_i$, $r_i = y_i - x_i'\hat{\beta}_{\text{LS}}$ is the i th residual, and $h_i = x_i'(X'X)^{-1}x_i$. Consequently, in the case of $g(\beta) = c'\beta$ for a fixed $c \in \mathbb{R}^p$,

$$v_{\text{JACK}} = \frac{n-1}{n} \sum_{i=1}^n \frac{(c'z_i)^2 r_i^2}{(1-h_i)^2} - \frac{n-1}{n^2} \left(\sum_{i=1}^n \frac{c'z_i r_i}{1-h_i} \right)^2 \quad (7.10)$$

and

$$b_{\text{JACK}} = \frac{n-1}{n} \sum_{i=1}^n \frac{c'z_i r_i}{1-h_i} = \frac{n-1}{n} \sum_{i=1}^n \frac{c'z_i h_i r_i}{1-h_i}, \quad (7.11)$$

where the last equality holds because of

$$\sum_{i=1}^n z_i r_i = 0. \quad (7.12)$$

While Miller (1974) showed that v_{JACK} is a consistent variance estimator under fairly weak conditions, the bias estimator b_{JACK} may be inconsistent. Note that $c'\hat{\beta}_{\text{LS}}$ is unbiased for $c'\beta$; but (7.11) indicates that b_{JACK} is not exactly 0 if the h_i are different and may not vanish to 0 fast enough. More precisely, it is shown in Section 7.5.2 that

$$\frac{n-1}{n} \sum_{i=1}^n \frac{c'z_i h_i r_i}{1-h_i} = O_p \left(h_{\max} [\text{bias}(\hat{\theta}_{\text{LS}})]^{1/2} \right),$$

where

$$h_{\max} = \max_{i \leq n} h_i. \quad (7.13)$$

Thus, b_{JACK} is consistent if $h_{\max} = o([\text{bias}(\hat{\theta}_{\text{LS}})]^{1/2})$ and may be inconsistent if this condition is not satisfied. This result also holds for the general case of $\hat{\theta}_{\text{LS}} = g(\hat{\beta}_{\text{LS}})$.

The inconsistency of the jackknife bias estimator is caused by the imbalance of model (7.1), i.e., the $E(y_i) = x_i'\beta$ are not the same. The imbalance of model (7.1) can be assessed by imbalance measures such as the h_{\max} in (7.13) and h_i , $i = 1, \dots, n$. The smaller h_{\max} is, the better performance the jackknife estimators have.

In view of the unbalanced nature of the model, some *weighted jackknife* procedures were proposed to provide some improvements. Hinkley (1977) defined weighted pseudovalues

$$\tilde{\theta}_i = \hat{\theta}_{\text{LS}} + n(1-h_i)(\hat{\theta}_{\text{LS}} - \hat{\theta}_{\text{LS},i}), \quad i = 1, \dots, n,$$

which lead to the following weighted jackknife variance and bias estimators:

$$v_{\text{HJACK}} = \frac{1}{n(n-p)} \sum_{i=1}^n \left(\tilde{\theta}_i - \frac{1}{n} \sum_{i=1}^n \tilde{\theta}_i \right)^2 \quad (7.14)$$

and

$$b_{\text{HJACK}} = \frac{1}{n} \sum_{i=1}^n (\hat{\theta}_{\text{LS}} - \tilde{\theta}_i). \quad (7.15)$$

For $\hat{\theta}_{\text{LS}} = c' \hat{\beta}_{\text{LS}}$,

$$v_{\text{HJACK}} = \frac{n}{n-p} \sum_{i=1}^n (c' z_i)^2 r_i^2, \quad (7.16)$$

and it follows from (7.9), (7.12), and (7.15) that $b_{\text{HJACK}} \equiv 0$ and the weighted jackknife provides a correct bias estimator.

Motivated by the representation

$$\hat{\beta}_{\text{LS}} = \sum_{i=1}^n \frac{1-h_i}{n-p} \hat{\beta}_{\text{LS},i},$$

Wu (1986) proposed the following weighted jackknife variance and bias estimators:

$$v_{\text{WJACK}} = \sum_{i=1}^n (1-h_i) (\hat{\theta}_{\text{LS},i} - \hat{\theta}_{\text{LS}})^2 \quad (7.17)$$

and

$$b_{\text{WJACK}} = \sum_{i=1}^n (1-h_i) (\hat{\theta}_{\text{LS},i} - \hat{\theta}_{\text{LS}}).$$

Note that $b_{\text{WJACK}} \equiv b_{\text{HJACK}}$, but v_{WJACK} and v_{HJACK} are generally different. Wu (1986) also proposed weighted delete-d jackknife estimators. For $\hat{\theta}_{\text{LS}} = c' \hat{\beta}_{\text{LS}}$,

$$v_{\text{WJACK}} = \sum_{i=1}^n \frac{(c' z_i)^2 r_i^2}{1-h_i}, \quad (7.18)$$

and v_{WJACK} reduces to v_{HJACK} if the h_i in (7.18) are replaced by their average p/n .

It is shown in Section 7.5.1 that both weighted jackknife methods provide consistent variance and bias estimators. Some comparisons among the two weighted jackknife estimators v_{HJACK} and v_{WJACK} and the unweighted jackknife estimator v_{JACK} are also given. In general, v_{JACK} is slightly upward-biased and v_{HJACK} is slightly downward-biased; the order of the bias of v_{WJACK} is always no larger than that of v_{JACK} or v_{HJACK} ; and the differences among the three jackknife estimators become negligible when h_{\max} is small. Finite sample properties of these jackknife variance estimators are studied by simulation in Wu (1986) (see Section 7.2.3). Here, we consider an example for illustration.

Example 7.1. Two-sample problem. Suppose that

$$y_{ij} = \beta_j + \varepsilon_{ij}, \quad i = 1, \dots, n_i, \quad j = 1, 2,$$

where the ε_{ij} are independent with $E\varepsilon_{ij} = 0$ and $E\varepsilon_{ij}^2 = \sigma_j^2$. The LSE of β_j is $\bar{y}_j = \sum_{i=1}^{n_j} y_{ij}/n_j$ with variance σ_j^2/n_j , $j = 1, 2$. Let $n = n_1 + n_2$. Then, for \bar{y}_j ,

$$v_{\text{JACK}} = \frac{(n-1)\hat{\sigma}_j^2}{n(n_j-1)}, \quad v_{\text{HJACK}} = \frac{n(n_j-1)\hat{\sigma}_j^2}{(n-2)n_j^2}, \quad \text{and} \quad v_{\text{WJACK}} = \frac{\hat{\sigma}_j^2}{n_j},$$

where $\hat{\sigma}_j^2 = \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2/(n_j - 1)$. Note that v_{WJACK} is the same as the customary variance estimator for the case of unequal variances and is exactly unbiased. The biases for v_{JACK} and v_{HJACK} are

$$\text{bias}(v_{\text{JACK}}) = \frac{(n-n_j)\sigma_j^2}{nn_j(n_j-1)} \quad \text{and} \quad \text{bias}(v_{\text{HJACK}}) = \frac{(2n_j-n)\sigma_j^2}{(n-2)n_j^3}.$$

Thus, in terms of their biases, v_{WJACK} is better than v_{HJACK} , and v_{HJACK} is better than v_{JACK} .

Since all three jackknife variance estimators are proportional to $\hat{\sigma}_j^2$, comparisons of their variances can be made by comparing the coefficients in front of $\hat{\sigma}_j^2$. In terms of the variances, v_{JACK} is always worse than v_{WJACK} , since $\frac{n-1}{n(n_j-1)} > \frac{1}{n_j}$. Note that $\frac{n(n_j-1)}{(n-2)n_j} \leq 1$ if and only if $2n_j \leq n$. Therefore, $\text{var}(v_{\text{HJACK}}) \leq \text{var}(v_{\text{WJACK}})$ for one j and $\text{var}(v_{\text{HJACK}}) \geq \text{var}(v_{\text{WJACK}})$ for the other j .

It seems that the unweighted jackknife estimator can be improved by multiplying v_{JACK} by $(n-2)/(n-1)$, which leads to

$$\text{bias}(v_{\text{JACK}}) = \frac{(n-2n_j)\sigma_j^2}{nn_j(n_j-1)},$$

so that v_{JACK} is comparable to v_{HJACK} .

Two issues in this example deserve to be mentioned. First, for the unweighted jackknife variance estimator, it is probably better to change the factor $\frac{n-1}{n}$ in (7.7) to $\frac{n-p}{n}$. Second, the jackknife variance estimators work well even when the σ_j^2 are unequal. This robustness (against unequal variances) property of the jackknife will be explored further in Section 7.2.3.

The applications of the jackknife for other estimators of θ , such as the WLSE, the M-estimator, and the L_1 -norm estimator, are similar, but it is not clear in general how to construct the weights for the weighted jackknife.

For the WLSE in (7.3), the matrix of weights, W , may have to be updated when calculating $\hat{\beta}_{\text{WLS},i}$ for each i ; that is,

$$\hat{\beta}_{\text{WLS},i} = \left(\sum_{j \neq i} w_{ij} x_j x_j' \right)^{-1} \sum_{j \neq i} x_j w_{ij} y_j,$$

where w_{ij} is the j th weight (estimate of σ_j^{-2}) after deleting the i th pair (y_i, x'_i) , not simply w_j . If the w_j , instead of the w_{ij} , are used, the resulting jackknife variance estimator may be inconsistent, unless one can show that the differences $w_j - w_{ij}$ are very small (Shao, 1989c; Shao and Rao, 1993).

7.2.2 Three types of bootstraps

Bootstrap variance and bias estimators are defined as in the i.i.d. case. Let $\hat{\theta}^*$ be the bootstrap analog of $\hat{\theta}$ based on bootstrap data. The bootstrap variance and bias estimators for $\hat{\theta}$ are

$$v_{\text{BOOT}} = \text{var}_*(\hat{\theta}^*) \quad \text{and} \quad b_{\text{BOOT}} = E_*\hat{\theta}^* - \hat{\theta},$$

respectively, where E_* and var_* are the expectation and variance with respect to bootstrap sampling. Monte Carlo or other methods described in Chapter 5 can be applied to compute bootstrap estimators when they have no explicit form. Under model (7.1), however, there are different ways to generate bootstrap data.

The bootstrap based on residuals

The *bootstrap based on residuals*, abbreviated as RB, was proposed by Efron (1979) and has been described in Example 1.7. Following the general idea of the bootstrap illustrated by diagram (1.25), we first identify the model by a “parameter” and then generate bootstrap data from the model with the parameter replaced by its estimate based on the original data. When the x_i are nonrandom and the ε_i are i.i.d., model (7.1) can be identified as (β, F_ε) , where F_ε is the unknown distribution of ε_i . Let β be estimated by $\hat{\beta}_{\text{LS}}$ and F_ε be estimated by the empirical distribution \hat{F}_ε putting mass n^{-1} to $r_i - \bar{r}$, $i = 1, \dots, n$, where $r_i = y_i - x'_i \hat{\beta}_{\text{LS}}$ is the i th residual and $\bar{r} = n^{-1} \sum_{i=1}^n r_i$. Note that \hat{F}_ε is centered at 0, since F_ε has mean 0. The bootstrap data can then be generated from the model with (β, F_ε) replaced by $(\hat{\beta}_{\text{LS}}, \hat{F}_\varepsilon)$. A convenient way to do this is to generate i.i.d. data $\varepsilon_1^*, \dots, \varepsilon_n^*$ from \hat{F}_ε and define $y_i^* = x'_i \hat{\beta}_{\text{LS}} + \varepsilon_i^*$. The bootstrap analog of $\hat{\beta}_{\text{LS}}$ is then $\hat{\beta}_{\text{LS}}^* = g(\hat{\beta}_{\text{LS}})$ with $\hat{\beta}_{\text{LS}}^*$ being the LSE of β calculated based on the data $(y_1^*, x'_1), \dots, (y_n^*, x'_n)$, i.e.,

$$\hat{\beta}_{\text{LS}}^* = (X'X)^{-1}X'y^*, \quad (7.19)$$

where $y^* = (y_1^*, \dots, y_n^*)'$. Since $E_*(\varepsilon_i^*) = 0$ for all i ,

$$E_*(\hat{\beta}_{\text{LS}}^*) = (X'X)^{-1}X'E_*(y^*) = \hat{\beta}_{\text{LS}}. \quad (7.20)$$

Also,

$$\text{var}_*(\hat{\beta}_{\text{LS}}^*) = (X'X)^{-1}X'\text{var}_*(y^*)X(X'X)^{-1} = \hat{\sigma}^2(X'X)^{-1}, \quad (7.21)$$

where $\hat{\sigma}^2 = \text{var}_*(\varepsilon_i^*) = n^{-1} \sum_{i=1}^n (r_i - \bar{r})^2$. From (7.20)-(7.21), this bootstrap procedure produces consistent variance and bias estimators for $\hat{\beta}_{\text{LS}}$ and, consequently, for $\hat{\theta}_{\text{LS}}$ with smooth function g under some moment conditions (see, e.g., Shao, 1988c).

An important assumption for the RB is that the ε_i are i.i.d. Even if this assumption holds, the empirical distribution \hat{F}_ε is not based on exactly i.i.d. data. The estimator in (7.21) has an explicit form but is not the same as the customary unbiased estimator given in (7.6). Also, the estimator in (7.21) is downward-biased, since

$$E(\hat{\sigma}^2) = (1 - \frac{k_n}{n})\sigma^2,$$

where $k_n = p + 1 - n^{-1} \sum_{i=1}^n \sum_{j=1}^n x_i' (X' X)^{-1} x_j \geq p > 0$. To remove this negative bias, Efron (1982) suggested that bootstrap data be drawn from the empirical distribution based on $(r_i - \bar{r})/\sqrt{1 - p/n}$, $i = 1, \dots, n$. This adjustment, however, still leads to a downward-biased variance estimator since $k_n > p$ when $\bar{r} \neq 0$. We suggest that bootstrap data be generated from the empirical distribution \tilde{F}_ε putting mass n^{-1} to the adjusted residual $(r_i - \bar{r})/\sqrt{1 - k_n/n}$, $i = 1, \dots, n$. Then (7.21) holds with $\hat{\sigma}^2$ replaced by

$$\tilde{\sigma}^2 = \frac{1}{n - k_n} \sum_{i=1}^n (r_i - \bar{r})^2. \quad (7.22)$$

If $\bar{r} = 0$ (e.g., the first component of x_i is 1 for all i), then $k_n = p$, $\tilde{\sigma}^2 = (n - p)^{-1} \sum_{i=1}^n r_i^2$ and the bootstrap variance estimator for $\hat{\beta}_{\text{LS}}$ is the same as the customary unbiased variance estimator in (7.6). The adjustment $1 - k_n/n$ has no substantial impact on variance estimation if n is very large. However, in Section 7.3.4 we show that this adjustment has a significant effect on bootstrap estimation of prediction errors.

The RB can also be applied to cases where β is estimated by using other estimators. For M-estimators, $\hat{\beta}_M$ and $\hat{\sigma}_M$, that are solutions of the equations defined by (7.4), Shorack (1982) proposed that the bootstrap data $\varepsilon_1^*, \dots, \varepsilon_n^*$ be generated from the empirical distribution putting mass n^{-1} to each $(y_i - x_i' \hat{\beta}_M)/\hat{\sigma}_M$, $i = 1, \dots, n$, and that $\hat{\beta}_M^*$, the bootstrap analog of $\hat{\beta}_M$, be then obtained by solving

$$\sum_{i=1}^n x_i \left[\psi \left(\frac{y_i^* - x_i' \beta}{\hat{\sigma}_M} \right) - \frac{1}{n} \sum_{j=1}^n \psi \left(\frac{y_j - x_j' \hat{\beta}_M}{\hat{\sigma}_M} \right) \right] = 0,$$

where $y_i^* = x_i' \hat{\beta}_M + \hat{\sigma}_M \varepsilon_i^*$. This bootstrap procedure has been shown to produce consistent bootstrap estimators. Similar results for other types of M-estimators and L_1 -norm estimators can be found in He (1987), Stangenhaus (1987), and DeAngelis, Hall and Young (1993).

The paired bootstrap

The *paired bootstrap*, abbreviated as PB, seems to be a very natural procedure when the x_i are random and (y_i, x'_i) , $i = 1, \dots, n$, are i.i.d. In this case, the model can be identified by the joint distribution of (y_i, x'_i) and estimated by the empirical distribution function putting mass n^{-1} to (y_i, x'_i) , $i = 1, \dots, n$. The bootstrap data are generated from this empirical distribution. This bootstrap method is very similar to what we have studied in the previous chapters, and its application is obviously not restricted to the LSE.

Note that the bootstrap analog X^* of X may not be of full rank even if X is. Although we may use any generalized inverse of $X^{*'}X^*$ in computing $\hat{\beta}_{\text{LS}}^*$, it is computationally more convenient to adopt the following modification:

$$\hat{\beta}_{\text{LS}}^* = \begin{cases} (X^{*'}X^*)^{-1}X^{*'}y^* & \text{if } \lambda^* \geq \lambda/2 \\ \hat{\beta}_{\text{LS}} & \text{otherwise,} \end{cases} \quad (7.23)$$

where λ^* and λ are the smallest eigenvalues of $X^{*'}X^*$ and $X'X$, respectively. Since $P_*\{\lambda^* < \lambda/2\}$ tends to 0 at an exponential rate, this modification is very minor, but it simplifies the theoretical studies of the bootstrap estimators. It can be shown that $E_*(\hat{\beta}_{\text{LS}}^*) \approx \hat{\beta}_{\text{LS}}$ and $\text{var}_*(\hat{\beta}_{\text{LS}}^*)$ is approximately equal to

$$(X'X)^{-1}\text{var}_*\left[\sum_{i=1}^n x_i^*(y_i^* - x_i^{*'}\hat{\beta}_{\text{LS}})\right](X'X)^{-1} = \sum_{i=1}^n z_i z'_i r_i^2,$$

which, apart from a factor of $\frac{n}{n-p}$, is the same as Hinkley's weighted jackknife variance estimator (7.16). Hence, the variance and bias estimators for $\hat{\theta}_{\text{LS}}$ based on the PB are consistent, under some moment conditions.

As a counterpart of the weighted jackknife, Wu (1986) also proposed a *weighted bootstrap variance estimator* for $\hat{\theta}_{\text{LS}}$:

$$E_*[w^*(\hat{\theta}_{\text{LS}}^* - \hat{\theta}_{\text{LS}})^2]/E_*(w^*),$$

where w^* is the determinant of $X^{*'}X^*$ and $\hat{\theta}_{\text{LS}}^* = g(\hat{\beta}_{\text{LS}}^*)$. The choice of the weights w^* was motivated by the fact that $\hat{\beta}_{\text{LS}} = E_*(w^*\hat{\beta}_{\text{LS}}^*)/E_*(w^*)$ (Wu, 1986, Theorem 2). This weighted bootstrap has a better finite sample performance than the unweighted bootstrap (see the simulation results in Section 7.2.3).

The external bootstrap

This method was first proposed by Wu (1986) for the LSE. Let e_i^* , $i = 1, \dots, n$, be i.i.d. from a distribution with mean 0 and variance 1. Then the

bootstrap data are generated by setting

$$y_i^* = x_i' \hat{\beta}_{\text{LS}} + \frac{|r_i|}{\sqrt{1-h_i}} e_i^*. \quad (7.24)$$

The bootstrap analog of $\hat{\beta}_{\text{LS}}$ is $\hat{\beta}_{\text{LS}}^* = (X'X)^{-1} X' y^*$. Since the distribution of e_i^* can be independent of the original data, this procedure is called *external bootstrap* (EB) or *wild bootstrap*. For the second order accuracy of the bootstrap estimators based on this method, Liu (1988) suggested that another restriction be imposed on e_i^* : $E_*(e_i^{*3}) = 1$.

For $\hat{\beta}_{\text{LS}}$, it is easy to see that $E_*(\hat{\beta}_{\text{LS}}^*) = \hat{\beta}_{\text{LS}}$ and

$$\text{var}_*(\hat{\beta}_{\text{LS}}^*) = (X'X)^{-1} \sum_{i=1}^n x_i x_i' \frac{r_i^2}{1-h_i} \text{var}_*(e_i^*)(X'X)^{-1},$$

which provides the same variance estimator (7.18) as Wu's (1986) weighted jackknife. Thus, the variance and bias estimators based on this bootstrap procedure are also consistent, and the variance estimator has behavior similar to that of the weighted jackknife estimator v_{WJACK} .

7.2.3 Robustness and efficiency

As we have discussed in the previous chapters, the jackknife and bootstrap estimators may not have any apparent superiority over the traditional estimators in terms of asymptotic performance. However, in many cases the traditional method rests upon some model assumptions that are not explicitly required by the jackknife or the bootstrap. Therefore, the performances of the jackknife and bootstrap estimators are less susceptible to violation of the model assumptions. This robustness property of the jackknife and bootstrap has been recognized by many researchers (Hinkley, 1977; Wu, 1986; Shao and Wu, 1987; Shao, 1988a; Liu and Singh, 1992a).

Under model (7.1), the LSE is based on the assumption that $\sigma_i^2 = \sigma^2$ for all i . This equal variance assumption, however, can be violated in many practical problems. Furthermore, it is often difficult to verify this assumption. It is well known that, in the general case where σ_i^2 are unequal,

$$\text{var}(\hat{\beta}_{\text{LS}}) = (X'X)^{-1} X' \text{var}(y) X (X'X)^{-1} = \sum_{i=1}^n \sigma_i^2 z_i z_i',$$

which reduces to (7.5) when $\sigma_i^2 = \sigma^2$ for all i . Note that $\hat{\beta}_{\text{LS}}$ is still consistent and asymptotically normal in the unequal variance case under mild conditions. This stable performance of an estimator designed primarily for the equal variance case is called *robustness against heteroscedasticity*. We now study the robustness against heteroscedasticity of the jackknife

and bootstrap variance estimators for the LSE. The bias estimators can be studied similarly.

The weighted and unweighted jackknife are not based on the equal variance assumption and, therefore, are robust against heteroscedasticity. More precisely, irrespective of whether the σ_i^2 are equal or not, the weighted and unweighted jackknife variance estimators are consistent and asymptotically unbiased for the variance or the asymptotic variance of the LSE. This has been shown in Example 7.1 and can be heuristically shown as follows. Under some moment conditions,

$$r_i = \sqrt{1 - h_i} \varepsilon_i + o_p(1).$$

Hence, in the case of $\hat{\theta}_{\text{LS}} = c' \hat{\beta}_{\text{LS}}$,

$$v_{\text{WJACK}} = \sum_{i=1}^n (c' z_i)^2 [\varepsilon_i + o_p(1)]^2 = \text{var}(c' \hat{\beta}_{\text{LS}}) + o_p(\text{var}(c' \hat{\beta}_{\text{LS}})),$$

assuming that the σ_i^2 are bounded away from 0 and ∞ . Similar results hold for v_{HJACK} and v_{JACK} when h_{\max} in (7.13) tends to 0. More rigorous treatment is given in Section 7.5.1.

Robustness of the bootstrap procedures described in Section 7.2.2 rests upon whether equal variance is assumed in the process of generating bootstrap data. It is clear that the EB is not based on the equal variance assumption. An important assumption for the PB is that (y_i, x'_i) are i.i.d.; $\sigma_i^2 = \sigma^2$ is not crucial. Hence, these two bootstrap procedures are robust against heteroscedasticity. In fact, we argued in Section 7.2.2 that the EB variance estimator is close to v_{WJACK} and the PB variance estimator is close to v_{HJACK} . The RB, however, does not provide a robust variance estimator, since it rests upon the i.i.d. assumption of the errors ε_i . It is clear that the estimator in (7.21), which is almost the same as the traditional variance estimator for the LSE based on the equal variance assumption, cannot be valid when the σ_i^2 are unequal.

However, nonrobust estimators are usually more efficient than robust estimators when the model assumptions are valid. Liu and Singh (1992a) showed that, for $\hat{\theta}_{\text{LS}} = g(\hat{\beta}_{\text{LS}})$, the ratio of the asymptotic variance of the variance estimator based on the RB over that of the variance estimator based on the jackknife or the other two bootstrap methods is

$$\left\{ \sum_{i=1}^n [\nabla g(\beta)' z_i]^2 \right\}^2 / \left\{ n \sum_{i=1}^n [\nabla g(\beta)' z_i]^4 \right\},$$

which is always less than or equal to 1 by the Cauchy-Schwarz inequality.

In Example 7.1, we studied finite sample properties of the weighted and unweighted jackknife variance estimators in a special case. The study

of finite sample properties of the jackknife and bootstrap estimators in general cases are limited to empirical simulations. Wu (1986) examined by simulation the relative biases of several variance estimators for estimating the variances and covariances of the components of $\hat{\beta}_{\text{LS}}$ under the model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad i = 1, \dots, 12,$$

with independent $\varepsilon_i \sim N(0, \sigma_i^2)$ and $x_i = 1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 6, 7, 8$, and 10. The results, based on 3000 simulations and listed in Table 7.1, show the robustness against heteroscedasticity of the estimators based on the weighted jackknife and the weighted PB, and the nonrobustness of the RB variance estimator. The results also indicate that the unweighted jackknife estimator is upward-biased and Hinkley's weighted jackknife estimator is downward-biased, which are supported by the asymptotic results given in Section 7.5.1. The estimators based on the unweighted jackknife and the unweighted PB perform poorly, although they are robust asymptotically. Simulation results for the mean squared error of some variance estimators can be found in Wu (1986, Rejoinder).

Table 7.1. Relative biases of variance estimators (i, j) : estimation of the covariance between the i th and the j th components of $\hat{\beta}_{\text{LS}}$ [Adapted from Wu (1986), by permission of Institute of Mathematical Statistics]

		(0, 0)	(0, 1)	(0, 2)	(1, 1)	(1, 2)	(2, 2)
$\sigma_i^2 = 1$	v_{JACK}	0.61	-0.78	1.03	0.93	-1.18	1.53
	v_{HJACK}	-0.13	0.16	-0.21	-0.17	0.22	-0.29
	v_{WJACK}	-0.01	0.01	-0.00	-0.00	-0.00	0.00
	v_{RBOOT}	-0.01	0.01	-0.01	-0.01	0.01	0.00
	v_{PBOOT}	0.63	-0.85	1.22	1.04	-1.49	2.18
	v_{WPBOOT}	-0.07	0.07	-0.08	-0.06	0.07	-0.06
$\sigma_i^2 = \frac{x_i}{2}$	v_{JACK}	0.97	-1.07	1.29	1.10	-1.29	1.45
	v_{HJACK}	-0.16	0.24	-0.35	-0.29	0.39	-0.47
	v_{WJACK}	0.02	0.04	-0.09	-0.08	0.12	-0.16
	v_{RBOOT}	0.39	-0.09	-0.04	-0.11	0.20	-0.29
	v_{PBOOT}	1.02	-0.98	1.17	0.91	-1.13	1.39
	v_{WPBOOT}	0.03	0.07	-0.14	-0.13	0.19	-0.28

v_{RBOOT} : variance estimator based on the RB.

v_{PBOOT} : variance estimator based on the PB.

v_{WPBOOT} : variance estimator based on the weighted PB.

In some problems, there are repeated measurements at each x_i , i.e.,

$$y_{ij} = x_i' \beta + \varepsilon_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, n.$$

We may treat $y_i = (y_{i1}, \dots, y_{in_i})'$ as a cluster or a group. The y_i are usually independent, but within each group the observations may be correlated. The customary estimator $\hat{\beta}_{\text{LS}}$, designed for the case where y_{i1}, \dots, y_{in_i} are independent, is robust against within-group correlation. If we apply the weighted and unweighted jackknife, the PB, and the EB by treating y_i as the primary unit of data, then the variance estimators based on these resampling procedures are robust against within-group correlation (see Section 7.5.1), but they are not as efficient as variance estimators derived under the within-group independence assumption when this assumption is indeed valid (Shao and Rao, 1993).

7.3 Inference and Prediction Using the Bootstrap

In addition to variance and bias estimation, the bootstrap can be applied to make other statistical inferences. In this section, applications of the bootstrap to confidence sets, simultaneous confidence sets, hypothesis tests, and predictions under model (7.1) will be discussed.

As in the i.i.d. case, the bootstrap can be applied to estimate the entire distribution of a given statistic. In fact, many of the procedures described in this section are justified by the consistency of the bootstrap distribution estimators for $\hat{\theta}$ (Section 7.5.3).

7.3.1 Confidence sets

Many results for bootstrap confidence sets described in Chapter 4 can be extended to the case of $\hat{\theta} = \hat{\theta}_{\text{LS}}$ under model (7.1) with i.i.d. ε_i . It was shown by Hall (1989c) that, in terms of the asymptotic accuracy, the confidence sets based on the RB may even perform better than they do in the i.i.d. case—a somewhat surprising result.

Assume that the first component of x_i equals 1 for all i so that $x_i = (1, t_i')'$. Define $\beta = (\beta_0, \beta_1)'$, where β_0 is a scalar and β_1 is a $(p - 1)$ -vector. Then model (7.1) can be written as

$$y_i = \beta_0 + t_i' \beta_1 + \varepsilon_i, \quad i = 1, \dots, n. \quad (7.25)$$

Note that β_0 is the intercept parameter and β_1 can be viewed as a “slope” parameter. The LSE $\hat{\beta}_{\text{LS}}$ can be written as $(\hat{\beta}_0, \hat{\beta}_1')'$ with

$$\hat{\beta}_0 = \bar{y} - \bar{t}' \hat{\beta}_1 \quad \text{and} \quad \hat{\beta}_1 = S_{tt}^{-1} S_{yt},$$

where $S_{tt} = \sum_{i=1}^n (t_i - \bar{t})(t_i - \bar{t})'$, $S_{yt} = \sum_{i=1}^n (y_i - \bar{y})(t_i - \bar{t})$, $\bar{t} = n^{-1} \sum_{i=1}^n t_i$, and $\bar{y} = n^{-1} \sum_{i=1}^n y_i$. Under model (7.25), $\bar{r} = 0$. Hence, for the RB, the bootstrap data $\varepsilon_1^*, \dots, \varepsilon_n^*$ are i.i.d. from the empirical distribution of r_1, \dots, r_n (or $r_1/\sqrt{1-p/n}, \dots, r_n/\sqrt{1-p/n}$).

We first consider confidence intervals or bounds for $\theta = c'\beta_1$ with a fixed $(p-1)$ -vector c . Define $\hat{\theta} = c'\hat{\beta}_1$, $\hat{\theta}^* = c'\hat{\beta}_1^*$,

$$K_{\text{BOOT}}(x) = P_*\{\hat{\theta}^* \leq x\}, \quad H_{\text{BOOT}}(x) = P_*\{\sqrt{n}(\hat{\theta}^* - \hat{\theta}) \leq x\},$$

and

$$G_{\text{BOOT}}(x) = P_*\{(\hat{\sigma}^{*2} c' S_{tt}^{-1} c)^{-1/2}(\hat{\theta}^* - \hat{\theta}) \leq x\},$$

where $\hat{\beta}_0^*$, $\hat{\beta}_1^*$, and $\hat{\sigma}^{*2} = n^{-1} \sum_{i=1}^n (y_i^* - \hat{\beta}_0^* - t_i^* \hat{\beta}_1^*)^2$ are the bootstrap analogs of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$ in (7.21), respectively. For a given α , the bootstrap percentile, the bootstrap BC, the hybrid bootstrap, and the bootstrap-t lower confidence bounds for $\theta = c'\beta_1$ are, respectively,

$$\underline{\theta}_{\text{BP}} = K_{\text{BOOT}}^{-1}(\alpha),$$

$$\underline{\theta}_{\text{BC}} = K_{\text{BOOT}}^{-1}(\Phi(z_\alpha + 2z_0))$$

with $z_0 = \Phi^{-1}(K_{\text{BOOT}}(\hat{\theta}))$,

$$\underline{\theta}_{\text{HB}} = \hat{\theta} - n^{-1/2} H_{\text{BOOT}}^{-1}(1-\alpha),$$

and

$$\underline{\theta}_{\text{BT}} = \hat{\theta} - \hat{\sigma}(c' S_{tt}^{-1} c)^{1/2} G_{\text{BOOT}}^{-1}(1-\alpha).$$

The bootstrap BC_a confidence bounds are difficult to construct because the acceleration constant a is not easy to estimate under the general model (7.25). In the special case where $p = 2$, i.e., t_i and β_1 are scalars, the bootstrap BC_a lower confidence bound for β_1 is

$$\underline{\theta}_{\text{BC}_a} = K_{\text{BOOT}}^{-1}\left(\Phi\left(z_0 + \frac{z_\alpha + z_0}{1 - \hat{a}(z_\alpha + z_0)}\right)\right),$$

where $z_0 = \Phi^{-1}(K_{\text{BOOT}}(\hat{\theta}))$ and

$$\hat{a} = -\frac{1}{6n^{7/2}\hat{\sigma}^3 S_{tt}^{3/2}} \sum_{i=1}^n r_i^3 \sum_{i=1}^n t_i^3.$$

The bootstrap upper confidence bounds can be constructed similarly, and equal-tail bootstrap confidence intervals can be obtained by combining the bootstrap upper and lower confidence bounds. Monte Carlo or other methods described in Chapter 5 can be used to approximate the percentiles of K_{BOOT} , H_{BOOT} , or G_{BOOT} .

Hall (1989c) investigated the asymptotic properties of these bootstrap confidence sets. Let G be the distribution of $(c' S_{tt}^{-1} c)^{-1/2}(\hat{\theta} - \theta)/\hat{\sigma}$ and

$$\underline{\theta}_{\text{EXACT}} = \hat{\theta} - \hat{\sigma}(c' S_{tt}^{-1} c)^{1/2}G^{-1}(1 - \alpha),$$

which is an exact lower confidence bound for θ . Under some conditions (see Section 7.5.4), Hall (1989c) showed that

$$\underline{\theta}_{\text{BT}} - \underline{\theta}_{\text{EXACT}} = O_p(n^{-3/2}), \quad \underline{\theta}_{\text{HB}} - \underline{\theta}_{\text{EXACT}} = O_p(n^{-3/2}),$$

$$\underline{\theta}_{\text{BC}_a} - \underline{\theta}_{\text{EXACT}} = O_p(n^{-3/2}) \quad (\text{for the special case of } p = 2),$$

$$\underline{\theta}_{\text{BP}} - \underline{\theta}_{\text{EXACT}} = O_p(n^{-1}) \quad \text{and} \quad \underline{\theta}_{\text{BC}} - \underline{\theta}_{\text{EXACT}} = O_p(n^{-1}).$$

The results for the bootstrap-t, the bootstrap percentile, the bootstrap BC, and the bootstrap BC_a are the same as those in the i.i.d. case studied in Chapter 4, i.e., $\underline{\theta}_{\text{BT}}$ and $\underline{\theta}_{\text{BC}_a}$ are closer to $\underline{\theta}_{\text{EXACT}}$ than $\underline{\theta}_{\text{BC}}$ and $\underline{\theta}_{\text{BP}}$. The result for $\underline{\theta}_{\text{HB}}$, however, is unusual in view of result (4.44): the hybrid bootstrap performs better than it does in the i.i.d. case.

For the coverage probability, Hall (1989c) showed that

$$P\{\underline{\theta}_{\text{BT}} \leq \theta\} = 1 - \alpha + O(n^{-3/2}), \quad P\{\underline{\theta}_{\text{HB}} \leq \theta\} = 1 - \alpha + O(n^{-1}),$$

$$P\{\underline{\theta}_{\text{BC}_a} \leq \theta\} = 1 - \alpha + O(n^{-1}) \quad (\text{for the special case of } p = 2),$$

$$P\{\underline{\theta}_{\text{BP}} \leq \theta\} = 1 - \alpha + O(n^{-1/2}) \quad \text{and} \quad P\{\underline{\theta}_{\text{BC}} \leq \theta\} = 1 - \alpha + O(n^{-1/2}).$$

We have some more unusual results! The bootstrap-t confidence bound is third order accurate and the hybrid bootstrap confidence bound is second order accurate. Note that in Chapter 4 we showed that, in the i.i.d. case, $[\underline{\theta}_{\text{BT}}, \infty)$ is second order accurate and $[\underline{\theta}_{\text{HB}}, \infty)$ is only first order accurate [see (4.39) and (4.45)]. The results for the bootstrap percentile, bootstrap BC, and BC_a confidence bounds are the same as those in Chapter 4: $[\underline{\theta}_{\text{BC}_a}, \infty)$ is second order accurate and $[\underline{\theta}_{\text{BP}}, \infty)$ and $[\underline{\theta}_{\text{BC}}, \infty)$ are first order accurate.

Hall (1989c) argued that the reason why we have these unusual results is that model (7.25) is equivalent to

$$y_i = (\beta_0 + \bar{t}' \beta_1) + (t_i - \bar{t})' \beta_1 + \varepsilon_i, \quad i = 1, \dots, n,$$

and because of $\sum_{i=1}^n (t_i - \bar{t}) = 0$, some terms in the Cornish-Fisher expansions for the quantile functions of $\sqrt{n}c'(\hat{\beta}_1^* - \hat{\beta}_1)$ and $\sqrt{n}c'(\hat{\beta}_1 - \beta_1)$ vanish. Hall (1989c) further showed that when the t_i are scalars, all of the bootstrap confidence bounds are at least second order accurate if $\sum_{i=1}^n t_i^3 = O(\sqrt{n})$ (e.g., the t_i are equally spaced).

For the equal-tail confidence intervals, the bootstrap-t confidence interval has been shown to be fourth order accurate, but all of the other four bootstrap confidence intervals are second order accurate. The bootstrap-t confidence interval is longer than the other intervals.

We now consider the intercept parameter β_0 or, more generally, $\theta = \beta_0 + t'\beta_1$, the mean of a future observation at t . The bootstrap confidence sets can be constructed similarly using $\hat{\theta} = \hat{\beta}_0 + t'\hat{\beta}_1$, $\hat{\theta}^* = \hat{\beta}_0^* + t'\hat{\beta}_1^*$, and the percentiles of K_{BOOT} , H_{BOOT} , and G_{BOOT} [with $c'S_{tt}^{-1}c$ changed to $S_{yy} = n^{-1} + (t - \bar{t})'S_{tt}^{-1}(t - \bar{t})$]. The bootstrap BC_a confidence sets can be constructed using

$$\hat{a} = \frac{1}{6n^{3/2}\hat{\sigma}^3} \sum_{i=1}^n r_i^3 \left(3S_{yy}^{-1/2} - \frac{2}{n} \sum_{i=1}^n y_i^3 \right).$$

Hall (1989c) showed that the asymptotic properties of the bootstrap confidence sets in this case are the same as those in the i.i.d. case: the bootstrap-t and BC_a confidence bounds are second order accurate and the other three bootstrap confidence bounds are first order accurate.

In regression analysis, one may want to construct a confidence region for the vector β under model (7.1). Adkins and Hill (1990) used the bootstrap-t confidence region

$$\{\beta : n(\hat{\beta}_{\text{LS}} - \beta)'X'X(\hat{\beta}_{\text{LS}} - \beta)/\hat{\sigma}^2 \leq J_{\text{BOOT}}^{-1}(1 - \alpha)\},$$

where $J_{\text{BOOT}}(x) = P_*\{n(\hat{\beta}_{\text{LS}}^* - \hat{\beta}_{\text{LS}})'X'X(\hat{\beta}_{\text{LS}}^* - \hat{\beta}_{\text{LS}})/\hat{\sigma}^{*2} \leq x\}$.

So far we have only considered the RB. Confidence sets based on the other two bootstraps, the PB and the EB, can be similarly constructed, except for the bootstrap BC_a , for which the estimation of the acceleration constant a may not be simple. For the PB, if $(y_1, x'_1), \dots, (y_n, x'_n)$ are i.i.d., then results from the previous chapters can be directly applied.

Because of the nature of the bootstrap procedures, the confidence sets based on the PB or the EB are robust against heteroscedasticity and the confidence sets based on the RB are not robust, whereas the latter are more efficient than the former in the equal variance case.

7.3.2 Simultaneous confidence intervals

In regression analysis, simultaneous confidence intervals are often required. Let $\mathbb{C} \subset \mathbb{R}^p$ be an index set. For a given α , intervals \mathcal{I}_c , $c \in \mathbb{C}$, are called $1 - \alpha$ simultaneous confidence intervals for $c'\beta$, $c \in \mathbb{C}$, if

$$P\{c'\beta \in \mathcal{I}_c \text{ for all } c \in \mathbb{C}\} = 1 - \alpha.$$

When the errors ε_i in model (7.1) have a common normal distribution, Scheffé (1953) constructed exact $1 - \alpha$ simultaneous confidence intervals based on the pivotal quantity

$$\mathfrak{R}(c'\beta) = |c'(\hat{\beta}_{\text{LS}} - \beta)|/\hat{\sigma}_c,$$

where $\hat{\sigma}_c = [c'(X'X)^{-1}c]^{1/2}\hat{\sigma}$. Scheffé's intervals, however, are not valid and have very low accuracy when the ε_i are not normal.

Beran (1988b) suggested a bootstrap prepivoting method to construct simultaneous confidence intervals for $c'\beta$, $c \in \mathbb{C}$. Let $H_c(\cdot, \beta, F_\varepsilon)$ be the distribution of $\mathfrak{R}(c'\beta)$, $H(\cdot, \beta, F_\varepsilon)$ be the distribution of

$$\mathfrak{R} = \sup_{c \in \mathbb{C}} H_c(\mathfrak{R}(c'\beta), \beta, F_\varepsilon),$$

and \hat{F}_ε be defined as in Section 7.2.2. Then $H_c(\cdot, \beta, F_\varepsilon)$ and $H(\cdot, \beta, F_\varepsilon)$ can be estimated by $\hat{H}_c(\cdot, \hat{\beta}_{\text{LS}}, \hat{F}_\varepsilon)$ and $\hat{H}(\cdot, \hat{\beta}_{\text{LS}}, \hat{F}_\varepsilon)$, respectively, and the bootstrap simultaneous confidence intervals can be defined as

$$\mathcal{I}_c = [c'\hat{\beta}_{\text{LS}} - \hat{\sigma}_c d_c, c'\hat{\beta}_{\text{LS}} + \hat{\sigma}_c d_c], \quad c \in \mathbb{C}, \quad (7.26)$$

where $d_c = \hat{H}_c^{-1}(\hat{H}^{-1}(1 - \alpha))$. If the ε_i are normally distributed, then it can be shown that the intervals in (7.26) are Scheffé's intervals. In general, under some assumptions, Beran (1988b) showed that

$$P\{c'\beta \in \mathcal{I}_c \text{ for all } c \in \mathbb{C}\} \rightarrow 1 - \alpha$$

and

$$\sup_{c \in \mathbb{C}} |P\{c'\beta \in \mathcal{I}_c\} - H^{-1}(1 - \alpha, \beta, F_\varepsilon)| \rightarrow 0.$$

That is, the intervals in (7.26) are approximate $1 - \alpha$ simultaneous confidence intervals, and for each fixed c , \mathcal{I}_c is an approximate $H^{-1}(1 - \alpha, \beta, F_\varepsilon)$ confidence interval for $c'\beta$. The second property is called the balancedness of simultaneous confidence intervals. Beran (1990) showed that the error in the coverage probability of this method can be further reduced using the double bootstrap.

The computation of d_c , however, is not simple. Beran (1988b) suggested the following stochastic approximation algorithm.

- (1) Draw i.i.d. bootstrap data $\varepsilon_1^{*b}, \dots, \varepsilon_n^{*b}$, $b = 1, \dots, B$, from \hat{F}_ε and set $y_i^{*b} = x'_i \hat{\beta}_{\text{LS}} + \varepsilon_i^{*b}$. Let $\hat{\beta}_{\text{LS}}^{*b}$ and $\hat{\sigma}_c^*$ be the bootstrap analogs of $\hat{\beta}_{\text{LS}}$ and $\hat{\sigma}_c$, respectively, based on the data $(y_1^{*b}, x'_1), \dots, (y_n^{*b}, x'_n)$, and let \hat{H}_c^B be the empirical distribution of $\mathfrak{R}_c^{*b} = |c'(\hat{\beta}_{\text{LS}}^{*b} - \hat{\beta}_{\text{LS}})|/\hat{\sigma}_c^*$, $b = 1, \dots, B$.

- (2) Assume that \mathbb{C} is a finite set. Let \hat{d}^B be the $(1 - \alpha)$ th quantile of the empirical distribution of

$$\max_{c \in \mathbb{C}} \hat{H}_c^B(\mathfrak{R}_c^{*b}) = \frac{1}{B} \max_{c \in \mathbb{C}} [\text{rank}(\mathfrak{R}_c^{*b}) - 1], \quad b = 1, \dots, B.$$

Then d_c can be approximated by the \hat{d}^B th quantile of \hat{H}_c^B .

- (3) For an infinite set \mathbb{C} , let U be a probability distribution having full support on \mathbb{C} and $\mathbb{C}^* = \{c_1, \dots, c_k\}$ be an i.i.d. sample from U . Then obtain an approximation to d_c by performing (1)-(2) with \mathbb{C} replaced by \mathbb{C}^* .

In Section 7.3.1, we discussed bootstrap confidence sets for $\beta_0 + t'\beta_1$ for a fixed t . It may be of interest to construct simultaneous confidence intervals (confidence band) for $\beta_0 + t'\beta_1$ for $t \in \mathbb{T} \subset \mathbb{R}^p$. When the ε_i are normal, simultaneous confidence intervals for $\beta_0 + t'\beta_1$, $t \in \mathbb{T}$, are

$$\mathcal{I}_t = [\hat{\beta}_0 + t'\hat{\beta}_1 - \hat{\sigma}u_-f_-(t), \hat{\beta}_0 + t'\hat{\beta}_1 + \hat{\sigma}u_+f_+(t)], \quad t \in \mathbb{T}, \quad (7.27)$$

where $(\hat{\beta}_0, \hat{\beta}_1')' = \hat{\beta}_{\text{LS}}$, f_+ and f_- are two nonnegative functions on \mathbb{T} , and u_+ and u_- are two scalars satisfying

$$P\{-u_+f_+(t) \leq \Delta(t) \leq u_-f_-(t) \text{ for all } t \in \mathbb{T}\} = 1 - \alpha$$

with $\Delta(t) = (\hat{\beta}_0 + t'\hat{\beta}_1 - \beta_0 - t'\beta_1)/\hat{\sigma}$. Common choices of f_+ and f_- are: (1) $f_\pm \equiv 1$, and (2) $f_\pm(t) = f_p(t) = [1 + n(t - \bar{t})'S_{tt}^{-1}(t - \bar{t})]^{1/2}$, which leads to a Working-Hotelling type confidence band.

When the ε_i are not normal, the traditional approach approximates u_\pm by normal approximations, which leads to first order accurate confidence intervals.

Hall and Pittelkow (1990) considered simultaneous confidence intervals for $\beta_0 + t'\beta_1$, $t \in \mathbb{T}$, based on the RB. Let $\Delta^*(t) = (\hat{\beta}_0^* + t'\hat{\beta}_1^* - \hat{\beta}_0 - t'\hat{\beta}_1)/\hat{\sigma}^*$. Then the bootstrap simultaneous intervals are given by (7.27) with u_\pm estimated by \hat{u}_\pm satisfying

$$P_*\{-\hat{u}_+f_+(t) \leq \Delta^*(t) \leq \hat{u}_-f_-(t), \text{ for all } t \in \mathbb{T}\} = 1 - \alpha. \quad (7.28)$$

Some other restrictions may be imposed to solve (7.28). For example, $\hat{u}_+ = \hat{u}_-$, which leads to a symmetric confidence band; selecting \hat{u}_\pm by minimizing $\hat{u}_+ + \hat{u}_-$ subject to (7.28) provides the narrowest width confidence band; and selecting \hat{u}_\pm satisfying

$$\begin{aligned} 1 - \alpha/2 &= P_*\{-\hat{u}_+f_+(t) \leq \Delta^*(t) \text{ for all } t \in \mathbb{T}\} \\ &= P_*\{\Delta^*(t) \leq \hat{u}_-f_-(t) \text{ for all } t \in \mathbb{T}\} \end{aligned}$$

leads to an equal-tail confidence band. Monte Carlo or other approximations may be used to compute \hat{u}_\pm .

Table 7.2. Empirical coverage probabilities of 95% bootstrap simultaneous confidence bands [Adapted from Hall and Pittelkow (1990), by permission of Gordon and Breach Science Publishers]

n	Error Type	$f_{\pm} = f_p, \mathbb{T} = \mathbb{R}$		$f_{\pm} = f_p \equiv 1, \mathbb{T} = \mathbb{R}$	
		$u_+ = u_-$	$u_+ + u_- = \min$	$u_+ = u_-$	$u_+ + u_- = \min$
10	(1)	0.967	0.955	0.965	0.957
	(2)	0.940	0.935	0.950	0.947
	(3)	0.909	0.902	0.918	0.923
	(4)	0.953	0.947	0.949	0.945
15	(1)	0.958	0.948	0.954	0.951
	(2)	0.973	0.956	0.948	0.940
	(3)	0.925	0.924	0.937	0.932
	(4)	0.951	0.950	0.949	0.948
20	(1)	0.955	0.943	0.943	0.942
	(2)	0.953	0.951	0.953	0.952
	(3)	0.917	0.917	0.921	0.923
	(4)	0.949	0.947	0.949	0.939

Hall and Pittelkow (1990) showed that the bootstrap confidence band is second order accurate. They also provided some empirical results under the model

$$y_i = \beta_0 + \frac{i}{n}\beta_1 + \varepsilon_i, \quad i = 1, \dots, n,$$

with $\beta_0 = 0$, $\beta_1 = 1$, and i.i.d. $\varepsilon_i \sim F_\varepsilon$, where F_ε is the distribution of the following four types of random variables: (1) $N(0, 1)$; (2) standardized $|N(0, 1)|$; (3) standardized $|N(0, 1)|^2$; and (4) standardized squared uniform over $[0, 1]$. The results in Table 7.2 are based on 1000 simulations and $B = 499$ in the bootstrap Monte Carlo approximations.

7.3.3 Hypothesis tests

The most commonly encountered testing problem under model (7.1) is testing the following linear hypothesis:

$$H_0 : C\beta = h \quad \text{versus} \quad H_1 : C\beta \neq h, \quad (7.29)$$

where C is a given $q \times p$ matrix of rank q and h is a given q -vector. The test statistic used in the traditional approach, based on the likelihood ratio principle and the normality assumption of the error distribution, is given by

$$T = \frac{(n-p)(C\hat{\beta}_{\text{LS}} - h)'[C(X'X)^{-1}C']^{-1}(C\hat{\beta}_{\text{LS}} - h)}{q[y'y - y'X(X'X)^{-1}X'y]}. \quad (7.30)$$

When the ε_i are i.i.d. from a normal distribution, T has an F-distribution with q and $n-p$ degrees of freedom. If the ε_i are not normal, the traditional approach uses an F-distribution approximation.

A test for (7.29) based on the RB can be constructed. According to the basic principle of bootstrap hypothesis testing discussed in Chapter 4, the bootstrap data should be generated from the model under the hypothesis H_0 :

$$y_i = x'_i \beta + \varepsilon_i, \quad i = 1, \dots, n, \quad \text{subject to } C\beta = h. \quad (7.31)$$

Under model (7.31), the LSE of β is the restricted LSE of β under model (7.1):

$$\tilde{\beta}_{LS} = \hat{\beta}_{LS} + (X'X)^{-1}C'[C(X'X)^{-1}C']^{-1}(h - C\hat{\beta}_{LS}). \quad (7.32)$$

Then we have the following bootstrap test for (7.29):

- (1) Draw i.i.d. data $\varepsilon_1^*, \dots, \varepsilon_n^*$ from \hat{F}_ε and define $y_i^* = x'_i \tilde{\beta}_{LS} + \varepsilon_i^*$, $y^* = (y_1^*, \dots, y_n^*)'$, $\tilde{\beta}_{LS}^* = (X'X)^{-1}X'y^*$, and

$$T^* = \frac{(n-p)(C\tilde{\beta}_{LS}^* - h)'[C(X'X)^{-1}C']^{-1}(C\tilde{\beta}_{LS}^* - h)}{q[y^{*\prime}y^* - y^{*\prime}X(X'X)^{-1}X'y^*]}. \quad (7.33)$$

- (2) Reject H_0 if $T \geq \hat{t}_\alpha$, where \hat{t}_α is the $(1-\alpha)$ th quantile of the bootstrap distribution of T^* (calculated by Monte Carlo or other approximations discussed in Chapter 5). A bootstrap estimate of the P -value for testing (7.29) is $P_*\{T^* \geq T\}$.

Mammen (1993) proved the asymptotic correctness (see Definition 4.3) of the bootstrap test described by (1)-(2). Rayner (1990) showed that, in linear models with correlated normal errors, this bootstrap test is second order correct in the sense that its type I error, the probability of rejecting H_0 when H_0 is true, differs from α by a term of order $O(n^{-1})$.

Mammen (1993) also defined bootstrap tests based on the PB and the EB. The EB test can be obtained by changing $\varepsilon_1^*, \dots, \varepsilon_n^*$ in step (1) to an independent sample with ε_i^* having mean 0, variance r_i^2 , and third moment r_i^3 , $i = 1, \dots, n$. For the PB test, T^* in (7.33) is replaced by

$$\frac{(n-p)(C\tilde{\beta}_{LS}^* - h)'[C(X^{*\prime}X^*)^{-1}C']^{-1}(C\tilde{\beta}_{LS}^* - h)}{q[y^{*\prime}y^* - y^{*\prime}X^*(X^{*\prime}X^*)^{-1}X^{*\prime}y^*]},$$

where $\tilde{\beta}_{LS}^*$ is given by (7.32), based on $(y_1^*, x_1^{*\prime}), \dots, (y_n^*, x_n^{*\prime})$ that are i.i.d. from the empirical distribution putting mass n^{-1} to $(y_i - x'_i \hat{\beta}_{LS} + x'_i \tilde{\beta}_{LS}, x'_i)$, $i = 1, \dots, n$. These bootstrap tests are shown to be asymptotically correct and robust against heteroscedasticity. In fact, Mammen (1993) also proved that the bootstrap tests are still asymptotically correct when the dimension

p of the parameter β tends to ∞ as a function of n , a situation in which the traditional approximate F-test may not work. Details can be found in Section 7.5.4.

7.3.4 Prediction

An important application of model (7.1) is the prediction of the future response y_f at a given value x_f of the explanatory variable. There are two types of prediction procedures in general: point prediction and interval prediction. The bootstrap can be used to assess the prediction error in a point prediction problem and to construct the prediction interval in an interval prediction problem.

Under model (7.1) with i.i.d. errors ε_i , a customary point prediction for y_f is

$$\hat{y}_f = x'_f \hat{\beta}_{LS},$$

and it is evaluated by its mean squared prediction error

$$mse(x_f) = E(y_f - \hat{y}_f)^2,$$

where the expectation is over the joint distribution of y_f and \hat{y}_f . Usually, y_f and y_1, \dots, y_n are independent. Therefore,

$$mse(x_f) = var(y_f) + var(x'_f \hat{\beta}_{LS}) = \sigma^2 + \sigma^2 x'_f (X' X)^{-1} x_f.$$

An estimator of $mse(x_f)$ based on the RB can be obtained as follows. Let $\varepsilon_1^*, \dots, \varepsilon_n^*$, and ε_f^* be i.i.d. from \tilde{F}_ε , the empirical distribution of the adjusted residuals $(r_i - \bar{r})/\sqrt{1 - k_n/n}$, $i = 1, \dots, n$, and let $\hat{\beta}_{LS}^*$ be defined as in Section 7.2.2. Let $y_f^* = x'_f \hat{\beta}_{LS} + \varepsilon_f^*$ be the bootstrap future value and $\hat{y}_f^* = x'_f \hat{\beta}_{LS}^*$ be the bootstrap prediction of y_f^* . Then the bootstrap estimator of $mse(x_f)$ is

$$\begin{aligned} \widehat{mse}_{BOOT}(x_f) &= E_*(y_f^* - \hat{y}_f^*)^2 = var_*(y_f^*) + var_*(x'_f \hat{\beta}_{LS}^*) \\ &= \tilde{\sigma}^2 + \tilde{\sigma}^2 x'_f (X' X)^{-1} x_f, \end{aligned}$$

where $\tilde{\sigma}^2$ is given by (7.22). Since $E(\tilde{\sigma}^2) = \sigma^2$, $\widehat{mse}_{BOOT}(x_f)$ is exactly unbiased, i.e., $E[\widehat{mse}_{BOOT}(x_f)] = mse(x_f)$.

Here we may appreciate the effect of generating bootstrap data from the adjusted residuals. If $\varepsilon_1^*, \dots, \varepsilon_n^*$ and ε_f^* are i.i.d. from $r_i - \bar{r}$, $i = 1, \dots, n$, then the bootstrap estimator of $mse(x_f)$ is

$$\hat{\sigma}^2 + \hat{\sigma}^2 x'_f (X' X)^{-1} x_f,$$

which has a negative bias

$$-\frac{k_n}{n} [\sigma^2 + \sigma^2 x'_f (X' X)^{-1} x_f].$$

If $(X'X)^{-1}$ is of the order $O(n^{-1})$, then this bias is of the same order as $\sigma^2 x_f'(X'X)^{-1}x_f$, the second term in $\text{mse}(x_f)$.

Sometimes we need to predict future values for a set \mathcal{X} of x_f . Let ν be a probability measure on \mathcal{X} . Then the average mean squared prediction error is

$$\int \text{mse}(x_f) d\nu(x_f) = \sigma^2 + \sigma^2 \text{tr}[(X'X)^{-1}\Sigma_x],$$

where $\text{tr}(A)$ is the trace of the matrix A and $\Sigma_x = \int x_f x_f' d\nu(x_f)$ is a known matrix, and its bootstrap estimator is

$$\int \widehat{\text{mse}}_{\text{BOOT}}(x_f) d\nu(x_f) = \tilde{\sigma}^2 + \tilde{\sigma}^2 \text{tr}[(X'X)^{-1}\Sigma_x]. \quad (7.34)$$

For example, if $\mathcal{X} = \{x_1, \dots, x_n\}$ and ν assigns mass n^{-1} to each x_i , then the average mse and its bootstrap estimator are, respectively,

$$\sigma^2 + \sigma^2 p/n \quad \text{and} \quad \tilde{\sigma}^2 + \tilde{\sigma}^2 p/n.$$

When x_1, \dots, x_n, x_f are random and i.i.d., the average mean squared prediction error is

$$\overline{\text{mse}} = \sigma^2 + \sigma^2 \text{tr}[E(X'X)^{-1}E(x_f x_f')] = \sigma^2 + \sigma^2 p/n + o(n^{-1}).$$

Based on the PB, Efron (1983) proposed a bootstrap estimator of $\overline{\text{mse}}$. Define the expected excess error by

$$e = E \left[(y_f - \hat{y}_f)^2 - \frac{1}{n} \sum_{i=1}^n r_i^2 \right]$$

and its bootstrap estimator by

$$\hat{e} = E_* \left[\frac{1}{n} \sum_{i=1}^n (y_i - x_i' \hat{\beta}_{\text{LS}}^*)^2 - \frac{1}{n} \sum_{i=1}^n (y_i^* - x_i^* \hat{\beta}_{\text{LS}}^*)^2 \right],$$

where $\hat{\beta}_{\text{LS}}^*$ is given by (7.23). The bootstrap estimator of $\overline{\text{mse}}$ given by Efron (1983) is then

$$\widehat{\overline{\text{mse}}}_{\text{BOOT}} = \frac{1}{n} \sum_{i=1}^n r_i^2 + \hat{e}. \quad (7.35)$$

It can be shown (see the proof of Theorem 7.12 in Section 7.5.4) that

$$E(\widehat{\overline{\text{mse}}}_{\text{BOOT}}) = \sigma^2 + \sigma^2 p/n + o_p(n^{-1}).$$

Hence, this bootstrap estimator is almost unbiased.

Some other bootstrap estimators of prediction errors can be found in Bunke and Droege (1984) and Kipnis (1992). The average mean squared prediction error $\overline{\text{mse}}$ can also be estimated using cross-validation (see Section 7.4.1).

We now turn to interval prediction. An interval $\mathcal{I} = \mathcal{I}(y, X)$ is said to be a $1 - \alpha$ prediction interval for y_f if

$$P\{y_f \in \mathcal{I}\} = 1 - \alpha,$$

where P is the joint probability of y_f and y (conditional on x_f and X if they are random). When the errors ε_i are i.i.d. and normal, an exact $1 - \alpha$ prediction interval is

$$[\hat{y}_f - t_{n-p,\alpha/2}\hat{\sigma}_f, \hat{y}_f + t_{n-p,\alpha/2}\hat{\sigma}_f], \quad (7.36)$$

where $t_{n-p,\alpha/2}$ is the $1 - \alpha/2$ quantile of the t-distribution with $n - p$ degrees of freedom and

$$\hat{\sigma}_f^2 = \frac{1}{n-p} \sum_{i=1}^n r_i^2 [1 + x'_f(X'X)^{-1}x_f].$$

If the ε_i are not normal, then the traditional approach replaces $t_{n-p,\alpha/2}$ in (7.36) by $z_{\alpha/2}$, which results in an approximate $1 - \alpha$ prediction interval.

The following prediction interval, based on the RB, was proposed by Stine (1985):

$$\mathcal{I}_{\text{BOOT}} = [\hat{y}_f + (H_{\text{BOOT}}^f)^{-1}(\frac{\alpha}{2}), \hat{y}_f + (H_{\text{BOOT}}^f)^{-1}(1 - \frac{\alpha}{2})],$$

where

$$H_{\text{BOOT}}^f(x) = P_*\{y_f^* - \hat{y}_f^* \leq x\},$$

$y_f^* = x'_f \hat{\beta}_{\text{LS}} + \varepsilon_f^*$, $\hat{y}_f^* = x'_f \hat{\beta}_{\text{LS}}^*$, and $\hat{\beta}_{\text{LS}}^*$ is given by (7.19). Note that this prediction interval is obtained by estimating the percentile of $y_f - \hat{y}_f$ by that of $y_f^* - \hat{y}_f^*$ and, therefore, is a hybrid bootstrap prediction interval. Under some conditions, Stine (1985) proved the consistency of $\mathcal{I}_{\text{BOOT}}$, i.e., $P\{y_f \in \mathcal{I}_{\text{BOOT}}\} \rightarrow 1 - \alpha$. To compute H_{BOOT}^f by Monte Carlo, we may use the fact that H_{BOOT}^f is the convolution of ε_f^* and $-x'_f(X'X)^{-1}X'\varepsilon^*$, since ε_f^* and $\varepsilon^* = (\varepsilon_1^*, \dots, \varepsilon_n^*)'$ are independent under the bootstrap sampling. Since ε_f^* is distributed as \tilde{F}_ε (or \hat{F}_ε), $H_{\text{BOOT}}^f(x)$ can be approximated by

$$\frac{1}{B} \sum_{b=1}^B \tilde{F}_\varepsilon(x + x'_f(X'X)^{-1}X'\varepsilon^{*b}),$$

where ε^{*b} is the b th bootstrap sample from \tilde{F}_ε .

We may obtain a bootstrap-t prediction interval by estimating the distribution of $(y_f - \hat{y}_f)/\{\hat{\sigma}[1 + x'_f(X'X)^{-1}x_f]^{1/2}\}$ by the bootstrap distribution of $(y_f^* - \hat{y}_f^*)/\{\hat{\sigma}^*[1 + x'_f(X'X)^{-1}x_f]^{1/2}\}$. However, Bai and Olshen (1988) and Bai, Bickel and Olshen (1990) showed that in a special case, the one-sided bootstrap-t prediction interval is only first order accurate, unlike the one-sided bootstrap-t confidence intervals. Beran (1990) suggested an approach similar to the preprinting method for confidence sets (Section 4.3.1) to construct a prediction region whose coverage probability converges to $1 - \alpha$ at the rate n^{-2} . Some further discussions on designing bootstrap prediction regions can be found in Beran (1992).

7.4 Model Selection

In Section 7.3.4, we discussed the prediction of future response y values using an explanatory variable x . Each component of x is called a predictor. Since some of the components of x may not be actually related to y , the use of all the p components of x as predictors does not necessarily produce an accurate prediction. For example, suppose that α is a subset of $\{1, \dots, p\}$ and, under model (7.1), the components of β indexed by integers not in α are equal to 0. Let β_α and $x_{i\alpha}$ be the subvectors containing components of β and x_i , respectively, indexed by the integers in α . Then a model that is more compact than model (7.1) and produces more efficient predictions is

$$y_i = x'_{i\alpha}\beta_\alpha + \varepsilon_i, \quad i = 1, \dots, n. \quad (7.37)$$

Of course, the optimal model is the model (7.37) with α_o such that β_{α_o} contains exactly all nonzero components of β .

However, the optimal model is unknown since β is unknown. This leads to the problem of model selection. That is, one wants to select a model of the form (7.37) based on the data $(y_1, x'_1), \dots, (y_n, x'_n)$.

Since each model of the form (7.37) corresponds to a subset α of the integers $1, \dots, p$, in what follows, each α represents a model. Let \mathcal{A} be a class of models to be selected. For each $\alpha \in \mathcal{A}$, the size of α is defined to be the size (dimension) of the model α . Without loss of generality, we assume that model (7.1), the model with the largest size p , is always in \mathcal{A} . Note that the number of models in \mathcal{A} can be as small as 2 and as large as 2^p .

Assume throughout this section that $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ has i.i.d. components with mean 0 and variance σ^2 . Then, under model α , a customary estimator of β_α is the LSE

$$\hat{\beta}_\alpha = (X'_\alpha X_\alpha)^{-1} X'_\alpha y,$$

where $X_\alpha = (x_{1\alpha}, \dots, x_{n\alpha})'$, and the corresponding least squares prediction

of y_f is

$$\hat{y}_{f\alpha} = x'_{f\alpha} \hat{\beta}_\alpha.$$

The mean squared prediction error is

$$\text{mse}(x_f, \alpha) = E(y_f - \hat{y}_{f\alpha})^2 = \sigma^2 + \sigma^2 x'_{f\alpha} (X'_\alpha X_\alpha)^{-1} x_{f\alpha} + \Delta(x_f, \alpha),$$

where

$$\Delta(x_f, \alpha) = [x'_f \beta - x'_{f\alpha} (X'_\alpha X_\alpha)^{-1} X \beta]^2. \quad (7.38)$$

If model α is correct in the sense that the components of β indexed by integers not in α are all equal to 0, then $X\beta = X_\alpha \beta_\alpha$, $x'_f \beta = x'_{f\alpha} \beta_\alpha$, and $\Delta(x_f, \alpha)$ in (7.38) is equal to 0. The optimal model is the α with $\Delta(x_f, \alpha) = 0$ and the smallest size. When $\Delta(x_f, \alpha) > 0$, it is usually of a larger order than $\sigma^2 x'_{f\alpha} (X'_\alpha X_\alpha)^{-1} x_{f\alpha}$. Thus, if $\text{mse}(x_f, \alpha)$ is known, then the optimal model can be obtained by minimizing $\text{mse}(x_f, \alpha)$ over $\alpha \in \mathcal{A}$. Note that the optimal model can also be obtained by minimizing the average of $\text{mse}(x_f, \alpha)$ over $\mathcal{X} = \{x_1, \dots, x_n\}$:

$$\overline{\text{mse}}(\alpha) = \frac{1}{n} \sum_{i=1}^n \text{mse}(x_i, \alpha) = \sigma^2 + \frac{\sigma^2 p}{n} + \Delta(\alpha),$$

where

$$\Delta(\alpha) = \frac{1}{n} \beta' X' (I - H_\alpha) X \beta, \quad (7.39)$$

$H_\alpha = X_\alpha (X'_\alpha X_\alpha)^{-1} X'_\alpha$, and I is the $p \times p$ identity matrix.

However, both $\text{mse}(x_f, \alpha)$ and $\overline{\text{mse}}(\alpha)$ are unknown. A simple idea is to estimate $\overline{\text{mse}}(\alpha)$ [which is easier than to estimate $\text{mse}(x_f, \alpha)$] by $\widehat{\overline{\text{mse}}}(\alpha)$ and then select a model by minimizing $\widehat{\overline{\text{mse}}}(\alpha)$ over $\alpha \in \mathcal{A}$. This leads to a number of model selection procedures, such as the information criterion (AIC) (Akaike, 1970), the C_p criterion (Mallows, 1973), the Bayesian information criterion (BIC) (Schwartz, 1978), and the generalized information criterion (GIC) (Rao and Wu, 1989). In this section, we study two model selection methods based on data-resampling, namely, the cross-validation (jackknife) and the bootstrap methods.

7.4.1 Cross-validation

Allen (1974) and Stone (1974) proposed a model selection method, which is referred to as the *cross-validation* (CV) method and is essentially a method based on the idea of the delete-1 jackknife. Let $\hat{\beta}_{\alpha,i}$ be the LSE of β under model α after removing the pair (y_i, x'_i) , i.e.,

$$\hat{\beta}_{\alpha,i} = \left(\sum_{j \neq i} x_{j\alpha} x'_{j\alpha} \right)^{-1} \sum_{j \neq i} x_{j\alpha} y_j, \quad i = 1, \dots, n.$$

Since y_i and $\hat{\beta}_{\alpha,i}$ are independent, $\overline{\text{mse}}(\alpha)$ can be estimated by

$$\widehat{\overline{\text{mse}}}_{\text{cv}}(\alpha) = \frac{1}{n} \sum_{i=1}^n (y_i - x'_i \hat{\beta}_{\alpha,i})^2.$$

The CV method selects a model by minimizing $\widehat{\overline{\text{mse}}}_{\text{cv}}(\alpha)$ over $\alpha \in \mathcal{A}$.

An essential asymptotic requirement for any given model selection procedure is its consistency in the sense that

$$\lim_{n \rightarrow \infty} P\{\hat{\alpha} = \alpha_0\} = 1, \quad (7.40)$$

where $\hat{\alpha}$ is the model selected by using the given procedure. Although $\widehat{\overline{\text{mse}}}_{\text{cv}}(\alpha)$ is an almost unbiased estimator of $\overline{\text{mse}}(\alpha)$, it is shown in Section 7.5.5 that, if $\hat{\alpha}_{\text{cv}}$ is the model selected using the CV, then

$$\lim_{n \rightarrow \infty} P\{\hat{\alpha}_{\text{cv}} = \text{an incorrect model}\} = 0 \quad (7.41)$$

and

$$\lim_{n \rightarrow \infty} P\{\hat{\alpha}_{\text{cv}} = \alpha_0\} < 1 \quad (7.42)$$

unless the only correct model is $\alpha = \{1, \dots, p\}$. This means that the CV is *inconsistent* (unless all components of β are nonzero) and is too conservative in the sense that it tends to pick a model with an unnecessarily large size.

The simulation results shown in Table 7.3 indicate that the inconsistency of the CV can be quite serious: the probability in (7.42) can be as low as 0.5.

The reason for the inconsistency of the CV can be explained as follows. First, the consistency of any model selection method based on minimizing $\overline{\text{mse}}(\alpha)$ over $\alpha \in \mathcal{A}$ rests upon the consistency of $\widehat{\overline{\text{mse}}}(\alpha) - \widehat{\overline{\text{mse}}}(\gamma)$ as an estimator of the difference

$$\overline{\text{mse}}(\alpha) - \overline{\text{mse}}(\gamma) = \frac{\sigma^2(p_\alpha - p_\gamma)}{n} + \Delta(\alpha) - \Delta(\gamma), \quad \alpha, \gamma \in \mathcal{A}, \quad (7.43)$$

where p_α is the size of α and $\Delta(\alpha)$ is given by (7.39). Second, when both α and γ are correct models [$\Delta(\alpha) = \Delta(\gamma) = 0$],

$$\widehat{\overline{\text{mse}}}_{\text{cv}}(\alpha) - \widehat{\overline{\text{mse}}}_{\text{cv}}(\gamma) = \frac{2\sigma^2(p_\alpha - p_\gamma)}{n} - \frac{\varepsilon'(H_\alpha - H_\gamma)\varepsilon}{n} + o(n^{-1})$$

is an almost unbiased but not consistent estimator of the difference given in (7.43) (see Section 7.5.5 for the proof).

Like the delete-d jackknife variance estimators for sample quantiles discussed in Section 2.3, a delete-d type of CV can be used to rectify the

inconsistency of the CV. The delete-d CV, first discussed by Geisser (1975) and then studied by Burman (1989), Shao (1993b), and Zhang (1993), is an extension of the delete-1 CV previously discussed. Suppose that we split the $n \times (1 + p)$ matrix (y, X) into two distinct submatrices: a $d \times (1 + p)$ matrix (y_s, X_s) containing the rows of (y, X) indexed by the integers in s , a subset of $\{1, \dots, n\}$ of size d , and an $(n - d) \times (1 + p)$ matrix (y_{s^c}, X_{s^c}) containing the rows of (y, X) indexed by the integers in s^c , the complement of s . For any $\alpha \in \mathcal{A}$, we estimate β_α by $\hat{\beta}_{\alpha, s^c}$, the LSE based on (y_{s^c}, X_{s^c}) under model α . Hence, (y_{s^c}, X_{s^c}) is called the construction data. The prediction error is assessed by $\|y_s - X_{\alpha, s} \hat{\beta}_{\alpha, s^c}\|^2$, where $X_{\alpha, s}$ is a $d \times p_\alpha$ matrix containing the columns of X_s indexed by the integers in α . Thus, (y_s, X_s) is called the validation data.

Let \mathcal{S} be a collection of subsets of $\{1, \dots, n\}$ of size $d < n$. The delete-d CV method selects a model by minimizing

$$\widehat{\text{mse}}_{\text{cv-d}}(\alpha) = \frac{1}{B} \sum_{s \in \mathcal{S}} \|y_s - X_{\alpha, s} \hat{\beta}_{\alpha, s^c}\|^2 \quad (7.44)$$

over $\alpha \in \mathcal{A}$, where B is the number of subsets in \mathcal{S} . The set \mathcal{S} can be obtained by using a balanced incomplete block design described in Section 5.2.1 or by taking a simple random sample from the collection of all possible subsets of $\{1, \dots, n\}$ of size d . Note that the CV discussed earlier is the special case of the delete-d CV with $d = 1$ and $\mathcal{S} = \{1, \dots, n\}$.

It was shown in Shao (1993b) that under some weak conditions (see Section 7.5.5) $\hat{\alpha}_{\text{cv-d}}$, the model selected by using the delete-d CV, is consistent in the sense of (7.40) if and only if $d/n \rightarrow 1$ and $n - d \rightarrow \infty$. This is a somewhat shocking discovery since it means that d , the size of the validation data set, should be much larger than $n - d$, the size of the construction data set, which is totally opposite to the delete-1 CV.

Technically, the condition $d/n \rightarrow 1$ is needed because it is necessary and sufficient for the consistency of $\widehat{\text{mse}}_{\text{cv-d}}(\alpha) - \widehat{\text{mse}}_{\text{cv-d}}(\gamma)$ as an estimator of $\text{mse}(\alpha) - \text{mse}(\gamma)$. Here, we explain heuristically the reason why we need this condition.

Note that $\text{mse}(\alpha) = \text{mse}_n(\alpha)$ depends on the sample size n , although the subscript n is omitted in the previous discussions. It can be seen from (7.44) that $\widehat{\text{mse}}_{\text{cv-d}}(\alpha)$ is an estimator of $\text{mse}_{n-d}(\alpha)$, not $\text{mse}_n(\alpha)$. When α is a correct model $[\Delta(\alpha) = 0]$,

$$\text{mse}_{n-d}(\alpha) = \sigma^2 + \frac{\sigma^2 p_\alpha}{n - d}.$$

Note that α_0 minimizes $\text{mse}_m(\alpha)$ for any fixed m . As a function of α , $\text{mse}_{n-d}(\alpha)$ is flat if d is small. Therefore, with a small d , it is difficult to find the minimum of $\text{mse}_{n-d}(\alpha)$ over all of the correct $\alpha \in \mathcal{A}$. Using

the delete-1 CV can be compared to using a telescope to see some objects 10,000 meters away, whereas using the delete-d CV is more like using the same telescope to see the same objects only 100 meters away. Of course, the latter method can see the differences among these objects more clearly. Thus, it is wise to use a relatively large d , although we still need $n - d \rightarrow \infty$ to ensure the consistency of the model fitting in the CV.

We end the discussion of the CV method by showing some empirical results. The finite sample performance of some selection procedures are studied by simulation under model (7.1) with $p = 5$, $n = 40$, and i.i.d. standard normal errors ε_i . The first component of each x_i is 1 and the values of other components of x_i are listed in Table 1 of Shao (1993b). The delete-1 CV and the delete-25 CV, as well as three non-resampling model

Table 7.3. Empirical model selection probabilities

True β'	α	AIC	C_p	BIC	CV_1^\dagger	CV_{25}^\dagger
(2, 0, 0, 4, 0)	1, 4 [‡]	.567	.594	.804	.484	.934
	1, 2, 4	.114	.110	.049	.133	.025
	1, 3, 4	.126	.113	.065	.127	.026
	1, 4, 5	.101	.095	.057	.138	.012
	1, 2, 3, 4	.030	.028	.009	.049	.000
	1, 2, 4, 5	.030	.027	.007	.029	.001
	1, 3, 4, 5	.022	.026	.008	.030	.002
	1, 2, 3, 4, 5	.010	.007	.001	.009	.000
(2, 0, 0, 4, 8)	1, 4, 5 [‡]	.683	.690	.881	.641	.947
	1, 2, 4, 5	.143	.129	.045	.158	.032
	1, 3, 4, 5	.116	.142	.067	.138	.020
	1, 2, 3, 4, 5	.058	.039	.007	.063	.001
(2, 9, 0, 4, 8)	1, 4, 5	.000	.000	.000	.005	.016
	1, 2, 4, 5 [‡]	.794	.817	.939	.801	.965
	1, 3, 4, 5	.000	.000	.000	.005	.002
	1, 2, 3, 4, 5	.206	.183	.061	.189	.017
(2, 9, 6, 4, 8)	1, 2, 3, 5	.000	.000	.000	.000	.002
	1, 2, 4, 5	.000	.000	.000	.000	.005
	1, 3, 4, 5	.000	.000	.000	.015	.045
	1, 2, 3, 4, 5 [‡]	1.00	1.00	1.00	.985	.948

[†] CV_d stands for the delete-d CV.

[‡] The optimal model.

selection procedures, the AIC, the C_p , and the BIC, are examined. It was shown in Shao (1995a) that the AIC, C_p , and the delete-1 CV are asymptotically equivalent, whereas the BIC and the delete-d CV are asymptotically equivalent when $d = n[1 - (\log n - 1)^{-1}]$. The \mathcal{S} in the delete-25 CV is obtained by taking a random sample of size $2n = 80$ from all possible subsets of $\{1, \dots, 40\}$ of size 25. For these five selection procedures, the empirical probabilities (based on 1000 simulations) of selecting each model are reported in Table 7.3.

The results in Table 7.3 can be summarized as follows.

- (1) In terms of the probability of selecting the optimal model, the delete-25 CV is much better than the delete-1 CV, except for the case where the model with the largest size is optimal.
- (2) For all of the methods, the empirical probabilities of selecting an incorrect model are negligible (equal to 0 in most cases, based on 1000 simulations).
- (3) The delete-1 CV tends to select models with unnecessarily large sizes. The more 0 components the β has, the worse performance the delete-1 CV has. On the other hand, the performance of the delete-25 CV is stable.
- (4) The AIC and the C_p have similar performances to the delete-1 CV. The BIC is better than the delete-1 CV but is still too conservative in some cases and worse than the delete-25 CV.

7.4.2 The bootstrap

Bootstrap model selection procedures can be derived from bootstrap estimators of $\widehat{\text{mse}}(\alpha)$. The bootstrap estimator of the form (7.34), however, cannot be used for model selection since it is not a consistent estimator when α is an incorrect model. Thus, we consider the bootstrap estimator of the form (7.35):

$$\widehat{\text{mse}}_{\text{BOOT}}(\alpha) = \frac{1}{n} \sum_{i=1}^n (y_i - x'_{i\alpha} \hat{\beta}_\alpha)^2 + \hat{e}(\alpha), \quad (7.45)$$

where $\hat{e}(\alpha)$ is a bootstrap estimator of the expected excess error

$$e(\alpha) = E \left[\frac{1}{n} \sum_{i=1}^n (y_{f,i} - x'_{i\alpha} \hat{\beta}_\alpha)^2 - \frac{1}{n} \sum_{i=1}^n (y_i - x'_{i\alpha} \hat{\beta}_\alpha)^2 \right] = \frac{2\sigma^2 p_\alpha}{n}$$

and $y_{f,i}$ is a future response at x_i independent of y_i . Let $\hat{\beta}_\alpha^*$ be the bootstrap analog of $\hat{\beta}_\alpha$, i.e., $\hat{\beta}_\alpha^*$ is given by (7.23) with X and X^* replaced by X_α and X_α^* , respectively, for the PB, and

$$\hat{\beta}_\alpha^* = (X'_\alpha X_\alpha)^{-1} X'_\alpha y_\alpha^*$$

for the RB, where $y_\alpha^* = (y_{1\alpha}^*, \dots, y_{n\alpha}^*)'$, $y_{i\alpha}^* = x'_{i\alpha} \hat{\beta}_\alpha + \varepsilon_i^*$, and the ε_i^* are i.i.d. from \hat{F}_ε defined in Section 7.2.2. Then the bootstrap estimator of $e(\alpha)$ is given by

$$\hat{e}(\alpha) = E_* \left[\frac{1}{n} \sum_{i=1}^n (y_i - x'_{i\alpha} \hat{\beta}_\alpha^*)^2 - \frac{1}{n} \sum_{i=1}^n (y_i^* - x'^*_{i\alpha} \hat{\beta}_\alpha^*)^2 \right] \quad (7.46)$$

for the PB, or is given by (7.46) with $x'_{i\alpha}$ changed to $x_{i\alpha}$ for the RB. Bootstrap estimators of $\overline{\text{mse}}(\alpha)$ are then given by (7.45) with an appropriate $\hat{e}(\alpha)$. An estimator of $\overline{\text{mse}}(\alpha)$ based on the EB can be similarly defined.

For the RB, a straightforward calculation leads to

$$\hat{e}(\alpha) = \frac{2\hat{\sigma}^2 p_\alpha}{n}, \quad (7.47)$$

which is an asymptotically unbiased and consistent estimator of $e(\alpha)$. Note that this estimator can also be derived by substituting $\hat{\sigma}^2$ for σ^2 in the expression of $e(\alpha) = 2\sigma^2 p_\alpha/n$. The $\hat{e}(\alpha)$ obtained by using the PB is approximately equal to the right-hand side of (7.47), but the derivation is more complicated and is given in Section 7.5.5.

Although $\widehat{\text{mse}}_{\text{BOOT}}(\alpha)$ is a reasonably good estimator of $\overline{\text{mse}}(\alpha)$, the model selection procedure based on minimizing $\widehat{\text{mse}}_{\text{BOOT}}(\alpha)$ over $\alpha \in \mathcal{A}$ is asymptotically equivalent to the delete-1 CV: it is inconsistent and too conservative (see Section 7.5.5). The reason for this inconsistency is the same as that for the delete-1 CV explained in Section 7.4.1.

Note that $e(\alpha) = e_n(\alpha)$ depends on the sample size n . Using the idea discussed in Section 7.4.1, we may similarly obtain a consistent bootstrap model selection procedure by first estimating $e_m(\alpha)$ by $\hat{e}_m(\alpha)$, where $m/n \rightarrow 0$, and then minimizing

$$\widehat{\text{mse}}_{\text{BOOT-}m}(\alpha) = \frac{1}{n} \sum_{i=1}^n (y_i - x'_i \hat{\beta}_\alpha)^2 + \hat{e}_m(\alpha) \quad (7.48)$$

over $\alpha \in \mathcal{A}$. We first consider the PB. To estimate $e_m(\alpha)$, we may generate m i.i.d. pairs $(y_1^*, x_1^{*\prime}), \dots, (y_m^*, x_m^{*\prime})$ from the empirical distribution putting mass n^{-1} to each (y_i, x_i') , and then use

$$\hat{e}_m(\alpha) = E_* \left[\frac{1}{n} \sum_{i=1}^n (y_i - x'_{i\alpha} \hat{\beta}_{m,\alpha}^*)^2 - \frac{1}{m} \sum_{i=1}^m (y_i^* - x'^*_{i\alpha} \hat{\beta}_{m,\alpha}^*)^2 \right], \quad (7.49)$$

where $\hat{\beta}_{m,\alpha}^*$ is defined the same way as $\hat{\beta}_\alpha^*$, except that it is based on m pairs of bootstrap data. It was shown in Shao (1995b) that if m is chosen so that $m/n \rightarrow 0$ and $m \rightarrow \infty$, then

$$\hat{e}_m(\alpha) = \frac{2\hat{\sigma}^2 p_\alpha}{m} + o(m^{-1}).$$

Furthermore, model selection by minimizing the quantity given by (7.48)-(7.49) is a consistent selection procedure (see Section 7.5.5). This result is analogous to the result for the CV method described in Section 7.4.1. In fact, using bootstrap sample size $m = n$ is asymptotically equivalent to using the delete-1 CV, whereas using m satisfying $m/n \rightarrow 0$ is asymptotically equivalent to using the delete-d CV with $d = n - m$.

For the RB, we can directly use (7.47) and estimate $e_m(\alpha)$ by $2\hat{\sigma}^2 p_\alpha/m$, which also leads to a consistent bootstrap model selection procedure.

We have seen in Chapter 3 that using a bootstrap sample size m with $m/n \rightarrow 0$ and $m \rightarrow \infty$ produces consistent bootstrap estimators in some nonregular cases. Here, this choice of the bootstrap sample size helps us once again in deriving consistent bootstrap model selection procedures.

7.5 Asymptotic Theory

The asymptotic theory in this section forms a theoretical base for the results in the previous sections of this chapter. Following the order of the discussion in Sections 7.2-7.4, we present some existing asymptotic results for variance estimation, bias estimation, inference, prediction, and model selection.

Unless otherwise specified, we assume model (7.1) with deterministic x_i and independent ε_i throughout this section. It is also assumed that ε_i has mean 0 and variance σ_i^2 , and $0 < \liminf_i \sigma_i^2 \leq \limsup_i \sigma_i^2 < \infty$. The results in this section are still valid in the almost sure sense when the x_i are random, provided that the conditions involving X hold almost surely.

7.5.1 Variance estimators

We mainly consider the case where $\theta = g(\beta)$ and $\hat{\theta}$ is the LSE $\hat{\theta}_{\text{LS}} = g(\hat{\beta}_{\text{LS}})$. The first result is for the three jackknife variance estimators defined in (7.7), (7.14), and (7.17).

Theorem 7.1. *Suppose that g is continuously differentiable at β with a nonzero $\nabla g(\beta)$ and that*

$$X'X \rightarrow \infty \quad \text{and} \quad h_{\max} \rightarrow 0 \tag{7.50}$$

and

$$\limsup_n \max_{i \leq n} E|\varepsilon_i|^{2+\delta} < \infty \tag{7.51}$$

for some $\delta > 0$. Then

$$v / \nabla g(\beta)' \text{var}(\hat{\beta}_{\text{LS}}) \nabla g(\beta) \rightarrow_p 1, \tag{7.52}$$

where $v = v_{\text{JACK}}$ in (7.7), v_{HJACK} in (7.14), or v_{WJACK} in (7.17).

Proof. We only show (7.52) for $v = v_{\text{WJACK}}$. Assume $g(\beta) = c'\beta$. Let $a_i = (c'z_i)^2\sigma_i^2 / \sum_{i=1}^n (c'z_i)^2\sigma_i^2$, where $z_i = (X'X)^{-1}x_i$. Note that $r_i = \varepsilon_i + u_i$, where $u_i = x_i'(\beta - \hat{\beta}_{\text{LS}})$. Then, by (7.18),

$$\frac{v_{\text{WJACK}}}{\text{var}(c'\hat{\beta}_{\text{LS}})} = \sum_{i=1}^n \frac{a_i \varepsilon_i^2}{(1-h_i)\sigma_i^2} + \sum_{i=1}^n \frac{a_i u_i^2}{(1-h_i)\sigma_i^2} + \sum_{i=1}^n \frac{2a_i \varepsilon_i u_i}{(1-h_i)\sigma_i^2}. \quad (7.53)$$

Under conditions (7.50) and (7.51), the first term on the right-hand side of (7.53) $\rightarrow_p 1$, by the law of large numbers (Appendix A.5). The second term on the right-hand side of (7.53) is bounded by

$$(1-h_{\max})^{-1} \max_{i \leq n} \sigma_i^{-2} u_i^2 = O_p(h_{\max}) \rightarrow_p 0,$$

since

$$\begin{aligned} u_i^2 &= [x_i'(X'X)^{-1/2}(X'X)^{1/2}(\beta - \hat{\beta}_{\text{LS}})]^2 \\ &\leq h_i(\beta - \hat{\beta}_{\text{LS}})'(X'X)(\beta - \hat{\beta}_{\text{LS}}) = O_p(h_i). \end{aligned}$$

Then, by the Cauchy-Schwarz inequality, the last term on the right-hand side of (7.53) $\rightarrow_p 0$. This proves (7.52) for the linear case where $g(\beta) = c'\beta$.

Using the arguments in Theorems 2.1 and 2.6, for nonlinear g we only need to show

$$\max_{i \leq n} \|\hat{\beta}_{\text{LS},i} - \hat{\beta}_{\text{LS}}\| \rightarrow_p 0,$$

which follows from

$$E\left(\max_{i \leq n} \|\hat{\beta}_{\text{LS},i} - \hat{\beta}_{\text{LS}}\|^2\right) \leq \sum_{i=1}^n \frac{z_i' z_i E(r_i^2)}{(1-h_i)^2} = O(\text{tr}[(X'X)^{-1}]) \rightarrow 0. \quad \square$$

Note that the result in Theorem 7.1 holds irrespective of whether the σ_i^2 are equal or not. Hence, all three jackknife variance estimators are robust against heteroscedasticity. From the proof of Theorem 7.1, one can conclude that the jackknife variance estimators are also strongly consistent if $\max_{i \leq n} u_i^2 \rightarrow_{a.s.} 0$, which holds, for example, when all of the x_i are within a bounded set.

The consistency of the jackknife and weighted jackknife variance estimators for $\hat{\theta}$ other than the LSE can be established similarly. For example, the consistency of the jackknife variance estimators for the WLSE of the form (7.3) was proved by Shao (1989c) and Shao and Rao (1993); the consistency of the jackknife variance estimators for M-estimators defined by (7.4) was established in Shao (1992c).

The next theorem provides the asymptotic order of the bias and mean squared error of the jackknife variance estimators in the linear case where $\hat{\theta} = c'\hat{\beta}_{\text{LS}}$. Similar results for nonlinear cases can be found in Shao (1988a).

Theorem 7.2. Suppose that $\sup_n h_{\max} < 1$. Let v_{JACK} , v_{HJACK} , and v_{WJACK} be given by (7.10), (7.16), and (7.18), respectively. Then we have the following conclusions:

(i) $\text{bias}(v_{\text{WJACK}}) = 0$ if $\sigma_i^2 = \sigma^2$ for all i , and, in general, $\text{bias}(v_{\text{WJACK}}) = O(\xi_n)$, where $\xi_n = \sum_{i=1}^n (c' z_i)^2 h_i$.

(ii) $\text{bias}(v_{\text{JACK}}) = O(\xi_n) + O(\zeta_n)$, where $\zeta_n = n^{-1} \sum_{i=1}^n (c' z_i)^2$. If $\zeta_n = o(\xi_n)$, then

$$\liminf_n \frac{\text{bias}(v_{\text{JACK}})}{\xi_n} > 0. \quad (7.54)$$

(iii) $\text{bias}(v_{\text{HJACK}}) = O(\xi_n) + O(\zeta_n)$. If $\zeta_n = o(\xi_n)$ and

$$\liminf_n (2 \min_{i \leq n} \sigma_i^2 - \max_{i \leq n} \sigma_i^2) > 0,$$

then

$$\liminf_n \frac{\text{bias}(v_{\text{HJACK}})}{\xi_n} < 0. \quad (7.55)$$

(iv) If $\limsup_n \max_{i \leq n} E|\varepsilon_i|^4 < \infty$, then $\text{var}(v) = O(\eta_n) + O(\xi_n^2)$, where $\eta_n = \sum_{i=1}^n (c' z_i)^4$ and $v = v_{\text{JACK}}$, v_{HJACK} , or v_{WJACK} .

Proof. (i) Let $h_{ij} = x'_i (X' X)^{-1} x_j$. Then

$$E(r_i^2) = (1 - h_i)^2 \sigma_i^2 + \sum_{j \neq i} h_{ij}^2 \sigma_j^2 = (1 - h_i) \sigma_i^2 + \sum_{j=1}^n h_{ij}^2 (\sigma_j^2 - \sigma_i^2).$$

Hence,

$$\text{bias}(v_{\text{WJACK}}) = \sum_{i=1}^n \frac{(c' z_i)^2}{1 - h_i} \sum_{j=1}^n h_{ij}^2 (\sigma_j^2 - \sigma_i^2),$$

which is 0 if $\sigma_i^2 = \sigma^2$ for all i , and is of the order $O(\xi_n)$ in general since $\sum_{j=1}^n h_{ij}^2 = h_i$.

(ii) From the proof of (i), we conclude that

$$E \left[\frac{1}{n} \sum_{i=1}^n \frac{(c' z_i)^2 r_i^2}{(1 - h_i)^2} \right] = O(\zeta_n) \quad (7.56)$$

and

$$E \left[\frac{n-1}{n^2} \left(\sum_{i=1}^n \frac{c' z_i r_i}{1 - h_i} \right)^2 \right] = O(\zeta_n).$$

Then, by (7.10),

$$\begin{aligned} E(v_{\text{JACK}}) &= \sum_{i=1}^n \frac{(c' z_i)^2 E(r_i^2)}{(1 - h_i)^2} + O(\zeta_n) \\ &= \text{var}(c' \hat{\beta}_{\text{LS}}) + \sum_{i=1}^n \frac{(c' z_i)^2}{(1 - h_i)^2} \sum_{j \neq i} h_{ij}^2 \sigma_j^2 + O(\zeta_n). \end{aligned} \quad (7.57)$$

The results in (ii) follow from (7.57) and

$$\frac{1}{\xi_n} \sum_{i=1}^n \frac{(c' z_i)^2}{(1-h_i)^2} \sum_{j \neq i} h_{ij}^2 \sigma_j^2 \geq (1-h_{\max}) \min_{i \leq n} \sigma_i^2.$$

(iii) By (7.16) and (7.56),

$$E(v_{\text{HJACK}}) = \sum_{i=1}^n (c' z_i)^2 E(r_i^2) + O(\zeta_n). \quad (7.58)$$

The results in (iii) follow from (7.58) and

$$\sigma_i^2 - E(r_i^2) = 2h_i \sigma_i^2 - \sum_{j=1}^n h_{ij}^2 \sigma_j^2 \geq h_i (2 \min_{i \leq n} \sigma_i^2 - \max_{i \leq n} \sigma_i^2).$$

(iv) The results in (iv) follow from the fact that $\text{var}(r_i^2) = O(1)$ and $\text{cov}(r_i^2, r_j^2) = O(\sqrt{h_i h_j})$ for $i \neq j$. \square

Some direct conclusions from Theorem 7.2 are (1) if $h_{\max} \rightarrow 0$, then all three jackknife variance estimators are asymptotically unbiased and consistent in mse; (ii) if $\zeta_n = o(\xi_n)$, then v_{JACK} is upward-biased [result (7.54)] and v_{HJACK} is downward-biased [result (7.55)]; (iii) if ζ_n and ξ_n are of the same order, then so are the biases of the three jackknife variance estimators.

Intuitively, the asymptotic properties of the three jackknife variance estimators are similar when model (7.1) is balanced, i.e., the h_i are close to each other so that ζ_n and ξ_n are almost the same. Results (7.54) and (7.55) occur in situations where the h_i are quite different. The following is an example.

Example 7.2. Orders of ξ_n and ζ_n . Consider model (7.1) with $p = 1$ and $x_1 = \sqrt{n}$, $x_i = 1$, $i = 2, \dots, n$. Then $z_1 = \sqrt{n}/(2n-1)$, $h_1 = n/(2n-1)$, $z_i = 1/(2n-1)$, and $h_i = 1/(2n-1)$, $i = 2, \dots, n$. Thus, $\xi_n = \sum_{i=1}^n z_i^2 h_i = (n^2 + n - 1)/(2n-1)^3$ is of the order n^{-1} and $\zeta_n = n^{-1} \sum_{i=1}^n z_i^2 = [n(2n-1)]^{-1}$ is of the order n^{-2} .

We now consider the bootstrap variance estimators described in Section 7.2.2. We focus on the linear case. The discussion for the nonlinear case is similar to that in Section 3.2.2. For $c' \hat{\beta}_{\text{LS}}$, the RB variance estimator is

$$v_{\text{RBOOT}} = \hat{\sigma}^2 c'(X'X)^{-1} c \text{ or } \tilde{\sigma}^2 c'(X'X)^{-1} c,$$

where $\hat{\sigma}^2$ (or $\tilde{\sigma}^2$) is defined in (7.21) [or (7.22)]; the PB variance estimator is

$$v_{\text{PBOOT}} = c' \text{var}_*(\hat{\beta}_{\text{LS}}^*) c,$$

where $\hat{\beta}_{\text{LS}}^*$ is defined in (7.23); and the EB produces the same variance estimator as Wu's weighted jackknife, v_{WJACK} .

Theorem 7.3. Assume condition (7.50).

(i) If $\sigma_i^2 = \sigma^2$ for all i , then $v_{\text{RBOOT}}/\text{var}(c'\hat{\beta}_{\text{LS}}) \rightarrow_p 1$. Furthermore, $\text{bias}(v_{\text{RBOOT}})/\text{var}(c'\hat{\beta}_{\text{LS}}) = O(n^{-1})$ and $\text{var}(v_{\text{RBOOT}})/[\text{var}(c'\hat{\beta}_{\text{LS}})]^2 = O(n^{-1})$ [assuming $E(\varepsilon_i^4) < \infty$].

(ii) If condition (7.51) holds, then

$$v_{\text{PBOOT}} = \sum_{i=1}^n (c' z_i)^2 r_i^2 + o_p(\text{var}(c'\hat{\beta}_{\text{LS}})). \quad (7.59)$$

Proof. (i) Note that $c'(X'X)^{-1}c$ and $\text{var}(c'\hat{\beta}_{\text{LS}})$ are of the same order, and

$$v_{\text{RBOOT}} - \text{var}(c'\hat{\beta}_{\text{LS}}) = \sum_{i=1}^n (c' z_i)^2 (\bar{\sigma}^2 - \sigma_i^2) + o_p(c'(X'X)^{-1}c), \quad (7.60)$$

where $\bar{\sigma}^2 = n^{-1} \sum_{i=1}^n \sigma_i^2$. Hence, the first assertion in (i) follows. The results for the bias and variance of v_{RBOOT} can be obtained by directly calculating the mean and variance of $\sum_{i=1}^n (r_i - \bar{r})^2$.

(ii) On the set $\{\lambda^* \geq \lambda/2\}$, we have

$$\begin{aligned} \text{var}_*(\hat{\beta}_{\text{LS}}^*) &= \text{var}_*[(X^{*\prime} X^*)^{-1} X^{*\prime} y^* - \hat{\beta}_{\text{LS}}] \\ &= \text{var}_*[(X^{*\prime} X^*)^{-1} \sum_{i=1}^n x_i^* (y^* - x_i^{*\prime} \hat{\beta}_{\text{LS}})] \\ &= \text{var}_*[(X'X)^{-1} \sum_{i=1}^n x_i^* (y^* - x_i^{*\prime} \hat{\beta}_{\text{LS}})] + o_p((X'X)^{-1}) \\ &= (X'X)^{-1} \sum_{i=1}^n x_i x_i' r_i^2 (X'X)^{-1} + o_p((X'X)^{-1}). \end{aligned}$$

Therefore, (7.59) follows from $P\{\lambda^* \geq \lambda/2\} \rightarrow 1$. \square

Result (7.59) shows that v_{PBOOT} is asymptotically equivalent to v_{HJACK} . Hence, it is consistent and robust against heteroscedasticity. From (7.60), we conclude that the bootstrap estimator v_{RBOOT} is inconsistent if the σ_i^2 are different. But the results in Theorem 7.3(i) indicate that v_{RBOOT} is consistent and, in fact, more efficient than v_{PBOOT} when $\sigma_i^2 = \sigma^2$ for all i (see also the discussion in Section 7.2.3).

Finally, we consider the model with repeated measurements:

$$y_{ij} = x_i' \beta + \varepsilon_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, n, \quad (7.61)$$

where $y_i = (y_{i1}, \dots, y_{in_i})'$ are independent and $\text{var}(y_{ij}) = \sigma_i^2$, but the components of y_i may be correlated. The jackknife, the weighted jackknife,

and the PB can be applied by treating (y'_i, x'_i) as pairs. The weighted jackknife variance estimator v_{WJACK} can be defined by (7.17) with $h_i = n_i x'_i (X' X)^{-1} x_i$. The following theorem shows the consistency and robustness (against heteroscedasticity and within-group correlations) of v_{WJACK} . Results for the unweighted jackknife and the PB variance estimators can be similarly established.

Theorem 7.4. *Assume model (7.61) and the conditions in Theorem 7.1. Then*

$$v_{\text{WJACK}} / \nabla g(\beta)' \text{var}(\hat{\beta}_{\text{LS}}) \nabla g(\beta) \rightarrow_p 1. \quad (7.62)$$

Proof. Similar to the proof of Theorem 7.1, we only need to show (7.62) for the linear case where $g(\beta) = c'\beta$. A straightforward calculation shows that $\hat{\beta}_{\text{LS}} - \hat{\beta}_{\text{LS},i} = z_i \tilde{r}_i / (1 - h_i)$, where $\tilde{r}_i = \sum_{j=1}^{n_i} (y_{ij} - x'_{ij} \hat{\beta}_{\text{LS}})$. Then

$$v_{\text{WJACK}} = \sum_{i=1}^n \frac{(c' z_i)^2 \tilde{r}_i^2}{1 - h_i}.$$

The rest of the proof is the same as that of Theorem 7.1. \square

7.5.2 Bias estimators

In this section, we show the consistency of the weighted jackknife bias estimator (7.15) and the possible inconsistency of the unweighted jackknife bias estimator (7.8) for the LSE $\hat{\theta}_{\text{LS}} = g(\hat{\beta}_{\text{LS}})$. The consistency of the bootstrap bias estimators can be similarly established (Shao, 1988c).

Assuming that g is second order continuously differentiable at β and using Taylor's expansion, we obtain that

$$\hat{\theta}_{\text{LS}} - \theta = \nabla g(\beta)' (\hat{\beta}_{\text{LS}} - \beta) + \frac{1}{2} (\hat{\beta}_{\text{LS}} - \beta)' \nabla^2 g(\beta) (\hat{\beta}_{\text{LS}} - \beta) + R_n,$$

where $R_n = o_p(\|\hat{\beta}_{\text{LS}} - \beta\|^2)$. Thus, according to the discussion in Section 2.4.1, the asymptotic bias of $\hat{\theta}_{\text{LS}}$ is

$$b_n = \frac{1}{2} \text{tr}[\nabla^2 g(\beta) \text{var}(\hat{\beta}_{\text{LS}})] = \frac{1}{2} \sum_{i=1}^n z'_i \nabla^2 g(\beta) z_i \sigma_i^2. \quad (7.63)$$

The following result is similar to Theorem 2.15.

Theorem 7.5. *Assume the conditions in Theorem 7.1. Then $b_{\text{WJACK}}/b_n \rightarrow_p 1$, where b_n is given by (7.63).*

Proof. By (7.9) and the continuity of $\nabla^2 g$,

$$\begin{aligned} \hat{\theta}_{\text{LS},i} - \hat{\theta}_{\text{LS}} &= \nabla g(\hat{\beta}_{\text{LS}}) (\hat{\beta}_{\text{LS},i} - \hat{\beta}_{\text{LS}}) \\ &\quad + \frac{1}{2} (\hat{\beta}_{\text{LS},i} - \hat{\beta}_{\text{LS}})' \nabla^2 g(\hat{\beta}_{\text{LS}}) (\hat{\beta}_{\text{LS},i} - \hat{\beta}_{\text{LS}}) + R_{n,i}, \end{aligned} \quad (7.64)$$

where $R_{n,i} = o_p(\|\hat{\beta}_{\text{LS},i} - \hat{\beta}_{\text{LS}}\|^2)$ uniformly in i . From (7.9), (7.12), and (7.64), we conclude that

$$b_{\text{WJACK}} = \frac{1}{2} \sum_{i=1}^n \frac{z_i' \nabla^2 g(\hat{\beta}_{\text{LS}}) z_i r_i^2}{1 - h_i} + \sum_{i=1}^n (1 - h_i) R_{n,i}.$$

By the same argument used in the proof of Theorem 7.1 and the continuity of $\nabla^2 g$,

$$\sum_{i=1}^n \frac{z_i' \nabla^2 g(\hat{\beta}_{\text{LS}}) z_i r_i^2}{1 - h_i} / \sum_{i=1}^n z_i' \nabla^2 g(\beta) z_i \sigma_i^2 \rightarrow_p 1.$$

The result follows from

$$\sum_{i=1}^n (1 - h_i) R_{n,i} = o_p \left(\sum_{i=1}^n (1 - h_i) (\hat{\beta}_{\text{LS},i} - \hat{\beta}_{\text{LS}})' \nabla^2 g(\hat{\beta}_{\text{LS}}) (\hat{\beta}_{\text{LS},i} - \hat{\beta}_{\text{LS}}) \right). \quad \square$$

Using expansion (7.64), we can analyze the asymptotic behavior of the unweighted jackknife bias estimator b_{JACK} . From (7.8), (7.64), and the proof of Theorem 7.5,

$$b_{\text{JACK}} = \frac{n-1}{n} \nabla g(\hat{\beta}_{\text{LS}}) \sum_{i=1}^n (\hat{\beta}_{\text{LS},i} - \hat{\beta}_{\text{LS}}) + \tilde{b}_n,$$

where \tilde{b}_n satisfies $\tilde{b}_n/b_n \rightarrow_p 1$. Therefore, b_{JACK} is consistent if and only if

$$b_{\text{JACK}} - \tilde{b}_n = -\frac{n-1}{n} \nabla g(\hat{\beta}_{\text{LS}}) \sum_{i=1}^n \frac{h_i r_i z_i}{1 - h_i} = o_p(b_n). \quad (7.65)$$

Let $\tau_n = \text{tr}(X'X)^{-1}$. Since $E(r_i^2) = O(1)$ and $E(r_i r_j) = O(h_{ij})$ for $i \neq j$,

$$E \left\| \sum_{i=1}^n \frac{h_i r_i z_i}{1 - h_i} \right\|^2 = E \left[\sum_{i=1}^n \sum_{j=1}^n \frac{h_i h_j r_i r_j z_i' z_j}{(1 - h_i)(1 - h_j)} \right] = O(h_{\max}^2 \tau_n).$$

Hence, $b_{\text{JACK}} - \tilde{b}_n = O_p(h_{\max} \tau_n^{1/2})$ and (7.65) holds if $h_{\max} \tau_n^{1/2}/b_n \rightarrow 0$. In general, the order of b_n in (7.63) is τ_n ; e.g., if $p = 1$, then

$$\frac{1}{2} \min_{i \leq n} \sigma_i^2 |\nabla^2 g(\beta)| \tau_n \leq |b_n| \leq \frac{1}{2} \max_{i \leq n} \sigma_i^2 |\nabla^2 g(\beta)| \tau_n.$$

Thus, $b_{\text{JACK}} - \tilde{b}_n = O_p(h_{\max} b_n^{1/2})$ and (7.65) holds if $h_{\max}/b_n^{1/2} \rightarrow 0$.

Further discussion about the jackknife bias estimators can be found in Shao (1988a).

7.5.3 Bootstrap distribution estimators

We first study bootstrap estimators of the distribution of the LSE. Since $\hat{\beta}_{\text{LS}}$ is linear, we may use Mallows' distance $\tilde{\rho}_2$ defined in Section 3.1.2. For two random vectors U and V with i.i.d. components and a deterministic matrix A , it was shown by Freedman (1981) that

$$[\tilde{\rho}_2(AU, AV)]^2 \leq \text{tr}(AA')[\tilde{\rho}_2(U_1, V_1)]^2, \quad (7.66)$$

where U_1 and V_1 are the first components of U and V , respectively. This inequality and other properties of Mallows' distance discussed in Section 3.1.2 are used in the proof of the following result for the consistency of the distribution estimator based on the RB.

Theorem 7.6. *Assume that (7.50) holds and the ε_i are i.i.d. Then*

$$\tilde{\rho}_2(H_{\text{BOOT}}, H_n) \rightarrow_{a.s.} 0, \quad (7.67)$$

where H_n is the distribution of $(X'X)^{1/2}(\hat{\beta}_{\text{LS}} - \beta)$, H_{BOOT} is the bootstrap distribution of $(X'X)^{1/2}(\hat{\beta}_{\text{LS}}^* - \hat{\beta}_{\text{LS}})$, and $\hat{\beta}_{\text{LS}}^*$ is given by (7.19).

Proof. Let $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ and $\varepsilon^* = (\varepsilon_1^*, \dots, \varepsilon_n^*)'$. Using inequality (7.66), we obtain that

$$\tilde{\rho}_2(H_{\text{BOOT}}, H_n) = \tilde{\rho}_2((X'X)^{-1/2}X'\varepsilon^*, (X'X)^{-1/2}X'\varepsilon) \leq \tilde{\rho}_2(\varepsilon_1^*, \varepsilon_1).$$

Hence, (7.67) follows from

$$\tilde{\rho}_2(\varepsilon_1^*, \varepsilon_1) = \tilde{\rho}_2(\hat{F}_\varepsilon, F_\varepsilon) \rightarrow_{a.s.} 0. \quad (7.68)$$

Let $\tilde{\varepsilon}_1$ be a random variable drawn from the empirical distribution of $\varepsilon_1, \dots, \varepsilon_n$ and $\bar{\varepsilon} = n^{-1} \sum_{i=1}^n \varepsilon_i$. From (3.3),

$$\tilde{\rho}_2(\tilde{\varepsilon}_1, \varepsilon_1) \rightarrow_{a.s.} 0.$$

Using equality (3.5) twice, we have

$$\begin{aligned} [\tilde{\rho}_2(\varepsilon_1^*, \tilde{\varepsilon}_1)]^2 &= \bar{\varepsilon}^2 + [\tilde{\rho}_2(\varepsilon_1^*, \tilde{\varepsilon}_1 - E\tilde{\varepsilon}_1)]^2 \\ &= \bar{\varepsilon}^2 + [\tilde{\rho}_2(\varepsilon_1^* + \bar{\varepsilon}, \tilde{\varepsilon}_1)]^2 - (\bar{r} - \bar{\varepsilon})^2. \end{aligned}$$

From the strong law of large numbers, $\bar{\varepsilon} \rightarrow_{a.s.} 0$. By the definition of Mallows' distance,

$$[\tilde{\rho}_2(\varepsilon_1^* + \bar{\varepsilon}, \tilde{\varepsilon}_1)]^2 \leq \frac{1}{n} \sum_{i=1}^n (r_i - \varepsilon_i)^2 = \frac{\varepsilon' X(X'X)^{-1}X'\varepsilon}{n} \rightarrow_{a.s.} 0$$

(see Lai, Robbins and Wei, 1979). Also,

$$(\bar{r} - \bar{\varepsilon})^2 \leq \frac{1}{n} \sum_{i=1}^n (r_i - \varepsilon_i)^2 \rightarrow_{a.s.} 0.$$

Therefore, (7.68) holds and the proof is completed. \square

Navidi (1989) studied the asymptotic accuracy of the estimator of the distribution of $c'_n(\hat{\beta}_{\text{LS}} - \beta)$ based on the RB, where $c_n \in \mathbb{R}^p$. We still denote the distribution of $c'_n(\hat{\beta}_{\text{LS}} - \beta)$ by H_n and its bootstrap estimator by H_{BOOT} . Without loss of generality, we assume that $c'_n(X'X)^{-1}c_n = 1$. Denote the i th component of $c'_n(X'X)^{-1}X'$ by a_i .

Theorem 7.7. *Assume the conditions of Theorem 7.6 and that F_ε is non-lattice with a finite eighth moment. Assume further that*

$$\max\{i : 1 \leq i \leq n, |a_i| \sqrt{\lambda_n} > \delta\} / \log \lambda_n \rightarrow \infty$$

for some $\delta > 0$, where $\lambda_n = (\max_{i \leq n} a_i^2)^{-1}$. Then

$$\sup_x |H_{\text{BOOT}}(x) - H_n(x)| = O_p(n^{-1/2}) + o_p(\lambda_n^{-1/2}).$$

Proof. The proof is similar to that of Theorem 3.11 but more difficult. We only sketch its main steps. First, we can show that H_n has the expansion

$$H_n(x) = \Phi\left(\frac{x}{\sigma}\right) + \frac{\gamma_n \mu_3}{16\sigma^3} \left(1 - \frac{x^2}{\sigma^2}\right) \varphi\left(\frac{x}{\sigma}\right) + o(\lambda_n^{-1/2}),$$

where $\mu_3 = E(\varepsilon_1^3)$ and $\gamma_n = \sum_{i=1}^n a_i^3$. Second, we can show that H_{BOOT} has a similar expansion

$$H_{\text{BOOT}}(x) = \Phi\left(\frac{x}{\hat{\sigma}}\right) + \frac{\gamma_n \hat{\mu}_3}{16\hat{\sigma}^3} \left(1 - \frac{x^2}{\hat{\sigma}^2}\right) \varphi\left(\frac{x}{\hat{\sigma}}\right) + o(\lambda_n^{-1/2}) \quad a.s.,$$

where $\hat{\mu}_3$ is the third moment of \hat{F}_ε . Then the result follows from the fact that $\hat{\mu}_3$ and $\hat{\sigma}$ converge to μ_3 and σ , respectively, at the rate $n^{-1/2}$. \square

As a comparison, we consider the traditional normal approximation to $H_n(x)$, given by $\Phi\left(\frac{x}{\sigma}\right)$. It can be shown that

$$\sup_x |H_n(x) - \Phi\left(\frac{x}{\sigma}\right)| = O_p(n^{-1/2}) + O_p(\gamma_n) + o_p(\lambda_n^{-1/2}).$$

Thus, the bootstrap approximation to H_n is always as accurate as the normal approximation and is better in situations where $\lambda_n/n \rightarrow 0$ and γ_n and $\lambda_n^{-1/2}$ are of the same order. For example, suppose that the elements in the first $n - \sqrt{n}$ rows of X are within a bounded interval and the elements

in the last \sqrt{n} rows are of the order $n^{1/4}$. Then, for most choices of contrast vector c_n , the first $n - \sqrt{n}$ of a_i are of the order $n^{-1/2}$ and the rest of a_i are of the order $n^{-1/4}$. Thus, both $\lambda_n^{-1/2}$ and γ_n are of the order $n^{-1/4}$; the convergence rate of the bootstrap estimator is $o_p(n^{-1/4})$, whereas the convergence rate of the normal approximation is $O_p(n^{-1/4})$.

It is expected that the bootstrap estimator of the distribution of a studentized statistic, which is based on $c'_n \hat{\beta}_{\text{LS}}$ and its variance estimator, is more accurate than the traditional normal approximation (see the discussion in Section 3.3). But only a few rigorous results of this type exist (Huet and Jolivet, 1989) because of the technical difficulties.

For the PB, consistency of the bootstrap distribution estimator for $\hat{\beta}_{\text{LS}}$ can be established in the same manner as in Chapter 3, under the assumption that (y_i, x'_i) are i.i.d. The following result was proved in Freedman (1981). Also, see Stute (1990) for the same result with weaker conditions.

Theorem 7.8. Suppose that (y_i, x'_i) are i.i.d. with $E\|(y_1, x'_1)\|^4 < \infty$, $E(x_1 x'_1) > 0$, and $E(\varepsilon_1 | x_1) = 0$. Then $\|H_{\text{BOOT}} - H_n\|_\infty \rightarrow_{a.s.} 0$, where H_n is given in Theorem 7.6 and H_{BOOT} is the conditional distribution of $(X' X)^{1/2} (\hat{\beta}_{\text{LS}}^* - \hat{\beta}_{\text{LS}})$ under the PB sampling.

The consistency of the EB distribution estimator can be established by using the imitation method (Section 3.1.4).

Theorem 7.9. Assume (7.50) and (7.51). Then $\|H_{\text{BOOT}} - H_n\|_\infty \rightarrow_p 0$, where H_n is given in Theorem 7.6 and H_{BOOT} is the conditional distribution of $(X' X)^{1/2} (\hat{\beta}_{\text{LS}}^* - \hat{\beta}_{\text{LS}})$ under the EB sampling.

Proof. Using conditions (7.50) and (7.51), we can show that Lindeberg's condition holds for $(X' X)^{1/2} (\hat{\beta}_{\text{LS}} - \beta) = (X' X)^{-1/2} X' \varepsilon$, treated as a linear combination of ε . Thus, by the central limit theorem (see Appendix A.8),

$$\{\text{var}[(X' X)^{-1/2} X' \varepsilon]\}^{1/2} (X' X)^{-1/2} X' \varepsilon \rightarrow_d N(0, I).$$

Similarly, under the bootstrap sampling, Lindeberg's condition holds for $(X' X)^{1/2} (\hat{\beta}_{\text{LS}}^* - \hat{\beta}_{\text{LS}}) = (X' X)^{-1/2} X' \varepsilon^*$, where $\varepsilon^* = (\hat{\sigma}_1 e_1^*, \dots, \hat{\sigma}_n e_n^*)'$, $\hat{\sigma}_i = |r_i|/\sqrt{1 - h_i}$, and e_i^* are given in (7.24). Hence,

$$\{\text{var}_*[(X' X)^{-1/2} X' \varepsilon^*]\}^{1/2} (X' X)^{-1/2} X' \varepsilon^* \rightarrow_d N(0, I).$$

Since

$$\text{var}[(X' X)^{-1/2} X' \varepsilon] = (X' X)^{-1/2} \sum_{i=1}^n x_i x'_i \sigma_i^2 (X' X)^{-1/2}$$

and

$$\text{var}_*[(X'X)^{-1/2} X' \varepsilon^*] = (X'X)^{-1/2} \sum_{i=1}^n x_i x'_i \hat{\sigma}_i^2 (X'X)^{-1/2},$$

the result follows from the consistency of $\text{var}_*[(X'X)^{-1/2} X' \varepsilon^*]$. \square

From the proof of Theorem 7.9, the EB estimator is also consistent for the distribution of standardized or studentized $\hat{\beta}_{\text{LS}}$, and it may be more accurate than the traditional normal approximation. Liu (1988) argued heuristically that the EB estimator is second order accurate if $E_*(e_i^{*3}) = 1$. In fact, the EB is almost equivalent to a random weighting procedure discussed in Chapter 10 that was shown to be second order accurate.

Neither Theorem 7.8 nor 7.9 requires that the σ_i^2 are equal. Hence, the PB and the EB distribution estimators are robust against heteroscedasticity.

We now briefly discuss bootstrap distribution estimators for other $\hat{\theta}$. Apparently, the consistency of the bootstrap distribution estimators can be established using Theorems 7.6, 7.8, and 7.9 and Taylor's expansion when $\hat{\theta} = g(\hat{\beta}_{\text{LS}})$ with a differentiable g . For M-estimators given by (7.4), Shorack (1982) proved the consistency of the bootstrap distribution estimator described in Section 7.2.2 when the ε_i are i.i.d. For M-estimator $\hat{\beta}_{\text{M}}$ defined as a solution of

$$\sum_{i=1}^n x_i \psi(y_i - x'_i \beta) = 0, \quad (7.69)$$

where ψ is differentiable, $E\psi(\varepsilon_i) = 0$, and the ε_i are i.i.d., we can define $\hat{\beta}_{\text{M}}^*$ as a solution of

$$\sum_{i=1}^n x_i \left[\psi(y_i^* - x'_i \beta) - \frac{1}{n} \sum_{j=1}^n \psi(y_j - x'_j \hat{\beta}_{\text{M}}) \right] = 0,$$

where $y_i^* = x'_i \hat{\beta}_{\text{M}} + \varepsilon_i^*$, the ε_i^* are i.i.d. from the empirical distribution putting mass n^{-1} to $y_i - x'_i \hat{\beta}_{\text{M}} - t_n$, $i = 1, \dots, n$, and t_n is a solution of

$$\sum_{i=1}^n \psi(y_i - x'_i \hat{\beta}_{\text{M}} - t_n) = 0.$$

The consistency of the bootstrap distribution estimator based on this procedure was established by He (1987) and Mammen (1989). Lahiri (1992c) showed the second order accuracy of the bootstrap estimators of the distributions of M-estimators.

To conclude this section, we discuss some results for the case where p in model (7.1) depends on n and $p \rightarrow \infty$ as $n \rightarrow \infty$. Bickel and

Freedman (1983) proved that the RB produces a consistent estimator of the distribution of $\hat{\beta}_{\text{LS}}$ in situations where the distribution of $\hat{\beta}_{\text{LS}}$ has no asymptotic limit. Mammen (1989) extended their result to M-estimators defined by (7.69). Suppose that the ε_i are i.i.d. and that ψ in (7.69) is bounded and has three bounded derivatives. Let $\mathfrak{R}_n = (X'X)^{1/2}(\hat{\beta}_{\text{M}} - \beta)$ and $\mathfrak{R}_n^* = (X'X)^{1/2}(\hat{\beta}_{\text{M}}^* - \hat{\beta}_{\text{M}})$. If $h_{\max} n^{1/3} (\log n)^{2/3} \rightarrow 0$, then

$$\tilde{\rho}(c_n' \mathfrak{R}_n^*, c_n' \mathfrak{R}_n) \rightarrow_p 0$$

for any bounded p -vector c_n , where $\tilde{\rho}$ is a modified Mallows distance defined by

$$\tilde{\rho}(F, G) = \inf_{T_{X,Y}} (E \min\{\|X - Y\|^2, 1\})^{1/2}$$

and $T_{X,Y}$ is the collection of all possible joint distributions of the pairs (X, Y) whose marginal distributions are F and G , respectively. If, in addition, $p^2/n \rightarrow 0$, then

$$\tilde{\rho}(\mathfrak{R}_n^*, \mathfrak{R}_n) \rightarrow_p 0.$$

If $p^2/n \rightarrow 0$ but $p h_{\max}$ does not converge to 0, then it can be shown that the distribution of \mathfrak{R}_n does not have a limit under the $\tilde{\rho}$ distance. Thus, the bootstrap works in some cases where the traditional normal approximation fails.

7.5.4 Inference and prediction

Bootstrap confidence sets and simultaneous confidence intervals described in Sections 7.3.1 and 7.3.2 can be justified by the consistency of the bootstrap distribution estimators. The accuracy of these confidence sets, however, is difficult to study (it involves Cornish-Fisher expansions) and requires more conditions on X and the distribution of ε_i . To illustrate these conditions, we state without proof the following theorem in Hall (1989c).

Theorem 7.10. *Consider model (7.25) with scalar t_i and i.i.d. ε_i . Then the assertions in Section 7.3.1 hold if $E|\varepsilon_i|^{48+\delta} < \infty$, S_{tt}/n is bounded away from 0, $|t_i - \bar{t}|$ is bounded, and $n_{\pm} > n^c$ for all sufficiently large n and some $c > 0$, where n_{\pm} is the number of i such that $\pm(t_i - \bar{t}) \geq c$.*

The next theorem concerns the bootstrap tests described in Section 7.3.3. Define

$$T_{\beta} = \frac{(n-p)\|(H - H_C)(y - X\beta)\|^2}{q[y'y - y'X(X'X)^{-1}X'y]},$$

where H and H_C are the projection matrices of model (7.1) and model (7.31), respectively.

Theorem 7.11. (i) Assume that the ε_i are i.i.d., $E(\varepsilon_i^4) < \infty$, and $p/n \rightarrow 0$. Let H_n be the distribution of $\sqrt{q}T_\beta$ and H_{BOOT} be the bootstrap distribution of $\sqrt{q}T^*$, where T^* is given by (7.33) under the RB sampling described in Section 7.3.3. Then

$$\|H_{\text{BOOT}} - H_n\|_\infty \rightarrow_p 0. \quad (7.70)$$

(ii) Assume that $\sqrt{qp}/n \rightarrow 0$ and that $p^{1+\delta}/n \rightarrow 0$ for a fixed $\delta > 0$. Assume further that (y_i, x'_i) are i.i.d. with $E\|(y_1, x'_1)\|^2 < \infty$, $E(\varepsilon_1|x_1) = 0$,

$$\sup_n \sup_{\|d\|=1, Cd \neq h} E[\varepsilon_i^4 (d' x_i)^4] < \infty,$$

and

$$\sup_n \sup_{\|d\|=1} E[(1 + \varepsilon_i^2)(d' x_i)^{4k}] < \infty,$$

where k is the smallest integer greater than or equal to $2/\delta$. Let H_n be the distribution of T_β and H_{BOOT} be the bootstrap distribution of T^* under the PB or the EB sampling described in Section 7.3.3. Then (7.70) holds.

The proof can be found in Mammen (1993). Note that $T_\beta = T$ in (7.30) under the hypothesis H_0 . Hence, Theorem 7.11 shows the asymptotic correctness of the bootstrap tests. Part (ii) of the theorem also shows the robustness of the PB and EB tests against heteroscedasticity. When p is fixed, the conditions $p/n \rightarrow 0$ in part (i) and $p^{1+\delta}/n \rightarrow 0$ and $\sqrt{qp}/n \rightarrow 0$ in part (ii) hold automatically, but the theorem remains valid when $p \rightarrow \infty$. Note that the validity of the traditional F-distribution approximation to H_n requires the additional condition that the maximal diagonal element of $H - H_C$ converges to 0 for the situations described by Theorem 7.11(i), or that $E(\varepsilon_i^2|x_i) = E(\varepsilon_i^2)$ for the situations described by Theorem 7.11(ii). Therefore, the bootstrap tests are asymptotically valid even when the traditional F-approximation is not.

The last result in this section shows some asymptotic properties of the bootstrap estimator of the mean squared prediction error given by (7.35).

Theorem 7.12. Suppose that the ε_i are i.i.d. and that (7.50) holds.

(i) If model (7.1) is correct, then

$$\widehat{\text{mse}}_{\text{BOOT}} = \frac{\|\varepsilon\|^2}{n} + \frac{2\sigma^2 p}{n} - \frac{\varepsilon' H \varepsilon}{n} + o_p(n^{-1}).$$

(ii) If model (7.1) is incorrect in the sense that $E(y) \neq X\beta$, then

$$\widehat{\text{mse}}_{\text{BOOT}} = \frac{\|\varepsilon\|^2}{n} + \frac{\|(I - H)E(y)\|^2}{n} + o_p(1).$$

Proof. From the proof of Theorem 7.3(ii) we know that

$$\text{var}_*(\hat{\beta}_{\text{LS}}^*) = (X'X)^{-1} \sum_{i=1}^n x_i x_i' r_i^2 (X'X)^{-1} + o_p((X'X)^{-1}).$$

From the definition of $\hat{\beta}_{\text{LS}}^*$ and \hat{e} , we obtain that

$$\begin{aligned} \hat{e} &= E_* \left\{ \frac{2}{n} \sum_{i=1}^n (y_i^* - x_i^{*\prime} \hat{\beta}_{\text{LS}}) x_i^{*\prime} (\hat{\beta}_{\text{LS}} - \hat{\beta}_{\text{LS}}^*) \right. \\ &\quad \left. + \frac{1}{n} (\hat{\beta}_{\text{LS}} - \hat{\beta}_{\text{LS}}^*)' [(X'X) - (X^{*\prime} X^*)] (\hat{\beta}_{\text{LS}} - \hat{\beta}_{\text{LS}}^*) \right\} \\ &= \frac{2}{n} E_* [(\hat{\beta}_{\text{LS}} - \hat{\beta}_{\text{LS}}^*)' (X^{*\prime} X^*) (\hat{\beta}_{\text{LS}} - \hat{\beta}_{\text{LS}}^*)] + o_p(n^{-1}) \\ &= \frac{2}{n} \text{tr}[(X'X) \text{var}_*(\hat{\beta}_{\text{LS}}^*)] + o_p(n^{-1}) \\ &= \frac{2}{n} \sum_{i=1}^n h_i r_i^2 + o_p(n^{-1}), \end{aligned}$$

which is equal to $2\sigma^2 p/n + o_p(n^{-1})$ when model (7.1) is correct and is equal to $O_p(h_{\max})$ when model (7.1) is incorrect. The results follow from

$$\frac{1}{n} \sum_{i=1}^n r_i^2 = \frac{\|\varepsilon\|^2}{n} + \frac{\|(I - H)E(y)\|^2}{n} - \frac{\varepsilon' H \varepsilon}{n} + \frac{2\varepsilon'(I - H)E(y)}{n}$$

and the fact that $E(\varepsilon' H \varepsilon) = \sigma^2 p$ and $(I - H)E(y) = 0$ when model (7.1) is correct. \square

If model (7.1) is incorrect, then $(I - H)E(y) \neq 0$ and the average mean squared prediction error, given x_1, \dots, x_n , is

$$\overline{\text{mse}} = \sigma^2 + \frac{\sigma^2 p}{n} + \frac{\|(I - H)E(y)\|^2}{n}.$$

Usually $\|(I - H)E(y)\|^2/n$ does not tend to 0. Thus, the bootstrap estimator (7.35) is asymptotically valid when model (7.1) is incorrect.

7.5.5 Model selection

We adopt the notation given in Section 7.4.

Theorem 7.13. Assume that the ε_i are i.i.d. and that

$$\max_{i \leq n} h_{i\alpha} \rightarrow 0 \quad \text{for all } \alpha \in \mathcal{A}, \tag{7.71}$$

where $h_{i\alpha} = x'_{i\alpha}(X'_\alpha X_\alpha)^{-1}x_{i\alpha}$.

(i) Consider the delete-1 CV. When α is an incorrect model,

$$\widehat{\text{mse}}_{\text{cv}}(\alpha) = \overline{\text{mse}}(\alpha) + o_p(1); \quad (7.72)$$

when α is a correct model,

$$\widehat{\text{mse}}_{\text{cv}}(\alpha) = \frac{\|\varepsilon\|^2}{n} + \frac{2\sigma^2 p_\alpha}{n} - \frac{\varepsilon' H_\alpha \varepsilon}{n} + o_p(n^{-1}).$$

(ii) Consider the delete- d CV with \mathcal{S} in (7.44) formed by using a balanced incomplete block design. Assume further that d is selected so that

$$\frac{d}{n} \rightarrow 1, \quad \frac{n}{n-d} \max_{i \leq n} h_{i\alpha} \rightarrow 0 \quad \text{for all } \alpha \in \mathcal{A}, \quad (7.73)$$

and

$$\lim_{n \rightarrow \infty} \max_{\mathbf{s} \in \mathcal{S}} \left\| \frac{1}{d} X'_{\alpha, \mathbf{s}} X_{\alpha, \mathbf{s}} - \frac{1}{n-d} X'_{\alpha, \mathbf{s}^c} X_{\alpha, \mathbf{s}^c} \right\| = 0. \quad (7.74)$$

Then, when α is an incorrect model, (7.72) holds with $\widehat{\text{mse}}_{\text{cv}}$ replaced by $\widehat{\text{mse}}_{\text{cv-d}}$; when α is a correct model,

$$\widehat{\text{mse}}_{\text{cv-d}}(\alpha) = \frac{\|\varepsilon\|^2}{n} + \frac{\sigma^2 p_\alpha}{n-d} + o_p\left(\frac{1}{n-d}\right).$$

(iii) Consider the delete- d CV with \mathcal{S} in (7.44) formed by taking a simple random sample of size B from the collection of all subsets of $\{1, \dots, n\}$. Assume all the conditions in part (ii) and $n^2/[B(n-d)^2] \rightarrow 0$. Then the results in part (ii) hold with $\|\varepsilon\|^2/n$ changed to $\sum_{\mathbf{s} \in \mathcal{S}} \|\varepsilon_{\mathbf{s}}\|^2/[B(n-d)]$.

(iv) Assume further that

$$\liminf_n \inf_{\alpha \text{ is incorrect}} \Delta(\alpha) > 0. \quad (7.75)$$

Then (7.41) and (7.42) hold; and $\hat{\alpha}_{\text{cv-d}}$ is consistent, i.e., (7.40) holds for $\hat{\alpha}_{\text{cv-d}}$.

Proof. We only prove part (i). The proofs for (ii) and (iii) can be found in Shao (1993b), and part (iv) follows directly from (i)-(iii) and condition (7.75). Let $r_{i\alpha} = y_i - x'_{i\alpha}\hat{\beta}_\alpha$. Then, by (7.9),

$$\widehat{\text{mse}}_{\text{cv}}(\alpha) = \frac{1}{n} \sum_{i=1}^n \frac{r_{i\alpha}^2}{(1-h_{i\alpha})^2} = \frac{1}{n} \sum_{i=1}^n r_{i\alpha}^2 + \frac{1}{n} \sum_{i=1}^n [2h_{i\alpha} + O(h_{i\alpha}^2)] r_{i\alpha}^2,$$

which is equal to

$$\frac{1}{n} \sum_{i=1}^n r_{i\alpha}^2 + \frac{2\sigma^2 p_\alpha}{n} + o_p(n^{-1})$$

if α is a correct model, and is equal to

$$\frac{1}{n} \sum_{i=1}^n r_{i\alpha}^2 + o_p(1)$$

if α is incorrect. The results then follow from

$$\frac{1}{n} \sum_{i=1}^n r_{i\alpha}^2 = \frac{\|\varepsilon\|^2}{n} + \Delta(\alpha) - \frac{\varepsilon' H_\alpha \varepsilon}{n} + \frac{2\varepsilon'(I - H_\alpha)X\beta}{n}$$

by the proof of Theorem 7.12. \square

This theorem justifies the assertions in Section 7.4.1. Even when \mathcal{A} does not contain any correct model, the results in Theorem 7.13 show that the CV still select the optimal model that minimizes $\widehat{\text{mse}}(\alpha)$ over \mathcal{A} .

Condition (7.75) means that an incorrect model is asymptotically worse than a correct model in terms of the mean squared prediction error. The second condition in (7.73) implies that $n-d \rightarrow \infty$ and is satisfied if $n-d \rightarrow \infty$ and $\{\|x_i\| : i = 1, 2, \dots\}$ is bounded.

Theorem 7.14. Assume that the ε_i are i.i.d. and that (7.71) holds.

(i) Consider the bootstrap estimator $\widehat{\text{mse}}_{\text{BOOT}}(\alpha)$ in (7.45) with $\hat{e}(\alpha)$ given by (7.46) (the PB) or (7.47) (the RB). When α is an incorrect model,

$$\widehat{\text{mse}}_{\text{BOOT}}(\alpha) = \overline{\text{mse}}(\alpha) + o_p(1); \quad (7.76)$$

when α is a correct model,

$$\widehat{\text{mse}}_{\text{BOOT}}(\alpha) = \frac{\|\varepsilon\|^2}{n} + \frac{2\sigma^2 p_\alpha}{n} - \frac{\varepsilon' H_\alpha \varepsilon}{n} + o_p(n^{-1}).$$

(ii) Let $\widehat{\text{mse}}_{\text{BOOT}-m}$ be defined by (7.48) with $\hat{e}_m(\alpha)$ given by (7.49) (the PB) or $2\hat{\sigma}^2 p_\alpha/m$ (the RB). Assume further that the bootstrap sample size m is selected so that

$$\frac{m}{n} \rightarrow 0, \quad \text{and} \quad \frac{n}{m} \max_{i \leq n} h_{i\alpha} \rightarrow 0 \quad \text{for all } \alpha \in \mathcal{A}. \quad (7.77)$$

Then, when α is an incorrect model, (7.76) holds with $\widehat{\text{mse}}_{\text{BOOT}}$ replaced by $\widehat{\text{mse}}_{\text{BOOT}-m}$; when α is a correct model,

$$\widehat{\text{mse}}_{\text{BOOT}-m}(\alpha) = \frac{\|\varepsilon\|^2}{n} + \frac{\sigma^2 p_\alpha}{m} + o_p\left(\frac{1}{m}\right).$$

(iii) If, in addition, (7.75) holds, then

$$\lim_{n \rightarrow \infty} P\{\hat{\alpha}_{\text{BOOT}} = \text{an incorrect model}\} = 0,$$

and

$$\lim_{n \rightarrow \infty} P\{\hat{\alpha}_{\text{BOOT}} = \alpha_0\} < 1$$

unless $\alpha = \{1, \dots, p\}$ is the only correct model; and $\hat{\alpha}_{\text{BOOT}-m}$ is consistent, i.e., (7.40) holds for $\hat{\alpha}_{\text{BQOT}-m}$, where $\hat{\alpha}_{\text{BQOT}}$ and $\hat{\alpha}_{\text{BOOT}-m}$ are the models selected by minimizing $\text{mse}_{\text{BOOT}}(\alpha)$ and $\widehat{\text{mse}}_{\text{BOOT}-m}(\alpha)$, respectively.

Proof. Part (iii) follows from (i) and (ii) and condition (7.75). In (i) and (ii), the proof for the RB is straightforward. The proof for the PB in (i) is the same as that of Theorem 7.12. It remains to show (ii) for the PB, but it follows from the proof of Theorem 7.12 and the fact that

$$\text{var}_*(\hat{\beta}_{m,\alpha}^*) = \frac{n}{m} \text{var}_*(\hat{\beta}_\alpha^*). \quad \square$$

7.6 Conclusions and Discussions

- (1) For variance estimation, all three jackknife estimators in Section 7.2.1 are asymptotically valid. For the LSE, the weighted jackknife estimators are better than the unweighted jackknife estimators, especially when the model is not balanced (the h_i are different). For bias estimation, the unweighted jackknife may be inconsistent, whereas both weighted jackknife methods produce the same consistent bias estimator.
- (2) Unlike the i.i.d. case, the jackknife variance estimators for the LSE of β (the linear case) are not the same as the traditional variance estimator (7.6); the latter is more efficient when the error variances are the same. The jackknife variance estimators are, however, robust against heteroscedasticity.
- (3) There are three different bootstrap procedures, the bootstrap based on residuals, the paired bootstrap, and the external bootstrap. The bootstrap based on residuals rests upon the i.i.d. assumption of the errors and is more efficient than the paired bootstrap when this assumption is valid. The paired bootstrap is the same as the bootstrap in the i.i.d. case if the pairs (y_i, x'_i) are i.i.d. and is robust against heteroscedasticity. The external bootstrap is robust against heteroscedasticity and may also be efficient if the distribution for bootstrap sampling is suitably chosen. One has to choose a bootstrap procedure that fits the model under consideration, is easy to implement, and provides accurate results. Recall that in our earlier discussion in Section 1.4, we emphasized that the bootstrap is not model-free.

- (4) For estimating the variance of the LSE, bootstrapping residuals leads to a variance estimator that is almost the same as the traditional variance estimator (7.6). The paired bootstrap and the external bootstrap produce variance estimators that are almost the same as the weighted jackknife variance estimators.
- (5) As in the i.i.d. case, the bootstrap can be applied to estimating the entire distribution of a statistic, constructing confidence sets, and testing hypotheses. The bootstrap distribution estimators and confidence sets may be more accurate than those obtained using the traditional normal approximation.
- (6) The bootstrap and the cross-validation estimators of the mean squared prediction error are asymptotically valid. For model selection based on minimizing the estimated mean squared prediction errors, however, one has to use the delete-d cross-validation with $d/n \rightarrow 1$, or the bootstrap of size m with $m/n \rightarrow 0$, in order to obtain a consistent model selection procedure. The reason has been given in Section 7.4.
- (7) In some situations, the explanatory variables may not be measured without error, and a linear error-in-variables model may be used to describe the relationship between the response and the explanatory variables. Applications of the bootstrap to this model are discussed in Booth and Hall (1993b) and Linder and Babu (1994).

Chapter 8

Applications to Nonlinear, Nonparametric, and Multivariate Models

Applications of the jackknife and the bootstrap to various nonlinear parametric (or semiparametric) models, nonparametric curve estimation problems, and multivariate analysis are discussed in this chapter. We mainly focus on the jackknife variance estimation, bootstrap distribution estimation, bootstrap confidence sets, cross-validation (jackknife) and bootstrap model selection, bandwidth (smoothing parameter) selection, and misclassification rate estimation. Other applications, such as the jackknife bias estimation and bias reduction, and the bootstrap variance estimation, can be similarly discussed but are omitted.

Except for some additional regularity conditions and technicalities, the theoretical justification for the jackknife and the bootstrap can be established along the lines of Chapters 2–4 and 7. Therefore, we will concentrate on the basic issues, formulas, derivations, and properties, without providing rigorous proofs. Technical details can always be found in the references cited thereafter.

8.1 Nonlinear Regression

A useful extension of the linear model (7.1) is the following parametric nonlinear model:

$$y_i = f(x_i, \beta) + \varepsilon_i, \quad i = 1, \dots, n, \tag{8.1}$$

where β is a p -vector of unknown parameters, f is a known function but nonlinear in β , and the x_i are q -vectors (q is not necessarily the same as p). When the x_i are deterministic, the ε_i are independent errors with mean 0 and variances σ_i^2 . When the x_i are random, the (y_i, x'_i) are i.i.d. with a finite second moment and $E(\varepsilon_i | x_i) = 0$. A customary estimator of β is $\hat{\beta}_{\text{LS}}$, the least squares estimator (LSE) defined as a solution of

$$L_n(\hat{\beta}_{\text{LS}}) = \min_{\gamma \in \mathbb{B}} L_n(\gamma), \quad (8.2)$$

where

$$L_n(\gamma) = \frac{1}{2} \sum_{i=1}^n [y_i - f(x_i, \gamma)]^2$$

and \mathbb{B} contains all possible values of β . If $\sigma_i^2 = \sigma^2$ for all i and the ε_i are normal, then $\hat{\beta}_{\text{LS}}$ is the maximum likelihood estimator of β ; otherwise, $\hat{\beta}_{\text{LS}}$ may not be an optimal estimator, but it is still used because of its popularity and simplicity. Throughout this section, we discuss only the LSE. When the parameter of interest is $\theta = g(\beta)$ for a given function g from \mathbb{R}^p to \mathbb{R} , the LSE of θ is $\hat{\theta}_{\text{LS}} = g(\hat{\beta}_{\text{LS}})$.

Unlike in the linear model case, (8.2) may not have a solution and, therefore, the LSE may not be well defined. In terms of practical applications, without loss of generality we may assume that \mathbb{B} is a compact subset of \mathbb{R}^p , which ensures the existence of $\hat{\beta}_{\text{LS}}$ for any n as well as the consistency of $\hat{\beta}_{\text{LS}}$ (Jennrich, 1969; Wu, 1981).

Suppose that the σ_i^2 are bounded away from 0 and ∞ ;

$$\nabla f(x, \gamma) = \frac{\partial f(x, \gamma)}{\partial \gamma} \quad \text{and} \quad \nabla^2 f(x, \gamma) = \frac{\partial^2 f(x, \gamma)}{\partial \gamma \partial \gamma'}$$

are continuous functions in x and γ ; and that

$$0 < \liminf_n \lambda_{\min}(M_n(\beta)/n) \leq \limsup_n \lambda_{\max}(M_n(\beta)/n) < \infty,$$

where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are the minimum and maximum eigenvalues of a matrix A , respectively, and

$$M_n(\gamma) = \sum_{i=1}^n \nabla f(x_i, \gamma) \nabla f(x_i, \gamma)'.$$

Assuming that g is differentiable at β , we can show that

$$(\hat{\theta}_{\text{LS}} - \theta) / \sqrt{\nabla g(\beta)' D_n \nabla g(\beta)} \rightarrow_d N(0, 1) \quad (8.3)$$

(see, e.g., Jennrich, 1969; Wu, 1981; Shao, 1992f), where

$$D_n = \sum_{i=1}^n \sigma_i^2 M_n^{-1}(\beta) \nabla f(x_i, \beta) \nabla f(x_i, \beta)' M_n^{-1}(\beta) \quad (8.4)$$

is the asymptotic covariance matrix of $\hat{\beta}_{\text{LS}}$.

8.1.1 Jackknife variance estimators

The jackknife can be applied to estimate the asymptotic variance of $\hat{\beta}_{\text{LS}}$ or $\hat{\theta}_{\text{LS}}$ given in (8.3)-(8.4). Let $\hat{\beta}_{\text{LS},i}$ be the LSE of β obtained after deleting the i th pair (y_i, x'_i) and $\hat{\theta}_{\text{LS},i} = g(\hat{\beta}_{\text{LS},i})$, $i = 1, \dots, n$. The (unweighted) jackknife variance estimator for $\hat{\theta}_{\text{LS}}$ is

$$v_{\text{JACK}} = \frac{n-1}{n} \sum_{i=1}^n \left(\hat{\theta}_{\text{LS},i} - \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{\text{LS},i} \right)^2.$$

This is a straightforward extension of (7.7). We may also consider weighted jackknives, but it is not clear what types of weights should be used.

In computing v_{JACK} , one has to repeat an iterative process n times to solve

$$L_{ni}(\hat{\beta}_{\text{LS},i}) = \min_{\gamma \in \mathbb{B}} L_{ni}(\gamma), \quad i = 1, \dots, n,$$

where

$$L_{ni}(\gamma) = \frac{1}{2} \sum_{j \neq i} [y_j - f(x_j, \gamma)]^2.$$

Although this may require a large amount of computation, the computation is routine if we adopt repeatedly the algorithm for computing $\hat{\beta}_{\text{LS}}$ to compute $\hat{\beta}_{\text{LS},i}$, $i = 1, \dots, n$. In addition, we can take $\hat{\beta}_{\text{LS}}$ as the initial point to reduce the number of iterations required in computing $\hat{\beta}_{\text{LS},i}$, $i = 1, \dots, n$, since $\max_{i \leq n} \|\hat{\beta}_{\text{LS},i} - \hat{\beta}_{\text{LS}}\| \rightarrow_{a.s.} 0$ under the conditions previously stated.

One may also apply the one-step jackknife discussed in Section 5.1.1 to reduce the computation. Suppose that the $\hat{\beta}_{\text{LS},i}$ are calculated using Newton's method; that is,

$$\hat{\beta}_{\text{LS},i} = \lim_{k \rightarrow \infty} \hat{\beta}_{\text{LS},i}^{[k]}, \quad \hat{\beta}_{\text{LS},i}^{[k+1]} = \hat{\beta}_{\text{LS},i}^{[k]} - [\nabla^2 L_{ni}(\hat{\beta}_{\text{LS},i}^{[k]})]^{-1} \nabla L_{ni}(\hat{\beta}_{\text{LS},i}^{[k]}).$$

If we take $\hat{\beta}_{\text{LS},i}^{[0]} = \hat{\beta}_{\text{LS}}$ for all i , then the one-step jackknife variance estimator is given by

$$v_{\text{JACK}}^{[1]} = \frac{n-1}{n} \sum_{i=1}^n \left(\hat{\theta}_{\text{LS},i}^{[1]} - \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{\text{LS},i}^{[1]} \right)^2$$

with $\hat{\theta}_{\text{LS},i}^{[1]} = g(\hat{\beta}_{\text{LS},i}^{[1]})$ and

$$\begin{aligned}\hat{\beta}_{\text{LS},i}^{[1]} &= \hat{\beta}_{\text{LS}} - [\nabla^2 L_{ni}(\hat{\beta}_{\text{LS}})]^{-1} \nabla L_{ni}(\hat{\beta}_{\text{LS}}) \\ &= \hat{\beta}_{\text{LS}} - [\nabla^2 L_{ni}(\hat{\beta}_{\text{LS}})]^{-1} r_i \nabla f(x_i, \hat{\beta}_{\text{LS}}),\end{aligned}\quad (8.5)$$

where $r_i = y_i - f(x_i, \hat{\beta}_{\text{LS}})$ and the last equality follows from the fact that $\nabla L_{ni}(\gamma) = \nabla L_n(\gamma) + [y_i - f(x_i, \gamma)] \nabla f(x_i, \gamma)$ and $\nabla L_n(\hat{\beta}_{\text{LS}}) = 0$. To compute $v_{\text{JACK}}^{[1]}$, however, one has to compute the inverses of

$$\nabla^2 L_{ni}(\hat{\beta}_{\text{LS}}) = \sum_{j \neq i} \nabla f(x_j, \hat{\beta}_{\text{LS}}) \nabla f(x_j, \hat{\beta}_{\text{LS}})' - \sum_{j \neq i} r_j \nabla^2 f(x_j, \hat{\beta}_{\text{LS}}) \quad (8.6)$$

for $i = 1, \dots, n$. We may replace $\nabla^2 L_{ni}(\hat{\beta}_{\text{LS}})$ in (8.5) by its leading term, the first term on the right-hand side of (8.6), which can be further approximated by $M_n(\hat{\beta}_{\text{LS}})$. This leads to

$$\tilde{\beta}_{\text{LS},i} = \hat{\beta}_{\text{LS}} - M_n^{-1}(\hat{\beta}_{\text{LS}}) r_i \nabla f(x_i, \hat{\beta}_{\text{LS}}), \quad i = 1, \dots, n,$$

and the linear jackknife variance estimator (Fox, Hinkley and Larntz, 1980)

$$v_{\text{LJACK}} = \frac{n-1}{n} \sum_{i=1}^n \left(\tilde{\theta}_{\text{LS},i} - \frac{1}{n} \sum_{i=1}^n \tilde{\theta}_{\text{LS},i} \right)^2$$

with $\tilde{\theta}_{\text{LS},i} = g(\tilde{\beta}_{\text{LS},i})$. Neither $v_{\text{JACK}}^{[1]}$ nor v_{LJACK} requires any iteration; the computation of v_{LJACK} is even simpler than that of $v_{\text{JACK}}^{[1]}$.

In view of the discussion in Section 7.2.1, we may replace the factor $\frac{n-1}{n}$ in v_{JACK} , $v_{\text{JACK}}^{[1]}$, and v_{LJACK} by $\frac{n-p}{n}$ for better finite sample performances of the variance estimators.

Shao (1992c, 1992f) established the consistency and robustness (against heteroscedasticity) of v_{JACK} , $v_{\text{JACK}}^{[1]}$, and v_{LJACK} . That is, irrespective of whether the σ_i^2 are equal or not,

$$v / \nabla g(\beta)' D_n \nabla g(\beta) \rightarrow_p 1$$

for $v = v_{\text{JACK}}$, $v_{\text{JACK}}^{[1]}$, or v_{LJACK} . Shao (1992c) actually established the same result for $\hat{\beta}_{\text{LS}}$ being regression M-estimators.

Some empirical results for the performances of the jackknife variance estimators can be found in Duncan (1978) and Fox, Hinkley and Larntz (1980).

The traditional approach estimates the variance of $\hat{\theta}_{\text{LS}}$ by linearization and substitution. If $\sigma_i^2 = \sigma^2$ is assumed, then $D_n = \sigma^2 M_n^{-1}(\beta)$, and the traditional variance estimator is $(n-p)^{-1} \sum_{i=1}^n r_i^2 \nabla g(\hat{\beta}_{\text{LS}})' M_n^{-1}(\hat{\beta}_{\text{LS}}) \nabla g(\hat{\beta}_{\text{LS}})$.

This estimator, however, is not robust against heteroscedasticity. A robust substitution variance estimator can be obtained by replacing β and σ_i^2 in (8.4) with $\hat{\beta}_{\text{LS}}$ and r_i^2 , respectively, which leads to the variance estimator

$$\sum_{i=1}^n r_i^2 [\nabla f(x_i, \hat{\beta}_{\text{LS}})' M_n^{-1}(\hat{\beta}_{\text{LS}}) \nabla g(\hat{\beta}_{\text{LS}})]^2. \quad (8.7)$$

This estimator is robust against heteroscedasticity and is almost the same as the linear jackknife estimator when g is linear.

The jackknife variance estimators are asymptotically equivalent to the variance estimator in (8.7). Like the i.i.d. case, the choice between the jackknife and the substitution estimators depends on the feasibility of their implementation.

8.1.2 Bootstrap distributions and confidence sets

All three bootstrap procedures described in Section 7.2.2 can be extended to model (8.1) in a straightforward manner. We discuss the bootstrap based on residuals (RB) for illustration, assuming that the ε_i are i.i.d.

Let $r_i = y_i - f(x_i, \hat{\beta}_{\text{LS}})$ be the i th residual, $\bar{r} = n^{-1} \sum_{i=1}^n r_i$, and \hat{F}_ε be the empirical distribution putting mass n^{-1} to $r_i - \bar{r}$, $i = 1, \dots, n$. Let $\varepsilon_1^*, \dots, \varepsilon_n^*$ be i.i.d. from \hat{F}_ε and $y_i^* = f(x_i, \hat{\beta}_{\text{LS}}) + \varepsilon_i^*$, $i = 1, \dots, n$. The bootstrap analog of $\hat{\theta}_{\text{LS}}$ is then $\hat{\theta}_{\text{LS}}^* = g(\hat{\beta}_{\text{LS}}^*)$ with $\hat{\beta}_{\text{LS}}^*$ being a solution of

$$L_n^*(\hat{\beta}_{\text{LS}}^*) = \min_{\gamma \in \mathbb{B}} L_n^*(\gamma),$$

where

$$L_n^*(\gamma) = \frac{1}{2} \sum_{i=1}^n [y_i^* - f(x_i, \gamma)]^2.$$

The bootstrap distribution estimators and confidence sets can be defined similarly to those in the previous chapters. Consistency of the bootstrap distribution estimators and asymptotic validity of the bootstrap confidence sets can be established under the conditions previously stated.

The bootstrap distribution estimators for studentized statistics and the related bootstrap-t confidence sets are expected to be asymptotically more accurate than those obtained by using the normal approximation based on result (8.3). Consider the studentized statistic

$$\mathfrak{R}_n = \frac{c'(\hat{\beta}_{\text{LS}} - \beta)}{\hat{\sigma}[c'M_n^{-1}(\hat{\beta}_{\text{LS}})c]^{1/2}},$$

where $\hat{\sigma}^2 = \text{var}_*(\varepsilon_1) = n^{-1} \sum_{i=1}^n (r_i - \bar{r})^2$. Under some complicated conditions on the function f and the error distribution, Huet and Jolivet (1989)

developed the Edgeworth and Cornish-Fisher expansions for the distribution and the quantile function, respectively, of \mathfrak{R}_n and its bootstrap analog

$$\mathfrak{R}_n^* = \frac{c'(\hat{\beta}_{\text{LS}}^* - \hat{\beta}_{\text{LS}})}{\hat{\sigma}^*[c'M_n^{-1}(\hat{\beta}_{\text{LS}}^*)c]^{1/2}},$$

where $\hat{\sigma}^{*2} = n^{-1} \sum_{i=1}^n (r_i^* - \bar{r}^*)^2$, $r_i^* = y_i^* - f(x_i, \hat{\beta}_{\text{LS}}^*)$, and $\bar{r}^* = n^{-1} \sum_{i=1}^n r_i^*$. Using these expansions, they showed that

$$\sup_t |P_*\{\mathfrak{R}_n^* \leq t\} - P\{\mathfrak{R}_n \leq t\}| = o(n^{-1/2}) \text{ a.s.}$$

Furthermore, if \hat{t}_α is the $(1 - \alpha)$ th quantile of the bootstrap distribution of \mathfrak{R}_n^* and $c'\hat{\beta}_{\text{LS}} - \hat{t}_\alpha \hat{\sigma}[c'M_n^{-1}(\hat{\beta}_{\text{LS}})c]^{1/2}$ is taken as a lower confidence bound for $c'\beta$, then

$$P\{c'\beta \geq c'\hat{\beta}_{\text{LS}} - \hat{t}_\alpha \hat{\sigma}[c'M_n^{-1}(\hat{\beta}_{\text{LS}})c]^{1/2}\} = 1 - \alpha + o(n^{-1/2}).$$

Huet, Jolivet and Messean (1990) studied by simulation finite sample performances of the bootstrap estimators. Some computation issues were addressed by Gruet, Huet and Jolivet (1993).

Similar to the one-step jackknife, we may apply the one-step bootstrap described in Section 5.4.7 to avoid the iterations in computing $\hat{\beta}_{\text{LS}}^*$. If $\hat{\beta}_{\text{LS}}^*$ is computed by using Newton's method, then the one-step bootstrap estimator of the distribution of $\sqrt{n}(\hat{\beta}_{\text{LS}} - \beta)$ is the bootstrap distribution of $\sqrt{n}(\hat{\beta}_{\text{LS}}^{*[1]} - \hat{\beta}_{\text{LS}})$, where

$$\hat{\beta}_{\text{LS}}^{*[1]} = \hat{\beta}_{\text{LS}} - [\nabla^2 L_n^*(\hat{\beta}_{\text{LS}})]^{-1} \nabla L_n^*(\hat{\beta}_{\text{LS}}).$$

Since

$$\begin{aligned} \frac{1}{n} \nabla^2 L_n^*(\hat{\beta}_{\text{LS}}) &= \frac{1}{n} \sum_{i=1}^n \nabla f(x_i, \hat{\beta}_{\text{LS}}) \nabla f(x_i, \hat{\beta}_{\text{LS}})' - \frac{1}{n} \sum_{i=1}^n \varepsilon_i^* \nabla^2 f(x_i, \hat{\beta}_{\text{LS}}) \\ &= \frac{M_n(\hat{\beta}_{\text{LS}})}{n} + o_p(1), \end{aligned} \tag{8.8}$$

we may replace $\nabla^2 L_n^*(\hat{\beta}_{\text{LS}})$ by $M_n(\hat{\beta}_{\text{LS}})$ and use

$$\tilde{\beta}_{\text{LS}}^* = \hat{\beta}_{\text{LS}} - M_n^{-1}(\hat{\beta}_{\text{LS}}) \nabla L_n^*(\hat{\beta}_{\text{LS}})$$

instead of $\hat{\beta}_{\text{LS}}^{*[1]}$.

It is actually easy to show the consistency of the bootstrap distribution of $\sqrt{n}(\tilde{\beta}_{\text{LS}}^* - \hat{\beta}_{\text{LS}})$ as an estimator of the distribution of $\sqrt{n}(\hat{\beta}_{\text{LS}} - \beta)$. Since

$$\nabla L_n^*(\hat{\beta}_{\text{LS}}) = - \sum_{i=1}^n [y_i^* - f(x_i, \hat{\beta}_{\text{LS}})] \nabla f(x_i, \hat{\beta}_{\text{LS}}) = - \sum_{i=1}^n \varepsilon_i^* \nabla f(x_i, \hat{\beta}_{\text{LS}}),$$

we have

$$\tilde{\beta}_{\text{LS}}^* - \hat{\beta}_{\text{LS}} = M_n^{-1}(\hat{\beta}_{\text{LS}}) \sum_{i=1}^n \varepsilon_i^* \nabla f(x_i, \hat{\beta}_{\text{LS}}), \quad (8.9)$$

and we can apply Lindeberg's condition to the right-hand side of (8.9) to establish the asymptotic normality of $\tilde{\beta}_{\text{LS}}^* - \hat{\beta}_{\text{LS}}$. The consistency follows from the fact that

$$\text{var}_*(\tilde{\beta}_{\text{LS}}^*) = M_n^{-1}(\hat{\beta}_{\text{LS}}) \sum_{i=1}^n \text{var}_*(\varepsilon_i^*) \nabla f(x_i, \hat{\beta}_{\text{LS}}) \nabla f(x_i, \hat{\beta}_{\text{LS}})' M_n^{-1}(\hat{\beta}_{\text{LS}})$$

is consistent for D_n .

By (8.8), we know that the one-step bootstrap distribution estimator for $\hat{\beta}_{\text{LS}} - \beta$ is also consistent. In fact, the same argument can be applied to the case where $\varepsilon_1^*, \dots, \varepsilon_n^*$ are obtained by external bootstrap sampling (see Section 7.2.2).

8.1.3 Cross-validation for model selection

Results for model selection by cross-validation (CV) described in Section 7.5.1 can be extended to the selection of a model from a class of models $\alpha \in \mathcal{A}$, where each α represents a nonlinear model

$$y_i = f_\alpha(x_i, \beta_\alpha) + \varepsilon_i, \quad i = 1, \dots, n, \quad (8.10)$$

β_α is a subvector of β , f_α is the function that relates x_i to β_α , and the ε_i are i.i.d. errors with mean 0 and variance σ^2 .

To define the delete-d CV, we adopt the notation in Section 7.5.1. Let \mathbf{s} be a subset of $\{1, \dots, n\}$ of size d and \mathbf{s}^c be its complement. Let $X' = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_n)'$, $(y_{\mathbf{s}}, X_{\mathbf{s}})$ be the $d \times (1+q)$ matrix containing the rows of (y, X) indexed by the integers in \mathbf{s} , and $(y_{\mathbf{s}^c}, X_{\mathbf{s}^c})$ be the $(n-d) \times (1+q)$ matrix containing the rows of (y, X) indexed by the integers in \mathbf{s}^c . For $\alpha \in \mathcal{A}$, let $\hat{\beta}_{\alpha, \mathbf{s}^c}$ be the LSE of β_α based on the data $(y_{\mathbf{s}^c}, X_{\mathbf{s}^c})$ under model (8.10) and let

$$\widehat{\text{mse}}_{\text{cv-d}}(\alpha) = \frac{1}{Bd} \sum_{\mathbf{s} \in \mathcal{S}} \sum_{i \in \mathbf{s}} [y_i - f(x_i, \hat{\beta}_{\alpha, \mathbf{s}^c})]^2,$$

where \mathcal{S} is the same as that in (7.44).

The model selected by using the delete-d CV, denoted by $\hat{\alpha}_{\text{cv-d}}$, is obtained by minimizing $\widehat{\text{mse}}_{\text{cv-d}}(\alpha)$ over $\alpha \in \mathcal{A}$. If $d/n \rightarrow 1$ and $n-d \rightarrow \infty$, then $\hat{\alpha}_{\text{cv-d}}$ is consistent, i.e.,

$$\lim_{n \rightarrow \infty} P\{\hat{\alpha}_{\text{cv-d}} = \text{the optimal model}\} = 1,$$

under some regularity conditions (Shao, 1993b). Also, the delete-1 CV is inconsistent when there is more than one correct model.

Results for model selection by bootstrap described in Section 7.5.2 can also be extended similarly to the selection of nonlinear models.

8.2 Generalized Linear Models

A generalized linear model is characterized by the following structure: the responses y_1, \dots, y_n are independent and

$$E(y_i) = \mu(\eta_i) \quad \text{and} \quad \text{var}(y_i) = \phi_i \mu'(\eta_i), \quad (8.11)$$

where the $\phi_i > 0$ are unknown parameters, $\mu(\eta)$ is a known differentiable function with $\mu'(\eta) > 0$, the η_i are related to the values of explanatory variables x_i by a known injective and third order continuously differentiable *link* function f :

$$f(\mu(\eta_i)) = x'_i \beta, \quad (8.12)$$

and β is a p -vector of unknown parameters.

Similar to the nonlinear model (8.1), the generalized linear model specified by (8.11)-(8.12) is a useful extension of the linear model (7.1) [which can be specified by (8.11)-(8.12) with $f \equiv 1$ and $\mu(\eta) = \eta$]. Examples of generalized linear models, including models such as logit models, log-linear models, gamma-distributed data models, and survival data models, can be found in Nelder and Wedderburn (1972) and McCullagh and Nelder (1989).

The main parameter of interest in a generalized linear model is β . The parameters ϕ_i are called dispersion parameters [$\phi_i = \sigma_i^2$ in the linear model (7.1)] and are often assumed the same. A customary estimator of β is the weighted least squares estimator $\hat{\beta}_{\text{WLS}}$, which is a solution of

$$\sum_{i=1}^n x_i h(x'_i \gamma) [y_i - \mu_i(\gamma)] = 0, \quad (8.13)$$

where $h(t)$ is the first order derivative of $\mu^{-1}(f^{-1}(t))$ and $\mu_i(\gamma) = f^{-1}(x'_i \gamma)$.

In biostatistics, the following extension of model (8.11)-(8.12) is very useful: $y_i = (y_{i1}, \dots, y_{in_i})'$ is a vector of n_i responses, $i = 1, \dots, n$; each y_{it} satisfies

$$E(y_{it}) = \mu(\eta_{it}) \quad \text{and} \quad f(\mu(\eta_{it})) = x'_{it} \beta; \quad (8.14)$$

and the y_i are independent, but for each i , y_{i1}, \dots, y_{in_i} may be dependent and the covariance matrix $\text{var}(y_i)$ is an unknown $n_i \times n_i$ matrix. Data sets of this type are called longitudinal data in biostatistics. Usually, the n_i are small but n is large. For longitudinal data, the covariance matrices

$\text{var}(y_i)$ are hard to model, and the estimation of β is often based on the quasi-likelihood approach (Liang and Zeger, 1986). Assume a “working” covariance matrix [which is not necessarily the same as $\text{var}(y_i)$]

$$\phi_i A_i R_i(\rho) A_i,$$

where ϕ_i is the same as that in (8.11), A_i is the $n_i \times n_i$ diagonal matrix with $\sqrt{\mu'(\eta_{it})}$ as the t th diagonal element, and $R_i(\rho)$ is an $n_i \times n_i$ “working” correlation matrix that has a known form but depends on ρ , an unknown vector of parameters. Let $\hat{\rho}$ be an estimator of ρ satisfying $\hat{\rho} - \rho_0 = O_p(n^{-1/2})$ for some ρ_0 (not necessarily the same as ρ). Then, the maximum quasi-likelihood estimator $\hat{\beta}_{\text{MQ}}$ of β is obtained by solving

$$\sum_{i=1}^n B_i(\gamma) [A_i(\gamma) R_i(\hat{\rho}) A_i(\gamma)]^{-1} [y_i - \mu_i(\gamma)] = 0, \quad (8.15)$$

where $B_i(\gamma) = \partial \mu_i(\gamma) / \partial \gamma$, $A_i(\gamma)$ is defined the same as A_i but with β replaced by γ , and

$$\mu_i(\gamma) = (f^{-1}(x'_{i1}\gamma), \dots, f^{-1}(x'_{in_i}\gamma))', \quad i = 1, \dots, n.$$

Although the working covariance matrix may not be correctly specified and $\hat{\rho}$ may not converge to ρ , the estimator $\hat{\beta}_{\text{MQ}}$ is still consistent ($n \rightarrow \infty$) and asymptotically normal. Assume that the ϕ_i are bounded away from 0 and ∞ ; the n_i are bounded; the x_i are deterministic and within a compact subset of \mathbb{R}^p ; $\lambda_{\min}(X'X) \rightarrow \infty$, where

$$X = (x_{11}, \dots, x_{1n_1}, \dots, x_{n1}, \dots, x_{nn_n})';$$

and

$$\limsup_n [\lambda_{\max}(X'X)]^{(1+\delta)/2} / \lambda_{\min}(X'X) < \infty.$$

Then $\hat{\beta}_{\text{MQ}} - \beta$ is asymptotically normal with mean 0 and asymptotic covariance matrix

$$D_n = M_n^{-1} \left[\sum_{i=1}^n B_i V_i^{-1} \text{var}(y_i) V_i^{-1} B_i' \right] M_n^{-1}, \quad (8.16)$$

where $B_i = B_i(\beta)$, $V_i = A_i R_i(\rho_0) A_i$, $M_n = M_n(\beta, \rho_0)$, and

$$M_n(\gamma, \rho) = \sum_{i=1}^n B_i(\gamma) [A_i(\gamma) R_i(\rho) A_i(\gamma)]^{-1} B_i(\gamma)'$$

(see, e.g., Liang and Zeger, 1986). Furthermore, a consistent estimator of D_n , irrespective of whether the covariance matrices are correctly specified

or not, is given by replacing $\text{var}(y_i)$ in (8.16) by $[y_i - \mu_i(\beta)][y_i - \mu_i(\beta)]'$ and then substituting β and ρ_0 with $\hat{\beta}_{MQ}$ and $\hat{\rho}$, respectively. Some still weaker conditions can be used to replace the stated conditions (e.g., the x_i may be unbounded; see Shao, 1992d).

A simple example of $R_i(\rho)$ is $R_i(\rho) = I_{n_i}$, the identity matrix, in which case (8.15) reduces to (8.13), the independence model. But other choices of $R_i(\rho)$ may provide more efficient estimators of β (Liang and Zeger, 1986).

If the parameter of interest is $\theta = g(\beta)$ for a given function g differentiable at β , then $\hat{\theta}_{MQ} = g(\hat{\beta}_{MQ})$ and $\hat{\theta}_{MQ} - \theta$ is asymptotically normal with mean 0 and asymptotic variance $\nabla g(\beta)'D_n\nabla g(\beta)$.

8.2.1 Jackknife variance estimators

We consider the jackknife variance estimator for $\hat{\theta}_{MQ} = g(\hat{\beta}_{MQ})$ with $\hat{\beta}_{MQ}$ being a solution of (8.15) under the longitudinal data model specified by (8.14). For each i , let $\hat{\beta}_{MQ,i}$ be the maximum quasi-likelihood estimator of β after deleting the pair (y'_i, x'_i) , i.e., $\hat{\beta}_{MQ,i}$ is a solution of

$$\sum_{j \neq i} B_j(\gamma)[A_j(\gamma)R_j(\hat{\rho})A_j(\gamma)]^{-1}[y_j - \mu_j(\gamma)] = 0. \quad (8.17)$$

Then, the jackknife variance estimator for $\hat{\theta}_{MQ}$ is

$$v_{JACK} = \frac{n-1}{n} \sum_{i=1}^n \left(\hat{\theta}_{MQ,i} - \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{MQ,i} \right)^2,$$

where $\hat{\theta}_{MQ,i} = g(\hat{\beta}_{MQ,i})$. This is an extension of the jackknife estimator discussed at the end of Section 7.2 and in Theorem 7.4. Another special case is when $n_i = 1$ for all i , in which case v_{JACK} is a straightforward extension of (7.7).

Let $s_n(\gamma)$ and $s_{ni}(\gamma)$ be the quantities on the left-hand sides of (8.15) and (8.17), respectively. When $\hat{\beta}_{MQ,i}$ are computed using Newton's method with $\hat{\beta}_{MQ}$ as the initial point,

$$\hat{\beta}_{MQ,i}^{[1]} = \hat{\beta}_{MQ} - [\nabla s_{ni}(\hat{\beta}_{MQ})]^{-1} s_{ni}(\hat{\beta}_{MQ}) = \hat{\beta}_{MQ} + [\nabla s_{ni}(\hat{\beta}_{MQ})]^{-1} \hat{B}_i \hat{V}_i^{-1} r_i,$$

where $\hat{B}_i = B_i(\hat{\beta}_{MQ})$, $\hat{V}_i = A_i(\hat{\beta}_{MQ})R_i(\hat{\rho})A_i(\hat{\beta}_{MQ})$, $r_i = y_i - \mu_i(\hat{\beta}_{MQ})$, and the last equality follows from $s_n(\hat{\beta}_{MQ}) = 0$ and $s_{ni}(\hat{\beta}_{MQ}) = s_n(\hat{\beta}_{MQ}) - \hat{B}_i \hat{V}_i^{-1} r_i$. Consequently, the one-step jackknife variance estimator for $\hat{\theta}_{MQ}$ is

$$v_{JACK}^{[1]} = \frac{n-1}{n} \sum_{i=1}^n \left(\hat{\theta}_{MQ,i}^{[1]} - \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{MQ,i}^{[1]} \right)^2$$

with $\hat{\theta}_{MQ,i}^{[1]} = g(\hat{\beta}_{MQ,i}^{[1]})$. Since

$$\nabla s_{ni}(\hat{\beta}_{MQ}) \approx \nabla s_n(\hat{\beta}_{MQ}) \approx \sum_{i=1}^n \hat{B}_i \hat{V}_i^{-1} \hat{B}_i = M_n(\hat{\beta}_{MQ}, \hat{\rho}) = \hat{M}_n,$$

$v_{JACK}^{[1]}$ can be simplified further by

$$v_{LJACK} = \frac{n-1}{n} \sum_{i=1}^n \left(\tilde{\theta}_{MQ,i} - \frac{1}{n} \sum_{i=1}^n \tilde{\theta}_{MQ,i} \right)^2$$

with $\tilde{\theta}_{MQ,i} = g(\tilde{\beta}_{MQ,i})$ and

$$\tilde{\beta}_{MQ,i} = \hat{\beta}_{MQ} - \hat{M}_n^{-1} \hat{B}_i \hat{V}_i^{-1} r_i.$$

For a linear function g , v_{LJACK} is almost the same as the substitution estimator described previously. In the case of $R_i(\rho) = I_{n_i}$,

$$\tilde{\beta}_{MQ,i} = \hat{\beta}_{MQ} - \left\{ \sum_{i=1}^n \sum_{t=1}^{n_i} [h(x'_{it} \hat{\beta}_{MQ})]^2 v_{it} x_{it} x'_{it} \right\}^{-1} \sum_{i=1}^n \sum_{t=1}^{n_i} h(x'_{it} \hat{\beta}_{MQ}) r_{it} x_{it},$$

$v_{it} = \mu'(\hat{\eta}_{it})$, $\hat{\eta}_{it} = \mu^{-1}(f^{-1}(x'_{it} \hat{\beta}_{MQ}))$, and r_{it} is the t th component of r_i .

Under the conditions previously stated, it can be shown, along the lines of Shao (1992g), that v_{JACK} , $v_{JACK}^{[1]}$, and v_{LJACK} are consistent estimators of the asymptotic variance of $\hat{\theta}_{MQ}$ and are robust against the mis-specification of the covariance matrices.

8.2.2 Bootstrap procedures

The bootstrap based on residuals described in Section 7.2.2 is applicable to the generalized models specified by (8.11)-(8.12) with equal ϕ_i . One needs to take bootstrap data from Pearson's residuals

$$\tilde{r}_i = r_i / \sqrt{v_i}, \quad i = 1, \dots, n,$$

where $r_i = y_i - \mu_i(\hat{\beta}_{WLS})$, $v_i = \mu'(\hat{\eta}_i)$, and $\hat{\eta}_i = \mu^{-1}(f^{-1}(x'_i \hat{\beta}_{WLS}))$. Let $\varepsilon_1^*, \dots, \varepsilon_n^*$ be i.i.d. from the empirical distribution putting mass n^{-1} to $\tilde{r}_i - n^{-1} \sum_{i=1}^n \tilde{r}_i$, $i = 1, \dots, n$, and

$$y_i^* = \mu_i(\hat{\beta}_{WLS}) + \sqrt{v_i} \varepsilon_i^*.$$

The bootstrap analog $\hat{\beta}_{WLS}^*$ of $\hat{\beta}_{WLS}$ is obtained by solving the bootstrap analog of (8.13):

$$s_n^*(\gamma) = \sum_{i=1}^n x_i h(x'_i \gamma) [y_i^* - \mu_i(\gamma)] = 0.$$

The one-step bootstrap can be used to reduce the computation; that is, we compute

$$\hat{\beta}_{\text{WLS}}^{*[1]} = \hat{\beta}_{\text{WLS}} - [\nabla s_n^*(\hat{\beta}_{\text{WLS}})]^{-1} s_n^*(\hat{\beta}_{\text{WLS}}) \quad (8.18)$$

and replace $\hat{\beta}_{\text{WLS}}^*$ by $\hat{\beta}_{\text{WLS}}^{*[1]}$. One can simplify further the computation by replacing $\nabla s_n^*(\hat{\beta}_{\text{WLS}})$ with

$$\hat{M}_n = \sum_{i=1}^n [h(x_i' \hat{\beta}_{\text{WLS}})]^2 v_i x_i x_i'$$

and replacing $\hat{\beta}_{\text{WLS}}^{*[1]}$ with

$$\tilde{\beta}_{\text{WLS}}^* = \hat{\beta}_{\text{WLS}} - \hat{M}_n^{-1} s_n^*(\hat{\beta}_{\text{WLS}}).$$

Conditioned on the y_i and x_i ,

$$\tilde{\beta}_{\text{WLS}}^* = \hat{\beta}_{\text{WLS}} - \hat{M}_n^{-1} \sum_{i=1}^n x_i h(x_i' \hat{\beta}_{\text{WLS}}) \sqrt{v_i} \varepsilon_i^*$$

is a linear combination of i.i.d. random variables ε_i^* , $E_*(\tilde{\beta}_{\text{WLS}}^*) = \hat{\beta}_{\text{WLS}}$, and

$$\text{var}_*(\tilde{\beta}_{\text{WLS}}^*) = \hat{\phi} \hat{M}_n^{-1},$$

where

$$\hat{\phi} = \text{var}_*(\varepsilon_i^*) = \frac{1}{n} \sum_{i=1}^n \left(\tilde{r}_i^2 - \frac{1}{n} \sum_{i=1}^n \tilde{r}_i \right)^2.$$

If $\phi_i = \phi$ for all i , then $\hat{\phi} \rightarrow_{a.s} \phi$, and this bootstrap produces a consistent estimator of the distribution of $\hat{\beta}_{\text{WLS}}$ or $g(\hat{\beta}_{\text{WLS}})$.

For a longitudinal data model specified by (8.14), a bootstrap based on residuals may not be appropriate, especially when ϕ_i are unequal. But the paired bootstrap or the external bootstrap described in Section 7.2.2 can be extended to this model.

For the paired bootstrap, we first draw i.i.d. random integers j_1^*, \dots, j_n^* from the set of integers $\{1, \dots, n\}$; we then define

$$(y_i^{**}, x_i^{**}) = (y_{j_i^*}', x_{j_i^*}'), \quad i = 1, \dots, n,$$

and $\hat{\beta}_{\text{MQ}}^*$ as a solution of

$$\sum_{i=1}^n B_i^*(\gamma) [A_i^*(\gamma) R_i^*(\hat{\rho}) A_i^*(\gamma)]^{-1} [y_i^* - \mu_i^*(\gamma)] = 0, \quad (8.19)$$

where $B_i^*(\gamma) = B_{j_i^*}(\gamma)$, $A_i^*(\gamma) = A_{j_i^*}(\gamma)$, $R_i^*(\hat{\rho}) = R_{j_i^*}(\hat{\rho})$, and $\mu_i^*(\gamma) = \mu_{j_i^*}(\gamma)$. The one-step bootstrap estimator is based on $\hat{\beta}_{\text{MQ}}^{*[1]}$ calculated using

(8.18) with $s_n^*(\gamma)$ replaced by the quantity on the left-hand side of (8.19). We may also simply use

$$\tilde{\beta}_{MQ}^* = \hat{\beta}_{MQ} - \hat{M}_n^{-1} \sum_{i=1}^n B_i^*(\hat{\beta}_{MQ}) [A_i^*(\hat{\beta}_{MQ}) R_i^*(\hat{\rho}) A_i^*(\hat{\beta}_{MQ})]^{-1} [y_i^* - \mu_i^*(\hat{\beta}_{MQ})]$$

to replace $\hat{\beta}_{MQ}^*$.

The external bootstrap can be extended as follows. Let $\varepsilon_i^*, i = 1, \dots, n$, be independent random n_i -vectors with $E_*(\varepsilon_i^*) = 0$ and $\text{var}_*(\varepsilon_i^*) = r_i r_i'$. Define

$$y_i^* = \mu_i(\hat{\beta}_{MQ}) + \varepsilon_i^*, \quad i = 1, \dots, n,$$

and $\hat{\beta}_{MQ}^*$ as a solution of

$$\sum_{i=1}^n B_i(\gamma) [A_i(\gamma) R_i(\hat{\rho}) A_i(\gamma)]^{-1} [y_i^* - \mu_i(\gamma)] = 0.$$

To reduce the computation, $\hat{\beta}_{MQ}^*$ can be replaced by

$$\tilde{\beta}_{MQ}^* = \hat{\beta}_{MQ} - \hat{M}_n^{-1} \sum_{i=1}^n \hat{B}_i \hat{V}_i^{-1} \varepsilon_i^*.$$

The bootstrap estimators of the distribution of $\hat{\beta}_{MQ}$ or $\hat{\theta}_{MQ}$ based on these bootstrap procedures are consistent. Examples of applications of these bootstrap procedures in various specified models can be found, for example, in Simonoff and Tsai (1988), Moulton and Zeger (1989, 1991), Sauermann (1989), and Lee (1990).

8.2.3 Model selection by bootstrapping

In this section, we study the selection of a model from a class of models represented by $\alpha \in \mathcal{A}$:

$$E(y_i) = \mu(\eta_{i\alpha}) \quad \text{and} \quad \text{var}(y_i) = \phi \mu'(\eta_{i\alpha}), \quad (8.20)$$

where α is a subset of $\{1, \dots, p\}$, $f(\mu(\eta_{i\alpha})) = x'_{i\alpha} \beta_\alpha$, f is the given link function, and $x_{i\alpha}$ and β_α are defined the same as in Section 7.4. Note that the dispersion parameters in (8.20) are the same for all i and α . We assume that the model with $\alpha = \{1, \dots, p\}$ is correct and is in \mathcal{A} ; and the standardized variables $[y_i - E(y_i)]/\sqrt{\text{var}(y_i)}$, $i = 1, \dots, n$, are i.i.d.

In linear models or nonlinear regression models, we select a model by minimizing the average mean squared prediction errors when the future

response at x_i is predicted by $x'_i \hat{\beta}_{\text{LS}}$ or $f(x_i, \hat{\beta}_{\text{LS}})$. However, in some generalized linear models y_i takes integer values, but $\mu_i(\hat{\beta}_{\text{WLS}})$ is not an integer and is not an appropriate prediction. Thus, we need to adopt an alternative approach.

Consider the weighted average of squared error losses when $E(y_i)$ is estimated by $\mu(\hat{\eta}_{i\alpha})$ under model α :

$$L_\alpha = \frac{1}{n} \sum_{i=1}^n \frac{[E(y_i) - \mu(\hat{\eta}_{i\alpha})]^2}{\text{var}(y_i)/\phi},$$

where $\hat{\eta}_{i\alpha} = \mu^{-1}(f^{-1}(x'_{i\alpha} \hat{\beta}_\alpha))$ and $\hat{\beta}_\alpha$ is the weighted LSE estimator of β_α under model α . Suppose that we would like to select a model α so that $E(L_\alpha)$ is as small as possible. Since $E(L_\alpha)$ is unknown, we select a model by minimizing $E_*(L_\alpha^*)$ over α , where

$$L_\alpha^* = \frac{1}{n} \sum_{i=1}^n \frac{[y_i - \mu(\hat{\eta}_{i\alpha}^*)]^2}{v_{i\alpha}}$$

is a bootstrap analog of L_α , $v_{i\alpha} = \mu'(\hat{\eta}_{i\alpha})$, $\hat{\eta}_{i\alpha}^* = \mu^{-1}(f^{-1}(x'_{i\alpha} \hat{\beta}_\alpha^*))$, and $\hat{\beta}_\alpha^*$ is a bootstrap analog of $\hat{\beta}_\alpha$. To simplify the computation, $\hat{\beta}_\alpha^*$ can be replaced by $\tilde{\beta}_\alpha^*$ described in Section 8.2.2 and based on any of the three bootstrap sampling methods.

In view of the results in Section 7.4.2, however, the bootstrap model selection procedure is inconsistent if $\hat{\beta}_\alpha^*$ or $\tilde{\beta}_\alpha^*$ is based on a bootstrap sample of size n . For the paired bootstrap, the problem can be solved by changing n to m with $m/n \rightarrow 0$ and $m \rightarrow \infty$. For the bootstrap based on residuals or the external bootstrap, we need to simply change ε_i^* to $\frac{n}{m}\varepsilon_i^*$, $i = 1, \dots, n$. Assume that for any incorrect model α ,

$$\liminf_n \frac{1}{n} \sum_{i=1}^n \frac{[y_i - \mu(\hat{\eta}_{i\alpha})]^2}{v_{i\alpha}} > \phi \text{ a.s.,} \quad (8.21)$$

which means that the average of squared Pearson's residuals under an incorrect model is asymptotically larger than that of a correct model since, when α is correct,

$$\frac{1}{n} \sum_{i=1}^n \frac{[y_i - \mu(\hat{\eta}_{i\alpha})]^2}{v_{i\alpha}} \rightarrow_{a.s.} \phi$$

(Shao, 1992d). Then, under some regularity conditions,

$$E_*(L_\alpha^*) = \frac{1}{n} \sum_{i=1}^n \frac{[y_i - E(y_i)]^2}{\text{var}(y_i)} + \frac{\phi p_\alpha}{m} + o_p\left(\frac{1}{m}\right) \quad (8.22)$$

if α is a correct model, and

$$E_*(L_\alpha^*) = \frac{1}{n} \sum_{i=1}^n \frac{[y_i - \mu(\hat{\eta}_{i\alpha})]^2}{v_{i\alpha}} + o_p(1) \quad (8.23)$$

if α is an incorrect model. When α is correct,

$$E(L_\alpha) = \frac{\phi p_\alpha}{n} + o\left(\frac{1}{n}\right). \quad (8.24)$$

Hence, the consistency of the bootstrap selection procedure with $m/n \rightarrow 0$ and $m \rightarrow \infty$ follows from (8.21)-(8.24).

8.3 Cox's Regression Models

For a distribution function F with density f , its hazard function is defined to be $h(t) = f(t)/[1 - F(t)]$ and is an important component in survival analysis. There is often interest in assessing the effects of some explanatory variables on the hazard functions for individuals. The following model, called Cox's regression model or the proportional hazard model, has been used quite extensively for survival analysis since it was suggested by Cox (1972):

$$h(y|x) = h_0(y) \exp(x'\beta), \quad (8.25)$$

where y is the survival time for an individual, $h(y|x)$ is the hazard function with given explanatory variable $x \in \mathbb{R}^p$, β is a p -vector of unknown parameters, and $h_0(y)$ is an unknown baseline hazard function common to all individuals in the study.

We consider randomly censored observations $(t_1, x'_1, \delta_1), \dots, (t_n, x'_n, \delta_n)$, where $t_i = \min(y_i, z_i)$, y_i is the i th survival time, z_i is the i th censoring time, $\delta_i = I\{y_i \leq z_i\}$, and x_i is the i th value of the explanatory variable. The parameter of primary interest, β , is estimated by maximizing the partial likelihood function

$$p_n(\gamma) = \prod_{i=1}^n \left[\exp(x'_i \gamma) / \sum_{j \in R_i} \exp(x'_j \gamma) \right]^{\delta_i}.$$

(the product of conditional probabilities), where R_i is the set of indices corresponding to individuals who survived until time t_i and is called the risk set at time t_i . Since the partial likelihood function is differentiable, the maximum partial likelihood estimator $\hat{\beta}_{MP}$ of β is a solution of

$$s_n(\gamma) = \nabla \log(p_n(\gamma)) = 0,$$

where

$$s_n(\gamma) = \sum_{i=1}^n \delta_i \left[x_i - \sum_{j \in R_i} x_j \exp(x'_j \gamma) \right] \Bigg/ \sum_{j \in R_i} \exp(x'_j \gamma).$$

Assume that x_i is random and the (t_i, x'_i, δ_i) are i.i.d.; $E[x_i \exp(x'_i \gamma)]^2$ is bounded uniformly in a neighborhood of β ; and $z_i \leq T_0$ but $P\{t_i \geq T_0\} > 0$ for a specific T_0 . Let $\theta = g(\beta)$ and $\hat{\theta}_{MP} = g(\hat{\beta}_{MP})$ with a known function g differentiable at β . Then $\hat{\theta}_{MP} - \theta$ is asymptotically normal with mean 0 and asymptotic variance $n^{-1} \nabla g(\beta)' [\mathcal{I}(\beta)]^{-1} \nabla g(\beta)$ (Tsiatis, 1981), where

$$\mathcal{I}(\gamma) = \int \int v(t, \gamma) \exp(x' \gamma) G(t|x) h_0(t) dF(x) dt, \quad (8.26)$$

$G(t|x) = P\{t_i \geq t | x_i = x\}$, $F(x)$ is the distribution of x_i , $v(t, \gamma) = u_2(t, \gamma) - u(t, \gamma)u(t, \gamma)'$,

$$u(t, \gamma) = \int x \exp(x' \gamma) G(t|x) dF(x) \Bigg/ \int \exp(x' \gamma) G(t|x) dF(x),$$

and

$$u_2(t, \gamma) = \int xx' \exp(x' \gamma) G(t|x) dF(x) \Bigg/ \int \exp(x' \gamma) G(t|x) dF(x).$$

8.3.1 Jackknife variance estimators

Under model (8.25) with randomly censored data, the jackknife can be applied to estimate the asymptotic variance of $\hat{\theta}_{MP} = g(\hat{\beta}_{MP})$ by deleting the triple (t_i, x'_i, δ_i) at a time. For each i , let $\hat{\beta}_{MP,i}$ be a solution of

$$s_{ni}(\gamma) = \sum_{k \neq i} \delta_k \left[x_k - \sum_{j \in R_{k,i}} x_j \exp(x'_j \gamma) \right] \Bigg/ \sum_{j \in R_{k,i}} \exp(x'_j \gamma) = 0,$$

where $R_{k,i}$ is the risk set at t_k after (t_i, x'_i, δ_i) is deleted (may not be the same as R_k). The exact jackknife variance estimator is then defined the same as before.

We now consider simplifying the computation of the jackknife. The one-step jackknife variance estimator can be defined the same as before, based on

$$\hat{\beta}_{MP,i}^{[1]} = \hat{\beta}_{MP} - [\nabla s_{ni}(\hat{\beta}_{MP})]^{-1} s_{ni}(\hat{\beta}_{MP}), \quad i = 1, \dots, n.$$

A further simplification can be achieved by first replacing $R_{k,i}$ with R_k for all i and then approximating $\nabla s_{ni}(\hat{\beta}_{MP})$ by $\nabla s_n(\hat{\beta}_{MP})$. Using the fact that

$s_n(\hat{\beta}_{\text{MP}}) = 0$, we can replace $\hat{\beta}_{\text{MP},i}^{[1]}$ by

$$\tilde{\beta}_{\text{MP},i} = \hat{\beta}_{\text{MP}} + [\nabla s_n(\hat{\beta}_{\text{MP}})]^{-1} \delta_i \left[x_i - \sum_{j \in R_i} x_j \exp(x'_j \hat{\beta}_{\text{MP}}) \right] / \sum_{j \in R_i} \exp(x'_j \hat{\beta}_{\text{MP}}).$$

This leads to the linear jackknife estimator

$$v_{\text{LJACK}} = \frac{n-1}{n} \sum_{i=1}^n \left(\tilde{\theta}_{\text{MP},i} - \frac{1}{n} \sum_{i=1}^n \tilde{\theta}_{\text{MP},i} \right)^2$$

with $\tilde{\theta}_{\text{MP},i} = g(\tilde{\beta}_{\text{MP},i})$. In the linear case of $g(\beta) = c' \beta$,

$$v_{\text{LJACK}} = (n-1)c'[\nabla s_n(\hat{\beta}_{\text{MP}})]^{-1} \hat{\mathcal{I}}_n(\hat{\beta}_{\text{MP}})[\nabla s_n(\hat{\beta}_{\text{MP}})]^{-1}c,$$

where

$$\hat{\mathcal{I}}_n(\gamma) = \frac{1}{n} \sum_{i=1}^n \delta_i [x_i - \hat{u}(t_i, \gamma)][x_i - \hat{u}(t_i, \gamma)]'$$

and

$$\hat{u}(t_i, \gamma) = \sum_{j \in R_i} x_j \exp(x'_j \gamma) / \sum_{j \in R_i} \exp(x'_j \gamma).$$

The consistency of v_{LJACK} was established by Quan and Tsai (1992) and can be illustrated as follows. By Lemma A.1 in Tsiatis (1981),

$$\max_{i \leq n} \|\hat{u}(t_i, \hat{\beta}_{\text{MP}}) - u(t_i, \hat{\beta}_{\text{MP}})\| \rightarrow_{a.s.} 0$$

and

$$\max_{i \leq n} \|\hat{u}_2(t_i, \hat{\beta}_{\text{MP}}) - u_2(t_i, \hat{\beta}_{\text{MP}})\| \rightarrow_{a.s.} 0,$$

where $u(t, \gamma)$ and $u_2(t, \gamma)$ are defined in (8.26),

$$\hat{u}_2(t_i, \gamma) = \sum_{j \in R_i} x_j x'_j \exp(x'_j \gamma) / \sum_{j \in R_i} \exp(x'_j \gamma),$$

and $\|A\| = [\text{tr}(A'A)]^{1/2}$ for any matrix A . Using these results and the uniform strong law of large numbers (e.g., Corollary A in Wu, 1981, or Lemma 2 in Shao, 1992f), we obtain that

$$\begin{aligned} \hat{A}_{1n} &= \frac{1}{n} \sum_{i=1}^n \delta_i x_i \hat{u}(t_i, \hat{\beta}_{\text{MP}})' \\ &= \frac{1}{n} \sum_{i=1}^n \delta_i x_i u(t_i, \hat{\beta}_{\text{MP}})' + o(1) \quad a.s. \\ &= E[\delta_i x_i u(t_i, \beta)'] + o(1) \quad a.s. \\ &= \int \int u(t, \beta) u(t, \beta)' \exp(x' \beta) G(t|x) h_0(t) dF(x) dt + o(1) \quad a.s., \end{aligned}$$

$$\begin{aligned}
\hat{A}_{2n} &= \frac{1}{n} \sum_{i=1}^n \delta_i \hat{u}(t_i, \hat{\beta}_{\text{MP}}) \hat{u}(t_i, \hat{\beta}_{\text{MP}})' \\
&= \frac{1}{n} \sum_{i=1}^n \delta_i u(t_i, \hat{\beta}_{\text{MP}}) u(t_i, \hat{\beta}_{\text{MP}})' + o(1) \quad a.s. \\
&= E[\delta_i u(t_i, \beta) u(t_i, \beta)'] + o(1) \quad a.s. \\
&= \int \int u(t, \beta) u(t, \beta)' \exp(x' \beta) G(t|x) h_0(t) dF(x) dt + o(1) \quad a.s., \\
\hat{A}_{3n} &= \frac{1}{n} \sum_{i=1}^n \delta_i \hat{u}_2(t_i, \hat{\beta}_{\text{MP}}) \\
&= \frac{1}{n} \sum_{i=1}^n \delta_i u_2(t_i, \hat{\beta}_{\text{MP}}) + o(1) \quad a.s. \\
&= E[\delta_i x_i u_2(t_i, \beta)] + o(1) \quad a.s. \\
&= \int \int u_2(t, \beta) \exp(x' \beta) G(t|x) h_0(t) dF(x) dt + o(1) \quad a.s.,
\end{aligned}$$

and

$$\begin{aligned}
\hat{A}_{4n} &= \frac{1}{n} \sum_{i=1}^n \delta_i x_i x_i' \\
&= E(\delta_i x_i x_i') + o(1) \quad a.s. \\
&= \int \int u_2(t, \beta) \exp(x' \beta) G(t|x) h_0(t) dF(x) dt + o(1) \quad a.s.
\end{aligned}$$

By (8.26),

$$\hat{\mathcal{I}}_n(\hat{\beta}_{\text{MP}}) = A_{4n} - A_{1n} - A'_{1n} + A_{2n} \rightarrow_{a.s.} \mathcal{I}(\beta),$$

$$-n^{-1} \nabla s_n(\hat{\beta}_{\text{MP}}) = A_{3n} - A_{2n} \rightarrow_{a.s.} \mathcal{I}(\beta)$$

and, therefore,

$$\frac{n(n-1)c'[\nabla s_n(\hat{\beta}_{\text{MP}})]^{-1}\hat{\mathcal{I}}_n(\hat{\beta}_{\text{MP}})[\nabla s_n(\hat{\beta}_{\text{MP}})]^{-1}c}{c'[\mathcal{I}(\beta)]^{-1}c} \rightarrow_{a.s.} 1.$$

This shows the consistency of v_{LJACK} for the linear case. The result for nonlinear g can be proved by using an argument similar to the one in the proof of Theorem 7.1 and by showing

$$\max_{i \leq n} \|\tilde{\beta}_{\text{MP},i} - \hat{\beta}_{\text{MP}}\|^2 \leq n[\nabla s_n(\hat{\beta}_{\text{MP}})]^{-1}\hat{\mathcal{I}}_n(\hat{\beta}_{\text{MP}})[\nabla s_n(\hat{\beta}_{\text{MP}})]^{-1} \rightarrow_{a.s.} 0.$$

Consistency of the exact and one-step jackknife variance estimators can also be established.

8.3.2 Bootstrap procedures

For randomly censored data and random x_i , the paired bootstrap described in Section 7.2.2 can be extended in a straightforward manner (Efron and Tibshirani, 1986): we resample triples $(t_i^*, x_i^*, \delta_i^*)$, $i = 1, \dots, n$, i.i.d. from the empirical distribution putting mass n^{-1} to (t_i, x_i', δ_i) , $i = 1, \dots, n$, and define $\hat{\beta}_{\text{MP}}^*$ as a solution of

$$s_n^*(\gamma) = \sum_{i=1}^n \delta_i^* \left[x_i^* - \sum_{j \in R_i^*} x_j^* \exp(x_j'^* \gamma) \middle/ \sum_{j \in R_i^*} \exp(x_j'^* \gamma) \right] = 0,$$

where R_i^* is the risk set at t_i^* .

To avoid the computation of R_i^* , one may use the following alternative: let $\mathbf{P}^* = (P_1^*, \dots, P_n^*)'$ be an n -vector such that $n\mathbf{P}^*$ is distributed as multinomial with parameters n and $(\frac{1}{n}, \dots, \frac{1}{n})'$ and define $\hat{\beta}_{\text{MP}}^*$ as a solution of

$$\tilde{s}_n^*(\gamma) = \sum_{i=1}^n P_i^* \delta_i \left[x_i - \sum_{j \in R_i} x_j \exp(x_j' \gamma) \middle/ \sum_{j \in R_i} \exp(x_j' \gamma) \right] = 0.$$

The one-step bootstrap estimators can be obtained based on

$$\hat{\beta}_{\text{MP}}^{*[1]} = \hat{\beta}_{\text{MP}} - [\nabla s_n(\hat{\beta}_{\text{MP}})]^{-1} \tilde{s}_n^*(\hat{\beta}_{\text{MP}}).$$

One may further simplify the computation by replacing $\hat{\beta}_{\text{MP}}^{*[1]}$ with

$$\tilde{\beta}_{\text{MP}}^* = \hat{\beta}_{\text{MP}} - [\nabla s_n(\hat{\beta}_{\text{MP}})]^{-1} \tilde{s}_n^*(\hat{\beta}_{\text{MP}}).$$

The consistency of the distribution estimator for $\hat{\beta}_{\text{MP}}$ based on the bootstrap distribution of $\tilde{\beta}_{\text{MP}}^*$ can be easily established. First, define

$$\xi_i = \delta_i \left[x_i - \sum_{j \in R_i} x_j \exp(x_j' \hat{\beta}_{\text{MP}}) \middle/ \sum_{j \in R_i} \exp(x_j' \hat{\beta}_{\text{MP}}) \right]$$

and let ξ_1^*, \dots, ξ_n^* be i.i.d. from $\{\xi_1, \dots, \xi_n\}$. Then

$$\tilde{\beta}_{\text{MP}}^* - \hat{\beta}_{\text{MP}} = [\nabla s_n(\hat{\beta}_{\text{MP}})]^{-1} \sum_{i=1}^n \xi_i^*$$

is asymptotically normal (Theorem 3.1), conditioned on (t_i, x_i', δ_i) , $i = 1, 2, \dots$. Second,

$$E_*[\tilde{s}_n^*(\hat{\beta}_{\text{MP}})] = \sum_{i=1}^n E_*(\xi_i^*) = \sum_{i=1}^n \xi_i = s_n(\hat{\beta}_{\text{MP}}) = 0$$

so that the asymptotic mean of $\tilde{\beta}_{\text{MP}}^* - \hat{\beta}_{\text{MP}}$ is 0. Finally,

$$\begin{aligned}\text{var}_*(\tilde{\beta}_{\text{MP}}^*) &= [\nabla s_n(\hat{\beta}_{\text{MP}})]^{-1} \sum_{i=1}^n \text{var}_*(\xi_i^*) [\nabla s_n(\hat{\beta}_{\text{MP}})]^{-1} \\ &= [\nabla s_n(\hat{\beta}_{\text{MP}})]^{-1} \sum_{i=1}^n \xi_i \xi_i' [\nabla s_n(\hat{\beta}_{\text{MP}})]^{-1} \\ &= \frac{n}{n-1} v_{\text{LJACK}},\end{aligned}$$

which is a consistent estimator of $n^{-1}[\mathcal{I}(\beta)]^{-1}$, the asymptotic covariance matrix of $\hat{\beta}_{\text{MP}}$. Therefore, $\tilde{\beta}_{\text{MP}}^* - \hat{\beta}_{\text{MP}}$ has the same asymptotic distribution as $\hat{\beta}_{\text{MP}} - \beta$.

The consistency of the distribution estimator for $\hat{\beta}_{\text{MP}}$ based on the bootstrap distribution of $\hat{\beta}_{\text{MP}}^*$ was proved by Gu (1992). Using the Eggedworth expansions, Gu (1992) also showed the second order accuracy of the bootstrap distribution of $[\mathcal{I}(\hat{\beta}_{\text{MP}}^*)]^{1/2}(\hat{\beta}_{\text{MP}}^* - \hat{\beta}_{\text{MP}})$ as an estimator of the distribution of the studentized variable $[\mathcal{I}(\hat{\beta}_{\text{MP}})]^{1/2}(\hat{\beta}_{\text{MP}} - \beta)$.

Although the above bootstrap procedure can also be applied to the case where the x_i are deterministic, more efficient bootstrap procedures may be available when the x_i are nonrandom. For example, we may apply the bootstrap or the external bootstrap based on the generalized residuals (Cox and Snell, 1968). Loughin and Koehler (1993) proposed generating bootstrap data from the empirical distribution putting mass n^{-1} to $(\hat{S}(t_i|x_i), \delta_i)$, $i = 1, \dots, n$, where $\hat{S}(t_i|x_i)$ is an estimator of $S(t_i|x_i) = [S_0(t_i)]^{\exp(x_i'\beta)}$, the probability of survival beyond time t_i for the i th individual with explanatory variables x_i . In the case where censoring is nonrandom, some parametric or semiparametric bootstrap procedures may also be employed (see Loughin and Koehler, 1993). Burr (1994) compared several bootstrap confidence intervals for β .

There are other applications of the bootstrap to regression models, which are closely related to model (8.25), in survival analysis. See Barlow and Sun (1989) and Garrison (1990) for results in a linear relative risk model and a piecewise exponential model, respectively.

8.4 Kernel Density Estimation

Kernel density curve estimation is the simplest among the nonparametric curve estimation problems studied in the rest of this chapter, but many results established in this section hold in more complicated curve estimation problems to be discussed later. Let X_1, \dots, X_n be i.i.d. from a univariate

distribution F with density f . A kernel estimator of $f(x)$ is

$$\hat{f}_\lambda(x) = \frac{1}{n\lambda} \sum_{i=1}^n \kappa\left(\frac{x - X_i}{\lambda}\right),$$

where $\kappa(x)$ is a given kernel function assumed to be symmetric about 0, $\int \kappa(x)dx = 1$ and $\kappa_2 = \int x^2 \kappa(x)dx \neq 0$, and $\lambda \geq 0$ is the bandwidth that determines the smoothness of the estimated density curve. Note that both \hat{f}_λ and λ depend on n , but the subscript n is omitted for simplicity. Detailed discussions about kernel density estimation can be found, for example, in Praksa Rao (1983) or Silverman (1986).

The results in this section can be generalized to the case where X_i is multivariate and/or f is estimated by other methods (e.g., Sain, Baggerly and Scott, 1994).

8.4.1 Bandwidth selection by cross-validation

An important step in kernel density estimation is the selection of the bandwidth or the smoothing parameter λ ; that is, we would like to select a λ so that the mean integrated squared error, defined by

$$\text{mise}(\hat{f}_\lambda) = E \int [\hat{f}_\lambda(x) - f(x)]^2 dx,$$

is as small as possible. Assume that f is twice differentiable and f'' is Lipschitz continuous. The asymptotic approach approximates $\text{mise}(\hat{f}_\lambda)$ by

$$\frac{\kappa_2^2 \lambda^4}{4} \int [f''(x)]^2 dx + \frac{1}{n\lambda} \int [\kappa(x)]^2 dx,$$

which is minimized by

$$\lambda_{\text{AM}} = \left\{ \frac{\int [\kappa(x)]^2 dx}{\kappa_2^2 n \int [f''(x)]^2 dx} \right\}^{1/5}. \quad (8.27)$$

Since $\int [f''(x)]^2 dx$ is unknown, it has to be estimated. The derivation of an estimate of $\int [f''(x)]^2 dx$ using the asymptotic approach involves many complicated analytic arguments, and the details can be found, for example, in Park and Marron (1990). Denote the estimated λ_{AM} by $\hat{\lambda}_{\text{AM}}$ and the minimizer of $\text{mise}(\hat{f}_\lambda)$ by λ_{MISE} . Assume that κ is a probability density with a bounded support and has four Lipschitz continuous derivatives; f'' is Lipschitz continuous and square integrable; and

$$|f^{(2+l)}(x) - f^{(2+l)}(y)| \leq M|x - y|^\eta \quad \text{for all } x, y,$$

where $f^{(k)}$ is the k th derivative of f , l is an integer, M and $\eta \in (0, 1]$ are constants, and $l + \eta > 2$. Then

$$n^{4/13} \left(\frac{\hat{\lambda}_{\text{AM}}}{\lambda_{\text{MISE}}} - 1 \right) \rightarrow_d N(\mu_{\text{AM}}, \tau_{\text{AM}}) \quad (8.28)$$

with some parameters μ_{AM} and $\tau_{\text{AM}} > 0$ (Park and Marron, 1990).

Rudemo (1982) and Bowman (1984) applied the cross-validation (CV) method described in Section 7.4.1 for bandwidth selection. Let

$$\hat{f}_{\lambda,i}(x) = \frac{1}{(n-1)\lambda} \sum_{j \neq i} \kappa\left(\frac{x - X_j}{\lambda}\right),$$

the density estimator after deleting X_i , and let f_i be an estimator of f based on X_i and used to validate $\hat{f}_{\lambda,i}$, $i = 1, \dots, n$. The CV selects a λ by minimizing

$$\frac{1}{n} \sum_{i=1}^n \int [f_i(x) - \hat{f}_{\lambda,i}(x)]^2 dx \quad (8.29)$$

over λ . With only one observation, however, it is difficult to construct a reasonable density estimator f_i . Noting that minimizing the quantity in (8.29) is the same as minimizing

$$\frac{1}{n} \sum_{i=1}^n \int [\hat{f}_{\lambda,i}(x)]^2 dx - \frac{2}{n} \sum_{i=1}^n \int \hat{f}_{\lambda,i}(x) dF_i(x),$$

where F_i is the distribution function corresponding to f_i , we can substitute F_i with the degenerate distribution at X_i , which leads to selecting a λ by minimizing

$$\widehat{\text{mise}}_{\text{cv}}(\lambda) = \frac{1}{n} \sum_{i=1}^n \int [\hat{f}_{\lambda,i}(x)]^2 dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{\lambda,i}(X_i).$$

In view of (8.27), in practice, the search for the minimum of $\widehat{\text{mise}}_{\text{cv}}(\lambda)$, denoted by $\hat{\lambda}_{\text{cv}}$, is often confined to the interval $[\epsilon_1 n^{-1/5}, \epsilon_2 n^{-1/5}]$ with two constants ϵ_1 and ϵ_2 . Hall (1983b) proved that

$$\hat{\lambda}_{\text{cv}} / \lambda_{\text{MISE}} \rightarrow_p 1 \quad \text{and} \quad \frac{\text{mise}(\hat{f}_{\hat{\lambda}_{\text{cv}}})}{\min_{\lambda} \text{mise}(\hat{f}_{\lambda})} \rightarrow_p 1, \quad (8.30)$$

under the assumption that κ is nonnegative and of bounded variation on \mathbb{R} and has two bounded derivatives; $\int x^2 [\kappa(x) + |\kappa'(x)| + |\kappa''(x)|] dx < \infty$; $\int |x|^{5/2} \kappa(x) dx < \infty$; $\int |x \kappa'(x)| dx < \infty$; f has two bounded derivatives with f'' uniformly continuous and square integrable; $\int |f'(x)| dx < \infty$; and

$\int \{F(x)[1 - F(x)]\}^{1/2} dx < \infty$. This shows that $\hat{\lambda}_{cv}$ minimizes $\text{mise}(\hat{f}_\lambda)$ asymptotically.

The following formula, which is asymptotically equivalent to $\widehat{\text{mise}}_{cv}(\lambda)$, is often preferred by many researchers:

$$\int [\hat{f}_\lambda(x)]^2 dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{\lambda,i}(X_i). \quad (8.31)$$

Stone (1984) suggested another approximation:

$$\frac{1}{n^2 \lambda} \sum_{i=1}^n \sum_{j=1}^n \kappa_c \left(\frac{X_i - X_j}{\lambda} \right) + \frac{2\kappa(0)}{n\lambda}, \quad (8.32)$$

where $\kappa_c(t) = \int \kappa(x-t)d\kappa(x) - 2\kappa(t)$. The λ selected by minimizing the quantity in (8.31) or (8.32) has the same asymptotic properties as $\hat{\lambda}_{cv}$ and, therefore, is still denoted by $\hat{\lambda}_{cv}$.

Under similar conditions for result (8.28), Park and Marron (1990) showed that

$$n^{1/10} \left(\frac{\hat{\lambda}_{cv}}{\lambda_{\text{MISE}}} - 1 \right) \xrightarrow{d} N(0, \tau_{cv}) \quad (8.33)$$

with some $\tau_{cv} > 0$. Hence, the convergence rate of $\hat{\lambda}_{cv}$ is $n^{-1/10}$. They further showed that a bias adjusted CV method (Scott and Terrell, 1987), which produces $\hat{\lambda}_{bcv}$ by minimizing

$$\int [\hat{f}_\lambda(x)]^2 dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{\lambda,i}(X_i) - \frac{\int [\kappa''(x)]^2 dx}{n\lambda^5},$$

satisfies

$$n^{1/10} \left(\frac{\hat{\lambda}_{bcv}}{\lambda_{\text{MISE}}} - 1 \right) \xrightarrow{d} N(0, \tau_{bcv}),$$

where τ_{bcv} may be much smaller than τ_{cv} .

In view of (8.28), $\hat{\lambda}_{am}$ has a better convergence rate than $\hat{\lambda}_{cv}$ or $\hat{\lambda}_{bcv}$. Thus, the use of the CV is purely because of its easier implementation. Some improvements over $\hat{\lambda}_{cv}$ were proposed by Feluch and Koronachi (1992) and Stute (1992).

8.4.2 Bandwidth selection by bootstrapping

Bootstrap bandwidth selection is carried out by minimizing a bootstrap estimator of $\text{mise}(\hat{f}_\lambda)$ (Taylor, 1989; Faraway and Jhun, 1990):

$$E_* \int [\hat{f}_\lambda^*(x) - \hat{f}_\lambda(x)]^2 dx,$$

where

$$\hat{f}_\lambda^*(x) = \frac{1}{n\lambda} \sum_{i=1}^n \kappa\left(\frac{x - X_i^*}{\lambda}\right) \quad (8.34)$$

and X_1^*, \dots, X_n^* are bootstrap data i.i.d. from an estimator \hat{F} of F .

If we take the empirical distribution of X_1, \dots, X_n as \hat{F} , then

$$E_*[\hat{f}_\lambda^*(x)] = \hat{f}_\lambda(x)$$

for all x , but the squared bias $[E\hat{f}_\lambda(x) - f(x)]^2$ is not a negligible term in $\text{mise}(\hat{f}_\lambda)$ when $\lambda n^{1/5} \not\rightarrow 0$. In view of (8.27), we have to modify the bootstrap procedure. Since the problem is that \hat{f}_λ^* is too close to \hat{f}_λ , Hall (1990d) proposed generating m , instead of n , bootstrap data from the empirical distribution (also, see Sections 3.6 and 7.4.2) and showed that if $m = \sqrt{n}$, then under some conditions the selected bandwidth converges at the same rate as that of the bandwidth selected by the CV method.

It is more natural in this case to consider the smoothed bootstrap (see Section 3.5) based on bootstrap data generated from a kernel density estimator, since the population density exists and is estimated by a kernel density estimator. Suppose that X_1^*, \dots, X_n^* are generated from the estimated density

$$\hat{f}_\gamma(x) = \frac{1}{n\gamma} \sum_{i=1}^n \kappa\left(\frac{x - X_i}{\gamma}\right),$$

where γ is a bandwidth that may be different from λ . Then the bootstrap method selects a λ by minimizing

$$\widehat{\text{mise}}_{\text{BOOT}}(\lambda) = E_* \int [\hat{f}_\lambda^*(x) - \hat{f}_\gamma(x)]^2 dx, \quad (8.35)$$

where \hat{f}_λ^* is given by (8.34). In this case, $\widehat{\text{mise}}_{\text{BOOT}}(\lambda)$ has an explicit form:

$$\begin{aligned} \widehat{\text{mise}}_{\text{BOOT}}(\lambda) &= \frac{1}{n\lambda} \int [\kappa(x)]^2 dx + \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \kappa_\gamma * \kappa_\gamma(X_i - X_j) \\ &\quad - \frac{2}{n^2} \sum_{i=1}^n \sum_{j=1}^n \kappa_\lambda * \kappa_\gamma * \kappa_\gamma(X_i - X_j) \\ &\quad + \frac{n+1}{n^3} \sum_{i=1}^n \sum_{j=1}^n \kappa_\lambda * \kappa_\lambda * \kappa_\gamma * \kappa_\gamma(X_i - X_j), \end{aligned}$$

where $\kappa_a(x) = a^{-1} \kappa(x/a)$ and $*$ is the convolution operator.

Taylor (1989) considered $\gamma = \lambda$. If $\kappa(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$, the standard normal density, then

$$\widehat{\text{mise}}_{\text{BOOT}}(\lambda) = \frac{1}{2n\lambda\sqrt{\pi}} + \frac{1}{2n^2\lambda} \sum_{i=1}^n \sum_{j=1}^n \left[\varphi\left(\frac{X_i - X_j}{2\lambda}\right) - \frac{4}{\sqrt{3}} \varphi\left(\frac{X_i - X_j}{\sqrt{3}\lambda}\right) + \sqrt{2} \varphi\left(\frac{X_i - X_j}{\sqrt{2}\lambda}\right) \right], \quad (8.36)$$

which has a bias of the order $n^{-1}\lambda^{-1}$ as an estimator of $\text{mise}(\hat{f}_\lambda)$. To correct the bias, Taylor (1989) suggested replacing the summations over i and j in (8.36) by the summation over all $i \neq j$ and showed by simulation that the resulting bandwidth, denoted by $\hat{\lambda}_{\text{BOOT}}$, may be closer to λ_{MISE} than the bandwidth selected by the CV. Hall, Marron and Park (1991) showed that

$$\frac{\hat{\lambda}_{\text{BOOT}}}{\lambda_{\text{MISE}}} - 1 = O_p(n^{-1/10}).$$

Faraway and Jhun (1990) proposed taking $\gamma = \hat{\lambda}_{\text{cv}}$, the bandwidth selected using the CV. Let $\hat{\lambda}_{\text{BOOT}}^{[1]}$ be the selected bandwidth using this method. Jones (1991) argued that $\hat{\lambda}_{\text{BOOT}}^{[1]}$ has almost the same asymptotic properties as $\hat{\lambda}_{\text{BOOT}}$; however, $\hat{f}_{\hat{\lambda}_{\text{BOOT}}}$ has a tendency to gross over smoothing, but $\hat{f}_{\hat{\lambda}_{\text{BOOT}}^{[1]}}$ may correct this somewhat. This method can be applied iteratively; that is, we can let $\gamma = \hat{\lambda}_{\text{BOOT}}^{[1]}$ and repeat the bootstrap selection procedure to get a new bandwidth $\hat{\lambda}_{\text{BOOT}}^{[2]}$. Faraway and Jhun (1990) compared by a simulation of size 400 the performances of $\hat{\lambda}_{\text{BOOT}}^{[1]}$, $\hat{\lambda}_{\text{BOOT}}^{[2]}$, and $\hat{\lambda}_{\text{cv}}$ in terms of the measure

$$\rho(\hat{\lambda}) = \frac{\text{mise}(\hat{f}_{\hat{\lambda}})}{\min_{\lambda} \text{mise}(\hat{f}_{\lambda})}.$$

The kernel $\kappa(x)$ was $\frac{3}{4}(1 - x^2)I\{|x| < 1\}$ and six different $f(x)$ were considered: (1) the standard normal density; (2) the density of the bimodal normal $\frac{1}{2}N(-1, \frac{1}{4}) + \frac{1}{2}N(1, \frac{1}{4})$; (3) the density of the mixture normal $\frac{1}{2}N(0, 4) + \frac{1}{2}N(0, \frac{1}{4})$; (4) the standard lognormal density; (5) the standard Cauchy density; and (6) the Beta density of parameters 2 and 2. The bootstrap estimator $\widehat{\text{mise}}_{\text{BOOT}}(\lambda)$ is approximated by Monte Carlo with size 100.

The results shown in Table 8.1 indicate that the bootstrap selection procedures are uniformly better than the CV, and the iterative bootstrap produces some further improvement. But both bootstrap procedures require substantially more computations than the CV.

Table 8.1. Simulation estimates of $\rho(\hat{\lambda})$ [Adapted from Faraway and Jhun (1990), by permission of American Statistical Association]

n	$\hat{\lambda}$	Population density					
		Normal	Bimodal	Mixture	Lognormal	Cauchy	Beta
50	$\hat{\lambda}_{\text{CV}}$	2.05	1.39	1.69	1.45	1.57	1.84
	$\hat{\lambda}_{\text{BOOT}}^{[1]}$	1.69	1.34	1.51	1.37	1.41	1.50
	$\hat{\lambda}_{\text{BOOT}}^{[2]}$	1.50	1.38	1.47	1.38	1.35	1.38
400	$\hat{\lambda}_{\text{CV}}$	1.51	1.24	1.49	1.29	1.40	1.48
	$\hat{\lambda}_{\text{BOOT}}^{[1]}$	1.29	1.13	1.28	1.17	1.24	1.28
	$\hat{\lambda}_{\text{BOOT}}^{[2]}$	1.22	1.10	1.20	1.16	1.19	1.20

All of the bootstrap procedures discussed so far produce estimators of λ_{MISE} with a convergence rate $n^{-1/10}$. Observing that if f has more than four derivatives and κ is a probability density,

$$\frac{\hat{\lambda}_{\text{BOOT}}}{\lambda_{\text{MISE}}} - 1 \approx \frac{c_1 Z}{n\gamma^{9/2}} + \frac{c_2}{\gamma^2} + \frac{c_3}{n\gamma^5}, \quad (8.37)$$

where Z is a standard normal random variable and the c_k are constants depending on f and κ , Hall, Marron and Park (1991) and Jones, Marron and Park (1991) suggested that γ be taken as the minimizer of the quantity on the right-hand side of (8.37); that is, $\gamma = c_4 n^{-1/7}$ for some c_4 depending on f and κ . With this choice of γ , it was shown that

$$\frac{\hat{\lambda}_{\text{BOOT}}}{\lambda_{\text{MISE}}} - 1 = O_p(n^{-5/14}).$$

This rate is better than the convergence rate of $\hat{\lambda}_{\text{AM}}$ given by (8.28). Data based methods for selecting c_4 have been suggested, but they depend on some heavy analytic arguments. If we allow γ to be a function of λ , i.e., $\gamma = c_5 n^u \lambda^{-2}$, then for a suitable u the convergence rate of $\hat{\lambda}_{\text{BOOT}}$ is $O_p(n^{-1/2})$. However, the selection of c_5 is not easy. Details can be found in the references previously cited. Falk (1992b) investigated the asymptotic influence of the choice of γ on the performance of the corresponding bootstrap estimator of $\text{misen}(\hat{f}_\lambda)$. Léger and Romano (1990) proposed a bootstrap method to select a tuning parameter in a certain class and applied their results to the bandwidth selection in density estimation.

8.4.3 Bootstrap confidence sets

We first consider the confidence sets for $f(x)$ with a fixed x , based on the kernel estimator \hat{f}_λ . We mainly discuss lower confidence bounds, since the

discussions for upper confidence bounds and equal-tail two-sided confidence intervals are similar. An application of the central limit theorem yields that

$$[\hat{f}_\lambda(x) - \mu(x)]/\sigma_\lambda(x) \rightarrow_d N(0, 1)$$

(e.g., Silverman, 1986), where

$$\mu(x) = E[\hat{f}_\lambda(x)] = \int \kappa(y)f(x - \lambda y)dy$$

and

$$\sigma_\lambda^2(x) = \text{var}[\hat{f}_\lambda(x)] = \frac{1}{n\lambda} \left\{ \int [\kappa(y)]^2 f(x - \lambda y)dy - \lambda[\mu(x)]^2 \right\}.$$

Furthermore, a consistent estimator of $\sigma_\lambda^2(x)$ is

$$\hat{\sigma}_\lambda^2(x) = \frac{1}{n\lambda} \left\{ \frac{1}{n\lambda} \sum_{i=1}^n \kappa^2\left(\frac{x - X_i}{\lambda}\right) - \lambda[\hat{f}_\lambda(x)]^2 \right\}.$$

Thus, the traditional asymptotic approach produces the following level $1-\alpha$ lower confidence bound for $\mu(x)$:

$$\underline{\mu}_{\text{NOR}}(x) = \hat{f}_\lambda(x) - z_{1-\alpha} \hat{\sigma}_\lambda(x).$$

To obtain a lower confidence bound for $f(x)$, one may simply use $\underline{\mu}_{\text{NOR}}(x)$ or a bias-corrected lower confidence bound

$$\underline{f}_{\text{NOR}}(x) = \underline{\mu}_{\text{NOR}}(x) - \hat{b}(x),$$

where $\hat{b}(x)$ is an estimator of the bias of $\hat{f}_\lambda(x)$. Assume that $\int x^t \kappa(x)dx = 0$ for $t = 1, \dots, r-1$ and $\int x^r \kappa(x)dx \neq 0$. Then we may adopt the bias estimator

$$\hat{b}(x) = \lambda^r \tilde{f}_\lambda^{(r)}(x) \int y^r \kappa(y)dy,$$

where

$$\tilde{f}_\lambda^{(r)}(x) = \frac{1}{n\tilde{\lambda}^{r+1}} \sum_{i=1}^n \tilde{\kappa}^{(r)}\left(\frac{x - X_i}{\tilde{\lambda}}\right),$$

$\tilde{\kappa}^{(r)}$ is the r th derivative of a new kernel $\tilde{\kappa}$, and $\tilde{\lambda}$ is a new bandwidth.

Bootstrap confidence sets for $f(x)$ were discussed in Hall (1991b, 1992b). Let X_1^*, \dots, X_n^* be i.i.d. from the empirical distribution of X_1, \dots, X_n , \hat{f}_λ^* be defined by (8.34),

$$\hat{\sigma}_\lambda^{*2}(x) = \frac{1}{n\lambda} \left\{ \frac{1}{n\lambda} \sum_{i=1}^n \kappa^2\left(\frac{x - X_i^*}{\lambda}\right) - \lambda[\hat{f}_\lambda^*(x)]^2 \right\},$$

$$H_{\text{BOOT}}(y) = P_*\{\hat{f}_\lambda^*(x) - \hat{f}_\lambda(x) \leq y\},$$

and

$$G_{\text{BOOT}}(y) = P_*\{[\hat{f}_\lambda^*(x) - \hat{f}_\lambda(x)]/\hat{\sigma}_\lambda^*(x) \leq y\}.$$

Then, for a given level $1 - \alpha$, the bootstrap percentile (see Section 4.1) lower confidence bound for $\mu(x)$ is

$$\underline{\mu}_{\text{BP}}(x) = \hat{f}_\lambda(x) + H_{\text{BOOT}}^{-1}(\alpha);$$

the hybrid bootstrap lower confidence bound for $\mu(x)$ is

$$\underline{\mu}_{\text{HB}}(x) = \hat{f}_\lambda(x) - H_{\text{BOOT}}^{-1}(1 - \alpha);$$

and the bootstrap-t lower confidence bound for $\mu(x)$ is

$$\underline{\mu}_{\text{BT}}(x) = \hat{f}_\lambda(x) - G_{\text{BOOT}}^{-1}(1 - \alpha)\hat{\sigma}_\lambda(x).$$

The bootstrap BC and BC_a confidence sets are not easy to compute because of the difficulty in computing z_0 and a .

The bootstrap percentile lower confidence bound for $f(x)$ can be either $\underline{\mu}_{\text{BP}}(x)$ or $f_{\text{BP}}(x) = \underline{\mu}_{\text{BP}}(x) - \hat{b}(x)$. The hybrid bootstrap and bootstrap-t lower confidence bounds can be similarly defined.

Hall (1991b) showed that, if f has $r + s$ derivatives, then

$$P\{\underline{f}_k(x) \leq f(x)\} = 1 - \alpha + cn^{-v_k} + o(n^{-v_k}),$$

where $v_k = 2b_k[r(2r+2s+1)+s]/[(2r+2b_k+1)(2r+2s+1)]$, $k = \text{BP, HB or BT}$, $b_k = \frac{1}{2}$ if $k = \text{BP or HB}$, and $b_k = 1$ if $k = \text{BT}$. The best convergence rate can be obtained by using $\lambda = c(n^{s/(2r+2s+1)-(b_k+1/2)})^{2/(2r+2b_k+1)}$ and $\tilde{\lambda} = cn^{-1/(2r+2s+1)}$. Similarly, if $\int x^t \kappa(x) dx = 0$ for $t = 1, \dots, r+s-1$ and $\int x^{r+s} \kappa(x) dx \neq 0$, then

$$P\{\underline{\mu}_k(x) \leq f(x)\} = 1 - \alpha + cn^{-u_k} + o(n^{-u_k}),$$

where $u_k = 2b_k(r+s)/(2r+2s+2b_k+1)$. The best convergence rate can be obtained by using $\lambda = cn^{-(2b_k+1)/(2r+2s+2b_k+1)}$. Since $u_k > v_k$, we conclude that the bootstrap method without bias correction, which is called the undersmoothing method, provides a greater degree of accuracy in terms of the coverage probability. The same conclusion can be drawn for the equal-tail two-sided bootstrap confidence intervals.

Hall (1992b) contains some simulation results for the performances of the bootstrap confidence intervals. The results indicate that the bootstrap confidence sets perform well if c in λ and $\tilde{\lambda}$ is suitably chosen and that the undersmoothing method can be substantially better than the method using bias correction.

Since f is a function of x , it is of interest to construct simultaneous confidence intervals (bands) for $f(x)$, $x \in \mathbb{R}$. The traditional asymptotic approach derives confidence bands by using the limit distribution of $\sup_x |\hat{f}_\lambda(x) - f(x)|$ (Bickel and Rosenblatt, 1973). However, the resulting formulas are quite complicated. Jhun (1988) proposed the following bootstrap confidence bands. Let X_1^*, \dots, X_n^* be i.i.d. from the empirical distribution of X_1, \dots, X_n , \hat{f}_λ^* be defined by (8.34),

$$M_n^* = \sup_x |[n\lambda\hat{f}_\lambda^{-1}(x)]^{1/2}[\hat{f}_\lambda^*(x) - \hat{f}_\lambda(x)]|,$$

and, for $\alpha \in (0, 1)$, c_α be any value between

$$\inf_t \{t : P_*\{M_n^* > t\sqrt{n\lambda}\} \leq \alpha\} \quad \text{and} \quad \inf_t \{t : P_*\{M_n^* \geq t\sqrt{n\lambda}\} \geq \alpha\}.$$

Then the level $1 - \alpha$ bootstrap confidence bands for $f(x)$, $x \in \mathbb{R}$, are

$$[\frac{1}{4}\{[c_\alpha + 4\hat{f}_\lambda(x)]^{1/2} - c_\alpha\}^2, \frac{1}{4}\{[c_\alpha + 4\hat{f}_\lambda(x)]^{1/2} + c_\alpha\}^2], \quad x \in \mathbb{R}. \quad (8.38)$$

Jhun (1988) proved the consistency of the bootstrap confidence bands (8.38) under the following conditions: $\kappa(x) \rightarrow 0$ as $|x| \rightarrow \infty$; $\int x^2\kappa(x)dx < \infty$; κ' exists and $\int |\kappa'(x)|^k dx < \infty$, $k = 1, 2$; $\int |x|^{3/2} \log(\log|x|)^{1/2} |\kappa'(x) + \kappa(x)| I\{|x| \geq 3\} dx < \infty$; f is positive and bounded; f'' and $f'/f^{1/2}$ are bounded; and $\lambda = n^{-\delta}$ with $\frac{1}{5} < \delta < \frac{1}{2}$. Hall (1993) applied a bias correction to the confidence bands (8.38) and compared the original and the bias-corrected bootstrap confidence bands.

Faraway and Jhun (1990) suggested another bootstrap procedure to cover the situation where λ is proportional to $n^{-1/5}$ (e.g., λ is selected using the methods in Sections 8.4.1-8.4.2) as a by-product of the bootstrap bandwidth selection. Let λ_0 be an initial bandwidth (e.g., $\lambda_0 = \hat{\lambda}_{cv}$). Generate B independent sets of bootstrap data from \hat{f}_{λ_0} . Let \hat{f}_b^* be the kernel density estimate based on the b th bootstrap data set and the bandwidth $\hat{\lambda}_b = \hat{\lambda}_{boot}^{[1]}$ defined in Section 8.4.2, and

$$M_b = \sup_x |\hat{f}_b^*(x) - \hat{f}_{\lambda_0}(x)|, \quad b = 1, \dots, B.$$

Then the proposed level $1 - 2\alpha$ bootstrap confidence bands are

$$[\hat{f}_{\hat{\lambda}_{boot}}(x) - M_{(\alpha)}, \hat{f}_{\hat{\lambda}_{boot}}(x) + M_{(\alpha)}], \quad x \in \mathbb{R}, \quad (8.39)$$

where $M_{(\alpha)}$ is the α th quantile of M_b , $b = 1, \dots, B$.

Since the convergence rates of the coverage probability of the bootstrap confidence bands and the confidence bands obtained from the asymptotic theory are slow, these methods require a very large sample size for their accuracy. Jhun (1988) showed by simulation that the bootstrap confidence bands (8.38) work well when $n = 100$.

8.5 Nonparametric Regression

Model (8.1) is an extension of model (7.1) from linear to nonlinear. The following model is a further extension from parametric to nonparametric:

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \dots, n, \quad (8.40)$$

where y_i is the i th response, x_i is the i th value of the explanatory variable, ε_i is the random error, and f is an unknown function. In general, there are two types of x_i : deterministic x_i (fixed design) and random x_i (random regressor). In the former case, the ε_i are assumed i.i.d. with mean 0 and variance σ^2 ; in the latter case, the (y_i, x_i) are assumed i.i.d. and $E(\varepsilon_i | x_i) = 0$.

There are similarities between this problem and the density estimation discussed in Section 8.4, since in both problems we need to estimate a curve $f(x)$.

There are different methods for model fitting, i.e., the estimation of f . We consider the methods of kernel, nearest neighbor, and smoothing spline. For simplicity, we assume that the x_i are univariate. The discussion for multivariate x_i is similar.

8.5.1 Kernel estimates for fixed design

Consider the case where the x_i are deterministic. Without loss of generality, we assume that $x_i \in [0, 1]$ for all i . A kernel estimator of f is given by

$$\hat{f}_\lambda(x) = \frac{1}{n\lambda} \sum_{i=1}^n y_i \kappa\left(\frac{x - x_i}{\lambda}\right), \quad (8.41)$$

where $\kappa(x)$ is a given kernel function and $\lambda \geq 0$ is a bandwidth (similar to those in Section 8.4).

The cross-validation (CV) method is often used to select a bandwidth that minimizes a global measure of the accuracy of \hat{f}_λ : the mean average squared errors defined by

$$\text{mase}(\hat{f}_\lambda) = E \left\{ \frac{1}{n} \sum_{i=1}^n w(x_i) [\hat{f}_\lambda(x_i) - f(x_i)]^2 \right\},$$

where $w(x)$ is a weight function used to eliminate boundary effects. If f'' exists and is uniformly continuous, then $\text{mase}(\hat{f}_\lambda)$ can be approximated by

$$\frac{\kappa_2^2 \lambda^4}{4} \int w(x) [f''(x)]^2 dx + \frac{\sigma^2}{n\lambda} \int [\kappa(x)]^2 dx \int w(x) dx,$$

where $\kappa_2 = \int x^2 \kappa(x) dx$ is assumed nonzero. The approximated $\text{mase}(\hat{f}_\lambda)$ is minimized by

$$\lambda_{\text{AM}} = \left\{ \frac{\sigma^2 \int [\kappa(x)]^2 dx \int w(x) dx}{\kappa_2^2 n \int w(x) [f''(x)]^2 dx} \right\}^{1/5}.$$

A substitution estimator of λ_{AM} can be obtained by replacing σ^2 and $\int w(x) [f''(x)]^2 dx$ by their estimators. But $\int w(x) [f''(x)]^2 dx$ is usually difficult to estimate. Another estimator of $\text{mase}(\hat{f}_\lambda)$ is the residual sum of squares

$$R(\hat{f}_\lambda) = \frac{1}{n} \sum_{i=1}^n w(x_i) [y_i - \hat{f}_\lambda(x_i)]^2.$$

But since $R(\hat{f}_\lambda)$ uses the same data to construct \hat{f}_λ and to assess \hat{f}_λ , it tends to be too optimistic and is negatively biased. Therefore, the λ selected by minimizing $R(\hat{f}_\lambda)$ is not a consistent estimator of λ_{AM} or λ_{MASE} , the minimizer of $\text{mase}(\hat{f}_\lambda)$. One way to overcome this difficulty is to multiply $R(\hat{f}_\lambda)$ by a correction factor of the form $c(n\lambda)$. Assuming some moment conditions for the distribution of ε_i and that κ is a probability density with a compact support, has a Lipschitz continuous second order derivative, and is symmetric about 0, Härdle, Hall and Marron (1988) proved that

$$n^{1/10} \left(\frac{\hat{\lambda}}{\lambda_{\text{MASE}}} - 1 \right) \rightarrow_d N(0, \tau) \quad (8.42)$$

for $\hat{\lambda}$ minimizing $c(n\lambda)R(\hat{f}_\lambda)$ with

$$c(n\lambda) = 1 + \frac{2\kappa(0)}{n\lambda} + O\left(\frac{1}{n^2\lambda^2}\right).$$

The CV method described in Section 7.4.1 can be directly extended to this problem. Let

$$\hat{f}_{\lambda,i}(x) = \frac{1}{(n-1)\lambda} \sum_{j \neq i} y_j \kappa\left(\frac{x - x_j}{\lambda}\right)$$

be the kernel estimator of $f(x)$ after deleting the pair (y_i, x_i) , $i = 1, \dots, n$. Since y_i and $\hat{f}_{\lambda,i}$ are independent, we may use y_i to assess the “prediction” $\hat{f}_{\lambda,i}(x_i)$. Hence, the CV method selects a λ by minimizing

$$\widehat{\text{mase}}_{\text{CV}}(\lambda) = \frac{1}{n} \sum_{i=1}^n w(x_i) [y_i - \hat{f}_{\lambda,i}(x_i)]^2.$$

Härdle, Hall and Marron (1988) showed that

$$\widehat{\text{mase}}_{\text{CV}}(\lambda) = \left[1 + \frac{2\kappa(0)}{n\lambda} + O\left(\frac{1}{n^2\lambda^2}\right) \right] R(\hat{f}_\lambda)$$

uniformly over $\lambda \in [n^{-1+\delta}, n]$ with a small $\delta > 0$. Therefore, the bandwidth selected by the CV method also satisfies (8.42). This result is parallel to result (8.33) in Section 8.4.1.

Some empirical comparisons among the CV and other bandwidth selection methods can be found in Härdle, Hall and Marron (1988).

For a fixed design, one is often interested in minimizing the mean squared error of \hat{f}_λ at a particular x :

$$\text{mse}(\hat{f}_\lambda(x)) = E[\hat{f}_\lambda(x) - f(x)]^2.$$

The minimizer of $\text{mse}(\hat{f}_\lambda(x))$, denoted by $\lambda_{\text{MSE}}(x)$, is called the locally optimal bandwidth.

The bootstrap can be applied to this problem. We first need to find a bootstrap estimator of $\text{mse}(\hat{f}_\lambda(x))$. Since the x_i are nonrandom, we apply the bootstrap based on residuals introduced in Section 7.2.2. A naive bootstrap estimator of $\text{mse}(\hat{f}_\lambda(x))$ is

$$E_*[\hat{f}_\lambda^*(x) - \hat{f}_\lambda(x)]^2,$$

where

$$\hat{f}_\lambda^*(x) = \frac{1}{n\lambda} \sum_{i=1}^n y_i^* \kappa\left(\frac{x - x_i}{\lambda}\right), \quad (8.43)$$

$y_i^* = \hat{f}_\lambda(x_i) + \varepsilon_i^*, \varepsilon_1^*, \dots, \varepsilon_n^*$ are i.i.d. from the empirical distribution putting mass $[(1 - 2\eta)n]^{-1}$ to $r_i - \bar{r}$, $\eta n + 1 \leq i \leq (1 - \eta)n - 1$ ($0 < \eta < \frac{1}{2}$ is used to exclude boundary effects), $r_i = y_i - \hat{f}_\lambda(x_i)$ is the i th residual, and \bar{r} is the average of r_i , $\eta n + 1 \leq i \leq (1 - \eta)n - 1$. But this bootstrap estimator is inconsistent when λ is of the order $n^{-1/5}$, because

$$E[\hat{f}_\lambda(x)] = f(x) + \frac{\lambda^2}{2} f''(x) + o(\lambda^2), \quad x \in [\eta, 1 - \eta],$$

whereas

$$\begin{aligned} E_*[\hat{f}_\lambda^*(x)] &= \frac{1}{n\lambda} \sum_{i=1}^n \hat{f}_\lambda(x_i) \kappa\left(\frac{x - x_i}{\lambda}\right) \\ &= \hat{f}_\lambda(x) + \frac{\lambda^2}{2} \hat{f}_\lambda''(x) + o(\lambda^2) \quad a.s., \quad x \in [\eta, 1 - \eta]. \end{aligned}$$

When λ is of order $n^{-1/5}$, $\hat{f}_\lambda'(x) \not\rightarrow f''(x)$, although $\hat{f}_\lambda(x) \rightarrow f(x)$ (Härdle, 1990, p. 33). Since the $\lambda_{\text{MSE}}(x)$ is of order $n^{-1/5}$, this bootstrap procedure does not produce a consistent estimator of $\lambda_{\text{MSE}}(x)$.

Härdle and Bowman (1988) proposed a remedy for this bootstrap procedure, but their method is complicated and involves the estimation of $f''(x)$. We now introduce some simple bootstrap procedures that yield consistent bandwidth estimators.

For a periodic function f with period 1, Faraway (1990) proposed to modify the residual r_i to

$$\tilde{r}_i = [y_i - \hat{f}_{\lambda_0}(x_i)] / \left[1 - \frac{\kappa(0)}{n\lambda} \right],$$

where λ_0 is an initial bandwidth satisfying $\lambda_0 \rightarrow 0$ and $n^{1/5}\lambda_0 \rightarrow \infty$. Let $\varepsilon_1^*, \dots, \varepsilon_n^*$ be i.i.d. from the centered empirical distribution of the modified residuals \tilde{r}_i , $y_i^* = \hat{f}_{\lambda_0}(x_i) + \varepsilon_i^*$, and \hat{f}_λ^* be defined by (8.43). The distribution of $\sqrt{n\lambda}[\hat{f}_\lambda(x) - f(x)]$ can be estimated by the bootstrap distribution of $\sqrt{n\lambda}[\hat{f}_\lambda^*(x) - \hat{f}_{\lambda_0}(x)]$. Then the bootstrap estimator of $\text{mse}(\hat{f}_\lambda(x))$ is

$$\widehat{\text{mse}}_{\text{BOOT}}(\lambda, x) = E_*[\hat{f}_\lambda^*(x) - \hat{f}_{\lambda_0}(x)]^2.$$

The bootstrap bandwidth $\hat{\lambda}_{\text{BOOT}}(x)$ is selected by minimizing $\widehat{\text{mse}}_{\text{BOOT}}(\lambda, x)$ and is shown to be consistent under the conditions (1) $x_i = i/n$; (2) f'' is continuous; and (3) κ is a probability density with a bounded support and $\int x^2\kappa(x)dx = 1$. This method can be used iteratively.

Hall (1990d) suggested a method to estimate $\lambda_{\text{MSE}}(x)$ by varying the bootstrap sample size. His idea can be summarized as follows. For simplicity, we assume that $x_i = i/n$. Generate $\varepsilon_1^*, \dots, \varepsilon_m^*$ i.i.d. from the centered empirical distribution of the residuals $y_i - \hat{f}_{\lambda_0}(\frac{i}{n})$, $i = 1, \dots, n$, where \hat{f}_{λ_0} is the estimator of f based on data $(y_i, \frac{i}{n})$, $i = 1, \dots, n$, and a bandwidth λ_0 of the order $n^{-1/5}$, and define $y_i^* = \hat{f}_{\lambda_0}(\frac{i}{m}) + \varepsilon_i^*$, $i = 1, \dots, m$. Let \hat{f}_λ^* be the estimator of f based on the data $(y_i^*, \frac{i}{m})$, $i = 1, \dots, m$, and $\hat{\lambda}_{\text{BOOT}-m}(x)$ be the minimizer of $E_*[\hat{f}_\lambda^*(x) - \hat{f}_{\lambda_0}(x)]^2$. Then the bandwidth selected using this method is

$$\hat{\lambda}_{\text{BOOT}}(x) = \left(\frac{m}{n}\right)^{1/5} \hat{\lambda}_{\text{BOOT}-m}(x).$$

For an estimator \hat{f}_λ slightly different from the kernel estimator defined by (8.41), Hall (1990d) showed that this bootstrap bandwidth is a consistent estimator of $\lambda_{\text{MSE}}(x)$, provided that $c_1 n^\delta \leq m \leq c_2 n^{1-\delta}$ for some positive constants c_1 , c_2 , and $\delta < 1$.

To avoid the boundary effects, we may generate bootstrap data based on trimmed residuals.

With some minor modifications, these bootstrap procedures can be applied to select a globally optimal bandwidth. For Hall's (1990d) method, it was suggested to take $m = \sqrt{n}$ in a global setting.

Bootstrap confidence sets for $f(x)$ with a fixed x can be obtained along the lines of the previous sections, provided that the bootstrap estimator of the distribution of $\sqrt{n\lambda}[\hat{f}_\lambda(x) - f(x)]$ is consistent. If $\lambda n^{1/5} \rightarrow 0$, then the simple bootstrap procedure drawing bootstrap data from $r_i - \bar{r}$ provides a consistent bootstrap distribution estimator. If $\lambda n^{1/5} \not\rightarrow 0$, e.g., $\lambda = \hat{\lambda}_{\text{CV}}$ or $\hat{\lambda}_{\text{BOOT}}(x)$, then the modified bootstrap procedures discussed earlier and

suggested by Härdle and Bowman (1988), Faraway (1990), and Hall (1990d) can be applied to construct bootstrap confidence sets for $f(x)$. Some other bootstrap procedures can be found in Section 8.5.2. Hall (1992c) suggested bootstrap-t confidence sets for $f(x)$ that have high accuracy in terms of coverage probability.

Bootstrap simultaneous confidence intervals for $f(x)$, $x \in \mathbb{R}$, can be obtained by using (8.39) with \hat{f}_λ being understood to be given by (8.41) and \hat{f}_b^* being given by (8.43), based on the b th bootstrap Monte Carlo sample. For a fixed design, we may only want simultaneous confidence intervals for $f(x)$ with x varying in $\{x_1, \dots, x_n\}$. Then, a set of improved and computationally simpler bootstrap simultaneous confidence intervals is given by

$$[\hat{f}_{\hat{\lambda}_{\text{BOOT}}}(x) - \bar{M}_{(\alpha)}, \hat{f}_{\hat{\lambda}_{\text{BOOT}}}(x) + \underline{M}_{(\alpha)}],$$

where $\bar{M}_{(\alpha)}$ and $\underline{M}_{(\alpha)}$ are $1 - \alpha$ sample quantiles of \bar{M}_b and \underline{M}_b , respectively,

$$\bar{M}_b = \max_{i \leq n} [\hat{f}_b^*(x_i) - \hat{f}_{\lambda_0}(x_i)] \quad \text{and} \quad \underline{M}_b = \min_{i \leq n} [\hat{f}_{\lambda_0}(x_i) - \hat{f}_b^*(x_i)].$$

8.5.2 Kernel estimates for random regressor

When the x_i are random, the (y_i, x_i) are i.i.d. and the function f in (8.40) is defined by

$$f(x) = E(y_i | x_i = x).$$

For a given kernel κ and bandwidth $\lambda \geq 0$, the kernel estimator of $f(x)$ is

$$\hat{f}_\lambda(x) = \sum_{i=1}^n y_i \kappa\left(\frac{x - x_i}{\lambda}\right) / \sum_{i=1}^n \kappa\left(\frac{x - x_i}{\lambda}\right). \quad (8.44)$$

Note that this estimator is slightly different from that in (8.41) and can also be used in the fixed design case. The denominator on the right-hand side of (8.44) is proportional to the kernel estimator of the common density of x_i . Estimators (8.41) and (8.44) have the same asymptotic properties, but estimator (8.44) may have better small sample performance.

The CV method introduced in Section 8.5.1 can be directly applied to bandwidth selection in this case. We shall not repeat the discussion.

Similar to the case of a fixed design, the validity of the bootstrap bandwidth selection and confidence sets rests upon the method of generating bootstrap data. Thus, we now concentrate on the discussion of how to draw bootstrap data.

At first glance, it seems natural to use the paired bootstrap (Section 7.2.2), since the x_i are random and the (y_i, x_i) are i.i.d. We have seen that

the paired bootstrap generating (y_i^*, x_i^*) from the empirical distribution of the pairs $(y_1, x_1), \dots, (y_n, x_n)$ works for linear models, nonlinear models, generalized linear models, and Cox's regression models, but it does not work for nonparametric regression models when $\lambda n^{1/5} \not\rightarrow 0$. Similar to the situations in the previous sections, a bandwidth selected by minimizing the mean average squared errors of \hat{f}_λ is usually of the order $n^{-1/5}$. If λ is of the order $n^{-1/5}$ and we define

$$\hat{f}_\lambda^*(x) = \sum_{i=1}^n y_i^* \kappa\left(\frac{x - x_i^*}{\lambda}\right) \Big/ \sum_{i=1}^n \kappa\left(\frac{x - x_i^*}{\lambda}\right),$$

then

$$E_*[\hat{f}_\lambda^*(x)] = \hat{f}_\lambda(x)$$

(Härdle, 1989), whereas the bias

$$E[\hat{f}_\lambda(x)] - f(x) \approx \frac{\lambda^2 f''(x)}{4} \int u^2 \kappa(u) du \quad (8.45)$$

is of a nonnegligible order. Therefore, the bootstrap distribution estimator based on $\hat{f}_\lambda^* - \hat{f}_\lambda$ is inconsistent. To get correct bootstrap estimators, we either have to make a bias correction as Härdle and Bowman (1988) did in the case of a fixed design or use a smoothed bootstrap as we did in Section 8.4.2, i.e., generate bootstrap data from a smoothed estimator of the joint distribution of (y_i, x_i) . However, the former approach requires estimating $f''(x)$ and the latter approach needs further development.

Härdle (1989) considered the application of the external bootstrap (Section 7.2.2). Let $r_i = y_i - \hat{f}_\lambda(x_i)$ be the i th residual, $i = 1, \dots, n$, and $\varepsilon_1^*, \dots, \varepsilon_n^*$ be generated independently and satisfy $E_*(\varepsilon_i^*) = 0$, $E_*(\varepsilon_i^{*2}) = r_i^2$, and $E_*(\varepsilon_i^{*3}) = r_i^3$, $i = 1, \dots, n$. Härdle (1989) provided an example: we can generate ε_i^* , $i = 1, \dots, n$, from a two-point distribution

$$\frac{5+\sqrt{5}}{10} \delta_{(1-\sqrt{5})r_i/2} + \frac{5-\sqrt{5}}{10} \delta_{(1+\sqrt{5})r_i/2}.$$

Define

$$\hat{f}_\lambda^*(x) = \sum_{i=1}^n y_i^* \kappa\left(\frac{x - x_i}{\lambda}\right) \Big/ \sum_{i=1}^n \kappa\left(\frac{x - x_i}{\lambda}\right), \quad (8.46)$$

where $y_i^* = \hat{f}_{\lambda_0}(x_i) + \varepsilon_i^*$, $i = 1, \dots, n$, and λ_0 is another bandwidth satisfying $\lambda_0/\lambda \rightarrow \infty$. The bootstrap distribution of $\sqrt{n\lambda}[\hat{f}_\lambda^*(x) - \hat{f}_{\lambda_0}(x)]$ can then be used to estimate the sampling distribution of $\sqrt{n\lambda}[\hat{f}_\lambda(x) - f(x)]$. The reason why we take a different bandwidth λ_0 in defining y_i^* is that

$$E_*[\hat{f}_\lambda^*(x)] - \hat{f}_{\lambda_0}(x) \approx \frac{\lambda^2 \hat{f}_{\lambda_0}''(x)}{4} \int u^2 \kappa(u) du$$

and, in view of (8.45), $\hat{f}_\lambda^*(x) - \hat{f}_{\lambda_0}(x)$ has a similar behavior to $\hat{f}_\lambda(x) - f(x)$ if $\hat{f}_{\lambda_0}''(x) \rightarrow f''(x)$, which requires that λ_0 tend to 0 slower than λ .

Under the conditions that (1) f'' exists, (2) $\int |\kappa(x)|^{2+\epsilon} dx < \infty$ for some $\epsilon > 0$, (3) λ is of the order $n^{-1/5}$, (4) the density of x_i is positive at x and is twice differentiable, and (5) $\text{var}(y_i|x_i = x)$ and $E(|y_i|^{2+\epsilon}|x_i = x)$ are continuous at x , Härdle (1990) showed that

$$\sup_t |P_*\{\sqrt{n\lambda}[\hat{f}_\lambda^*(x) - \hat{f}_{\lambda_0}(x)] \leq t\} - H_{y|X}(t)| \rightarrow_{a.s.} 0, \quad (8.47)$$

where $H_{y|X}$ is the conditional distribution of $\sqrt{n\lambda}[\hat{f}_\lambda(x) - f(x)]$, given x_1, \dots, x_n . Cao-Abad (1991) showed that the convergence rate of the left-hand side of (8.47) is $O_p(n^{-2/9})$ if λ_0 is of the order $n^{-1/9}$, under some further conditions that (6) $f(x)$ is four times continuously differentiable, (7) κ is nonnegative and satisfies $\int x^2 \kappa(x) dx < \infty$ and $\int [\kappa(x)]^3 dx < \infty$, (8) $\sup_x E[|y_i - f(x_i)|^3|x_i = x] < \infty$, and (9) the density of x_i , $p(x)$, is four times continuously differentiable in its support D and $\inf_{x \in D} p(x) > 0$.

Result (8.47) ensures the consistency of the bootstrap confidence sets constructed by using the external bootstrap. Härdle (1990) provided an algorithm for the construction of hybrid bootstrap confidence sets for $f(x)$ with a fixed x . Härdle and Marron (1991) applied the external bootstrap in constructing simultaneous confidence intervals for $f(x)$, $x \in \{x_1, \dots, x_n\}$. These methods can also be used in the fixed design case.

Note that the conditional distribution of y_i , given $x_i = x$, can be estimated by

$$\hat{F}_\lambda(y|x) = \sum_{i=1}^n \kappa\left(\frac{x - x_i}{\lambda}\right) I\{y_i \leq y\} \Big/ \sum_{i=1}^n \kappa\left(\frac{x - x_i}{\lambda}\right).$$

Tu (1988a) proposed generating bootstrap data y_1^*, \dots, y_n^* independently from $\hat{F}_\lambda(\cdot|x)$. Let \hat{f}_λ^* be given by (8.46). Then the bootstrap distribution of $\sqrt{n\lambda}[\hat{f}_\lambda^*(x) - \hat{f}_\lambda(x)]$ is a consistent estimator of the sampling distribution of $\sqrt{n\lambda}[\hat{f}_\lambda(x) - f(x)]$, provided that $\lambda n^{1/5} \rightarrow 0$. In this approach, we can also generate y_i^* from $\hat{F}_\lambda(\cdot|x_i)$ (Rutherford and Yakowitz, 1991), $i = 1, \dots, n$, and/or generate bootstrap data from $\hat{F}_{\lambda_0}(\cdot|\cdot)$ with a λ_0 satisfying $\lambda_0/\lambda \rightarrow \infty$ so that we can select a bandwidth and use it for constructing bootstrap confidence sets.

8.5.3 Nearest neighbor estimates

Consider model (8.40) with fixed or random x_i ($x_i \in [0, 1]$ is assumed when x_i is nonrandom). A λ -nearest neighbor estimator of $f(x_i)$ depends only on the λ observations whose x values are closest to x_i , $\lambda = 2, \dots, n$. Let

$x_{i(j)}$ be the j th nearest neighbor of x_i according to the distance $|x_i - x_j|$, $j \neq i$. For a given weight function $w_{n,\lambda}(\cdot)$, the λ -nearest neighbor estimator of $f(x_i)$ is

$$\hat{f}_\lambda(x_i) = \sum_{j=1}^{\lambda} w_{n,\lambda}(j)y_{i(j)},$$

where $y_{i(j)}$ is the response observed at $x_{i(j)}$. An example of the weight function $w_{n,\lambda}$ is the uniform weight $w_{n,\lambda}(i) = \lambda^{-1}$. Let $y = (y_1, \dots, y_n)'$. Then $(\hat{f}_\lambda(x_1), \dots, \hat{f}_\lambda(x_n))'$ is of the form $M_\lambda y$ with each row of M_λ being some permutation of the vector $(w_{n,\lambda}(1), \dots, w_{n,\lambda}(\lambda), 0, \dots, 0)$.

Asymptotic properties of the nearest neighbor estimators are given, for example, in Stone (1977). The first decision we have to make is the number of neighbors that should enter into our estimate, i.e., the determination of the smoothing parameter λ that plays a similar role to the bandwidth for the kernel method discussed in the previous sections. Usually, λ is selected to minimize the average squared errors or the loss

$$L(\hat{f}_\lambda) = \frac{1}{n} \sum_{i=1}^n [f(x_i) - \hat{f}_\lambda(x_i)]^2.$$

A difference between the selection of λ in this problem and those in the previous sections is that λ in this problem takes a value in $\{2, \dots, n\}$, whereas previously λ has a continuous domain.

The CV method can be applied to the selection of λ . Let

$$\hat{f}_{\lambda,i}(x_i) = \sum_{j=1}^n w_{n,\lambda}(j)y_{i(j+1)} \quad (8.48)$$

be the λ -nearest neighbor estimator of $f(x_i)$ after deleting y_i , $i = 1, \dots, n$. Note that the weights in (8.48) are unchanged after deleting y_i . Let \tilde{M}_λ be the $n \times n$ matrix with 0 diagonals and the $(i, i(j))$ th entry ($j \neq 1$) being $w_{n,\lambda}(j-1)$. Then $(\hat{f}_{\lambda,1}(x_1), \dots, \hat{f}_{\lambda,n}(x_n))' = \tilde{M}_\lambda y$ and the CV method selects a bandwidth $\hat{\lambda}_{\text{cv}}$ that minimizes

$$\frac{\|y - \tilde{M}_\lambda y\|^2}{n} \quad (8.49)$$

over $\lambda = 2, \dots, n$. A modified CV procedure, called the generalized cross-validation (GCV) (Craven and Wahba, 1979), selects a bandwidth $\hat{\lambda}_{\text{gcv}}$ that minimizes

$$\frac{\|y - M_\lambda y\|^2}{n[1 - \text{tr}(M_\lambda)/n]^2}. \quad (8.50)$$

From the definition of M_λ , $\text{tr}(M_\lambda)/n = w_{n,\lambda}(1)$. Thus, $\hat{\lambda}_{\text{gcv}}$ minimizes $\|y - M_\lambda y\|^2/[1 - w_{n,\lambda}(1)]^2$ over λ .

Li (1987) showed that the CV and GCV are consistent in the sense that

$$L(\hat{f}_\lambda)/ \min_{\lambda=2,\dots,n} L(\hat{f}_\lambda) \rightarrow_p 1$$

for $\hat{\lambda} = \hat{\lambda}_{\text{CV}}$ or $\hat{\lambda}_{\text{GCV}}$, under the following conditions: (1) $w_{n,\lambda}(1) \leq c < 1$ for any $n, \lambda \geq 2$; (2) $\sum_{i=1}^\lambda w_{n,\lambda}(i) = 1$; (3) for any n, λ , and i , $w_{n,\lambda}(i) \geq w_{n,\lambda}(i+1) \geq 0$; (4) there exist positive constants c_1 and c_2 such that $w_{n,\lambda}(1) \leq c_1 \lambda^{-(1/2+c_2)}$ for any $n, \lambda \geq 2$; (5) there is a positive integer p such that $E|\varepsilon_1|^{4p} < \infty$ and $\lim_{n \rightarrow \infty} \{\inf_\lambda E[L(\hat{f}_\lambda)] n^{1-1/p}\} = \infty$; and (6) $\inf_\lambda L(\hat{f}_\lambda) \rightarrow_p 0$. These conditions can be easily satisfied by most commonly used weights and nearest neighbor estimators.

When the x_i are random, the following smoothed nearest neighbor estimator of the whole curve $f(x)$ can be used:

$$\hat{f}_\lambda(x) = \sum_{i=1}^n w_{n,\lambda}(x, x_i) y_i,$$

where

$$w_{n,\lambda}(x, z) = \kappa\left(\frac{F_n(x) - F_n(z)}{\lambda}\right) / \sum_{i=1}^n \kappa\left(\frac{F_n(x) - F_n(x_i)}{\lambda}\right)$$

is a weight function, κ is a kernel, $\lambda \geq 0$ is a bandwidth (smoothing parameter), and F_n is the empirical distribution of x_1, \dots, x_n .

The CV and GCV can still be used to select a bandwidth. For the GCV, we still minimize the quantity in (8.50) with M_λ being the matrix whose (i, j) th element is $w_{n,\lambda}(x_i, x_j)$. For the CV, we minimize (8.49) with \tilde{M}_λ being the matrix whose diagonal elements are 0 and the (i, j) th off-diagonal element is

$$\kappa\left(\frac{F_{n-1,i}(x_i) - F_{n-1,i}(x_j)}{\lambda}\right) / \sum_{l \neq i} \kappa\left(\frac{F_{n-1,i}(x_i) - F_{n-1,i}(x_l)}{\lambda}\right).$$

The results for the CV and GCV derived for the unsmoothed nearest neighbor estimators are expected to carry over.

Under appropriate conditions (Stute, 1984),

$$\sqrt{n\lambda}[\hat{f}_\lambda(x) - f(x)] \rightarrow_d N(b(x), \sigma^2(x)),$$

where

$$b(x) = \sqrt{r} 2(f \circ F^{-1})''(F(x)) \int u^2 \kappa(u) du,$$

$r = \lim_{n \rightarrow \infty} \lambda^5 n$, F is the common distribution function of x_i , and

$$\sigma^2(x) = \text{var}(y_i | x_i = x) \int [\kappa(u)]^2 du.$$

If $r = 0$, then one can construct a confidence set for $f(x)$ with a fixed x by using a normal approximation and estimating $\sigma^2(x)$ by

$$\hat{\sigma}^2(x) = \sum_{i=1}^n w_{n,\lambda}(x, x_i)[y_i - \hat{f}_\lambda(x)]^2 \int [\kappa(u)]^2 du.$$

If $r > 0$, then one has to estimate the bias function $b(x)$.

Dikta (1990) studied the distribution estimators and confidence sets based on the paired bootstrap. Let $(y_1^*, x_1^*), \dots, (y_n^*, x_n^*)$ be i.i.d. from the empirical distribution of the pairs $(y_1, x_1), \dots, (y_n, x_n)$. Then the bootstrap analog of \hat{f}_λ is

$$\hat{f}_\lambda^*(x) = \sum_{i=1}^n w_{n,\lambda}^*(x, x_i^*) y_i^*,$$

where

$$w_{n,\lambda}^*(x, x_i^*) = \kappa\left(\frac{F_n^*(x) - F_n^*(x_i^*)}{\lambda}\right) \Bigg/ \sum_{i=1}^n \kappa\left(\frac{F_n^*(x) - F_n^*(x_i^*)}{\lambda}\right)$$

is the bootstrap analog of $w_{n,\lambda}(x, x_i)$ and F_n^* is the empirical distribution of x_1^*, \dots, x_n^* . Assume that F is continuous; $E(y_i^4) < \infty$; $f \circ F^{-1}$ is twice continuously differentiable in an open neighborhood of $F(x)$; and κ is a symmetric, continuously differentiable, and strictly decreasing probability density with a compact support. If the bandwidth λ is chosen so that $\lambda n^{1/5} \rightarrow 0$, $n\lambda^4/\log(n) \rightarrow \infty$, $\sum_{n=1}^\infty \lambda^{-4} n^{-2} < \infty$, and $\sum_{n=1}^\infty [n^{-1} \lambda^{-1} \log(n)]^{3/2} < \infty$, then the bootstrap distribution of $\sqrt{n\lambda}[\hat{f}_\lambda^*(x) - \hat{f}_\lambda(x)]$ is strongly consistent for the distribution of $\sqrt{n\lambda}[\hat{f}_\lambda(x) - f(x)]$ in terms of the distance generated by $\|\cdot\|_\infty$.

Similar to the situation where f is estimated by a kernel estimator, the optimal bandwidth is usually of the order $n^{-1/5}$. If $\lambda n^{1/5} \not\rightarrow 0$, then this bootstrap procedure produces an inconsistent distribution estimator and is also not appropriate for bandwidth selection. These results are parallel to those for the case of kernel estimation (Section 8.5.2).

To obtain correct bootstrap estimators when λ is of the order $n^{-1/5}$, Dikta (1990) proposed estimating the bias $b(x)$ by $\hat{b}(x)$ and then using the bootstrap distribution of $\sqrt{n\lambda}[\hat{f}_\lambda^*(x) - \hat{f}_\lambda(x)] - \hat{b}(x)$. The methods suggested by Härdle (1989) and Hall (1990d) for kernel estimation may be applied here to solve the problem. Further theoretical investigation is needed.

Dikta (1990) compared by simulation the performances of the 90% hybrid bootstrap (HB) confidence interval and the 90% normal approximation (NOR) confidence interval. The following two models were considered:

Table 8.2. Empirical coverage probabilities (CP) and expected lengths (EL) of the 90% NOR and HB confidence intervals for $f(x)$
 [Adapted from Dikta (1990), by permission of Academic Press]

x	Model 1				Model 2			
	NOR		HB		NOR		HB	
	CP	EL	CP	EL	CP	EL	CP	EL
-0.7	0.14	0.42	0.57	0.52	0.00	0.38	0.31	0.51
-0.4	0.33	0.91	0.58	0.90	0.35	0.84	0.65	0.85
-0.2	0.82	1.34	0.85	1.28	0.85	1.34	0.87	1.24
0.0	0.97	1.80	0.91	1.70	0.97	1.87	0.93	1.78
0.2	0.99	2.26	0.86	2.00	0.97	2.32	0.90	2.05
0.4	0.99	2.62	0.90	1.99	0.99	2.66	0.90	2.05
0.7	0.32	2.48	0.32	2.06	0.48	2.53	0.50	2.07

Model 1: $y_i = 5x_i^2 + 7x_i + e_i$; x_i and e_i are independent and have uniform distributions on $[-1, 1]$ and $[-\frac{1}{2}, \frac{1}{2}]$, respectively.

Model 2: $y_i = 5x_i^2 + 7x_i + (x_i + 1)e_i$; x_i and e_i are the same as those in model 1.

Under both models, $f(x) = 5x^2 + 7x$, while $\text{var}(y_i|x_i = x) = \frac{1}{12}$ under model 1 and $\text{var}(y_i|x_i = x) = \frac{1}{12}(x+1)^2$ under model 2.

In the simulation, $\kappa(x) = \frac{3}{4}(1-x)^2 I\{|x| \leq 1\}$; $n = 50$; $\lambda = [n \log(n)]^{-1/5} = 0.35$; and the bootstrap percentiles are approximated by Monte Carlo with size 1000. The simulation size is 100.

The results in Table 8.2 indicate that the performance of the normal approximation confidence interval is poor and the bootstrap confidence interval is much better in terms of both coverage probability and the expected length. When x is near the boundary, both methods do not perform well, due to boundary effects.

8.5.4 Smoothing splines

Consider model (8.40) with nonrandom $x_i \in [0, 1]$. Suppose that f is a member of $W_2^k[0, 1]$, the collection of all functions f on $[0, 1]$ that has absolutely continuous derivatives f' , f'' , \dots , $f^{(k-1)}$ and $\int_0^1 [f^{(k)}(x)]^2 dx < \infty$. The smoothing spline estimator \hat{f}_λ of f is the minimizer of

$$\frac{1}{n} \sum_{i=1}^n [y_i - f(x_i)]^2 + \lambda \int_0^1 [f^{(k)}(x)]^2 dx \quad (8.51)$$

over $f \in W_2^k[0, 1]$, where $\lambda \geq 0$ is the bandwidth or smoothing parameter that controls the trade-off between the infidelity to the data as measured by the first term in (8.51) and the smoothness $\int_0^1 [f^{(k)}(x)]^2 dx$ of the estimated solution. If $\lambda = \infty$, then \hat{f}_λ is the least squares estimator. The optimal bandwidth is usually defined to be the minimizer of the average squared errors or the loss

$$L(\hat{f}_\lambda) = \frac{1}{n} \sum_{i=1}^n [f(x_i) - \hat{f}_\lambda(x_i)]^2.$$

Since $L(\hat{f}_\lambda)$ is unknown, the bandwidth has to be selected based on the data.

The GCV described in Section 8.5.3 has become the most popular method for bandwidth selection since it was proposed by Craven and Wahba (1979). Let $y = (y_1, \dots, y_n)'$ and M_λ be the $n \times n$ matrix satisfying

$$(\hat{f}_\lambda(x_1), \dots, \hat{f}_\lambda(x_n))' = M_\lambda y.$$

Then the GCV selects a bandwidth $\hat{\lambda}_{\text{GCV}}$ that minimizes

$$\frac{\|y - M_\lambda y\|^2}{n[1 - \text{tr}(M_\lambda)/n]^2}.$$

This is the same as the GCV bandwidth selection for the nearest neighbor estimator [see (8.50)], except that the minimization is over $\lambda \in [0, \infty)$ for smoothing splines.

Properties of $\hat{\lambda}_{\text{GCV}}$ have been studied by many researchers (see the references provided by Wahba, 1985). Li (1986) proved the following strongest result:

$$L(\hat{f}_{\hat{\lambda}_{\text{GCV}}}) / \inf_{\lambda} L(\hat{f}_\lambda) \rightarrow_p 1,$$

provided that f is not a polynomial of order $k-1$ or less. Weaker properties of $\hat{\lambda}_{\text{GCV}}$, such as

$$E[L(\hat{f}_{\hat{\lambda}_{\text{GCV}}})] / \inf_{\lambda} E[L(\hat{f}_\lambda)] \rightarrow 1,$$

have been established (Craven and Wahba, 1979; Wahba, 1985) even for f being a polynomial of order $k-1$ or less.

There are other methods for bandwidth selection. For example, when the ε_i are normal, the bandwidth selected by using a generalized maximum likelihood (GML) method, denoted by $\hat{\lambda}_{\text{GML}}$, is the minimizer of

$$\frac{\|y - M_\lambda y\|^2}{q_\lambda^{1/(n-k)}},$$

where q_λ is the product of the $n-k$ nonzero eigenvalues of the matrix $I_n - M_\lambda$.

Wahba (1985) compared asymptotic properties of $\hat{\lambda}_{\text{GML}}$ and $\hat{\lambda}_{\text{GCV}}$. Some empirical comparisons were also made. The results indicate that the GCV method is preferred. We described here the simulation results in Wahba (1985). Table 8.3 shows the estimates, based on 10 simulations, of $\gamma_{\text{GML}} = L(\hat{f}_{\hat{\lambda}_{\text{GML}}})/\inf_{\lambda} L(\hat{f}_{\lambda})$, $\gamma_{\text{GCV}} = L(\hat{f}_{\hat{\lambda}_{\text{GCV}}})/\inf_{\lambda} L(\hat{f}_{\lambda})$, and ν = the number of times that $\gamma_{\text{GML}} < \gamma_{\text{GCV}}$ in 10 simulations (a tie is counted as 0.5). The model under consideration was

$$y_i = f\left(\frac{i}{n}\right) + \varepsilon_i, \quad i = 1, \dots, n,$$

with $\varepsilon_i \sim N(0, \sigma^2)$ and three different f given by

$$(i) \quad f_I(x) = \frac{1}{3}\beta_{10,5}(x) + \frac{1}{3}\beta_{7,7}(x) + \frac{1}{3}\beta_{5,10}(x),$$

$$(ii) \quad f_{II}(x) = \frac{6}{10}\beta_{30,17}(x) + \frac{4}{10}\beta_{3,11}(x),$$

and

$$(iii) \quad f_{III}(x) = \frac{1}{3}\beta_{20,5}(x) + \frac{1}{3}\beta_{12,12}(x) + \frac{1}{3}\beta_{7,30}(x),$$

where $\beta_{p,q}(x)$ is the Beta function.

A more comprehensive survey on the GCV and smoothing spline can be found in Wahba (1990).

Table 8.3. Empirical comparison of the GML and GCV [Adapted from Wahba (1985), by permission of Institute of Mathematical Statistics]

n	σ	f_I			f_{II}			f_{III}		
		γ_{GML}	γ_{GCV}	ν	γ_{GML}	γ_{GCV}	ν	γ_{GML}	γ_{GCV}	ν
32	0.0125	1.49	1.41	1	1.38	1.24	3	1.50	1.45	4
	0.025	1.32	1.33	3	1.83	1.27	1	1.51	1.13	0
	0.05	1.22	1.94	8	1.41	1.09	1.5	1.62	1.43	1
	0.1	1.25	1.20	2	1.23	1.07	2.5	1.17	1.11	3
	0.2	1.39	1.40	5	1.05	1.07	6	1.12	2.02	5.5
64	0.0125	1.40	1.07	2	2.09	1.31	1	1.94	1.10	0
	0.025	1.40	1.09	0	1.49	1.33	2	1.51	1.06	0
	0.05	1.23	1.05	0	1.43	1.16	1	1.30	1.14	1
	0.1	1.48	1.83	5.5	1.24	1.05	2	1.20	1.21	1
	0.2	1.22	1.32	8	1.18	1.06	2	1.12	1.45	4
128	0.0125	1.67	1.06	1	1.75	1.03	0	1.69	1.09	1
	0.025	1.39	1.09	1	1.59	1.07	0	1.34	1.07	0
	0.05	1.29	1.06	2	1.38	1.03	0	1.28	1.07	2
	0.1	1.32	1.16	2	1.26	1.06	1	1.20	1.04	0
	0.2	1.07	1.50	3	1.30	1.18	0	1.23	1.19	3

8.6 Multivariate Analysis

Statistical procedures in classical multivariate analysis are based on the normality assumption on the distribution of the sample and may be sensitive to slight violations of the normality assumption. Even under the normality assumption, an exact and tractable form of the distribution of the statistic of interest may not be available and asymptotic approximations are required. The bootstrap provides a nonparametric alternative for approximating the distributions of statistics in multivariate analysis under almost no distributional assumption. The cross-validation and the bootstrap can also be conveniently applied to get nonparametric misclassification rate estimators in discriminant analysis.

Many results previously discussed can be immediately extended to the multivariate case. In fact, in some of the previous discussions, the data are considered to be multivariate. For example, in Chapter 4 we discussed the construction of confidence sets for functions of multivariate sample means; the bootstrap test given in Example 4.8 is a bootstrap version of Hotelling's T^2 test for a multivariate mean. In this section, we present some topics that are unique in the field of multivariate analysis and have not been discussed previously.

8.6.1 Analysis of covariance matrix

Let X_1, \dots, X_n be i.i.d. random p -vectors from a distribution F with unknown mean μ and covariance matrix Σ . The analysis of the covariance matrix is an important component in multivariate analysis. The sample covariance matrix is defined as

$$\mathbf{S}_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)(X_i - \bar{X}_n)',$$

where $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ is the sample mean. Many statistics in multivariate analysis can be expressed as functions of \mathbf{S}_n . For example, the sample correlation coefficient of the i th and j th components of X_1 is $s_{ij}/\sqrt{s_{ii}s_{jj}}$, where s_{ij} is the (i, j) th element of \mathbf{S}_n ; the eigenvalues and eigenvectors of \mathbf{S}_n are important statistics in principal component analysis.

Nagao (1985, 1988) and Romanazzi (1993) discussed applications of the jackknife to the eigenvalues and eigenvectors of \mathbf{S}_n . The details are omitted since they are similar to those discussed in Chapter 2.

Beran and Srivastava (1985) studied the bootstrap estimation of H_n , the distribution of $\sqrt{n}(\mathbf{S}_n - \Sigma)$. The bootstrap procedure is a straightforward extension of the one described for the univariate case (Chapter 1). Let X_1^*, \dots, X_n^* be i.i.d. random vectors from the empirical distribution F_n of

$X_1, \dots, X_n, \bar{X}_n^* = n^{-1} \sum_{i=1}^n X_i^*$, and

$$\mathbf{S}_n^* = \frac{1}{n-1} \sum_{i=1}^n (X_i^* - \bar{X}_n^*)(X_i^* - \bar{X}_n^*)'.$$

Given X_1, \dots, X_n , the conditional distribution of $\sqrt{n}(\mathbf{S}_n^* - \hat{\Sigma}_n)$, denoted by H_{BOOT} , is the bootstrap estimator of H_n , where $\hat{\Sigma}_n = E_* \mathbf{S}_n^* = \frac{n-1}{n} \mathbf{S}_n$ (which can be replaced by \mathbf{S}_n).

If F has a finite fourth moment, then

$$\|H_{\text{BOOT}} - H_n\|_\infty \rightarrow_{a.s.} 0, \quad (8.52)$$

i.e., the bootstrap estimator is consistent. Result (8.52) can be proved using the same arguments for proving the consistency of the bootstrap distribution estimators introduced in Chapter 3.

An immediate application of result (8.52) is the construction of bootstrap confidence regions for functions of Σ and bootstrap tests for Σ . We first consider bootstrap confidence regions. In what follows, we use $g(A)$ to denote a k -vector-valued function of the elements in the diagonal and the upper triangle of a symmetric matrix A . Thus, g is from $\mathbb{R}^{p(p+1)/2}$ to \mathbb{R}^k . A bootstrap confidence region for $g(\Sigma)$ can be constructed as follows. Let ψ be a continuous function from \mathbb{R}^k to \mathbb{R} and let J_n be the distribution function of $\psi(\sqrt{n}[g(\mathbf{S}_n) - g(\Sigma)])$. For $\alpha \in (0, 1)$, let

$$c_{n1}(F) = \inf\{t : J_n(t) \geq 1 - \alpha\}, \quad c_{n2}(F) = \sup\{t : J_n(t) \leq 1 - \alpha\},$$

and $c_n(F_n)$ be a random variable satisfying

$$c_{n1}(F_n) \leq c_n(F_n) \leq c_{n2}(F_n).$$

Then a bootstrap confidence region for $g(\Sigma)$ with level $1 - \alpha$ is

$$\mathcal{R} = \{x \in \mathbb{R}^k : \psi(\sqrt{n}[g(\mathbf{S}_n) - x]) \leq c_n(F_n)\}.$$

If $\psi(ax) = a\psi(x)$ for any $x \in \mathbb{R}^k$ and positive constant a , then

$$\mathcal{R} = \{x \in \mathbb{R}^k : \psi(g(\mathbf{S}_n) - x)) \leq c_\alpha\},$$

where c_α is the $1 - \alpha$ quantile of the distribution of $\psi(g(\mathbf{S}_n^*) - g(\hat{\Sigma}_n))$, conditioned on X_1, \dots, X_n . c_α can be calculated by Monte Carlo.

By (8.52), this bootstrap confidence region is asymptotically correct, i.e.,

$$P\{g(\Sigma) \in \mathcal{R}\} \rightarrow 1 - \alpha,$$

provided that g is continuously differentiable at Σ and $\nabla g(\Sigma) \neq 0$.

As an example, consider the case where Σ has p different eigenvalues $\lambda_1(\Sigma) > \dots > \lambda_p(\Sigma) > 0$. Let

$$g(\Sigma) = (\log(\lambda_1(\Sigma)), \dots, \log(\lambda_p(\Sigma)))'.$$

The reason we take the logarithm is to stabilize the variance. It can be shown that $g(\Sigma)$ is continuously differentiable. Therefore, we can apply the bootstrap procedure described above. An example of the function ψ is $\psi(x_1, \dots, x_p) = \max_{j \leq p} |x_j|$.

We now consider bootstrap tests. Consider the null hypothesis

$$H_0 : \Sigma = \tau(\Sigma),$$

where τ is a continuous, nonidentity function from $\mathbb{R}^{p(p+1)/2}$ to $\mathbb{R}^{p(p+1)/2}$ (an example is given later). Suppose that a test rule is given as follows: H_0 is rejected if $T_n = ng(\mathbf{S}_n)$ is large, where g is a real-valued function twice continuously differentiable at Σ with $g(\Sigma) = 0$ and $\nabla g(\Sigma) = 0$ whenever $\Sigma = \tau(\Sigma)$ holds. Define

$$V_n(F) = [\tau(\Sigma)]^{1/2} \Sigma^{-1/2} \mathbf{S}_n \Sigma^{-1/2} [\tau(\Sigma)]^{1/2},$$

which is a functional of F having covariance matrix Σ . Under the null hypothesis H_0 , the distribution of T_n can be approximated by the conditional distribution of $ng(V_n^*)$, where

$$V_n^* = V_n(F_n) = [\tau(\hat{\Sigma}_n)]^{1/2} \hat{\Sigma}_n^{-1/2} \mathbf{S}_n^* \hat{\Sigma}_n^{-1/2} [\tau(\hat{\Sigma}_n)]^{1/2}.$$

Then a bootstrap test of level α rejects H_0 if

$$T_n > d_\alpha,$$

where d_α is the $1 - \alpha$ quantile of the conditional distribution of $ng(V_n^*)$, given X_1, \dots, X_n . d_α can be approximated by Monte Carlo. It can be shown that

$$P\{T_n > d_\alpha\} \rightarrow \alpha$$

if H_0 holds.

As an example, consider the null hypothesis

$$H_0 : \Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix},$$

where Σ_1 is a diagonal matrix of order r and Σ_2 is an arbitrary covariance matrix of order $p - r$. The function τ can be specified by

$$\tau \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} \text{diag}(\Sigma_{11}) & 0 \\ 0 & \Sigma_{22} \end{pmatrix},$$

where $\text{diag}(A)$ is the diagonal matrix containing the diagonal elements of A . Under the normality assumption on F , the likelihood ratio test rejects H_0 if

$$g(\mathbf{S}_n) = \log(|\tau(\mathbf{S}_n)|) - \log(|\mathbf{S}_n|)$$

is large, where $|A|$ is the determinant of A . The function g is twice continuously differentiable at $\tau(\Sigma)$ with both g and ∇g vanishing at $\tau(\Sigma)$. Thus, we can apply the bootstrap test procedure described above.

Some more examples can be found in Beran and Srivastava (1985). Chan and Srivastava (1988) and Nagao and Srivastava (1992) used similar methods to find the powers of the sphericity tests under general and local alternatives, respectively. Alemayehu (1988) applied Beran's prepivoting method (Section 4.3.1) to the problem of constructing simultaneous bootstrap confidence sets for parameters that are latent roots of some random matrices, including canonical correlation coefficients, measures of association between vectors, and principal components. Zhang, Pantula and Boos (1991) suggested a pooled bootstrap to modify the bootstrap test introduced in this section and showed by simulation that the modified bootstrap test has a better performance. Zhang and Boos (1992, 1993) proposed some bootstrap tests for the hypotheses about covariance matrices.

8.6.2 Multivariate linear models

A multivariate linear model is an extension of model (7.1) to multivariate responses; that is,

$$y_i = x'_i B + \varepsilon_i, \quad i = 1, \dots, n,$$

where y'_i is a q -vector of responses, x_i is a p -vector of explanatory variables, B is a $p \times q$ matrix of unknown parameters, and $\varepsilon'_1, \dots, \varepsilon'_n$ are i.i.d. random q -vectors with mean 0 and an unknown covariance matrix Σ . For simplicity, we assume that $X' = (x_1, \dots, x_n)$ is of full rank. The parameter matrix B can still be estimated by the least squares estimator

$$\hat{B} = (X' X)^{-1} X' Y,$$

where $Y = (y'_1, \dots, y'_n)'$. The covariance matrix Σ can be estimated by

$$\hat{\Sigma} = \frac{Y' Y - Y' X (X' X)^{-1} X' Y}{n}.$$

Let C and D be two given matrices of orders $r \times p$ and $q \times k$, respectively. In multivariate analysis, we are often interested in constructing simultaneous confidence intervals for linear combinations of CBD and tests for hypotheses related to CBD .

Consider first the construction of simultaneous confidence intervals for $aCDBb'$, $(a, b) \in \mathbb{C} \subset \mathbb{R}^r \times \mathbb{R}^k$. Define

$$T_{n,a,b} = \frac{(aC\hat{B}Db' - aCDBb')^2}{naC(X'X)^{-1}C'a'b\hat{\Sigma}b'} \quad \text{and} \quad T_n = \sup_{(a,b) \in \mathbb{C}} T_{n,a,b}.$$

When $\mathbb{C} = \mathbb{R}^r \times \mathbb{R}^k$, T_n is the largest eigenvalue of the matrix

$$nD'(Y - XB)'H_C(Y - XB)D\hat{\Sigma}^{-1}, \quad (8.53)$$

where $H_C = X(X'X)^{-1}C'[C(X'X)^{-1}C']^{-1}C(X'X)^{-1}X'$. If G_n , the distribution of T_n , is known, then simultaneous confidence intervals of level $1 - \alpha$ are given by

$$aCDBb' : T_{n,a,b} \leq G_n^{-1}(1 - \alpha), \quad (a, b) \in \mathbb{C}.$$

The bootstrap can be applied to approximate $G_n^{-1}(1 - \alpha)$ when G_n is unknown. We consider the bootstrap based on residuals introduced in Section 7.2.2. Let F_n be the empirical distribution of $y_i - x_i'\hat{B}$, $i = 1, \dots, n$, recentered so that F_n has mean 0. Let $\varepsilon_1^*, \dots, \varepsilon_n^*$ be i.i.d. from F_n , $y_i^* = x_i'\hat{B} + \varepsilon_i^*$, and \hat{B}^* and $\hat{\Sigma}^*$ be the bootstrap analogs of \hat{B} and $\hat{\Sigma}$, respectively. Then the bootstrap-t simultaneous confidence intervals of level $1 - \alpha$ are given by

$$aCDBb' : T_{n,a,b} \leq G_{\text{BOOT}}^{-1}(1 - \alpha), \quad (a, b) \in \mathbb{C},$$

where G_{BOOT} is the conditional distribution of

$$T_n^* = \sup_{(a,b) \in \mathbb{C}} \frac{(aC\hat{B}^*Db' - aC\hat{B}Db')^2}{naC(X'X)^{-1}C'a'b\hat{\Sigma}^*b'}.$$

Alemayeha (1987) pointed out that

$$P\{aCDBb' : T_{n,a,b} \leq G_{\text{BOOT}}^{-1}(1 - \alpha) \text{ for all } (a, b) \in \mathbb{C}\} \rightarrow 1 - \alpha,$$

provided that $\max_{i \leq n} x_i'(X'X)^{-1}x_i \rightarrow 0$ and $X'X/n \rightarrow V > 0$. He also used Beran's prepivoting method to construct balanced simultaneous confidence sets for $aCDBb'$, $(a, b) \in \mathbb{C}$.

We next consider bootstrap tests for hypotheses of the form

$$H_0 : CBD = M,$$

where M is a given $r \times k$ matrix. There are many test statistics derived under the normality assumption on F , and most of them are functions of the eigenvalues of the matrix given in (8.53). However, almost all of these tests require special tables for their critical values. The bootstrap can be applied to approximate the critical value for a given test statistic T_n . Let

\hat{B}_0 be an estimator of B under the null hypothesis H_0 , e.g., the restricted least squares estimator of B under the restriction $CBD = M$. Define $y_{i0}^* = x_i' \hat{B}_0 + \varepsilon_i^*$, where the ε_i^* are bootstrap data i.i.d. from F_n . Then a bootstrap test rejects H_0 if

$$T_n \geq c_\alpha,$$

where c_α is the $1 - \alpha$ quantile of the conditional distribution of the bootstrap analog T_n^* based on y_{i0}^* and x_i , $i = 1, \dots, n$. Alemayeha (1987) argued that the level of this bootstrap test is approximately α .

Finally, we present some simulation results for a one-way MANOVA model:

$$y_{ij} = \mu + \beta_j + \varepsilon_{ij}, \quad j = 1, \dots, 4, \quad i = 1, \dots, 10,$$

where $\beta_j \in \mathbb{R}^4$ and four-dimensional disturbance terms ε_{ij} are from either $N(0, I_4)$ or $CN = 0.95N(0, I_4) + 0.05N(0, 10I_4)$. Table 8.4 gives the empirical coverage probabilities of the bootstrap simultaneous confidence intervals for pairwise contrasts $\beta_i - \beta_j$. Table 8.5 shows the empirical levels of the bootstrap tests for H_0 : the second component of β_2 is 0. All results are based on 1000 simulations and the bootstrap quantiles are approximated by Monte Carlo with size 200.

Table 8.4. Empirical coverage probabilities of the bootstrap simultaneous confidence intervals [Adapted from Alemayeha (1987), by permission of American Statistical Association]

F	$1 - \alpha = 0.90$			$1 - \alpha = 0.95$		
	$n = 20$	$n = 30$	$n = 60$	$n = 20$	$n = 30$	$n = 60$
$N(0, I_4)$	0.885	0.889	0.899	0.944	0.945	0.947
CN	0.885	0.889	0.899	0.944	0.945	0.947

Table 8.5. Empirical levels of the bootstrap tests [Adapted from Alemayeha (1987), by permission of American Statistical Association]

F	$\alpha = 0.10$				$\alpha = 0.05$			
	T_{n1}	T_{n2}	T_{n3}	T_{n4}	T_{n1}	T_{n2}	T_{n3}	T_{n4}
$N(0, I_4)$	0.099	0.099	0.099	0.099	0.048	0.045	0.053	0.051
CN	0.099	0.091	0.098	0.099	0.047	0.051	0.049	0.051

T_{n1} : Wilks' test; T_{n2} : Lawley-Hotelling's test; T_{n3} : Roy's largest test;
 T_{n4} : Pillai's trace test.

8.6.3 Discriminant analysis

The primary aim in discriminant analysis is to assign an individual to one of $k \geq 2$ distinct groups on the basis of measurements on some characteristics of the individual. This problem is sometimes referred to as classification or pattern recognition. An allocation or classification rule is constructed based on a training sample $\{X_1, \dots, X_n\}$ for which the group membership of each observation X_i is known. It is important to estimate the misclassification rates of a given classification rule in allocating a future observation that is randomly from one of the k groups. There is a rich literature discussing the cross-validation (CV) and bootstrap estimators of the misclassification rates, e.g., Lachenbruch and Mickey (1968), Lachenbruch (1975), Efron (1979, 1983), McLachlam (1980), Chatterjee and Chatterjee (1983), Läuter (1985), Wang (1986), Chen and Tu (1987), Snapinn and Knoke (1988), Konishi and Honda (1990), Ganeshanandam and Krzanowski (1990), and Davison and Hall (1992). Here we only discuss the situation where an individual observation is to be classified into one of $k = 2$ p -dimensional populations, and the classification rule is constructed by using Fisher's linear discriminant function.

Let the two population distributions be $F^{(1)}$ and $F^{(2)}$ and $\{X_i^{(t)}, i = 1, \dots, n_t\}$ be a random sample from $F^{(t)}$, $t = 1, 2$. Define

$$\bar{X}^{(t)} = \frac{1}{n_t} \sum_{i=1}^{n_t} X_i^{(t)} \quad t = 1, 2,$$

and

$$\mathbf{S} = \frac{1}{n_1 + n_2 - 2} \sum_{t=1}^2 \sum_{i=1}^{n_t} (X_i^{(t)} - \bar{X}^{(t)})(X_i^{(t)} - \bar{X}^{(t)})'.$$

Fisher's linear discriminant function is defined by

$$d(x) = [x - \frac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})]' \mathbf{S}^{-1} (\bar{X}^{(1)} - \bar{X}^{(2)}).$$

The classification rule based on $d(x)$ allocates an individual with measurement X_0 to population $F^{(1)}$ if $d(X_0) > c$ or to population $F^{(2)}$ if $d(X_0) \leq c$, where c is a constant depending on the costs of misclassification and the prior probabilities of $X_0 \sim F^{(t)}$. In particular, $c = 0$ for equal prior probabilities and equal costs of misclassification. This classification rule has some optimality properties when $F^{(t)}$ are multivariate normal with a common covariance matrix. The conditional misclassification rates are defined by

$$e(F^{(1)}) = P\{ d(X_0) \leq c | \bar{X}^{(1)}, \bar{X}^{(2)}, \mathbf{S} \}, \quad X_0 \sim F^{(1)},$$

and

$$e(F^{(2)}) = P\{ d(X_0) > c | \bar{X}^{(1)}, \bar{X}^{(2)}, \mathbf{S} \}, \quad X_0 \sim F^{(2)}.$$

The unconditional misclassification rates are $E[e(F^{(t)})]$, $t = 1, 2$.

When $F^{(1)}$ and $F^{(2)}$ are normal and have a common covariance matrix, an asymptotic expansion of $E[e(F^{(t)})]$ can be obtained and estimated by substituting the unknown parameters with their estimators. A simple nonparametric estimator of $e(F^{(t)})$, called the resubstitution estimator and denoted by $\hat{e}_{\text{R}}^{(t)}$, can be obtained by applying the classification rule to every observation in the training sample; that is, $\hat{e}_{\text{R}}^{(t)}$ is the proportion of $\{X_i^{(t)}, i = 1, \dots, n_t\}$ incorrectly classified. Since the same data set is used for constructing and validating the classification rule, the resubstitution method is too optimistic and leads to serious underestimation of the misclassification rates.

The CV and the bootstrap can be applied to get better nonparametric estimators of misclassification rates. The delete-1 CV estimator of $e(F^{(1)})$ can be constructed as follows. Omit $X_i^{(1)}$ and use the remaining $n_1 + n_2 - 1$ observations in the training sample to construct a new linear discriminant function $d_i(x)$; assess the classification rule based on $d_i(x)$ by classifying $X_i^{(1)}$, i.e., calculating $d_i(X_i^{(1)})$; repeat the same process for $i = 1, \dots, n_1$; the CV estimator is defined by

$$\hat{e}_{\text{CV}}^{(1)} = \frac{1}{n_1} \sum_{i=1}^{n_1} I\{d_i(X_i^{(1)}) \leq c\}.$$

Some algebraic calculations lead to

$$d_i(X_i^{(1)}) = [X_i^{(1)} - \frac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)} - Z_i^{(1)})]' \mathbf{S}_i^{-1} (\bar{X}^{(1)} - \bar{X}^{(2)} - Z_i^{(1)})$$

(e.g., Morrison, 1990), where $Z_i^{(1)} = (n_1 - 1)^{-1}(X_i^{(1)} - \bar{X}^{(1)})$,

$$\mathbf{S}_i^{-1} = \frac{n_1 + n_2 - 3}{n_1 + n_2 - 2} \left(\mathbf{S}^{-1} + \frac{c_1 \mathbf{S}^{-1} Z_i^{(1)} Z_i^{(1)'} \mathbf{S}^{-1}}{1 - c_1 Z_i^{(1)'} \mathbf{S}^{-1} Z_i^{(1)}} \right),$$

and $c_1 = n_1(n_1 - 1)/(n_1 + n_2 - 2)$. The CV estimator of $e(F^{(2)})$ can be similarly defined.

Efron (1983) suggested the following bootstrap estimator of $e(F^{(1)})$:

$$\hat{e}_{\text{BOOT}}^{(1)} = \hat{e}_{\text{R}}^{(1)} - b_{\text{BOOT}},$$

where

$$b_{\text{BOOT}} = E_* \left(\frac{1}{n_1} \sum_{i=1}^{n_1} I\{d^*(X_i^{(1)*}) \leq c\} - \frac{1}{n_1} \sum_{i=1}^{n_1} I\{d^*(X_i^{(1)}) \leq c\} \right)$$

is a bootstrap estimator of the bias of the resubstitution estimator $\hat{e}_{\text{R}}^{(1)}$, $\{X_1^{(t)*}, \dots, X_{n_t}^{(t)*}\}$, $t = 1, 2$, are independent bootstrap data i.i.d. from the

empirical distributions of $X_1^{(t)}, \dots, X_{n_t}^{(t)}$, $t = 1, 2$, respectively, and $d^*(x)$ is the linear discriminant function based on the bootstrap data. The bootstrap estimators $\hat{e}_{\text{BOOT}}^{(t)}$ have to be approximated by Monte Carlo.

Davison and Hall (1992) provided some theoretical results showing that asymptotically the bootstrap estimators have lower variability but higher bias than the CV estimators. Efron (1983) considered some improvements over the bootstrap estimators by randomization and double bootstrapping.

The following is a simulation example due to Konishi and Honda (1990). Consider

$$F^{(1)} = (1 - \epsilon)N(0, I_3) + \epsilon N(\nu, \sigma^2 I_3)$$

and

$$F^{(2)} = (1 - \epsilon)N(\mu, I_3) + \epsilon N(\mu + \nu, \sigma^2 I_3).$$

These two distributions can be identified by parameters ϵ , $\delta = \mu' \Sigma^{-1} \mu$, the skewness γ , and the kurtosis κ (defined in Konishi and Honda, 1990). Table 8.6 lists the means (ME) and the standard deviations (SD) of the CV and bootstrap estimators of the misclassification rates, based on $n_1 = n_2 = 30$ and 1000 simulations. The bootstrap estimators are approximated by Monte Carlo with size 100. The true misclassification rates (TMR), calculated by simulation, are also included in Table 8.6.

It can be seen from Table 8.6 that the bootstrap estimators perform well and are less variable than the CV estimators. The CV estimators get worse as δ decreases.

A more extensive simulation comparisons of estimators of misclassification rates can be found in Ganeshanandam and Krzanowski (1990).

The CV and bootstrap methods can be easily extended to the general case of $k > 2$ populations. They can also be extended to the situa-

Table 8.6. Simulation means and standard deviations of the CV and bootstrap estimators of misclassification rates [Adapted from Konishi and Honda (1990), by permission of Gordon and Breach Science Publishers]

ϵ	δ	γ	κ	TMR	CV		BOOT	
					ME	SD	ME	SD
0.0	3.0	0	15.0	0.074	0.077	0.034	0.074	0.033
0.1	2.2	0	41.7	0.104	0.109	0.041	0.105	0.040
0.3	1.6	0	32.4	0.155	0.161	0.050	0.155	0.048
0.5	1.3	0	24.6	0.205	0.195	0.047	0.207	0.054
0.7	1.2	0	19.6	0.254	0.265	0.062	0.255	0.060
0.9	1.1	0	16.3	0.302	0.313	0.066	0.301	0.064
1.0	1.0	0	15.0	0.325	0.338	0.068	0.324	0.065

tions where other classification rules are used. For example, Läuter (1985) considered quadratic discriminant functions. Applications of the CV and bootstrap in classification and regression trees can be found in Breiman *et al.* (1984).

8.6.4 Factor analysis and clustering

Factor analysis and clustering, together with principal component analysis, are the main statistical techniques used in multivariate analysis for data reduction. Since the statistics used in principal component analysis are functions of eigenvalues and eigenvectors of the sample covariance matrix, the bootstrap procedure introduced in Section 8.6.1 can be used to approximate their distributions and to make other statistical inferences. Here we briefly discuss some applications of the bootstrap in factor analysis and clustering.

In factor analysis, the p -dimensional observation vector X is assumed to satisfy the following model:

$$X - \mu = LF + E,$$

where μ is a p -vector (the population mean), L is a deterministic $p \times m$ matrix of factor loadings, F is a random m -vector of unobservable common factors, $m < p$, and E is a random p -vector of unobservable specific factors. It is assumed that F and E are independent, $E(F) = 0$, $\text{cov}(F) = I_m$, $E(E) = 0$, and $\text{cov}(E) = \Psi$, a diagonal matrix. These assumptions imply that

$$\Sigma = \text{cov}(X) = LL' + \Psi.$$

For a given m , the main problem in factor analysis is to estimate the loading matrix L and Ψ . Many estimation procedures have been proposed, e.g., the method of principal component or the principal factor, the maximum likelihood method, the generalized least squares, etc. However, the asymptotic variances and covariances of the estimators based on these methods are very complicated and difficult to derive, even though we assume that X is normally distributed. See Lawley and Maxwell (1971) for some complicated expressions. In addition, results based on these expressions are sensitive to departures from the normal distribution. The bootstrap provides a simple nonparametric method to estimate the sampling distributions, variances, and covariances of these estimators. We only need to repeatedly draw bootstrap data from the original data set and compute the estimators of L and Ψ based on the bootstrap data. Chatterjee (1984) provided some empirical results when the principal component method is used for estimation. His results indicate that as long as the sample size is moderate, the bootstrap estimators perform well. To use the bootstrap

one imposes almost no distributional assumption, at the expense of a large amount of computation.

Clustering is to partition n p -dimensional observations into k ($\ll n$) clusters (groups) so that the data within a cluster are “similar” and the data in different clusters are not similar. After clustering, the data set can be represented by k cluster descriptors (one from each cluster), thereby greatly reducing the size of the data set and ideally losing little information.

Peck, Fisher and van Ness (1989) proposed the following method for clustering. Let $\{X_1, \dots, X_n\}$ be the data set. For any clustering rule ξ , which determines the number of clusters k , allocates X_1, \dots, X_n to k clusters, and constructs cluster descriptors Y_1, \dots, Y_k , we define its loss function by

$$L_n(\xi) = c_1(k) + \frac{1}{n} \sum_{i=1}^k \sum_{X_j \in R_i} c_2(X_j, Y_i),$$

where c_1 is the cost of having k clusters, c_2 is the cost associated with describing an individual by its cluster descriptor, and R_i contains the data that are classified into the i th cluster. A clustering rule minimizing L_n is called the optimal sample clustering. Some numerical algorithms were suggested by Peck, Fisher and van Ness (1989) to find the optimal sample clustering.

One problem in clustering is to construct a confidence interval for k , the number of clusters. Peck, Fisher and van Ness (1989) applied the bootstrap for this purpose. Let X_1^*, \dots, X_n^* be i.i.d. from the empirical distribution of X_1, \dots, X_n . Find a clustering rule ξ^* that minimizes

$$L_n^*(\xi) = c_1(k) + \frac{1}{n} \sum_{i=1}^k \sum_{X_j^* \in R_i} c_2(X_j^*, Y_i).$$

Let k^* be the number of clusters determined by ξ^* . Repeat the above process B times independently to obtain k_1^*, \dots, k_B^* . We can then use the following bootstrap percentile confidence interval for k :

$$[c_\alpha^*, c_{1-\alpha}^*],$$

where c_α^* and $c_{1-\alpha}^*$ are the α th and $(1 - \alpha)$ th quantiles of the empirical distribution of k_1^*, \dots, k_B^* , respectively. Some simulation results are shown in Peck, Fisher and van Ness (1989).

Jhun (1990) discussed the bootstrap approximation to the distribution of cluster descriptors constructed by the k -mean clustering rule (e.g., Morrison, 1990).

8.7 Conclusions and Discussions

- (1) In nonlinear regression models, generalized linear models, and Cox's regression problems, the delete-one-pair (response and covariates) jackknife provides consistent estimators of the asymptotic variances of functions of regression parameters. Since the regression estimates are calculated by iteration in these nonlinear problems, the one-step jackknife or the linear jackknife can be applied to reduce the computation for the jackknife estimators.
- (2) The three bootstrap procedures studied in linear models, the bootstrap based on residuals, the paired bootstrap, and the external bootstrap, can all be applied to the nonlinear models and have properties similar to what they have in linear models. Again, the one-step bootstrap or the linear bootstrap can be used to reduce the computation for the bootstrap estimators.
- (3) Model selection using the cross-validation or the bootstrap in nonlinear or generalized linear models has a similar tendency to that in linear models; that is, the delete-1 cross-validation or the bootstrap with bootstrap sample size n may not lead to a consistent model selection procedure, whereas the delete-d cross-validation with $d/n \rightarrow 1$ or the bootstrap with bootstrap sample size m satisfying $m/n \rightarrow 0$ provides consistent model selection procedures.
- (4) In Section 8.3, we discussed the applications of the jackknife and bootstrap in Cox's regression models. In survival analysis, the jackknife and bootstrap can also be applied to estimate the variance and distribution of the Kaplan-Meier estimator (Efron, 1981b; Gaver and Miller, 1983; Akritas, 1986; Lo and Singh, 1986; Horvath and Yandell, 1987; Singh and Liu, 1990; Stute and Wang, 1994). Rao and Tu (1991) and Tu and Gross (1995) applied the jackknifed Edgeworth expansion (Section 4.4.2) and bootstrap to construct accurate confidence intervals for the ratio of specific occurrence/exposure rates in survival analysis.
- (5) In nonparametric curve estimation problems (density estimation and nonparametric regression), the cross-validation (or the generalized cross-validation) and the bootstrap can be applied to bandwidth selection. The cross-validation is simple and usually provides asymptotically valid results. The bootstrap may be more accurate than the cross-validation, but it requires more computations and involves some complicated steps such as bias estimation or an initial bandwidth selection. Furthermore, the asymptotic validity of the bootstrap rests upon a correct way of generating bootstrap data.
- (6) The bootstrap (or the one-step bootstrap) can be applied to set confidence sets or simultaneous confidence intervals in nonlinear or non-

parametric models. If a studentized statistic is available, then the bootstrap-t confidence sets are more accurate than the other confidence intervals in terms of the coverage probability. Similar to the case of linear models, the bootstrap is not model-free and one has to be careful in applying this method. When the models and/or the problems are complex, some bootstrap estimators may be inconsistent. Modifications by taking account of the special feature of the model under consideration are often necessary.

- (7) In Sections 8.4-8.5, we mainly focused on the application of the resampling methods when the quantity of interest is $f(x)$ (density or regression function). Sometimes it is of interest to estimate a given functional of f . The bootstrap can be used to estimate the distribution of an estimator of this functional. For example, Padgett and Thombs (1986) considered the kernel estimator of a quantile with censored data; Romano (1988b) studied the kernel estimator of the mode of a density; Zheng (1985), Liu and Tu (1987), and Gangopadhyay and Sen (1990) discussed the kernel and nearest neighbor estimators for the conditional quantile functional in nonparametric regression with random regressor; and Tu (1988a,b, 1989) investigated the kernel and nearest neighbor estimators for conditional L -functionals in the same setting.
- (8) There are other important nonparametric models that are not discussed here, e.g., the projection pursuit and the generalized additive model. Applications of the cross-validation and bootstrap to these models can be found, for example, in Hastie and Tibshirani (1985, 1990).
- (9) The bootstrap can be applied to almost all branches of multivariate analysis. The cross-validation and the bootstrap can be effectively applied to estimate misclassification rates in discriminant analysis. The resampling schemes are almost the same as those for the univariate case. But the computations of the bootstrap estimators in multivariate cases may be very heavy, especially when one wants to apply some computer-intensive methods in multivariate analysis such as the projection pursuit. The theory behind the proposed bootstrap and cross-validation procedures, however, has not yet been completely developed. Some theoretical studies and developments of methods that reduce the computations required by the bootstrap are called for.
- (10) To give the reader some idea about the regularity conditions required for the asymptotic validity of the resampling methods, some key regularity conditions are described when theoretical properties of the resampling methods are stated. More details and discussions about these regularity conditions and rigorous proofs of the results can be found in the references cited in this chapter.

Chapter 9

Applications to Time Series and Other Dependent Data

In the last three chapters, we discussed many applications of the jackknife and bootstrap methods in problems with non-i.i.d. data. In most of the problems previously studied, however, the data were either independent or had a cluster structure in which the observations from different clusters were independent and the number of clusters was large so that the jackknife and bootstrap could be applied to clusters instead of the original units that were correlated within each cluster (e.g., the survey data described in Chapter 6 or the longitudinal data described in Chapter 8). In this chapter, we study the application of the jackknife and bootstrap to time series and other dependent data. A time series is a sequence of observations that are indexed by time and are usually correlated. Other dependent data include m -dependent data, Markov chains, α -mixing data, and other stationary stochastic processes that are not indexed by time.

Since the original jackknife and bootstrap were developed for i.i.d. data, they may not be applicable to a non-i.i.d. problem, yet in the previous chapters we found from time to time that the jackknife and bootstrap had some robustness properties against the violation of the i.i.d. assumption (e.g., robustness against heteroscedasticity). For general dependent data, however, the jackknife and bootstrap fail to capture the dependence structure of the data and require nontrivial modifications in order to produce valid variance estimators and other inference procedures. Examples will be given as we proceed.

Similar to Chapter 8, we shall omit most of the technical proofs that are too complicated or require high level mathematics.

9.1 **m**-Dependent Data

A sequence of random variables $\{X_t, t = 0, \pm 1, \pm 2, \dots\}$ is (strongly) stationary if, for any integers q and $p > 0$, $\{X_1, \dots, X_p\}$ has the same distribution as $\{X_{1+q}, \dots, X_{p+q}\}$. All sequences of random variables considered in this section are assumed to be stationary.

m-dependence is the simplest dependence structure in statistical applications. A sequence of random variables $\{X_t, t = 0, \pm 1, \pm 2, \dots\}$ is said to be *m*-dependent if there is a fixed nonnegative integer *m* such that for all integers *t*, $\{\dots, X_{t-1}, X_t\}$ and $\{X_{t+m+1}, X_{t+m+2}, \dots\}$ are independent. From the definition, X_i are i.i.d. if *m* = 0. If *m* ≥ 1, then X_i are dependent. The following is an example.

Example 9.1. Moving average models. A moving average model is a time series model in which the data $\{y_t, t = 0, \pm 1, \pm 2, \dots\}$ can be represented as

$$y_t = \mu + \varepsilon_t - \phi_1 \varepsilon_{t-1} - \phi_2 \varepsilon_{t-2} - \dots - \phi_m \varepsilon_{t-m}, \quad (9.1)$$

where *m* is a fixed positive integer, μ and ϕ_j , $j = 1, \dots, m$, are unknown parameters, and the ε_t are i.i.d. random variables with mean 0 and variance σ^2 . It can be easily shown that $\{y_t, t = 0, \pm 1, \pm 2, \dots\}$ is an *m*-dependent series. In fact,

$$\text{cov}(y_t, y_{t+p}) = \sigma^2 \sum_{j=p}^m \phi_j \phi_{j-p}$$

for $1 \leq p \leq m$ ($\phi_0 = -1$) and $\text{cov}(y_t, y_{t+p}) = 0$ for $p > m$.

As an example of the inconsistency of the bootstrap, Singh (1981) considered the sample mean \bar{X}_n of *m*-dependent X_1, \dots, X_n . It is known that if $E(X_1) = \mu$ and $\text{var}(X_1) = \sigma^2$ exist and

$$\text{var}(\bar{X}_n) = \frac{\sigma^2}{n} + \frac{2}{n} \sum_{t=1}^m \text{cov}(X_1, X_{1+t}) > 0,$$

then

$$(\bar{X}_n - \mu)/[\text{var}(\bar{X}_n)]^{1/2} \rightarrow_d N(0, 1)$$

(e.g., Fuller, 1976). However, if \bar{X}_n^* is the sample mean of X_1^*, \dots, X_n^* that are i.i.d. from the empirical distribution of X_1, \dots, X_n , then

$$\text{var}_*(\bar{X}_n^*) = \frac{1}{n^2} \sum_{i=1}^n (X_i - \bar{X}_n)^2;$$

and using the methods in Chapter 3 we can prove that, for almost all given X_1, X_2, \dots ,

$$(\bar{X}_n^* - \bar{X}_n)/[\text{var}_*(\bar{X}_n^*)]^{1/2} \rightarrow_d N(0, 1).$$

From the strong law of large numbers for m -dependent data (Stout, 1974),

$$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \rightarrow_{a.s.} \sigma^2.$$

Therefore, the bootstrap distribution estimator is inconsistent unless

$$\sum_{t=1}^m \text{cov}(X_1, X_{1+t}) = 0. \quad (9.2)$$

Also, the bootstrap variance estimator for \bar{X}_n , which is equal to $\frac{n}{n-1}$ the jackknife variance estimator, is inconsistent unless (9.2) holds. The main reason for this inconsistency is that X_1^*, \dots, X_n^* are independent, whereas X_1, \dots, X_n are not.

Since the discovery of the inconsistency of the bootstrap, there have been several proposals to rectify the inconsistency by modifying the method of generating bootstrap data. The earliest proposal is the grouping method given by Carlstein (1986) and Shi (1986b). Without loss of generality, we assume that $n = kh$ for two integers k and h satisfying $k \rightarrow \infty$ and $h \rightarrow \infty$ as $n \rightarrow \infty$. We group the data into k groups with h observations in each group: $\{X_{(j-1)h+1}, X_{(j-1)h+2}, \dots, X_{jh}\}$, $j = 1, \dots, k$. When h is large, the groups are approximately independent. We can then apply the jackknife and bootstrap by treating the groups as the primary units of the data set.

The idea of grouping has been used in reducing the amount of computation required by the jackknife (Section 5.1.2). But there are two important distinctions between the grouping for computational savings and the grouping for m -dependent data: (1) for m -dependent data, the size of the group, h , must tend to ∞ as n does, but for computational savings h can be fixed; (2) for m -dependent data, the observations are grouped according to their natural order, whereas for computational savings in the i.i.d. case the groups are formed randomly.

Example 9.2. Variance estimation for \bar{X}_n . Assume that $\{X_1, X_2, \dots\}$ is m -dependent, $E(X_1) = \mu$ ($\mu = 0$ with out loss of generality), and $\text{var}(X_1) = \sigma^2$. Let $n = kh$ with $h = O(n^\delta)$ for a $\delta \in (0, \frac{1}{9})$ and Z_i be the sample mean of the data in the i th group, i.e.,

$$Z_i = \frac{1}{h} \sum_{j=1}^h X_{(i-1)h+j}, \quad i = 1, \dots, k. \quad (9.3)$$

Then the *grouped jackknife variance estimator* for \bar{X}_n is given by

$$\frac{1}{k(k-1)} \sum_{i=1}^k \left(Z_i - \frac{1}{k} \sum_{i=1}^k Z_i \right)^2 = \frac{1}{k(k-1)} \sum_{i=1}^k (Z_i - \bar{X}_n)^2, \quad (9.4)$$

and the *grouped bootstrap variance estimator* is given by (9.4) with $\frac{1}{k(k-1)}$ replaced by $\frac{1}{k^2}$. Define

$$\zeta_{nt} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{h-t} X_{(i-1)h+j} X_{(i-1)h+j+t}, \quad t = 1, \dots, m,$$

and

$$R_n = \frac{2}{n} \sum_{i=1}^k \sum_{l=j>m, l \leq h} X_{(i-1)h+l} X_{(i-1)h+j}.$$

Then

$$\begin{aligned} \frac{n}{k^2} \sum_{i=1}^k (Z_i - \bar{X}_n)^2 &= \frac{h}{k} \sum_{i=1}^k Z_i^2 - h\bar{X}_n^2 \\ &= \frac{1}{n} \sum_{i=1}^n X_i^2 + 2 \sum_{t=1}^m \zeta_{nt} + R_n - h\bar{X}_n^2 \\ &= \sigma^2 + 2 \sum_{t=1}^m \zeta_{nt} + R_n + o(1) \text{ a.s.}, \end{aligned}$$

where the last equality follows from $n^{-1} \sum_{i=1}^n X_i^2 \rightarrow_{a.s.} \sigma^2$ and $h\bar{X}_n^2 \rightarrow_{a.s.} 0$ (the strong law of large numbers and Marcinkiewicz strong law of large numbers for *m*-dependent random variables; see Stout, 1974). For each t ,

$$\begin{aligned} \zeta_{nt} &= \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^h X_{(i-1)h+j} X_{(i-1)h+j+t} \\ &\quad - \frac{1}{n} \sum_{i=1}^k \sum_{j=h-t+1}^h X_{(i-1)h+j} X_{(i-1)h+j+t} \\ &= \frac{1}{n} \sum_{i=1}^n X_i X_{i+t} - \frac{1}{h} \sum_{l=1}^t \left(\frac{1}{k} \sum_{i=1}^k X_{ih-t+l} X_{ih+l} \right). \end{aligned}$$

Since $n^{-1} \sum_{i=1}^n X_i X_{i+t}$ is an average of $(m+t)$ -dependent random variables, it converges almost surely to $E(X_1 X_{1+t}) = \text{cov}(X_1, X_{1+t})$ by the strong law of large numbers. For each fixed l and sufficiently large h ,

$k^{-1} \sum_{i=1}^k X_{ih-t+l} X_{ih+l}$ is an average of i.i.d. random variables and converges almost surely to $E(X_1 X_{1+t})$. Since $h \rightarrow \infty$, we conclude that $\zeta_{nt} \rightarrow_{a.s.} \text{cov}(X_1, X_{1+t})$. Similarly, we can show that $R_n \rightarrow_{a.s.} 0$. Thus,

$$\frac{n}{k^2} \sum_{i=1}^k (Z_i - \bar{X}_n)^2 \rightarrow_{a.s.} \sigma^2 + 2 \sum_{t=1}^m \text{cov}(X_1, X_{1+t}), \quad (9.5)$$

which implies the strong consistency of the grouped jackknife and bootstrap variance estimators for \bar{X}_n .

Shi and Shao (1988) suggested the external bootstrap (Section 7.2.2) based on the group means Z_i . Let e_1^*, \dots, e_n^* be i.i.d. from a distribution with mean 0 and variance 1. Define

$$X_{(i-1)h+l}^* = \bar{X}_n + \sqrt{h}(Z_i - \bar{X}_n)e_{(i-1)h+l}^*, \quad l = 1, \dots, h, \quad i = 1, \dots, k,$$

where Z_i is defined by (9.3). We can take X_1^*, \dots, X_n^* as the bootstrap data. Apparently, this procedure can still be applied when the X_i are vectors. The bootstrap estimator of the distribution of $\sqrt{n}[g(\bar{X}_n) - g(\mu)]$ is then the bootstrap distribution of $\sqrt{n}[g(\bar{X}_n^*) - g(\bar{X}_n)]$, where \bar{X}_n^* is the average of X_1^*, \dots, X_n^* . The variance estimator produced by this procedure is the same as the grouped bootstrap variance estimator.

Example 9.3. Distribution estimation for \bar{X}_n . Assume the conditions in Example 9.2 and that $E|X_1|^3 < \infty$. Then the consistency of the grouped external bootstrap distribution estimator for $\sqrt{n}(\bar{X}_n - \mu)$ can be shown as follows. Let \hat{a}_n^2 be the quantity on the left-hand side of (9.5). Since $X_i^* - \bar{X}_n$ are independent conditionally on X_1, \dots, X_n , an application of the Berry-Esséen inequality (Appendix A.9) leads to

$$\sup_x \left| P_* \left\{ \frac{\sqrt{n}(\bar{X}_n^* - \bar{X}_n)}{\hat{a}_n} \leq x \right\} - \Phi(x) \right| \leq \frac{ch}{k^{3/2} \hat{a}_n^3} \sum_{i=1}^k |Z_i - \bar{X}_n|^3$$

(c is a constant and Φ is the standard normal distribution), which is bounded by $ch^{-1/2}$ according to the definition of \hat{a}_n . From the Berry-Esséen inequality for m -dependent data (Tikhomirov, 1980),

$$\sup_x \left| P \left\{ \frac{\bar{X}_n - \mu}{[\text{var}(\bar{X}_n)]^{1/2}} \leq x \right\} - \Phi(x) \right| \rightarrow 0.$$

Therefore, the consistency of the bootstrap estimator follows from (9.5).

This example indicates that the grouped external bootstrap produces a consistent distribution estimator for $g(\bar{X}_n)$ with a smooth g . However, the convergence rate of this bootstrap distribution estimator is slower than

the normal approximation. By varying the size of the bootstrap sample, Yu and Tu (1987) demonstrated that the convergence rate of the grouped external bootstrap estimator can be improved to be the same as that of the normal approximation.

Note that the groups in the previous discussions are not overlapped. By allowing overlapped groups, Künsch (1989) and Liu and Singh (1992b) introduced the *moving block jackknife and bootstrap*. Let h be an integer less than n and $B_i = \{X_i, X_{i+1}, \dots, X_{i+h-1}\}$ be the i th block (group) of size h , $i = 1, \dots, n-h+1$. The moving block jackknife variance estimator for a statistic $T_n = T_n(X_1, \dots, X_n)$ of interest is given by

$$\frac{h}{n(n-h+1)} \sum_{i=1}^{n-h+1} \left(\tilde{T}_{n,i} - \frac{1}{n-h+1} \sum_{i=1}^{n-h+1} \tilde{T}_{n,i} \right)^2, \quad (9.6)$$

where $\tilde{T}_{n,i} = [nT_n - (n-h)T_{n-h,i}]/h$ is the i th pseudovalue and $T_{n-h,i}$ is the statistic T_{n-h} based on the data without the i th block B_i , $i = 1, \dots, n-h+1$. The moving block bootstrap can be defined as follows. Let B_1^*, \dots, B_k^* be i.i.d. from $\{B_1, \dots, B_{n-h+1}\}$, $\ell = kh$, and T_ℓ^* be the statistic T_ℓ based on the bootstrap data B_1^*, \dots, B_k^* . Then the bootstrap distribution estimator for $\sqrt{n}(T_n - ET_n)$ is the bootstrap distribution of $\sqrt{\ell}(T_\ell^* - E_*T_\ell^*)$.

For m -dependent $\{X_1, X_2, \dots\}$ and $T_n = \bar{X}_n$, Liu and Singh (1992b) showed that if $E(X_1^4) < \infty$, $h \rightarrow \infty$, and $h/n \rightarrow 0$, then the jackknife variance estimator defined in (9.6) is consistent; if $E|X_1|^{4+\delta}$ for some $\delta > 0$, $h \rightarrow \infty$, and $h/n \rightarrow 0$, then

$$\sup_x |P_*\{\sqrt{\ell}(\bar{X}_\ell^* - E_*\bar{X}_\ell^*) \leq x\} - P\{\sqrt{n}(\bar{X}_n - \mu) \leq x\}| \rightarrow_p 0. \quad (9.7)$$

Furthermore, if $h/\sqrt{n} \rightarrow 0$, then $E_*\bar{X}_\ell^*$ in (9.7) can be replaced by \bar{X}_n . The idea of the proof of these results is the same as that in Examples 9.2 and 9.3, but the derivations are more involved.

Applications of these methods for more general dependent data are discussed later in Section 9.4.

Künsch (1989) showed that the moving block bootstrap variance estimator, which is almost the same as the moving block jackknife variance estimator (9.6), has mean squared error of order hn^{-3} when $T_n = \bar{X}_n$. The same conclusion can be drawn for the grouped jackknife and bootstrap variance estimators. Since $h \rightarrow \infty$ for the consistency of these estimators, their convergence rate is slower than n^{-3} , which is the convergence rate of the original jackknife and bootstrap variance estimators when the data are actually independent. This is the price one has to pay in order to adapt a method designed for independent data to the case of dependent data.

On the other hand, given a particular dependence structure, it is possible to find a consistent jackknife or bootstrap variance estimator that has the same efficiency as the original jackknife and bootstrap variance estimators in the i.i.d. case. For example, Shao (1992e) provided the following modified jackknife variance estimator for m -dependent data:

$$\frac{n-1}{n} \sum_{i,j, |j-i| \leq m} (T_{n-1,i} - \bar{T}_n)(T_{n-1,j} - \bar{T}_n),$$

where $\bar{T}_n = n^{-1} \sum_{i=1}^n T_{n-1,i}$. This estimator is strongly consistent for m -dependent data and T_n generated by differentiable functionals (see Section 2.2), and has mean squared error of order n^{-3} when $T_n = \bar{X}_n$.

9.2 Markov Chains

A sequence of random variables $\{X_0, X_1, X_2, \dots\}$ taking values in Ω , a countable subset of \mathbb{R} , is said to be a Markov chain if, for any integer i ,

$$P\{X_i = x_i | X_{i-1} = x_{i-1}, \dots, X_0 = x_0\} = P\{X_i = x_i | X_{i-1} = x_{i-1}\},$$

where x_0, x_1, \dots, x_i are elements of Ω . By its definition, a Markov chain is just one step away from an i.i.d. sequence.

For convenience, we assume that $\Omega = \{1, 2, \dots\}$. If $\{X_0, X_1, X_2, \dots\}$ is stationary, then the probabilities

$$p_{ij} = P\{X_t = j | X_{t-1} = i\}, \quad i, j \in \Omega,$$

are independent of t and $\{X_0, X_1, X_2, \dots\}$ is said to be homogeneous. The distribution of a homogeneous Markov chain is determined by its transition probability matrix \mathbf{P} whose (i, j) th element is p_{ij} .

Let $\hat{\mathbf{P}}$ be an estimator of \mathbf{P} based on the observations X_0, X_1, \dots, X_n . In this section, we discuss the bootstrap distribution estimators for quantities such as $\sqrt{n}(\hat{\mathbf{P}} - \mathbf{P})$.

Assume that Ω is a finite set. Then $\Omega = \{1, 2, \dots, k\}$ for a fixed integer k , and \mathbf{P} is a $k \times k$ matrix. The maximum likelihood estimator of \mathbf{P} is $\hat{\mathbf{P}}$, whose (i, j) th element is

$$\hat{p}_{ij} = \begin{cases} n_{ij}/n_i & \text{if } n_i > 0 \\ 1 & \text{if } n_i = 0 \text{ and } i = j \\ 0 & \text{if } n_i = 0 \text{ and } i \neq j, \end{cases}$$

where

$$n_{ij} = \sum_{t=1}^n I\{X_{t-1} = i, X_t = j\} \quad \text{and} \quad n_i = \sum_{t=1}^n I\{X_t = i\}.$$

According to the basic idea of the bootstrap illustrated by diagram (1.25), a bootstrap Markov chain $\{X_0^*, X_1^*, \dots, X_n^*\}$ can be generated by using $\hat{\mathbf{P}}$ as the transition probability matrix. Let $\hat{\mathbf{P}}^*$ be the maximum likelihood estimator based on the bootstrap data $\{X_0^*, X_1^*, \dots, X_n^*\}$. The bootstrap estimator of the distribution of $\sqrt{n}(\hat{\mathbf{P}} - \mathbf{P})$ is then the bootstrap distribution of $\sqrt{n}(\hat{\mathbf{P}}^* - \hat{\mathbf{P}})$, conditioned on X_0, X_1, \dots, X_n .

Kulperger and Prakasa Rao (1989) proved that if $\{X_0, X_1, X_2, \dots\}$ is a homogeneous ergodic¹ Markov chain with $\Omega = \{1, \dots, k\}$, then the conditional limit distribution of $\sqrt{n}(\hat{\mathbf{P}}^* - \hat{\mathbf{P}})$ is the same as the limit distribution of $\sqrt{n}(\hat{\mathbf{P}} - \mathbf{P})$; that is, the bootstrap distribution estimator is consistent. They also proved the consistency of the bootstrap distribution estimators for quantities such as the estimators of the distribution and the expectation of first hitting time of state i defined by $t_i = \inf\{t \geq 0 : X_t = i\}$.

What will happen if $\{X_0^*, X_1^*, \dots, X_n^*\}$ is drawn i.i.d. from $\{X_1, \dots, X_n\}$ (i.e., the dependence structure is ignored)? It is not difficult to show that $E_*(n_{ij}^*) = n_i n_j / n$ and $E_*(n_i^*) = n_i$ when the X_t^* are i.i.d. generated from $\{X_1, \dots, X_n\}$. Hence, the bootstrap procedure that works for the i.i.d. data is inconsistent for Markov chains.

Motivated by a multinomial-type representation of the Markov chain, Basawa *et al.* (1990) proposed a conditional bootstrap for estimating the distribution of $\sqrt{n}(\hat{\mathbf{P}} - \mathbf{P})$, conditioned on n_i , $i = 1, \dots, k$. Given the data and the maximum likelihood estimates \hat{p}_{ij} , the conditional bootstrap data X_{it}^* , $t = 1, \dots, n_i$, $i = 1, \dots, k$, are generated independently according to the probabilities

$$P_*\{X_{it}^* = j\} = \hat{p}_{ij}, \quad j = 1, \dots, k.$$

Define

$$n_{ij}^* = \sum_{t=1}^{n_i} I\{X_{it}^* = j\} \quad \text{and} \quad \hat{p}_{ij}^* = n_{ij}^*/n_i, \quad 1 \leq i, j \leq k.$$

Then the distribution of $\sqrt{n_i}(\hat{p}_{ij}^* - \hat{p}_{ij})$, conditioned on n_i , can be estimated by the bootstrap distribution of $\sqrt{n_i}(\hat{p}_{ij}^* - \hat{p}_{ij})$. Basawa *et al.* (1990) proved the consistency of this bootstrap estimator. Datta and McCormick (1992) further showed that if $\{X_0, X_1, X_2, \dots\}$ is a homogeneous ergodic Markov chain with $\Omega = \{1, \dots, k\}$ and $0 < p_{ij} < 1$, then

$$\sup_x \left| P_* \left\{ \frac{\sqrt{n_i}(\hat{p}_{ij}^* - \hat{p}_{ij})}{\sqrt{\hat{p}_{ij}(1 - \hat{p}_{ij})}} \leq x \right\} - P \left\{ \frac{\sqrt{n_i}(\hat{p}_{ij} - p_{ij})}{\sqrt{c_i p_{ij}(1 - p_{ij})}} \leq x \right\} \right| = O(n^{-1/2}) \text{ a.s.},$$

where $c_i = n P\{X_t = i\}/n_i$. The proof of this result is a direct application of the Berry-Esséen inequalities. Conditioned on X_0, X_1, \dots, X_n , the n_{ij}^*

¹See Feller (1968, p. 389).

are independent and distributed as binomial with parameters n_i and \hat{p}_{ij} . Hence,

$$\sup_x \left| P_* \left\{ \frac{\sqrt{n_i}(\hat{p}_{ij}^* - \hat{p}_{ij})}{\sqrt{\hat{p}_{ij}(1 - \hat{p}_{ij})}} \leq x \right\} - \Phi(x) \right| \leq cn_i^{-1/2} [\hat{p}_{ij}(1 - \hat{p}_{ij})]^{-3/2},$$

which is $O(n^{-1/2})$ a.s., since $n_i/n \rightarrow_{a.s.} p_i > 0$ and $\hat{p}_{ij}(1 - \hat{p}_{ij}) \rightarrow_{a.s.} p_{ij}(1 - p_{ij}) > 0$. On the other hand, from the Berry-Esséen inequality for homogeneous Markov chains (Nagaev, 1961), we obtain that

$$\sup_x \left| P \left\{ \frac{\sqrt{n_i}(\hat{p}_{ij} - p_{ij})}{\sqrt{c_i p_{ij}(1 - p_{ij})}} \leq x \right\} - \Phi(x) \right| = O(n^{-1/2}).$$

Datta and McCormick (1992) found, however, that the $n^{-1/2}$ -terms in the Edgeworth expansions of the bootstrap distribution of $\sqrt{n_i}(\hat{p}_{ij}^* - \hat{p}_{ij})$ and the distribution of $\sqrt{n_i}(\hat{p}_{ij} - p_{ij})$ do not match. Hence, the conditional bootstrap distribution estimator cannot be second order accurate. A possible reason is that the i.i.d. bootstrap sampling cannot account for a part of the skewness term arising from the dependence structure of the data. Some improvements that achieve the second order accuracy were proposed by Datta and McCormick (1992).

The situation where Ω is an infinite set is more complicated. Athreya and Fuh (1992a,b) proposed some bootstrap methods for estimating the distribution of $\hat{\mathbf{P}} - \mathbf{P}$, where $\hat{\mathbf{P}}$ is a consistent estimator of \mathbf{P} . Datta and McCormick (1993) suggested a cycling bootstrap for a Markov chain with some types of Ω .

9.3 Autoregressive Time Series

A sequence of dependent random variables $\{y_t, t = 0, \pm 1 \pm 2, \dots\}$ is often called a time series, even if the random variables are not indexed by time. In practical applications, many time series can be represented (at least approximately) as linear combinations of independent random variables. The moving average series given by (9.1) in Example 9.1 is one example. In this section we study another important type of time series: the autoregressive time series.

A time series $\{y_t, t = 0, \pm 1, \pm 2, \dots\}$ is called an autoregressive time series of order p if

$$y_t = \mu + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \dots + \theta_p y_{t-p} + \varepsilon_t, \quad (9.8)$$

where p is a fixed nonnegative integer, μ and θ_i , $i = 1, \dots, p$, are unknown parameters, and the ε_t are i.i.d. random variables with mean 0 and variance σ^2 . An autoregressive time series is stationary if the roots of $1 + \theta_1 z + \theta_2 z^2 + \dots + \theta_p z^p = 0$ are outside of the unit circle.

9.3.1 Bootstrapping residuals

The parameters of primary interest in an autoregressive model are θ_i , $i = 1, \dots, p$. To introduce the idea, we assume $\mu = 0$. Suppose that we observe $\{y_1, \dots, y_n\}$ and some initial observations $\{y_{1-p}, \dots, y_0\}$. Customary estimators of θ_i , $i = 1, \dots, p$, are the least squares estimators $\hat{\theta}_1, \dots, \hat{\theta}_p$ that minimize

$$\sum_{t=1}^n \left(y_t - \sum_{i=1}^p \theta_i y_{t-i} \right)^2.$$

Let $\beta = (\theta_1, \dots, \theta_p)'$ and $\hat{\beta} = (\hat{\theta}_1, \dots, \hat{\theta}_p)'$. Then

$$\hat{\beta} = S_n^{-1} \left(\sum_{t=1}^n y_{t-1} y_t, \dots, \sum_{t=1}^n y_{t-p} y_t \right)', \quad (9.9)$$

where S_n is a $p \times p$ matrix whose (i, j) th element is $\sum_{t=1}^n y_{t-i} y_{t-j}$.

To derive a bootstrap distribution estimator for $\hat{\beta} - \beta$, we have to be aware that $\{y_1, \dots, y_n\}$ are dependent. A careless way of generating bootstrap data may lead to inconsistent results. The following is an example.

Example 9.4. Autoregressive model of order 1. Consider the special case of (9.8) with $p = 1$ and $\mu = 0$. Write $\theta_1 = \theta$. Then

$$\hat{\theta} = \frac{\sum_{t=1}^n y_{t-1} y_t}{\sum_{t=1}^n y_t^2}.$$

Recognizing that $\hat{\theta}$ is a function of the sample mean $\bar{X}_n = n^{-1} \sum_{t=1}^n X_t$ with $X_t = (y_{t-1} y_t, y_{t-1}^2)'$, we might think that taking bootstrap data X_1^*, \dots, X_n^* i.i.d. from the empirical distribution of X_1, \dots, X_n produces a consistent bootstrap distribution estimator for $\hat{\theta}$. However, X_t and X_s are dependent when $t \neq s$. Using the same argument as that in Section 9.1, where we showed the inconsistency of this bootstrap procedure for m -dependent data, we can show that this bootstrap procedure is also inconsistent in this example.

Although one may apply the moving block bootstrap in an autoregressive model, a more efficient bootstrap procedure can be derived using the special structure of autoregressive models. Since the residuals

$$r_t = y_t - \sum_{i=1}^p \hat{\theta}_i y_{t-i}, \quad t = 1, \dots, n,$$

have behaviors similar to the errors ε_t , we can extend the bootstrap based on residuals described in Section 7.2.2 to autoregressive models. Let ε_t^* ,

$t = 0, \pm 1, \pm 2, \dots$, be i.i.d. from the empirical distribution putting mass n^{-1} to $r_t - \bar{r}$, $t = 1, \dots, n$, $\bar{r} = n^{-1} \sum_{t=1}^n r_t$. The bootstrap analog $\hat{\beta}^*$ of $\hat{\beta}$ is defined by (9.9) with y_t replaced by

$$y_t^* = \sum_{i=1}^p \hat{\theta}_i y_{t-i}^* + \varepsilon_i^*, \quad t = 1-p, \dots, 0, 1, \dots, n, \quad (9.10)$$

where $\{y_{1-2p}^*, \dots, y_{-p}^*\} = \{y_{1-p}, \dots, y_0\}$ are initial bootstrap observations (Efron and Tibshirani, 1986; Holbert and Son, 1986). We can then estimate the distribution of $\hat{\beta} - \beta$ by the bootstrap distribution of $\hat{\beta}^* - \hat{\beta}$.

Let Σ be the $p \times p$ matrix whose (i, j) th element is $\text{cov}(y_i, y_j)/\sigma^2$ and $\hat{\Sigma}$ be the $p \times p$ matrix whose (i, j) th element is $\text{cov}_*(y_i^*, y_j^*)/\text{var}_*(y_i^*)$. Assume that the roots of $1 + \theta_1 z + \theta_2 z^2 + \dots + \theta_p z^p = 0$ are outside of the unit circle,

$$E|\varepsilon_1|^{2(s+1)} < \infty \quad \text{for some } s \geq 3, \quad (9.11)$$

and that $(\varepsilon_1, \varepsilon_1^2)$ satisfies Cramér's condition, i.e., for every $c > 0$, there exists $\delta_c > 0$ such that

$$\sup_{\|u\| \geq c} |\psi(u)| \leq e^{-\delta_c}, \quad (9.12)$$

where $\psi(u)$ is the characteristic function of $(\varepsilon_1, \varepsilon_1^2)$. Using the Edgeworth expansion for sums of dependent random variables, Bose (1988) proved that

$$\sup_x |H_{\text{BOOT}}(x) - H_n(x)| = o(n^{-1/2}) \quad a.s., \quad (9.13)$$

where H_n is the distribution of $\sqrt{n}\Sigma^{-1/2}(\hat{\beta} - \beta)$ and H_{BOOT} is its bootstrap estimator: the bootstrap distribution of $\sqrt{n}\hat{\Sigma}^{-1/2}(\hat{\beta}^* - \hat{\beta})$.

When μ is not necessarily 0 and has to be estimated, the least squares estimators of μ and θ_i , $i = 1, \dots, p$, are obtained by minimizing

$$\sum_{t=1}^n \left(y_t - \mu - \sum_{i=1}^p \theta_i y_{t-i} \right)^2.$$

The bootstrap procedure discussed for the case of $\mu = 0$ can be extended to the general case by using the residuals

$$r_t = y_t - \hat{\mu} - \sum_{i=1}^p \hat{\theta}_i y_{t-i}, \quad t = 1, \dots, n,$$

and

$$y_t^* = \hat{\mu} + \sum_{i=1}^p \hat{\theta}_i y_{t-i}^* + \varepsilon_i^*, \quad t = 1-p, \dots, 0, 1, \dots, n.$$

Then result (9.13) is still valid.

Stine (1987), Thombs and Schucany (1990), and Kabala (1993) discussed the construction of prediction intervals and the estimation of the mean squared prediction error by using the bootstrap procedure previously discussed. The methodology is similar to that in linear models (Section 7.3.4) and will not be discussed here. Kim, Haddock and Willemain (1993) suggested a binary bootstrap to construct a confidence interval for the success probability in autoregressive binary time series. Janas (1993) provides a survey on the application of the bootstrap in time series analysis.

Some numerical results can be found in the following papers. Efron and Tibshirani (1986) and Holbert and Son (1986) applied the bootstrap in analyzing the data of annual sunspot numbers for the years 1770-1889 and 95 daily readings of viscosity of a chemical product XB-75-5. Holbert and Son (1986) and Bose (1990) contain some simulation results.

Note that result (9.13) is valid when $\{y_t, t = 0, \pm 1, \pm 2, \dots\}$ is stationary. When the time series is not stationary, the least squares estimator $\hat{\beta}$ may not be asymptotically normal, and it is of interest to know whether the bootstrap based on residuals can still provide a consistent distribution estimator for $\hat{\beta}$. We consider the special case considered in Example 9.4:

$$y_t = \theta y_{t-1} + \varepsilon_t, \quad t = 1, \dots, n. \quad (9.14)$$

Under model (9.14), the time series is stationary if and only if $|\theta| < 1$. Basawa *et al.* (1989) showed that when $|\theta| > 1$, the bootstrap based on residuals produces a consistent estimator of the distribution of $\hat{\theta}$ that is not asymptotically normal. When $|\theta| = 1$, however, Basawa *et al.* (1991a) showed that the same bootstrap estimator is inconsistent. Basawa *et al.* (1991b) established the asymptotic validity of a bootstrap estimator for the null distribution of a test for $\theta = 1$. They also suggested a sequential bootstrap procedure that produces a consistent estimator of the distribution of a sequential estimator of θ . It is not clear whether these results can be extended to a nonstationary autoregressive time series of order $p > 1$.

9.3.2 Model selection

In many practical problems, the order of an autoregressive model is unknown and has to be estimated by using the data. The estimation of the order can be formulated as a model selection problem in which we select a model α from $\{1, \dots, p\}$ and each α corresponds to the autoregressive model of order α :

$$y_t = \theta_1 y_{t-1} + \dots + \theta_\alpha y_{t-\alpha} + \varepsilon_t. \quad (9.15)$$

We assume that $\alpha = p$ is the largest possible model. The optimal model is defined to be $\alpha_o = \max\{j : 1 \leq j \leq p, \theta_j \neq 0\}$.

Because of the dependence among the data, the cross-validation method introduced in Section 7.4.1 may not be suitable for this problem. Rissanen (1986) proposed the predictive least squares (PLS) principle, which produces a selection procedure using a similar idea of the cross-validation. For an integer j , let $\hat{\beta}_{\alpha}^{(j)}$ be the least squares estimator of $\beta_{\alpha} = (\theta_1, \dots, \theta_{\alpha})'$ based on the subseries $\{y_0, y_1, \dots, y_j\}$ under model α , $\alpha = 1, \dots, p$, and $j = t_0, t_0 + 1, \dots, n$, where t_0 is the first integer so that $\hat{\beta}_p^{(t_0)}$ is uniquely defined. The PLS procedure selects an α by minimizing

$$\sum_{t=t_0+1}^n (y_t - z'_{t\alpha} \hat{\beta}_{\alpha}^{(t-1)})^2 \quad (9.16)$$

over $\alpha \in \{1, \dots, p\}$, where $z'_{t\alpha} = (y_{t-1}, \dots, y_{t-\alpha})$.

Note that each term $(y_t - z'_{t\alpha} \hat{\beta}_{\alpha}^{(t-1)})^2$ in (9.16) can be viewed as a delete-($n - t + 1$) cross-validation based on one data splitting. We do not need “cross-validation” here because $\{y_0, y_1, \dots, y_n\}$ has a natural order.

Let $\hat{\alpha}_{\text{PLS}}$ be the model selected by the PLS. Wax (1988) showed that the PLS procedure is consistent in the sense that

$$P\{\hat{\alpha}_{\text{PLS}} = \alpha_o\} \rightarrow 1, \quad (9.17)$$

assuming that the autoregressive series is stationary. Hannan, McDougall and Poskitt (1989) and Hemerly and Davis (1989) showed that (9.17) can be strengthened to

$$P\{\hat{\alpha}_{\text{PLS}} = \alpha_o \text{ eventually}\} = 1. \quad (9.18)$$

Wei (1992) established (9.18) for nonstationary autoregressive series and some other time series models. Wei (1992) also studied the asymptotic equivalence between the PLS and the BIC proposed by Schwartz (1978).

We next apply the bootstrap based on residuals described previously to model selection. Let y_t^* be the bootstrap data given by (9.10). Under model α , let $\hat{\beta}_{\alpha}^*$ be the bootstrap analog of the least squares estimator of β_{α} . Then the model selected by the bootstrap, denoted by $\hat{\alpha}_{\text{BOOT}}$, is the minimizer of

$$B_{\alpha} = E_* \left[\frac{1}{n} \sum_{t=1}^n (y_t - z'_{t\alpha} \hat{\beta}_{\alpha}^*)^2 \right], \quad (9.19)$$

where $z_{t\alpha}$ is defined in (9.16).

Due to the same reason discussed in Section 7.4.2, $\hat{\alpha}_{\text{BOOT}}$ is inconsistent unless the only correct model is $\alpha = p$. Therefore, we need to modify the bootstrap procedure. The method of changing the bootstrap sample size introduced in Section 7.4.2 can be used here; that is, we replace $\hat{\beta}_{\alpha}^*$ in (9.19)

by $\hat{\beta}_{\alpha,m}^*$, the least squares estimator of β_α under model α , but based on the subseries $\{y_t^*, t \leq m\}$, where m is an integer satisfying $m \rightarrow \infty$ and $m/n \rightarrow 0$ as $n \rightarrow \infty$. Let the model selected by this bootstrap method be $\hat{\alpha}_{\text{BOOT-}m}$. We now illustrate the consistency of this bootstrap procedure.

When $\alpha \geq \alpha_o$, using result (9.13) we obtain that

$$\begin{aligned} B_\alpha &= \hat{\sigma}_\alpha^2 + \frac{1}{n} \sum_{t=1}^n z'_{t\alpha} \text{var}_*(\hat{\beta}_{\alpha,m}^*) z_{t\alpha} + o_p\left(\frac{1}{m}\right) \\ &= \hat{\sigma}_\alpha^2 + \frac{\text{tr}(\hat{\Sigma}_\alpha^{-1} S_{n\alpha})}{m} + o_p\left(\frac{1}{m}\right) \\ &= \frac{1}{n} \sum_{t=1}^n \varepsilon_t^2 + \frac{\text{tr}(\hat{\Sigma}_\alpha^{-1} S_{n\alpha})}{m} + o_p\left(\frac{1}{m}\right), \end{aligned}$$

where

$$\hat{\sigma}_\alpha^2 = \frac{1}{n} \sum_{t=1}^n (y_t - z'_{t\alpha} \hat{\beta}_\alpha)^2,$$

$$S_{n\alpha} = \frac{1}{n} \sum_{t=1}^n z_{t\alpha} z'_{t\alpha},$$

$\hat{\Sigma}_\alpha$ is the $\alpha \times \alpha$ matrix whose (i,j) th element is $\text{cov}_*(y_i^*, y_j^*)/\text{var}_*(y_i^*)$, $i, j \leq \alpha$, and the last equality follows from the fact that, when $\alpha \geq \alpha_o$,

$$z'_{t\alpha} \beta_\alpha = \theta_1 y_{t-1} + \cdots + \theta_p y_{t-p} = y_t - \varepsilon_t$$

and

$$\hat{\beta}_\alpha - \beta_\alpha = O_p(n^{-1/2}).$$

Since both $\hat{\Sigma}_\alpha$ and $S_{n\alpha}/\sigma^2$ converge to Σ_α , the $\alpha \times \alpha$ matrix whose (i,j) th element is $\text{cov}(y_i, y_j)/\sigma^2$, $i, j \leq \alpha$ (Bose, 1988), we have

$$B_\alpha = \frac{1}{n} \sum_{t=1}^n \varepsilon_t^2 + \frac{\sigma^2 \alpha}{m} + o_p\left(\frac{1}{m}\right). \quad (9.20)$$

When $\alpha < \alpha_o$,

$$B_\alpha \geq \hat{\sigma}_\alpha^2 \quad (9.21)$$

and

$$\liminf_n (\hat{\sigma}_\alpha^2 - \hat{\sigma}_{\alpha_o}^2) > 0 \quad a.s. \quad (9.22)$$

(Wei, 1992). It follows from (9.20)-(9.22) that

$$P\{\hat{\alpha}_{\text{BOOT-}m} = \alpha_o\} \rightarrow 1.$$

The results in this section can be easily extended to the case where a constant term μ is added to models given by (9.15).

9.4 Other Time Series

In this section, we consider more general time series. The following are some examples.

Example 9.5. Autoregressive moving average models. A sequence of random variables $\{y_t, t = 0, \pm 1, \pm 2, \dots\}$ is called an autoregressive moving average time series of order (p, q) , denoted by ARMA(p, q), if

$$\begin{aligned} y_t &= \mu + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \cdots + \theta_p y_{t-p} \\ &\quad + \varepsilon_t - \phi_1 \varepsilon_{t-1} - \cdots - \phi_q \varepsilon_{t-q}, \end{aligned} \tag{9.23}$$

where p and q are fixed nonnegative integers, μ , θ_i , and ϕ_j are unknown parameters, and the ε_t are i.i.d. random variables with mean 0 and variance σ^2 . An ARMA(p, q) series is stationary if the roots of

$$1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_p z^p = 0$$

and the roots of

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_q z^q = 0$$

are outside of the unit circle.

Note that the series in (9.23) is a combination of the two series, the moving average series in (9.1) and the autoregressive series in (9.8). If $p = 0$ ($q = 0$), then the series in (9.23) reduces to a moving average (autoregressive) series. The autoregressive, the moving average, and, more generally, the ARMA(p, q) models are the most useful time series models in applications.

Example 9.6. Dynamical linear regression. A dynamical linear regression model is of the form

$$y_t = Ay_t + By_{t-1} + Cx_t + \varepsilon_t, \quad t = 1, \dots, n, \tag{9.24}$$

where y_t is the vector of endogenous variables at time t , x_t is the vector of exogenous variables at time t , A , B , and C are unknown coefficient matrices of appropriate orders, and the ε_t are i.i.d. errors with mean 0. The x_t can be random or deterministic. In the former case, the (x'_t, ε'_t) are i.i.d. with $E(x_t \varepsilon_t) = 0$.

Another example of a time series is a linear regression model with time series errors.

Similar to the case of autoregressive models, we can use the special structure of these time series models to find some bootstrap procedures that are more efficient than the grouped bootstrap or the moving block bootstrap introduced in Section 9.1.

9.4.1 ARMA(p, q) models

To illustrate the idea, we first consider a special case of $\mu = 0$, $p = 0$, and $q = 1$, i.e., a moving average model of order 1:

$$y_t = \varepsilon_t - \phi \varepsilon_{t-1}. \quad (9.25)$$

In addition, we assume $\text{var}(\varepsilon_1) = 1$ so that an unbiased moment estimator of ϕ is

$$\hat{\phi} = -\frac{1}{n} \sum_{t=1}^n y_t y_{t-1},$$

based on observations $\{y_0, y_1, \dots, y_n\}$.

Since a moving average time series of order q is q -dependent, the grouped bootstrap or the moving block bootstrap can be applied to estimate the sampling distribution of $\hat{\phi}$. But it is possible to develop a bootstrap procedure incorporating the specific structure of the model. Define

$$\tilde{\varepsilon}_t = \sum_{j=0}^{t-1} \phi^j y_{t-j}, \quad t = 1, \dots, n.$$

By (9.25),

$$\tilde{\varepsilon}_t = \varepsilon_t - \phi^t \sum_{j=0}^{\infty} \phi^j y_j.$$

Hence, $\tilde{\varepsilon}_t$ and ε_t are close for all large t if $|\phi| < 1$ (the series is stationary), and we may generate bootstrap data from estimated $\tilde{\varepsilon}_t$ given by

$$r_t = \sum_{j=0}^{t-1} \hat{\phi}^j y_{t-j}, \quad t = 1, \dots, n.$$

Let $\varepsilon_{-1}^*, \dots, \varepsilon_n^*$ be i.i.d. from the empirical distribution of r_1, \dots, r_n centered at $\bar{r} = n^{-1} \sum_{t=1}^n r_t$,

$$y_t^* = \varepsilon_t^* - \hat{\phi} \varepsilon_{t-1}^*, \quad t = 0, 1, \dots, n,$$

and

$$\hat{\phi}^* = -\frac{1}{n} \sum_{t=1}^n y_t^* y_{t-1}^*.$$

Assuming (9.11), (9.12), and $|\phi| < 1$, Bose (1990) showed that

$$\sup_x \left| P_* \left\{ \frac{\hat{\phi}^* - \hat{\phi}}{[\text{var}_*(\hat{\phi}^*)]^{1/2}} \leq x \right\} - P \left\{ \frac{\hat{\phi} - \phi}{[\text{var}(\hat{\phi})]^{1/2}} \leq x \right\} \right| = o(n^{-1/2}) \quad a.s.$$

The same conclusion can be drawn when $\mu \neq 0$ and ε_t has an unknown variance σ^2 .

We now consider the general ARMA(p, q) models described in Example 9.5. Let B be the backward shift operator satisfying $B^j y_t = y_{t-j}$ for any t and nonnegative integer j , and let $\theta(B) = 1 - \theta_1 B - \cdots - \theta_p B^p$ and $\phi(B) = 1 - \phi_1 B - \cdots - \phi_q B^q$. Then the ARMA(p, q) model (9.23) can be written as

$$\theta(B)y_t = \mu + \phi(B)\varepsilon_t.$$

If the roots of $\phi(z) = 0$ lie outside the unit circle, then $\phi^{-1}(B)$ is well defined and

$$\varepsilon_t = \phi^{-1}(B)[\theta(B)y_t - \mu]. \quad (9.26)$$

The estimation of μ , θ_i , and ϕ_j is not simple. Box and Jenkins (1970) suggested some estimators by solving some nonlinear equations. Suppose that we have estimators $\hat{\mu}$, $\hat{\theta}_i$, and $\hat{\phi}_j$, $i = 1, \dots, p$, $j = 1, \dots, q$. Using equation (9.26), we can estimate ε_t by

$$r_t = \hat{\phi}^{-1}(B)[\hat{\theta}(B)y_t - \hat{\mu}],$$

where $\hat{\theta}(B) = 1 - \hat{\theta}_1 B - \cdots - \hat{\theta}_p B^p$ and $\hat{\phi}(B) = 1 - \hat{\phi}_1 B - \cdots - \hat{\phi}_q B^q$, and generate bootstrap data $\varepsilon_{-q}^*, \dots, \varepsilon_n^*$ i.i.d. from the empirical distribution of r_1, \dots, r_n centered at \bar{r} . The bootstrap analogs of $\hat{\mu}$, $\hat{\theta}_i$, and $\hat{\phi}_j$ can be obtained by using the same algorithms in calculating $\hat{\mu}$, $\hat{\theta}_i$, and $\hat{\phi}_j$ but with the data replaced by the bootstrap analogs

$$y_t^* = \hat{\theta}^{-1}(B)[\hat{\mu} + \hat{\phi}(B)\varepsilon_t^*]$$

[assuming that the roots of $\theta(z) = 0$ lie outside the unit circle, i.e., the series is stationary]. Some initial y^* values may be generated similarly to the case of autoregressive models. This procedure is similar to that for model (9.25) discussed earlier.

Kreiss and Franke (1992) proved the consistency of the above bootstrap procedure when the $\hat{\theta}_i$ and $\hat{\phi}_j$ are M-estimators based on the sample. Chatterjee (1986) provided some empirical comparisons of the standard deviation estimators for $\hat{\theta}$ or $\hat{\phi}$ constructed using the bootstrap and a parametric asymptotic approach. The model was ARMA(1, 1) with $\theta = 0.5$, $\phi = 0.9$, and the errors ε_t were generated from a symmetric distribution with characteristic function $\exp(-|t|^a)$. When $a = 2$, the ε_t are standard normal; when $a = 1$, the ε_t are from the standard Cauchy distribution. The bootstrap estimators are approximated by Monte Carlo with size 200. The results in Table 9.1 are based on 100 simulations.

Table 9.1. Simulation mean of the standard deviation estimators
 [Adapted from Chatterjee (1986), by permission of IEEE]

a	Method [†]	$n = 30$		$n = 40$		$n = 50$		$n = 100$	
		$\hat{\theta}$	$\hat{\phi}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\theta}$	$\hat{\phi}$	$\hat{\theta}$	$\hat{\phi}$
1	PARA	0.16	0.04	0.14	0.10	0.13	0.07	0.14	0.02
	SIMU	0.15	0.33	0.27	0.28	0.14	0.31	0.09	0.09
	BOOT	0.25	0.73	0.40	0.45	0.25	0.53	0.17	0.13
1.2	PARA	0.18	0.12	0.14	0.10	0.15	0.10	0.13	0.02
	SIMU	0.14	0.35	0.13	0.38	0.12	0.31	0.08	0.09
	BOOT	0.26	0.57	0.34	0.62	0.39	0.40	0.15	0.10
1.6	PARA	0.21	0.11	0.15	0.08	0.14	0.06	0.09	0.02
	SIMU	0.22	0.32	0.16	0.15	0.13	0.10	0.11	0.41
	BOOT	0.47	0.57	0.30	0.33	0.49	0.56	0.11	0.09
1.8	PARA	0.20	0.07	0.14	0.09	0.14	0.06	0.09	0.04
	SIMU	0.27	0.20	0.21	0.34	0.15	0.23	0.08	0.27
	BOOT	0.40	0.43	0.34	0.47	0.31	0.40	0.23	0.18
2	PARA	0.14	0.07	0.17	0.07	0.15	0.05	0.09	0.04
	SIMU	0.30	0.28	0.19	0.26	0.18	0.27	0.09	0.27
	BOOT	0.56	0.79	0.23	0.46	0.23	0.29	0.23	0.21

[†]PARA, SIMU, and BOOT refer to the parametric asymptotic, simulated, and the bootstrap standard deviation estimates, respectively.

9.4.2 Linear regression with time series errors

In Chapter 7, we considered linear models

$$y_i = x_i' \beta + \varepsilon_i, \quad i = 1, \dots, n,$$

with independent errors ε_i . In applications, often the errors are related and can be modeled as a time series such as the ARMA(p, q). Thus, we consider the following linear regression model with time series errors:

$$y_t = x_t' \beta + \eta_t, \quad t = 1, \dots, n, \tag{9.27}$$

where β is a vector of regression parameters, x_t is a vector of deterministic explanatory variables, and $\{\eta_t, t = 0, \pm 1, \pm 2, \dots\}$ is an ARMA(p, q) series with mean 0. $\{y_t, t = 0, \pm 1, \pm 2, \dots\}$ itself is a time series. For simplicity, we assume that β is two-dimensional and model (9.27) can be written as

$$y_t = \beta_0 + \beta_1(x_t - \bar{x}) + \eta_t, \quad t = 1, \dots, n, \tag{9.28}$$

for a scalar explanatory variable x_t , where $\bar{x} = n^{-1} \sum_{t=1}^n x_t$.

Under model (9.28), the primary parameters of interest are β_0 and β_1 . Eriksson (1983) suggested that when $\{\eta_t, t = 0, \pm 1, \pm 2, \dots\}$ is ARMA(1, 0) (autoregressive of order 1) or ARMA(0, 1) (moving average of order 1), we can still estimate β_0 and β_1 by the least squares estimators

$$\hat{\beta}_0 = \frac{1}{n} \sum_{t=1}^n y_t \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{t=1}^n (x_t - \bar{x}) y_t}{\sum_{t=1}^n (x_t - \bar{x})^2}.$$

Assuming that

$$\eta_t = \theta \eta_{t-1} + \varepsilon_t, \quad t = 0, \pm 1, \pm 2, \dots, \quad (9.29)$$

where the ε_t are i.i.d. with mean 0 and variance σ^2 , and θ is unknown, Eriksson (1983) showed that the asymptotic variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ can be estimated by

$$v_0 = \frac{s^2(s^2 + \hat{c}_1)}{n(s^2 - \hat{c}_1)} \quad \text{and} \quad v_1 = \frac{12s^2(s^2 + \hat{c}_1)}{(n^3 - n)(s^2 - \hat{c}_1)},$$

respectively, where

$$s^2 = \frac{1}{n-2} \sum_{t=1}^n \hat{\eta}_t^2, \quad \hat{c}_1 = \frac{1}{n-1} \sum_{t=1}^{n-1} \hat{\eta}_t \hat{\eta}_{t+1},$$

and $\hat{\eta}_t = y_t - \hat{\beta}_0 - \hat{\beta}_1(x_t - \bar{x})$, $t = 1, \dots, n$. Similarly, if

$$\eta_t = \varepsilon_t + \phi \varepsilon_{t-1}, \quad t = 0, \pm 1, \pm 2, \dots, \quad (9.30)$$

then the variance estimators are

$$v_0 = \frac{s^2 + 2\hat{c}_1}{n} \quad \text{and} \quad v_1 = \frac{12(s^2 + 2\hat{c}_1)}{n^3 - n}.$$

Based on the normal approximation and the variance estimators v_0 and v_1 , we can obtain approximate level $1 - 2\alpha$ confidence intervals

$$[\hat{\beta}_k - z_{1-\alpha} \sqrt{v_k}, \hat{\beta}_k + z_{1-\alpha} \sqrt{v_k}], \quad k = 0, 1, \quad (9.31)$$

for β_0 and β_1 , respectively.

DeWet and van Wyk (1986) found that these confidence intervals can be improved by using the bootstrap. Since the errors η_t are dependent, we cannot directly apply the bootstrap methods introduced in Chapter 7. But we may apply the technique described in this chapter for ARMA(p, q) models. If the errors follow model (9.29), then ε_t can be estimated by

$$\hat{\varepsilon}_t = \hat{\eta}_t - \hat{\theta} \hat{\eta}_{t-1}, \quad t = 1, \dots, n,$$

where $\hat{\theta} = \hat{c}_1/s^2$. If the errors follow model (9.30), then ε_t can be estimated by

$$\hat{\varepsilon}_t = \hat{\eta}_t - \hat{\phi}\hat{\varepsilon}_{t-1}, \quad t = 1, \dots, n,$$

where $\hat{\phi}$ is defined by $\hat{c}_1(1 + \hat{\phi}^2) = s^2\hat{\phi}$ and $\hat{\varepsilon}_0$ is a random variable generated from $N(0, s^2(1 + \hat{\phi}^2)^{-1})$. In both cases, we can generate bootstrap data $\varepsilon_1^*, \dots, \varepsilon_n^*$ from the empirical distribution putting mass n^{-1} to $\hat{\varepsilon}_t - \bar{\varepsilon}$, $t = 1, \dots, n$, $\bar{\varepsilon} = n^{-1} \sum_{t=1}^n \hat{\varepsilon}_t$. Define $\eta_1^* = \hat{\eta}_1$,

$$\eta_t^* = \hat{\theta}\eta_{t-1}^* + \varepsilon_t^*, \quad t = 2, \dots, n,$$

when the errors η_t follow model (9.29); and

$$\eta_t^* = \varepsilon_t^* + \hat{\phi}\varepsilon_{t-1}^*, \quad t = 2, \dots, n,$$

when the errors follow model (9.30). Calculate $\hat{\beta}_0^*$, $\hat{\beta}_1^*$, v_0^* , and v_1^* , the bootstrap analogs of $\hat{\beta}_0$, $\hat{\beta}_1$, v_0 , and v_1 , respectively, using the bootstrap data

$$y_t^* = \hat{\beta}_0 + \hat{\beta}_1(x_t - \bar{x}) + \eta_t^*, \quad t = 1, \dots, n.$$

We can then obtain the bootstrap-t confidence interval for β_k by replacing the $z_{1-\alpha}$ in (9.31) with the $(1 - \alpha)$ th quantile of the bootstrap distribution of $(\hat{\beta}_k^* - \hat{\beta}_k)/\sqrt{v_k^*}$, $k = 0, 1$.

DeWet and van Wyk (1986) compared by simulation the bootstrap-t and the normal approximation 95% confidence intervals for β_1 . The true values of β_0 and β_1 were 0. The errors η_t were either ARMA(1, 0) or ARMA(0, 1) with ε_t distributed as $N(0, 1)$, $CN(\epsilon, \tau) = (1 - \epsilon)N(0, 1) + \epsilon N(0, \tau)$, or t_3 , the t-distribution with 3 degrees of freedom. The bootstrap estimators were calculated by Monte Carlo with size 200. The results shown in Table 9.2 were based on $n = 20$ and 500 simulations. It is clear that the coverage probability of the confidence interval based on normal approximation falls below the nominal level much quicker and more severely than that of the bootstrap-t confidence interval.

There is no theoretical justification for this bootstrap procedure. It is also of interest to apply the bootstrap in the case where the errors η_t follow a general ARMA(p, q) model. The main difficulty is the estimation of the θ_i and ϕ_j in the ARMA(p, q) model.

Sometimes we would like to test whether there is a serial correlation in the errors η_t in model (9.27). Srivastava (1987) and DeBeer and Swanepoel (1989) proposed some bootstrap procedures to estimate the null distribution of the Durbin-Watson test statistic for $\theta = 0$ versus $\theta > 0$ when (9.29) is assumed for the errors. The proposed procedures are more powerful and more robust to the deviations from normality than the procedure based on normal approximations.

Table 9.2. Simulation coverage probabilities (in %) of the 95% confidence intervals for β_1 [Adapted from DeWet and van Wyk (1986), by permission of Gordon and Breach Science Publishers]

		θ in ARMA(1, 0) model								
		-0.8	-0.6	-0.4	-0.2	0.0	0.2	0.4	0.6	0.8
1 [†]	A [‡]	93.6	90.8	92.8	90.0	90.2	86.4	82.8	81.8	68.0
	B	95.2	93.6	96.2	94.0	95.0	93.6	89.2	89.2	84.0
2	A	92.4	92.2	92.6	92.4	91.4	86.0	83.4	80.8	66.4
	B	95.0	93.6	94.2	94.8	94.6	91.4	91.4	90.4	79.8
3	A	92.6	92.6	93.4	93.4	92.4	86.6	83.2	80.0	64.8
	B	95.2	94.8	95.6	95.4	95.4	91.2	90.6	89.0	80.0
4	A	93.0	93.4	94.4	93.6	92.4	85.6	83.4	81.4	65.0
	B	95.8	96.0	95.4	95.8	96.6	92.0	90.2	90.0	79.0
5	A	95.4	93.6	93.0	92.4	90.0	85.8	83.6	79.4	67.8
	B	96.4	95.2	95.0	95.2	93.8	91.2	91.2	87.6	81.4
		ϕ in ARMA(0, 1) model								
		-0.8	-0.6	-0.4	-0.2	0.0	0.2	0.4	0.6	0.8
1	A	93.8	93.8	88.6	86.2	88.2	88.8	86.4	88.8	90.0
	B	95.0	95.8	92.0	93.6	93.8	94.2	94.8	96.0	95.8
2	A	94.4	91.6	90.6	87.6	86.0	87.6	87.4	87.6	88.2
	B	95.6	94.2	94.4	95.2	94.4	95.0	95.4	96.4	94.4
3	A	94.6	89.6	87.4	87.2	86.8	87.0	87.2	88.0	89.0
	B	95.0	93.2	93.4	95.0	94.6	96.0	95.6	96.8	95.2
4	A	94.2	92.0	90.2	87.6	89.6	87.0	89.0	89.2	87.4
	B	96.4	95.2	95.2	94.8	96.8	95.6	96.2	96.0	95.8
5	A	93.0	90.8	86.0	87.0	90.4	87.6	87.0	90.0	86.4
	B	95.2	94.4	92.0	93.6	96.6	95.6	96.4	97.2	95.0

[†] 1-5 refer to five distributions for ε_t : $N(0, 1)$, $CN(0.1, 9)$, $CN(0.1, 36)$, $CN(0.1, 81)$, and t_3 , respectively.

[‡] A and B refer to the normal approximation and bootstrap methods, respectively.

9.4.3 Dynamical linear regression

We now consider the dynamical linear regression model given by (9.24) in Example 9.6. Suppose that we observe (y_t, x_t) , $t = 1, \dots, n$, and y_0 , and we estimate A , B , and C by using the two-stage least squares estimators (Freedman, 1984) \hat{A} , \hat{B} , and \hat{C} , respectively. Let

$$r_t = y_t - \hat{A}y_t - \hat{B}y_{t-1} - \hat{C}x_t, \quad t = 1, \dots, n,$$

be residuals. Then we can draw bootstrap data $(x_1^{*'}, \varepsilon_1^{*'}), \dots, (x_n^{*'}, \varepsilon_n^{*'})$ i.i.d. from the empirical distribution putting mass n^{-1} to (x_t', r_t') , $t = 1, \dots, n$. The bootstrap distribution estimators of $\hat{A} - A$, $\hat{B} - B$, and $\hat{C} - C$ can be obtained by using the bootstrap analogs \hat{A}^* , \hat{B}^* , and \hat{C}^* based on $(y_1^{*'}, x_1^{*'}), \dots, (y_n^{*'}, x_n^{*'})$ with

$$y_t^* = (I - \hat{A})^{-1}(\hat{B}y_{t-1}^* + \hat{C}x_t^* + \varepsilon_t^*).$$

Consistency and other properties of these bootstrap estimators were established by Freedman (1984). Some empirical results about these estimators can be found in Freedman and Peters (1984).

When the x_t are deterministic, this bootstrap can still be used with the following modification: the ε_t^* are generated from the empirical distribution of r_1, \dots, r_n and

$$y_t^* = (I - \hat{A})^{-1}(\hat{B}y_{t-1}^* + \hat{C}x_t + \varepsilon_t^*).$$

There are many other applications of the jackknife and bootstrap to various types of dynamical linear regression models raised in econometrics. See the review provided by Veall (1989).

9.5 Stationary Processes

There are dependent data that cannot be modeled as one of the time series models described in the previous sections. It is therefore of interest to develop some general jackknife and bootstrap procedures that can be applied to general stationary processes with short-range dependence. In this section, we discuss two such procedures, the *moving block* and the *circular block*, and study their consistency and accuracy.

9.5.1 Moving block and circular block

Carlstein (1986) and Shi (1986b) proposed the use of nonoverlapping subseries values of a statistic to estimate its variance. A more sophisticated method is the moving block that was discussed in Section 9.1.

Consider a statistic $T_n = T_n(X_1, \dots, X_n)$ that can be expressed as a functional of the empirical p -dimensional marginal distribution defined by

$$F_n^{(p)} = \frac{1}{n-p+1} \sum_{t=1}^{n-p+1} \delta_{Z_t},$$

where δ_{Z_t} is the point mass at $Z_t = (X_t, \dots, X_{t+p-1})'$, i.e., $T_n = T(F_n^{(p)})$ for some given functional T . Note that if $p = 1$, then $F_n^{(1)}$ is simply the

empirical distribution of X_1, \dots, X_n . A simple example of T_n is the least squares estimator $\hat{\theta}$ of θ in the autoregressive model of order 1 (Example 9.4):

$$\hat{\theta} = \frac{\iint xy dF_n^{(2)}(x, y)}{\iint x^2 dF_n^{(2)}(x, y)}$$

is a functional of $F_n^{(2)}$.

Künsch (1989) introduced a class of variance estimators for T_n , which includes the moving block jackknife variance estimator (9.6) as a special case, by downweighting blocks of p -tuples in the marginal empirical distribution. Let $w_n(i)$ be weights satisfying $0 \leq w_n(i) \leq 1$ and $w_n(i) > 0$ if and only if $1 \leq i \leq h$, where h is the length of the downweighted block. Define

$$F_{n-h,i}^{(p)} = \frac{1}{n - \|w_n\|_1} \sum_{t=1}^n [1 - w_n(t-i+1)] \delta_{Z_t}$$

and

$$T_{n-h,i} = T(F_{n-h,i}^{(p)}), \quad i = 1, \dots, n-h+1,$$

where $\|w_n\|_1 = \sum_{t=1}^h w_n(t)$. Then the proposed variance estimator is

$$\frac{(n - \|w_n\|_1)^2}{n(n-h+1)\|w_n\|_2^2} \sum_{i=1}^{n-h+1} \left(T_{n-h,i} - \frac{1}{n-h+1} \sum_{i=1}^{n-h+1} T_{n-h,i} \right)^2, \quad (9.32)$$

where $\|w_n\|_2^2 = \sum_{t=1}^h [w_n(t)]^2$. If $w_n(t) = I\{\frac{1}{2} \leq t \leq \frac{1}{2}h\}$, then the estimator (9.32) reduces to the moving block jackknife variance estimator (9.6).

Assume that T_n converges almost surely to $T(F^{(p)})$, where $F^{(p)}$ is the marginal distribution of X_1, \dots, X_p . We can use the moving block bootstrap described in Section 9.1 to estimate the distribution of $T_n - T(F^{(p)})$ in the general case of $p > 1$. Define the blocks

$$B_t = \{Z_t, Z_{t+1}, \dots, Z_{t+h-1}\}, \quad t = 1, \dots, n-h+1.$$

Let B_1^*, \dots, B_k^* be i.i.d. from $\{B_1, \dots, B_{n-h+1}\}$ and $T_\ell^* = T(F_\ell^{(p)*})$, where $\ell = kh$ and

$$F_\ell^{(p)*} = \frac{1}{\ell} \sum_{i=1}^k \sum_{Z_t \in B_i^*} \delta_{Z_t}.$$

The distribution of $\sqrt{n}[T_n - T(F^{(p)})]$ can then be estimated by the bootstrap distribution of $\sqrt{\ell}(T_\ell^* - T_n)$.

The asymptotic properties of the moving block jackknife and bootstrap estimators were discussed by Künsch (1989). In particular, he proved that

the moving block jackknife variance estimator and the moving block bootstrap distribution estimator are consistent under some heavy conditions that are not easy to verify. Empirical results were also provided to compare the moving block jackknife and bootstrap variance estimators with the variance estimators obtained by using Carlstein's subseries method and the asymptotic method, when T_n is the least squares estimator $\hat{\theta}$ in Example 9.4. Shi and Liu (1992) showed the strong consistency of the moving block jackknife variance estimator and the moving block bootstrap variance and distribution estimators for functions of the sample mean based on stationary m -dependent or ϕ -mixing observations (see Example 9.7). Bühlmann (1994) applied the block bootstrap to the empirical process of a stationary sequence.

There is one small defect in this moving block procedure; that is, when $T_n = \bar{X}_n = n^{-1} \sum_{t=1}^n X_t$,

$$E_* \bar{X}_\ell^* = E_* \left(\frac{1}{h} \sum_{X_t \in B_1^*} X_t \right) = \frac{1}{(n-h+1)h} \sum_{i=1}^{n-h+1} \sum_{t=1}^h X_{i+t-1},$$

which is not exactly equal to \bar{X}_n . This does not lead to the inconsistency of the bootstrap estimators since $E_* \bar{X}_\ell^* - \bar{X}_n$ converges to 0 at a certain rate, but it destroys the second order accuracy property of the bootstrap distribution estimators. Lahiri (1991, 1992b) suggested recentering the bootstrap distribution at $E_* \bar{X}_\ell^*$, and we shall discuss his result in Section 9.5.3.

Noting that the reason $E_* \bar{X}_\ell^* \neq \bar{X}_n$ is that, in the bootstrap sampling, the first and last few observations in the series do not have the same chance to be drawn as the observations in the middle part of the series, Politis and Romano (1992a) and Shao and Yu (1993) proposed a circular block method by wrapping the observations X_1, \dots, X_n around in a circle and then generating consecutive blocks of bootstrap data from the circle. More precisely, let $Z_t = (X_t, \dots, X_{t+p-1})'$,

$$\tilde{Z}_t = \begin{cases} Z_t & t = 1, \dots, n, \\ Z_{t-n} & t = n+1, \dots, n+h-1, \end{cases}$$

and define the following blocks of size h :

$$\tilde{B}_t = \{\tilde{Z}_t, \dots, \tilde{Z}_{t+h-1}\}, \quad t = 1, \dots, n.$$

The circular block bootstrap data $\tilde{B}_1^*, \dots, \tilde{B}_k^*$ are then generated i.i.d. from $\{\tilde{B}_1, \dots, \tilde{B}_n\}$. This method has the following property:

$$P_* \{\tilde{Z}_t^* = Z_s\} = n^{-1}$$

for any t and s , which implies $E_* \bar{X}_\ell^* = \bar{X}_n$.

A circular block jackknife can also be defined by deleting the blocks $\tilde{B}_1, \dots, \tilde{B}_n$ consecutively.

9.5.2 Consistency of the bootstrap

In this section, we discuss the kinds of stationary processes in which the bootstrap distribution estimators are consistent when $T_n = \bar{X}_n$. Because of $E_*\bar{X}_\ell^* \neq \bar{X}_n$ for the moving block bootstrap, its consistency requires stronger conditions than the circular block bootstrap and is also much harder to prove. Thus, we focus on the circular block bootstrap.

Let $\ell = kh$, \bar{X}_n be the sample mean of a stationary process with mean μ and a finite variance, and \bar{X}_ℓ^* be the sample mean of the circular block bootstrap data. Shao and Yu (1993) proved that, as $n \rightarrow \infty$ and $k \rightarrow \infty$,

$$\sup_x |P_*\{\sqrt{\ell}(\bar{X}_\ell^* - \bar{X}_n) \leq x\} - P\{\sqrt{n}(\bar{X}_n - \mu) \leq x\}| \rightarrow_{a.s.} 0, \quad (9.33)$$

provided that the following conditions hold:

$$\sqrt{n}(\bar{X}_n - \mu) \rightarrow_d N(0, \tau) \quad \text{and} \quad \text{var}(\sqrt{n}\bar{X}_n) \rightarrow \tau$$

for some $\tau > 0$;

$$\begin{aligned} \frac{1}{nh} \sum_{j=0}^{n-1} \left[\left(\sum_{i=1}^h (X_{i+j} - \mu) \right)^2 - E \left(\sum_{i=1}^h (X_i - \mu) \right)^2 \right] &\rightarrow_{a.s.} 0; \\ \frac{1}{nh} \sum_{j=0}^{n-1} \left(\sum_{i=1}^h (X_{i+j} - \mu) \right)^2 I \left\{ \left(\sum_{i=1}^h (X_{i+j} - \mu) \right)^2 \geq \epsilon \ell \right\} &\rightarrow_{a.s.} 0 \end{aligned} \quad (9.34)$$

for any $\epsilon > 0$; and

$$h \rightarrow \infty \quad \text{and} \quad h(\log n)^4/n \rightarrow 0. \quad (9.35)$$

They also showed that result (9.33) still holds when conditions (9.34) and (9.35) are replaced by

$$\frac{1}{n} \sum_{j=0}^{n-1} \left[I \left\{ \frac{1}{\sqrt{h}} \sum_{i=1}^h (X_{i+j} - \mu) \leq x \right\} - P \left\{ \frac{1}{\sqrt{h}} \sum_{i=1}^h (X_i - \mu) \leq x \right\} \right] \rightarrow_{a.s.} 0$$

for each fixed x , $h \rightarrow \infty$, and $h/n \rightarrow 0$.

In what follows, we show some examples of stationary processes for which (9.33) holds. The proofs are given in Shao and Yu (1993).

Example 9.7. ϕ -mixing sequences. Let $\{X_t, t = 1, 2, \dots\}$ be a stationary sequence of random variables. Let \mathcal{F}_t^s be the σ -field generated by the random variables $\{X_i, t \leq i \leq s\}$,² $1 \leq t \leq s$, and

$$\phi(n) = \sup_{t \geq 1} \sup_{A \in \mathcal{F}_1^t, B \in \mathcal{F}_{n+t}^\infty} |P(B|A) - P(B)|, \quad n = 1, 2, \dots$$

²A σ -field \mathcal{F} is said to be generated by a set of random variables $\{X, Y, Z, \dots\}$ if \mathcal{F} is the smallest σ -field such that the random variables X, Y, Z, \dots are \mathcal{F} -measurable.

The sequence $\{X_t, t = 1, 2, \dots\}$ is said to be ϕ -mixing if $\phi(n) \rightarrow 0$ as $n \rightarrow \infty$. Assume that $EX_1 = \mu$, $\text{var}(X_1) < \infty$, $\text{var}(n\bar{X}_n) \rightarrow \infty$, h and n/h are nondecreasing as n increases, $h \rightarrow \infty$ and $h(\log n)^\epsilon/n \rightarrow 0$ for some $\epsilon > 0$, and $\sum_{n=1}^{\infty} \sqrt{\phi(2^n)} < \infty$. Then (9.33) holds.

Example 9.8. α -mixing sequences. A sequence of stationary random variables $\{X_t, t = 1, 2, \dots\}$ is said to be α -mixing if $\alpha(n) \rightarrow 0$ as $n \rightarrow \infty$, where

$$\alpha(n) = \sup_{t \geq 1} \sup_{A \in \mathcal{F}_1^t, B \in \mathcal{F}_{n+t}^{\infty}} |P(A \cap B) - P(A)P(B)|, \quad n = 1, 2, \dots$$

Assume that $EX_1 = \mu$, $E|X_1|^{2+\delta} < \infty$ for some $\delta > 0$, h and n/h are nondecreasing as n increases, $h \rightarrow \infty$ and $h \leq n^{1-\epsilon}$ for some $\epsilon > 0$, $\alpha(n) \leq cn^{-r}$ for some $c > 0$ and $r > (2 + \delta)/\delta$, and

$$\text{var}(X_1) + 2 \sum_{t=2}^{\infty} \text{cov}(X_1, X_t) > 0.$$

Then (9.33) holds. If the last two conditions are changed to $\alpha(n) \leq cn^{-r}$ for some $c > 0$ and $r > 2(2 + \delta)/\delta$ and $\text{var}(n\bar{X}_n) \rightarrow \infty$, then (9.33) still holds.

Example 9.9. ρ -mixing sequences. A sequence of stationary random variables $\{X_t, t = 1, 2, \dots\}$ is said to be ρ -mixing if $\rho(n) \rightarrow 0$ as $n \rightarrow \infty$, where

$$\rho(n) = \sup_{t \geq 1} \sup_{Y \in L_2(\mathcal{F}_1^t), Z \in L_2(\mathcal{F}_{n+t}^{\infty})} \left| \frac{\text{cov}(Y, Z)}{\sqrt{\text{var}(Y)\text{var}(Z)}} \right|, \quad n = 1, 2, \dots$$

and $L_2(\mathcal{F})$ is the class of \mathcal{F} -measurable random variables that are squared integrable. Assume that $EX_1 = \mu$, $\text{var}(X_1) < \infty$, $\text{var}(n\bar{X}_n) \rightarrow \infty$, h and n/h are nondecreasing as n increases, $h \rightarrow \infty$ and $h(\log n)^4/n \rightarrow 0$, and $\sum_{n=1}^{\infty} \sqrt{\rho(2^n)} < \infty$. Then (9.33) holds.

Shao and Yu (1993) also proved the consistency of the bootstrap distribution estimator for the empirical distribution process.

9.5.3 Accuracy of the bootstrap

As we discussed earlier, $E_*\bar{X}_\ell^* \neq \bar{X}_n$ for the moving block bootstrap. For estimating the distribution of $\bar{X}_n - \mu$, Lahiri (1991, 1992b) considered using the bootstrap distribution of $\bar{X}_\ell^* - E_*\bar{X}_\ell^*$ instead of $\bar{X}_\ell^* - \bar{X}_n$. He proved the following result. Let $\{X_1, X_2, \dots\}$ be a sequence of stationary random

variables with mean μ and a finite variance, and let $\{\mathcal{D}_t, t = 0, \pm 1, \pm 2, \dots\}$ be a given sequence of σ -fields. Assume the following conditions:

- (1) $E|X_1|^4 < \infty$ and $\text{var}(\sqrt{n}\bar{X}_n)$ has a positive limit.
- (2) There is a positive constant d such that for $n, m = 1, 2, \dots$, with $m > d^{-1}$, there is a \mathcal{F}_{n-m}^{n+m} -measurable random variable $W_{n,m}$ satisfying

$$E|X_n - W_{n,m}| \leq d^{-1}e^{-dm}$$

and

$$E(|W_{n,a_n}|^{12}I\{|W_{n,a_n}| < n^{1/4}\}) < d^{-1},$$

where \mathcal{F}_t^s is the smallest σ -field containing $\{\mathcal{D}_i, t \leq i \leq s\}$, $t \leq s$, and $\{a_n\}$ is a sequence of real numbers satisfying $a_n = O((\log n)^{1+1/d})$ and $\log n = o(a_n)$.

- (3) There is $d > 0$ such that for all $m, n = 1, 2, \dots$, $A \in \mathcal{F}_{-\infty}^n$, and $B \in \mathcal{F}_{n+m}^\infty$,

$$|P(A \cap B) - P(A)P(B)| \leq d^{-1}e^{-dm}.$$

- (4) There is $d > 0$ such that for all $m, n = 1, 2, \dots$, $d^{-1} < m < n$, and all u with $|u| \geq d$,

$$E\left|E\left[\exp\left(iu \sum_{t=-m}^m X_{n-t}\right) \middle| \mathcal{D}_j : j \neq n\right]\right| \leq e^{-d},$$

where $i = \sqrt{-1}$.

- (5) There is $d > 0$ such that for all $m, n, q = 1, 2, \dots$, and $A \in \mathcal{F}_{m-q}^{m+q}$,

$$E|P(A|\mathcal{D}_j : j \neq n) - P(A|\mathcal{D}_j : 0 < |n-j| \leq m+q)| \leq d^{-1}e^{-dm}.$$

Assume further that the block size h is selected to be of the order n^a with $0 < a < \frac{1}{4}$. Then

$$\sup_x \left| P_* \left\{ \frac{\bar{X}_\ell^* - E_* \bar{X}_\ell^*}{[\text{var}_*(\bar{X}_\ell^*)]^{1/2}} \leq x \right\} - P \left\{ \frac{\bar{X}_n - \mu}{[\text{var}(\bar{X}_n)]^{1/2}} \leq x \right\} \right| = o(n^{-1/2}) \text{ a.s.}$$

The result for the circular block bootstrap was obtained by Politis and Romano (1992a) under similar conditions. Lahiri (1991, 1992b) also proved a similar result for the sample mean of a multivariate stationary process and for T_n being a smooth function of the sample mean. He also extended some results to nonstationary processes.

In applications, the selection of the block size h is crucial. But it is not clear how to choose the block size. Léger, Politis and Romano (1992) studied the effect of different h on the bias of the moving block bootstrap

variance estimator for \bar{X}_n in the following model:

$$X_t = \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2},$$

where the ε_t are i.i.d. $N(0, 1)$. When $n = 100$, it was found that the bootstrap estimator with $h = 1$ (the bootstrap estimator for independent data) seriously underestimates the true variance; h near 10 provides the most accurate estimate, but a large value of h ($h \geq 20$) produces an estimate with a larger bias than the choice of $h = 1$.

9.5.4 Remarks

As we discussed in Chapter 1, an appropriate bootstrap procedure should mimic the true sampling mechanism that generates the original data. For example, when the original data are observed from a stationary process, the bootstrap data should also be stationary. However, it can be shown that the moving block bootstrap method does not have this property. Politis and Romano (1990, 1994) proposed a bootstrap method by resampling blocks of random lengths, which produces stationary bootstrap data. Their method can be described as follows. First, we randomly select a bootstrap sample X_1^* from the original n observations. Suppose that $X_1^* = X_j$. The second bootstrap sample is equal to X_{j+1} with probability $1 - p$; otherwise, it is picked at random from the original n observations, where p is a given positive constant. Having the i th bootstrap sample $X_i^* = X_{i^*}$, we take the next bootstrap sample according to

$$X_{i+1}^* = \begin{cases} X_{i^*} & \text{with prob. } 1 - p \\ \text{a random sample from } \{X_1, \dots, X_n\} & \text{with prob. } p, \end{cases}$$

$i = 1, \dots, n-1$. For the case of $T_n = \bar{X}_n$, it can be shown that $E_* \bar{X}_n^* = \bar{X}_n$. Some other properties of this method were discussed in Politis and Romano (1990, 1994). The selection of the constant p is the major barrier to the application of this bootstrap method.

Setting confidence intervals or bands for parameters associated with the infinite-dimensional distribution of the process $\{X_1, X_2, \dots\}$ is an interesting topic that has been discussed in recent years. Examples of such parameters are the spectral density function of $\{X_1, X_2, \dots\}$ and its functionals. Swanepoel and van Wyk (1986) proposed approximating a stationary process by an autoregressive model of finite order and then using the bootstrap-t method to set a confidence band for the spectral density function of the autoregressive process, which can be taken as an approximate confidence band for the spectral density function of the original process. The bootstrap method introduced for the autoregressive models can be applied for this purpose. Ramos (1988) assumed that the process is Gaussian and used a method similar to the parametric bootstrap to assess the

sampling distributions of estimators of functionals of the spectral density. Franke and Härdle (1992) applied the idea of Härdle and Bowman (1988) for bootstrapping a nonparametric regression to the problem of kernel-smoothed spectral density estimation. Politis and Romano (1992b, 1993b) introduced a generalized moving block method to set a confidence interval for the spectral density function evaluated at a fixed point. Politis, Romano and Lai (1992) proposed bootstrap confidence bands for spectra and cross-spectra of a stationary and weakly dependent time series.

There are applications of the jackknife and bootstrap to other dependent data problems. For example, Gray and Schucany (1972) introduced some generalized jackknife methods to reduce the bias and to estimate the variance for an estimator of a parameter related to a piecewise continuous stochastic process on an interval $[a, b]$. Lele (1991) showed the consistency of a modified jackknife variance estimator for the maximum pseudolikelihood estimator of the data from a spatial process. Stoffer and Wall (1991) proposed a bootstrap method for assessing the precision of Gaussian maximum likelihood estimators of the parameters of linear state space models, which include the ARMA(p, q) models as special cases. Cressie (1991) applied the jackknife and bootstrap to spatial lattice data. Mykland (1992) discussed the bootstrap approximation to the distribution of a sequence of martingales that can be applied to approximate the distribution of the estimators in asymptotic ergodic differential equation models. Burman, Chow and Nolan (1994) used a block cross-validation method to estimate the expected prediction error for a general stationary process.

9.6 Conclusions and Discussions

- (1) The applications of the jackknife and bootstrap to dependent data are not straightforward. Modifications to the procedures that work well for independent data are necessary. There are in general two types of resampling methods for dependent data. The first type of method takes resamples from appropriately defined residuals, whereas the second type applies resampling to groups or blocks of the original data to keep the dependence structure of the data. The resampling methods based on residuals are usually based on some model assumptions and, therefore, are not robust against the violation of the model assumptions. The other type of method is less model-dependent, but the applications of the resampling methods are still not automatic, e.g., one has to carefully examine the data in order to determine the size of the blocks in the moving block bootstrap procedure. It should be noted, however, that the above procedures may not work for some long range dependent data (Lahiri, 1993b).

- (2) Although it is much more difficult to apply the jackknife and bootstrap to dependent data than to apply them to independent data, there are still needs and advantages of using the resampling methods. For example, the asymptotic variance of an estimator based on dependent data may be very difficult to estimate by using the traditional asymptotic approach; applications of the jackknife and bootstrap do not require difficult analytic derivations.
- (3) Once a valid resampling plan is used, the jackknife and bootstrap procedures can be carried out in a manner similar to the case of independent data, and their properties are also similar: the bootstrap distribution estimator for a standardized or studentized statistic is more accurate than the estimator based on normal approximation; bootstrap-t confidence intervals are usually more accurate than confidence intervals based on the normal approximation or the hybrid bootstrap; and model selection using the bootstrap based on residuals provides consistent results, provided that the bootstrap sample size is much smaller than the size of the original data. Davison and Hall (1993) addressed the importance of studentizing for implementing the block bootstrap.
- (4) The most efficient bootstrap method is the one based on residuals, but it requires some model assumptions. The moving block bootstrap is better than the grouped bootstrap, while the circular block bootstrap seems better than the moving block bootstrap.

Chapter 10

Bayesian Bootstrap and Random Weighting

Bayesian bootstrap and *random weighting* are two variants of the bootstrap. The former is aimed at simulating the posterior distribution of a parameter, given the observations. The latter is still used to estimate the sampling distribution of a random variable but adopts a resampling plan different from the bootstrap: instead of generating resamples from data, the random weighting method assigns a random weight to each observation. Random weighting can be regarded as a smoothing for the bootstrap. These two methods are pooled together and introduced in one chapter because their prime forms are exactly the same and they can be unified into a general framework.

10.1 Bayesian Bootstrap

Let $X = (X_1, \dots, X_n)$ be an i.i.d. sample from a univariate distribution F and θ be a vector-valued functional of F . Suppose that θ is estimated by $\hat{\theta}_n = \hat{\theta}_n(X)$. Let $\mathfrak{R}(X, F)$ be a functional of X and F , e.g., $\mathfrak{R}(X, F) = \hat{\theta}_n - \theta$. In frequentist statistical analysis, we are often concerned about the sampling distribution of $\mathfrak{R}(X, F)$, where θ is considered to be unknown but nonrandom. In Bayesian statistical analysis, however, we assume that θ or F itself is random and has a prior distribution. After X is taken from F (considered as a conditional distribution of X given θ or F), Bayesian analysis is carried out based on the posterior distribution of $\mathfrak{R}(X, F)$, i.e., the conditional distribution of $\mathfrak{R}(X, F)$ given X .

When F is in a parametric family and θ is the index of the family,

the posterior distribution of $\mathfrak{R}(X, F)$ can be derived explicitly only when the prior distribution has a specific form (e.g., Berger, 1985). In general, posterior distributions are expressible only in terms of complicated analytical functions; we often cannot readily compute the marginal densities and moments of the posterior distribution. Because of this, normal approximation to the posterior distribution is often suggested. As in the context of frequentist statistics, the normal approximation lacks high order accuracy, and the posterior variances are sometimes difficult to calculate. Rubin (1981) suggested a nonparametric Bayesian bootstrap method to address this problem under a specific assumption on the prior distribution of F . We introduce in Section 10.1.1 the basic idea and asymptotic properties of Rubin's method. Some other developments in the Bayesian bootstrap methodology will be discussed in Sections 10.1.2 and 10.1.3.

10.1.1 Bayesian bootstrap with a noninformative prior

Let U_1, \dots, U_{n-1} be $n - 1$ i.i.d. random variables from $U(0, 1)$, the uniform distribution on $[0, 1]$, $U_{(1)} < \dots < U_{(n-1)}$ be the order statistics of U_1, \dots, U_{n-1} , $U_{(0)} = 0$, $U_{(n)} = 1$, $w_i = U_{(i)} - U_{(i-1)}$, $i = 1, \dots, n$, and

$$D_n(x) = \sum_{i=1}^n w_i I\{X_i \leq x\}. \quad (10.1)$$

Rubin (1981) suggested that we use the conditional distribution of $\mathfrak{R}(X, D_n)$ (given X) to approximate the posterior distribution of $\mathfrak{R}(X, F)$.

In practice, it is seldom that the conditional distribution of $\mathfrak{R}(X, D_n)$ can be calculated analytically. As for the bootstrap, the Monte Carlo simulation can be used to approximate the Bayesian bootstrap distribution of $\mathfrak{R}(X, D_n)$. Therefore, an integral Bayesian bootstrap procedure can be summarized as follows (Lo, 1987):

- (1) Simulate $n - 1$ i.i.d. $U(0, 1)$ random variables (independent of X) and denote their order statistics by $0 = U_{(0)} < U_{(1)} < \dots < U_{(n-1)} < U_{(n)} = 1$. Let $w_i = U_{(i)} - U_{(i-1)}$, $i = 1, \dots, n$. Construct a random discrete distribution function D_n with weight w_i at X_i for $i = 1, \dots, n$.
- (2) Repeat step (1) independently B times to obtain D_{n1}, \dots, D_{nB} and compute $\mathfrak{R}_b = \mathfrak{R}(X, D_{nb})$, $b = 1, \dots, B$.
- (3) The empirical distribution function of $\mathfrak{R}_1, \dots, \mathfrak{R}_B$ approximates the conditional distribution of $\mathfrak{R}(X, D_n)$ for large B .

From the above introduction, it is not clear what prior distribution of F is used and why the gaps of ordered $U(0, 1)$ random variables are used to construct D_n . The second question is closely associated with the first one.

Rubin (1981) justified that if X_1 has a discrete distribution of the form

$$P\{X_1 = d_i|\theta\} = \theta_i, \quad i = 1, \dots, k, \quad \sum_{i=1}^k \theta_i = 1,$$

$\theta = (\theta_1, \dots, \theta_k)'$ and the prior distribution of θ is proportional to $\prod_{i=1}^k \theta_i^{-1}$ (a noninformative prior), then the posterior distribution of θ is in agreement with the Bayesian bootstrap distribution. For example, if $k = 2$, $d_1 = 1$, and $d_2 = 0$, then both the posterior distribution and the Bayesian bootstrap distribution of θ_1 are the Beta distribution with parameters n_1 and $n - n_1$, where $n_1 = \sum_{i=1}^n X_i$. Therefore, the Bayesian bootstrap method simulates the posterior distribution of a parameter when it has a noninformative prior distribution.

Some asymptotic properties of the Bayesian bootstrap distribution were studied by Lo (1987). For convenience of exposition, we assume that the X_i are one-dimensional. Ferguson (1973) suggested that the prior knowledge of F can be represented by a measure α on the real line \mathbb{R} such that, for any partition of \mathbb{R} into disjoint intervals B_1, \dots, B_k , the joint prior distribution of p_1, \dots, p_k , where $p_i = P\{X_1 \in B_i\}$, is $\text{Diri}(\alpha_1, \dots, \alpha_k)$, the Dirichlet distribution with parameters $\alpha_i = \alpha(B_i)$, $i = 1, \dots, k$ (e.g., Hogg and Craig, 1970); that is, the joint density of (p_1, \dots, p_k) is given by

$$\pi(p_1, \dots, p_k) = \frac{\Gamma(\alpha(\mathbb{R}))}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k p_i^{\alpha_i-1}, \quad p_i \geq 0, \quad \sum_{i=1}^k p_i = 1,$$

where $\Gamma(x)$ is the Gamma function. This actually assigns a noninformative Dirichlet prior to F . If such a prior is specified, then the posterior knowledge given X can be summarized by the measure $\bar{\alpha} = \alpha + \sum_{j=1}^n \delta_{X_j}$, i.e., the joint posterior distribution of p_1, \dots, p_k , given X , is $\text{Diri}(\bar{\alpha}(B_1), \dots, \bar{\alpha}(B_k))$. It can be shown that the conditional distribution of $D_n(x)$ given X can be represented by the measure $\sum_{j=1}^n \delta_{X_j}$. Thus, Theorem 1 of Ferguson (1973) implies that one may view the conditional distribution of $D_n(x)$ as the posterior distribution of $F(x)$ when F has a “flat” Dirichlet prior. It is expected that the Bayesian bootstrap distribution is a consistent estimator of the posterior distribution. In fact, Lo (1987) established the following result. Let $\mathfrak{R}(X, F) = \theta(F) - \bar{X}_n$, where $\theta(F) = \int x dF(x)$ and $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ is the sample mean.

Theorem 10.1. Suppose that F_0 is the true distribution of X and F is assumed to have a Dirichlet prior with measure α . If $\int x^2 dF_0(x) < \infty$ and $\int x^2 d\alpha < \infty$, then the Bayesian bootstrap approximation to the posterior distribution of $\mathfrak{R}(X, F) = \theta(F) - \bar{X}_n$ is strongly consistent, i.e., for almost all given X , the conditional distribution of $\sqrt{n}[\theta(D_n) - \bar{X}_n]$ and the posterior

distribution of $\sqrt{n}[\theta(F) - \bar{X}_n]$ tend to $N(0, \sigma^2(F_0))$, where $\sigma^2(F)$ is the variance of F .

Proof. Note that

$$\mathfrak{R}(X, D_n) = \theta(D_n) - \bar{X}_n = \sum_{i=1}^n w_i(X_i - \bar{X}_n)$$

and (w_1, \dots, w_n) has the same distribution as $(Z_1, \dots, Z_n) / \sum_{j=1}^n Z_j$, where Z_1, \dots, Z_n are i.i.d. exponential random variables with mean 1 and independent of X . Thus, we have

$$\mathfrak{R}(X, D_n) = \sum_{i=1}^n Z_i(X_i - \bar{X}_n) \Big/ \sum_{j=1}^n Z_j.$$

It is easy to show that $n^{-1} \sum_{j=1}^n Z_j \rightarrow_{a.s.} 1$. By the central limit theorem for the weighted sum of random variables (see Appendix A.8 or Chow and Teicher, 1988, p. 308),

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i(X_i - \bar{X}_n) \rightarrow_d N(0, \sigma^2(F_0)) \text{ conditioned on } X \text{ a.s.}$$

By introducing a “posterior Dirichlet process”, Lo (1987) showed that the posterior distribution of $\sqrt{n}[\theta(F) - \bar{X}_n]$ also tends to $N(0, \sigma^2(F_0))$, conditioned on X , if $\int x^2 d\alpha < \infty$. We omit the details that involve some knowledge of stochastic processes. \square

Lo (1987) also proved a similar result for the variance functional $\theta = \sigma^2(F) = \frac{1}{2} \iint (x - y)^2 dF(x) dF(y)$.

Weng (1989) discussed the second order asymptotic property of the Bayesian bootstrap approximation to the posterior distribution of the mean functional. He developed Edgeworth expansions for the posterior distribution and the Bayesian bootstrap distribution of the normalized mean functional using the Edgeworth expansions for linear functions of uniform order statistics obtained by van Zwet (1979). By comparing these two expansions, he concluded that, if $\int |x|^3 dF_0(x) < \infty$ and $\int |x|^3 d\alpha < \infty$, then the Bayesian bootstrap approximation is second order accurate and thus better than the normal approximation. An alternative method to derive the Edgeworth expansion for the weighted sum of i.i.d. random variables is discussed in Section 10.2.3.

Lo (1987) and Banks (1988) also studied smoothing of the Bayesian bootstrap distribution. Lo (1988, 1993a,b) applied the basic idea of the Bayesian bootstrap to the estimation of finite populations, survival analysis, and weighted sampling.

After obtaining the Bayesian bootstrap approximation to the posterior distribution of θ , we can use the methods discussed in Chapter 4 to construct credible intervals for θ . For example, suppose that $\Re(X, F) = \sup_x |F(x) - F_n(x)|$, where F_n is the empirical distribution of X_1, \dots, X_n . For a sample of size $n = 15$ (3.39, 3.30, 2.81, 3.03, 3.44, 3.07, 3.00, 3.43, 3.36, 3.13, 3.12, 2.74, 2.76, 2.88, 2.96), 1000 Bayesian bootstrap Monte Carlo replications yield the following level 95% hybrid Bayesian bootstrap credible bands: $F_n \pm 0.31$ (Lo, 1987).

10.1.2 Bayesian bootstrap using prior information

When there is useful prior information, it is desired to incorporate the prior information into the Bayesian bootstrap procedure. Unfortunately, in general it is difficult to express given prior information as a Dirichlet prior used in Rubin's Bayesian bootstrap. Stewart (1986) suggested a variation of Rubin's Bayesian bootstrap to incorporate partial information about the parameter of interest. We introduce this method in the special case of $\theta = E[g(X_1)]$, where g is a given vector-valued function.

Let a_1, \dots, a_{m-1} be a set of equally spaced points on the real line and $\phi_i = F(a_i) - F(a_{i-1})$, $i = 1, \dots, m$, where $F(a_0) = 0$ and $F(a_m) = 1$. Assume that the joint distribution of (ϕ_1, \dots, ϕ_m) is $\text{Dir}(\alpha_1, \dots, \alpha_m)$ (the α_i will be specified later). Suppose that the prior distribution of F can be summarized by the vector (ϕ_1, \dots, ϕ_m) and the parameter θ can be approximated by the parameter value ϑ corresponding to a discrete distribution with probability mass given by $P\{X_1 = \bar{a}_i\} = \phi_i$, $i = 1, \dots, m$, where \bar{a}_i is some typical value representing the interval $(a_{i-1}, a_i]$. For example, since $\theta = E[g(X_1)]$, it can be approximated by $\vartheta = \sum_{i=1}^m \phi_i g(\bar{a}_i)$. When m is large and F has a compact support, the choice of \bar{a}_i is not particularly critical and usually we can take $\bar{a}_i = (a_{i-1} + a_i)/2$.

The posterior distribution of (ϕ_1, \dots, ϕ_m) is $\text{Dir}(\bar{\alpha}_1, \dots, \bar{\alpha}_m)$, where $\bar{\alpha}_i$ is the sum of α_i and the number of observations falling in the interval $(a_{i-1}, a_i]$. Hence, the posterior distribution of ϑ can be obtained by generating independent B random samples of size m from $\text{Dir}(\bar{\alpha}_1, \dots, \bar{\alpha}_m)$ and computing the ϑ^* corresponding to each vector sampled. The empirical distribution of these ϑ^* can be taken as an approximation to the posterior distribution of ϑ as well as the posterior distribution of θ .

The α_i are specified by incorporating the given prior information. For example, if the prior mean and covariance matrix of θ are known to be $\bar{\theta}$ and V , respectively, then we may determine the α_i by forcing $E\vartheta = \bar{\theta}$ and $\text{var}(\vartheta) = V$. A general rule, however, is difficult to obtain. A further investigation is deserved.

Simulation results indicate that this method is substantially less sensi-

tive to incorrect prior specification than parametric Bayesian procedures. The major difficulty with this method is its implementation and generalization to other types of θ .

Boos and Monahan (1986) proposed a different bootstrap method to estimate the posterior distribution with an emphasis on dropping the normality assumption for the distribution of X_1 . Assume that the conditional density $f(x|\theta)$ of X (given a parameter θ) exists. $f(x|\theta)$ is called the likelihood function of X when it is viewed as a function of θ . Let $\pi(\theta)$ be the prior density of the parameter θ . From Bayes' theorem, we know that the posterior density of θ is

$$\pi(\theta|X) \propto \pi(\theta)f(X|\theta).$$

Suppose that θ is a location parameter and $\hat{\theta}_n$ is an estimate of θ based on X . The bootstrap introduced in the previous chapters can be used to simulate the density $p(u)$ of a pivotal quantity $\hat{\theta}_n - \theta$ as follows. Let F_n be the empirical distribution of X_1, \dots, X_n . Generate B random samples of size n from F_n and calculate $\hat{\theta}_{nb}^*$ for the b th sample. Using the B simulated estimates $\hat{\theta}_{n1}^*, \dots, \hat{\theta}_{nB}^*$, we compute the kernel density estimate

$$\hat{p}(u) = \frac{1}{Bh_B} \sum_{b=1}^B \kappa\left(\frac{u - (\hat{\theta}_{nb}^* - \hat{\theta}_n)}{h_B}\right),$$

where h_B is a bandwidth and κ is a kernel function. Since $\hat{p}(u - \theta)$ is an estimate of the sampling density of $\hat{\theta}_n$ given θ , the likelihood function of $\hat{\theta}_n$ can be estimated by

$$\hat{p}(\hat{\theta}_n - \theta) = \frac{1}{Bh_B} \sum_{b=1}^B \kappa\left(\frac{2\hat{\theta}_n - \theta - \hat{\theta}_{nb}^*}{h_B}\right).$$

The Bayesian bootstrap estimate for the posterior distribution of θ is

$$\hat{\pi}(\theta|X) = c(X)\pi(\theta)\hat{p}(\hat{\theta}_n - \theta),$$

where the normalizing constant $c(X)$ can be found by numerical integration.

The simulation results in Boos and Monahan (1986) show that if the conditional distribution of X_1 is the Laplace or t-distribution, then in terms of the mean squared error, $\hat{\pi}(\theta|X)$ is substantially better than the posterior distribution derived based on the assumption that X_1 is normally distributed, when a robust estimate of θ is used.

It can be seen from the above discussion that the central idea of this method is to use the bootstrap to estimate the likelihood function. Recently, there have been some other computer-intensive methods proposed

to estimate the likelihood function, e.g., the bootstrap likelihood method by Davison, Hinkley and Worton (1992) and the empirical likelihood method by Owen (1988), which can also be used to estimate the posterior distributions when prior information is incorporated. But the numerical calculation of the normalizing constant $c(X)$, a basic difficulty in Bayesian statistical analysis, cannot be avoided when these methods are used.

10.1.3 The weighted likelihood bootstrap

When the distribution function of X_1 is in a parametric family, LeCam (1956) proved that if the joint density function $f(x|\theta)$ of X and the prior density $\pi(\theta)$ satisfy some regularity conditions, then the limiting posterior distribution of θ given X is normal, i.e.,

$$\hat{\Sigma}^{-1/2}(\theta - \hat{\theta}_n) \rightarrow_d N(0, I_k) \text{ conditioned on } X \text{ a.s.}, \quad (10.2)$$

where $\hat{\theta}_n$ denotes the maximum likelihood estimator of θ , I_k is the identity matrix of order k (the dimension of θ), and $\hat{\Sigma}^{-1} = \partial^2 \log f(X|\theta) / \partial \theta \partial \theta'$ evaluated at $\theta = \hat{\theta}_n$.

It is interesting to note that the large sample posterior distribution of θ does not depend upon the prior. That is, whatever the prior distribution of θ is (of course, it should satisfy some regularity conditions), we can always use the normal distribution with an estimated covariance matrix to approximate the posterior distribution of θ . An explanation for this result is that, in large samples, the data totally dominate the prior beliefs.

To use the normal approximation, we have to know the form of the information matrix $\partial^2 \log f(X|\theta) / \partial \theta \partial \theta'$. Its derivation is sometimes difficult. The bootstrap method can be applied to approximate consistently the posterior distributions without calculating the information matrix. Newton and Raftery (1994) suggested the *weighted likelihood bootstrap* (WLB) method, which is an extension of Rubin's Bayesian bootstrap. We now introduce this method.

Assume that X_1, \dots, X_n are independent and, for given θ , each X_i has a density function $f_i(x_i|\theta)$. Suppose that θ is estimated by the maximum likelihood estimate $\hat{\theta}_n$, which maximizes

$$L(\theta) = \prod_{i=1}^n f_i(X_i|\theta).$$

To simulate the posterior distribution of θ for any prior density $\pi(\theta)$, we generate a weight vector $w = (w_1, \dots, w_n)$, which has some probability distribution determined by statisticians. For example, w can be the uniform Dirichlet, i.e., $w_i = Z_i / \sum_{j=1}^n Z_j$, or over- (down-) dispersed relative to the

Dirichlet distribution, i.e., $w_i \propto Z_i^\alpha$, $\alpha > 1$ ($\alpha < 1$), where Z_1, \dots, Z_n are i.i.d. from the exponential distribution with mean 1 and are independent of X . Define the weighted likelihood function by

$$\tilde{L}(\theta) = \prod_{i=1}^n f_i(X_i|\theta)^{w_i}.$$

Let $\tilde{\theta}_n$ be any value maximizing $\tilde{L}(\theta)$. The basic idea of the WLB is that the conditional distribution of $\tilde{\theta}_n$ given the data X can provide a good approximation of the posterior distribution of θ . The conditional distribution of $\tilde{\theta}_n$ can be found by repeatedly generating weight vectors and maximizing $\tilde{L}(\theta)$. Let $\tilde{\theta}_{nb}$ be the value maximizing the weighted likelihood function based on the b th weight vector, $b = 1, \dots, B$. Then the empirical distribution of $\tilde{\theta}_{nb}$, $b = 1, \dots, B$, can be used to approximate the posterior distribution of θ .

Newton and Raftery (1994) argued that, with uniform Dirichlet weights and under some regularity conditions, we have

$$\hat{\Sigma}^{-1/2}(\tilde{\theta}_n - \hat{\theta}_n) \rightarrow_d N(0, I_k) \text{ conditioned on } X \text{ a.s.}$$

Combining this result and result (10.2), we know that the WLB provides a consistent estimator of the posterior distribution.

Since the WLB with uniform weights does not use information from the prior, it is doubtful that the WLB estimates the posterior distribution with high order accuracy. In fact, when $k = 1$ and the X_i are i.i.d., Newton and Raftery (1994) calculated that

$$P\{\hat{\Sigma}^{-1/2}(\tilde{\theta}_n - \hat{\theta}_n) \leq t|X\} = \Phi(t) + \frac{\tilde{h}(\theta_0)\varphi(t)(t^2 - 1)}{6\sqrt{n}\mathcal{I}(\theta_0)^{3/2}} + o\left(\frac{1}{\sqrt{n}}\right) \text{ a.s.},$$

where $\tilde{h}(\theta) = 2E[\partial \log f(X_1|\theta)/\partial \theta]^3$, $\mathcal{I}(\theta) = -E[\partial^2 \log f(X_1|\theta)/\partial \theta^2]$, and θ_0 is the true parameter value. But Johnson (1970) showed that the posterior distribution function of $\hat{\Sigma}^{-1/2}(\theta - \hat{\theta}_n)$ is equal to

$$\Phi(t) + \frac{\varphi(t)}{6\sqrt{n}} \left[\frac{h(\theta_0)(t^2 + 2)}{\mathcal{I}(\theta_0)^{3/2}} + \frac{6\pi'(\theta_0)}{\pi(\theta_0)\mathcal{I}(\theta_0)^{1/2}} \right] + o\left(\frac{1}{\sqrt{n}}\right) \text{ a.s.},$$

where $h(\theta) = E[\partial^3 \log f(X_1|\theta)/\partial \theta^3]$. Therefore, by comparing these two expansions, we find that the WLB with uniform Dirichlet weight cannot be second order accurate for any given prior density $\pi(\theta)$. Newton and Raftery (1994) found that the second order accuracy can be achieved only if $h(\theta) = -2\mathcal{I}'(\theta)$ and $\pi(\theta) \propto \mathcal{I}(\theta)$.

The above discussion indicates that for a given prior the WLB with uniform Dirichlet weights provides a consistent approximation of the posterior distribution of θ , but the approximation may be poor if the sample size n

is not very large. Newton and Raftery (1994) suggested that we can use Rubin's sampling importance resampling (Rubin, 1988; Smith and Gelfand, 1992) to adjust the WLB so that the adjusted WLB is simulation consistent; that is, it produces exact answers as the amount of computations increases without bound. The basic idea is that we first use $\tilde{\theta}_{n1}, \dots, \tilde{\theta}_{nB}$ produced by the WLB to construct a kernel density estimate $\hat{\pi}$, which is an approximation of the posterior density of θ . Define the importance weights as

$$u_i \propto \pi(\tilde{\theta}_{ni})L(\tilde{\theta}_{ni})/\hat{\pi}(\tilde{\theta}_{ni}), \quad i = 1, \dots, B,$$

such that the weights sum to 1. Then we resample iteratively from the distribution putting mass u_i to $\tilde{\theta}_{ni}$, $i = 1, \dots, B$. The empirical distribution of the final sample is a simulation consistent estimator of the true posterior distribution of θ , given X_1, \dots, X_n , for a fixed prior $\pi(\theta)$.

Tu (1994) suggested that a transformation method used to improve the accuracy of the random weighting method introduced in Section 10.2 may also be applied to improve the WLB with uniform Dirichlet weights. The detailed form of the transformation, however, is still under study.

The applications of the WLB are very broad. Newton and Raftery (1994) gave examples of applications to nonlinear regression models, generalized linear models, and some dependent data problems.

10.1.4 Some remarks

For pure Bayesian statistical analysis, the prior distribution of a parameter is subjectively specified. A compromise between the frequentist and Bayesian statistical analysis is to assume that the parameter has an unknown prior distribution which can be estimated from the data. This is called empirical Bayesian statistical analysis. Laird and Louis (1987) suggested a bootstrap method to estimate the uncertainty of the prior distribution in constructing empirical Bayesian confidence sets.

In recent years, there have been many computer-intensive methods proposed to approximate posterior distributions in Bayesian statistical analysis, for example, the data augmentation method (Tanner and Wong, 1987), the sampling importance resampling method (Rubin, 1988), and the Gibbs sampler (Casella and George, 1992). A detailed discussion about these methods is certainly out of the scope of this book. Interested readers are referred to, for example, Tanner (1991). A further investigation in understanding the relationship between the Bayesian bootstrap and these methods will be interesting.

The study of applications of the Bayesian bootstrap to complex problems is still in a primary stage. Some deeper investigations are called for.

10.2 Random Weighting

The basic idea and properties of the random weighting method are introduced in this section. Applications of random weighting to linear models and statistical functionals are given in Section 10.3, and some empirical results are shown in Section 10.4.

10.2.1 Motivation

To motivate the random weighting method, we first look at a phenomenon found by Rubin (1981). Consider the sample correlation coefficient $\hat{\rho}$ discussed in Section 3.4.2 based on 12 bivariate observations $X_i = (Y_i, Z_i)$ given in Efron (1979). The sampling distribution of $\hat{\rho} - \rho$ can be estimated by H_{BOOT} , the bootstrap distribution of $\hat{\rho}^* - \hat{\rho}$ with

$$\begin{aligned}\hat{\rho}^* &= \frac{\sum Y_i^* Z_i^* - \sum Y_i^* \sum Z_i^*}{\{[\sum Y_i^{*2} - (\sum Y_i^*)^2][\sum Z_i^{*2} - (\sum Z_i^*)^2]\}^{1/2}} \\ &= \frac{\sum P_i Y_i Z_i - \sum P_i Y_i \sum P_i Z_i}{\{[\sum P_i Y_i^2 - (\sum P_i Y_i)^2][\sum P_i Z_i^2 - (\sum P_i Z_i)^2]\}^{1/2}},\end{aligned}$$

where $\{(Y_i^*, Z_i^*), i = 1, \dots, 12\}$ is a bootstrap sample from the empirical distribution of (Y_i, Z_i) , $i = 1, \dots, 12$, P_i is the frequency of $(Y_j^*, Z_j^*) = (Y_i, Z_i)$, $j = 1, \dots, 12$, and \sum is the summation over $i = 1, \dots, 12$. Using Rubin's Bayesian bootstrap, the posterior distribution of $\rho - \hat{\rho}$, given X , can be approximated by H_{BB} , the conditional distribution of $\bar{\rho}^* - \hat{\rho}$, where

$$\bar{\rho}^* = \frac{\sum w_i Y_i Z_i - \sum w_i Y_i \sum w_i Z_i}{\{[\sum w_i Y_i^2 - (\sum w_i Y_i)^2][\sum w_i Z_i^2 - (\sum w_i Z_i)^2]\}^{1/2}}$$

and w_i , $i = 1, \dots, 12$, are the weights defined in (10.1). The histograms corresponding to H_{BOOT} and H_{BB} based on 1000 Monte Carlo replications are given in Figure 10.1, where the dashed line represents the bootstrap histogram and the boldfaced line represents the Bayesian bootstrap histogram.

Although H_{BOOT} and H_{BB} estimate different quantities, it can be seen from Figure 10.1 that the two histograms are close, and that the Bayesian bootstrap histogram is smoother than the bootstrap histogram.

The reason why the Bayesian bootstrap histogram is smoother can be explained as follows. By comparing $\hat{\rho}^*$ with $\bar{\rho}^*$, we find that both of them are based on weighted averages of $(Y_i, Z_i, Y_i^2, Z_i^2, Y_i Z_i)$, $i = 1, \dots, 12$. The weights (P_1, \dots, P_{12}) for the bootstrap are multinomial, whereas the Bayesian bootstrap method adopts smooth weights (w_1, \dots, w_{12}) that are jointly distributed as $\text{Diri}(1, \dots, 1)$.

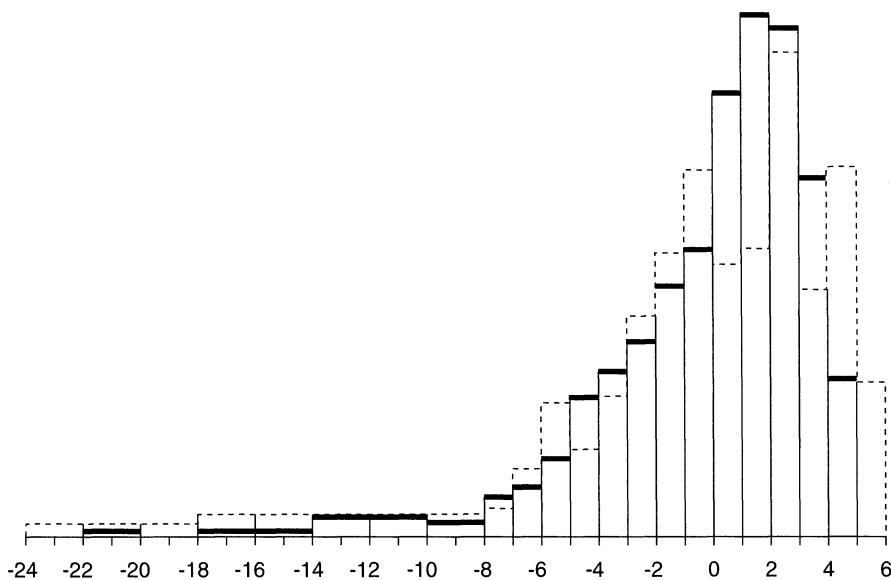


Figure 10.1. Comparison of the bootstrap and Bayesian bootstrap histograms [From Rubin (1981), Copyright© (1981) by Institute of Mathematical Statistics, All rights reserved]. Dashed line: the bootstrap histogram. Boldfaced line: the Bayesian bootstrap histogram.

This indicates that the conditional distribution of $\bar{\rho}^* - \hat{\rho}$ can be taken as an approximation to the sampling distribution of $\hat{\rho} - \rho$ and can be viewed as a smoothed alternative to the bootstrap distribution. If the sampling distribution of $\hat{\rho} - \rho$ is smooth, e.g., $\hat{\rho} - \rho$ has a density, then this alternative method may provide more information about the sampling distribution of $\hat{\rho} - \rho$.

Partially based on these observations, Zheng (1987a) argued that the conditional distribution obtained from the Bayesian bootstrap operation can also be used to approximate the sampling distribution in frequentist statistical analysis. He called this method the random weighting method. Let $w = (w_1, \dots, w_n)$ be a vector of random weights independent of X and define

$$\bar{F}_n^*(x) = \sum_{i=1}^n w_i I\{X_i \leq x\}. \quad (10.3)$$

To estimate the distribution of $T(F_n) - T(F)$ for a given functional T , we can use the conditional distribution of $T(\bar{F}_n^*) - T(F_n)$, given X . The

actual calculation of the conditional distribution can be done by repeatedly sampling weights from the distribution of w , e.g., the $\text{Diri}(1, \dots, 1)$. Note that \bar{F}_n^* has the same form as D_n in (10.1).

The distribution of the random weight vector does not have to be restricted to the $\text{Diri}(1, \dots, 1)$. Later investigations found that the weights having a scaled $\text{Diri}(4, \dots, 4)$ distribution give better approximations (Tu and Zheng, 1987; Weng, 1989). Detailed results will be given later. In fact, the bootstrap can also be viewed as a random weighting method with weights distributed as a scaled multinomial distribution. Lo (1991) suggested that we choose w_i as $Z_i / \sum_{j=1}^n Z_j$, where Z_1, \dots, Z_n are i.i.d. non-negative random variables from a known distribution G . He called this method the bootstrap clone method. Mason and Newton (1992) proposed using a vector of random weights (w_1, \dots, w_n) satisfying only $w_i \geq 0$ and $\sum_{i=1}^n w_i = 1$. They called this method the general bootstrap method.

The random weighting method is designed to approximate the sampling distribution of a random variable. Therefore, it can also be applied to construct confidence intervals, using the methods introduced in Chapter 4 with some modifications. A detailed discussion about the properties of random weighted confidence intervals is not yet available. Some simulation results are shown in Section 10.4.

Since the selection of the random weights is flexible, for a practical problem we need to know what types of weights we may use. Unfortunately, a general rule does not exist, but the asymptotic theory presented next provides some limited rules.

10.2.2 Consistency

In the following discussion, a minimum assumption on the weight vector is that it has a density. Thus, the bootstrap (the random weighting method with multinomial weights) is excluded.

Let H_n be the sampling distribution of $\sqrt{n}(\bar{X}_n - \mu)$, where \bar{X}_n is the sample mean of i.i.d. random variables X_1, \dots, X_n with mean μ and variance σ^2 . The definition of consistency is the same as that given in Definition 3.1, except that the bootstrap distribution H_{BOOT} is replaced by the random weighting distribution H_{RW} , the conditional distribution of $\sqrt{n}(\bar{X}_n^* - \bar{X}_n)$ given $X = (X_1, \dots, X_n)$, where $\bar{X}_n^* = \int x d\bar{F}_n^*(x)$ with \bar{F}_n^* given by (10.3). The consistency of H_{RW} with w distributed as $\text{Diri}(1, \dots, 1)$ was proved by Zheng (1987a). The same result for scaled $\text{Diri}(4, \dots, 4)$ weights can be derived from the results given by Tu and Zheng (1987) and Weng (1989). When $w_i = Z_i / \sum_{j=1}^n Z_j$, Lo (1991) showed that H_{RW} is consistent if $\sqrt{\text{var}(\bar{Z}_1)} / E\bar{Z}_1 = 1$. Mason and Newton (1992) provided the following general result.

Theorem 10.2. Assume that the weights w_1, \dots, w_n are exchangeable, independent of X , and satisfy $w_i \geq 0$, $\sum_{i=1}^n w_i = 1$, $\sum_{i=1}^n (w_i - \frac{1}{n})^2 \rightarrow_{a.s.} 1$, and $\sqrt{n} \max_{1 \leq i \leq n} |w_i - \frac{1}{n}| \rightarrow_{a.s.} 0$. Then

$$\|H_{\text{RW}} - H_n\|_\infty \rightarrow_{a.s.} 0.$$

Proof. The proof of Theorem 10.2 involves a difficult rank statistics approach and is rather complicated. We instead describe the proof for a special type of weight (but general enough to include the Dirichlet weights) considered by Lo (1991). Let Z_1, \dots, Z_n be i.i.d. nonnegative random variables independent of X and $w_i = Z_i / \sum_{j=1}^n Z_j$. It is easy to check that the conditions on w_i are satisfied if Z_1 has a finite second moment and $\sqrt{\text{var}(Z_1)} / EZ_1 = 1$. Let \bar{Z}_n be the average of Z_1, \dots, Z_n . Then

$$\begin{aligned} \sqrt{n}(\bar{X}_n^* - \bar{X}_n) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{Z_i}{\bar{Z}_n} (X_i - \bar{X}_n) \\ &= \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{Z_i - EZ_i}{\sqrt{\text{var}(Z_1)}} (X_i - \bar{X}_n) \right] \frac{\sqrt{\text{var}(Z_1)}}{\bar{Z}_n}. \end{aligned}$$

Note that

$$\max_{i \leq n} (X_i - \bar{X}_n)^2 \Big/ \sum_{i=1}^n (X_i - \bar{X}_n)^2 \rightarrow_{a.s.} 0$$

and

$$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \rightarrow_{a.s.} \sigma^2.$$

Therefore, for almost all given X ,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{Z_i - EZ_1}{\sqrt{\text{var}(Z_1)}} (X_i - \bar{X}_n) \rightarrow_d N(0, \sigma^2)$$

(e.g., Chow and Teicher, 1988). Under the given condition on Z_i ,

$$\sqrt{\text{var}(Z_1)} / \bar{Z}_n \rightarrow_{a.s.} 1.$$

Hence, the result follows from

$$\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \rightarrow_d N(0, \sigma^2) \text{ conditioned on } X \text{ a.s. } \square$$

From the definition of the Dirichlet distribution, we know that if w is distributed as $\text{Diri}(1, \dots, 1)$, then $w_i = Z_i / \sum_{j=1}^n Z_j$, where Z_1, \dots, Z_n are i.i.d. exponential random variables with mean 1. It is easy to check that $EZ_1 = \text{var}(Z_1) = 1$. Thus, by Theorem 10.2, we know that the random weighting method with $\text{Diri}(1, \dots, 1)$ weight vector is consistent.

It is interesting to note that if w has the $\text{Diri}(4, \dots, 4)$ distribution, then $\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \not\rightarrow_d N(0, \sigma^2)$ and H_{RW} is inconsistent. In fact, w_i can be written as $w_i = Z_i / \sum_{j=1}^n Z_j$, $i = 1, \dots, n$, where Z_1, \dots, Z_n are i.i.d. as $\text{Gamma}(4)$ (the gamma distribution with density $[\Gamma(4)]^{-1}t^3e^{-t}I\{t > 0\}$), and $EZ_1 = 4 \neq \sqrt{\text{var}(Z_1)} = 2$. Hence, if we rescale the weight vector w so that $w/2$ has the $\text{Diri}(4, \dots, 4)$ distribution, then H_{RW} is consistent. Thus, the selection of the weight vector in the random weighting method is a delicate issue. More discussion about this will be given later.

The consistency of random weighting approximations for some statistics other than the sample mean will be discussed in Section 10.3.

10.2.3 Asymptotic accuracy

We now study the asymptotic accuracy of the random weighting method by considering the estimation of $\tilde{H}_n(x) = P\{\sqrt{n}(\bar{X}_n - \mu)/\sigma \leq x\}$, the distribution of the standardized sample mean of i.i.d. random variables; that is, we want to know how close the random weighting estimator

$$\tilde{H}_{\text{RW}}(x) = P_*\{(\bar{X}_n^* - \bar{X}_n)/S_w \leq x\}$$

can approximate $\tilde{H}_n(x)$, where $S_w^2 = \text{var}_*(\bar{X}_n^*)$. Note that

$$S_w^2 = \frac{1}{n(n+1)} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

for $\text{Diri}(1, \dots, 1)$ weights and

$$S_w^2 = \frac{1}{n(4n+1)} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

for $\text{Diri}(4, \dots, 4)$ weights.

By developing an asymptotic expansion for $\tilde{H}_{\text{RW}}(x)$, Tu and Zheng (1987) found that the random weighting method with $\text{Diri}(1, \dots, 1)$ weights cannot be second order accurate. They also proved the following result.

Theorem 10.3. *If F is nonlattice and $E|X_1|^3 < \infty$, then the random weighting estimator $\tilde{H}_{\text{RW}}(x)$ using the weight vector having $\text{Diri}(4, \dots, 4)$ distribution satisfies*

$$\|\tilde{H}_{\text{RW}} - \tilde{H}_n\|_\infty = o(n^{-1/2}) \quad a.s.; \tag{10.4}$$

that is, \tilde{H}_{RW} is second order accurate.

Because the weight vector has a continuous distribution, the proof of Theorem 10.3 is easier than that for the accuracy of the bootstrap estimator (for which we have to deal with the asymptotic expansion of a discrete distribution). The proof relies on the following expansion of \tilde{H}_{RW} .

Lemma 10.1. If $E|X_1|^3 < \infty$ and the vector of weights is distributed as $\text{Diri}(4, \dots, 4)$, then uniformly in x ,

$$\tilde{H}_{\text{RW}}(x) = \Phi(\beta(x)) - \frac{1}{6}\varphi(\beta(x))[\beta^2(x) - 1]\gamma(x) + o(n^{-1/2}) \quad a.s.,$$

where

$$\beta(x) = 2nxS_w / \left[\sum_{i=1}^n (X_i - \bar{X}_n - xS_w)^2 \right]^{1/2}$$

and

$$\gamma(x) = \sum_{i=1}^n (X_i - \bar{X}_n - xS_w)^3 / \left[\sum_{i=1}^n (X_i - \bar{X}_n - xS_w)^2 \right]^{3/2}.$$

Proof. From the definition of the Dirichlet distribution, we know that (w_1, \dots, w_n) has the same distribution as $(Z_1, \dots, Z_n)/\sum_{i=1}^n Z_i$, where the Z_i are i.i.d. as Gamma(4), $EZ_1 = 4$, $\text{var}(Z_1) = 4$, and $E(Z_1 - 4)^3 = 8$. Thus,

$$\begin{aligned} \tilde{H}_{\text{RW}}(x) &= P_* \left\{ \sum_{i=1}^n Z_i (X_i - \bar{X}_n) \leq xS_w \sum_{i=1}^n Z_i \right\} \\ &= P_* \left\{ \sum_{i=1}^n (Z_i - 4)(X_i - \bar{X}_n - xS_w) \leq 4xS_w \right\} \\ &= P_* \left\{ \sum_{i=1}^n Y_i / \left(\sum_{j=1}^n E_* Y_j^2 \right)^{1/2} \leq \beta(x) \right\}, \end{aligned}$$

where the $Y_i = (Z_i - 4)(X_i - \bar{X}_n - xS_w)$ are independent for given X_1, \dots, X_n . From the result on the asymptotic expansions for sums of independent random variables in Bai and Zhao (1986) (see Appendix A.10), we have

$$|\tilde{H}_{\text{RW}}(x) - \Phi(\beta(x)) + \frac{1}{6}\varphi(\beta(x))[\beta^2(x) - 1]\gamma(x)| \leq c(J_1 + J_2 + J_3),$$

where c is a constant,

$$J_1 = \frac{1}{\mu_2^{3/2}} \sum_{j=1}^n E_* |U_j|^3, \quad J_2 = \frac{1}{\mu_2^2} \sum_{j=1}^n E_* |Y_j|^4,$$

$$J_3 = n^6 \left[\sup_{|t| \geq \delta_n} \frac{1}{n} \sum_{j=1}^n |\psi_j(t)| + \frac{1}{2n} \right]^n, \quad \mu_2 = 2 \sum_{i=1}^n (X_i - \bar{X}_n - xS_w)^2,$$

$$U_j = (Z_j - 4)(X_j - \bar{X}_n - xS_w) I\{|(Z_j - 4)(X_j - \bar{X}_n - xS_w)| > \mu_2^{1/2}\},$$

$$\psi_j(t) = E_* \exp(itY_j) = \frac{\exp[-4it(X_j - \bar{X}_n - xS_w)]}{[1 - t(X_j - \bar{X}_n - xS_w)]^4},$$

and $\delta_n = \frac{\mu_2}{12} (\sum_{j=1}^n E_* |Y_j|^3)^{-1}$. Define

$$\alpha_i(x) = (X_i - \bar{X}_n - xS_w) / \left[\sum_{i=1}^n (X_i - \bar{X}_n - xS_w)^2 \right]^{1/2}.$$

Then, some probabilistic arguments yield that

$$\max_{i \leq n} \sup_x |\alpha_i(x)| \rightarrow_{a.s.} 0.$$

Thus, for J_1 , we have

$$\begin{aligned} J_1 &\leq \sum_{j=1}^n |\alpha_i(x)|^3 E_* \left\{ |Z_1 - 4|^3 I\{|Z_1 - 4| > \left[\max_{i \leq n} \sup_x |\alpha_i(x)| \right]^{-1}\} \right\} \\ &= o\left(\sum_{i=1}^n |\alpha_i(x)|^3\right) \text{ a.s.} \end{aligned}$$

Similarly,

$$J_2 = \sum_{i=1}^n |\alpha_i(x)|^4 E_*(Z_1 - 4)^4 = o\left(\sum_{i=1}^n |\alpha_i(x)|^3\right) \text{ a.s.}$$

Some more difficult arguments on $\psi_j(t)$ yield that [details can be found in Tu and Zheng (1987) or Tu (1988c)]

$$J_3 = o\left(\sum_{i=1}^n |\alpha_i(x)|^3\right) \text{ a.s.}$$

From the strong law of large numbers and $E|X_1|^3 < \infty$, we obtain that

$$\limsup_{n \rightarrow \infty} \sqrt{n} \sup_x \sum_{i=1}^n |\alpha_i(x)|^3 < \infty \text{ a.s.}$$

Hence, $J_1 + J_2 + J_3 = o(n^{-1/2})$ a.s. and the lemma is proved. \square

Proof of Theorem 10.3. If F is nonlattice and $E|X_1|^3 < \infty$, then

$$\tilde{H}_n(x) = \Phi(x) - \varphi(x)(x^2 - 1) \frac{E(X_1 - \mu)^3}{6\sqrt{n}\sigma^3} + o\left(\frac{1}{\sqrt{n}}\right)$$

uniformly in x (see Appendix A.10). By some elementary analysis, we can show that

$$\sup_x |\Phi(\beta(x)) - \Phi(x)| = O(n^{-1}),$$

$$\sup_x |\varphi(\beta(x))[\beta^2(x) - 1] - \varphi(x)(x^2 - 1)| = O(n^{-1}),$$

and

$$\sup_x \left| \varphi(x)(x^2 - 1) \left[\sum_{i=1}^n \alpha_i^3(x) - \frac{E(X_1 - \mu)^2}{\sqrt{n}\sigma^3} \right] \right| = o\left(\frac{1}{\sqrt{n}}\right) \text{ a.s.}$$

Therefore, result (10.4) follows from Lemma 10.1. \square

From the proofs of Theorem 10.3 and Lemma 10.1, we can see that if we use

$$\tilde{H}_{RW2}(x) = P_* \left\{ \sum_{i=1}^n \xi_i (X_i - \bar{X}_n) \middle/ \sum_{i=1}^n (X_i - \bar{X}_n)^2 \leq x \right\}$$

to approximate $\tilde{H}_n(x)$, where ξ_1, \dots, ξ_n are i.i.d. random variables with Gamma(4) distribution, then (10.4) also holds for \tilde{H}_{RW2} (Zheng, 1987a). Technically, the proof of (10.4) is a bit easier; and practically this approximation may be better for tail probability estimation since the weights ξ_1, \dots, ξ_n are unbounded. Simulation studies will be given in Section 10.4 to compare this approach with those based on other weights.

Zheng (1987b) showed that Theorem 10.3 is still true if we only assume that X_1, \dots, X_n are independent with the same mean μ (they are not necessarily identically distributed). Of course, some additional assumptions on the distributions of X_j , $j = 1, \dots, n$, should be made. This is a robustness property of random weighting, like the external bootstrap (Section 7.2.2).

Let G_n be the distribution of the studentized sample mean

$$\sqrt{n}(\bar{X}_n - \mu) \middle/ \left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \right]^{1/2}.$$

The estimation of G_n is more desirable in practice. We should use the conditional distribution of

$$(\bar{X}_n^* - \bar{X}_n) \middle/ \left[\sum_{i=1}^n (w_i X_i - \bar{X}_n^*)^2 \right]^{1/2}$$

as an estimate of G_n . The asymptotic properties of this estimator are, however, technically difficult to discuss. We may use $\tilde{H}_{RW2}(x)$ to approximate G_n , but the second order accuracy is lost. Zheng (1989) suggested that if we use the following transformation

$$\phi_n(u) = u - \frac{\hat{\mu}_3}{2\sqrt{n}\hat{\mu}_2^{3/2}} u^2 + \frac{\hat{\mu}_3^2}{16n\hat{\mu}_2^2} u^3,$$

where $\hat{\mu}_k = n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^k$, $k = 2, 3$, then we have

$$\sup_x \left| G_n(x) - P_* \left\{ \phi_n \left[\sum_{i=1}^n \xi_i (X_i - \bar{X}_n) / \sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \right] \leq x \right\} \right| = o\left(\frac{1}{\sqrt{n}}\right) \text{ a.s.}$$

Now we discuss the accuracy of random weighting estimators with general weights. In principle, if w_i can be written as $w_i = \xi_i / \sum_{i=1}^n \xi_i$, where ξ_1, \dots, ξ_n are i.i.d. random variables independent of X , then Theorem 10.3 can be generalized if $n \text{var}_*(\bar{X}_n^* - \bar{X}_n) \rightarrow_{a.s.} c$ for some $c > 0$. Haeusler, Mason and Newton (1991) considered the approximation to $\tilde{H}_n(x)$ by using the conditional distribution

$$\tilde{H}_\xi(x) = P_* \left\{ \rho_\xi \sum_{i=1}^n w_i (X_i - \bar{X}_n) / S_n \leq x \right\},$$

where $\rho_\xi = n^{-1} \sum_{i=1}^n \xi_i / [(n-1)^{-1} \sum_{i=1}^n (\xi_i - n^{-1} \sum_{j=1}^n \xi_j)^2]^{1/2}$ and $S_n^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. Using a result of Schneller (1989) on the Edgeworth expansion for linear rank statistics, they showed that

$$\|\tilde{H}_\xi - \tilde{H}_n\|_\infty = o(n^{-1/2}) \text{ a.s.},$$

if (1) $E|\xi_1 - E\xi_1|^{2s} < \infty$ for some $s > 24/7$; (2) the density of $|\xi_1 - \xi_2|$ is bounded near 0; (3) $E(\xi_1 - E\xi_1)^3 / [\text{var}(\xi_1)]^{3/2} = 1$; and (4) X_1 is nonlattice and $E|X_1|^3 < \infty$.

Condition (3) restricts the selection of the weight vector w . Some examples are: ξ_1 has the Gamma(4) distribution; ξ_1 has the Gamma distribution with parameters 4 and $\frac{1}{2}$; and $\xi_1 = \zeta_1 \eta_1$, where $\zeta_1 \sim N((\sqrt{17}+1)/\sqrt{24}, \frac{1}{2})$, $\eta_1 \sim N((\sqrt{17}-1)/\sqrt{6}, \frac{1}{2})$, and ζ_1 and η_1 are independent (Liu, 1988).

Haeusler, Mason and Newton (1991) also considered the second order accuracy of the random weighting estimators with a sequence of unbounded independent weights as discussed in Zheng (1987a). They also developed a two-term Edgeworth expansion for $\tilde{H}_\xi(x)$. To match the second order term in the Edgeworth expansion with the bootstrap approximation so that these two methods are equivalent up to the order of $O(n^{-3/2})$, we have to allow the distributions of the weights depending on X_1, \dots, X_n . Some examples of these weights can be found in Haeusler, Mason and Newton (1991). They also compared by simulation the random weighting method using data-dependent weights with the bootstrap. The results show that the median quantile of the random weighting distribution is closer to the true quantile of interest than the median quantile of the bootstrap distribution, although the variability of the random weighting estimator is generally larger.

10.3 Random Weighting for Functionals and Linear Models

In this section, we introduce applications of the random weighting method to problems of estimating the sampling distributions of statistical functionals and the least squares estimator in linear regression. There are also other applications of random weighting. For example, Yu (1988) considered applications in sample surveys and Fan and Mei (1991) discussed the accuracy of the random weighting estimators in autoregressive models. Rao and Zhao (1992) suggested the random weighting approximation for M-estimators in linear models.

10.3.1 Statistical functionals

In this section, we study the random weighting estimator of H_n , the distribution of $\sqrt{n}[T(F_n) - T(F)]$, where T is a given functional and F_n is the empirical distribution of i.i.d. random variables X_1, \dots, X_n with a common but unknown distribution F . From the basic idea of the random weighting method, we can define a randomly weighted empirical distribution \bar{F}_n^* according to (10.3) and estimate H_n by H_{RW} , the conditional distribution of $\sqrt{n}[T(\bar{F}_n^*) - T(F_n)]$, given $X = (X_1, \dots, X_n)$. Zheng (1987c) and Mason and Newton (1992) showed that the randomly weighted empirical process $\sqrt{n}[\bar{F}_n^*(x) - F_n(x)]$ approximates consistently the empirical process $\sqrt{n}[F_n(x) - F(x)]$ for Diri(1, ..., 1) weights and general exchangeable weights, respectively. Tu (1988d) proved the consistency of the random weighting estimator H_{RW} .

Theorem 10.4. *Assume that there exists a functional $\phi(x, F)$ such that, for any G and F ,*

$$T(G) - T(F) = \int \phi(x, F) dG(x) + R(G, F),$$

where $R(G, F)$ satisfies (i) $\sqrt{n}R(F_n, F) \rightarrow_p 0$; (ii) $P_*\{\sqrt{n}R(\bar{F}_n^*, F_n)| > \epsilon\} \rightarrow_{a.s.} 0$ for any $\epsilon > 0$; (iii) $\int [\phi(x, F_n) - \phi(x, F)]^2 dF_n(x) \rightarrow_{a.s.} 0$; and (iv) $\int \phi(x, F) dF(x) = 0$ and $0 < \sigma^2(F) = \int [\phi(x, F)]^2 dF(x) < \infty$. Then

$$\|H_{\text{RW}} - H_n\|_\infty \rightarrow_{a.s.} 0,$$

if the distribution of (w_1, \dots, w_n) is Diri(1, ..., 1).

Proof. Under conditions (i) and (iv),

$$\sqrt{n}[T(F_n) - T(F)] \rightarrow_d N(0, \sigma^2(F)).$$

Since

$$T(\bar{F}_n^*) - T(F_n) = \sum_{i=1}^n w_i \phi(X_i, F_n) + R(\bar{F}_n^*, F_n)$$

and by (ii), $\sqrt{n}R(\bar{F}_n^*, F_n) \rightarrow_p^* 0$ a.s. (\rightarrow_p^* denotes convergence in P_* probability), we only need to show that

$$\sqrt{n} \sum_{i=1}^n w_i \phi(X_i, F_n) \rightarrow_d N(0, \sigma^2(F)) \text{ a.s.}$$

From Theorem 10.2, we have

$$\sqrt{n} \sum_{i=1}^n w_i \phi(X_i, F) \rightarrow_d N(0, \sigma^2(F)) \text{ a.s.}$$

Thus, it remains to show that

$$A_n = \sqrt{n} \sum_{i=1}^n w_i [\phi(X_i, F_n) - \phi(x, F)] \rightarrow_p^* 0 \text{ a.s.}$$

It is easy to calculate that

$$\begin{aligned} E_* A_n^2 &= \frac{n}{n+1} \int [\phi(x, F_n) - \phi(x, F)]^2 dF_n(x) \\ &\quad + \frac{1}{n+1} \left[\int \phi(x, F_n) dF_n(x) - \int \phi(x, F) dF_n(x) \right]^2 \rightarrow_{a.s.} 0 \end{aligned}$$

under condition (iii). \square

The following are some examples of statistical functionals for which we can apply Theorem 10.4 (Tu, 1988d).

Example 10.1. U-statistics. Let $U_n = (\frac{n}{2})^{-1} \sum_{i < j} h(X_i, X_j)$. Suppose that $0 < E[h^2(X_1, X_2)] < \infty$ and $E[h^2(X_1, X_1)] < \infty$. Then Theorem 10.4 can be applied to U_n .

Example 10.2. Smooth L-statistics. Define $T(F_n) = \int_0^1 F_n^{-1}(t) J(t) dt$. If $0 < \text{var}[\phi(X, F)] < \infty$, where $\phi(x, F) = -\int [I\{y \geq x\} - F(y)] J(F(y)) dy$, then Theorem 10.4 can be applied under either of the following two conditions: (a) J is bounded and continuous a.e. Lebesgue measure and F^{-1} and vanishes outside of $(\alpha, \beta]$, where $0 < \alpha < \beta < 1$; (b) J is continuous on $[0, 1]$ and satisfies the Lipschitz condition of order λ ($0 < \lambda < 1$) and F satisfies $\int \{F(y)[1 - F(y)]\}^{1/2} dy < \infty$.

Huskova and Janssen (1993a,b) and Janssen (1994) discussed the random weighting approximation for studentized U-statistics. Zheng (1988c)

considered the random weighting approximation to the distribution of the sample median, a nonsmooth statistical functional. Zheng (1987c) and Mason and Newton (1992) proved the consistency of the random weighting approximation for the sample quantile process.

Now we consider the second order accuracy of the random weighting estimator of \tilde{H}_n , the distribution of $[T(F_n) - T(F)]/\{\text{var}[T(F_n)]\}^{1/2}$. A natural way to apply the random weighting method is to use \tilde{H}_{RW} , the conditional distribution of $[T(\bar{F}_n^*) - T(F_n)]/\{\text{var}_*[T(\bar{F}_n^*)]\}^{1/2}$. But Tu (1986b) found that for the minimum contrast estimator defined by

$$\hat{\theta}_n = \arg \min_{\theta} \sum_{i=1}^n \rho(X_i, \theta),$$

\tilde{H}_{RW} cannot estimate \tilde{H}_n with second order accuracy, even if the distribution of the weight vector is $\text{Dir}(4, \dots, 4)$. Zheng (1988d) and Tu and Shi (1988) suggested some transformations of $[T(\bar{F}_n^*) - T(F_n)]/\{\text{var}_*[T(\bar{F}_n^*)]\}^{1/2}$ and showed that the transformed random weighting distribution estimators are second order accurate when $T(F_n)$ is an M-estimator or a U-statistic. These transformations, however, depend on the influence functions of $T(F)$ and may be difficult to estimate if $T(F_n)$ is complicated. Tu (1992a) proposed a method using the jackknife to construct the transformations. Let $F_{n-1,i}$ be the empirical distribution of $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$, $T_n = T(F_n)$, $T_{n-1,i} = T(F_{n-1,i})$, $\tilde{T}_{n,i} = nT_n - (n-1)T_{n-1,i}$, $\tilde{T} = n^{-1} \sum_{i=1}^n \tilde{T}_{n,i}$, and

$$T_n^* = \frac{1}{n} \sum_{i=1}^n \xi_i (\tilde{T}_{n,i} - \tilde{T}),$$

where ξ_1, \dots, ξ_n are i.i.d. random variables independent of X_1, \dots, X_n and satisfy $\text{var}(\xi_1) = 1$ and $E(\xi_1 - E\xi_1)^3 = 1$. The computation of the distribution of T_n^* is easier than that of $T(\bar{F}_n^*)$. For each weight vector generated, we do not need to recalculate the statistic $T(\bar{F}_n^*)$, which is an advantage if the computation of T_n involves an iterative process.

However, if we estimate \tilde{H}_n by using the conditional distribution of $T_n^*/[\text{var}_*(T_n^*)]^{1/2}$, we cannot have a second order accurate estimator (Tu, 1992a). A polynomial transformation with coefficients constructed from the jackknife pseudovalues $\tilde{T}_{n,i}$ was suggested by Tu (1992a) to improve the accuracy. Let

$$\psi_n(y) = \left(y - \frac{\beta_{0n}}{\sqrt{n}} \right) - \frac{\beta_{2n}}{\sqrt{n}} \left(y - \frac{\beta_{0n}}{\sqrt{n}} \right)^2 + \frac{\beta_{2n}}{3n} \left(y - \frac{\beta_{0n}}{\sqrt{n}} \right)^3,$$

where

$$\beta_{0n} = \frac{sk_{\text{JACK}}}{6} - \frac{\hat{k}}{2} - \frac{\sqrt{n}}{6} \sum_{j=1}^n r_{nj}^3, \quad \beta_{2n} = -\frac{sk_{\text{JACK}}}{6} + \frac{\sqrt{n}}{6} \sum_{j=1}^n r_{nj}^3,$$

$$r_{nj} = (\tilde{T}_{n,j} - \tilde{T}_\cdot) \Big/ \left[\sum_{i=1}^n (\tilde{T}_{n,i} - \tilde{T}_\cdot)^2 \right]^{1/2},$$

$$\hat{k} = \frac{n-1}{[\text{var}_*(T_n^*)]^{1/2}} \sum_{i=1}^n (T_{n-1,i} - T_n),$$

and sk_{JACK} is defined by (3.60). Tu (1992a) proved that for U-statistics, M-estimators, and some second order differentiable statistical functionals,

$$\sup_x |P_*\{\psi_n(T_n^*/[\text{var}_*(T_n^*)]^{1/2}) \leq x\} - \tilde{H}_n(x)| = o(n^{-1/2}) \text{ a.s.},$$

provided that $E|T(F_n)|^4 < \infty$ and the distribution of $\phi(X_1, F)$ is nonlattice.

For the studentized pivot $[T(F_n) - T(F)]/\hat{\sigma}_n$, where $\hat{\sigma}_n^2$ is a variance estimate of $T(F_n)$, e.g., the jackknife variance estimator v_{JACK} , we can modify the transformation ψ_n based on the asymptotic expansion of the studentized pivot so that the distribution of the studentized pivot can be estimated with second order accuracy. Zhang and Tu (1990) proposed a procedure using $\hat{\sigma}_n^2 = v_{\text{JACK}}$ and the following transformation (see also Section 10.4):

$$\tilde{\psi}_n(y) = (y - b_n) + a_n(y - b_n)^2 + \frac{a_n^2}{3}(y - b_n)^2, \quad (10.5)$$

where

$$a_n = \frac{1}{2(nv_{\text{JACK}})^{3/2}} \left[-\sum_{j=1}^n \Delta_j^3 + n(n-1)^2 \sum_{j \neq k} \Delta_j \Delta_k \Delta_{jk} \right],$$

$$b_n = \frac{1}{(nS_{nJ})^{3/2}} \left[\sum_{j=1}^n \Delta_j + \frac{n(n-1)^2}{2} \sum_{j \neq k} \Delta_j \Delta_k \delta_{jk} \right],$$

$\Delta_j = (n-1)(T_n - T_{n-1,j})$, and Δ_{jk} is defined in (3.60).

10.3.2 Linear models

Consider the linear model described in (7.1):

$$y_i = x'_i \beta + \varepsilon_i, \quad i = 1, \dots, n,$$

where β is a p -vector of unknown parameters, $x_i \in \mathbb{R}^p$ is deterministic, and the ε_i are i.i.d. with mean 0 and unknown variance σ^2 . Let $X' = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)'$. Assume that X is of full rank for simplicity. Then the least squares estimator of β is $\hat{\beta} = (X'X)^{-1}X'y$.

In this section, we discuss the application of random weighting in estimating the distributions of some quantities related to $\hat{\beta} - \beta$. Let c be a fixed p -vector and

$$\mathfrak{R}_n = c'(\hat{\beta} - \beta)/[\text{var}(c'\hat{\beta})]^{1/2}.$$

Define $l_i = c'(X'X)^{-1}x_i$, $a_i = l_i/(\sum_{i=1}^n l_i^2)^{1/2}$, and $e_i = \varepsilon_i/\sigma$. Since $\text{var}(c'\hat{\beta}) = \sigma^2 c'(X'X)^{-1}c$, $\mathfrak{R}_n = \sum_{i=1}^n a_i e_i$, which is a linear combination of a vector of i.i.d. random variables e_1, \dots, e_n . e_i can be estimated by $\hat{e}_i = r_i/\hat{\sigma}$, where $r_i = y_i - x'_i\hat{\beta}$ and $\hat{\sigma}^2 = (n-p)^{-1}\sum_{i=1}^n r_i^2$. This suggests that we define the random weighting analog of \mathfrak{R}_n by

$$\mathfrak{R}_n^* = d_n^*/[\text{var}_*(d_n^*)]^{1/2},$$

where $d_n^* = \sum_{i=1}^n w_i a_i \hat{e}_i$ and $w = (w_1, \dots, w_n)$ is a vector of weights independent of y . Since

$$\text{var}_*(d_n^*) = \frac{1}{n(4n+1)} \sum_{i=1}^n a_i^2 \hat{e}_i^2$$

when w has the $\text{Diri}(4, \dots, 4)$ distribution, \mathfrak{R}_n^* can be written as

$$\mathfrak{R}_n^* = \sum_{i=1}^n w_i a_i r_i / \left[\frac{1}{n(4n+1)} \sum_{i=1}^n a_i^2 r_i^2 \right]^{1/2}.$$

For the consistency of the random weighting distribution estimator, we have the following theorem.

Theorem 10.5. *Suppose that w is distributed as $\text{Diri}(4, \dots, 4)$. If $E\varepsilon_i^2 < \infty$, $\sup_i \|x_i\| < \infty$, and $X'X/n \rightarrow V > 0$ (positive definite), then*

$$\sup_x |P_*\{\mathfrak{R}_n^* \leq x\} - P\{\mathfrak{R}_n \leq x\}| \rightarrow_{a.s.} 0.$$

The proof of Theorem 10.5 is similar to that of Theorem 10.2, but some asymptotic properties of the residuals r_1, \dots, r_n need to be utilized. The details can be found in Zheng and Tu (1988).

The second order accuracy of the random weighting estimator was also discussed by Zheng and Tu (1988). By applying the asymptotic expansions for sums of independent random variables given by Bai and Zhao (1986), Zheng and Tu (1988) showed that

$$\sqrt{n} \sup_x \left| P\{\mathfrak{R}_n \leq x\} - \Phi(x) + \varphi(x)(x^2 - 1) \frac{E\varepsilon_1^3}{6\sigma^3} \sum_{i=1}^n a_i^3 \right| \rightarrow 0, \quad (10.6)$$

provided that $E|\varepsilon_i|^3 < \infty$, $\max_{i \leq n} a_i^2 = O(n^{-1})$, and ε_1 satisfies Cramér's condition, i.e., for any $\delta > 0$, there is a $\lambda < 1$ such that $\sup_{|t| \geq \delta} |\psi(t)| \leq \lambda$, where $\psi(t)$ is the characteristic function of ε_1 .

Theorem 10.6. Suppose that the weight vector w has the $\text{Dir}(4, \dots, 4)$ distribution and that ε_1 satisfies Cramér's condition and one of the following two conditions holds:

- (i) $E\varepsilon_1^6 < \infty$ and $\max_{i \leq n} a_i^2 = O(n^{-1})$;
- (ii) $E|\varepsilon_1|^3 < \infty$, $\sup_i \|x_i\| < \infty$, and $X'X/n \rightarrow V > 0$.

Then

$$\sup_x |P_*\{\mathfrak{R}_n^* \leq x\} - P\{\mathfrak{R}_n \leq x\}| = o(n^{-1/2}) \quad a.s.$$

Proof. The same technique used in the proof of Lemma 10.1 yields that

$$\sqrt{n} \sup_x |P_*\{\mathfrak{R}_n^* \leq x\} - \Phi(\rho_n(x)) + \varphi(\rho_n(x))[\rho_n^2(x) - 1]\gamma_n(x)| \rightarrow_{a.s.} 0,$$

where $\rho_n(x) = 2\sqrt{nx}/\sqrt{4n+1+x^2}$ and

$$\gamma_n(x) = \frac{1}{6} \sum_{i=1}^n [a_i \hat{e}_i - x \sqrt{\text{var}_*(d_n^*)}]^3 / \{[n(4n+1) + nx^2] \text{var}_*(d_n^*)\}^{3/2}.$$

The result follows by combining this result and (10.6) and using the following results, which can be proved by some elementary analysis:

$$\sup_x |\Phi(\rho_n(x)) - \Phi(x)| = O(n^{-1}),$$

$$\sup_x |\varphi(\rho_n(x))[\rho_n^2(x) - 1] - \varphi(x)(x^2 - 1)| = O(n^{-1}),$$

and

$$\sqrt{n} \sup_x \left| \Phi(x)(x^2 - 1) \left[\gamma_n(x) - \frac{E\varepsilon_1^3}{6\sigma^3} \sum_{j=1}^n a_j^3 \right] \right| \rightarrow_{a.s.} 0. \quad \square$$

The assumption on ε_1 in condition (i) of Theorem 10.6 is stronger than that in condition (ii). However, it is known that $X'X/n \rightarrow V > 0$ and $\sup_i \|x_i\| < \infty$ imply $\max_{i \leq n} a_i^2 = O(n^{-1})$ (Wu, 1981). Therefore, the assumption on the matrix X in condition (i) is weaker than that in condition (ii). The condition $\max_{i \leq n} a_i^2 = O(n^{-1})$ is satisfied if X is E -, G -, A -, or D -optimal (Schmidt, 1979).

A random weighting estimator of the variance of $c'\hat{\beta}$ can be derived:

$$v_{\text{RW}} = \sum_{i=1}^n l_i^2 r_i^2$$

(Zheng and Tu, 1988). Note that this estimator is $\frac{n-p}{n}$ times Hinkley's weighted jackknife variance estimator (7.16) and is robust against heteroscedasticity.

Zheng (1988a) also showed that the random weighting distribution estimators are robust against heteroscedasticity.

With the variance estimator v_{RW} , it is sometimes desired to estimate the distribution of the studentized quantity

$$c'(\hat{\beta} - \beta)/\sqrt{v_{\text{RW}}}.$$

Zheng (1988b) suggested the use of a third order polynomial transformation [similar to that given by (10.5)] of

$$\sum_{i=1}^n \zeta_i a_i r_i / \left(\sum_{i=1}^n a_i^2 r_i^2 \right)^{1/2},$$

where $\zeta_1/2, \dots, \zeta_n/2$ are i.i.d. with the Gamma(4) distribution. The random weighting distribution estimator based on the transformed randomly weighted statistic is shown to be second order accurate.

A similar discussion about the random weighting estimator of the distribution of the variance estimator $\hat{\sigma}^2$ can be found in Tu (1992b).

It is interesting to note that the external bootstrap developed by Wu (1986) and Liu (1988) (Section 7.2.2) is also a random weighting method with weights being i.i.d. random variables.

10.4 Empirical Results for Random Weighting

In this section, we present two simulation examples to compare the random weighting method with the bootstrap and some other methods. The problem considered in both examples is the construction of confidence intervals for the population mean μ based on the sample mean. The first example compares methods of only first order accuracy while the second example considers methods of second order accuracy. The results are based on those in Zhang and Tu (1990) and Tu and Zheng (1991).

Simulation study 1

Let X_1, \dots, X_n be independent with a common mean μ . We are interested in constructing confidence intervals for μ based on $\sqrt{n}(\bar{X}_n - \mu)$, where \bar{X}_n is the sample mean.

From Section 4.1.5, the hybrid bootstrap confidence interval for μ with approximate confidence coefficient $1 - \alpha$ is

$$[\bar{X}_n - n^{-1/2} H_{\text{BOOT}}^{-1}(1 - \alpha/2), \bar{X}_n - n^{-1/2} H_{\text{BOOT}}^{-1}(\alpha/2)],$$

where $H_{\text{BOOT}}(x) = P_*\{\sqrt{n}(\bar{X}_n^* - \bar{X}_n) \leq x\}$ and \bar{X}_n^* is the sample mean of X_1^*, \dots, X_n^* i.i.d. from the empirical distribution of X_1, \dots, X_n .

Two hybrid random weighting confidence intervals for μ are considered. The first one, denoted by RW1, is

$$[\bar{X}_n - n^{-1/2} H_{\text{RW1}}^{-1}(1 - \alpha/2), \bar{X}_n - n^{-1/2} H_{\text{RW1}}^{-1}(\alpha/2)],$$

where $H_{\text{RW1}}(x) = P_*\{2\sqrt{n}\sum_{i=1}^n w_i(X_i - \bar{X}_n) \leq x\}$ and the joint distribution of (w_1, \dots, w_n) is $\text{Diri}(4, \dots, 4)$. H_{RW1} can be approximated by repeatedly generating random vectors (w_1, \dots, w_n) and multiplying w_i with $X_i - \bar{X}_n$. The $\text{Diri}(4, \dots, 4)$ weight vector (w_1, \dots, w_n) can be generated as follows: first generate $4(n-1)$ i.i.d. $U(0, 1)$ random variables $U_1, U_2, \dots, U_{4(n-1)}$; let $U_{(1)} < \dots < U_{(4(n-1))}$ be their order statistics; then we can take $w_1 = U_{(1)}$, $w_2 = U_{(8)} - U_{(4)}$, \dots , $w_{n-1} = U_{(4(n-1))} - U_{(4(n-2))}$, $w_n = 1 - U_{(4(n-1))}$. The second method, denoted by RW2, is

$$[\bar{X}_n - n^{-1/2} H_{\text{RW2}}^{-1}(1 - \alpha/2), \bar{X}_n - n^{-1/2} H_{\text{RW2}}^{-1}(\alpha/2)],$$

where $H_{\text{RW2}}(x) = P_*\{\frac{1}{2\sqrt{n}}\sum_{i=1}^n \xi_i(X_i - \bar{X}_n) \leq x\}$ and ξ_1, \dots, ξ_n are i.i.d. from $\text{Gamma}(4)$. H_{RW2} can be approximated by repeatedly sampling the ξ_i from the $\text{Gamma}(4)$ distribution.

In the simulation study, we take $\alpha = 0.05$. We first assume that

Table 10.1. Empirical coverage probabilities of the bootstrap and random weighting confidence intervals for μ [Adapted from Tu and Zheng (1991), by permission of JCISS]

n	Method	Distribution F					
		(1)	(2)	(3)	(4)	(5)	(6)
8	BOOT	0.818	0.694	0.802	0.790	0.822	0.814
	RW1	0.878	0.744	0.872	0.848	0.874	0.858
	RW2	0.870	0.748	0.876	0.848	0.870	0.836
10	BOOT	0.862	0.748	0.862	0.832	0.840	0.858
	RW1	0.888	0.778	0.892	0.844	0.890	0.890
	RW2	0.898	0.786	0.892	0.856	0.892	0.892
15	BOOT	0.864	0.792	0.864	0.848	0.880	0.852
	RW1	0.888	0.832	0.892	0.904	0.928	0.892
	RW2	0.880	0.838	0.898	0.904	0.926	0.890
20	BOOT	0.856	0.806	0.866	0.862	0.874	0.884
	RW1	0.892	0.840	0.926	0.912	0.930	0.924
	RW2	0.890	0.842	0.916	0.916	0.924	0.926
25	BOOT	0.864	0.790	0.862	0.868	0.880	0.888
	RW1	0.886	0.840	0.898	0.912	0.922	0.920
	RW2	0.888	0.840	0.904	0.914	0.922	0.912

X_1, \dots, X_n are i.i.d. from a distribution F , where F is taken to be (1) $N(0, 1)$, (2) Gamma(1), (3) the distribution with density

$$\int_0^1 \frac{1}{\sqrt{2\pi(1+t^2)}} \exp[-x^2/(1+t^2)] dt,$$

(4) the mixture distribution $0.9N(0, 1) + 0.1N(0, 9)$, and (5) the mixture distribution $0.95N(0, 1) + 0.05N(0, 9)$. To examine the robustness of the bootstrap and random weighting methods, we also consider (6) the situation where X_1, \dots, X_n are independent but not identically distributed: X_1, X_3, \dots, X_{n-1} are i.i.d. $N(1, 1)$ and X_2, X_4, \dots, X_n are i.i.d. Gamma(1). The simulation size is 500. Both the bootstrap and the random weighting Monte Carlo replications are 200. The empirical coverage probabilities of the three confidence intervals in various cases are given in Table 10.1.

The results show that the random weighting method has better coverage probability than the bootstrap, especially when the sample size is small and the underlying distribution of the sample deviates largely from the normal distribution. Between the two random weighting methods, the RW2 is slightly better than the RW1 when the underlying distribution of the sample is not close to the normal distribution. This pattern is also true for non-i.i.d. model [distribution (6)].

Simulation study 2

In this study, we consider confidence intervals for μ based on the studentized pivot $(\bar{X}_n - \mu)/S_n$, where $S_n^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ is also a jackknife variance estimator. Although in this case the random weighting method for statistical functionals suggested in Section 10.3.1 can be substantially simplified, we still consider the general approach and compare it with the bootstrap-t (Section 4.1.1) and the jackknifed Edgeworth expansion (Hinkley and Wei, 1984) confidence intervals (see Section 4.2.2).

The random weighting method under consideration is the one suggested at the end of Section 10.3.1, due to Zhang and Tu (1990). A detailed description of this method is given as follows. For $b = 1, \dots, B$, let $(\xi_i^{(b)}, \dots, \xi_n^{(b)})$ be the b th independent weight vector from a distribution with density $4x^3 e^{-2x} I\{x > 0\}$. Define

$$t_{nb} = -\frac{n-1}{n} \sum_{i=1}^n \xi_i^{(b)} \left(\bar{X}_{n-1,i} - \frac{1}{n} \sum_{j=1}^n \bar{X}_{n-1,j} \right) / S_n,$$

where $\bar{X}_{n-1,i}$ is the sample mean of $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$. Let $\tilde{\psi}_n$ be the transformation given by (10.5). Then the empirical distribution of $\tilde{\psi}_n(t_{nb})$, $b = 1, \dots, B$, is a random weighting approximation to the distribution of $(\bar{X}_n - \mu)/S_n$, and the random weighting confidence interval for μ

Table 10.2. Comparisons of the JE, BT, and RW confidence intervals for μ (normal distribution case)

n	$1 - \alpha$	JE		BT		RW	
		CP	EL	CP	EL	CP	EL
10	0.990	0.967	1.572	0.990	2.252	0.968	1.566
	0.975	0.934	1.368	0.972	1.820	0.939	1.345
	0.950	0.900	1.196	0.939	1.482	0.894	1.148
	0.925	0.870	1.087	0.914	1.304	0.865	1.038
	0.900	0.856	1.004	0.886	1.174	0.839	0.949
	0.875	0.831	0.936	0.866	1.082	0.819	0.883
	0.850	0.814	0.879	0.844	0.998	0.784	0.823
20	0.990	0.969	1.136	0.979	1.297	0.969	1.141
	0.975	0.953	0.989	0.964	1.108	0.952	0.988
	0.950	0.921	0.864	0.942	0.940	0.927	0.850
	0.925	0.901	0.785	0.918	0.846	0.905	0.770
	0.900	0.881	0.725	0.893	0.772	0.883	0.706
	0.875	0.858	0.677	0.876	0.717	0.860	0.659
	0.850	0.833	0.635	0.850	0.668	0.828	0.615
30	0.990	0.984	0.993	0.991	1.010	0.985	0.936
	0.975	0.967	0.812	0.975	0.870	0.968	0.813
	0.950	0.936	0.710	0.947	0.745	0.941	0.701
	0.925	0.918	0.645	0.927	0.674	0.917	0.638
	0.900	0.893	0.596	0.909	0.617	0.892	0.585
	0.875	0.876	0.556	0.884	0.576	0.864	0.547
	0.850	0.846	0.521	0.856	0.537	0.844	0.511

with approximate confidence coefficient $1 - \alpha$ is

$$[\bar{X}_n - S_n \gamma_{(1-\alpha/2)}, \bar{X}_n - S_n \gamma_{(\alpha/2)}],$$

where $\gamma_{(p)}$ is the $[(B+1)p]$ th order statistic of $\tilde{\psi}_n(t_{nb})$, $b = 1, \dots, B$.

Note that this method can also be used for general statistics T_n . One simply needs to replace \bar{X}_n , $\bar{X}_{n-1,i}$, and S_n^2 by T_n , $T_{n-1,i}$, and a variance estimator for T_n (e.g., v_{JACK}), respectively.

From the above description, we can see that random weighting requires computation of the original statistic $n(n+2)$ times, which is the same as that needed in the jackknifed Edgeworth expansion approach. The bootstrap with the jackknife variance estimator requires calculating the original

Table 10.3. Comparisons of the JE, BT, and RW confidence intervals for μ (double exponential distribution case)

n	$1 - \alpha$	JE		BT		RW	
		CP	EL	CP	EL	CP	EL
10	0.990	0.954	2.211	0.989	0.3178	0.984	2.226
	0.975	0.930	1.924	0.967	2.629	0.965	1.903
	0.950	0.899	1.683	0.936	2.188	0.935	1.624
	0.925	0.866	1.529	0.907	1.928	0.909	1.464
	0.900	0.845	1.412	0.880	1.754	0.865	1.336
	0.875	0.817	1.317	0.861	1.625	0.839	1.241
	0.850	0.789	1.236	0.825	1.504	0.814	1.155
	0.990	0.958	1.570	0.982	1.801	0.991	1.593
20	0.975	0.935	1.366	0.957	1.550	0.978	1.374
	0.950	0.899	1.194	0.924	1.325	0.954	1.176
	0.925	0.877	1.085	0.896	1.198	0.917	1.063
	0.900	0.961	1.002	0.873	1.095	0.889	0.973
	0.875	0.834	0.935	0.852	1.019	0.864	0.907
	0.850	0.816	0.877	0.830	0.949	0.834	0.845
	0.990	0.972	1.287	0.986	1.139	0.994	1.305
	0.975	0.947	1.120	0.966	1.207	0.981	1.129
30	0.950	0.922	0.980	0.932	1.039	0.948	0.970
	0.925	0.886	0.890	0.905	0.943	0.921	0.880
	0.900	0.860	0.822	0.876	0.865	0.889	0.807
	0.875	0.837	0.767	0.848	0.807	0.867	0.752
	0.850	0.805	0.719	0.824	0.753	0.831	0.702

statistic $B(n+1)$ times with B around 1000. The random weighting method is less computationally intensive if n is not very large.

In the simulation study, X_1, \dots, X_n are i.i.d. from either the standard normal distribution or the standard double exponential distribution. The empirical coverage probability (CP) and the empirical expected length (EL) of the bootstrap-t (BT), the jackknifed Edgeworth expansion (JE), and the random weighting (RW) confidence intervals are shown in Tables 10.2 and 10.3. The simulation size is 1000. The bootstrap and random weighting Monte Carlo sizes are also 1000.

It can be seen from the two tables that when X_i is distributed as $N(0, 1)$ the bootstrap has the most accurate empirical coverage probability. In this case, the random weighting performs almost the same as the jackknifed

Edgeworth expansion when $n = 10$ and becomes better as n increases for small α . For the double exponential case, the random weighting gives substantially more accurate confidence intervals than the jackknifed Edgeworth expansion method for all n and the bootstrap when $n = 20, 30$, and α is small. That is, the random weighting is more appropriate for approximating the tail probabilities of heavy tailed distribution. In almost all cases, the random weighting confidence intervals have the shortest expected lengths.

10.5 Conclusions and Discussions

- (1) The Bayesian bootstrap method provides a way to simulate the posterior distribution of a parameter in Bayesian statistical analysis. Some further investigations are still needed so that it can approximate the posterior distribution with a higher order accuracy for any given prior distribution. Compared with other simulation methods, for example, the Markov chain method, the adaptive rejection sampling, and the Gibbs samplers, the Bayesian bootstrap procedure is easier to implement and program (Newton and Raftery, 1994). A comprehensive study to compare the advantages and disadvantages of the Bayesian bootstrap and other recently proposed computer-intensive methods is still warranted.
- (2) The random weighting method was motivated by the idea of smoothing the bootstrap. Unlike the bootstrap, the generalization of the random weighting method to statistics other than the sample mean is not direct. For example, the extension to vector-valued statistics is still not available. When generalizing this method to some complex models, sometimes we need to take the model structure into account. As we discussed before, however, “serious data analysis requires serious consideration of the effect of model assumptions” (Rubin, 1981). Thus, this may be seen as an advantage of the random weighting method.
- (3) Technically, the asymptotic properties of random weighting are easier to study than those of the bootstrap because a continuous weight vector is used. This is why, in this chapter, some detailed technical discussions are presented. The basic ideas in proving the consistency and asymptotic accuracy of both the bootstrap and random weighting are essentially the same. We hope the readers gain a better understanding of the materials presented in Chapter 3 after reading this chapter.
- (4) There are some other generalizations and alternations to the original bootstrap. Some examples are:
 - (i) Stochastic procedures by Beran and Millar (1987). Stochastic procedures are estimates, tests, and confidence sets with two pro-

perties: (a) they are functions of the original sample and one or more artificially constructed auxiliary samples; and (b) they become nearly nonrandomized when the auxiliary samples increase in size. These procedures are useful as approximations to numerically intractable procedures. The classical Monte Carlo techniques, the bootstrap, and the random weighting methods are all examples of stochastic procedures. Other examples can be found in Beran and Millar (1987, 1989). They present some basic tools for investigating the asymptotic properties of stochastic procedures.

- (ii) Empirical likelihood by Owen (1988). The empirical likelihood method is an alternative to likelihood-type bootstrap methods (Hall, 1987; see Section 4.3.5) for constructing confidence regions. It amounts to computing the profile likelihood of a general multinomial distribution that has its atoms at data points. It does not require the construction of a pivotal statistic. A summary on this method can be found in Hall and La Scala (1990).
- (iii) Pao-Zhuan Yin-Yu by Fu and Li (1992). The name of this method comes from a Chinese phrase, which means “taking a humble initiative in hoping that some excellent results will be generated from the initiation”. It is a computer-intensive method that starts with a simple unbiased estimator and then generates an improved estimator in the sense of having smaller variance numerically by interpreting the Rao-Blackwell theorem (Hogg and Craig, 1970) empirically via sequential resampling. The variance of the improved estimator can also be obtained from this method. The purpose of using this method is slightly different from the bootstrap: it looks for an optimal estimator of a parameter via resampling. The applications are still limited to some parametric models.

Appendix A

Asymptotic Results

We list here some basic and important asymptotic concepts and results that are used throughout this book. The detailed proofs of these results can be found, for example, in Rao (1973), Chung (1974), Petrov (1975), Bai and Zhao (1986), Chow and Teicher (1988), Bai and Rao (1991), and Hall (1992d).

A.1 Modes of Convergence

Let X_1, X_2, \dots and X be random vectors.

(1) Convergence almost surely. If

$$P\{X_n \rightarrow X\} = 1,$$

then $\{X_n\}$ converges to X *a.s.* and we write $X_n \rightarrow_{a.s.} X$.

(2) Convergence in probability. If, for every $\epsilon > 0$,

$$P\{\|X_n - X\| > \epsilon\} \rightarrow 0,$$

then $\{X_n\}$ converges to X in probability and we write $X_n \rightarrow_p X$.

(3) Convergence in distribution. Let F_{X_n} be the distribution function of X_n , $n = 1, 2, \dots$ and F_X be the distribution function of X . If, for each continuity point x of F ,

$$F_{X_n}(x) \rightarrow F_X(x),$$

then $\{X_n\}$ converges to X in distribution and we write $X_n \rightarrow_d X$.

(4) Pólya's theorem. If $X_n \rightarrow_d X$ and F_X is continuous, then

$$\lim_{n \rightarrow \infty} \sup_x |F_{X_n}(x) - F_X(x)| = 0.$$

(5) Cramér and Wald's theorem. A sequence of random vectors $\{X_n\}$ satisfies $X_n \rightarrow_d X$ if and only if each linear combination of the components of X_n converges in distribution to the same linear combination of the components of X .

(6) Relationships among the three different modes of convergence. Convergence almost surely implies convergence in probability, and convergence in probability implies convergence in distribution.

A.2 Convergence of Transformations

(1) Continuous mapping. Let X_1, X_2, \dots and X be random p -vectors and g be a vector-valued continuous function on \mathbb{R}^p . Then

- (i) $X_n \rightarrow_{a.s.} X$ implies $g(X_n) \rightarrow_{a.s.} g(X)$;
- (ii) $X_n \rightarrow_p X$ implies $g(X_n) \rightarrow_p g(X)$;
- (iii) $X_n \rightarrow_d X$ implies $g(X_n) \rightarrow_d g(X)$.

(2) Slutsky's theorem. Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables satisfying $X_n \rightarrow_d X$ (a random variable) and $Y_n \rightarrow_p c$ (a constant). Then

- (i) $X_n + Y_n \rightarrow_d X + c$;
- (ii) $X_n Y_n \rightarrow_d cX$;
- (iii) $X_n / Y_n \rightarrow_d X/c$ if $c \neq 0$.

(3) The delta method. Let $\{X_n\}$ be a sequence of random p -vectors, $\{a_n\}$ be a sequence of numbers, and g be a function from \mathbb{R}^p to \mathbb{R} . If

$$a_n(X_n - \mu) \rightarrow_d N(0, I_p)$$

for a vector μ , and g is differentiable at μ , then

$$a_n[g(X_n) - g(\mu)] \rightarrow_d N(0, \|\nabla g(\mu)\|^2).$$

A.3 $O(\cdot)$, $o(\cdot)$, and Stochastic $O(\cdot)$, $o(\cdot)$

(1) For two sequences of numbers $\{a_n\}$ and $\{b_n\}$,

$$a_n = O(b_n)$$

if $|a_n| \leq c|b_n|$ for all n and a constant c ; and

$$a_n = o(b_n)$$

if $a_n/b_n \rightarrow 0$ as $n \rightarrow \infty$.

(2) For a sequence of $k \times k$ matrices $\{A_n\}$,

$$A_n = O(b_n)$$

if $a_n^{(ij)} = O(b_n)$ for any i and j , where $a_n^{(ij)}$ is the (i, j) th element of A_n ; and

$$A_n = o(b_n)$$

if $a_n^{(ij)} = o(b_n)$ for any i and j .

(3) For two sequences of random variables $\{X_n\}$ and $\{Y_n\}$,

$$X_n = O(Y_n) \quad a.s.$$

if $X_n = O(Y_n)$ for almost all sequences $\{X_1, X_2, \dots\}$ and $\{Y_1, Y_2, \dots\}$; and

$$X_n = o(Y_n) \quad a.s.$$

if $X_n/Y_n \rightarrow_{a.s.} 0$.

(4) For two sequences of random variables $\{X_n\}$ and $\{Y_n\}$,

$$X_n = O_p(Y_n)$$

if, for any $\epsilon > 0$, there is a constant $C_\epsilon > 0$ such that $P\{|X_n| \geq C_\epsilon |Y_n|\} < \epsilon$ for all n ; and

$$X_n = o_p(Y_n)$$

if $X_n/Y_n \rightarrow_p 0$.

(5) If $X_n \rightarrow_d X$ for a random variable X , then $X_n = O_p(1)$.

A.4 The Borel-Cantelli Lemma

Let $\{B_n\}$ be a sequence of events. If $\sum_{n=1}^{\infty} P(B_n) < \infty$, then

$$P\{B_n \text{ occurs infinitely often}\} = 0.$$

If the B_n are independent and $\sum_{n=1}^{\infty} P(B_n) = \infty$, then

$$P\{B_n \text{ occurs infinitely often}\} = 1.$$

A.5 The Law of Large Numbers

(1) The strong law of large numbers. If X_1, \dots, X_n are i.i.d. random variables with $E|X_1| < \infty$, then

$$\frac{1}{n} \sum_{i=1}^n c_i(X_i - EX_1) \rightarrow_{a.s.} 0,$$

where $\{c_i\}$ is a sequence of bounded numbers. If X_1, \dots, X_n are independent (not necessarily identically distributed) with finite means $\mu_i = EX_i$ and finite variances $\sigma_i^2 = \text{var}(X_i)$, and

$$\sum_{i=1}^{\infty} \frac{\sigma_i^2}{i^2} < \infty,$$

then

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mu_i) \rightarrow_{a.s.} 0.$$

(2) The weak law of large numbers. If X_1, \dots, X_n are independent random variables with finite means $\mu_i = EX_i$ and

$$\frac{1}{n^{1+\delta}} \sum_{i=1}^n E|X_i|^{1+\delta} \rightarrow 0$$

for a positive $\delta \leq 1$, then

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mu_i) \rightarrow_p 0.$$

(3) The Marcinkiewicz strong law of large numbers. If X_1, \dots, X_n are i.i.d. random variables and $E|X_1|^\delta < \infty$ for a $\delta \in (0, 1)$, then

$$\frac{1}{n^{1/\delta}} \sum_{i=1}^n |X_i| \rightarrow_{a.s.} 0.$$

A.6 The Law of the Iterated Logarithm

If X_1, \dots, X_n are i.i.d. random variables with a finite variance σ^2 , then

$$\limsup_n \frac{1}{\sqrt{2\sigma^2 n \log \log n}} \sum_{i=1}^n (X_i - EX_1) = 1 \quad a.s.$$

A.7 Uniform Integrability

(1) A sequence of random variables $\{X_n\}$ is uniformly integrable if

$$\lim_{c \rightarrow \infty} \sup_n E(|X_n| I\{|X_n| > c\}) = 0.$$

(2) A sufficient condition for the uniform integrability of $\{X_n\}$ is that

$$\sup_n E|X_n|^{1+\delta} < \infty$$

for a $\delta > 0$.

(3) Suppose that $X_n \rightarrow_d X$. Then

$$\lim_{n \rightarrow \infty} E|X_n|^r = E|X|^r < \infty$$

if and only if $\{|X_n|^r\}$ is uniformly integrable, where $r > 0$.

A.8 The Central Limit Theorem

(1) Let $\{X_{nj}, j = 1, \dots, k_n, n = 1, 2, \dots\}$ be a double array with independent random variables within rows. Define $\mu_{nj} = EX_{nj}$ and $B_n^2 = \text{var}(\sum_{j=1}^{k_n} X_{nj})$. If Lindeberg's condition

$$\lim_{n \rightarrow \infty} \frac{1}{B_n^2} \sum_{j=1}^{k_n} E[(X_{nj} - \mu_{nj})^2 I\{|X_{nj} - \mu_{nj}| > \epsilon B_n\}] = 0 \quad \text{for any } \epsilon > 0$$

holds, then

$$\frac{1}{B_n} \sum_{j=1}^{k_n} (X_{nj} - \mu_{nj}) \rightarrow_d N(0, 1).$$

Lindeberg's condition is implied by Liapunov's condition

$$\sum_{j=1}^{k_n} E|X_{nj} - \mu_{nj}|^v = o(B_n^v)$$

for some $v > 2$.

(2) Let $\{X_i\}$ be i.i.d. random vectors with a common mean μ and a common covariance matrix Σ . Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu) \rightarrow_d N(0, \Sigma).$$

A.9 The Berry-Esséen Theorem

If X_1, X_2, \dots are independent random variables with $EX_i = 0$, $\text{var}(X_i) = \sigma_i^2 < \infty$, and $B_n^2 = \sum_{i=1}^n \sigma_i^2 > 0$, then

$$\sup_x \left| P\left\{ \frac{1}{B_n} \sum_{i=1}^n X_i \leq x \right\} - \Phi(x) \right| \leq \frac{c}{B_n^3} \sum_{i=1}^n E|X_i|^3,$$

where c is a constant independent of n and the distributions of X_i .

A.10 Edgeworth Expansions

(1) Let $\{Z_n\}$ be a sequence of i.i.d. random p -vectors and g be a real-valued Borel measurable function on \mathbb{R}^p . Suppose that (i) Z_1 has a finite $m \geq 3$ moment, (ii) g is $m - 1$ times continuously differentiable in a neighborhood of $\mu = EZ_1$, (iii) $\nabla g(\mu) \neq 0$ and, without loss of generality, the first component of $\nabla g(\mu)$ is assumed to be positive, and (iv) $\limsup_{|t| \rightarrow \infty} E|\psi_1(t)| < 1$, where $\psi_1(t) = E[\exp(itz_{11})|z_{12}, \dots, z_{1p}]$, $i = \sqrt{-1}$, and z_{1j} is the j th component of Z_1 . Then

$$\sup_x \left| P\{W_n \leq x\} - \Phi(x) - \sum_{t=1}^{m-2} \frac{h_t(x)\varphi(x)}{n^{t/2}} \right| = o\left(\frac{1}{n^{(m-2)/2}}\right),$$

where

$$W_n = \sqrt{n}[g(\bar{Z}_n) - g(\mu)]/[\nabla g(\mu)' \Sigma \nabla g(\mu)]^{1/2},$$

$\bar{Z}_n = n^{-1} \sum_{i=1}^n Z_i$, $\Sigma = \text{var}(Z_1)$, $\varphi(x) = d\Phi(x)/dx$, and $h_t(x)$ is a polynomial of degree at most $3t$ whose coefficients are functions of the first $t + 2$ moments of Z_1 and the partial derivatives of g at μ .

For example, if $Z_i = (X_i, X_i^2)$, $i = 1, \dots, n$, $g(x, y) = (x - \mu)/\sqrt{(y - x^2)}$, where X_1, \dots, X_n are i.i.d. random variables with mean μ , and the above conditions on g and Z_1 are satisfied, then the result holds with

$$W_n = \sqrt{n}(\bar{X}_n - \mu) \Big/ \left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \right]^{1/2},$$

the studentized sample mean. In particular, we know that

$$h_1(x) = \frac{\gamma(2x^2 + 1)}{6}$$

and

$$h_2(x) = x \left[\frac{\kappa(x^2 - 3)}{12} - \frac{\gamma^2(x^4 + 2x^2 - 3)}{18} - \frac{x^2 + 3}{4} \right],$$

where $\gamma = E(X_1 - \mu)^3/\sigma^3$, $\sigma^2 = \text{var}(X_1)$, and $\kappa = E(X_1 - \mu)^4/\sigma^4 - 3$.

(2) Let $\{X_j\}$ be a sequence of independent random variables with $EX_j = 0$ and $\text{var}(X_j) = \sigma_j^2$. Let $\psi_j(t)$ be the characteristic function of X_j and $\gamma_{\nu j}$ be its ν th cumulant defined as the coefficient of the ν th term in the expansion

$$\log[\psi_j(t)] = \sum_{\nu=2}^{\infty} \frac{i^{\nu} t^{\nu} \gamma_{\nu j}}{\nu!},$$

where $i = \sqrt{-1}$. Define

$$B_n^2 = \sum_{i=1}^n \sigma_i^2, \quad \lambda_{\nu n} = \frac{n^{(\nu-2)/2}}{B_n^{\nu}} \sum_{j=1}^n \gamma_{\nu j},$$

$\nu \geq 2$, and

$$Q_{\nu n}(x) = \sum (-1)^{\nu+2s} \frac{d^{\nu+2s}\Phi(x)}{dx^{\nu+2s}} \prod_{m=1}^{\nu} \frac{1}{k_m!} \left[\frac{\lambda_{(m+2)n}}{(m+2)!} \right]^{k_m},$$

where \sum is the sum over all nonnegative integers k_1, \dots, k_ν such that $\sum_{m=1}^{\nu} m k_m = \nu$ and $\sum_{m=1}^{\nu} k_m = s$. If $E|X_j|^k < \infty$ for some $k \geq 3$ and all j , then

$$\begin{aligned} \left| P\left\{ \frac{1}{B_n} \sum_{i=1}^n X_i \leq x \right\} - \sum_{\nu=1}^{k-2} \frac{Q_{\nu n}(x)}{n^{\nu/2}} \right| &\leq c_k \left[\frac{1}{B_n^k (1+|x|)^k} \sum_{j=1}^n E|W_{nj}(x)|^k \right. \\ &\quad \left. + \frac{1}{B_n^{k+1} (1+|x|)^{k+1}} \sum_{j=1}^n E|Z_{nj}(x)|^{k+1} + \frac{n^{k(k+1)/2}}{(1+|x|)^{k+1}} U_n \right] \end{aligned}$$

uniformly in x , where c_k is a constant that depends only on k ,

$$W_{nj}(x) = X_j I\{|X_j| > B_n(1+|x|)\},$$

$$Z_{nj}(x) = X_j I\{|X_j| \leq B_n(1+|x|)\},$$

$$U_n = \left(\sup_{|t| \geq \delta_n} \frac{1}{n} \sum_{j=1}^n |\psi_j(t)| + \frac{1}{2n} \right)^n,$$

and $\delta_n = B_n^2 / (12 \sum_{j=1}^n E|X_j| I\{|X_j| \leq B_n\})$. Consider the case where $k = 3$. Suppose that (i) there are constants p and p_1 satisfying $0 \leq p_1 < \min[p, \frac{3}{8}(1+p)]$ and a constant c such that $E|X_n|^{3(1+p)} \leq cn^{p_1}$; (ii) $\limsup_n \frac{1}{n} \sum_{i=1}^n E|X_i|^3 < \infty$; (iii) $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \text{var}(X_i) = \sigma_0^2 > 0$; and (iv) for any $\epsilon > 0$, there exists $\delta_\epsilon > 0$ such that $\liminf_n (n_{\epsilon, \delta_\epsilon}/n) > 0$, where $n_{\epsilon, \delta_\epsilon}$ is the number of integers j such that $1 \leq j \leq n$, $|\psi_j(t)| \leq 1 - \delta_\epsilon$ for all $|t| \geq \epsilon$. Then

$$\sup_x \left| P\left\{ \frac{1}{B_n} \sum_{i=1}^n X_i \leq x \right\} - \Phi(x) + \frac{\varphi(x)(x^2 - 1)}{6B_n^3} \sum_{i=1}^n EX_i^3 \right| = o\left(\frac{1}{\sqrt{n}}\right).$$

(3) Let $\{X_i\}$ be a sequence of i.i.d. random variables with $EX_1 = 0$, γ_ν be the ν th cumulant of X_1 , $\sigma^2 = \text{var}(X_1)$, and $Q_\nu(x)$ be the function $Q_{\nu n}(x)$ defined above with $\lambda_{\nu n}$ replaced by γ_ν/σ^ν . If $E|X_1|^k < \infty$ for some integer $k \geq 3$ and X_1 satisfies Cramér's condition, i.e., $\limsup_{|t| \rightarrow \infty} |\psi(t)| < 1$, then

$$\sup_x \left| P\left\{ \frac{1}{\sqrt{n}\sigma} \sum_{i=1}^n X_i \leq x \right\} - \Phi(x) - \sum_{\nu=1}^{k-2} \frac{Q_\nu(x)}{n^{\nu/2}} \right| = o\left(\frac{1}{n^{(k-2)/2}}\right).$$

In particular,

$$Q_1(x) = \frac{\gamma}{6}(1 - x^2)\varphi(x)$$

and

$$Q_2(x) = -x \left[\frac{\kappa}{24}(x^2 - 3) + \frac{\gamma^2}{72}(x^4 - 10x^2 + 15) \right] \varphi(x),$$

where $\gamma = EX_1^3/\sigma^3$ and $\kappa = EX_1^4/\sigma^4 - 3$.

A.11 Cornish-Fisher Expansions

Let $\{X_i\}$ be a sequence of i.i.d. random p -vectors with $\mu = EX_1$ and $\Sigma = \text{var}(X_1)$. Let $\xi_n = \sqrt{n}A(\bar{X}_n)$, where A is a function from \mathbb{R}^p to \mathbb{R} satisfying $A(\mu) = 0$, $\sigma^2 = \nabla A(\mu)' \Sigma \nabla A(\mu)$, and $\hat{\sigma}^2 = \nabla A(\bar{X}_n)' \hat{\Sigma} \nabla A(\bar{X}_n)$ (assuming that ∇A is a known function). Suppose that the following Edgeworth expansions hold:

$$P\{\xi_n/\sigma \leq x\} = \Phi(x) + \sum_{j=1}^k \frac{p_j(x)\varphi(x)}{n^{j/2}} + O\left(\frac{1}{n^{(k+1)/2}}\right)$$

and

$$P\{\xi_n/\hat{\sigma} \leq x\} = \Phi(x) + \sum_{j=1}^k \frac{q_j(x)\varphi(x)}{n^{j/2}} + O\left(\frac{1}{n^{(k+1)/2}}\right)$$

uniformly in x , where $p_j(x)$ and $q_j(x)$ are polynomials of degree $3j - 1$ and are odd (even) functions if $3j - 1$ is even (odd). Then the following (inverse) Cornish-Fisher expansions hold:

$$x_\alpha = z_\alpha + \sum_{j=1}^k \frac{p_{j1}(z_\alpha)}{n^{j/2}} + O\left(\frac{1}{n^{(k+1)/2}}\right)$$

and

$$y_\alpha = z_\alpha + \sum_{j=1}^k \frac{q_{j1}(z_\alpha)}{n^{j/2}} + O\left(\frac{1}{n^{(k+1)/2}}\right),$$

where $z_\alpha = \Phi^{-1}(\alpha)$, x_α and y_α are defined by

$$\alpha = P\{\xi_n/\sigma \leq x_\alpha\} = P\{\xi_n/\hat{\sigma} \leq y_\alpha\},$$

and $p_{j1}(x)$ and $q_{j1}(x)$ are odd (even) polynomials of degree $j + 1$ when j is even (odd) and can be defined in terms of $p_i(x)$ and $q_i(x)$ with $i \leq j$. In particular,

$$p_{11}(x) = -p_1(x) \quad \text{and} \quad p_{21}(x) = p_1(x)p'_1(x) - \frac{1}{2}xp_1^2(x) - p_2(x)$$

[with similar formulas for $q_{11}(x)$ and $q_{21}(x)$].

Appendix B

Notation

\mathbb{R} and \mathbb{R}^p : the real line and the p -dimensional Euclidean space.

$\{a, b\}$: the set consisting of the elements a and b .

$\{a_n\}$: a sequence of vectors or random vectors a_1, a_2, \dots

$a_n \rightarrow a$: $\{a_n\}$ converges to a as n increases to ∞ .

$a_n \not\rightarrow a$: $\{a_n\}$ does not converge to a .

$I\{A\}$: the indicator function of the set A .

A^c : the complement of the set A .

(a, b) and $[a, b]$: the open and closed intervals from a to b .

g' , g'' , and $g^{(k)}$: the first, second, and k th order derivatives of a function g on \mathbb{R} .

∇g and $\nabla^2 g$: the first and second order derivatives of a function g on \mathbb{R}^p .

$g(x+)$ or $g(x-)$: the right or the left limit of the function g at x .

$\|h\|_\infty$: the sup-norm of a function h on \mathbb{R}^p , $\|h\|_\infty = \sup_x |h(x)|$.

F^{-1} : the quantile function of a distribution F , $F^{-1}(t) = \inf\{x : F(x) \geq t\}$, $t \in (0, 1)$.

$\binom{n}{k}$: defined as $n!/[k!(n - k)!]$, $n! = n \times (n - 1) \times \cdots \times 2 \times 1$.

$[t]$: the integer part of $t \in \mathbb{R}$.

A' , $|A|$, A^{-1} , and $\text{tr}(A)$: the transpose, determinant, inverse, and trace of a matrix A .

$A \geq B$: the matrix $A - B$ is nonnegative definite.

$A > B$: the matrix $A - B$ is positive definite.

$A^{1/2}$: the square root of a nonnegative definite matrix A , $A^{1/2}A^{1/2} = A$.

$A^{-1/2}$: the inverse of $A^{1/2}$.

I_p : the $p \times p$ identity matrix.

$\|x\|$: the Euclidean norm of a vector $x \in \mathbb{R}^p$, $\|x\|^2 = x'x$.

$P\{A\}$ or $P(A)$: the probability of the set A .

$E(X)$ or EX : the expectation of a random variable (vector) X .

$\text{var}(X)$: the variance (covariance matrix) of a random variable (vector) X .

$\text{cov}(X, Y)$: the covariance between random variables X and Y .

P_* , E_* , var_* , and cov_* : the bootstrap probability, expectation, variance, and covariance, conditioned on the original observed data.

$N_p(\mu, \Sigma)$ or $N(\mu, \Sigma)$: the p -variate normal random vector with mean vector μ and covariance matrix Σ .

$X \sim N(\mu, \Sigma)$: X is distributed as $N(\mu, \Sigma)$.

$\varphi(x)$: the standard normal density, $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

$\Phi(x)$: the standard normal distribution, $\Phi(x) = \int_{-\infty}^x \varphi(t)dt$.

$\Gamma(x)$: the gamma function, $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$.

z_α : the α th quantile of Φ , $z_\alpha = \Phi^{-1}(\alpha)$.

$U(0, 1)$: the uniform random variable on $(0, 1)$.

δ_x : the distribution degenerated at $x \in \mathbb{R}^p$.

$X_{(i)}$: the i th order statistic of X_1, \dots, X_n .

\bar{X}_n : the sample mean of X_1, \dots, X_n , $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$.

F_n : the empirical distribution of X_1, \dots, X_n , $F_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$.

$\hat{\theta}$, $\hat{\beta}$, and $\hat{\gamma}$: estimators of parameters θ , β , and γ .

H_0 and H_1 : null and alternative hypotheses.

i.i.d.: independent and identically distributed.

a.s.: almost surely.

$\rightarrow_{a.s.}$: convergence almost surely.

\rightarrow_p : convergence in probability.

\rightarrow_d : convergence in distribution.

\square : end of a proof.

References

- Abramovitch, L. and Singh, K. (1985). Edgeworth corrected pivotal statistics and the bootstrap, *Ann. Statist.*, **13**, 116–132.
- Adkins, L. C. and Hill, R. C. (1990). An improved confidence ellipsoid for the linear regression models, *J. Statist. Compu. Simul.*, **36**, 9–18.
- Akaike, H. (1970). Statistical predictor identification, *Ann. Inst. Statist. Math.*, **22**, 203–217.
- Akritas, M. G. (1986). Bootstrapping the Kaplan–Meier estimator, *J. Amer. Statist. Assoc.*, **81**, 1032–1038.
- Alemayeha, D. (1987). Bootstrap method for multivariate analysis, *Proceedings of Statistical Computations*, 321–324, American Statistical Association.
- Alemayeha, D. (1988). Bootstrapping the latest roots of certain random matrices, *Comm. Statist. B*, **17**, 857–869.
- Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction, *Technometrics*, **16**, 125–127.
- Arcones, M. A. and Giné, E. (1989). The bootstrap of the mean with arbitrary bootstrap sample size, *Ann. Inst. Henri Poincaré*, **25**, 457–481.
- Arcones, M. A. and Giné, E. (1991). Some bootstrap tests of symmetry for univariate continuous distributions, *Ann. Statist.*, **19**, 1496–1551.
- Arvesen, J. N. (1969). Jackknifing U-statistics, *Ann. Math. Statist.*, **40**, 2076–2100.
- Athreya, K. B. (1987). Bootstrap of the mean in the infinite variance case, *Ann. Statist.*, **14**, 724–731.
- Athreya, K. B. and Fuh, C. D. (1992a). Bootstrapping Markov chains: Countable case, *J. Statist. Plan. Inference*, **33**, 311–331.

- Athreya, K. B. and Fuh, C. D. (1992b). Bootstrapping Markov chains, *Exploring the Limit of Bootstrap*, R. LePage and L. Billard eds., 49–64, Wiley, New York.
- Babu, G. J. (1984). Bootstrapping statistics with linear combinations of chi-square as weak limit, *Sankhyā A*, **46**, 85–93.
- Babu, G. J. (1986). A note on bootstrapping the variance of sample quantiles, *Ann. Inst. Statist. Math.*, **38**, 439–443.
- Babu, G. J. and Bose, A. (1989). Bootstrap confidence intervals, *Statist. Prob. Letters*, **7**, 151–160.
- Babu, G. J. and Singh, K. (1983). Inference on means using the bootstrap, *Ann. Statist.*, **11**, 999–1003.
- Babu, G. J. and Singh, K. (1984a). Asymptotic representations related to jackknifing and bootstrapping L-statistics, *Sankhyā A*, **46**, 195–206.
- Babu, G. J. and Singh, K. (1984b). On one term Edgeworth correction by Efron's bootstrap, *Sankhyā A*, **46**, 219–232.
- Babu, G. J. and Singh, K. (1989). On Edgeworth expansions in the mixture cases, *Ann. Statist.*, **17**, 443–447.
- Bahadur, R. R. and Savage, L. J. (1956). The nonexistence of certain statistical procedures in nonparametric problems, *Ann. Math. Statist.*, **27**, 1115–1122.
- Bai, C. and Olshen, B. A. (1988). Discussion of “Theoretical comparison of bootstrap confidence intervals” by P. Hall, *Ann. Statist.*, **16**, 953–956.
- Bai, C., Bickel, P. J. and Olshen, R. A. (1990). Hyperaccuracy of bootstrap based prediction, *Probability in Banach Spaces VII*, E. Eberlein, J. Kuelbs and M. B. Marcus eds., 31–42, Birkhäuser, Boston.
- Bai, Z. and Rao, C. R. (1991). Edgeworth expansion of a function of sample means, *Ann. Statist.*, **19**, 1295–1315.
- Bai, Z. and Zhao, L. (1986). Edgeworth expansions of distribution function of independent random variables, *Scientia Sinica A*, **29**, 1–22.
- Banks, D. L. (1988). Histogram smoothing the Bayesian bootstrap, *Biometrika*, **75**, 673–684.
- Barlow, W. E. and Sun, W. H. (1989). Bootstrapped confidence intervals for linear relative risk models, *Statist. Medicine*, **8**, 927–935.
- Basawa, I. V., Green, T. A., McCormick, W. P. and Taylor, R. L. (1990). Asymptotic bootstrap validity for finite Markov chains, *Comm. Statist. A*, **19**, 1493–1510.
- Basawa, I. V., Mallik, A. K. McCormick, W. P. and Taylor, R. L. (1989). Bootstrapping explosive autoregressive processes, *Ann. Statist.*, **17**, 1479–1486.

- Basawa, I. V., Mallik, A. K., McCormick, W. P., Reeves, J. H. and Taylor, R. L. (1991a). Bootstrapping unstable first order autoregressive processes, *Ann. Statist.*, **19**, 1098–1101.
- Basawa, I. V., Mallik, A. K., McCormick, W. P., Reeves, J. H. and Taylor, R. L. (1991b). Bootstrap test of significance and sequential bootstrap estimation for unstable first order autoregressive processes, *Comm. Statist. A*, **20**, 1015–1026.
- Beran, R. (1984a). Jackknife approximations to bootstrap estimates, *Ann. Statist.*, **12**, 101–118.
- Beran, R. (1984b). Bootstrap in statistics, *Jber. d. Dt. Math-verin*, **86**, 14–30.
- Beran, R. (1986). Simulated power functions, *Ann. Statist.*, **14**, 151–173.
- Beran, R. (1987). Preprinting to reduce level error of confidence sets, *Biometrika*, **74**, 151–173.
- Beran, R. (1988a). Preprinting test statistics: A bootstrap view of asymptotic refinements, *J. Amer. Statist. Assoc.*, **83**, 687–697.
- Beran, R. (1988b). Balanced simultaneous confidence sets, *J. Amer. Statist. Assoc.*, **83**, 679–686.
- Beran, R. (1990). Refining bootstrap simultaneous confidence sets, *J. Amer. Statist. Assoc.*, **85**, 417–428.
- Beran, R. (1992). Designing bootstrap prediction regions, *Bootstrapping and Related Techniques*, K. H. Jöckel, G. Rothe and W. Sendle eds., 23–30, Springer-Verlag, Berlin.
- Beran, R. and Ducharme, G. R. (1991). *Asymptotic Theory for Bootstrap Methods in Statistics*, Les Publications Centre de Recherches Mathématiques, Université de Montréal, Montréal.
- Beran, R. and Millar, P. W. (1985). Asymptotic theory of confidence sets, *Proceedings of Berkeley Conference in Honor of J. Neyman and J. Kiefer*, L. M. LeCam and R. A. Olshen eds., vol. **2**, 865–887, Wadsworth, Monterey, CA.
- Beran, R. and Millar, P. W. (1986). Confidence sets for a multivariate distribution, *Ann. Statist.*, **14**, 431–443.
- Beran, R. and Millar, P. W. (1987). Stochastic estimation and testing, *Ann. Statist.*, **15**, 1131–1154.
- Beran, R. and Millar, P. W. (1989). A stochastic minimum distance test for multivariate parametric models, *Ann. Statist.*, **17**, 125–140.
- Beran, R. and Srivastava, M. S. (1985). Bootstrap tests and confidence regions for functions of covariance matrix, *Ann. Statist.*, **13**, 95–115.
- Beran, R., LeCam, L. and Millar, P. W. (1987). Convergence of stochastic empirical measures, *J. Multivariate Anal.*, **23**, 159–168.

- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, second edition, Springer-Verlag, New York.
- Bhattacharya, R. N. and Ghosh, J. K. (1978). On the validity of the formal Edgeworth expansion, *Ann. Statist.*, **6**, 434–485.
- Bhattacharya, R. N. and Qumsiyeh, M. (1989). Second order and L^p -comparisons between the bootstrap and empirical Edgeworth expansion methodologies, *Ann. Statist.*, **17**, 160–169.
- Bickel, P. J. (1992). Theoretical comparison of different bootstrap-t confidence bounds, *Exploring the Limit of Bootstrap*, R. LePage and L. Billard eds., 65–76, Wiley, New York.
- Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics*, Holden-Day, San Francisco.
- Bickel, P. J. and Freedman, D. A. (1980). On Edgeworth expansions for bootstrap, Tech. Report, Dept. of Statist., University of California, Berkeley.
- Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap, *Ann. Statist.*, **9**, 1196–1217.
- Bickel, P. J. and Freedman, D. A. (1983). Bootstrapping regression models with many parameters, *A Festschrift for Erich L. Lehmann*, P. J. Bickel, K. Doksum and J. L. Hodges eds., 28–48, Wadsworth, Belmont, CA.
- Bickel, P. J. and Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling, *Ann. Statist.*, **12**, 470–482.
- Bickel, P. J. and Ghosh, J. K. (1990). A decomposition for the likelihood ratio statistic and the Bartlett correction—a Bayesian argument, *Ann. Statist.*, **18**, 1070–1090.
- Bickel, P. J. and Krieger, A. M. (1989). Confidence bands for a distribution function using the bootstrap, *J. Amer. Statist. Assoc.*, **84**, 95–1000.
- Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates, *Ann. Statist.*, **1**, 1071–1095.
- Bickel, P. J. and Yahav, J. A. (1988). Richardson extrapolation and the bootstrap, *J. Amer. Statist. Assoc.*, **83**, 381–393.
- Billingsley, P. (1979). *Probability and Measure*, Wiley, New York.
- Bloch, D. A. and Gastwirth, J. L. (1968). On a simple estimate of reciprocal of the density function, *Ann. Math. Statist.*, **39**, 1083–1085.
- Boos, D. D. and Brownie, C. (1989). Bootstrap methods for testing homogeneity of variances, *Technometrics*, **31**, 69–82.
- Boos, D. D. and Monahan, J. F. (1986). Bootstrap methods using prior information, *Biometrika*, **73**, 77–83.

- Boos, D. D., Janssen, P. and Veraverbeke, N. (1989). Resampling from centered data in the two sample problem, *J. Statist. Plan. Inference*, **21**, 327–345.
- Booth, J. G. and Hall, P. (1993a). An improvement of the jackknife distribution function estimator, *Ann. Statist.*, **21**, 1476–1485.
- Booth, J. G. and Hall, P. (1993b). Bootstrap confidence regions for functional relationships in error-in-variables models, *Ann. Statist.*, **21**, 1780–1791.
- Booth, J. G. and Hall, P. (1994). Monte-Carlo approximation and the iterated bootstrap, *Biometrika*, **81**, 331–340.
- Booth, J. G., Butler, R. W. and Hall, P. (1994). Bootstrap methods for finite populations, *J. Amer. Statist. Assoc.*, **89**, 1282–1289.
- Booth, J. G., Hall, P. and Wood, A. T. A. (1993). Balanced importance resampling for the bootstrap. *Ann. Statist.*, **21**, 286–298.
- Bose, A. (1988). Edgeworth correction by bootstrap in autoregressions, *Ann. Statist.*, **16**, 1709–1722.
- Bose, A. (1990). Bootstrap in moving average models, *Ann. Inst. Statist. Math.*, **42**, 753–768.
- Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates, *Biometrika*, **71**, 353–360.
- Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco.
- Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*, Wadsworth, Belmont, CA.
- Bühlmann, P. (1994). Blockwise bootstrap empirical process for stationary sequences, *Ann. Statist.*, **22**, 995–1012.
- Bunke, O. and Droege, B. (1984). Bootstrap and cross validation estimates of the prediction error for linear regression models, *Ann. Statist.*, **12**, 1400–1424.
- Bunke, O. and Riemer, S. (1983). A note on bootstrap and other empirical procedures for testing linear hypotheses without normality, *Statistics*, **14**, 517–526.
- Burman, P. (1989). A comparative study of ordinary cross-validation, vhold cross-validation and the repeated learning-testing methods, *Biometrika*, **76**, 503–514.
- Burman, P., Chow, E. and Nolan, D. (1994). A cross-validatory method for dependent data, *Biometrika*, **81**, 351–358.
- Burr, D. (1994). A comparison of certain bootstrap confidence intervals in the Cox model, *J. Amer. Statist. Assoc.*, **89**, 1290–1302.

- Cao-Abad, R. (1991). Rate of convergence for the wild bootstrap in non-parametric regression, *Ann. Statist.*, **19**, 2226–2231.
- Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence, *Ann. Statist.*, **14**, 1171–1194.
- Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler, *Amer. Statist.*, **46**, 167–174.
- Chan, Y. M. and Srivastava, M. S. (1988). Comparison of powers for the sphericity tests using both the asymptotic distribution and the bootstrap method, *Comm. Statist. A*, **17**, 671–690.
- Chao, M. T. and Lo, S. H. (1985). A bootstrap method for finite populations, *Sankhyā A*, **47**, 399–405.
- Chatterjee, S. (1984). Variance estimation in factor analysis, an application of bootstrap, *British J. Math. Statist. Psycho.*, **37**, 252–262.
- Chatterjee, S. (1986). Bootstrapping ARMA models: Some simulations, *IEEE Trans. Systems, Man and Cybernetics*, SMC-**16**, 294–299.
- Chatterjee, S. and Chatterjee, S. (1983). Estimation of misclassification probabilities by bootstrap methods, *Comm. Statist. B*, **12**, 645–656.
- Chen, H. and Loh, W. Y. (1991). Consistency of bootstrap for the transformed two sample t-tests, *Comm. Statist. A*, **20**, 997–1014.
- Chen, J. and Sitter, R. R. (1993). Edgeworth expansions and the bootstrap for stratified sampling without replacement from a finite population, *Canadian J. Statist.*, **21**, 347–357.
- Chen, Y. and Tu, D. (1987). Estimating the error rate in discriminant analysis: By the delta, jackknife and bootstrap methods, *Chinese J. Appl. Prob. Statist.*, **3**, 203–210.
- Chen, Z. and Do, K. A. (1992). Importance resampling for the smoothed bootstrap, *J. Statist. Compu. Simul.*, **40**, 107–124.
- Chow, Y. S. and Teicher, H. (1988). *Probability Theory*, second edition, Springer-Verlag, New York.
- Chung, K. L. (1974). *A Course in Probability Theory*, second edition, Academic Press, New York.
- Clarke, B. R. (1983). Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations, *Ann. Statist.*, **11**, 1196–1205.
- Clarke, B. R. (1986). Nonsmooth analysis and Fréchet differentiability of M-functionals, *Prob. Theory and Related Fields*, **73**, 197–209.
- Cochran, W. G. (1977). *Sampling Techniques*, third edition, Wiley, New York.

- Cox, D. R. (1972). Regression models and life tables, *J. R. Statist. Soc. B*, **34**, 187–220.
- Cox, D. R. and Snell, E. J. (1968). A general definition of residuals (with discussions), *J. R. Statist. Soc. B*, **30**, 248–265.
- Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation, *Numer. Math.*, **31**, 377–403.
- Cressie, N. A. C. (1991). *Statistics for Spatial Data*, Wiley, New York.
- Csörgő, M. (1983). *Quantile Processes With Statistical Applications*, SIAM, Philadelphia.
- Csörgő, S. and Mason, D. M. (1989). Bootstrap empirical functions, *Ann. Statist.*, **17**, 1447–1471.
- Daniels, H. E. and Young, G. A. (1991). Saddle point approximation for studentized mean, with applications to the bootstrap, *Biometrika*, **78**, 169–179.
- Datta, S. (1992). A note on continuous Edgeworth expansions and the bootstrap, *Sankhyā A*, **54**, 171–182.
- Datta, S. and McCormick, W. P. (1992). Bootstrap for a finite state Markov chain based on i.i.d. resampling, *Exploring the Limit of Bootstrap*, R. LePage and L. Billard eds., 77–97, Wiley, New York.
- Datta, S. and McCormick, W. P. (1993). Regeneration-based bootstrap for Markov chains, *Canadian J. Statist.*, **21**, 181–193.
- Davison, A. C. and Hall, P. (1992). On the bias and variability of bootstrap and cross-validation estimates of error rate in discriminant analysis, *Biometrika*, **79**, 279–284.
- Davison, A. C. and Hall, P. (1993). On studentizing and blocking methods for implementing the bootstrap with dependent data, *Austral. J. Statist.*, **35**, 215–224.
- Davison, A. C. and Hinkley, D. V. (1988). Saddle point approximations in resampling methods, *Biometrika*, **75**, 417–431.
- Davison, A. C., Hinkley, D. V. and Schechtman, E. (1986). Efficient bootstrap simulations, *Biometrika*, **73**, 555–566.
- Davison, A. C., Hinkley, D. V. and Worton, B. J. (1992). Bootstrap likelihoods, *Biometrika*, **79**, 113–130.
- DeAngelis, D. and Young, G. A. (1992). Smoothing the bootstrap, *Inter. Statist. Review*, **60**, 45–56.
- DeAngelis, D., Hall, P. and Young, G. A. (1993). Analytical and bootstrap approximations to estimator distributions in L_1 regression, *J. Amer. Statist. Assoc.*, **88**, 1310–1316.

- DeBeer, C. F. and Swanepoel, J. W. H. (1989). A modified Durbin-Watson test for serial correlation in multiple regression under non-normality using the bootstrap, *J. Statist. Compu. Simul.*, **33**, 75–82.
- Deheuvels, P., Mason, D. M. and Shorack, G. R. (1993). Some results on the influence of extremes on the bootstrap, *Ann. Inst. Henri Poincaré*, **29**, 83–103.
- DeWet, T. and van Wyk, J. W. J. (1986). Bootstrap confidence intervals for regression coefficients when the residuals are dependent, *J. Statist. Compu. Simul.*, **23**, 317–327.
- Diaconis, P. and Efron, B. (1983). Computer intensive methods in statistics, *Scientific American*, May, 116–130.
- DiCiccio, T. J. and Efron, B. (1992). More accurate confidence intervals in exponential families, *Biometrika*, **79**, 231–245.
- DiCiccio, T. J. and Romano, J. P. (1988a). Discussion of “Theoretical comparison of bootstrap confidence intervals” by P. Hall, *Ann. Statist.*, **16**, 965–969.
- DiCiccio, T. J. and Romano, J. P. (1988b). A review of bootstrap confidence intervals (with discussions), *J. R. Statist. Soc. B*, **50**, 338–354.
- DiCiccio, T. J. and Romano, J. P. (1989). The automatic percentile method: Accurate confidence limits in parametric models, *Canadian J. Statist.*, **17**, 155–169.
- DiCiccio, T. J. and Romano, J. P. (1990). Nonparametric confidence limits by resampling methods and least favorable families, *Inter. Statist. Review*, **58**, 59–76.
- DiCiccio, T. J. and Tibshirani, R. J. (1987). Bootstrap confidence intervals and bootstrap approximations, *J. Amer. Statist. Assoc.*, **82**, 163–170.
- DiCiccio, T. J., Martin, M. A. and Young, G. A. (1992a). Fast and accurate approximate double bootstrap confidence intervals, *Biometrika*, **79**, 285–295.
- DiCiccio, T. J., Martin, M. A. and Young, G. A. (1992b). Analytic approximation for iterated bootstrap confidence intervals, *Statist. Computing*, **2**, 161–171.
- Dikta, G. (1990). Bootstrap approximation of nearest neighbor regression function estimates, *J. Multivariate Anal.*, **32**, 213–229.
- Do, K. A. (1992). A simulation study of balanced and antithetic bootstrap resampling methods, *J. Statist. Compu. Simul.*, **40**, 153–156.
- Do, K. A. and Hall, P. (1991). On importance resampling for the bootstrap, *Biometrika*, **78**, 161–167.

- Do, K. A. and Hall, P. (1992). Distribution estimation using concomitants of order statistics with applications to Monte Carlo simulation for the bootstrap, *J. R. Statist. Soc. B*, **54**, 595–607.
- Dodge, Y. (ed.) (1987). *Statistical Data Analysis Based on L_1 Norm and Related Topics*, North-Holland, Amsterdam.
- Doss, H. and Chiang, Y. C. (1994). Choosing the resampling scheme when bootstrapping: A case study in reliability, *J. Amer. Statist. Assoc.*, **89**, 298–308.
- Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*, second edition, Wiley, New York.
- Ducharme, G. R. and Jhun, M. (1986). A note on the bootstrap procedure in testing linear hypotheses, *Statistics*, **17**, 527–531.
- Ducharme, G. R., Jhun, M., Romano, J. P. and Truong, K. N. (1985). Bootstrap confidence cones for directional data, *Biometrika*, **72**, 637–645.
- Dudley, R. M. (1978). Central limit theorem for empirical measures, *Ann. Prob.*, **6**, 899–929.
- Duncan, G. T. (1978). An empirical study of jackknife-constructed confidence region in nonlinear regression, *Technometrics*, **20**, 123–129.
- Duttweiler, D. L. (1973). The mean-square error of Bahadur's orderstatistic approximation, *Ann. Statist.*, **1**, 446–453.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife, *Ann. Statist.*, **7**, 1–26.
- Efron, B. (1981a). Nonparametric standard errors and confidence intervals (with discussions), *Canadian J. Statist.*, **9**, 139–172.
- Efron, B. (1981b). Censored data and bootstrap, *J. Amer. Statist. Assoc.*, **76**, 312–319.
- Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*, SIAM, Philadelphia.
- Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation, *J. Amer. Statist. Assoc.*, **78**, 316–331.
- Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems, *Biometrika*, **72**, 45–58.
- Efron, B. (1987). Better bootstrap confidence intervals (with discussions), *J. Amer. Statist. Assoc.*, **82**, 171–200.
- Efron, B. (1990). More efficient bootstrap computations, *J. Amer. Statist. Assoc.*, **85**, 79–89.
- Efron, B. (1992a). Jackknife-after-bootstrap standard errors and influence functions (with discussions), *J. R. Statist. Soc. B*, **54**, 83–127.

- Efron, B. (1992b). Six questions raised by the bootstrap, *Exploring the Limits of Bootstrap*, R. LePage and L. Billard eds., 99–126, Wiley, New York.
- Efron, B. (1994). Missing data, imputation, and the bootstrap (with discussion), *J. Amer. Statist. Assoc.*, **89**, 463–479.
- Efron, B. and Stein, C. (1981). The jackknife estimate of variance, *Ann. Statist.*, **9**, 586–596.
- Efron, B. and Tibshirani, R. J. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, *Statist. Science*, **1**, 54–77.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York.
- Eriksson, B. (1983). On the construction of confidence limits for the regression coefficients when the residuals are dependent, *J. Statist. Comput. Simul.*, **17**, 297–309.
- Falk, M. (1988). Weak convergence of the bootstrap process for large quantiles, *Statist. Decisions*, **6**, 385–396.
- Falk, M. (1992a). Bootstrapping the sample quantile: A survey, *Bootstrapping and Related Techniques*, K. H. Jöckel, G. Rothe and W. Sendle eds., 165–172, Springer-Verlag, Berlin.
- Falk, M. (1992b). Bootstrap optimal bandwidth selection for kernel density estimates, *J. Statist. Plan. Inference*, **30**, 13–32.
- Falk, M. and Kaufman, E. (1991). Coverage probabilities of bootstrap confidence intervals for quantiles, *Ann. Statist.*, **19**, 485–495.
- Falk, M. and Reiss, R. D. (1989). Weak convergence of smoothed and non-smoothed bootstrap quantile estimates, *Ann. Prob.*, **17**, 362–371.
- Fan, J. and Mei, C. (1991). The convergence rate of randomly weighted approximation for errors of estimated parameters of AR(1) models, *Xian Jiaotong DaXue Xuebao*, **25**, 1–6.
- Faraway, J. J. (1990). Bootstrap selection of bandwidth and confidence bands for nonparametric regression, *J. Statist. Comput. Simul.*, **37**, 37–44.
- Faraway, J. J. and Jhun, M. (1990). Bootstrap choice of bandwidth for density estimation (with discussions), *J. Amer. Statist. Assoc.*, **85**, 1119–1122.
- Fay, R. E. (1991). A design-based perspective on missing data variance, *Proceedings of the Seventh Annual Research Conference*, 429–440, Bureau of the Census, Washington, D.C.
- Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, vol. 1, third edition, Wiley, New York.

- Feluch, W. and Koronachi, J. (1992). A note on modified cross-validation in density estimates, *Compu. Statist. Data Anal.*, **13**, 143–151.
- Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems, *Ann. Statist.*, **1**, 209–230.
- Fernholz, L. T. (1983). *Von Mises Calculus for Statistical Functionals*, Lecture Notes in Statistics, **19**, Springer-Verlag, New York.
- Feuerverger, A. (1989). On the empirical saddle point approximations, *Biometrika*, **76**, 457–464.
- Fisher, N. I. and Hall, P. (1989). Bootstrap confidence regions for directional data, *J. Amer. Statist. Assoc.*, **84**, 996–1002.
- Fisher, N. I. and Hall, P. (1991). Bootstrap algorithms for small samples, *J. Statist. Plan. Inference*, **27**, 157–169.
- Folks, J. L. and Chhikara, R. S. (1978). The inverse Gaussian distribution and its statistical application—a review, *J. R. Statist. Soc. B*, **40**, 263–289.
- Fox, T., Hinkley, D. V. and Larntz, K. (1980). Jackknifing in nonlinear regression, *Technometrics*, **22**, 123–129.
- Francisco, C. A. and Fuller, W. A. (1991). Quantile estimation with a complex survey design, *Ann. Statist.*, **19**, 454–469.
- Frangos, C. C. and Schucany, W. R. (1990). Jackknife estimation of the bootstrap acceleration constant, *Compu. Statist. Data Anal.*, **9**, 271–282.
- Franke, J. and Härdle, W. (1992). On bootstrapping kernel spectral estimates, *Ann. Statist.*, **20**, 121–145.
- Freedman, D. A. (1981). Bootstrapping regression models, *Ann. Statist.*, **9**, 1218–1228.
- Freedman, D. A. (1984). On bootstrapping two-stage least squares estimates in stationary linear models, *Ann. Statist.*, **12**, 827–842.
- Freedman, D. A. and Peters, S. C. (1984). Bootstrapping a regression equation: Some empirical results, *J. Amer. Statist. Assoc.*, **79**, 97–106.
- Fu, J. C. and Li, L.-A. (1992). Method of Pao-Zhuan Yin-Yu: A method of stochastic point estimation, *Statist. Sinica*, **2**, 171–188.
- Fuller, W. A. (1976). *Introduction to Statistical Time Series*, Wiley, New York.
- Gaenssler, P. (1987). Bootstrapping empirical measures indexed by Vapnik-Cervoneckis class of sets, *Proceedings of the First World Congress of the Bernoulli Society*, Y. A. Prohorov and V. V. Sazonov eds., **1**, 467–487, VMU Science Press, Utrecht, Netherlands.

- Ganeshanandam, S. and Krzanowski, W. J. (1990). Error-rate estimation in two-group discriminant analysis using the linear discriminant function, *J. Statist. Compu. Simul.*, **36**, 157–176.
- Gangopadhyay, A. K. and Sen, P. K. (1990). Bootstrap confidence interval for conditional quantile functions, *Sankhyā A*, **52**, 346–363.
- Gaver, D. P. and Miller, R. G. (1983). Jackknifing the Kaplan-Meier survival estimator for censored data: Simulation results and asymptotic analysis, *Comm. Statist. A*, **15**, 1701–1718.
- Geisser, S. (1975). The predictive sample reuse method with applications, *J. Amer. Statist. Assoc.* **70**, 320–328.
- Ghosh, J. K. (1971). A new proof of the Bahadur representation of quantiles and an application, *Ann. Math. Statist.*, **42**, 1957–1961.
- Ghosh, M. (1985). Berry-Esséen bounds for functions of U-statistics, *Sankhyā A*, **47**, 255–270.
- Ghosh, M., Parr, W. C., Singh, K. and Babu, G. J. (1984). A note on bootstrapping the sample median, *Ann. Statist.*, **12**, 1130–1135.
- Gill, R. D. (1989). Non- and semi-parametric MLE and von-Mises methods (Part I), *Scan. J. Statist.*, **16**, 97–128.
- Giné, E. and Zinn, J. (1989). Necessary conditions for the bootstrap of the mean, *Ann. Statist.*, **17**, 684–691.
- Giné, E. and Zinn, J. (1990). Bootstrap general empirical measures, *Ann. Prob.*, **18**, 851–869.
- Gleason, J. R. (1988). Algorithms for balanced bootstrap simulations, *Amer. Statist.*, **42**, 263–266.
- Götze, F. (1989). Edgeworth expansions in functional limit theorems, *Ann. Prob.*, **17**, 1602–1634.
- Graham, R. L., Hinkley, D. V., John, P. W. M. and Shi, S. (1990). Balanced design of bootstrap simulations, *J. R. Statist. Soc. B*, **52**, 185–202.
- Gray, H. L. and Schucany, W. R. (1972). *The Generalized Jackknife Statistics*, Marcel Dekker, New York.
- Gross, S. (1980). Median estimation in sample surveys, *Proceedings of the Section on Survey Research Methods*, 181–184, American Statistical Association, Alexandria, VA.
- Gruet, M. A., Huet, S. and Jolivet, E. (1993). Practical use of bootstrap in regression, *Computer Intensive Methods in Statistics*, 150–166, Statist. Compu., Physica, Heidelberg.
- Gu, M. (1992). On the Edgeworth expansion and bootstrap approximation for the Cox regression model under random censorship, *Canadian J. Statist.*, **20**, 399–414.

- Gupta, V. K. and Nigam, A. K. (1987). Mixed orthogonal arrays for variance estimation with unequal numbers of primary selections per stratum, *Biometrika*, **74**, 735–742.
- Gurney, M. and Jewett, R. S. (1975). Constructing orthogonal replications for standard errors, *J. Amer. Statist. Assoc.*, **70**, 819–821.
- Haeusler, E., Mason, D. M. and Newton, M. A. (1991). Weighted bootstrapping of means, *Cent. Wisk. Inf. Q.*, **4**, 213–228.
- Hájek, J. (1960). Limiting distribution in simple random sampling from a finite population, *Publ. Math. Inst. Hungarian Acad. Sci.*, **5**, 361–374.
- Hall, P. (1983a). Inverting an Edgeworth expansion, *Ann. Statist.*, **11**, 569–576.
- Hall, P. (1983b). Large sample optimality of least squares cross-validation in density estimation, *Ann. Statist.*, **11**, 1156–1174.
- Hall, P. (1986). On the bootstrap and confidence intervals, *Ann. Statist.*, **14**, 1431–1452.
- Hall, P. (1987). On the bootstrap and likelihood-based confidence regions, *Biometrika*, **74**, 481–493.
- Hall, P. (1988a). Rate of convergence in bootstrap approximations, *Ann. Prob.*, **16**, 1165–1185.
- Hall, P. (1988b). Theoretical comparisons of bootstrap confidence intervals (with discussions), *Ann. Statist.*, **16**, 927–953.
- Hall, P. (1989a). On efficient bootstrap simulation, *Biometrika*, **76**, 613–617.
- Hall, P. (1989b). Antithetic resampling for the bootstrap, *Biometrika*, **76**, 713–724.
- Hall, P. (1989c). Unusual properties of bootstrap confidence intervals in regression problem, *Prob. Theory and Related Fields*, **81**, 247–273.
- Hall, P. (1990a). Asymptotic properties of the bootstrap for heavy-tailed distributions, *Ann. Prob.*, **18**, 1342–1360.
- Hall, P. (1990b). On the relative performance of bootstrap and Edgeworth approximations of a distribution function. *J. Multivariate Anal.*, **35**, 108–129.
- Hall, P. (1990c). Performance of bootstrap balanced resampling in distribution function and quantile problems, *Prob. Theory and Related Fields*, **85**, 239–267.
- Hall, P. (1990d). Using the bootstrap to estimate mean squared error and select smoothing parameters in nonparametric problems, *J. Multivariate Anal.*, **32**, 177–203.

- Hall, P. (1991a). Bahadur representations for uniform resampling and importance resampling with applications to asymptotic relative efficiency, *Ann. Statist.*, **19**, 1062–1072.
- Hall, P. (1991b). Edgeworth expansions for nonparametric density estimators with applications, *Statistics*, **22**, 215–232.
- Hall, P. (1992a). Efficient bootstrap simulations, *Exploring the Limits of Bootstrap*, R. LePage and L. Billard eds., 127–143, Wiley, New York.
- Hall, P. (1992b). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density, *Ann. Statist.*, **20**, 675–694.
- Hall, P. (1992c). On bootstrap confidence intervals in nonparametric regression, *Ann. Statist.*, **20**, 695–711.
- Hall, P. (1992d). *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.
- Hall, P. (1993). On Edgeworth expansion and bootstrap confidence bands in nonparametric curve estimation, *J. R. Statist. Soc. B*, **55**, 291–304.
- Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood, *Inter. Statist. Review*, **58**, 109–127.
- Hall, P. and Martin, M. A. (1988a). On the bootstrap and two sample problems, *Austral. J. Statist.*, **30**, 179–182.
- Hall, P. and Martin, M. A. (1988b). On bootstrap resampling and iteration, *Biometrika*, **75**, 661–671.
- Hall, P. and Martin, M. A. (1988c). Exact convergence rate of bootstrap quantile variance estimation, *Prob. Theory and Related Fields*, **80**, 261–268.
- Hall, P. and Martin, M. A. (1989). A note on the accuracy of bootstrap percentile method confidence intervals for a quantile, *Statist. Prob. Letters*, **8**, 197–200.
- Hall, P. and Pittelkow, Y. E. (1990). Simultaneous bootstrap confidence bands in regression, *J. Statist. Compu. Simul.*, **37**, 99–113.
- Hall, P. and Wilson, S. R. (1991). Two guide lines for bootstrap hypothesis testing, *Biometrics*, **47**, 757–762.
- Hall, P., DiCiccio, T. J. and Romano, J. P. (1989). On smoothing and the bootstraps, *Ann. Statist.*, **17**, 692–704.
- Hall, P., Marron, J. S. and Park, B. V. (1991). Smoothed cross-validation, *Prob. Theory and Related Fields*, **92**, 1–20.
- Hampel, F. R. (1974). The influence curve and its role in robust estimation, *J. Amer. Statist. Assoc.*, **69**, 383–393.

- Hannan, E. J., McDougall, A. L. and Poskit, D. S. (1989). Recursive estimation of autoregressions, *J. R. Statist. Soc. B*, **51**, 217–233.
- Härdle, W. (1989). Resampling for inference from curves, *Proceedings of the 47th Session of International Statistical Institute*, 53–63, Paris.
- Härdle, W. (1990). *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.
- Härdle, W. and Bowman, A. (1988). Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands, *J. Amer. Statist. Assoc.*, **83**, 102–110.
- Härdle, W. and Marron, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression, *Ann. Statist.*, **13**, 1465–1481.
- Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? (with discussions), *J. Amer. Statist. Assoc.*, **83**, 86–99.
- Hartigan, J. A. (1969). Using subsample values as typical value, *J. Amer. Statist. Assoc.*, **64**, 1303–1317.
- Hastie, T. J. and Tibshirani, R. J. (1985). Discussion of “Projection pursuit” by P. Huber, *Ann. Statist.*, **13**, 502–508.
- Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Methods*, Chapman and Hall, London.
- He, K. (1987). Bootstrapping linear M-regression models, *Acta Math. Sinica*, **29**, 613–617.
- Helmers, R. (1991). On the Edgeworth expansion and the bootstrap approximation for a studentized U-statistics, *Ann. Statist.*, **19**, 470–484.
- Helmers, R., Janssen, P. and Serfling, R. (1988). Glivenko-Cantelli properties of some generalized empirical DF's and strong convergence of generalized L-statistics, *Prob. Theory and Related Fields*, **79**, 75–93.
- Helmers, R., Janssen, P. and Serfling, R. (1990). Berry-Esséen and bootstrap results for generalized L-statistics, *Scand. J. Statist.*, **17**, 65–78.
- Helmers, R., Janssen, P. and Veraverbeke, N. (1992). Bootstrapping U-quantiles, *Exploring the Limit of Bootstrap*, R. LePage and L. Billard eds., 145–155, Wiley, New York.
- Hemerly, E. M. and Davis, M. H. A. (1989). Strong consistency of the PLS criterion for order determination of autoregressive processes, *Ann. Statist.*, **17**, 941–946.
- Hinkley, D. V. (1977). Jackknifing in unbalanced situations, *Technometrics*, **19**, 285–292.
- Hinkley, D. V. (1987). Bootstrap significance tests, *Proceedings of the 47th Session of International Statistical Institute*, 65–74, Paris.

- Hinkley, D. V. (1988). Bootstrap methods (with discussions), *J. R. Statist. Soc. B*, **50**, 321–337.
- Hinkley, D. V. (1994). Discussion of “Bootstrap: More than a stab in the dark?” by G. A. Young, *Statist. Science*, **9**, 400–403.
- Hinkley, D. V. and Schechtman, E. (1987). Conditional bootstrap methods in mean-shift models, *Biometrika*, **74**, 85–93.
- Hinkley, D. V. and Shi, S. (1989). Importance sampling and the nested bootstrap, *Biometrika*, **76**, 435–446.
- Hinkley, D. V. and Wang, S. (1988). Discussion of “Saddle point methods and statistical inference” by N. Reid, *Statist. Science*, **3**, 232–233.
- Hinkley, D. V. and Wei, B. C. (1984). Improvement of jackknife confidence limit methods, *Biometrika*, **71**, 331–339.
- Hogg, R. V. and Craig, A. T. (1970). *Introduction to Mathematical Statistics*, second edition, Macmillan, New York.
- Holbert, D. and Son, M. S. (1986). Bootstrapping a time series model: Some empirical results, *Comm. Statist. A*, **15**, 3669–3691.
- Horváth, L. and Yandell, B. S. (1987). Convergence rates for the bootstrapped product limit processes, *Ann. Statist.*, **15**, 1155–1173.
- Huang, J. S. (1991). Efficiency computation of the performance of bootstrap and jackknife estimators of the variance of L-statistics, *J. Statist. Compu. Simul.*, **38**, 45–56.
- Huang, J. S., Sen, P. K. and Shao, J. (1995). Bootstrapping a sample quantile when the density has a jump, *Statist. Sinica*, to appear.
- Huber, P. J. (1981). *Robust Statistics*, Wiley, New York.
- Huet, S. and Jolivet, E. (1989). Exactitude au second ordre des intervalles de confiance bootstrap pour les paramètres d’un modèle de régression non linéaire, *C. R. Acad. Sci. Paris Sér. I. Math.*, **308**, 429–432.
- Huet, S., Jolivet, E. and Messean, A. (1990). Some simulation results about confidence intervals and bootstrap methods in nonlinear regression, *Statistics*, **21**, 369–432.
- Huskova, M. and Janssen, P. (1993a). Generalized bootstrap for studentized U-statistics: A rank statistic approach, *Statist. Prob. Letters*, **16**, 225–233.
- Huskova, M. and Janssen, P. (1993b). Consistency of the generalized bootstrap for degenerate U-statistics, *Ann. Statist.*, **21**, 1811–1823.
- Ibragimov, Z. A. and Has’minskii, R. Z. (1981). *Statistical Estimators: Asymptotic Theory*, Springer-Verlag, New York.
- Jaeckel, L. (1972). The infinitesimal jackknife, *Memorandum*, MM 72–1215–11, Bell Lab., Murray Hill, NJ.

- Janas, D. (1993). *Bootstrap Procedures for Time Series*, Shaker, Aachen.
- Janssen, P. (1994). Weighted bootstrapping of U-statistics, *J. Statist. Plan. Inference*, **38**, 31–42.
- Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators, *Ann. Math. Statist.*, **40**, 633–643.
- Jhun, M. (1988). Bootstrapping density estimates, *Comm. Statist. A*, **17**, 61–78.
- Jhun, M. (1990). Bootstrapping k-means clustering, *J. Japanese Soc. Compu. Statist.*, **3**, 1–14.
- John, P. W. M. (1971). *Statistical Design and Analysis of Experiments*, Macmillan, New York.
- Johns, M. V. Jr. (1988). Importance sampling for bootstrap confidence intervals, *J. Amer. Statist. Assoc.*, **83**, 709–714.
- Johnson, R. A. (1970). Asymptotic expansions associated with posterior distributions, *Ann. Math. Statist.*, **41**, 851–864.
- Jones, M. C. (1991). Discussion of “Bootstrap choice of bandwidth for density estimation” by J. J. Faraway and M. Jhun, *J. Amer. Statist. Assoc.*, **86**, 1153–1154.
- Jones, M. C., Marron, J. S. and Park, B. U. (1991). A simple root n bandwidth selector, *Ann. Statist.*, **19**, 1919–1932.
- Jorgensen, M. A. (1987). Jackknifing fixed points of iterations, *Biometrika*, **74**, 207–211.
- Kabala, P. (1993). On bootstrap predictive inference for autoregressive process, *J. Time Ser. Anal.*, **14**, 473–484.
- Kalton, G. (1981). *Compensating for Missing Data*, ISR research report series, Survey Research Center, University of Michigan, Ann Arbor.
- Karrison, T. (1990). Bootstrapping censored data with covariates, *J. Statist. Compu. Simul.*, **36**, 195–207.
- Kendall, M. G. and Staurt, A. (1979). *The Advance Theory of Statistics*, vol. I, fourth edition, MacMillam, New York.
- Kim, Y. B., Haddock, J. and Willemain, T. R. (1993). The binary bootstrap: Inference with autocorrelated binary data, *Comm. Statist. B*, **22**, 205–216.
- Kinateder, J. G. (1992). An invariance principle applicable to the bootstrap, *Exploring the Limit of Bootstrap*, R. LePage and L. Billard eds., 157–181, Wiley, New York.
- Kipnis, V. (1992). Bootstrap assessment of prediction in exploratory regression analysis, *Exploring the Limit of Bootstrap*, R. LePage and L. Billard eds., 363–387, Wiley, New York.

- Kish, L. and Frankel, M. R. (1970). Balanced repeated replication for standard errors, *J. Amer. Statist. Assoc.*, **65**, 1071–1094.
- Kish, L. and Frankel, M. R. (1974). Inference from complex samples (with discussions), *J. R. Statist. Soc. B*, **36**, 1–37.
- Klenk, A. and Stute, W. (1987). Bootstrapping of L-statistics, *Statist. Decision*, **16**, 1696–1708.
- Knight, K. (1989). On the bootstrap of the sample mean in the infinite variance case, *Ann. Statist.*, **17**, 1168–1175.
- Koenker, R. and Bassett, G. (1978). Regression quantiles, *Econometrica*, **46**, 33–50.
- Koenker, R. and Portnoy, S. (1987). L-estimation for linear models, *J. Amer. Statist. Assoc.*, **82**, 851–857.
- Konishi, S. (1991). Normalizing transformations and bootstrap confidence intervals, *Ann. Statist.*, **19**, 2209–2225.
- Konishi, S. and Honda, M. (1990). Comparison of procedures for estimation of error rates in discriminant analysis under nonnormal populations, *J. Statist. Compu. Simul.*, **36**, 105–116.
- Kovar, J. G. (1985). Variance estimation of nonlinear statistics in stratified samples. Methodology Branch Working Paper #85-052E, Statistics Canada.
- Kovar, J. G. (1987). Variance estimation of medians in stratified samples. Methodology Branch Working Paper #87-004E, Statistics Canada.
- Kovar, J. G., Rao, J. N. K. and Wu, C. F. J. (1988). Bootstrap and other methods to measure errors in survey estimates, *Canadian J. Statist.*, **16**, Supplement, 25–45.
- Kreiss, J. P. and Franke, J. (1992). Bootstrapping stationary autoregressive moving average models, *J. Time Ser. Anal.*, **13**, 297–317.
- Krewski, D. (1978). On the stability of some replication variance estimators in the linear case, *J. Statist. Plan. Inference*, **2**, 45–51.
- Krewski, D. and Rao, J. N. K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods, *Ann. Statist.*, **9**, 1010–1019.
- Krewski, D., Smythe, R. T., Dewanji, A. and Szyszkowicz, M. (1991a). Bootstrapping an empirical Bayes estimator of the distribution of historical controls in carcinogen bioassay, preprint.
- Krewski, D., Smythe, R. T., Fung, K. Y. and Burnett, R. (1991b). Conditional and unconditional tests with historical controls, *Canadian J. Statist.*, **19**, 407–423.

- Kuk, A. Y. C. (1987). Bootstrap estimators of variance under sampling with proportional to aggregate size, *J. Statist. Compu. Simul.*, **28**, 303–311.
- Kuk, A. Y. C. (1989). Double bootstrap estimation of variance under systematic sampling with probability proportional to size, *J. Statist. Compu. Simul.*, **31**, 73–82.
- Kulperger, P. J. and Prakasa Rao, B. L. S. P. (1989). Bootstrapping a finite state Markov chain, *Sankhyā A*, **51**, 178–191.
- Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations, *Ann. Statist.*, **17**, 1217–1241.
- Lachenbruch, P. A. (1975). *Discriminant Analysis*, Hafuer, New York.
- Lachenbruch, P. A. and Mickey, M. R. (1968). Estimation of error rates in discriminant analysis, *Technometrics*, **10**, 1–11.
- Lahiri, S. N. (1991). Second order optimality of stationary bootstrap, *Statist. Prob. Letters*, **14**, 335–341.
- Lahiri, S. N. (1992a). On bootstrapping M-estimators, *Sankhyā A*, **54**, 157–170.
- Lahiri, S. N. (1992b). Edgeworth correction by moving block bootstrap for stationary and nonstationary data, *Exploring the Limit of Bootstrap*, R. LePage and L. Billard eds., 183–214, Wiley, New York.
- Lahiri, S. N. (1992c). Bootstrapping M-estimators of a multiple linear regression parameter, *Ann. Statist.*, **20**, 1548–1570.
- Lahiri, S. N. (1993a). Bootstrapping the studentized sample mean of lattice variables, *J. Multivariate Anal.*, **45**, 247–256.
- Lahiri, S. N. (1993b). On the moving block bootstrap under long range dependence, *Statist. Prob. Letters*, **18**, 405–413.
- Lai, T. L. and Wang, J. Q. Z. (1993). Edgeworth expansions for symmetric statistics with applications to bootstrap methods, *Statistica Sinica*, **3**, 517–542.
- Lai, T. L., Robbins, H. and Wei, C. (1979). Strong consistency of least squares estimates in multiple regression, *J. Multivariate Anal.*, **9**, 343–361.
- Laird, N. M. and Louis, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples (with discussions), *J. Amer. Statist. Assoc.*, **82**, 739–757.
- Läuter, H. (1985). An efficient estimate for the error rate in discriminant analysis, *Statistics*, **16**, 107–119.
- Lawley, D. M. and Maxwell, A. E. (1971). *Factor Analysis as A Statistical Method*, American Elsevier, New York.

- LeCam, L. (1956). On the asymptotic theory of estimation and testing hypotheses, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. I, 129–156, University of California Press, Berkeley.
- LeCam, L. (1983). A remark on empirical measures, *A Festschrift for Erich Lehmann*, P. J. Bickel, K. Doksum and J. L. Hodges eds., Wadsworth, Belmont, CA.
- Lee, K. W. (1990). Bootstrapping logistic regression models with random regressors, *Comm. Statist. A*, **19**, 2527–2539.
- Lee, S. (1994). Optimal choice between parametric and nonparametric bootstrap estimates, *Math. Proc. Cambridge Philos. Soc.*, **115**, 335–363.
- Lee, S. and Young, G. A. (1994). Practical higher-order smoothing of the bootstrap, *Statistica Sinica*, **4**, 445–460.
- Léger, C. and Romano, J. P. (1990). Bootstrap choice of tuning parameters, *Ann. Inst. Statist. Math.*, **42**, 709–735.
- Léger, C., Politis, D. N. and Romano, J. P. (1992). Bootstrap technology and applications, *Technometrics*, **34**, 378–398.
- Lehmann, E. L. (1983). *Theory of Point Estimation*, Wiley, New York.
- Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, second edition, Wiley, New York.
- Lele, S. (1991). Jackknifing linear estimating equations: Asymptotic theory and applications in stochastic processes, *J. R. Statist. Soc. B*, **53**, 253–267.
- LePage, R. (1992). Bootstrapping signs, *Exploring the Limit of Bootstrap*, R. LePage and L. Billard eds., 215–224, Wiley, New York.
- Li, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing, *Ann. Statist.*, **14**, 1101–1112.
- Li, K.-C. (1987). Asymptotic optimality for C_p , C_L , cross-validation and generalized cross-validation: Discrete index set, *Ann. Statist.*, **15**, 958–975.
- Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika*, **73**, 13–22.
- Linder, E. and Babu, G. J. (1994). Bootstrapping the linear functional relationship with known error variance ratio, *Scand. J. Statist.*, **21**, 21–39.
- Liu, J. (1992). *Inference from Stratified Samples: Application of Edgeworth Expansions*, Ph.D. thesis, Carleton University.
- Liu, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models, *Ann. Statist.*, **16**, 1697–1708.

- Liu, R. Y. and Singh, K. (1987). On a partial correction by the bootstrap, *Ann. Statist.*, **15**, 1713–1718.
- Liu, R. Y. and Singh, K. (1992a). Efficiency and robustness in resampling, *Ann. Statist.*, **20**, 370–384.
- Liu, R. Y. and Singh, K. (1992b). Moving blocks jackknife and bootstrap capture weak dependence, *Exploring the Limit of Bootstrap*, R. LePage and L. Billard eds., 225–248, Wiley, New York.
- Liu, R. Y., Singh, K. and Lo, S. H. (1989). On a representation related to bootstrap, *Sankhyā A*, **51**, 168–177.
- Liu, Z. and Tu, D. (1987). Kernel methods on conditional median estimation, *Kexue Tongbao (Chinese Bulletin of Science)*, **32**, 642–643.
- Lo, A. Y. (1987). A large sample study of the Bayesian bootstrap, *Ann. Statist.*, **15**, 360–375.
- Lo, A. Y. (1988). A Bayesian bootstrap for a finite population, *Ann. Statist.*, **16**, 1684–1695.
- Lo, A. Y. (1991). Bayesian bootstrap clones and a biometry function, *Sankhyā A*, **53**, 320–333.
- Lo, A. Y. (1993a). A Bayesian bootstrap for censored data, *Ann. Statist.*, **21**, 100–123.
- Lo, A. Y. (1993b). A Bayesian bootstrap for weighted sampling, *Ann. Statist.*, **21**, 2138–2148.
- Lo, S. H. and Singh, K. (1986). The product-limit estimator and the bootstrap: Some asymptotic representations, *Prob. Theory and Related Fields* **71**, 455–465.
- Loh, W. Y. (1984). Estimating an endpoint of a distribution with resampling methods, *Ann. Statist.*, **12**, 1543–1550.
- Loh, W. Y. (1985). A new method for testing separate families of hypotheses, *J. Amer. Statist. Assoc.*, **80**, 362–368.
- Loh, W. Y. (1987). Calibrating confidence coefficients, *J. Amer. Statist. Assoc.*, **82**, 155–162.
- Loh, W. Y. (1988). Discussion of “Theoretical comparison of bootstrap confidence intervals” by P. Hall, *Ann. Statist.*, **16**, 972–976.
- Loh, W. Y. (1991). Bootstrap calibration for confidence interval construction and selection, *Statist. Sinica*, **1**, 479–495.
- Loh, W. Y. and Wu, C. F. J. (1987). Discussion of “Better bootstrap confidence intervals” by B. Efron, *J. Amer. Statist. Assoc.*, **82**, 188–190.
- Lohse, K. (1987). Consistency of the bootstrap, *Statist. Decision*, **5**, 353–366.

- Loughin, T. M. and Koehler, K. (1993). Bootstrapping in proportional hazards models with fixed explanatory variables, preprint.
- Mallows, C. L. (1972). A note on asymptotic joint normality, *Ann. Math. Statist.*, **39**, 755–771.
- Mallows, C. L. (1973). Some comments on C_p , *Technometrics*, **15**, 661–675.
- Mammen, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap, *Ann. Statist.*, **17**, 382–400.
- Mammen, E. (1992). *When does Bootstrap Work? Asymptotic Results and Simulation*, Springer-Verlag, Heidelberg.
- Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models, *Ann. Statist.*, **21**, 255–285.
- Mantel, H. J. and Singh, A. C. (1991). Standard errors of estimates of low income proportions: A proposed methodology, preprint.
- Maritz, J. S. and Jarrett, R. G. (1978). A note on estimating the variance of the sample median, *J. Amer. Statist. Assoc.*, **73**, 194–196.
- Martin, M. A. (1990). On bootstrap iteration for converge correction in confidence intervals, *J. Amer. Statist. Assoc.*, **85**, 1105–1108.
- Mason, D. M. and Newton, M. A. (1992). A rank statistics approach to the consistency of a general bootstrap, *Ann. Statist.*, **20**, 1611–1624.
- McCarthy, P. J. (1969). Pseudo-replication: Half samples, *Rev. Internat. Statist. Inst.*, **37**, 239–264.
- McCarthy, P. J. and Snowden, C. B. (1985). The bootstrap and finite population sampling, *Vital and Health Statistics*, 2–95, Public Health Service Publication 85–1369, U.S. Government Printing Office, Washington, D.C.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, second edition, Chapman and Hall, London.
- McLachlam, G. J. (1980). The efficiency of Efron's "bootstrap" approach applied to error rate estimation in discriminate analysis, *J. Statist. Compu. Simul.*, **11**, 273–279.
- Miller, R. G. (1964). A trustworthy jackknife, *Ann. Math. Statist.*, **35**, 1594–1605.
- Miller, R. G. (1974). An unbalanced jackknife, *Ann. Statist.*, **2**, 880–891.
- Morrison, D. F. (1990). *Multivariate Statistical Methods*, third edition, McGraw-Hill, New York.
- Moulton, L. H. and Zeger, S. L. (1989). Analyzing repeated measures on generalized linear models via the bootstrap, *Biometrics*, **45**, 381–394.

- Moulton, L. H. and Zeger, S. L. (1991). Bootstrapping generalized linear models, *Compu. Statist. Data Anal.*, **11**, 53–63.
- Mykland, A. (1992). Asymptotic expansions and bootstrapping distribution for dependent variables: A martingale approach, *Ann. Statist.*, **20**, 623–654.
- Nagaev, S. V. (1961). More exact statements of limit theorems for homogeneous Markov chains, *Theory Prob. Appl.*, **6**, 62–80.
- Nagao, H. (1985). On the limiting distribution of the jackknife statistics for eigenvalues of a sample covariance matrix, *Comm. Statist. A*, **14**, 1547–1567.
- Nagao, H. (1988). On the jackknife statistics for eigenvalues and eigenvectors of a correlation matrix, *Ann. Inst. Statist. Math.*, **40**, 477–489.
- Nagao, H. and Srivastava, M. S. (1992). On the distributions of some test criteria for a covariance matrix under local alternatives and bootstrap approximations, *J. Multivariate Anal.*, **43**, 331–350.
- Navidi, W. (1989). Edgeworth expansions for bootstrapping regression models, *Ann. Statist.*, **17**, 1472–1478.
- Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *J. R. Statist. Soc. A*, **135**, 370–384.
- Newton, M. A. and Geyer, G. J. (1994). Bootstrap recycling: A Monte Carlo alternative to the nested bootstrap, *J. Amer. Statist. Assoc.*, **89**, 905–912.
- Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussions), *J. R. Statist. Soc. B*, **56**, 3–48.
- Noreen, E. W. (1989). *Computer Intensive Methods for Testing Hypotheses: An Introduction*, Wiley, New York.
- Owen, A. B. (1988). Empirical likelihood ratio confidence interval for a single functional, *Biometrika*, **75**, 237–249.
- Padgett, W. and Thombs, L. A. (1986). Smooth nonparametric quantile estimation under censoring: Simulation and bootstrap methods, *Comm. Statist. B*, **15**, 1003–1025.
- Park, B. and Marron, J. S. (1990). Comparison of data-driven bandwidth selectors, *J. Amer. Statist. Assoc.*, **85**, 66–72.
- Parr, W. C. (1983). A note on the jackknife, the bootstrap and the delta method estimates of bias and variance, *Biometrika*, **70**, 719–722.
- Parr, W. C. (1985). Jackknifing differentiable statistical functionals, *J. R. Statist. Soc. B*, **47**, 56–66.
- Parr, W. C. and Schucany, W. R. (1982). Jackknifing L-statistics with smooth weight functions, *J. Amer. Statist. Assoc.*, **77**, 629–638.

- Peck, R., Fisher, L. and van Ness, J. (1989). Bootstrap confidence intervals for the numbers of clusters in cluster analysis, *J. Amer. Statist. Assoc.*, **84**, 184–191.
- Peters, S. C. and Freedman, D. A. (1987). Balm for bootstrap confidence intervals, *J. Amer. Statist. Assoc.*, **82**, 186–187.
- Petrov, V. V. (1975). *Sums of Independent Random Variables*, Springer-Verlag, New York.
- Pfanzagel, J. (1985). *Asymptotic Expansions for General Statistical Models*, Springer-Verlag, Berlin.
- Politis, D. N. and Romano, J. P. (1990). A nonparametric resampling procedure for multivariate confidence regions in time series analysis, *Proceedings of INTERFACE'90, 22nd Symposium on the Interface of Computing Science and Statistics*, C. Page and R. LePage eds., Springer-Verlag, New York.
- Politis, D. N. and Romano, J. P. (1992a). A circular block-resampling procedure for stationary data, *Exploring the Limit of Bootstrap*, R. LePage and L. Billard eds., 263–270, Wiley, New York.
- Politis, D. N. and Romano, J. P. (1992b). A general resampling scheme for triangular arrays of α -mixing random variables with application to the problem of spectral density estimation, *Ann. Statist.*, **20**, 1985–2007.
- Politis, D. N. and Romano, J. P. (1993a). Estimating the distribution of a studentized statistic by subsampling, *Bull. Intern. Statist. Inst.*, **49th Session**, **2**, 315–316.
- Politis, D. N. and Romano, J. P. (1993b). Nonparametric resampling for homogeneous strong mixing random fields, *J. Multivariate Anal.*, **47**, 301–328.
- Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap, *J. Amer. Statist. Assoc.*, **89**, 1303–1313.
- Politis, D. N. and Romano, J. P. (1995). A general theory for large sample confidence regions based on subsamples under minimal assumptions, *Ann. Statist.*, **23**, in press.
- Politis, D. N., Romano, J. P. and Lai, T. L. (1992). Bootstrap confidence bands for spectra and cross-spectra, *IEEE Trans. Signal Process.*, **40**, 1206–1215.
- Prakasa Rao, B. L. S. P. (1983). *Nonparametric Functional Estimation*, Academic, New York.
- Quan, H. and Tsai, W.-Y. (1992). Jackknife for the proportional hazards model, *J. Statist. Compu. Simul.*, **43**, 163–176.

- Quenneville, B. (1986). Bootstrap procedures for testing linear hypothesis without normality, *Statistics*, **17**, 533–538.
- Quenouille, M. (1949). Approximation tests of correlation in time series, *J. R. Statist. Soc. B*, **11**, 18–84.
- Ramos, E. (1988). Resampling methods for time series, Tech. Report ONR-C-2, Dept. of Statistics, Harvard University.
- Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, second edition, Wiley, New York.
- Rao, C. R. (1989). *Statistics and Truth. Putting Chance to Work*, International Co-operative Publishing House, Burtonsville, Md.
- Rao, C. R. and Tu, D. (1991). Inference on the occurrence/exposure rate with mixed censoring models, *Calcutta Statist. Assoc. Bull.*, **40**, 65–87.
- Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem, *Biometrika*, **76**, 369–374.
- Rao, C. R. and Zhao, L. (1992). Approximation to the distribution of M-estimates in linear models by randomly weighted bootstrap, *Sankhyā A*, **54**, 323–331.
- Rao, J. N. K. (1979). On deriving mean square errors and their nonnegative unbiased estimators in finite population sampling, *J. Indian Statist. Assoc.*, **17**, 125–136.
- Rao, J. N. K. (1988). Variance estimation in sample surveys, *Handbook of Statistics*, P. K. Krishnaiah and C. R. Rao eds., vol. **6**, 427–447, Elsevier, North-Holland, Amsterdam.
- Rao, J. N. K. and Bellhouse, D. R. (1990). History and development of the theoretical foundations of survey based estimation and analysis, *Survey Methodology*, **16**, 3–29.
- Rao, J. N. K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation, *Biometrika*, **79**, 811–822.
- Rao, J. N. K. and Shao, J. (1995). On balanced half-sample variance estimation in stratified sampling, *J. Amer. Statist. Assoc.*, to appear.
- Rao, J. N. K. and Sitter, R. R. (1994). Variance estimation under two-phase sampling with application to imputation for missing data, preprint.
- Rao, J. N. K. and Wu, C. F. J. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics, *J. Amer. Statist. Assoc.*, **80**, 620–630.
- Rao, J. N. K. and Wu, C. F. J. (1987). Methods for standard errors and confidence intervals from sample survey data: Some recent work, *Bull. Intern. Statist. Inst., Proceedings of the 46th Session*, **3**, 5–19.

- Rao, J. N. K. and Wu, C. F. J. (1988). Resampling inference with complex survey data, *J. Amer. Statist. Assoc.*, **83**, 231–241.
- Rao, J. N. K., Hartley, H. O. and Cochran, W. G. (1962). On a simple procedure of unequal probability sampling without replacement, *J. R. Statist. Soc. B*, **24**, 482–491.
- Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys, *Survey Methodology*, **18**, 209–217.
- Rayner, R. E. (1990). Bootstrap tests for generalized least squares regression models, *Econom. Letters*, **34**, 261–265.
- Reeds, J. A. (1978). Jackknifing maximum likelihood estimates, *Ann. Statist.*, **6**, 727–739.
- Reid, N. (1988). Saddle point methods and statistical inference (with discussions), *Statist. Science*, **3**, 213–238.
- Rissanen, J. (1986). Stochastic complexity and modeling, *Ann. Statist.* **14**, 1080–1100.
- Robinson, J. (1978). An asymptotic expansion for samples from a finite population, *Ann. Statist.*, **5**, 1005–1011.
- Roche, D. M. (1993). Almost-exact parametric bootstrap calculation via the saddle point approximation, *Compu. Statist. Data Anal.*, **15**, 451–460.
- Romanazzi, M. (1993). Jackknife estimation of the eigenvalues of the covariance matrix, *Compu. Statist. Data Anal.*, **15**, 179–198.
- Romano, J. P. (1988a). A bootstrap revival of some nonparametric distance tests, *J. Amer. Statist. Assoc.*, **83**, 698–708.
- Romano, J. P. (1988b). Bootstrapping the mode, *Ann. Inst. Statist. Math.*, **40**, 565–586.
- Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses, *Ann. Statist.*, **17**, 141–159.
- Rubin, D. B. (1978). Multiple imputations in sample surveys—a phenomenological Bayesian approach to nonresponse, *Proceedings of the Survey Research Methods Section*, 20–34, American Statistical Association, Alexandria, VA.
- Rubin, D. B. (1981). The Bayesian bootstrap, *Ann. Statist.*, **9**, 130–134.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.
- Rubin, D. B. (1988). Using the SIR algorithm to simulate posterior distributions, *Bayesian Statistics* **3**, J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A F. M. Smith eds., 395–402, Oxford Univ. Press, Oxford.

- Rubin, D. B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse, *J. Amer. Statist. Assoc.*, **81**, 366–374.
- Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators, *Scand. J. Statist.*, **9**, 65–78.
- Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model, *J. Amer. Statist. Assoc.*, **75**, 828–838.
- Rutherford, B. and Yakowitz, S. (1991). Error inference for nonparametric regression, *Ann. Inst. Statist. Math.*, **43**, 115–129.
- Sain, S. R., Baggerly, K. A. and Scott, D. W. (1994). Cross-validation of multivariate densities, *J. Amer. Statist. Assoc.*, **89**, 807–817.
- Sauermann, W. (1989). Bootstrapping the maximum likelihood estimator in high dimensional log linear models, *Ann. Statist.*, **17**, 1198–1216.
- Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance, *Biometrika*, **40**, 87–104.
- Schenker, N. (1985). Qualms about bootstrap confidence intervals, *J. Amer. Statist. Assoc.*, **80**, 360–361.
- Schenker, N. and Welsh, A. H. (1988). Asymptotic results for multiple imputation, *Ann. Statist.*, **16**, 1550–1566.
- Schmidt, W. H. (1979). Normal theory approximations to tests for linear hypotheses, *Math. Oper. Statist.*, **10**, 353–365.
- Schneller, W. (1989). Edgeworth expansions for linear rank statistics, *Ann. Statist.*, **17**, 1103–1123.
- Schucany, W. R. and Sheather, S. J. (1989). Jackknifing R-estimators, *Biometrika*, **76**, 393–398.
- Schucany, W. R. and Wang, S. (1991). One-step bootstrapping for smooth iterative procedures, *J. R. Statist. Soc. B*, **53**, 587–596.
- Schuster, E. F. (1987). Identifying the closest symmetric distribution or density function, *Ann. Statist.*, **15**, 865–874.
- Schuster, E. F. and Barker, R. C. (1987). Using the bootstrap in testing symmetry versus asymmetry, *Comm. Statist. B*, **16**, 69–84.
- Schwartz, G. (1978). Estimating the dimensions of a model, *Ann. Statist.*, **6**, 461–464.
- Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation, *J. Amer. Statist. Assoc.*, **82**, 1131–1146.
- Searle, S. R. (1971). *Linear Models*, Wiley, New York.
- Sedransk, J. (1985). The objective and practice of imputation, *Proceedings of the First Annual Research Conference*, 445–452, Bureau of the Census, Washington, D.C.

- Sen, P. K. (1977). Some invariance principles relating to jackknifing and their role in sequential analysis, *Ann. Statist.*, **5**, 316–329.
- Sen, P. K. (1988). Functional jackknifing: Rationality and general asymptotics, *Ann. Statist.*, **16**, 450–469.
- Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*, Wiley, New York.
- Serfling, R. J. (1984). Generalized L-, M- and R-statistics, *Ann. Statist.*, **12**, 76–86.
- Shao, J. (1987). Sampling and resampling: An efficient approximation to jackknife variance estimators in linear models, *Chinese J. Appl. Prob. Statist.*, **3**, 368–379.
- Shao, J. (1988a). On resampling methods for variance and bias estimation in linear models, *Ann. Statist.*, **16**, 986–1008.
- Shao, J. (1988b). Consistency of jackknife estimators of the variance of sample quantiles, *Comm. Statist. A*, **17**, 3017–3028.
- Shao, J. (1988c). Bootstrap variance and bias estimation in linear models, *Canadian J. Statist.*, **16**, 371–382.
- Shao, J. (1989a). Bootstrapping for generalized L-statistics, *Comm. Statist. A*, **18**, 2005–2016.
- Shao, J. (1989b). The efficiency and consistency of approximations to the jackknife variance estimators, *J. Amer. Statist. Assoc.*, **84**, 114–119.
- Shao, J. (1989c). Jackknifing weighted least squares estimators, *J. R. Statist. Soc. B*, **51**, 139–156.
- Shao, J. (1990a). Influence function and variance estimation, *Chinese J. Appl. Prob. Statist.*, **6**, 309–315.
- Shao, J. (1990b). Bootstrap estimation of the asymptotic variance of statistical functionals, *Ann. Inst. Statist. Math.*, **42**, 737–752.
- Shao, J. (1991a). Second-order differentiability and jackknife, *Statist. Sinica*, **1**, 185–202.
- Shao, J. (1991b). Consistency of jackknife variance estimators, *Statistics*, **22**, 49–57.
- Shao, J. (1992a). Some results for differentiable statistical functionals, *Non-parametric Statistics and Related Topics*, A. K. Md. E. Saleh ed., 179–188, North-Holland, Amsterdam.
- Shao, J. (1992b). Bootstrap variance estimators with truncation, *Statist. Prob. Letter*, **15**, 95–101.
- Shao, J. (1992c). One-step jackknife for M-estimators computed using Newton's method, *Ann. Inst. Statist. Math.*, **44**, 687–701.

- Shao, J. (1992d). Asymptotic theory in generalized linear models with nuisance scale parameter, *Prob. Theory and Related Fields*, **91**, 25–41.
- Shao, J. (1992e). Jackknife variance estimator for m-dependent stationary process, *Acta Math. Appl. Sinica*, **8**, 115–123.
- Shao, J. (1992f). Consistency of least squares estimator and its jackknife variance estimator in nonlinear models, *Canadian J. Statist.*, **20**, 415–428.
- Shao, J. (1992g). Jackknifing in generalized linear models, *Ann. Inst. Statist. Math.*, **44**, 673–686.
- Shao, J. (1993a). Differentiability of statistical functionals and consistency of the jackknife, *Ann. Statist.*, **21**, 61–75.
- Shao, J. (1993b). Linear model selection by cross-validation, *J. Amer. Statist. Assoc.*, **88**, 486–494.
- Shao, J. (1994a). Bootstrap sample size in nonregular cases, *Proceedings of the Amer. Math. Soc.*, **122**, 1251–1262.
- Shao, J. (1994b). L-statistics in complex surveys, *Ann. Statist.*, **22**, 946–967.
- Shao, J. (1995a). An asymptotic theory for linear model selection, *Statist. Sinica*, to appear.
- Shao, J. (1995b). Bootstrap variable selection in regression, *J. Amer. Statist. Assoc.*, to appear.
- Shao, J. and Rao, J. N. K. (1993). Jackknife inference for heteroscedastic linear regression models, *Canadian J. Statist.*, **21**, 377–395.
- Shao, J. and Rao, J. N. K. (1994). Standard errors for low income proportions estimated from stratified multistage samples, *Sankhyā B*, Special Volume **55**, 393–414.
- Shao, J. and Wu, C. F. J. (1987). Heteroscedasticity-robustness of jackknife variance estimators in linear models, *Ann. Statist.*, **15**, 1563–1579.
- Shao, J. and Wu, C. F. J. (1989). A general theory for jackknife variance estimation, *Ann. Statist.*, **17**, 1176–1197.
- Shao, J. and Wu, C. F. J. (1992). Asymptotic properties of the balanced repeated replication method for sample quantiles, *Ann. Statist.*, **20**, 1571–1593.
- Shao, Q. and Yu, H. (1993). Bootstrapping the sample means for stationary mixing sequences, *Stochastic Processes and Their Appl.*, **48**, 175–190.
- Shi, X. (1984). The approximate independence of jackknife pseudo-values and the bootstrap methods, *J. Wuhan Inst. Hydra. Elect. Engin.*, **2**, 83–90.

- Shi, X. (1986a). A note on bootstrapping U-statistics, *Chinese J. Appl. Prob. Statist.*, **2**, 144–148.
- Shi, X. (1986b). Bootstrap estimate for m -dependent sample means, *Kexue Tongbao (Chinese Bulletin of Science)*, **31**, 404–407.
- Shi, X. (1987). Some asymptotic properties of bootstrapping U-statistics, *J. Sys. Sci. and Math. Sci.*, **7**, 23–26.
- Shi, X. (1991). Some asymptotic results for jackknifing the sample quantile, *Ann. Statist.*, **19**, 496–503.
- Shi, X. and Liu, K. (1992). Resampling method under dependent models, *Chinese Ann. Math. B*, **13**, 25–34.
- Shi, X. and Shao, J. (1988). Resampling estimation when the observations are m -dependent, *Comm. Statist. A*, **17**, 3923–3934.
- Shi, X., Chen, J. and Wu, C. F. J. (1990). Weak and strong representations for quantile processes from finite populations with applications to simulation size in resampling methods, *Canadian J. Statist.*, **18**, 141–148.
- Shorack, G. P. (1982). Bootstrapping robust regression, *Comm. Statist. A*, **11**, 961–972.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.
- Silverman, B. W. and Young, G. A. (1987). The bootstraps: To smooth or not to smooth, *Biometrika*, **74**, 469–479.
- Simonoff, J. S. and Tsai, C. (1988). Jackknifing and bootstrapping quasi-likelihood estimators, *J. Statist. Compu. Simul.*, **30**, 213–232.
- Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap, *Ann. Statist.*, **9**, 1187–1195.
- Singh, K. and Babu, G. J. (1990). On the asymptotic optimality of the bootstrap, *Scand. J. Statist.*, **17**, 1–9.
- Singh, K. and Liu, R. Y. (1990). On the validity of the jackknife procedure, *Scand. J. Statist.*, **17**, 11–21.
- Sitter, R. R. (1992a). A resampling procedure for complex survey data, *J. Amer. Statist. Assoc.*, **87**, 755–765.
- Sitter, R. R. (1992b). Comparing three bootstrap methods for survey data, *Canadian J. Statist.*, **20**, 135–154.
- Sitter, R. R. (1993). Balanced repeated replications based on orthogonal multi-arrays, *Biometrika*, **80**, 211–221.
- Smith, A. F. M. and Gelfand, A. E. (1992). Bayesian statistics without tears: A sampling-resampling perspective, *J. Amer. Statist. Assoc.*, **46**, 84–88.

- Snapinn, S. M. and Knoke, J. D. (1988). Bootstrap and smoothed classification error rate estimates, *Comm. Statist. B*, **17**, 1135–1153.
- Srivastava, M. S. (1987). Bootstrapping Durbin-Waston statistic, *Indian J. Math.*, **29**, 193–210.
- Srivastava, M. S. and Chan Y. M. (1989). A comparison of bootstrap method and Edgeworth expansion in approximating the distribution of sample variance—one sample and two sample cases, *Comm. Statist. B*, **18**, 339–361.
- Stangenhaus, G. (1987). Bootstrap and inference procedures for L_1 -regression, *Statistical Data Analysis Based on L_1 -Norm and Related Methods*, Y. Dodge ed., 323–332, North-Holland, Amsterdam.
- Stewart, T. J. (1986). Experience with a Bayesian bootstrap method incorporating proper prior information, *Comm. Statist. A*, **15**, 3205–3225.
- Stein, C. (1956). Efficient nonparametric testing and estimation, *Proceedings of the Third Berkeley Symposium*, 187–196, Univ. of California Press, Berkeley.
- Stine, R. A. (1985). Bootstrap prediction interval for regression, *J. Amer. Statist. Assoc.*, **80**, 1026–1031.
- Stine, R. A. (1987). Estimating properties of autoregressive forecasts, *J. Amer. Statist. Assoc.*, **82**, 1072–1078.
- Stoffer, D. S. and Wall, K. D. (1991). Bootstrapping state-space models: Gaussian maximum likelihood estimation and Kalman filter, *J. Amer. Statist. Assoc.*, **86**, 1024–1033.
- Stone, C. J. (1977). Consistent nonparametric regression (with discussions), *Ann. Statist.*, **5**, 595–645.
- Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates, *Ann. Statist.*, **12**, 1285–1297.
- Stone, M. (1974). Cross-validation choice and assessment of statistical predictions, *J. R. Statist. Soci. B* **36**, 111–147.
- Stout, W. F. (1974). *Almost Sure Convergence*, Academic Press, New York.
- Stute, W. (1984). Asymptotic normality of nearest neighbor regression function estimates, *Ann. Statist.*, **12**, 917–926.
- Stute, W. (1990). Bootstrap of the linear correlation model, *Statistics*, **21**, 433–436.
- Stute, W. (1992). Modified cross-validation in density estimation, *J. Statist. Plan. Inference*, **30**, 293–305.
- Stute, W. and Wang, J. (1994). The jackknife estimate of a Kaplan-Meier integral, *Biometrika*, **81**, 602–606.

- Stute, W., Manteiga, W. C. and Quindimil, M. P. (1993). Bootstrap based goodness-of-fit tests, *Metrika*, **40**, 243–256.
- Sutton, C. D. (1993). Computer-intensive methods for tests about the mean of an asymmetric distribution, *J. Amer. Statist. Assoc.*, **88**, 802–810.
- Swanepoel, J. W. H. (1986). A note on proving that the (modified) bootstrap works, *Comm. Statist. A*, **15**, 3193–3203.
- Swanepoel, J. W. H. and van Wyk, J. W. J. (1986). The bootstrap applied to spectral density function estimation, *Biometrika*, **73**, 135–142.
- Swanepoel, J. W. H., van Wyk, J. W. J. and Venter, J. H. (1983). Fixed width confidence intervals based on bootstrap procedures, *Sequential Anal.*, **2**, 289–310.
- Tanner, M. A. (1991). *Tools for Statistical Inference*, Springer-Verlag, New York.
- Tanner, M. A. and Wong, W. (1987). The calculation of posterior densities by data augmentation (with discussions), *J. Amer. Statist. Assoc.*, **82**, 528–550.
- Taylor, C. C. (1989). Bootstrap choice of the smoothing parameter in kernel density estimation, *Biometrika*, **76**, 705–712.
- Thombs, L. A. and Schucany, W. R. (1990). Bootstrap prediction intervals for autoregressions, *J. Amer. Statist. Assoc.*, **85**, 486–492.
- Thorburn, D. E. (1977). On the asymptotic normality of the jackknife, *Scand. J. Statist.*, **4**, 113–118.
- Tibshirani, R. J. (1988). Variance stabilization and the bootstrap, *Biometrika*, **75**, 433–444.
- Tikhomirov, A. N. (1980). On the convergence rate in the central limit theorem for weakly dependent random variables, *Theory Prob. Appl.*, **25**, 790–809.
- Tong, Y. L. (1990). *The Multivariate Normal Distribution*, Springer-Verlag, New York.
- Tsiatis, A. A. (1981). A large sample study of Cox's regression model, *Ann. Statist.*, **9**, 93–108.
- Tu, D. (1986a). Bootstrapping of L-statistics, *Kexue Tongbao (Chinese Bulletin of Science)*, **31**, 965–969.
- Tu, D. (1986b). On the asymptotic expansions relating to the random weighting statistics of minimum contrast estimator, Tech. Report, Inst. Systems Sci., Academia Sinica, Beijing.
- Tu, D. (1988a). The kernel estimator of conditional L-functional and its bootstrapping statistics, *Acta Math. Appl. Sinica*, **11**, 53–68.

- Tu, D. (1988b). The nearest neighbor estimate of the conditional L-functional and its bootstrapping statistics, *Chinese Ann. Math. A*, **8**, 345–357.
- Tu, D. (1988c). On the non-uniform bounds of asymptotic expansion for linear combination of uniform order statistics, *Acta Math. Sinica*, **31**, 729–735.
- Tu, D. (1988d). Randomly weighting the functional statistics, *J. Math. Res. Exp.*, **8**, 439–446.
- Tu, D. (1989). L-functional and nonparametric L-regression estimates: Asymptotic distributions and bootstrapping approximations, Tech. Report 89–51, Center for Multivariate Analysis, Penn State Univ.
- Tu, D. (1992a). Approximating the distribution of a general standardized functional statistic with that of jackknife pseudovalues, *Exploring the limits of Bootstrap*, R. LePage and L. Billard eds., 279–306, Wiley, New York.
- Tu, D. (1992b). Weighted stochastic approximation of estimators of error and variance in linear models, *Acta Math. Sci.*, **12**, 226–233.
- Tu, D. (1994). Discussion of “Approximate Bayesian inference with the weighted likelihood bootstrap” by M. A. Newton and A. E. Raftery, *J. R. Statist. Soc. B*, **56**, 43.
- Tu, D. and Cheng, P. (1989). Bootstrapping untrimmed L-statistics, *J. Sys. Sci. and Math. Sci.*, **9**, 14–23.
- Tu, D. and Gross, A. J. (1994). Bias reduction for jackknife skewness estimators, *Comm. Statist. A*, **23**, 2323–2341.
- Tu, D. and Gross, A. J. (1995). Accurate confidence intervals for the ratio of specific occurrence/exposure rates in risk and survival analysis, *Biometrical J.*, **37**, 611–626.
- Tu, D. and Shi, X. (1988). Bootstrapping and randomly weighting the U-statistics with jackknifed pseudo values, *Math. Statist. Appl. Prob.*, **3**, 205–212.
- Tu, D. and Zhang, L. (1992a). On the estimation of skewness of a statistic using the jackknife and the bootstrap, *Statistical Papers*, **33**, 39–56.
- Tu, D. and Zhang, L. (1992b). Jackknife approximations for some nonparametric confidence intervals of functional parameters based on normalizing transformations, *Computational Statist.*, **7**, 3–15.
- Tu, D. and Zheng, Z. (1987). On the Edgeworth’s expansion of random weighting method, *Chinese J. Appl. Prob. Statist.*, **3**, 340–347.
- Tu, D. and Zheng, Z. (1991). Random weighting: Another approach to approximate the unknown distributions of pivotal quantities, *J. Comb. Info. Systems Sciences*, **16**, 249–270.

- Tukey, J. (1958). Bias and confidence in not quite large samples, *Ann. Math. Statist.*, **29**, 614.
- Valliant, R. (1987). Some prediction properties of balanced half-sample variance estimators in single-stage sampling, *J. R. Statist. Soc. B*, **49**, 68–81.
- van Zwet, W. R. (1979). The Edgeworth expansion for linear combinations of uniform order statistics, *Proceedings of the 2nd Prague Symp. Asymptotic Statistics*, P. Mandl and M. Huskova eds., 93–101, North-Holland, Amsterdam.
- Veall, M. R. (1989). Applications of computationally-intensive methods to econometrics, *Proceedings of the 47th Session of International Statistical Institute*, 75–88, Paris.
- Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem, *Ann. Statist.*, **13**, 1378–1402.
- Wahba, G. (1990). *Spline Methods for Observational Data*, SIAM, Philadelphia.
- Wang, M. C. (1986). Resampling procedures for reducing bias of error rate estimation in multinomial classification, *Compu. Statist. Data Anal.*, **4**, 15–39.
- Wang, S. (1989). On the bootstrap and smoothed bootstrap, *Comm. Statist. A*, **18**, 3949–3962.
- Wang, S. (1992). General saddle point approximations in the bootstrap, *Statist. Prob. Letters*, **13**, 61–66.
- Wax, M. (1988). Order selection for AR models by predictive least squares, *IEEE Trans. Acoust. Speech Signal Process*, **36**, 581–588.
- Wei, C. Z. (1992). On predictive least squares principles, *Ann. Statist.*, **20**, 1–42.
- Weng, C. S. (1989). On a second order property of the Bayesian bootstrap, *Ann. Statist.*, **17**, 705–710.
- Wilson, E. B. and Hilfery, M. M. (1931). The distribution of chi-square, *Proc. Natl. Acad. Sci.*, **17**, 684–688.
- Wolter, K. M. (1985). *Introduction to Variance Estimation*, Springer-Verlag, New York.
- Woodruff, R. S. (1952). Confidence intervals for medians and other position measures, *J. Amer. Statist. Assoc.*, **47**, 635–646.
- Wu, C. F. J. (1981). Asymptotic theory of nonlinear least squares estimation, *Ann. Statist.*, **9**, 501–513.
- Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussions), *Ann. Statist.*, **14**, 1261–1350.

- Wu, C. F. J. (1990). On the asymptotic properties of the jackknife histogram, *Ann. Statist.*, **18**, 1438–1452.
- Wu, C. F. J. (1991). Balanced repeated replications based on mixed orthogonal arrays, *Biometrika*, **78**, 181–188.
- Yang, S. S. (1985). On bootstrapping a class of differentiable statistical functionals with applications to L- and M-estimates, *Statist. Neerlandica*, **39**, 375–385.
- Yang, S. S. (1988). A central limit theorem for the bootstrap mean, *Amer. Statist.*, **42**, 202–203.
- Young, G. A. (1988). Resampling tests of statistical hypotheses, *Proceedings of the Eighth Biannual Symposium on Computational Statistics*, D. Edwards and N. E. Raun eds., 233–238, Physica-Verlag, Heidelberg.
- Young, G. A. (1994). Bootstrap: More than a stab in the dark? (with discussions), *Statist. Science*, **9**, 382–415.
- Yu, K. (1988). The random weighting approximation of sample variance estimates with applications to sampling survey, *Chinese J. Appl. Prob. Statist.*, **3**, 340–347.
- Yu, Z. and Tu, D. (1987). On the convergence rate of bootstrapped and randomly weighted m -dependent means, Research Report, Institute of Systems Science, Academia Sinica, Beijing.
- Zhang, J. and Boos, D. D. (1992). Bootstrap critical values for testing homogeneity of covariance matrices, *J. Amer. Statist. Assoc.*, **87**, 425–429.
- Zhang, J. and Boos, D. D. (1993). Testing hypotheses about covariance matrices using bootstrap methods, *Comm. Statist. A*, **22**, 723–739.
- Zhang, J., Pantula, S. G. and Boos, D. D. (1991). Robust methods for testing the pattern of a single covariance matrix, *Biometrika*, **78**, 787–795.
- Zhang, L. and Tu, D. (1990). A comparison of some jackknife and bootstrap procedures in estimating sampling distributions of studentized statistics and constructing confidence intervals, Tech. Report 90–28, Center for Multivariate Analysis, Pennsylvania State Univ.
- Zhang, P. (1993). Model selection via multifold cross validation, *Ann. Statist.*, **21**, 299–313.
- Zheng, Z. (1985). Asymptotic behavior of the nearest neighbor estimate and its bootstrap statistics, *Scientia Sinica A*, XXVIII, 439–494.
- Zheng, Z. (1987a). Random weighting methods, *Acta Math. Appl. Sinica*, **10**, 247–253.

- Zheng, Z. (1987b). The expansion of random weighting distribution for sample mean—non i.i.d. situation, *Chinese J. Appl. Prob. Statist.*, **3**, 159–166.
- Zheng, Z. (1987c). Some results on random weighting methods, *Chinese J. Appl. Prob. Statist.*, **3**, 1–7.
- Zheng, Z. (1988a). The random weighting method in linear models with independent error distributions, *Chinese Ann. Math. A*, **9**, 724–735.
- Zheng, Z. (1988b). Applications of random weighting methods in regression models, *J. Sys. Sci. and Math. Sci.*, **8**, 151–166.
- Zheng, Z. (1988c). The expansion for the error distribution of sample median and its random weighting statistics, *Acta Math. Appl. Sinica* (English Series), **4**, 340–344.
- Zheng, Z. (1988d). M-estimation and random weighting methods, *Acta Sci. Nat. Uni. Pekinensis*, **24**, 277–285.
- Zheng, Z. (1989). Random weighting method for T -statistic, *Acta Math. Sinica* (New Series), **5**, 87–94.
- Zheng, Z. and Tu, D. (1988). Random weighting method in regression models, *Scientia Sinica A*, **31**, 1442–1459.

Author Index

- Abramovitch, L., 94
Adkins, L.C., 298
Akaike, H., 307
Akritas, M.G., 384
Alemayeha, D., 376–378
Allen, D.M., 307
Arcones, M.A., 81, 186
Arvesen, J.N., 27, 68
Athreya, K.B., 80, 394
Babu, G.J., 80, 82, 89, 94, 95, 99,
 101, 151, 210, 330
Baggerly, K.A., 351
Bahadur, R.R., 178
Bai, C., 306
Bai, Z., 430, 438, 447
Banks, D.L., 419
Barker, R.C., 185, 186
Barlow, W.E., 350
Basawa, I.V., 393, 397
Bassett, G., 284
Bellhouse, D.R., 233
Beran, R., 44, 85, 97, 98, 143, 152,
 155, 156, 161, 180–183, 188,
 203, 299, 306, 373, 376, 445,
 446
Berger, J.O., 417
Bhattacharya, R.N., 99, 101, 145
Bickel, P.J., 73, 74, 80–82, 85, 92,
 118, 120, 123, 168, 189, 231,
 237, 247, 250, 261, 268, 306,
 323, 359
Billingsley, P., 76
Bloch, D.A., 117
Boos, D.D., 186, 187, 376, 421
Booth, J.G., 60, 210, 231, 268, 330
Bose, A., 151, 396, 397, 399, 401
Bowman, A.W., 352, 362, 364, 365,
 414
Box, G.E.P., 402
Breiman, L., 382
Brownie, C., 187
Bühlmann, P., 409
Bunke, D., 188, 305
Burman, P., 309, 414
Burr, D., 350
Butler, R.W., 268
Cao-Abad, R., 366
Carlstein, E., 388, 407
Carroll, R.J., 284
Casella, G., 424
Chan, Y.M., 110, 112, 376
Chao, M.T., 250
Chatterjee, S., 379, 382, 402
Chen, H., 187,
Chen, J., 210, 268
Chen, Y., 379
Chen, Z., 231
Cheng, P., 82
Chhikara, R.S., 67
Chiang, Y.C., 189
Chow, E., 414
Chow, Y.S., 419, 428, 447
Chung, K.L., 76, 447
Clarke, B.R., 38
Cochran, W.G., 234, 235, 256, 270
Cox, D.R., 345, 350
Craig, A.T., 418, 446
Craven, P., 367, 371
Cressie, N.A.C., 414
Csörgő, M., 117
Csörgő, S., 85, 189
Daniels, H.E., 205

- Datta, S., 94, 393, 394
 Davis, M.H.A., 398
 Davison, A.C., 189, 204, 205,
 211–215, 221, 379, 381, 415,
 422
 DeAnglis, D., 118, 290
 DeBeer, C.F., 405
 Deheuvels, P., 124
 DeWet, T., 404, 405
 Diaconis, P., 108
 DiCiccio, T.J., 117, 135–137,
 161–164, 166, 168, 170, 189,
 206, 231
 Dikta, G., 369
 Do, K.A., 219, 223, 226, 227, 231
 Dodge, Y., 118
 Doss, H., 189
 Draper, N.R., 283
 Droge, B., 305
 Ducharme, G.R., 156, 188, 189
 Dudley, R.M., 143
 Duncan, G.T., 334
 Duttweiler, D.L., 53, 54
 Efron, B., 9, 12, 14, 15, 28, 32,
 49, 107, 108, 113, 131–137,
 139, 140, 170, 184, 189, 201,
 206, 208, 210, 215–219, 240,
 248, 280, 289, 290, 304, 349,
 379–381, 384, 396, 397, 425
 Eriksson, B., 404
 Falk, M., 82, 96, 116, 151, 356
 Fan, J., 434
 Faraway, J.J., 353, 355, 359, 363,
 364
 Fay, R.E., 278
 Feller, W., 393
 Feluch, W., 353
 Ferguson, T.S., 418
 Fernholz, L.T., 34, 85
 Feuerverger, A., 204, 206
 Fisher, L., 16, 383
 Fisher, N.I., 189
 Folks, J.L., 67
 Fox, T., 191, 334
 Francisco, C.A., 238, 268, 270
 Frangos, C.C., 140
 Franke, J., 402, 414
 Frankel, M.R., 18, 234, 239, 244
 Freedman, D.A., 73, 74, 80–82, 85,
 107, 118, 120, 123, 189, 237,
 247, 250, 261, 268, 320, 322,
 324, 406, 407
 Fu, J.C., 446
 Fuh, C.D., 394
 Fuller, W.A., 238, 268, 270, 387
 Gaenssler, P., 85
 Ganeshanandam, S., 379, 381
 Gangopadhyay, A.K., 385
 Gastwirth, J.L., 117
 Gaver, D.P., 384
 Geisser, S., 309
 Gelfand, A.E., 424
 George, E.I., 424
 Geyer, C.J., 157
 Ghosh, J.K., 53, 101, 145, 189
 Ghosh, M., 68, 86, 89
 Gleason, J.R., 212
 Gill, R.D., 83
 Giné, E., 80, 81, 85, 186
 Götze, F., 151
 Graham, R. L., 215, 219
 Gray, H.L., 65, 267, 414
 Gross, A.J., 203, 250, 384
 Gross, S., 250
 Gruet, M.A., 336
 Gu, M., 350
 Gupta, V.K., 243
 Gurney, M., 243
 Haddock, J., 397
 Haeusler, E., 433
 Hájek, J., 57
 Hall, P., 16, 60, 80, 91, 93, 97,
 102, 103, 117, 140, 141,
 144–146, 151, 152, 154, 155,
 157–159, 165, 166, 177, 186,
 188, 189, 210, 212–214, 216,
 219, 221–223, 226, 227, 231,
 268, 290, 295, 297, 298, 300,
 301, 324, 330, 352, 354–359,

- 361–364, 369, 379, 381, 415,
446, 447
- Hampel, F.R., 33
- Hannan, E.J., 398
- Härdle, W., 361, 362, 364–366, 369,
414
- Hartigan, J.A., 9
- Hartley, H.O., 235
- Has'minskii, R.Z., 97
- Hastie, T.J., 385
- He, K., 290, 323
- Helmers, R., 41, 84, 85, 95, 151
- Hemerly, E.M., 398
- Hilferty, M.M., 135
- Hill, R.C., 298
- Hinkley, D.V., 22, 68, 108, 110, 157,
173, 184, 187, 189, 191, 204,
205, 211–215, 221, 231, 286,
292, 334, 422, 442
- Hogg, R.V., 418, 446
- Holbert, D., 396, 397
- Honda, M., 379, 381
- Horváth, L., 384
- Huang, J.S., 128, 191, 200
- Huber, P.J., 37, 284
- Huet, S., 322, 335, 336
- Huskova, M., 435
- Ibragimov, Z.A., 97
- Jaeckel, L., 48, 202
- Janas, D., 397
- Janssen, P., 41, 84, 85, 186, 435
- Jarrett, R.G., 11
- Jenkins, G.M., 402
- Jennrich, R.I., 332, 333
- Jewett, R.S., 243
- Jhun, M., 188, 353, 355, 359, 383
- John, P.W.M., 197
- Johns, M.V.Jr., 224, 225
- Johnson, R.A., 423
- Jolivet, E., 322, 335, 336
- Jones, M.C., 355, 356
- Jorgensen, M.A., 191
- Kabala, P., 397
- Kalton, G., 270, 271
- Karrison, T., 350
- Kaufman, E., 151
- Kendall, M.G., 209
- Kim, Y.B., 397
- Kinateder, J.G., 81
- Kipnis, V., 305
- Kish, L., 18, 234, 239, 244
- Klenk, A., 82
- Knight, K., 80
- Knoke, J.D., 379
- Koehler, K., 350
- Koenker, R., 284
- Konishi, S., 135, 136, 138, 173, 206,
379, 381
- Koronachi, J., 353
- Kovar, J.G., 238, 239, 245, 251, 254,
255, 257
- Kreiss, J.P., 402
- Krewski, D., 186, 234, 237, 244, 261,
268
- Krieger, A.M., 189
- Krzanowski, W.J., 379, 381
- Kuk, A.Y.C., 281
- Kulperger, P.J., 393
- Künsch, H.R., 391, 408
- La Scala, B., 446
- Lachenbruch, P.A., 379
- Lahiri, S.N., 95, 96, 323, 409, 411,
412, 414
- Lai, T.L., 96, 321, 414
- Laird, N.M., 189, 424
- Larntz, K., 191, 334
- Läuter, H., 379, 382
- Lawley, D.M., 382
- LeCam, L., 85, 143, 422
- Lee, K.W., 343
- Lee, S., 117, 189
- Léger, C., 356, 412
- Lehmann, E.L., 2, 61, 132, 193
- Lele, S., 414
- LePage, R., 81
- Li, K.-C., 368, 371,
- Li, L.-A., 446
- Liang, K.-Y., 339, 340
- Linder, E., 330
- Liu, J., 268

- Liu, K., 409
 Liu, R.Y., 83, 99, 100, 153, 188, 292,
 293, 323, 384, 391, 433, 440
 Liu, Z., 385
 Lo, A.Y., 417–420, 427, 428
 Lo, S.H., 83, 250, 384
 Loh, W.-Y., 123, 138, 143, 144, 160,
 161, 168, 174, 187, 189
 Lohse, K., 84
 Loughin, T.M., 350
 Louis, T.A., 189, 424
 Mallows, C.L., 73, 307
 Mammen, E., 85, 302, 323–325
 Manteiga, W.C., 182
 Mantel, H.J., 30
 Maritz, J.S., 11
 Marron, J.S., 351–353, 355, 356,
 361, 362, 366
 Martin, M.A., 117, 140, 151, 158,
 159, 163, 173, 186, 189, 231
 Mason, D.M., 85, 124, 189, 427, 433,
 434, 436
 Maxwell, A.E., 382
 McCarthy, P.J., 241, 247, 250
 McCormick, W.P., 393, 394
 McCullagh, P., 338
 McDougall, A.L., 398
 McLachlam, G.J., 379
 Mei, C., 434
 Mesean, A., 336
 Mickey, M.R., 379
 Millar, P.W., 85, 143, 152, 182, 183,
 445, 446
 Miller, R.G., 18, 24, 68, 285, 286,
 384
 Monahan, J.F., 421
 Morrison, D.F., 380, 383
 Moulton, L.H., 343
 Mykland, A., 414
 Nagaev, S.V., 394
 Nagao, H., 373, 376
 Navidi, W., 321
 Nelder, J.A., 338
 Newton, M.A., 157, 422–424, 427,
 433, 434, 436, 445
 Nigam, A.K., 243
 Nolan, D., 414
 Noreen, E.W., 186
 Olshen, R.A., 306
 Owen, A.B., 185, 422, 446
 Padgett, W., 385
 Pantula, S.G., 376
 Park, B., 351–353, 355, 356
 Parr, W.C., 38, 39, 67, 105
 Peck, R., 383
 Peters, S.C., 107, 407
 Petrov, V.V., 91, 103, 447
 Pfanzagel, J., 136
 Pittelkow, Y.E., 300, 301
 Politis, D.N., 60, 70, 127, 128, 409,
 412–414
 Portnoy, S., 284
 Poskit, D.S., 398
 Prakasa Rao, B.L.S., 113, 351, 393
 Quan, H., 347
 Quenneville, B., 188
 Quenouille, M.H., 4
 Quindimil, M.P., 182
 Qumsiyeh, M., 99, 101
 Raftery, A.E., 422–424, 445
 Ramos, E., 413
 Rao, C.R., 1, 2, 307, 384, 434, 447
 Rao, J.N.K., 12, 18, 233–235,
 237–242, 244, 245, 248, 251,
 255–257, 261, 262, 268, 269,
 271, 273, 281, 289, 295, 314
 Rayner, R.E., 302
 Reeds, J.A., 38
 Reid, N., 203, 204
 Reiss, R.D., 96, 116
 Riemer, S., 188
 Rissanen, J., 398
 Robbins, H., 321
 Robinson, J., 58
 Roche, D.M., 205
 Romanazzi, M., 373
 Romano, J.P., 60, 70, 117, 127, 128,
 136, 137, 161–164, 166, 182,
 183, 356, 385, 409, 412–414
 Rosenblatt, M., 359

- Rubin, D.B., 19, 277, 417, 418, 424, 425, 445
Rudemo, M., 352
Ruppert, D., 284
Rutherford, B., 366
Sain, S.R., 351
Sauermann, W., 343
Savage, L.J., 178
Schechtman, E., 189, 211–215, 221
Scheffé, H., 299
Schenker, N., 135, 275, 277, 278, 280
Schmidt, W.H., 439
Schneller, W., 433
Schucany, W.R., 39, 54, 65, 67, 127, 140, 228, 267, 397, 414
Schuster, E.F., 185, 186
Schwartz, G., 307, 398
Scott, D.W., 351, 353
Searle, S.R., 283
Sedransk, J., 271
Sen, P.K., 28, 44, 48, 126, 385
Serfling, R., 27, 28, 35, 37, 38, 41, 51, 53, 57, 64, 77, 79, 81, 84, 85, 88, 96, 125, 208, 224
Shao, J., 29, 37, 39, 44–49, 53, 54, 61, 82, 83, 85, 88, 90, 121, 125, 126, 193, 196, 200, 237, 238, 244, 245, 256, 264, 268–271, 273, 281, 289, 290, 292, 295, 309–312, 314, 318, 319, 327, 333, 334, 338, 340, 341, 344, 347, 390, 392
Shao, Q., 409–411
Sheather, S.J., 54, 127
Shi, S., 157, 231
Shi, X., 60, 68, 81, 95, 210, 388, 390, 407, 409, 436
Shorack, G.R., 85, 124, 290, 323
Silverman, B.W., 113, 114, 351, 357
Simonoff, J.S., 343
Singh, A.C., 30
Singh, K., 75, 80, 82, 83, 85, 92–95, 99–101, 118, 153, 188, 210, 292, 293, 384, 387, 391
Sitter, R.R., 12, 238, 243, 249–251, 257, 268, 281
Smith, A.F.M., 424
Smith, H., 283
Snapinn, S.M., 379
Snell, E.J., 350
Snowden, C.B., 247, 250
Son, M.S., 396, 397
Srivastava, M.S., 110, 112, 373, 376, 405
Stangenhaus, G., 290
Staurt, A., 209
Stein, C., 28, 137
Stewart, T.J., 420
Stine, R.A., 305, 397
Stoffer, D.S., 414
Stone, C.J., 353, 367
Stone, M., 307
Stout, W.F., 388, 389
Stute, W., 82, 182, 322, 353, 368, 384
Sun, W.H., 350
Sutton, C.D., 188
Swanepoel, J.W.H., 124, 165, 405, 413
Tanner, M.A., 424
Taylor, C.C., 353, 355
Teicher, H., 419, 428, 447
Terrell, G.R., 353
Thombs, L.A., 385, 397
Thorburn, D.E., 68
Tibshirani, R.J., 15, 135, 163, 164, 168, 170, 189, 206, 210, 349, 385, 396, 397
Tikhomirov, A.N., 390
Tong, Y.L., 108
Tsai, C., 343
Tsai, W.-Y., 347
Tsiatis, A.A., 346, 347
Tu, D., 69, 82, 106, 140, 173, 203, 366, 379, 384, 385, 391, 424, 427, 429, 431, 434–440, 442
Tukey, J., 5, 6, 22, 68
Valliant, R., 244, 256
van Ness, J., 383

- van Wyk, J.W.J., 165, 404, 405, 413
van Zwet, W.R., 419
Veall, M.R., 407
Venter, J.H., 165
Veraverbeke, N., 85, 186
Wahba, G., 367, 371, 372
Wall, K.D., 414
Wang, J., 384
Wang, J.Q.Z., 96
Wang, M.C., 379
Wang, S., 115, 204, 205, 228
Wax, M., 398
Wedderburn, R.W.M., 338
Wei, B.C., 68, 108, 110, 173, 442
Wei, C.Z., 321, 398, 399
Welsh, A.H., 275, 280
Weng, C.S., 419, 427
Willemain, T.R., 397
Wilson, E.B., 135
Wilson, S.R., 177, 188
Wolter, K.M., 244, 245
Wong, W., 424
Wood, A.T.A., 231
Woodruff, R.S., 238
Worton, B.J., 189, 422
Wu, C.F.J., 12, 18, 22, 44, 49, 53,
57–59, 70, 138, 168, 210,
238–243, 245, 248, 251, 255,
257, 262, 264, 268, 270, 281,
287, 291, 292, 294, 332, 333,
347, 439, 440
Wu, Y., 307
Yahav, J.A., 231
Yakowitz, S., 366
Yandell, B.S., 384
Yang, S.S., 76, 83, 170
Young, G.A., 114, 117, 118, 185,
187, 189, 205, 231, 290
Yu, H., 409–411
Yu, K., 434
Yu, Z., 391
Yue, K., 238, 239
Zeger, S.L., 339, 340, 343
Zhang, J., 376
Zhang, L., 69, 106, 140, 173, 203,
437, 440, 442
Zhang, P., 309
Zhao, L., 430, 434, 438, 447
Zheng, Z., 19, 385, 426, 427, 429,
431–436, 438–440
Zinn, J., 80, 81, 85

Subject Index

A

ANOVA decomposition, 216
Acceleration constant, 135
calculation, 136–139
for function of means, 139
for ratio of the means, 138
for variance, 138
in linear models, 296, 298
in multiparametric models, 137
in nonparametric models, 139
in one parametric models,
136–137
jackknife approximation, 140
Accuracy, 21, *see also* Convergence
rates
Bayesian posterior distribution
estimators, 408
bootstrap confidence sets, 144–152
bootstrap confidence sets for
regression parameters,
295–298
bootstrap distribution estimators,
91
bootstrap distribution estimators
of regression parameters, 321
randomly weighted distribution
estimators, 429–433
smoothed bootstrap distribution
estimators for sample
quantiles, 116
Accuracy measures, 1
bootstrap estimators, 9–17
jackknife estimators, 4–9
traditional estimators, 2–4

Adjusted jackknife variance
estimators, 273
Adjusted one-step iteration, 229
 α -mixing sequence, 411
Antithetic bootstrap resampling,
221–223
for bias estimators, 222
for distribution estimators, 222
for quantile estimators, 223
for variance estimators, 222
Antithetic sampling, 221
Approximated BRR, 244–246
Asymptotic accuracy, *see* Accuracy
Asymptotic bias, 61
Asymptotic comparison
approximated jackknife variance
estimators, 196
first order accurate confidence
intervals, 153–154
Asymptotic correctness, 178
Asymptotic efficiency, 42
Asymptotic expansion, *see*
Edgeworth expansion
Asymptotic mean squared error, 99
bootstrap distribution estimators,
100–101
normal approximation, 100
one-term Edgeworth expansion,
100–101
Asymptotic minimaxity
bootstrap estimators, 98
confidence sets, 152
distribution estimators, 97
hybrid bootstrap confidence sets,
152

- Asymptotic (*cont.*)
 normal approximation, 98
 one-term Edgeworth expansion, 98
- Asymptotic variance, 21
 of function of sample means, 24
 of U-statistics, 27
- Autoregressive moving average
 (ARMA) time series, 400
 bootstrap procedures, 402–403
- Autoregressive (AR) time series, 394
 least square estimators, 395
 model selection, 397–399
 nonstationary, 397
 order estimation, 397
 prediction, 397
 spectral density, 413–414
 stationary, 394
- Average mean squared prediction
 error, 304
 bootstrap estimators in linear
 model, 304
- B**
- Balanced bootstrap resampling
 for bias estimators, 211–212
 for distribution and quantile
 estimators, 214
 for variance estimators, 213
- Balanced incomplete block design, 197, 309
- Balanced repeated replication
 (BRR) variance estimators, 241–246
 adjusted, 242
 approximation, 244–246
 consistency, 260–264
 construction, 243–244
 grouped, 244
 random sampling, 245
 repeatedly grouped, 244–245
- Balanced subsampling
 delete-d jackknife variance
 estimators, 197–198
- Bandwidth selection
 by bootstrapping, 353–356,
 362–363
 by cross-validation, 351–353,
 360–361
- Bayesian bootstrap
 accuracy, 419
 consistency, 419
 credible bands, 420
 procedures, 417
 using prior information, 420–422
 with a noninformative prior,
 416–425
- Bayesian statistical analysis, 416
- Berry-Esséen's inequality
 for finite sample means, 58
 for homogeneous Markov chains,
 394
 for independent sample means,
 74–75, 451
 for m-dependent data, 390
- Bias adjusted cross-validation, 353
- Bias of a statistic, 5
 bootstrap estimators, 14
 iterative bootstrap estimators,
 158
 jackknife estimators, 5
- Bias reduction, 64–66
- Bias-reduced jackknife estimators, 5
 comparison, 67–68
 mean squared errors, 65–66
- Block sizes, 412
- Bootstrap, 9
 based on residuals (RB), 17,
 289–290, 395–396
 circular block, 409–410
 external (EB), 291–292
 grouped external, 390
 linear, 219–221
 mirror-match (BMM), 249–250
 moving block, 391
 naive, 246–247
 nonparametric, 16
 one-step, 228–229
 paired (PB), 17, 291

- parametric, 16
relationship with jackknife, 11–12,
 202–203
rescaling (BRS), 248–249
smoothed, 113–118
weighted, 18, 291
with replacement (BWR),
 247–248
without replacement (BWO),
 250–251
without resampling, 206
- Bootstrap accelerated bias-corrected (BC_a) percentile, 136, *see also* Acceleration constant
accuracy, 149–150
consistency, 142
length, 154
- Bootstrap based on residuals (RB)
for autoregressive time series,
 395–396
in generalized linear models,
 341–342
in linear regression, 17, 289–290
in multivariate linear models, 377
in nonlinear regression, 335
- Bootstrap bias-corrected (BC)
percentile, 134
accuracy, 148–149
analytic approximation, 206
consistency, 142
error in coverage probability, 153
for sample variance, 136
length, 154
- Bootstrap bias estimation, 4, 105
fixed sample performance,
 105–107
- Bootstrap calibrating, 160
confidence intervals, 161
lower confidence bounds, 160
- Bootstrap clone method, *see*
 Random weighting method
- Bootstrap confidence sets
accuracy, 144–152
bootstrap BC, 148–149
bootstrap BC_a, 149–150
- bootstrap percentile, 148
bootstrap-t, 145–146
hybrid bootstrap, 146–148
asymptotic minimaxity, 152–153
comparisons by errors in coverage
 probability, 154–155
comparisons of lengths, 154–155
consistency, 141–144
empirical comparisons, 166–176
for covariance matrix, 374
for density function, 356–359
in linear models, 295–298, 324
in multivariate regression, 377
in nonlinear models, 336
in nonparametric regression, 363,
 366
- Bootstrap data sets, 11
- Bootstrap distribution estimation,
 12–13
accuracy, 91–97
asymptotic mean squared error,
 99–101
asymptotic minimaxity, 97–99
asymptotic relative error, 102–105
balanced bootstrap resampling
 approximations, 214
convergence rate, 91–97
delta method approximations,
 201–202
fixed sample performance,
 108–112
jackknife approximations, 202–203
Monte Carlo approximations,
 207–210
saddle point approximations,
 203–205
techniques in proving consistency,
 72–79
- Bootstrap distance tests, 182–184
- Bootstrap hypothesis tests, 176–188
distance tests, 182–183
for covariance matrix, 375
for generalized linear hypothesis,
 377–378
for independence, 182–183

- Bootstrap (*cont.*)
- for linear hypothesis, 301–303, 325
 - for multivariate normality, 181
 - for population mean, 180–181, 184–185
 - for scales, 187
 - for symmetry about the origin, 181–182, 185–186
 - for two-sided hypotheses with nuisance parameters, 179–180
 - goodness-of-fit test, 183
 - k-sample problems, 187
 - optimization approach, 184–186
 - p-values, 186
- Bootstrap inverting, 157
- Bootstrap model selection
- for autoregressive time series, 398–400
 - for general linear models, 311–313, 327–329
 - for generalized linear models, 343–345
- Bootstrap percentile, 132
- accuracy, 148
- consistency, 142
- for sample median, 132
- invariance, 132
- length, 154
- Bootstrap prediction intervals, 305
- Bootstrap prepivoting, 156
- Bootstrap rejection region, 180
- Bootstrap sample sizes
 - consistency, 118–125, 312
 - See also* Monte Carlo sample sizes
- Bootstrap skewness estimators, 105
- fixed sample performance, 105–107
- Bootstrap-t, 131–133
- accuracy, 145–146
 - consistency, 142
 - disadvantages, 131
- Bootstrap variance estimators, 9–11
- consistency, 86–90
 - fixed sample performance, 105–107
- Monte Carlo approximation, 11
- theoretical form, 10
- Bootstrapping pairs, *see* Paired bootstrap
- Bootstrapping residuals, *see* Bootstrap based on residuals
- Bootstrapping under imputation, 278–280
- Borel-Cantelli lemma, 87, 449
- C**
- Centered Monte Carlo approximation
- to bootstrap bias estimators, 216
 - to bootstrap quantiles, 218
 - to bootstrap variance estimators, 217
- Centering after Monte Carlo, *see* Centered Monte Carlo approximation
- Central limit theorem, 451
- Characteristic function, 76
- Circular block bootstrap, 409–410
- accuracy, 411–412
 - consistency, 410
- Cluster sampling, 234
- Clustering, 383
- Compact differentiability, *see* Hadamard differentiability
- Comparison of lengths, 154
- Confidence intervals
- length, 154
 - one-sided, 129
 - traditional approach, 130
 - two-sided, 130
 - See also* Confidence sets and lower confidence bounds
- Confidence sets, 129–131
- automatic percentile, 162–163
 - based on transformation, 173
 - bootstrap calibrating, 160–161
 - bootstrap inverting, 157–158

- bootstrap preprinting, 155–156
confidence coefficient, 129
exact, 130
fixed width bootstrap, 164–165
for a distribution, 143
for regression parameters, 295–298
iterative bootstrap, 155–159
level, 130
likelihood-based bootstrap,
 165–166
variance stabilizing bootstrap,
 163–164
See also Confidence intervals and
 lower confidence bounds
- Consistency
 Bayesian posterior distribution
 estimators, 407–408
 bootstrap distribution estimators,
 72–86
 bootstrap variance estimators,
 86–90
 confidence sets, 141–142
 distribution estimators, 72
 jackknife variance estimators,
 24–28
 model selection procedures, 308
 randomly weighted distribution
 estimators, 427–429
 strong, 20
 tests, 178
 weak, 20
- Convergence
 almost surely, 447
 in distribution, 447
 in moments, 79
 in probability, 447
 of transformations, 448
- Convergence rate
 jackknife variance estimators, 42
 smoothed bootstrap variance
 estimators for sample
 quantiles, 117
See also Accuracy
- Convolution functionals, 35
- Cornish-Fisher expansion, 145, 147,
 454
- Correct model, 307
- Corrected jackknife bias estimators,
 240
- Correlation coefficient, 133
 comparison of confidence
 intervals, 169–170
 See also Sample correlation
 coefficient
- Covariance matrix, 373
 bootstrap confidence regions, 374
 bootstrap tests, 375–376
- Cox's regression models, 345
- Cramér and Wald's theorem, 448
- Cramér's condition, 396
- Cramér-von Mises test statistic
 consistency of jackknife variance
 estimators, 39–40
 weighted, 121
- Critical values of a test, 178
 bootstrap estimators, 180, 182
- Cross-validation
 for bandwidth selection, 351–353,
 360–362, 367–368
 delete-d, 309, 337
 for linear model selection,
 307–311, 327
 for nonlinear regression model
 selection, 337–338
 generalized, 367–368, 371–372
- Cumulate adjustment formula, 218
- D**
- Darbin-Watson test statistic, 405
- Data resampling, 12, 16
- Degradation models, 30–31
- Delta method
 approximation for bootstrap
 estimators, 201–202
- Delete-d cross-validation
 for linear model selection,
 309–311, 327

- Delete-d (*cont.*)
 for nonlinear regression model selection, 337–338
- Delete-d jackknife, 49
 histograms, 56
 consistency, 57–58, 59–60
 convergence rates, 58
- variance estimators, 50
 balanced subsampling, 197–198
 comparison with delete-1 jackknife, 54–55
 consistency, 52–53
 random subsampling, 198–200
- Delete-1 jackknife variance estimators, *see* Jackknife variance estimators
- Density function
 bootstrap confidence sets, 356–358
 simultaneous bootstrap confidence bands, 359
- Design-based approach, 233
- Diagram for bootstrap, 15, 279
- Differentiable functionals, *see* Statistical functionals
- Dirichlet distribution, 418
- Dirichlet prior, 418
- Discriminate analysis, 379–382
- E**
- Edgeworth expansion, 158
 for dependent random variables, 396
 for functions of sample means, 138
 for sum of i.i.d. random variables, 93, 102–103, 452–453
 for sum of independent random variables, 430, 452–453
- Edgeworth expansion distribution estimators
 accuracy, 93
 asymptotic mean squared error, 100–101
 asymptotic minimaxity, 98
- asymptotic relative error, 102–103
 for ratio estimators, 108
 for sample variances, 110
- Efficiency of jackknife and bootstrap, 293
- Empirical comparison, 22
 accuracy of saddle point approximation, 204–205
 bias-reduced jackknife estimators and the original estimators, 67
- bootstrap BC_a and transformation, 172–173
 bootstrap and jackknife bias estimators, 107
 bootstrap and jackknife skewness estimators, 107
 bootstrap and jackknife standard deviation estimators, 106
 bootstrap and jackknife variance estimators, 107
 bootstrap histogram and normal density estimators, 108–109
 bootstrap, normal approximation and Edgeworth expansion estimators of distribution functions, 108–112
 bootstrap percentile, BC and BC_a, 167
 bootstrap percentile, BC, BC_a and bootstrap-t, 168–169
 bootstrap percentile, bootstrap-t, hybrid bootstrap and their one-step iterations, 173–174
- bootstrap procedures for ARMA time series, 403
- bootstrap simultaneous confidence intervals in multivariate linear models, 378
- confidence intervals for correlation coefficient, 169
- confidence intervals in linear regression with time series errors, 406

- confidence intervals for nonparametric regression functions, 370
- coverage probabilities of bootstrap simultaneous confidence bands, 301
- delete-1 and -d jackknife variance estimators, 54
- efficiencies of random subsampling, 200
- estimated level of bootstrap tests, 184
- estimators of misclassification rates, 381
- GML and GCV, 372
- hybrid bootstrap, bootstrap-t and normal approximation, 171
- jackknife and linearization, 32
- levels of bootstrap tests, 378
- methods of bandwidth selection, 356
- model selection probabilities, 310
- normal approximation, bootstrap percentile, BC and BC_a , 167–168
- normal approximation, bootstrap-t and Edgeworth expansion, 170–173
- one-step jackknife variance estimators, 193–195
- powers of bootstrap tests, 185
- random weighting distribution estimators, 440–445
- relative biases of variance estimators in linear models, 294
- resampling variance estimators and confidence intervals in sample surveys, 251–258
- smoothed bootstrap standard deviation estimators, 114
- Empirical distribution, 4 randomly weighted, 426
- Empirical p -dimensional marginal distribution, 407
- Empirical processes consistency of bootstrap distribution estimators, 85 randomly weighted, 434 statistics based on, 4
- Empirical saddle point approximation, 204
- Equal probability of selecting, 233
- Equal variance assumption, 284
- Error in coverage probability, 153 bootstrap BC, 153 hybrid bootstrap, 153 normal approximation, 153
- Estimation of accuracy measures by bootstrap, 9–17 by jackknife, 4–9 by traditional approach, 2–4
- Estimation of poverty line, 29–30
- Estimators based on tests, 125–126
- Exact level of a test, 178
- Exact lower confidence bounds, 134, 136, 144, 162
- External bootstrap for m-dependent data, 390 in linear models, 291–292 in nonparametric regression, 365–366
- Extreme order statistics, 123–125, 143–144
- F**
- Factor analysis, 382
- Finite population distribution, 235 estimators, 237
- Finite sample correction, 56
- Fisher's linear discriminant function, 379
- Fix point of an iterative process, 191
- Fixed sample performance, *see* Empirical comparison
- Fixed width bootstrap confidence intervals, 165
- Fréchet differentiability, 33–34 continuous, 35–36

- Fréchet (*cont.*)
 second order, 45
 uniformly, 47
- Frequentist statistical analysis, 416
- Functionals, *see* Statistical functionals
- Functions of (weighted) averages, 237
 accuracy of bootstrap distribution estimators, 94
 consistency of bootstrap, 267–268
 consistency of jackknife and BRR, 262–264
 consistency of RGBRR and RSBRR, 264–267
 corrected jackknife bias estimators, 240
 jackknife variance estimators, 238–240
 linearization variance estimators, 237
 relation between linearization and jackknife, 240
See also Functions of sample means
- Functions of population totals, 236
- Functions of sample means
 bias reduction by jackknife, 7, 66
 bootstrap variance estimators
 consistency, 86–89
 calculation of acceleration constant, 139–140
 consistency of bootstrap distribution estimators, 78–79, 80–81
- jackknife bias estimators, 7
- jackknife variance estimators
 description, 6–7
 bias, 28–29
 consistency, 24–27
 convergence rates, 42–44
 delete-d, 50–53
 nondifferentiable, 122
 variance estimation by bootstrap, 10
- variance estimation by traditional approach, 2–3, 29
 with null derivatives, 119–120
- G**
- Gâteaux differential, 33
 continuous, 35
 uniformly, 47
- General bootstrap, *see* Random weighting
- General iterative bootstrap, 158–159
 bias estimation, 158
 computational algorithm, 159
 interval estimation, 158
- General linear models, 16–17, 283–284, *see also* Linear regression models
- Generalized cross validation (GCV), 367, 371
- Generalized empirical distributions, 41
 statistics based on
 consistency of bootstrap distribution estimators, 84–85
 consistency of jackknife variance estimators, 41–42
- Generalized linear models, 338
 bootstrap procedures, 341–343
 model selection, 343–345
 jackknife variance estimators, 340–341
- Generalized maximum likelihood method, 371
- Gini's mean difference, 38, 106
 comparison of jackknife and bootstrap estimators, 107
- Goodness of fit test, 183
- Grouped bootstrap variance estimators
 for sample mean of m-dependent data, 389
- Grouped balanced repeated replication (GBRR), 244

Grouped external bootstrap
distribution estimators for
m-dependent data, 390–391

Grouped jackknife variance
estimators
approximations, 195
for sample mean of m-dependent
data, 389

Grouping and random subsampling, 195–196

Grouping method, 388

H

Hadamard differentiability, 33
second order, 45

Hodges-Lehmann estimators, 41

Hot deck imputation, 271–272
multiple bootstrap, 277–278

Hybrid bootstrap, 141
accuracy, 146–148
asymptotic minimaxity, 152–153
consistency, 142
error in coverage probability, 153
length, 154

Hypothesis testing
alternative hypothesis, 177
asymptotic correctness, 178
complex null hypothesis, 177
consistency, 178
critical value, 178
general description, 177
level, 178
null hypothesis, 177
rejective region, 178
simple null hypothesis, 177
type I error, 178

I

Imitation method, 76–78, 322

Importance bootstrap resampling, 223–227

Imputation, 270

Inconsistency

bootstrap distribution estimators
for m-dependent data, 387–388
for sample mean, 80
for U-statistics, 81
in nonregular cases, 119–126

bootstrap variance estimators, 86

jackknife variance estimators for
sample quantiles, 49

Infinitesimal jackknife, 48, 191, 202, 206

Information matrix, 422

Interquartile range, 14

Inverse of a distribution function, 4,
see also Quantile function

Iterative bootstrap, 155–159

J

Jackknife, 4
adjusted, 273–277
delete-d, 49
grouped, 389
linear, 334, 347
moving block, 391
one-step, 191–195
relationship with bootstrap,
11–12, 202–203
weighted, 285–289

Jackknife and Edgeworth
approximation, 203

Jackknife approximation, 203

Jackknife bias estimators, 5
asymptotic properties, 61–64
fixed sample performance,
105–107

Jackknife histograms, 56
consistency, 57–58, 59–60
convergence rates, 58

Jackknife pseudovalues, 6

Jackknife skewness estimators, 69
fixed sample performance,
106–107
modification, 203

Jackknife variance estimators, 6
asymptotic properties, 23–55
bias, 28

Jackknife (*cont.*)

- computations, 190–200
- consistency, 24–28
- fixed sample performance, 107
- for functions of weighted average, 238–240, 260–264
- grouped approximation, 195
- explicit formula for L-statistics, 190–191
- one-step approximation, 191
- random subsampling approximation, 196

K

Kernel estimators

- for density function, 351
- for nonparametric regression function, 360, 364

L

L-statistics

- consistency
 - of bootstrap distribution estimators, 82–83
 - of bootstrap variance estimators, 89
 - of jackknife variance estimators, 38–39
 - of random weighting, 435
 - explicit formula of jackknife variance estimators, 190–191
 - smooth, 38
 - trimmed, 82
 - untrimmed, 83
- L-functional, 38, *see also* L-statistics
- Law of iterated logarithm, 450
- Law of large numbers, 449–450
- Law school data, 108, 170, 414
- Least squares estimators (LSE)
 - in autoregressive time series, 395
 - bootstrap distribution estimators, 395–396
- in linear regression, 284

bootstrap bias estimators,

289–292, 318

bootstrap distribution

estimators, 320–323

bootstrap variance estimators,

289–292, 316–318

jackknife bias estimators,

285–289, 318–319

jackknife variance estimators,

285–289, 313–316, 333–335

randomly weighted distribution estimators, 438–440

randomly weighted variance estimators, 428

unbiased variance estimators, 285

weighted jackknife bias estimators, 287

weighted jackknife variance estimators, 286–287

weighted delete-d jackknife variance estimators, 287

in linear regression with time series errors, 404

asymptotic variance, 404

bootstrap procedures, 404–406

two stage, 406–407

in multivariate linear regression, 376

in nonlinear regression, 332

bootstrap distribution estimators, 335–337

jackknife variance estimators, 333

one-step bootstrap distribution estimator, 335–336

one-step jackknife variance estimator, 333

Length of a confidence interval, 154

Liapunov's condition, 451

Lindeberg's condition, 76, 337, 451

Likelihood based bootstrap

confidence sets, 165–166

Linear bootstrap

for bootstrap bias estimators, 220

- for bootstrap variance estimators, 220
Linear jackknife variance estimators
 for maximum partial likelihood estimators, 347
 for nonlinear LSE, 334
Linear rank statistics
 consistency of jackknife variance estimators, 39
Linear regression models
 dynamical, 400, 406–407
 with independent errors, 16–17, 283–284
 with repeated measurements, 295
 with time series errors, 403
Linear statistic, 26
Linearization method
 for variance estimators, 29, 237
 in proving the consistency of bootstrap, 78–79
Lipschitz condition, 43, 351, 361
Location-scale family, 13
Longitudinal data models, 338–340
 bootstrap procedures, 342–343
 jackknife variance estimators, 340–341
Lower confidence bounds, 130
 automatic percentile, 162
 bootstrap BC_a, 135–140
 bootstrap BC, 133–135
 bootstrap calibrating, 160
 bootstrap-t, 131
 for density function, 356–358
 variance stabilizing bootstrap, 163
- M**
- m-dependent data, 387
 grouped external bootstrap, 389–390
 grouped jackknife, 389
inconsistency of naive bootstrap and jackknife, 387–388
- moving block jackknife and bootstrap, 391
modified jackknife variance estimators, 392
M-estimators
 consistency of jackknife variance estimators, 37–38
 of regression parameters, 284, 290, 323
one-step jackknife variance estimators, 192–195
Mallows' distance, 73–74, 320
 modified, 324
Marcinkiewicz strong law of large numbers
 for i.i.d. random variables, 75–76, 450
 for m-dependent random variables, 389
Markov chains, 392
 conditional bootstrap, 393
 homogenous, 392
 transition probability matrix, 392
Maximum likelihood estimators, 422
 of transition probabilities, 392
 bootstrap distribution estimators, 393
 conditional bootstrap distribution estimators, 393
Maximum partial likelihood estimators, 345
 bootstrap procedures, 349–350
 jackknife variance estimators, 346–348
Mean average squared errors, 360
Mean integrated squared errors, 351
Mean parameter
 bootstrap confidence sets, 145, 147, 153–154
 bootstrap hypothesis test, 180–181, 184–185
 comparison
 of bootstrap percentile, BC, BC_a and bootstrap-t, 168–169

Mean (*cont.*)

- of bootstrap percentile, hybrid bootstrap, bootstrap-t and their one-step iterations, 173–174
- of Bootstrap-t, calibrated normal approximation and calibrated Edgeworth expansion, 174–175
- of hybrid bootstrap, bootstrap-t and normal approximation based on the trimmed sample mean, 171

See also Sample mean

- Mean squared errors, 15
 bootstrap estimators, 15
- Mean squared prediction errors, 303
 bootstrap estimators in linear models, 303

Mirror-match bootstrap (BMM) 249

- Misclassification rates
 conditional, 379
 estimation by bootstrap, 380–381
 estimation by cross-validation, 380
 unconditional, 380

- Model selection
 for autoregressive time series, 397–399

in generalized linear models, 343–345

in linear models, 306–313, 326–329

in nonlinear regression, 337–338

- Modified jackknife variance
 estimators for m-dependent data, 392

- Monte Carlo approximation, 11
 to Bayesian bootstrap distributions, 417

to bootstrap bias estimators, 14, 207

to bootstrap distribution estimators, 13, 207

to bootstrap estimators of mean squared errors, 15

to bootstrap standard deviation estimators, 208

to bootstrap variance estimators, 11, 207

Monte Carlo method, *see* Monte Carlo approximation

Monte Carlo sample size
 for construction of confidence intervals, 209–210
 for estimation of standard deviation, 208–209
 rule of thumb, 210

Moving average time series, 387
 bootstrap procedures, 401

Moving block bootstrap
 for m-dependent data, 391
 for stationary processes, 408–409

accuracy, 411–412
 selection of block size, 412–413

Moving block jackknife
 for m-dependent data, 391
 for stationary processes, 408–409

Multiple bootstrap hot deck imputation, 277–278

Multivariate analysis, 373

Multivariate linear models, 376

N

N-smoothed bootstrap, 113

Naive bootstrap

in sample surveys, 246–247
 for dependent data, 387

Nested bootstrap algorithm, 180

Newton's method, 192, 333

Noninformative Dirichlet prior, 418

Noninformative prior, 418

Nonlinear regression model, 331–332

Nonparametric bootstrap, 16
 confidence sets, 168

Nonparametric minimum distance test statistics, 183

Nonparametric regression models,
360
kernel estimators
fixed design, 360
random regressor, 364
nearest neighbor estimates, 366
smoothing splines, 370
Normal approximation
accuracy, 93
asymptotic relative error, 102
asymptotic mean squared error,
100–101
asymptotic minimaxity, 98, 150
confidence intervals, 165
for the sampling distribution, 93
in constructing confidence
intervals, 150, 153, 154
Normalizing transformation, 135
Number of clusters, 383

O

$O(\cdot)$ and $o(\cdot)$, 448–449
One-step bootstrap, 228–229,
336–337, 342–343
One-step jackknife variance
estimators, 191
consistency, 193
fixed sample performance,
193–194
for M-estimators, 192
for nonlinear LSE, 333–334
in Cox’s regression model, 346
in generalized linear models,
340–341
One-term Edgeworth expansion
estimators, *see* Edgeworth
expansion distribution
estimators
Optimal model, 306
Optimization approach in bootstrap
hypothesis testing, 184–186
Orthogonal decomposition, 216

P

P-value, 186
bootstrap estimators, 186
Paired bootstrap
in Cox’s regression models, 349
in generalized linear models,
342–343
in linear models, 17, 291
in nonparametric regression
models, 365, 369
Parametric bootstrap, 16
confidence sets, 166, 189
hypothesis tests, 189
Partial likelihood function, 345
Pearson’s residuals, 341
 ϕ -mixing sequence, 410–411
Pivotal quantities, 130
Polya’s theorem, 57, 447
Population quantiles, 236
Woodruff’s confidence interval,
238
Population total, 236
Posterior distribution, 416
Bayesian bootstrap estimators,
417–422
weighted likelihood bootstrap
estimators, 422–424
Poverty line, 29
Poverty proportion
delete-d jackknife variance
estimators, 53–54
Predictive least square (PLS)
principle, 398
Prediction intervals, 305
bootstrap estimators, 305–306
Prediction of future responses,
303–306, 325–326, 397
Prepivoting, 156
Prior, 416–424, 445
Probability sampling, 233
Proportional hazard model, *see*
Cox’s regression model

Q

Quantile function, 4

Quantile processes

- consistency of bootstrap distribution estimators, 85

Quasi-likelihood, 339

R

R-estimators, 40–41

Random subsampling

- approximation to delete-d jackknife variance estimators, 198–200
- approximation to jackknife variance estimators, 196

Random subsampling BRR, 245

Random weighting method, 425–427

- accuracy, 429–433
- consistency, 427–429
- fixed sample properties, 440–445
- for linear models, 437–440
- for statistical functionals, 434–437

Randomization approach, 233

Randomly censored observations, 345

Randomly weighted empirical distribution, 426

Ratio estimator, 24

- bootstrap, normal approximation and Edgeworth expansion distribution estimators, 180–109

- calculation of acceleration constant, 138

- delta method approximation to jackknife estimators, 201–202

- jackknife and linearization estimators of standard deviation, 32

Ratio parameter

- comparison of bootstrap percentile, BC, and BC_a , 167
- comparison of normal approximation, bootstrap-t

and Edgeworth expansion, 170–173

See also Ratio estimator

Rectifying inconsistency of bootstrap, 118–126

Regression analysis, 18

Regression parameters in linear models

- least squares estimators (LSE), 284

M-estimators, 284

- weighted least squares estimators, 284

Repeated measurements, 295

Repeated random group (RRG), 245

Repeatedly grouped BRR, 244–245

Resampling after linear approximation, *see* Linear bootstrap

Resampling blocks of random lengths, 413

Resampling estimators, 12

Resampling methods, 12

Resampling under imputation, 270–280

Resamples, 11

Rescaling bootstrap (BRS), 248

Restricted least squares estimators in linear models, 302

ρ -mixing sequences, 411

Robustness against heteroscedasticity, 292, 334

S

Saddle point approximation, 203–206

Sample correlation coefficient, 24, 32, 106, 108

- Bayesian bootstrap estimators, 414

- bootstrap standard deviation estimators, 105

- empirical comparison of density estimators, 108–109

- of standard deviation estimators, 106
- jackknife
 - standard deviation estimators, 105
 - variance estimators, 24, 32
- normal theory standard deviation estimators, 105
- smoothed bootstrap estimators of standard deviation, 113–114
- Sample covariance matrix, 373
 - bootstrap distribution estimators, 373–374
- Sample fraction, 50
- Sample means
 - accuracy of Bayesian bootstrap, 418
 - of bootstrap distribution estimation, 92–93
 - of random weighting, 418–422
- comparison of bootstrap and random weighting, 440–445
- consistency
 - of Bayesian bootstrap, 407–408
 - of bootstrap distribution estimation, 74–76, 81
 - of random weighting, 416–418
- necessary condition for consistency of bootstrap, 81
- smoothed bootstrap density estimators, 115
- Sample median
 - bootstrap percentile confidence bounds, 132
 - bootstrap variance estimators, 10
- Sample quantiles
 - accuracy of bootstrap confidence intervals, 151
 - of bootstrap distribution estimation, 95–96
- consistency of bootstrap distribution estimation, 77–78, 268–270
- of bootstrap variance estimators, 89
- of BRR variance estimators, 268–269
- of delete-d jackknife variance estimators, 53
- in nonregular cases, 126
- in sample surveys, 237
- smoothed bootstrap estimators, 116–117
- Woodruff's variance estimators, 238
- Sample surveys, 18, 232
- Sample variances
 - bootstrap and two-term Edgeworth expansion estimators of distributions, 110–112
 - calculation of acceleration constant, 138
 - normalizing and variance stabilizing transformation, 135
- Sampling distributions, 1–2
 - estimation by bootstrap, 12–13
 - estimation by delete-d jackknife, 56
- Sampling with replacement, 198
- Sampling without replacement, 198
- Second order differentiability, 44–45
- Second order accuracy, 144, 323, *see also Accuracy*
- Sequential method, 165
- Shortest bootstrap-t confidence intervals, 154–155
- Simulation comparison, *see empirical comparison*
- Simulation consistency, 424
- Simultaneous confidence intervals
 - for density function, 359
 - for linear combinations in multivariate linear models, 377
- for nonparametric regression function, 364, 377

- Simultaneous (*cont.*)
 for regression parameters in linear models, 298–301
- Single stage simple random sampling, 233–234
- Size of a model, 306
- Skewness of a statistic, 105
 bootstrap estimators, 105
 jackknife estimators, 106
- Slutsky's theorem, 448
- Smoothing parameter, 367, 371
- Smoothed bootstrap
 for density of sample means, 114
 for distribution and variance of sample quantiles, 116
 for standard deviation estimators, 113–114
- Smoothed estimator of a distribution function, 113–114
- Spectral density function, 413–414
- Standard deviation of a statistics
 bootstrap estimators, 105
 jackknife estimators, 105
 normal theory estimators, 105
See also Variance of a statistic
- Stationary processes, 407–414
 circular block bootstrap, 409
 moving block jackknife and bootstrap, 408–409
 resampling blocks of random lengths, 413
- Stationary random variables, 387
- Statistical functionals, 32
 accuracy of random weighting, 425–426
 consistency
 of bootstrap distribution estimators, 83–84, 121–122
 of bootstrap variance estimators, 89–90
 of jackknife variance estimators, 36–37, 45–46, 48
 of random weighting, 423–424
- Fréchet differentiability, 33–34
 continuous, 35–36
- second order, 45
 uniform, 47
 uniformly second order, 63
- Gâteaux differentiability, 33
 continuous, 35
 uniform, 47
- Hadamard differentiability, 33
 second order, 45
- Lipschitz differentiability, 43
- Statistics, 1
 asymptotic bias, 60
 asymptotic variance, 20
 bias, 5
 sampling distribution, 1
 skewness, 105
 mean squared error, 15
 variance, 2
- Stochastic approximation algorithm, 182, 299–300
- Stochastic $O(\cdot)$ and $o(\cdot)$, 448–449
- Stratified sampling, 234
 one-stage simple random sampling, 236
 two stage unequal probability sampling, 236
 multistage sampling design, 234–236
- Strong consistency, 20
- Strong law of large numbers
 for i.i.d. random variables, 449–450
 for m-dependent data, 388–389
- Studentized sample means
 accuracy of bootstrap distribution estimators, 94–95
 comparison of bootstrap and random weighting, 442–445
- Studentized variables, 12, 13
- Survey weight, 235

T

- Taylor's expansion method, *see*
 Delta method
- Traditional approach

in estimating accuracy measures, 2
 in estimating sampling distributions, 12
 in constructing confidence intervals, 130
 in constructing simultaneous confidence intervals for density functions, 359
 in hypothesis testing, 178
 weaknesses and disadvantages, 4, 130
 Transition probability matrix, 392
 maximum likelihood estimators, 392
 Trimmed L-statistics, 82
 Trimmed sample means
 confidence interval based, 170
 jackknife variance estimators, 7–8
 variance estimation by traditional approach, 3–4
 Tukey's conjectures, 6, 68
 Two-sample linear rank statistics
 consistency of jackknife variance estimators, 40
 Two-sample Wilcoxon statistics, 40
 Two stage sampling, 234
 Two-term Edgeworth expansion estimators
 for ratio estimators, 108
 for sample variances, 112

U

U-smoothed bootstrap, 113
 U-statistics, 27
 accuracy of bootstrap distribution estimators, 95
 consistency
 of bootstrap distribution estimators, 81–82
 of jackknife variance estimators, 27
 of random weighting, 424
 Unequal probability sampling, 234

Uniform integrability, 54, 79, 450–451
 Untrimmed L-statistics, 83
 Upper confidence bounds, 130

V

V-statistics
 bias reduction by jackknife, 8–9, 66–67
 jackknife bias estimators, 8
 jackknife variance estimators, 8–9
 Variance of a statistic, 2
 bootstrap estimators, 9, 11
 estimation by traditional approach, 2
 jackknife estimators, 6
 Variance parameter comparison
 of bootstrap percentile, BC,
 BC_a and bootstrap-t, 168–169
 of BC_a and transformation, 173
 of normal approximation,
 bootstrap percentile, BC and BC_a, 167–168
 See also Sample variances
 Variance stabilizing bootstrap, 163–164
 Variance stabilizing transformation, 133, 135, 163

W

Weak consistency, 20
 Weak law of large numbers, 450
 Weighted average of squared errors, 344
 Weighted bootstrap variance estimators, 18, 291
 Weighted jackknife bias and variance estimators, 18, 286–287
 Weighted least squares estimators
 in generalized linear models, 338

Weighted (*cont.*)

- jackknife variance estimators,
340
- one-step jackknife variance
estimators, 340
- bootstrap procedures, 341–343
- in linear models, 284, 288
- Weighted likelihood bootstrap,
422–424
- adjusted, 424

Weighted likelihood function, 423

Wilcoxon signed rank statistic, 39

Wild bootstrap, *see* External
bootstrap

Winsorized signed rank statistic, 39

With replacement bootstrap
(BWR), 247

Without replacement bootstrap
(BWO), 250

Springer Series in Statistics

(continued from p. ii)

Pollard: Convergence of Stochastic Processes.

Pratt/Gibbons: Concepts of Nonparametric Theory.

Read/Cressie: Goodness-of-Fit Statistics for Discrete Multivariate Data.

Reinsel: Elements of Multivariate Time Series Analysis.

Reiss: A Course on Point Processes.

Reiss: Approximate Distributions of Order Statistics: With Applications to Non-parametric Statistics.

Rieder: Robust Asymptotic Statistics.

Rosenbaum: Observational Studies.

Ross: Nonlinear Estimation.

Sachs: Applied Statistics: A Handbook of Techniques, 2nd edition.

Särndal/Swensson/Wretman: Model Assisted Survey Sampling.

Schervish: Theory of Statistics.

Seneta: Non-Negative Matrices and Markov Chains, 2nd edition.

Shao/Tu: The Jackknife and Bootstrap.

Siegmund: Sequential Analysis: Tests and Confidence Intervals.

Simonoff: Smoothing Methods in Statistics.

Small: The Statistical Theory of Shape.

Tanner: Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, 3rd edition.

Tong: The Multivariate Normal Distribution.

van der Vaart/Wellner: Weak Convergence and Empirical Processes: With Applications to Statistics.

Vapnik: Estimation of Dependences Based on Empirical Data.

Weerahandi: Exact Statistical Methods for Data Analysis.

West/Harrison: Bayesian Forecasting and Dynamic Models.

Wolter: Introduction to Variance Estimation.

Yaglom: Correlation Theory of Stationary and Related Random Functions I: Basic Results.

Yaglom: Correlation Theory of Stationary and Related Random Functions II: Supplementary Notes and References.