Springer Texts in Statistics

Advisors:
George Casella Stephen Fienberg Ingram Olkin

Springer
New York
Berlin
Heidelberg
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Springer Texts in Statistics

Alfred: Elements of Statistics for the Life and Social Sciences
Berger: An Introduction to Probability and Stochastic Processes
Blom: Probability and Statistics: Theory and Applications
Brockwell and Davis: An Introduction to Times Series and Forecasting
Chow and Teicher: Probability Theory: Independence, Interchangeability, Martingales, Third Edition
Christensen: Plane Answers to Complex Questions: The Theory of Linear Models, Second Edition
Christensen: Linear Models for Multivariate, Time Series, and Spatial Data
Christensen: Log-Linear Models and Logistic Regression, Second Edition
Creighton: A First Course in Probability Models and Statistical Inference
Dean and Voss: Design and Analysis of Experiments
du Toit, Steyn, and Stumpf: Graphical Exploratory Data Analysis
Edwards: Introduction to Graphical Modelling
Finkelstein and Levin: Statistics for Lawyers
Flury: A First Course in Multivariate Statistics
Jobson: Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design
Jobson: Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods
Kalbfleisch: Probability and Statistical Inference, Volume I: Probability, Second Edition
Kalbfleisch: Probability and Statistical Inference, Volume II: Statistical Inference, Second Edition
Karr: Probability
Keyfitz: Applied Mathematical Demography, Second Edition
Kiefer: Introduction to Statistical Inference
Kokoska and Nevison: Statistical Tables and Formulae
Lehmann: Elements of Large-Sample Theory
Lehmann: Testing Statistical Hypotheses, Second Edition
Lehmann and Casella: Theory of Point Estimation, Second Edition
Lindman: Analysis of Variance in Experimental Design
Lindsey: Applying Generalized Linear Models
Madansky: Prescriptions for Working Statisticians
McPherson: Statistics in Scientific Investigation: Its Basis, Application, and Interpretation
Mueller: Basic Principles of Structural Equation Modeling
Nguyen and Rogers: Fundamentals of Mathematical Statistics: Volume I: Probability for Statistics
Nguyen and Rogers: Fundamentals of Mathematical Statistics: Volume II: Statistical Inference

(Continued after index)

Jun Shao

Mathematical Statistics

Springer

Jun Shao
Department of Statistics
University of Wisconsin, Madison
Madison, WI 53706-1685
USA

Editorial Board

George Casella
Biometrics Unit
Cornell University
Ithaca, NY 14853-7801
USA

Stephen Fienberg
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-3890
USA

Ingram Olkin
Department of Statistics
Stanford University
Stanford, CA 94305
USA

Library of Congress Cataloging-in-Publication Data
Shao, Jun.
Mathematical statistics / Jun Shao.
p. cm. — (Springer texts in statistics)
Includes bibliographical references and indexes.
ISBN 0-387-98674-X (alk. paper)
I. Mathematical statistics. I. Title. II. Series.
QA276.S458 1998
519.5—ddc21 98-45794

© 1999 Springer-Verlag New York, Inc.
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA), 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.

ISBN 0-387-98674-X Springer-Verlag New York Berlin Heidelberg SPIN 10699982

To Guang, Jason, and Annie

Preface

This book is intended for a course entitled *Mathematical Statistics* offered at the Department of Statistics, University of Wisconsin-Madison. This course, taught in a mathematically rigorous fashion, covers essential materials in statistical theory that a first or second year graduate student typically needs to learn as preparation for work on a Ph.D. degree in statistics. The course is designed for two 15-week semesters, with three lecture hours and two discussion hours in each week. Students in this course are assumed to have a good knowledge of advanced calculus. A course in real analysis or measure theory prior to this course is often recommended.

Chapter 1 provides a quick overview of important concepts and results in measure-theoretic probability theory that are used as tools in the rest of the book. Chapter 2 introduces some fundamental concepts in statistics, including statistical models, the principle of sufficiency in data reduction, and two statistical approaches adopted throughout the book: statistical decision theory and statistical inference. Each of Chapters 3 through 7 provides a detailed study of an important topic in statistical decision theory and inference; Chapter 3 introduces the theory of unbiased estimation; Chapter 4 studies theory and methods in point estimation under parametric models; Chapter 5 covers point estimation in nonparametric settings; Chapter 6 focuses on hypothesis testing; and Chapter 7 discusses interval estimation and confidence sets. The classical frequentist approach is adopted in this book, although the Bayesian approach is also introduced (§2.3.2, §4.1, §6.4.4, and §7.1.3). Asymptotic (large sample) theory, a crucial part of statistical inference, is studied throughout the book, rather than in a separate chapter.

About 85% of the book covers classical results in statistical theory that are typically found in textbooks of a similar level. These materials are in the Statistics Department's Ph.D. qualifying examination syllabus. This part of the book is influenced by several standard textbooks, such as Casella and Berger (1990), Ferguson (1967), Lehmann (1983, 1986), and Rohatgi (1976). The other 15% of the book covers some topics in modern statistical theory

that have been developed in recent years, including robustness of the least squares estimators, Markov chain Monte Carlo, generalized linear models, quasi-likelihoods, empirical likelihoods, statistical functionals, generalized estimation equations, the jackknife, and the bootstrap.

In addition to the presentation of fruitful ideas and results, this book emphasizes the use of important tools in establishing theoretical results. Thus, most proofs of theorems, propositions, and lemmas are provided or left as exercises. Some proofs of theorems are omitted (especially in Chapter 1), because the proofs are lengthy or beyond the scope of the book (references are always provided). Each chapter contains a number of examples. Part of them are designed as materials covered in the discussion section of this course, which is typically taught by a teaching assistant (a senior graduate student). The exercises in each chapter form an important part of the book. They provide not only practice problems for students, but also many additional results as complementary materials to the main text.

Appendices A and B provide lists of frequently used abbreviations and notation, respectively. Definitions, examples, theorems, propositions, corollaries, and lemmas are numbered according to chapters, and their page numbers can be found in the subject index.

The book is essentially based on (1) my class notes taken in 1983-84 when I was a student in this course, (2) the notes I used when I was a teaching assistant for this course in 1984-85, and (3) the lecture notes I prepared during 1997-98 as the instructor of this course. I would like to express my thanks to Dennis Cox, who taught this course when I was a student and a teaching assistant, and undoubtfully has influence on my teaching style and textbook for this course. I am also very grateful to students in my class who provided helpful comments; to Mr. Yonghee Lee, who helped me to prepare all the figures in this book; to Springer-Verlag Production and Copy Editors who helped to improve the presentation; and to my family members who provided support during the writing of this book.

Madison, Wisconsin Jun Shao
January 1999

Contents

Preface **vii**

Chapter 1. Probability Theory **1**

1.1 Probability Spaces and Random Elements 1

1.1.1 σ-fields and measures 1

1.1.2 Measurable functions and distributions 6

1.2 Integration and Differentiation 9

1.2.1 Integration 9

1.2.2 Radon-Nikodym derivative 14

1.3 Distributions and Their Characteristics 17

1.3.1 Useful probability densities 17

1.3.2 Moments and generating functions 25

1.4 Conditional Expectations 30

1.4.1 Conditional expectations 30

1.4.2 Independence 34

1.4.3 Conditional distributions 36

1.5 Asymptotic Theorems 38

1.5.1 Convergence modes and stochastic orders 38

1.5.2 Convergence of transformations 42

1.5.3 The law of large numbers 45

1.5.4 The central limit theorem 47

1.6 Exercises 49

Chapter 2. Fundamentals of Statistics **61**

2.1 Populations, Samples, and Models 61

2.1.1 Populations and samples 61

2.1.2 Parametric and nonparametric models 64
2.1.3 Exponential and location-scale families 66
2.2 Statistics and Sufficiency 70
2.2.1 Statistics and their distributions 70
2.2.2 Sufficiency and minimal sufficiency 73
2.2.3 Complete statistics 79
2.3 Statistical Decision Theory 83
2.3.1 Decision rules, loss functions, and risks 83
2.3.2 Admissibility and optimality 86
2.4 Statistical Inference 92
2.4.1 Point estimators 92
2.4.2 Hypothesis tests 95
2.4.3 Confidence sets 99
2.5 Asymptotic Criteria and Inference 101
2.5.1 Consistency 102
2.5.2 Asymptotic bias, variance, and mse 105
2.5.3 Asymptotic inference 109
2.6 Exercises 112

Chapter 3. Unbiased Estimation 127
3.1 The UMVUE 127
3.1.1 Sufficient and complete statistics 128
3.1.2 A necessary and sufficient condition 132
3.1.3 Information inequality 135
3.1.4 Asymptotic properties of UMVUE's 138
3.2 U-Statistics 140
3.2.1 Some examples 140
3.2.2 Variances of U-statistics 142
3.2.3 The projection method 144
3.3 The LSE in Linear Models 148
3.3.1 The LSE and estimability 148
3.3.2 The UMVUE and BLUE 152
3.3.3 Robustness of LSE's 155
3.3.4 Asymptotic properties of LSE's 159
3.4 Unbiased Estimators in Survey Problems 161

3.4.1 UMVUE's of population totals 161
3.4.2 Horvitz-Thompson estimators 165
3.5 Asymptotically Unbiased Estimators 170
3.5.1 Functions of unbiased estimators 170
3.5.2 The method of moments 173
3.5.3 V-statistics 176
3.5.4 The weighted LSE 179
3.6 Exercises 182

Chapter 4. Estimation in Parametric Models 193
4.1 Bayes Decisions and Estimators 193
4.1.1 Bayes actions 193
4.1.2 Empirical and hierarchical Bayes methods 198
4.1.3 Bayes rules and estimators 201
4.1.4 Markov chain Monte Carlo 207
4.2 Invariance 213
4.2.1 One-parameter location families 213
4.2.2 One-parameter scale families 217
4.2.3 General location-scale families 219
4.3 Minimaxity and Admissibility 223
4.3.1 Estimators with constant risks 223
4.3.2 Results in one-parameter exponential families 227
4.3.3 Simultaneous estimation and shrinkage estimators . . 229
4.4 The Method of Maximum Likelihood 235
4.4.1 The likelihood function and MLE's 235
4.4.2 MLE's in generalized linear models 241
4.4.3 Quasi-likelihoods and conditional likelihoods 245
4.5 Asymptotically Efficient Estimation 248
4.5.1 Asymptotic optimality 248
4.5.2 Asymptotic efficiency of MLE's and RLE's 252
4.5.3 Other asymptotically efficient estimators 257
4.6 Exercises 261

Chapter 5. Estimation in Nonparametric Models 277
5.1 Distribution Estimators 277
5.1.1 Empirical c.d.f.'s in i.i.d. cases 278

5.1.2 Empirical likelihoods 281
5.1.3 Density estimation 288
5.2 Statistical Functionals 291
5.2.1 Differentiability and asymptotic normality 291
5.2.2 L-, M-, R-estimators and rank statistics 296
5.3 Linear Functions of Order Statistics 304
5.3.1 Sample quantiles 304
5.3.2 Robustness and efficiency 308
5.3.3 L-estimators in linear models 311
5.4 Generalized Estimating Equations 312
5.4.1 The GEE method and its relationship with others . . 313
5.4.2 Consistency of GEE estimators 317
5.4.3 Asymptotic normality of GEE estimators 321
5.5 Variance Estimation 325
5.5.1 The substitution method 326
5.5.2 The jackknife 329
5.5.3 The bootstrap 334
5.6 Exercises 337

Chapter 6. Hypothesis Tests 345
6.1 UMP Tests 345
6.1.1 The Neyman-Pearson lemma 346
6.1.2 Monotone likelihood ratio 349
6.1.3 UMP tests for two-sided hypotheses 353
6.2 UMP Unbiased Tests 356
6.2.1 Unbiasedness and similarity 356
6.2.2 UMPU tests in exponential families 358
6.2.3 UMPU tests in normal families 362
6.3 UMP Invariant Tests 369
6.3.1 Invariance and UMPI tests 369
6.3.2 UMPI tests in normal linear models 374
6.4 Tests in Parametric Models 380
6.4.1 Likelihood ratio tests 380
6.4.2 Asymptotic tests based on likelihoods 383
6.4.3 χ^2-tests 387

6.4.4 Bayes tests 392
6.5 Tests in Nonparametric Models 394
6.5.1 Sign, permutation, and rank tests 394
6.5.2 Kolmogorov-Smirnov and Cramér-von Mises tests . . 398
6.5.3 Empirical likelihood ratio tests 401
6.5.4 Asymptotic tests 404
6.6 Exercises 406

Chapter 7. Confidence Sets 421

7.1 Construction of Confidence Sets 421
7.1.1 Pivotal quantities 421
7.1.2 Inverting acceptance regions of tests 427
7.1.3 The Bayesian approach 430
7.1.4 Prediction sets 432
7.2 Properties of Confidence Sets 434
7.2.1 Lengths of confidence intervals 434
7.2.2 UMA and UMAU confidence sets 438
7.2.3 Randomized confidence sets 441
7.2.4 Invariant confidence sets 443
7.3 Asymptotic Confidence Sets 445
7.3.1 Asymptotically pivotal quantities 445
7.3.2 Confidence sets based on likelihoods 447
7.3.3 Results for quantiles 451
7.4 Bootstrap Confidence Sets 453
7.4.1 Construction of bootstrap confidence intervals 453
7.4.2 Asymptotic correctness and accuracy 457
7.4.3 High-order accurate bootstrap confidence sets 463
7.5 Simultaneous Confidence Intervals 467
7.5.1 Bonferroni's method 468
7.5.2 Scheffé's method in linear models 469
7.5.3 Tukey's method in one-way ANOVA models 471
7.5.4 Confidence bands for c.d.f.'s 473
7.6 Exercises 475

Appendix A. Abbreviations **489**

Appendix B. Notation **491**

References **493**

Author Index **505**

Subject Index **509**

Chapter 1

Probability Theory

Mathematical statistics relies on probability theory, which in turn is based on measure theory. The present chapter provides some principal concepts and notational conventions of probability theory, and some important results that are essential tools used in this book. A more complete account of probability theory can be found in many standard textbooks, for example, Billingsley (1986) and Chung (1974). The reader is assumed to be familiar with set operations and set functions (mappings) in advanced calculus.

1.1 Probability Spaces and Random Elements

In an elementary probability course, one defines a *random experiment* to be an experiment for which the outcome of the experiment cannot be predicted with certainty, and the probability of A (a collection of possible outcomes) to be the fraction of times that the outcome of the random experiment results in A in a large number of trials of the random experiment. A rigorous and logically consistent definition of probability was given by A. N. Kolmogorov in his measure-theoretic fundamental development of probability theory in 1933.

1.1.1 σ-fields and measures

Let Ω be a set of elements of interest. For example, Ω can be a set of numbers, a subinterval of the real line, or all possible outcomes of a random experiment. In probability theory, Ω is often called the outcome space, whereas in statistical theory, Ω is called the *sample space*. This is because in probability and statistics, Ω is usually the set of all possible outcomes of a random experiment under study.

A *measure* is a natural mathematical extension of the length, area, or volume of subsets in one-, two-, or three-dimensional Euclidean space. In a given sample space Ω, a measure is a set function defined for certain subsets of Ω. It will be necessary for this collection of subsets to satisfy certain properties, which are given in the following definition.

Definition 1.1. Let $\mathcal{F}$ be a collection of subsets of a sample space Ω. $\mathcal{F}$ is called a σ-field (or σ-algebra) if and only if it has the following properties.
(i) The empty set $\emptyset \in \mathcal{F}$.
(ii) If $A \in \mathcal{F}$, then the complement $A^c \in \mathcal{F}$.
(iii) If $A_i \in \mathcal{F}$, $i = 1, 2, ...$, then their union $\cup A_i \in \mathcal{F}$. ∎

A pair $(\Omega, \mathcal{F})$ consisting of a set Ω and a σ-field $\mathcal{F}$ of subsets of Ω is called a *measurable space*. The elements of $\mathcal{F}$ are called measurable sets in measure theory or *events* in probability and statistics.

Since $\emptyset^c = \Omega$, it follows from (i) and (ii) in Definition 1.1 that $\Omega \in \mathcal{F}$ if $\mathcal{F}$ is a σ-field on Ω. Also, it follows from (ii) and (iii) that if $A_i \in \mathcal{F}$, $i = 1, 2, ...$, and $\mathcal{F}$ is a σ-field, then the intersection $\cap A_i \in \mathcal{F}$. This can be shown using DeMorgan's law: $(\cap A_i)^c = \cup A_i^c$.

For any given Ω, there are two trivial σ-fields. The first one is the collection containing exactly two elements, $\emptyset$ and Ω. This is the smallest possible σ-field on Ω. The second one is the collection of all subsets of Ω, which is called the power set and is the largest σ-field on Ω.

Let us now consider some nontrivial σ-fields. Let A be a nonempty proper subset of Ω ($A \subset \Omega$, $A \neq \Omega$). Then (verify)

$$\{\emptyset, A, A^c, \Omega\} \tag{1.1}$$

is a σ-field. In fact, this is the smallest σ-field containing A in the sense that if $\mathcal{F}$ is any σ-field containing A, then the σ-field in (1.1) is a subcollection of $\mathcal{F}$. In general, the smallest σ-field containing $\mathcal{C}$, a collection of subsets of Ω, is denoted by $\sigma(\mathcal{C})$ and is called the σ-field generated by $\mathcal{C}$. Hence, the σ-field in (1.1) is $\sigma(\{A\})$. Note that $\sigma(\{A, A^c\})$, $\sigma(\{A, \Omega\})$, and $\sigma(\{A, \emptyset\})$ are all the same as $\sigma(\{A\})$. Of course, if $\mathcal{C}$ itself is a σ-field, then $\sigma(\mathcal{C}) = \mathcal{C}$.

On the real line $\mathcal{R}$, there is a special σ-field that will be used almost exclusively. Let $\mathcal{C}$ be the collection of all finite open intervals on $\mathcal{R}$. Then $\mathcal{B} = \sigma(\mathcal{C})$ is called the *Borel σ-field*. The elements of $\mathcal{B}$ are called *Borel sets*. The Borel σ-field $\mathcal{B}^k$ on the k-dimensional Euclidean space $\mathcal{R}^k$ can be similarly defined. It can be shown that all intervals (finite or infinite), open sets, and closed sets are Borel sets. To illustrate, we now show that on the real line, $\mathcal{B} = \sigma(\mathcal{O})$, where $\mathcal{O}$ is the collection of all open sets. Typically, one needs to show that $\sigma(\mathcal{C}) \subset \sigma(\mathcal{O})$ and $\sigma(\mathcal{O}) \subset \sigma(\mathcal{C})$. Since an open interval is an open set, $\mathcal{C} \subset \mathcal{O}$ and, hence, $\sigma(\mathcal{C}) \subset \sigma(\mathcal{O})$ (see Exercise 3 in §1.6). Let U be an open set. Then U can be expressed as a union

of a sequence of finite open intervals (see Royden (1968, p.39)). Hence, $U \in \sigma(\mathcal{C})$ (Definition 1.1(iii)) and $\mathcal{O} \subset \sigma(\mathcal{C})$. By the definition of $\sigma(\mathcal{O})$, $\sigma(\mathcal{O}) \subset \sigma(\mathcal{C})$. This completes the proof.

Let $C \subset \mathcal{R}^k$ be a Borel set and let $\mathcal{B}_C = \{C \cap B : B \in \mathcal{B}^k\}$. Then $(C, \mathcal{B}_C)$ is a measurable space and $\mathcal{B}_C$ is called the Borel σ-field on C.

Now we can introduce the notion of a measure.

Definition 1.2. Let $(\Omega, \mathcal{F})$ be a measurable space. A set function ν defined on $\mathcal{F}$ is called a *measure* if and only if it has the following properties.
(i) $0 \le \nu(A) \le \infty$ for any $A \in \mathcal{F}$.
(ii) $\nu(\emptyset) = 0$.
(iii) If $A_i \in \mathcal{F}$, $i = 1, 2, ...$, and A_i's are disjoint, i.e., $A_i \cap A_j = \emptyset$ for any $i \ne j$, then

$$\nu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \nu(A_i). \quad \blacksquare$$

The triple $(\Omega, \mathcal{F}, \nu)$ is called a *measure space*. If $\nu(\Omega) = 1$, then ν is called a *probability measure* and we usually denote it by P instead of ν, in which case $(\Omega, \mathcal{F}, P)$ is called a *probability space*.

Although measure is an extension of length, area, or volume, sometimes it can be quite abstract. For example, the following set function is a measure:

$$\nu(A) = \begin{cases} \infty & A \in \mathcal{F}, A \ne \emptyset \\ 0 & A = \emptyset. \end{cases} \tag{1.2}$$

Since a measure can take ∞ as its value, we must know how to do arithmetic with ∞. In this book, it suffices to know that (1) for any $x \in \mathcal{R}$, $\infty + x = \infty$, $x\,\infty = \infty$ if $x > 0$, $x\,\infty = -\infty$ if $x < 0$, and $0\,\infty = 0$; (2) $\infty + \infty = \infty$; and (3) $\infty\,\infty = \infty$. However, $\infty - \infty$ or ∞/∞ is not defined.

The following examples provide two very important measures in probability and statistics.

Example 1.1 (Counting measure). Let Ω be a sample space, $\mathcal{F}$ the collection of all subsets, and $\nu(A)$ the number of elements in $A \in \mathcal{F}$ ($\nu(A) = \infty$ if A contains infinitely many elements). Then ν is a measure on $\mathcal{F}$ and is called the *counting measure*. $\blacksquare$

Example 1.2 (Lebesgue measure). There is a unique measure m on $(\mathcal{R}, \mathcal{B})$ that satisfies

$$m([a, b]) = b - a \tag{1.3}$$

for every finite interval $[a, b]$, $-\infty < a \le b < \infty$. This is called the *Lebesgue measure*. If we restrict m to the measurable space $([0, 1], \mathcal{B}_{[0,1]})$, then m is a probability measure. $\blacksquare$

If Ω is *countable* in the sense that there is a one-to-one correspondence between Ω and the set of all integers, then one can usually consider the trivial σ-field that contains all subsets of Ω and a measure that assigns a value to every subset of Ω. When Ω is uncountable (e.g., $\Omega = \mathcal{R}$ or $[0,1]$), it is not possible to define a reasonable measure for every subset of Ω; for example, it is not possible to find a measure on all subsets of $\mathcal{R}$ and still satisfy property (1.3). This is why it is necessary to introduce σ-fields that are smaller than the power set.

The following result provides some basic properties of measures. Whenever we consider $\nu(A)$, it is implicitly assumed that $A \in \mathcal{F}$.

Proposition 1.1. Let $(\Omega, \mathcal{F}, \nu)$ be a measure space.
(i) (Monotonicity). If $A \subset B$, then $\nu(A) \leq \nu(B)$.
(ii) (Subadditivity). For any sequence $A_1, A_2, ...,$

$$\nu\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \nu(A_i).$$

(iii) (Continuity). If $A_1 \subset A_2 \subset A_3 \subset \cdots$ (or $A_1 \supset A_2 \supset A_3 \supset \cdots$ and $\nu(A_1) < \infty$), then

$$\nu\left(\lim_{n\to\infty} A_n\right) = \lim_{n\to\infty} \nu\left(A_n\right),$$

where

$$\lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i \quad \left(\text{ or } = \bigcap_{i=1}^{\infty} A_i\right).$$

Proof. We prove (i) only. The proofs of (ii) and (iii) are left as exercises. Since $A \subset B$, $B = A \cup (A^c \cap B)$ and A and $A^c \cap B$ are disjoint. By Definition 1.2(iii), $\nu(B) = \nu(A) + \nu(A^c \cap B)$, which is no smaller than $\nu(A)$ since $\nu(A^c \cap B) \geq 0$ by Definition 1.2(i). ∎

There is a one-to-one correspondence between the set of all probability measures on $(\mathcal{R}, \mathcal{B})$ and a set of functions on $\mathcal{R}$. Let P be a probability measure. The *cumulative distribution function* (c.d.f.) of P is defined to be

$$F(x) = P\left((-\infty, x]\right), \quad x \in \mathcal{R}. \tag{1.4}$$

Proposition 1.2. (i) Let F be a c.d.f. on $\mathcal{R}$. Then
(a) $F(-\infty) = \lim_{x\to-\infty} F(x) = 0$;
(b) $F(\infty) = \lim_{x\to\infty} F(x) = 1$;
(c) F is nondecreasing, i.e., $F(x) \leq F(y)$ if $x \leq y$;
(d) F is right continuous, i.e., $\lim_{y\to x, y>x} F(y) = F(x)$.
(ii) Suppose that a real-valued function F on $\mathcal{R}$ satisfies (a)-(d) in part (i). Then F is the c.d.f. of a unique probability measure on $(\mathcal{R}, \mathcal{B})$. ∎

The *Cartesian product* of any k sets $A_1, ..., A_k$ (which may be subsets of different sample spaces) is defined as the set of all $(a_1, ..., a_k)$, $a_i \in A_i$, and is denoted by $A_1 \times \cdots \times A_k$. Let $(\Omega_i, \mathcal{F}_i, \nu_i)$, $i = 1, ..., k$, be k measure spaces. We now introduce a convenient way of constructing a σ-field and a measure on the product space $\Omega_1 \times \cdots \times \Omega_k$.

First, note that $\mathcal{F}_1 \times \cdots \times \mathcal{F}_k$ is not necessarily a σ-field. We define the σ-field $\sigma(\mathcal{F}_1 \times \cdots \times \mathcal{F}_k)$ as the *product* σ-field on $\Omega_1 \times \cdots \times \Omega_k$. As an example, consider $(\Omega_i, \mathcal{F}_i) = (\mathcal{R}, \mathcal{B})$ for all i. Then the product space is $\mathcal{R}^k$ and it can be shown that the product σ-field is the same as the Borel σ-field on $\mathcal{R}^k$, which is the σ-field generated by $\mathcal{O}$, all open sets in $\mathcal{R}^k$.

In Example 1.2, the usual length of an interval $[a, b] \subset \mathcal{R}$ is the same as the Lebesgue measure of $[a, b]$. Consider a rectangle $[a_1, b_1] \times [a_2, b_2] \subset \mathcal{R}^2$. The usual area of $[a_1, b_1] \times [a_2, b_2]$ is

$$(b_1 - a_1)(b_2 - a_2) = m([a_1, b_1])m([a_2, b_2]), \tag{1.5}$$

i.e., the product of the Lebesgue measures of two intervals $[a_1, b_1]$ and $[a_2, b_2]$. Note that $[a_1, b_1] \times [a_2, b_2]$ is a measurable set by the definition of the product σ-field. Is $m([a_1, b_1])m([a_2, b_2])$ the same as the value of a measure defined on the product σ-field? To answer this, we need the following technical definition. A measure ν on $(\Omega, \mathcal{F})$ is said to be *σ-finite* if there exists a sequence $\{A_1, A_2, ...\}$ such that $\cup A_i = \Omega$ and $\nu(A_i) < \infty$ for all i. Any finite measure (such as a probability measure) is clearly σ-finite. The Lebesgue measure in Example 1.2 is σ-finite, since $\mathcal{R} = \cup A_n$ with $A_n = (-n, n)$, $n = 1, 2, ...$ The counting measure in Example 1.1 is σ-finite if and only if Ω is countable. The measure defined by (1.2), however, is not σ-finite.

Proposition 1.3 (Product measure theorem). Let $(\Omega_i, \mathcal{F}_i, \nu_i)$, $i = 1, ..., k$, be measure spaces and ν_i be σ-finite measures. Then there exists a unique σ-finite measure on the product σ-field $\sigma(\mathcal{F}_1 \times \cdots \times \mathcal{F}_k)$, called the *product measure* and denoted by $\nu_1 \times \cdots \times \nu_k$, such that

$$\nu_1 \times \cdots \times \nu_k(A_1 \times \cdots \times A_k) = \nu_1(A_1) \cdots \nu_k(A_k)$$

for all $A_i \in \mathcal{F}_i$, $i = 1, ..., k$. ∎

Thus, in $\mathcal{R}^2$ there is a unique measure, the product measure $m \times m$, for which $m \times m([a_1, b_1] \times [a_2, b_2])$ is equal to the value given by (1.5). This measure is called the Lebesgue measure on $(\mathcal{R}^2, \mathcal{B}^2)$. Similarly, we can define the Lebesgue measure on $(\mathcal{R}^3, \mathcal{B}^3)$, which exactly equals the usual volume for subsets of the form $[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$.

In general, the product measure space generated by $(\Omega_i, \mathcal{F}_i, \nu_i)$, $i = 1, ..., k$, is denoted by $\prod_{i=1}^k (\Omega_i, \mathcal{F}_i, \nu_i)$.

The concept of c.d.f. can be extended to $\mathcal{R}^k$. Let P be a probability measure on $(\mathcal{R}^k, \mathcal{B}^k)$. The c.d.f. (or *joint* c.d.f.) of P is defined by

$$F(x_1, ..., x_k) = P\left((-\infty, x_1] \times \cdots \times (-\infty, x_k]\right), \quad x_i \in \mathcal{R}. \tag{1.6}$$

Note that P is not necessarily a product measure. Again, there is a one-to-one correspondence between probability measures and joint c.d.f.'s on $\mathcal{R}^k$. Some properties of a joint c.d.f. are given in Exercise 10 in §1.6.

1.1.2 Measurable functions and distributions

Since Ω can be quite arbitrary, it is often convenient to consider a function (mapping) f from Ω to a simpler space Λ (often $\Lambda = \mathcal{R}^k$, the k-dimensional Euclidean space). Let $B \subset \Lambda$. Then the *inverse image* of B under f is

$$f^{-1}(B) = \{f \in B\} = \{\omega \in \Omega : f(\omega) \in B\}.$$

The inverse function f^{-1} need not exist for $f^{-1}(B)$ to be defined. The reader is asked to verify the following properties:

(a) $f^{-1}(B^c) = (f^{-1}(B))^c$ for any $B \subset \Lambda$;

(b) $f^{-1}(\cup B_i) = \cup f^{-1}(B_i)$ for any $B_i \subset \Lambda$, $i = 1, 2, ...$

Let $\mathcal{C}$ be a collection of subsets of Λ. We define

$$f^{-1}(\mathcal{C}) = \{f^{-1}(C) : C \in \mathcal{C}\}.$$

Definition 1.3. Let $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ be measurable spaces and f a function from Ω to Λ. The function f is called a *measurable function* from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ if and only if $f^{-1}(\mathcal{G}) \subset \mathcal{F}$. ∎

If $\Lambda = \mathcal{R}$ and $\mathcal{G} = \mathcal{B}$ (Borel σ-field), then f is said to be *Borel measurable* or is called a *Borel function*.

In probability theory, a measurable function is called a *random element* and denoted by one of X, Y, Z,.... If X is measurable from $(\Omega, \mathcal{F})$ to $(\mathcal{R}, \mathcal{B})$, then it is called a *random variable*; if X is measurable from $(\Omega, \mathcal{F})$ to $(\mathcal{R}^k, \mathcal{B}^k)$, then it is called a *random k-vector* (as a notational convention in this book, any vector is considered to be a row vector). If $X_1, ..., X_k$ are random variables defined on a common probability space, then the vector $(X_1, ..., X_k)$ is a random k-vector.

If f is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$, then $f^{-1}(\mathcal{G})$ is a sub-σ-field of $\mathcal{F}$ (verify). It is called the σ-field generated by f and is denoted by $\sigma(f)$.

Now we consider some examples of measurable functions. If $\mathcal{F}$ is the collection of all subsets of Ω, then any function f is measurable. Let $A \subset \Omega$. The *indicator function* for A is defined as

$$I_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A. \end{cases}$$

For any $B \subset \mathcal{R}$,

$$I_A^{-1}(B) = \begin{cases} \emptyset & 0 \notin B, 1 \notin B \\ A & 0 \notin B, 1 \in B \\ A^c & 0 \in B, 1 \notin B \\ \Omega & 0 \in B, 1 \in B. \end{cases}$$

Then $\sigma(I_A)$ is the σ-field given in (1.1). If A is a measurable set, then I_A is a Borel function.

Note that $\sigma(I_A)$ is a much smaller σ-field than the original σ-field $\mathcal{F}$. This is another reason why we introduce the concept of measurable functions and random variables, in addition to the reason that it is easy to deal with numbers. Often the σ-field $\mathcal{F}$ (such as the power set) contains too many subsets and we are only interested in some of them. One can then define a random variable X with $\sigma(X)$ containing subsets that are of interest. In general, $\sigma(X)$ is between the trivial σ-field $\{\emptyset, \Omega\}$ and $\mathcal{F}$, and contains more subsets if X is more complicated. For the simplest function I_A, we have shown that $\sigma(I_A)$ contains only four elements.

The class of *simple functions* is obtained by taking linear combinations of indicators of measurable sets, i.e.,

$$\varphi(\omega) = \sum_{i=1}^{k} a_i I_{A_i}(\omega), \tag{1.7}$$

where $A_1, ..., A_k$ are measurable sets on Ω and $a_1, ..., a_k$ are real numbers. One can show directly that such a function is a Borel function, but it follows immediately from Proposition 1.4. Let $A_1, ..., A_k$ be a partition of Ω, i.e., A_i's are disjoint and $A_1 \cup \cdots \cup A_k = \Omega$. Then the simple function φ given by (1.7) with distinct a_i's exactly characterizes this partition and $\sigma(\varphi) = \sigma(\{A_1, ..., A_k\})$.

Proposition 1.4. Let $(\Omega, \mathcal{F})$ be a measurable space.
(i) If f and g are Borel, then so are fg and $af + bg$, where a and b are real numbers; also, f/g is Borel provided $g(\omega) \neq 0$ for any $\omega \in \Omega$.
(ii) f is Borel if and only if $f^{-1}(a, \infty) \in \mathcal{F}$ for all $a \in \mathcal{R}$.
(iii) If $f_1, f_2, ...$ are Borel, then so are $\sup_n f_n$, $\inf_n f_n$, $\limsup_n f_n$, and $\liminf_n f_n$. Furthermore, the set

$$A = \left\{\omega \in \Omega : \lim_{n\to\infty} f_n(\omega) \text{ exists}\right\}$$

is an event and the function

$$h(\omega) = \begin{cases} \lim_{n\to\infty} f_n(\omega) & \omega \in A \\ f_1(\omega) & \omega \notin A \end{cases}$$

is Borel.
(iv) Suppose that f is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ and g is measurable from $(\Lambda, \mathcal{G})$ to $(\Delta, \mathcal{H})$. Then the composite function $g \circ f$ is measurable from $(\Omega, \mathcal{F})$ to $(\Delta, \mathcal{H})$.
(v) Let Ω be a Borel set in $\mathcal{R}^p$. If f is a continuous function from Ω to $\mathcal{R}^q$, then f is measurable. ∎

Proposition 1.4 indicates that there are many Borel functions. In fact, it is hard to find a non-Borel function.

Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and f be a measurable function from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$. The *induced measure* by f, denoted by $\nu \circ f^{-1}$, is a measure on $\mathcal{G}$ defined as

$$\nu \circ f^{-1}(B) = \nu(f \in B) = \nu\left(f^{-1}(B)\right), \quad B \in \mathcal{G}. \tag{1.8}$$

It is usually easier to deal with $\nu \circ f^{-1}$ than to deal with ν since $(\Lambda, \mathcal{G})$ is usually simpler than $(\Omega, \mathcal{F})$. Furthermore, subsets not in $\sigma(f)$ are not involved in the definition of $\nu \circ f^{-1}$. As we discussed earlier, in some cases we are only interested in subsets in $\sigma(f)$.

If $\nu = P$ is a probability measure and X is a random variable or a random vector, then $P \circ X^{-1}$ is called the *law* or the *distribution* of X and is denoted by P_X. The c.d.f. of P_X defined by (1.4) or (1.6) is also called the c.d.f. or joint c.d.f. of X and is denoted by F_X. On the other hand, for any c.d.f. or joint c.d.f. F, there exists at least one random variable or vector (usually there are many) defined on some probability space for which $F_X = F$. The following are some examples of random variables and their c.d.f.'s. More examples can be found in §1.3.1.

Example 1.3 (Discrete c.d.f.'s). Let $a_1 < a_2 < \cdots$ be a sequence of real numbers and let p_n, $n = 1, 2, ...$, be a sequence of positive numbers such that $\sum_{n=1}^{\infty} p_n = 1$. Define

$$F(x) = \begin{cases} \sum_{i=1}^{n} p_i & a_n \le x < a_{n+1}, \quad n = 1, 2, ... \\ 0 & -\infty < x < a_1. \end{cases} \tag{1.9}$$

Then F is a *stepwise* c.d.f. It has a jump of size p_n at each a_n and is flat between a_n and a_{n+1}, $n = 1, 2, ...$ Such a c.d.f. is called a *discrete* c.d.f. and the corresponding random variable is called a *discrete random variable*. We can easily obtain a random variable having F in (1.9) as its c.d.f. For example, let $\Omega = \{a_1, a_2, ...\}$, $\mathcal{F}$ be the collection of all subsets of Ω,

$$P(A) = \sum_{i: a_i \in A} p_i, \quad A \in \mathcal{F}, \tag{1.10}$$

and $X(\omega) = \omega$. One can show that P is a probability measure and the c.d.f. of X is F in (1.9). ∎

Example 1.4 (Continuous c.d.f.'s). Opposite to the class of discrete c.d.f.'s is the class of continuous c.d.f.'s. Without the concepts of integration and differentiation introduced in the next section, we can only provide a few examples of continuous c.d.f.'s. One such example is the *uniform* c.d.f. on the interval $[a, b]$ defined as

$$F(x) = \begin{cases} 0 & -\infty < x < a \\ \frac{x-a}{b-a} & a \le x < b \\ 1 & b \le x < \infty. \end{cases}$$

Another example is the *exponential* c.d.f. defined as

$$F(x) = \begin{cases} 0 & -\infty < x < 0 \\ 1 - e^{-x/\theta} & 0 \le x < \infty, \end{cases}$$

where θ is a fixed positive constant. Note that both uniform and exponential c.d.f.'s are continuous functions. ∎

1.2 Integration and Differentiation

Differentiation and integration are two of the main components of calculus. This is also true in measure theory or probability theory, except that integration is introduced first whereas in calculus, differentiation is introduced first.

1.2.1 Integration

An important concept needed in probability and statistics is the *integration* of Borel functions with respect to (w.r.t.) a measure ν, which is a type of "average". The definition proceeds in several steps. First, we define the integral of a nonnegative simple function, i.e., a simple function φ given by (1.7) with $a_i \ge 0$, $i = 1, ..., k$.

Definition 1.4(a). The integral of a nonnegative simple function φ given by (1.7) w.r.t. ν is defined as

$$\int \varphi d\nu = \sum_{i=1}^{k} a_i \nu(A_i). \quad ∎ \tag{1.11}$$

The right-hand side of (1.11) is a weighted average of a_i's with $\nu(A_i)$'s as weights. Since $a\infty = \infty$ if $a > 0$ and $a\infty = 0$ if $a = 0$, the right-hand side of (1.11) is always well defined, although $\int \varphi d\nu = \infty$ is possible. Note

that different a_i's and A_i's may produce the same function φ; for example, with $\Omega = \mathcal{R}$,

$$2I_{(0,1)}(x) + I_{[1,2]}(x) = I_{(0,2]}(x) + I_{(0,1)}(x).$$

However, one can show that different representations of φ in (1.7) produce the same value for $\int \varphi d\nu$ so that the integral of a nonnegative simple function is well defined.

Next, we consider nonnegative Borel function f.

Definition 1.4(b). Let f be a nonnegative Borel function and let $\mathcal{S}_f$ be the collection of all nonnegative simple functions of the form (1.7) satisfying $\varphi(\omega) \leq f(\omega)$ for any $\omega \in \Omega$. The integral of f w.r.t. ν is defined as

$$\int f d\nu = \sup\left\{\int \varphi d\nu : \ \varphi \in \mathcal{S}_f\right\}. \quad \blacksquare$$

Hence, for any Borel function $f \geq 0$, there exists a sequence of simple functions $\varphi_1, \varphi_2, ...$ such that $0 \leq \varphi_i \leq f$ for all i and $\lim_{n\to\infty} \int \varphi_n d\nu = \int f d\nu$.

Finally, for a Borel function f, we first define the positive part of f by

$$f_+(\omega) = \max\{f(\omega), 0\}$$

and the negative part of f by

$$f_-(\omega) = \max\{-f(\omega), 0\}.$$

Note that f_+ and f_- are nonnegative Borel functions, $f(\omega) = f_+(\omega) - f_-(\omega)$, and $|f(\omega)| = f_+(\omega) + f_-(\omega)$.

Definition 1.4(c). Let f be a Borel function. We say $\int f d\nu$ exists if and only if at least one of $\int f_+ d\nu$ and $\int f_- d\nu$ is finite, in which case

$$\int f d\nu = \int f_+ d\nu - \int f_- d\nu. \tag{1.12}$$

Let A be a measurable set and I_A be its indicator function. The integral of f over A is defined as

$$\int_A f d\nu = \int I_A f d\nu. \quad \blacksquare$$

Note that the left-hand side of (1.12) is always well defined, although it can be ∞ or $-\infty$. When both $\int f_+ d\nu$ and $\int f_- d\nu$ are finite, we say that f is

integrable (for a nonnegative Borel function f, f is integrable if $\int f d\nu < \infty$). Note that the existence of $\int f d\nu$ is different from the integrability of f.

The integral of f may be denoted differently whenever there is a need to indicate the variable(s) to be integrated and the integration domain; for example, $\int_\Omega f d\nu$, $\int f(\omega) d\nu$, $\int f(\omega) d\nu(\omega)$, or $\int f(\omega)\nu(d\omega)$, and so on. In probability and statistics, $\int X dP$ is usually written as EX or $E(X)$ and called the *expectation* or *expected value* of X. If F is the c.d.f. of P on $(\mathcal{R}^k, \mathcal{B}^k)$, $\int f(x) dP$ is also denoted by $\int f(x) dF(x)$ or $\int f dF$.

Example 1.5. Let Ω be a countable set, $\mathcal{F}$ be all subsets of Ω, and ν be the counting measure given in Example 1.1. For any Borel function f, the integral of f w.r.t. ν is

$$\int f d\nu = \sum_{\omega \in \Omega} f(\omega). \tag{1.13}$$

This is obvious if f is a simple function. The proof for general f is left as an exercise. ∎

Example 1.6. If $\Omega = \mathcal{R}$ and ν is the Lebesgue measure, then the integral of f over an interval $[a, b]$ agrees with the Riemann integral in calculus when the latter is well defined, and is usually written as $\int_{[a,b]} f(x) dx = \int_a^b f(x) dx$. However, there are functions for which the Lebesgue integrals are defined but not the Riemann integrals. ∎

We now introduce some properties of integrals. The proof of the following result is left to the reader.

Proposition 1.5 (Linearity of integrals). Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and f and g be Borel functions.
(i) If $\int f d\nu$ exists and $a \in \mathcal{R}$, then $\int (af) d\nu$ exists and is equal to $a \int f d\nu$.
(ii) If both $\int f d\nu$ and $\int g d\nu$ exist and $\int f d\nu + \int g d\nu$ is well defined, then $\int (f + g) d\nu$ exists and is equal to $\int f d\nu + \int g d\nu$. ∎

If a statement holds for all ω in $\Omega - \mathcal{N}$ with $\nu(\mathcal{N}) = 0$, then the statement is said to hold a.e. (almost every where) ν (or simply a.e. if the measure ν is clear from the context). If ν is a probability measure, then a.e. may be replaced by a.s. (almost surely).

Proposition 1.6. Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and f and g be Borel.
(i) If $f \le g$ a.e., then $\int f d\nu \le \int g d\nu$, provided that the integrals exist.
(ii) If $f \ge 0$ a.e. and $\int f d\nu = 0$, then $f = 0$ a.e. ∎

Some direct consequences of Proposition 1.6(i) are: $|\int f d\nu| \le \int |f| d\nu$; if $f \ge 0$ a.e., then $\int f d\nu \ge 0$; and if $f = g$ a.e., then $\int f d\nu = \int g d\nu$.

We now prove part (ii) of Proposition 1.6 as an illustration. The proof for part (i) is left to the reader. Let $A = \{f > 0\}$ and $A_n = \{f \geq n^{-1}\}$, $n = 1, 2, \ldots$. Then $A_n \subset A$ for any n and $\lim_{n\to\infty} A_n = A$ (why?). By Proposition 1.1(iii), $\lim_{n\to\infty} \nu(A_n) = \nu(A)$. Using Proposition 1.5 and part (i) of Proposition 1.6, we obtain that

$$n^{-1}\nu(A_n) = \int n^{-1} I_{A_n} d\nu \leq \int f I_{A_n} d\nu \leq \int f d\nu = 0$$

for any n. Hence $\nu(A) = 0$ and $f = 0$ a.e.

It is sometimes required to know whether the following interchange of two operations is valid:

$$\int \lim_{n\to\infty} f_n d\nu = \lim_{n\to\infty} \int f_n d\nu, \tag{1.14}$$

where $f_1, f_2, \ldots$ is a sequence of Borel functions. Note that we only require $\lim_{n\to\infty} f_n$ exists a.e. Also, the limit of a sequence of Borel functions is Borel (Proposition 1.4). The following example shows that (1.14) is not always true.

Example 1.7. Consider $(\mathcal{R}, \mathcal{B})$ and the Lebesgue measure. Define $f_n(x) = nI_{[0,n^{-1}]}(x)$, $n = 1, 2, \ldots$. Then $\lim_{n\to\infty} f_n(x) = 0$ for all x but $x = 0$. Since the Lebesgue measure of a single point set is 0 (see Example 1.2), $\lim_{n\to\infty} f_n(x) = 0$ a.e. and $\int \lim_{n\to\infty} f_n(x)dx = 0$. On the other hand, $\int f_n(x)dx = 1$ for any n and, hence, $\lim_{n\to\infty} \int f_n(x)dx = 1$. ∎

The following result gives some sufficient conditions under which (1.14) holds.

Theorem 1.1. Let $f_1, f_2, \ldots$ be a sequence of Borel functions.
(i) (Dominated convergence theorem). If $\lim_{n\to\infty} f_n = f$ a.e. and there exists an integrable function g such that $|f_n| \leq g$ a.e., then (1.14) holds.
(ii) (Fatou's lemma). If $f_n \geq 0$, then

$$\int \liminf_n f_n d\nu \leq \liminf_n \int f_n d\nu.$$

(iii) (Monotone convergence theorem). If $0 \leq f_1 \leq f_2 \leq \cdots$ and $\lim_{n\to\infty} f_n = f$ a.e., then (1.14) holds. ∎

The proof is omitted. However, it can be seen that if f in (iii) is integrable, then part (iii) is a consequence of part (i). The following is an application of Theorem 1.1.

Example 1.8 (Interchange of differentiation and integration). Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and for any fixed θ, $f(\omega, \theta)$ be a Borel function on

Ω. Suppose that $\partial f(\omega,\theta)/\partial\theta$ exists a.e. for $\theta \in (a,b) \subset \mathcal{R}$ and that $|\partial f(\omega,\theta)/\partial\theta| \le g(\omega)$ a.e., where g is an integrable function on Ω. Then, for each $\theta \in (a,b)$, $\partial f(\omega,\theta)/\partial\theta$ is integrable and

$$\frac{d}{d\theta}\int f(\omega,\theta)d\nu = \int \frac{\partial f(\omega,\theta)}{\partial\theta}d\nu. \quad \blacksquare$$

Theorem 1.2 (Change of variables). Let f be measurable from $(\Omega,\mathcal{F},\nu)$ to $(\Lambda,\mathcal{G})$ and g be Borel on $(\Lambda,\mathcal{G})$. Then

$$\int_\Omega g\circ f d\nu = \int_\Lambda g d(\nu\circ f^{-1}), \tag{1.15}$$

i.e., if either integral exists, then so does the other, and the two are the same. $\blacksquare$

The proof is again omitted. Note that integration domains are indicated on both sides of (1.15). This result extends the change of variable formula for Riemann integrals, i.e., $\int g(y)dy = \int g(f(x))f'(x)dx$, $y = f(x)$.

Result (1.15) is very important in probability and statistics. Let X be a random variable on a probability space $(\Omega,\mathcal{F},P)$. If $EX = \int_\Omega XdP$ exists, then usually it is much simpler to compute $EX = \int_\mathcal{R} xdP_X$, where $P_X = P\circ X^{-1}$ is the law of X. Let Y be a random vector from Ω to $\mathcal{R}^k$ and g be Borel from $\mathcal{R}^k$ to $\mathcal{R}$. According to (1.15), $Eg(Y)$ can be computed as $\int_{\mathcal{R}^k} g(y)dP_Y$ or $\int_\mathcal{R} xdP_{g(Y)}$, depending on which of P_Y and $P_{g(Y)}$ is easier to handle. As a more specific example, consider $k=2$, $Y=(X_1,X_2)$, and $g(Y) = X_1+X_2$. Using Proposition 1.5(ii), $E(X_1+X_2) = EX_1+EX_2$ and, hence,

$$E(X_1+X_2) = \int_\mathcal{R} xdP_{X_1} + \int_\mathcal{R} xdP_{X_2}.$$

Then we need to handle two integrals involving P_{X_1} and P_{X_2}. On the other hand,

$$E(X_1+X_2) = \int_\mathcal{R} xdP_{X_1+X_2},$$

which involves one integral w.r.t. $P_{X_1+X_2}$. Unless we have some knowledge about the joint c.d.f. of (X_1,X_2), it is not easy to handle $P_{X_1+X_2}$.

The following theorem states how to evaluate an integral w.r.t. a product measure via iterated integration.

Theorem 1.3 (Fubini's theorem). Let ν_i be a σ-finite measure on $(\Omega_i,\mathcal{F}_i)$, $i=1,2$, and let f be a Borel function on $\prod_{i=1}^2(\Omega_i,\mathcal{F}_i)$ whose integral w.r.t. $\nu_1\times\nu_2$ exists. Then

$$g(\omega_2) = \int_{\Omega_1} f(\omega_1,\omega_2)d\nu_1$$

exists a.e. ν_2 and defines a Borel function on Ω_2 whose integral w.r.t. ν_2 exists, and

$$\int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2) d\nu_1 \times \nu_2 = \int_{\Omega_2} \left[\int_{\Omega_1} f(\omega_1, \omega_2) d\nu_1 \right] d\nu_2. \quad \blacksquare$$

This result can be naturally extended to the integral w.r.t. the product measure on $\prod_{i=1}^{k} (\Omega_i, \mathcal{F}_i)$ for any finite positive integer k.

Example 1.9. Let $\Omega_1 = \Omega_2 = \{0, 1, 2, ...\}$, and $\nu_1 = \nu_2$ be the counting measure (Example 1.1). A function f on $\Omega_1 \times \Omega_2$ defines a double sequence. If $\int f d\nu_1 \times \nu_2$ exists, then

$$\int f d\nu_1 \times \nu_2 = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} f(i, j) = \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} f(i, j) \tag{1.16}$$

(by Theorem 1.3 and Example 1.5). Thus, a double series can be summed in either order, if it is well defined. $\blacksquare$

1.2.2 Radon-Nikodym derivative

Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and f be a nonnegative Borel function. One can show that the set function

$$\lambda(A) = \int_A f d\nu, \quad A \in \mathcal{F} \tag{1.17}$$

is a measure on $(\Omega, \mathcal{F})$. Note that

$$\nu(A) = 0 \quad \text{implies} \quad \lambda(A) = 0. \tag{1.18}$$

If (1.18) holds for two measures λ and ν defined on the same measurable space, then we say λ is *absolutely continuous* w.r.t. ν, and write $\lambda \ll \nu$.

Formula (1.17) gives us not only a way of constructing measures, but also a method of computing measures of measurable sets. Let ν be a well-known measure (such as the Lebesgue measure or the counting measure) and λ a relatively unknown measure. If we can find a function f such that (1.17) holds, then computing $\lambda(A)$ can be done through integration. A necessary condition for (1.17) is clearly $\lambda \ll \nu$. The following result shows that $\lambda \ll \nu$ is also almost sufficient for (1.17).

Theorem 1.4 (Radon-Nikodym theorem). Let ν and λ be two measures on $(\Omega, \mathcal{F})$ and ν be σ-finite. If $\lambda \ll \nu$, then there exists a nonnegative Borel

function f on Ω such that (1.17) holds. Furthermore, f is unique a.e. ν, i.e., if $\lambda(A) = \int_A g d\nu$ for any $A \in \mathcal{F}$, then $f = g$ a.e. ν. ∎

The proof of this theorem is beyond our scope. If (1.17) holds, then the function f is called the Radon-Nikodym *derivative* or *density* of λ w.r.t. ν, and is denoted by $d\lambda/d\nu$.

A useful consequence of Theorem 1.4 is that if f is Borel on $(\Omega, \mathcal{F})$ and $\int_A f d\nu = 0$ for any $A \in \mathcal{F}$, then $f = 0$ a.e.

If $\int f d\nu = 1$ for an $f \geq 0$ a.e. ν, then λ given by (1.17) is a probability measure and f is called its *probability density function* (p.d.f.) w.r.t. ν. For any probability measure P on $\mathcal{R}^k$ corresponding to a c.d.f. F or a random vector X, if P has a p.d.f. f w.r.t. a measure ν, then f is also called the p.d.f. of F or X w.r.t. ν.

Example 1.10 (p.d.f. of a discrete c.d.f.). Consider the discrete c.d.f. F in (1.9) of Example 1.3 with its probability measure given by (1.10). Let $\Omega = \{a_1, a_2, ...\}$ and ν be the counting measure on the power set of Ω. By Example 1.5,

$$P(A) = \int_A f d\nu = \sum_{a_i \in A} f(a_i), \quad A \subset \Omega, \tag{1.19}$$

where $f(a_i) = p_i$, $i = 1, 2, ...$ That is, f is the p.d.f. of P or F w.r.t. ν. Hence any discrete c.d.f. has a p.d.f. w.r.t. counting measure. A p.d.f. w.r.t. counting measure is called a *discrete* p.d.f. ∎

Example 1.11. Let F be a c.d.f. Assume that F is differentiable in the usual sense in calculus. Let f be the derivative of F. From calculus,

$$F(x) = \int_{-\infty}^{x} f(y) dy, \quad x \in \mathcal{R}. \tag{1.20}$$

Let P be the probability measure corresponding to F. It can be shown that $P(A) = \int_A f dm$ for any $A \in \mathcal{B}$, where m is the Lebesgue measure on $\mathcal{R}$. Hence, f is the p.d.f. of P or F w.r.t. Lebesgue measure. In this case, the Radon-Nikodym derivative is the same as the usual derivative of F in calculus.

A continuous c.d.f. may not have a p.d.f. w.r.t. Lebesgue measure. A necessary and sufficient condition for a c.d.f. F having a p.d.f. w.r.t. Lebesgue measure is that F is *absolute continuous* in the sense that for any $\epsilon > 0$, there exists $\delta > 0$ such that for each finite collection of disjoint bounded open intervals (a_i, b_i), $\sum(b_i - a_i) < \delta$ implies $\sum[F(b_i) - F(a_i)] < \epsilon$. Absolute continuity is weaker than differentiability, but is stronger than continuity. Thus, any discontinuous c.d.f. (such as a discrete c.d.f.) is not absolute continuous. Note that every c.d.f. is differentiable a.e. Lebesgue

measure (Chung, 1974, Chapter 1). Hence, if f is the p.d.f. of F w.r.t. Lebesgue measure, then $f =$ the usual derivative of F a.e. Lebesgue measure and (1.20) holds. In such a case probabilities can be computed through integration. It can be shown that the uniform and exponential c.d.f.'s in Example 1.4 are absolute continuous and their p.d.f.'s are, respectively,

$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x < b \\ 0 & \text{otherwise} \end{cases}$$

and

$$f(x) = \begin{cases} 0 & -\infty < x < 0 \\ \theta^{-1} e^{-x/\theta} & 0 \le x < \infty. \end{cases}$$

A p.d.f. w.r.t. Lebesgue measure is called a Lebesgue p.d.f. ∎

More examples of p.d.f.'s are given in §1.3.1.

The following result provides some basic properties of Radon-Nikodym derivatives.

Proposition 1.7 (Calculus with Radon-Nikodym derivatives). Let ν be a σ-finite measure on a measure space $(\Omega, \mathcal{F})$. All other measures discussed in (i)-(iii) are defined on $(\Omega, \mathcal{F})$.
(i) If λ is a measure, $\lambda \ll \nu$, and $f \ge 0$, then

$$\int f d\lambda = \int f \frac{d\lambda}{d\nu} d\nu.$$

(Notice how the $d\nu$'s "cancel" on the right-hand side.)
(ii) If λ_i, $i = 1, 2$, are measures and $\lambda_i \ll \nu$, then $\lambda_1 + \lambda_2 \ll \nu$ and

$$\frac{d(\lambda_1 + \lambda_2)}{d\nu} = \frac{d\lambda_1}{d\nu} + \frac{d\lambda_2}{d\nu} \quad \text{a.e. } \nu.$$

(iii) (Chain rule). If τ is a measure, λ is a σ-finite measure, and $\tau \ll \lambda \ll \nu$, then

$$\frac{d\tau}{d\nu} = \frac{d\tau}{d\lambda} \frac{d\lambda}{d\nu} \quad \text{a.e. } \nu.$$

In particular, if $\lambda \ll \nu$ and $\nu \ll \lambda$ (in which case λ and ν are *equivalent*), then

$$\frac{d\lambda}{d\nu} = \left(\frac{d\nu}{d\lambda} \right)^{-1} \quad \text{a.e. } \nu \text{ or } \lambda.$$

(iv) Let $(\Omega_i, \mathcal{F}_i, \nu_i)$ be a measure space and ν_i be σ-finite, $i = 1, 2$. Let λ_i be a measure on $(\Omega_i, \mathcal{F}_i)$ and $\lambda_i \ll \nu_i$, $i = 1, 2$. Then $\lambda_1 \times \lambda_2 \ll \nu_1 \times \nu_2$ and

$$\frac{d(\lambda_1 \times \lambda_2)}{d(\nu_1 \times \nu_2)}(\omega_1, \omega_2) = \frac{d\lambda_1}{d\nu_1}(\omega_1) \frac{d\lambda_2}{d\nu_2}(\omega_2) \quad \text{a.e. } \nu_1 \times \nu_2. \quad ∎$$

1.3 Distributions and Their Characteristics

We now discuss some distributions useful in statistics, and their moments and generating functions.

1.3.1 Useful probability densities

It is often more convenient to work with p.d.f.'s than to work with c.d.f.'s. We now introduce some p.d.f.'s useful in statistics.

Most discrete p.d.f.'s are w.r.t. counting measure on the space of all nonnegative integers. Table 1.1 lists all discrete p.d.f.'s in elementary probability textbooks. For any discrete p.d.f. f, its c.d.f. $F(x)$ can be obtained using (1.19) with $A = (\infty, x]$. Values of $F(x)$ can be obtained from statistical tables or software.

Two Lebesgue p.d.f.'s are introduced in Example 1.11. Some other useful Lebesgue p.d.f.'s are listed in Table 1.2. Note that the exponential p.d.f. in Example 1.11 is a special case of that in Table 1.2 with $a = 0$. For any Lebesgue p.d.f., (1.20) gives its c.d.f. A few c.d.f.'s have explicit forms, whereas many others do not and they have to be evaluated numerically or computed using tables or software.

There are p.d.f.'s that are neither discrete nor Lebesgue.

Example 1.12. Let X be a random variable on $(\Omega, \mathcal{F}, P)$ whose c.d.f. F_X has a Lebesgue p.d.f. f_X and $F_X(c) < 1$, where c is a fixed constant. Let $Y = \min(X, c)$, i.e., Y is the smaller of X and c. Note that $Y^{-1}((-\infty, x]) = \Omega$ if $x \geq c$ and $Y^{-1}((-\infty, x]) = X^{-1}((\infty, x])$ if $x < c$. Hence Y is a random variable and the c.d.f. of Y is

$$F_Y(x) = \begin{cases} 1 & x \geq c \\ F_X(x) & x < c. \end{cases}$$

This c.d.f. is discontinuous at c, since $F(c) < 1$. Thus, it does not have a Lebesgue p.d.f. It is not discrete either. Does P_Y, the probability measure corresponding to F_Y, has a p.d.f. w.r.t. some measure? Define a probability measure on $(\mathcal{R}, \mathcal{B})$, called *point mass* at c, by

$$\delta_c(A) = \begin{cases} 1 & c \in A \\ 0 & c \notin A, \end{cases} \qquad A \in \mathcal{B} \tag{1.21}$$

(which is a special case of the discrete uniform distribution in Table 1.1). Then $P_Y \ll m + \delta_c$, where m is the Lebesgue measure, and the p.d.f. of P_Y is

$$\frac{dP_Y}{d(m+\delta_c)}(x) = \begin{cases} 0 & x > c \\ 1 - F_X(c) & x = c \\ f_X(x) & x < c. \end{cases} \qquad \blacksquare \tag{1.22}$$

Table 1.1. Discrete Distributions on $\mathcal{R}$

Uniform	p.d.f.	$1/m,\ x = a_1, ..., a_m$
	m.g.f.	$\sum_{j=1}^{m} e^{a_j t}/m,\ \ t \in \mathcal{R}$
$DU(a_1, ..., a_m)$	Expectation	$\sum_{j=1}^{m} a_j/m$
	Variance	$\sum_{j=1}^{m}(a_j - \bar{a})^2/m,\ \bar{a} = \sum_{j=1}^{m} a_j/m$
	Parameter	$a_i \in \mathcal{R},\ m = 1, 2, ...$
Binomial	p.d.f.	$\binom{n}{x} p^x (1-p)^{n-x},\ \ x = 0, 1, ..., n$
	m.g.f.	$(pe^t + 1 - p)^n,\ \ t \in \mathcal{R}$
$Bi(p, n)$	Expectation	np
	Variance	$np(1-p)$
	Parameter	$p \in [0, 1],\ \ n = 1, 2, ...$
Poisson	p.d.f.	$\theta^x e^{-\theta}/x!,\ \ x = 0, 1, 2, ...$
	m.g.f.	$e^{\theta(e^t - 1)},\ \ t \in \mathcal{R}$
$P(\theta)$	Expectation	θ
	Variance	θ
	Parameter	$\theta > 0$
Geometric	p.d.f.	$(1-p)^{x-1}p,\ \ x = 1, 2, ...$
	m.g.f.	$pe^t/[1 - (1-p)e^t],\ \ t < -\log(1-p)$
$G(p)$	Expectation	$1/p$
	Variance	$(1-p)/p^2$
	Parameter	$p \in [0, 1]$
Hyper-	p.d.f.	$\binom{n}{x}\binom{m}{r-x} / \binom{N}{r}$
geometric		$x = 0, 1, ..., \min(r, n),\ \ r - x \le m$
	m.g.f.	No explicit form
$HG(r, n, m)$	Expectation	rn/N
	Variance	$rnm(N-r)/[N^2(N-1)]$
	Parameter	$r, n, m = 1, 2, ...,\ N = n + m$
Negative	p.d.f.	$\binom{x-1}{r-1} p^r (1-p)^{x-r},\ \ x = r, r+1, ...$
binomial	m.g.f.	$p^r e^{rt}/[1 - (1-p)e^t]^r,\ \ t < -\log(1-p)$
	Expectation	r/p
$NB(p, r)$	Variance	$r(1-p)/p^2$
	Parameter	$p \in [0, 1],\ \ r = 1, 2, ...$
Log-	p.d.f.	$-(\log p)^{-1} x^{-1} (1-p)^x,\ \ x = 1, 2, ...$
distribution	m.g.f.	$\log[1 - (1-p)e^t]/\log p,\ \ t \in \mathcal{R}$
	Expectation	$-(1-p)/(p \log p)$
$L(p)$	Variance	$-(1-p)[1 + (1-p)/\log p]/(p^2 \log p)$
	Parameter	$p \in (0, 1)$

All p.d.f.'s are w.r.t. counting measure.

The random variable Y in Example 1.12 is a transformation of the random variable X. Transformations of random variables or vectors are frequently used in probability and statistics. For a random variable or vector X, $f(X)$ is a random variable or vector as long as f is measurable (Proposition 1.4). How do we find the c.d.f. (or p.d.f.) of $f(X)$ when the c.d.f. (or p.d.f.) of X is known? In many cases, the most effective method is direct computation. Example 1.12 is one example. The following is another one.

Example 1.13. Let X be a random variable with c.d.f. F_X and Lebesgue p.d.f. f_X, and let $Y = X^2$. Note that $Y^{-1}((-\infty, x]) = \emptyset$ if $x < 0$ and $Y^{-1}((-\infty, x]) = Y^{-1}([0, x]) = X^{-1}([-\sqrt{x}, \sqrt{x}\,])$ if $x \geq 0$. Hence

$$\begin{aligned} F_Y(x) &= P \circ Y^{-1}((-\infty, x]) \\ &= P \circ X^{-1}([-\sqrt{x}, \sqrt{x}\,]) \\ &= F_X(\sqrt{x}) - F_X(-\sqrt{x}) \end{aligned}$$

if $x \geq 0$ and $F_Y(x) = 0$ if $x < 0$. Clearly, the Lebesgue p.d.f. of F_Y is

$$f_Y(x) = \frac{1}{2\sqrt{x}}[f_X(\sqrt{x}) + f_X(-\sqrt{x})]I_{(0,\infty)}(x). \tag{1.23}$$

In particular, if

$$f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \tag{1.24}$$

which is the Lebesgue p.d.f. for the normal distribution $N(0, 1)$ (Table 1.2), then

$$f_Y(x) = \frac{1}{\sqrt{2\pi x}}e^{-x/2}I_{(0,\infty)}(x),$$

which is the Lebesgue p.d.f. for the chi-square distribution χ^2_1 (Table 1.2). This is actually an important result in statistics. ∎

In some cases one may apply the following general result.

Proposition 1.8. Let X be a random k-vector with a Lebesgue p.d.f. f_X and let $Y = g(X)$, where g is a Borel function from $(\mathcal{R}^k, \mathcal{B}^k)$ to $(\mathcal{R}^k, \mathcal{B}^k)$. Let $A_1, ..., A_m$ be disjoint subsets of $\mathcal{R}^k$ such that $\mathcal{R}^k - (A_1 \cup \cdots \cup A_m)$ has Lebesgue measure 0 and g on A_j is one-to-one with a nonvanishing Jacobian, i.e., $\mathrm{Det}(\partial g(x)/\partial x) \neq 0$ on A_j, $j = 1, ..., m$, where $\mathrm{Det}(M)$ is the determinant of a square matrix M. Then Y has the following Lebesgue p.d.f.:

$$f_Y(x) = \sum_{j=1}^{m} \left|\mathrm{Det}\left(\partial h_j(x)/\partial x\right)\right| f_X\left(h_j(x)\right),$$

where h_j is the inverse function of g on A_j, $j = 1, ..., m$. ∎

Table 1.2. Distributions on $\mathcal{R}$ with Lebesgue p.d.f.'s

Uniform	p.d.f.	$(b-a)^{-1}I_{(a,b)}(x)$
	m.g.f.	$(e^{bt}-e^{at})/(b-a),\ \ t\in\mathcal{R}$
$U(a,b)$	Expectation	$(a+b)/2$
	Variance	$(b-a)^2/12$
	Parameter	$a,\ b\in\mathcal{R},\ \ a<b$
Normal	p.d.f.	$\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}$
	m.g.f.	$e^{\mu t-\sigma^2t^2/2},\ \ t\in\mathcal{R}$
$N(\mu,\sigma^2)$	Expectation	μ
	Variance	σ^2
	Parameter	$\mu\in\mathcal{R},\ \ \sigma>0$
Exponential	p.d.f.	$\theta^{-1}e^{-(x-a)/\theta}I_{(a,\infty)}(x)$
	m.g.f.	$e^{at}(1-\theta t)^{-1},\ \ t<\theta^{-1}$
$E(a,\theta)$	Expectation	$\theta+a$
	Variance	θ^2
	Parameter	$\theta>0,\ \ a\in\mathcal{R}$
Chi-square	p.d.f.	$\frac{1}{\Gamma(k/2)2^{k/2}}x^{k/2-1}e^{-x/2}I_{(0,\infty)}(x)$
	m.g.f.	$(1-2t)^{-k/2},\ \ t<1/2$
χ^2_k	Expectation	k
	Variance	$2k$
	Parameter	$k=1,2,...$
Gamma	p.d.f.	$\frac{1}{\Gamma(\alpha)\gamma^\alpha}x^{\alpha-1}e^{-x/\gamma}I_{(0,\infty)}(x)$
	m.g.f.	$(1-\gamma t)^{-\alpha},\ \ t<\gamma^{-1}$
$\Gamma(\alpha,\gamma)$	Expectation	$\alpha\gamma$
	Variance	$\alpha\gamma^2$
	Parameter	$\gamma>0,\ \alpha>0$
Beta	p.d.f.	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}I_{(0,1)}(x)$
	m.g.f.	No explicit form
$B(\alpha,\beta)$	Expectation	$\alpha/(\alpha+\beta)$
	Variance	$\alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]$
	Parameter	$\alpha>0,\ \ \beta>0$
Cauchy	p.d.f.	$\frac{1}{\pi\sigma}\left[1+\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-1}$
	m.g.f.	Does not exist
$C(\mu,\sigma)$	Expectation	Does not exist
	Variance	Does not exist
	ch.f.	$e^{\sqrt{-1}\mu t-\sigma\lvert t\rvert}$
	Parameter	$\mu\in\mathcal{R},\ \ \sigma>0$

Table 1.2. (continued)

t-distribution	p.d.f.	$\frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\Gamma(n/2)}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}$
	m.g.f.	No explicit form
t_n	Expectation	$0,\ \ (n>1)$
	Variance	$n/(n-2),\ \ (n>2)$
	Parameter	$n=1,2,...$
F-distribution	p.d.f.	$\frac{n^{n/2}m^{m/2}\Gamma[(n+m)/2]x^{n/2-1}}{\Gamma(n/2)\Gamma(m/2)(m+nx)^{(n+m)/2}}I_{(0,\infty)}(x)$
	m.g.f.	No explicit form
$F_{n,m}$	Expectation	$m/(m-2),\ \ (m>2)$
	Variance	$2m^2(n+m-2)/[n(m-2)^2(m-4)],$ $(m>4)$
	Parameter	$n=1,2,...,\ \ m=1,2,...$
Log-normal	p.d.f.	$\frac{1}{\sqrt{2\pi}\sigma}x^{-1}e^{-(\log x-\mu)^2/2\sigma^2}I_{(0,\infty)}(x)$
	m.g.f.	Does not exist
$LN(\mu,\sigma^2)$	Expectation	$e^{\mu+\sigma^2/2}$
	Variance	$e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$
	Parameter	$\mu\in\mathcal{R},\ \ \sigma>0$
Weibull	p.d.f.	$\frac{\alpha}{\theta}x^{\alpha-1}e^{-x^\alpha/\theta}I_{(0,\infty)}(x)$
	m.g.f.	No explicit form
$W(\alpha,\theta)$	Expectation	$\theta^{1/\alpha}\Gamma(\alpha^{-1}+1)$
	Variance	$\theta^{2/\alpha}\left[\Gamma(2\alpha^{-1}+1)-\Gamma(\alpha^{-1}+1)\right]^2$
	Parameter	$\theta>0,\ \ \alpha>0$
Double	p.d.f.	$\frac{1}{2\theta}e^{-\lvert x-\mu\rvert/\theta}$
Exponential	m.g.f.	$e^{\mu t}/(1+\theta^2t^2),\ \ t\in\mathcal{R}$
	Expectation	μ
$DE(\mu,\theta)$	Variance	$2\theta^2$
	Parameter	$\mu\in\mathcal{R},\ \ \theta>0$
Pareto	p.d.f.	$\theta a^\theta x^{-(\theta+1)}I_{(a,\infty)}(x)$
	m.g.f.	No explicit form
$Pa(a,\theta)$	Expectation	$\theta a/(\theta-1),\ \ (\theta>1)$
	Variance	$\theta a^2/[(\theta-1)^2(\theta-2)],\ \ (\theta>2)$
	Parameter	$\theta>0,\ \ a>0$
Logistic	p.d.f.	$\sigma^{-1}e^{-(x-\mu)/\sigma}/[1+e^{-(x-\mu)/\sigma}]^2$
	m.g.f.	$e^{\mu t}\Gamma(1+\sigma t)\Gamma(1-\sigma t),\ \lvert t\rvert<\sigma$
$LG(\mu,\sigma)$	Expectation	μ
	Variance	$\sigma^2\pi^2/3$
	Parameter	$\mu\in\mathcal{R},\ \sigma>0$

One may apply Proposition 1.8 to obtain result (1.23) in Example 1.13, using $A_1 = (-\infty, 0)$, $A_2 = (0, \infty)$, and $g(x) = x^2$. Note that $h_1(x) = -\sqrt{x}$, $h_2(x) = \sqrt{x}$, and $|dh_j(x)/dx| = 1/(2\sqrt{x})$.

A p.d.f. corresponding to a joint c.d.f. is called a joint p.d.f. The following is an important joint Lebesgue p.d.f. in statistics:

$$f(x) = (2\pi)^{-k/2}[\mathrm{Det}(\Sigma)]^{-1/2}e^{-(x-\mu)\Sigma^{-1}(x-\mu)^\tau/2}, \quad x \in \mathcal{R}^k, \tag{1.25}$$

where $\mu \in \mathcal{R}^k$ is a vector of parameters, Σ is a positive definite $k \times k$ matrix of parameters, and A^τ denotes the transpose of a vector or matrix A. The p.d.f. in (1.25) and its c.d.f. are called the k-dimensional multivariate normal p.d.f. and c.d.f. and both are denoted by $N_k(\mu, \Sigma)$. The normal distribution $N(\mu, \sigma^2)$ in Table 1.2 is a special case of $N_k(\mu, \Sigma)$ with $k = 1$. The p.d.f. in (1.24) is $N(0, 1)$ and is called the *standard* normal p.d.f. Sometimes random vectors having the $N_k(\mu, \Sigma)$ distribution are also denoted by $N_k(\mu, \Sigma)$ for convenience. Useful properties of multivariate normal distributions can be found in Exercise 51.

Let X be a random k-vector having a c.d.f. F_X. Then the ith component of X is a random variable X_i having the following c.d.f.:

$$F_{X_i}(x) = \lim_{x_j \to \infty, j=1,\ldots,i-1,i+1,\ldots,k} F_X(x_1, ..., x_{i-1}, x, x_{i+1}, ..., x_k),$$

which is called the *marginal* c.d.f. of X_i. That is, the k marginal c.d.f.'s are determined by the joint c.d.f. If F_X has a Lebesgue p.d.f. f_X, then X_i has the following Lebesgue p.d.f.:

$$f_{X_i}(x) = \int \cdots \int f_X(x_1, ..., x_{i-1}, x, x_{i+1}, ..., x_k)dx_1 \cdots dx_{i-1}dx_{i+1} \cdots dx_k.$$

In general, a joint c.d.f. cannot be determined by k marginal c.d.f.'s. There is one special but important case in which the joint c.d.f. of a random k-vector is determined by its k marginal c.d.f.'s, i.e.,

$$F_X(x_1, ..., x_k) = F_{X_1}(x_1) \cdots F_{X_k}(x_k), \quad (x_1, ..., x_k) \in \mathcal{R}^k. \tag{1.26}$$

If (1.26) holds, then random variables $X_1, ..., X_k$ are said to be *independent*. The meaning of independence is further discussed in §1.4.2. If each X_i has a Lebesgue p.d.f. f_{X_i}, then $X_1, ..., X_k$ are independent if and only if the joint p.d.f. of X satisfies

$$f_X(x_1, ..., x_k) = f_{X_1}(x_1) \cdots f_{X_k}(x_k), \quad (x_1, ..., x_k) \in \mathcal{R}^k. \tag{1.27}$$

Example 1.14. Let $X = (X_1, X_2)$ be a random 2-vector having a joint Lebesgue p.d.f. f_X. Consider first the transformation $g(x) = (x_1, x_1 + x_2)$. Using Proposition 1.8, one can show that the joint p.d.f. of $g(X)$ is

$$f_{g(X)}(x_1, y) = f_X(x_1, y - x_1),$$

where $y = x_1 + x_2$ (note that the Jacobian equals 1). The marginal p.d.f. of $Y = X_1 + X_2$ is then

$$f_Y(y) = \int f_X(x_1, y - x_1)dx_1.$$

In particular, if X_1 and X_2 are independent, then

$$f_Y(y) = \int f_{X_1}(x_1)f_{X_2}(y - x_1)dx_1. \tag{1.28}$$

Next, consider the transformation $h(x_1, x_2) = (x_1/x_2, x_2)$, assuming that $X_2 \neq 0$ a.s. Using Proposition 1.8, one can show that the joint p.d.f. of $h(X)$ is

$$f_{h(X)}(z, x_2) = |x_2| f_X(zx_2, x_2),$$

where $z = x_1/x_2$. The marginal p.d.f. of $Z = X_1/X_2$ is

$$f_Z(z) = \int |x_2| f_X(zx_2, x_2)dx_2.$$

In particular, if X_1 and X_2 are independent, then

$$f_Z(z) = \int |x_2| f_{X_1}(zx_2)f_{X_2}(x_2)dx_2. \quad \blacksquare \tag{1.29}$$

A number of results can be derived from (1.28) and (1.29). For example, if X_1 and X_2 are independent and both have the standard normal p.d.f. given by (1.24), then, by (1.29), the Lebesgue p.d.f. of $Z = X_1/X_2$ is

$$\begin{aligned} f_Z(z) &= \frac{1}{2\pi}\int |x_2| e^{-(1+z^2)x_2^2/2}dx_2 \\ &= \frac{1}{\pi}\int_0^\infty e^{-(1+z^2)x}dx \\ &= \frac{1}{\pi(1+z^2)}, \end{aligned}$$

which is the p.d.f. of the Cauchy distribution $C(0, 1)$ in Table 1.2. Another application of formula (1.29) leads to the following important result in statistics.

Example 1.15 (t-distribution and F-distribution). Let X_1 and X_2 be independent random variables having the chi-square distributions $\chi^2_{n_1}$ and $\chi^2_{n_2}$ (Table 1.2), respectively. By (1.29), the p.d.f. of $Z = X_1/X_2$ is

$$\begin{aligned} f_Z(z) &= \frac{z^{n_1/2-1}I_{(0,\infty)}(z)}{2^{(n_1+n_2)/2}\Gamma(n_1/2)\Gamma(n_2/2)}\int_0^\infty x_2^{(n_1+n_2)/2-1}e^{-(1+z)x_2/2}dx_2 \\ &= \frac{\Gamma[(n_1+n_2)/2]}{\Gamma(n_1/2)\Gamma(n_2/2)}\frac{z^{n_1/2-1}}{(1+z)^{(n_1+n_2)/2}}I_{(0,\infty)}(z), \end{aligned}$$

where the last equality follows from the fact that

$$\frac{1}{2^{(n_1+n_2)/2}\Gamma[(n_1+n_2)/2]}x_2^{(n_1+n_2)/2-1}e^{-x_2/2}I_{(0,\infty)}(x_2)$$

is the p.d.f. of the chi-square distribution $\chi^2_{n_1+n_2}$. Using Proposition 1.8, one can show that the p.d.f. of $Y = (X_1/n_1)/(X_2/n_2) = (n_2/n_1)Z$ is the p.d.f. of the F-distribution F_{n_1,n_2} given in Table 1.2.

Let U_1 be a random variable having the standard normal distribution $N(0,1)$ and U_2 a random variable having the chi-square distribution χ^2_n. Using the same argument, one can show that if U_1 and U_2 are independent, then the distribution of $T = U_1/\sqrt{U_2/n}$ is the t-distribution t_n given in Table 1.2. This result can also be derived using the result given in this example as follows. Let $X_1 = U_1^2$ and $X_2 = U_2$. Then X_1 and X_2 are independent (which can be shown directly, but follows immediately from Proposition 1.13 in §1.4.2). By Example 1.13, the distribution of X_1 is χ^2_1. Then $Y = X_1/(X_2/n)$ has the F-distribution $F_{1,n}$ and its Lebesgue p.d.f. is

$$\frac{n^{n/2}\Gamma[(n+1)/2]x^{-1/2}}{\sqrt{n\pi}\Gamma(n/2)(n+x)^{(n+1)/2}}I_{(0,\infty)}(x).$$

Note that

$$T = \begin{cases} \sqrt{Y} & U_1 \geq 0 \\ -\sqrt{Y} & U_1 < 0. \end{cases}$$

The result follows from Proposition 1.8 and the fact that

$$P \circ T^{-1}\left((-\infty, -t]\right) = P \circ T^{-1}\left([t, \infty)\right), \quad t > 0. \quad \blacksquare \tag{1.30}$$

If a random variable T satisfies (1.30), then T and its c.d.f. and p.d.f. (if it exists) are said to be *symmetric* about 0. If T has a Lebesgue p.d.f. f_T, then T is symmetric about 0 if and only if $f_T(x) = f_T(-x)$ for any $x > 0$. T and its c.d.f. and p.d.f. are said to be symmetric about a (or symmetric for simplicity) if and only if $T - a$ is symmetric about 0 for a fixed $a \in \mathcal{R}$. The c.d.f.'s of t-distributions are symmetric about 0 and the normal, Cauchy, and double exponential c.d.f.'s are symmetric.

The chi-square, t-, and F-distributions in the previous examples are special cases of the following *noncentral* chi-square, t-, and F-distributions, which are useful in some statistical problems.

Let $X_1, ..., X_n$ be independent random variables and $X_i = N(\mu_i, \sigma^2)$, $i = 1, ..., n$. The distribution of the random variable $Y = (X_1^2 + \cdots + X_n^2)/\sigma^2$ is called the *noncentral chi-square* distribution and denoted by $\chi^2_n(\delta)$, where $\delta = (\mu_1^2 + \cdots + \mu_n^2)/\sigma^2$ is the noncentrality parameter. It can be shown

(exercise) that Y has the following Lebesgue p.d.f.:

$$e^{-\delta/2}\sum_{j=0}^{\infty}\frac{\delta^j}{2^j j!}f_{2j+n}(x), \tag{1.31}$$

where $f_k(x)$ is the Lebesgue p.d.f. of the chi-square distribution χ^2_k. It is easy to see that the chi-square distribution χ^2_k in Table 1.2 is a special case of the noncentral chi-square distribution $\chi^2_k(\delta)$ with $\delta = 0$ and, therefore, is called a *central* chi-square distribution.

The result for the t-distribution in Example 1.15 can be extended to the case where U_1 has a nonzero expectation (U_2 still has the χ^2_n distribution and is independent of U_1). The distribution of $T = U_1/\sqrt{U_2/n}$ is called the *noncentral* t-distribution and denoted by $t_n(\delta)$, where $\delta = \mu$ is the noncentrality parameter. Using the same argument as that in Example 1.15, one can show (exercise) that T has the following Lebesgue p.d.f.:

$$\frac{1}{2^{(n+1)/2}\Gamma(n/2)\sqrt{\pi n}}\int_0^{\infty} y^{(n-1)/2}e^{-[(x\sqrt{y/n}-\delta)^2+y]/2}dy. \tag{1.32}$$

The t-distribution t_n in Example 1.15 is called a *central* t-distribution, since it is a special case of the noncentral t-distribution $t_n(\delta)$ with $\delta = 0$.

Similarly, the result for the F-distribution in Example 1.15 can be extended to the case where X_1 has the noncentral chi-square distribution $\chi^2_{n_1}(\delta)$, X_2 has the central chi-square distribution $\chi^2_{n_2}$, and X_1 and X_2 are independent. The distribution of $Y = (X_1/n_1)/(X_2/n_2)$ is called the *noncentral* F-distribution and denoted by $F_{n_1,n_2}(\delta)$, where δ is the noncentrality parameter. It can be shown (exercise) that Y has the following Lebesgue p.d.f.:

$$\frac{e^{-\delta/2}n_1^{n_1/2}n_2^{n_2/2}}{\Gamma(n_2/2)}\sum_{j=0}^{\infty}\frac{(\delta n_1 x/2)^j\Gamma((n_1+n_2)/2+j)x^{n_1/2-1}}{j!\Gamma(n_1/2+j)(n_1x+n_2)^{(n_1+n_2)/2+j}}I_{(0,\infty)}(x). \tag{1.33}$$

The F-distribution F_{n_1,n_2} in Example 1.15 is called a *central* F-distribution, since it is a special case of the noncentral F-distribution $F_{n_1,n_2}(\delta)$ with $\delta = 0$.

1.3.2 Moments and generating functions

We have defined the expectation of a random variable in §1.2.1. It is an important characteristic of a random variable. In this section we introduce other important *moments* and two *generating functions* of a random vector.

Let X be a random variable. If EX^k is finite, where k is a positive integer, then EX^k is called the kth *moment* of X (or the distribution of

X). If $E|X|^a < \infty$ for some real number a, then $E|X|^a$ is called the ath *absolute moment* of X (or the distribution of X). If $\mu = EX$ and $E(X-\mu)^k$ are finite for a positive integer k, then $E(X-\mu)^k$ is called the kth *central moment* of X (or the distribution of X).

The expectation and the second central moment (if they exist) are two important characteristics of a random variable (or its distribution) in statistics. They are listed in Tables 1.1 and 1.2 for those useful distributions. The expectation, also called the *mean* in statistics, is a measure of the central location of the distribution of a random variable. The second central moment, also called the *variance* in statistics, is a measure of dispersion or spread of a random variable. The variance of a random variable X is denoted by $\mathrm{Var}(X)$. The variance is always nonnegative. If the variance of X is 0, then X is equal to its mean a.s. (Proposition 1.6). The squared root of the variance is called the *standard deviation*, another important characteristic of a random variable in statistics.

The concept of mean and variance can be extended to random vectors. The expectation of a random matrix M with (i,j)th element M_{ij} is defined to be the matrix whose (i,j)th element is EM_{ij}. Thus, for a random k-vector $X = (X_1, ..., X_k)$, its mean is $EX = (EX_1, ..., EX_k)$; the extension of variance is the *variance-covariance matrix* of X defined as

$$\mathrm{Var}(X) = E(X - EX)^\tau (X - EX),$$

which is a $k \times k$ symmetric matrix whose diagonal elements are variances of X_i's. The (i,j)th element of $\mathrm{Var}(X)$, $i \neq j$, is $E(X_i - EX_i)(X_j - EX_j)$, which is called the *covariance* of X_i and X_j and is denoted by $\mathrm{Cov}(X_i, X_j)$.

Let $c = (c_1, ..., c_k) \in \mathcal{R}^k$ and $X = (X_1, ..., X_k)$ be a random k-vector. Then $Y = cX^\tau = c_1X_1 + \cdots + c_kX_k$ is a random variable, and

$$EY = c_1EX_1 + \cdots + c_kEX_k = cEX^\tau$$

and

$$\begin{aligned}\mathrm{Var}(Y) &= E(cX^\tau - cEX^\tau)^2 \\ &= E[c(X - EX)^\tau (X - EX)c^\tau] \\ &= c[E(X - EX)^\tau (X - EX)]c^\tau \\ &= c\mathrm{Var}(X)c^\tau,\end{aligned}$$

assuming that all expectations exist. Since $\mathrm{Var}(Y) \geq 0$ for any $c \in \mathcal{R}^k$, the matrix $\mathrm{Var}(X)$ is nonnegative definite. Consequently,

$$[\mathrm{Cov}(X_i, X_j)]^2 \leq \mathrm{Var}(X_i)\mathrm{Var}(X_j), \quad i \neq j. \tag{1.34}$$

An important quantity in statistics is the *correlation coefficient* defined to be $\rho_{X_i,X_j} = \mathrm{Cov}(X_i, X_j)/\sqrt{\mathrm{Var}(X_i)\mathrm{Var}(X_j)}$, which is, by inequality (1.34),

always between -1 and 1. It is a measure of relationship between X_i and X_j; if ρ_{X_i,X_j} is positive (or negative), then X_i and X_j tend to be positively (or negatively) related; if $\rho_{X_i,X_j} = \pm 1$, then $P(X_i = c_1 \pm c_2 X_j) = 1$ with some constants c_1 and $c_2 > 0$; if $\rho_{X_i,X_j} = 0$ (i.e., $\mathrm{Cov}(X_i, X_j) = 0$), then X_i and X_j are said to be *uncorrelated*. One can show that if X_i and X_j are independent, then they are uncorrelated. But the converse is not necessarily true. Examples can be found in Exercises 48-49.

The following result indicates that if the rank of $\mathrm{Var}(X)$ is $r < k$, then X is in a subspace of $\mathcal{R}^k$ with dimension r. For any $k \times k$ symmetric matrix M, define $R_M = \{y \in \mathcal{R}^k : y = xM \text{ with some } x \in \mathcal{R}^k\}$.

Proposition 1.9. Let X be a random k-vector with a finite $\mathrm{Var}(X)$. Then we have the following conclusions.
(i) $P(X - EX \in R_{\mathrm{Var}(X)}) = 1$.
(ii) If $P_X \ll$ Lebesgue measure on $\mathcal{R}^k$, then the rank of $\mathrm{Var}(X)$ is k. ∎

Example 1.16. Let X be a random k-vector having the $N_k(\mu, \Sigma)$ distribution. It can be shown (exercise) that $EX = \mu$ and $\mathrm{Var}(X) = \Sigma$. Thus, μ and Σ in (1.25) are the mean vector and the variance-covariance matrix of X. If Σ is a diagonal matrix (i.e., all components of X are uncorrelated), then by (1.27), the components of X are independent. This shows an important property of random variables having normal distributions: they are independent if and only if they are uncorrelated. ∎

Moments are important characteristics of a distribution, but they do not determine a distribution in the sense that two different distributions may have the same moments of all orders. Functions that determine a distribution are introduced in the following definition.

Definition 1.5. Let X be a random k-vector.
(i) The *moment generating function* (m.g.f.) of X (or P_X) is defined as

$$\psi_X(t) = Ee^{tX^\tau}, \quad t \in \mathcal{R}^k.$$

(ii) The *characteristic function* (ch.f.) of X (or P_X) is defined as

$$\phi_X(t) = Ee^{\sqrt{-1}tX^\tau} = E[\cos(tX^\tau)] + \sqrt{-1}\,E[\sin(tX^\tau)], \quad t \in \mathcal{R}^k. \quad ∎$$

The ch.f. is complex-valued and always well defined. The m.g.f. is nonnegative but may be ∞ everywhere except at $t = 0$. If the m.g.f. is finite at some $t \neq 0$, then $\phi_X(t)$ can be obtained by replacing t in $\psi_X(t)$ by $\sqrt{-1}t$. Tables 1.1 and 1.2 contain the m.g.f. (or ch.f. when the m.g.f. is ∞ everywhere except at 0) for distributions useful in statistics. Some useful properties of the m.g.f. and ch.f. are given in the following result.

Proposition 1.10. Let X be a random k-vector with m.g.f. $\psi_X(t)$ and ch.f. $\phi_X(t)$.
(i) (Relation to moments). If EX is finite, then

$$\left.\frac{\partial \phi_X(t)}{\partial t}\right|_{t=0} = \sqrt{-1}EX.$$

If $\mathrm{Var}(X)$ is finite, then

$$\left.\frac{\partial^2 \phi_X(t)}{\partial t \partial t^\tau}\right|_{t=0} = -E(X^\tau X).$$

If $k = 1$ and EX^p is finite for a positive integer p, then

$$\left.\frac{d^p \phi_X(t)}{dt^p}\right|_{t=0} = (-1)^{p/2} EX^p.$$

If $\psi_X(t) < \infty$ for $t \in N_\epsilon = \{t \in \mathcal{R}^k : tt^\tau \leq \epsilon\}$, then the components of X have finite moments of all orders,

$$\left.\frac{\partial \psi_X(t)}{\partial t}\right|_{t=0} = EX,$$

$$\left.\frac{\partial^2 \psi_X(t)}{\partial t \partial t^\tau}\right|_{t=0} = E(X^\tau X),$$

and, when $k = 1$ and p is a positive integer,

$$\left.\frac{d^p \psi_X(t)}{dt^p}\right|_{t=0} = EX^p.$$

(ii) (Uniqueness). If Y is a random k-vector and $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathcal{R}^k$, then $P_X = P_Y$. If there is an $\epsilon > 0$ such that $\psi_X(t) = \psi_Y(t) < \infty$ for all $t \in N_\epsilon = \{t \in \mathcal{R}^k : tt^\tau \leq \epsilon\}$, then $P_X = P_Y$.
(iii) (Sums of independent random vectors). Let Y be a random k-vector independent of X. Then

$$\psi_{X+Y}(t) = \psi_X(t)\psi_Y(t), \quad t \in \mathcal{R}^k,$$

and the same result holds when ψ is replaced by ϕ.
(iv) (Linear transformations). Let $Y = XC^\tau + c$, where C is an $m \times k$ matrix and $c \in \mathcal{R}^m$. Then

$$\psi_Y(u) = e^{uc^\tau}\psi_X(uC), \quad u \in \mathcal{R}^m,$$

and

$$\phi_Y(u) = e^{\sqrt{-1}uc^\tau}\phi_X(uC), \quad u \in \mathcal{R}^m. \quad \blacksquare$$

Proposition 1.10(ii)-(iii) provides a useful tool to obtain distributions of sums of independent random vectors. The following example is an illustration.

Example 1.17. Let X_i, $i = 1, ..., k$, be independent random variables and X_i have the gamma distribution $\Gamma(\alpha_i, \gamma)$ (Table 1.2), $i = 1, ..., k$. From Table 1.2, X_i has the m.g.f. $\psi_{X_i}(t) = (1 - \gamma t)^{-\alpha_i}$, $t < \gamma^{-1}$, $i = 1, ..., k$. By Proposition 1.10(iii), the m.g.f. of $Y = X_1 + \cdots + X_k$ is equal to $\psi_Y(t) = (1 - \gamma t)^{-(\alpha_1 + \cdots + \alpha_k)}$, $t < \gamma^{-1}$. From Table 1.2, the gamma distribution $\Gamma(\alpha_1 + \cdots + \alpha_k, \gamma)$ has the m.g.f. $\psi_Y(t)$ and, hence, is the distribution of Y (by Proposition 1.10(ii)). ∎

Using Proposition 1.10 and a result from linear algebra, we can prove the following result useful in *analysis of variance* (Scheffé, 1959; Searle, 1971).

Theorem 1.5. (i) Suppose that $Y_1, ..., Y_k$ are independent random variables and that Y_i has the noncentral chi-square distribution $\chi^2_{n_i}(\delta_i)$, $i = 1, ..., k$. Then $Y = Y_1 + \cdots + Y_k$ has the noncentral chi-square distribution $\chi^2_{n_1 + \cdots + n_k}(\delta_1 + \cdots + \delta_k)$.
(ii) (Cochran's theorem). Suppose that $X = N_n(\mu, I_n)$ and

$$XX^\tau = XA_1X^\tau + \cdots + XA_kX^\tau, \tag{1.35}$$

where A_i is a nonnegative definite $n \times n$ matrix with rank n_i, $i = 1, ..., k$. Then a necessary and sufficient condition that XA_iX^τ has the noncentral chi-square distribution $\chi^2_{n_i}(\delta_i)$, $i = 1, ..., k$, and XA_iX^τ's are independent is $n = n_1 + \cdots + n_k$, in which case $\delta_i = \mu A_i \mu^\tau$ and $\delta_1 + \cdots + \delta_k = \mu\mu^\tau$.
Proof. (i) The ch.f. of Y_i is $(1 - 2\sqrt{-1}t)^{n_i/2} e^{\sqrt{-1}\delta_i t/(1 - 2\sqrt{-1}t)}$ (Exercise 53). Then, the result follows from Proposition 1.10(ii)-(iii).
(ii) The necessity follows from (i) and the fact that XX^τ has the noncentral chi-square distribution $\chi^2_n(\mu\mu^\tau)$ (by definition). We now prove the sufficiency.

Assume that $n = n_1 + \cdots + n_k$. We use the following fact from linear algebra: there exists an $n \times n$ matrix C such that $Y = XC$ and

$$XA_iX^\tau = \sum_{j=n_1+\cdots+n_{i-1}+1}^{n_1+\cdots+n_{i-1}+n_i} Y_j^2, \tag{1.36}$$

where Y_j is the jth component of Y. From (1.35) and (1.36), $XX^\tau = YY^\tau$, i.e., $CC^\tau = I_n$. Thus, $Y = N_n(\mu C, I_n)$ (Exercise 51); the independence of XA_iX^τ follows from (1.36); and the fact that XA_iX^τ has the noncentral chi-square distribution follows directly from the definition of the noncentral chi-square distribution and (1.36). ∎

1.4 Conditional Expectations

In elementary probability the conditional probability of an event B given an event A is defined as $P(B|A) = P(A \cap B)/P(A)$, provided that $P(A) > 0$. In probability and statistics, however, we sometimes need a notion of "conditional probability" even for A's with $P(A) = 0$; for example, $A = \{Y = c\}$, where Y is a random variable and $c \in \mathcal{R}$. General definitions of conditional probability, expectation, and distribution are introduced in this section, and they are shown to agree with those defined in elementary probability in special cases.

1.4.1 Conditional expectations

Definition 1.6. Let X be an integrable random variable on $(\Omega, \mathcal{F}, P)$.
(i) Let $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$. The *conditional expectation* of X given $\mathcal{A}$, denoted by $E(X|\mathcal{A})$, is the a.s.-unique random variable satisfying the following two conditions:

(a) $E(X|\mathcal{A})$ is measurable from $(\Omega, \mathcal{A})$ to $(\mathcal{R}, \mathcal{B})$;
(b) $\int_A E(X|\mathcal{A})dP = \int_A XdP$ for any $A \in \mathcal{A}$.

(ii) Let $B \in \mathcal{F}$. The *conditional probability* of B given $\mathcal{A}$ is defined to be $P(B|\mathcal{A}) = E(I_B|\mathcal{A})$.
(iii) Let Y be measurable from $(\Omega, \mathcal{F}, P)$ to $(\Lambda, \mathcal{G})$. The conditional expectation of X given Y is defined to be $E(X|Y) = E[X|\sigma(Y)]$. ∎

Essentially, the σ-field $\sigma(Y)$ contains "the information in Y". Hence, $E(X|Y)$ is the "expectation" of X given the information provided by $\sigma(Y)$. The following useful result shows that there is a Borel function h defined on the range of Y such that $E(X|Y) = h \circ Y$.

Theorem 1.6. Let Y be measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ and Z a function from $(\Omega, \mathcal{F})$ to $\mathcal{R}^k$. Then Z is measurable from $(\Omega, \sigma(Y))$ to $(\mathcal{R}^k, \mathcal{B}^k)$ if and only if there is a measurable function h from $(\Lambda, \mathcal{G})$ to $(\mathcal{R}^k, \mathcal{B}^k)$ such that $Z = h \circ Y$. ∎

The function h in $E(X|Y) = h \circ Y$ is a Borel function on $(\Lambda, \mathcal{G})$. Let $y \in \Lambda$. Then we define

$$E(X|Y = y) = h(y)$$

to be the conditional expectation of X given $Y = y$. Note that $h(y)$ is a function on Λ, whereas $h \circ Y$ is a function on Ω.

Example 1.18. Let X be an integrable random variable on $(\Omega, \mathcal{F}, P)$, $A_1, A_2, ...$ be disjoint events on $(\Omega, \mathcal{F}, P)$ such that $\cup A_i = \Omega$ and $P(A_i) > 0$

for all i, and let $a_1, a_2, ...$ be distinct real numbers. Define $Y = a_1 I_{A_1} + a_2 I_{A_2} + \cdots$. We now show that

$$E(X|Y) = \sum_{i=1}^{\infty} \frac{\int_{A_i} X dP}{P(A_i)} I_{A_i}. \tag{1.37}$$

We need to verify (a) and (b) in Definition 1.6 with $\mathcal{A} = \sigma(Y)$. Since $\sigma(Y) = \sigma(\{A_1, A_2, ...\})$, it is clear that the function on the right-hand side of (1.37) is measurable on $(\Omega, \sigma(Y))$. For any $B \in \mathcal{B}$, $Y^{-1}(B) = \cup_{i:a_i \in B} A_i$. Using properties of integrals, we obtain that

$$\begin{aligned}
\int_{Y^{-1}(B)} X dP &= \sum_{i:a_i \in B} \int_{A_i} X dP \\
&= \sum_{i=1}^{\infty} \frac{\int_{A_i} X dP}{P(A_i)} P\left(A_i \cap Y^{-1}(B)\right) \\
&= \int_{Y^{-1}(B)} \left[\sum_{i=1}^{\infty} \frac{\int_{A_i} X dP}{P(A_i)} I_{A_i} \right] dP.
\end{aligned}$$

This verifies (b) and thus (1.37) holds.

Let $A \in \mathcal{F}$ and $X = I_A$. Then

$$P(A|Y) = E(X|Y) = \sum_{i=1}^{\infty} \frac{P(A \cap A_i)}{P(A_i)} I_{A_i}.$$

Note that $\{Y = a_i\} = \{\omega \in \Omega : Y(\omega) = a_i\} = A_i$. If $\omega \in A_i$,

$$P(A|Y)(\omega) = \frac{P(A \cap A_i)}{P(A_i)} = P(A|A_i) = P(A|\{Y = a_i\}).$$

Hence the definition of conditional probability in Definition 1.6 agrees with that in elementary probability. More generally, let X be a discrete random variable whose range is $\{c_1, c_2, ...\}$, where c_i's are distinct real numbers. Let $C_i = X^{-1}(\{c_i\})$, $i = 1, 2, ...$ Then, by (1.37),

$$E(X|Y) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} c_j P(C_j|A_i) I_{A_i}.$$

If $\omega \in A_i$, then

$$E(X|Y)(\omega) = \sum_{j=1}^{\infty} c_j P(C_j|A_i) = \sum_{j=1}^{\infty} c_j P(C_j|\{Y = a_i\}),$$

which agrees with $E(X|Y = a_i)$ defined in elementary probability.

Let h be a Borel function on $\mathcal{R}$ satisfying $h(a_i) = \int_{A_i} X dP/P(A_i)$. Then, by (1.37), $E(X|Y) = h \circ Y$ and $E(X|Y = y) = h(y)$. ∎

The next result generalizes the result in Example 1.18 to conditional expectations of random variables having p.d.f's.

Proposition 1.11. Let X be a random n-vector and Y a random m-vector. Suppose that (X, Y) has a joint p.d.f. $f(x, y)$ w.r.t. $\nu \times \lambda$, where ν and λ are σ-finite measures on $(\mathcal{R}^n, \mathcal{B}^n)$ and $(\mathcal{R}^m, \mathcal{B}^m)$, respectively. Let $g(x, y)$ be a Borel function on $\mathcal{R}^{n+m}$ for which $E|g(X, Y)| < \infty$. Then

$$E[g(X,Y)|Y] = \frac{\int g(x,Y)f(x,Y)d\nu(x)}{\int f(x,Y)d\nu(x)} \quad \text{a.s.} \tag{1.38}$$

Proof. Denote the right-hand side of (1.38) by $h(Y)$. By Fubini's theorem, h is Borel. Then, by Theorem 1.6, $h(Y)$ is Borel on $(\Omega, \sigma(Y))$. Also, by Fubini's theorem, $f_Y(y) = \int f(x, y)d\nu(x)$ is the p.d.f. of Y w.r.t. λ. For $B \in \mathcal{B}^m$,

$$\begin{aligned}
\int_{Y^{-1}(B)} h(Y)dP &= \int_B h(y)dP_Y \\
&= \int_B \frac{\int g(x,y)f(x,y)d\nu(x)}{\int f(x,y)d\nu(x)} f_Y(y)d\lambda(y) \\
&= \int_{\mathcal{R}^n \times B} g(x,y)f(x,y)d\nu \times \lambda \\
&= \int_{\mathcal{R}^n \times B} g(X,Y)dP_{(X,Y)} \\
&= \int_{Y^{-1}(B)} g(X,Y)dP,
\end{aligned}$$

where the first and the last equalities follow from Theorem 1.2; the second and the next to last equalities follow from the definition of h and p.d.f.'s; and the third equality follows from Theorem 1.3 (Fubini's theorem). ∎

For a random vector (X, Y) with a joint p.d.f. $f(x, y)$ w.r.t. $\nu \times \lambda$, define the *conditional* p.d.f. of X given $Y = y$ to be

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}, \tag{1.39}$$

where $f_Y(y) = \int f(x, y)d\nu(x)$ is the marginal p.d.f. of Y w.r.t. ν. One can easily check that for each fixed y with $f_Y(y) > 0$, $f_{X|Y}(x|y)$ in (1.39) is a p.d.f. w.r.t. ν. Then equation (1.38) can be rewritten as

$$E[g(X,Y)|Y] = \int g(x,Y)f_{X|Y}(x|Y)d\nu(x).$$

Again, this agrees with the conditional expectation defined in elementary probability (i.e., the conditional expectation of $g(X,Y)$ given Y is equal to the expectation of $g(X,Y)$ w.r.t. the conditional p.d.f. of X given Y).

Now we list some useful properties of conditional expectations. The proof is left to the reader.

Proposition 1.12. Let X and Y be integrable random variables on $(\Omega, \mathcal{F}, P)$ and $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$.
(i) If $X = c$ a.s., $c \in \mathcal{R}$, then $E(X|\mathcal{A}) = c$ a.s.
(ii) If $X \le Y$ a.s., then $E(X|\mathcal{A}) \le E(Y|\mathcal{A})$ a.s.
(iii) If a and b are real numbers, then $E(aX+bY|\mathcal{A}) = aE(X|\mathcal{A})+bE(Y|\mathcal{A})$ a.s.
(iv) $E[E(X|\mathcal{A})] = EX$.
(v) $E[E(X|\mathcal{A})|\mathcal{A}_0] = E(X|\mathcal{A}_0) = E[E(X|\mathcal{A}_0)|\mathcal{A}]$ a.s., where $\mathcal{A}_0$ is a sub-σ-field of $\mathcal{A}$.
(vi) If $\sigma(Y) \subset \mathcal{A}$ and $E|XY| < \infty$, then $E(XY|\mathcal{A}) = YE(X|\mathcal{A})$ a.s.
(vii) If $E|g(X,Y)| < \infty$, then $E[g(X,Y)|Y=y] = E[g(X,y)|Y=y]$ a.s.
(viii) If $EX^2 < \infty$, then $[E(X|Y)]^2 \le E(X^2|Y)$ a.s. ∎

As an application, we consider the following example.

Example 1.19. Let X be a random variable on $(\Omega, \mathcal{F}, P)$ with $EX^2 < \infty$ and Y a measurable function from $(\Omega, \mathcal{F}, P)$ to $(\Lambda, \mathcal{G})$. One may wish to predict the value of X based on an observed value of Y. Let $g(Y)$ be a predictor, i.e., $g \in \aleph = \{$all Borel functions g with $E[g(Y)]^2 < \infty\}$. Each predictor is assessed by the "mean squared prediction error" $E[X-g(Y)]^2$. We now show that $E(X|Y)$ is the best predictor of X in the sense that

$$E[X - E(X|Y)]^2 = \min_{g \in \aleph} E[X - g(Y)]^2. \tag{1.40}$$

First, it follows from Proposition 1.12(viii) that $E(X|Y) \in \aleph$. Next, for any $g \in \aleph$,

$$\begin{aligned}
E[X - g(Y)]^2 &= E[X - E(X|Y) + E(X|Y) - g(Y)]^2 \\
&= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
&\quad + 2E[X - E(X|Y)][E(X|Y) - g(Y)] \\
&= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
&\quad + 2E\big\{E\{[X - E(X|Y)][E(X|Y) - g(Y)]|Y\}\big\} \\
&= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
&\quad + 2E\{[E(X|Y) - g(Y)]E[X - E(X|Y)|Y]\} \\
&= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
&\ge E[X - E(X|Y)]^2,
\end{aligned}$$

where the third equality follows from Proposition 1.12(iv); the fourth equality follows from Proposition 1.12(vi); and the last equality follows from Proposition 1.2(iii) and (vi). ∎

1.4.2 Independence

Definition 1.7. Let $(\Omega, \mathcal{F}, P)$ be a probability space.
(i) Events A_i, $i = 1, ..., n$, are said to be *independent* if and only if for *any* subset $\{i_1, i_2, ..., i_k\}$ of $\{1, ..., n\}$,

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}).$$

Events in an infinite (countable or uncountable) collection $\mathcal{C}$ are said to be independent if and only if events in each finite subcollection of $\mathcal{C}$ are independent.
(ii) Collections $\mathcal{C}_i \subset \mathcal{F}$, $i \in \mathcal{I}$ (an index set that can be uncountable), are said to be independent if and only if events in any collection of the form $\{A_i \in \mathcal{C}_i : i \in \mathcal{I}\}$ are independent.
(iii) Random vectors X_i, $i \in \mathcal{I}$, are said to be independent if and only if $\sigma(X_i)$, $i \in \mathcal{I}$, are independent. ∎

It is easy to see from Definition 1.7 that if X and Y are independent random vectors, then so are $g(X)$ and $h(Y)$, where g and h are Borel functions.

The following result is useful for checking the independence of several σ-fields.

Proposition 1.13. Let $\mathcal{C}_i$, $i \in \mathcal{I}$, be independent collections of events. Suppose that each $\mathcal{C}_i$ has the property that if $A \in \mathcal{C}_i$ and $B \in \mathcal{C}_i$, then $A \cap B \in \mathcal{C}_i$. Then $\sigma(\mathcal{C}_i)$, $i \in \mathcal{I}$, are independent. ∎

An immediate application of Proposition 1.13 is to show (exercise) that random variables X_i, $i = 1, ..., n$, are independent if and only if (1.26) holds. Hence, Definition 1.7 agrees with the concept of independence discussed in §1.3.1.

Independent random variables can be obtained using product measures introduced in §1.1.2. Let P_i be a probability measure on $(\mathcal{R}, \mathcal{B})$, $i = 1, ..., k$. Then any random vector whose law is the product measure $P_1 \times \cdots \times P_k$ on the space $(\mathcal{R}^k, \mathcal{B}^k)$ has independent components. If F_i is the c.d.f. of P_i, then the joint c.d.f. of $P_1 \times \cdots \times P_k$ is $F(x_1, ..., x_k) = F_1(x_1) \cdots F_k(x_k)$. Consequently, by Fubini's theorem, we obtain that if $X_1, ..., X_n$ are independent random variables and $E|X_1 \cdots X_n| < \infty$, then

$$E(X_1 \cdots X_n) = EX_1 \cdots EX_n. \tag{1.41}$$

When $n = 2$, this implies that if $X_1, ..., X_n$ are independent, then X_i and X_j are uncorrelated, $i \neq j$.

For two events A and B with $P(A) > 0$, A and B are independent if and only if $P(B|A) = P(B)$. This means that A provides no information about B. The following result is a useful extension.

Proposition 1.14. Let X, Y_1, and Y_2 be random variables with $E|X| < \infty$. Suppose that (X, Y_1) and Y_2 are independent. Then

$$E[X|(Y_1, Y_2)] = E(X|Y_1) \text{ a.s.}$$

Proof. First, $E(X|Y_1)$ is Borel on $(\Omega, \sigma(Y_1, Y_2))$, since $\sigma(Y_1) \subset \sigma(Y_1, Y_2)$. Next, we need to show that for any Borel set $B \subset \mathcal{R}^2$,

$$\int_{(Y_1,Y_2)^{-1}(B)} X dP = \int_{(Y_1,Y_2)^{-1}(B)} E(X|Y_1) dP. \tag{1.42}$$

If $B = B_1 \times B_2$, where B_i's are Borel sets in $\mathcal{R}$, then $(Y_1, Y_2)^{-1}(B) = Y_1^{-1}(B_1) \cap Y_2^{-1}(B_2)$ and

$$\begin{aligned}
\int_{Y_1^{-1}(B_1)\cap Y_2^{-1}(B_2)} E(X|Y_1) dP &= \int I_{Y_1^{-1}(B_1)} I_{Y_2^{-1}(B_2)} E(X|Y_1) dP \\
&= \int I_{Y_1^{-1}(B_1)} E(X|Y_1) dP \int I_{Y_2^{-1}(B_2)} dP \\
&= \int I_{Y_1^{-1}(B_1)} X dP \int I_{Y_2^{-1}(B_2)} dP \\
&= \int I_{Y_1^{-1}(B_1)} I_{Y_2^{-1}(B_2)} X dP \\
&= \int_{Y_1^{-1}(B_1)\cap Y_2^{-1}(B_2)} X dP,
\end{aligned}$$

where the second and the next to last equalities follow from result (1.41) and the independence of (X, Y_1) and Y_2; and the third equality follows from the fact that $E(X|Y_1)$ is the conditional expectation of X given Y_1. This shows that (1.42) holds for $B = B_1 \times B_2$. Let $\mathcal{H}$ be the collection of subsets of $\mathcal{R}^2$ for which (1.42) holds. Then we can show that $\mathcal{H}$ is a σ-field. Since we have shown that $\mathcal{B} \times \mathcal{B} \subset \mathcal{H}$, $\mathcal{B}^2 = \sigma(\mathcal{B} \times \mathcal{B}) \subset \mathcal{H}$ and thus the result follows. ∎

Let $X_1, ..., X_k$ be random variables. If X_i and X_j are independent for every pair $i \neq j$, then $X_1, ..., X_k$ are said to be *pairwise independent*. If $X_1, ..., X_k$ are independent, then clearly they are pairwise independent. However, the converse is not true. The following is an example.

Example 1.20. Let X_1 and X_2 be independent random variables each assuming the values 1 and -1 with probability 0.5, and $X_3 = X_1X_2$. Let $A_i = \{X_i = 1\}$, $i = 1, 2, 3$. Then $P(A_i) = 0.5$ for any i and $P(A_1)P(A_2)P(A_3) = 0.125$. However, $P(A_1 \cap A_2 \cap A_3) = P(A_1 \cap A_2) = P(A_1)P(A_2) = 0.25$. Hence X_1, X_2, X_3 are not independent. We now show that X_1, X_2, X_3 are pairwise independent. It is enough to show that X_1 and X_3 are independent. Let $B_i = \{X_i = -1\}$, $i = 1, 2, 3$. Note that $A_1 \cap A_3 = A_1 \cap A_2$, $A_1 \cap B_3 = A_1 \cap B_2$, $B_1 \cap A_3 = B_1 \cap B_2$, and $B_1 \cap B_3 = B_1 \cap A_2$. Then the result follows from the fact that $P(A_i) = P(B_i) = 0.5$ for any i and X_1 and X_2 are independent. ∎

1.4.3 Conditional distributions

The conditional p.d.f. was introduced in §1.4.1 for random variables having p.d.f.'s w.r.t. some measures. We now consider *conditional distributions* in general cases where we may not have any p.d.f.

Theorem 1.7 (Existence of conditional distributions). Let X be a random n-vector on a probability space $(\Omega, \mathcal{F}, P)$ and Y be measurable from $(\Omega, \mathcal{F}, P)$ to $(\Lambda, \mathcal{G})$. Then there exists $P_{X|Y}(A|y)$ such that

(a) $P_{X|Y}(A|y) = P[X^{-1}(A)|Y = y]$ a.s. P_Y for any fixed $A \in \mathcal{B}^n$, and

(b) $P_{X|Y}(\cdot|y)$ is a probability measure on $(\mathcal{R}^n, \mathcal{B}^n)$ for any fixed $y \in \Lambda$.

Furthermore, if $E|g(X, Y)| < \infty$, then

$$E[g(X, Y)|Y = y] = \int_{\mathcal{R}^n} g(x, y) dP_{X|Y}(x|y) \text{ a.s. } P_Y. \quad ∎$$

For a fixed y, $P_{X|Y=y} = P_{X|Y}(\cdot|y)$ is called the conditional distribution of X given $Y = y$. Under the conditions in Theorem 1.7, if Y is a random m-vector and (X, Y) has a p.d.f. w.r.t. $\nu \times \lambda$ (ν and λ are σ-finite measures on $(\mathcal{R}^n, \mathcal{B}^n)$ and $(\mathcal{R}^m, \mathcal{B}^m)$, respectively), then $f_{X|Y}(x|y)$ defined in (1.39) is the p.d.f. of $P_{X|Y=y}$ w.r.t. ν for any fixed y.

Given a collection of conditional distributions $\{P_{X|Y=y} : y \in \Lambda\}$ and a distribution P_Y on Λ, we can construct a joint distribution, as the following result states.

Proposition 1.15. Let $\mathcal{B}^n$ be the Borel σ-field on $\mathcal{R}^n$ and $(\Lambda, \mathcal{G}, P_2)$ be a probability space. Suppose that P_1 is a function from $\mathcal{B}^n \times \Lambda$ to $\mathcal{R}$ and satisfies

(a) $P_1(\cdot, y)$ is a probability measure on $(\mathcal{R}^n, \mathcal{B}^n)$ for any $y \in \Lambda$, and

(b) $P_1(B, \cdot)$ is Borel for any $B \in \mathcal{B}^n$.

Then there is a unique probability measure P on $(\mathcal{R}^n \times \Lambda, \sigma(\mathcal{B}^n \times \mathcal{G}))$ such

that, for $B \in \mathcal{B}$ and $C \in \mathcal{G}$,

$$P(B \times C) = \int_C P_1(B, y) dP_2(y). \tag{1.43}$$

Furthermore, if $(\Lambda, \mathcal{G}) = (\mathcal{R}^m, \mathcal{B}^m)$, $X(x, y) = x$, and $Y(x, y) = y$ define the coordinate random vectors, then $P_Y = P_2$, $P_{X|Y=y} = P_1(\cdot, y)$, and the probability measure in (1.43) is the joint distribution of (X, Y), which has the following joint c.d.f.:

$$F(x, y) = \int_{(-\infty, y]} P_{X|Y=z}\big((-\infty, x]\big) dP_Y(z), \quad x \in \mathcal{R}^n, y \in \mathcal{R}^m, \tag{1.44}$$

where $(-\infty, a]$ denotes $(-\infty, a_1] \times \cdots \times (-\infty, a_k]$ for $a = (a_1, ..., a_k)$. ∎

Proposition 1.15 is sometimes called the "two-stage experiment theorem" for the following reason. If $Y \in \mathcal{R}^m$ is selected in stage 1 of an experiment according to its marginal distribution $P_Y = P_2$, and X is chosen afterward according to its conditional distribution $P_{X|Y=y} = P_1(\cdot, y)$, then the combined two-stage experiment produces a jointly distributed pair (X, Y) with distribution $P_{(X,Y)}$ given by (1.43). The following is an example.

Example 1.21. A market survey is conducted to study whether a new product is preferred over the product currently available in the market (old product). The survey is conducted by mail. Questionnaires are sent along with the sample products (both new and old) to N customers randomly selected from a population, where N is a positive integer. Each customer is asked to fill out the questionnaire and return it. Responses from customers are either 1 (new is better than old) or 0 (otherwise). Some customers, however, do not return the questionnaires. Let X be the number of ones in the returned questionnaires. What is the distribution of X?

If every customer returns the questionnaire, then (from elementary probability) X has the binomial distribution $Bi(p, N)$ in Table 1.1 (assuming that the population is large enough so that customers respond independently), where $p \in (0, 1)$ is the overall rate of customers who prefer the new product. Now, let Y be the number of customers who respond. Then Y is random. Suppose that customers respond independently with the same probability $\pi \in (0, 1)$. Then P_Y is the binomial distribution $Bi(\pi, N)$. Given $Y = y$ (an integer between 0 and N), $P_{X|Y=y}$ is the binomial distribution $Bi(p, y)$ if $y \geq 1$ and the point mass at 0 if $y = 0$. Using (1.44) and the fact that binomial distributions have p.d.f.'s (Table 1.1) w.r.t counting measure, we obtain that the joint c.d.f. of (X, Y) is

$$F(x, y) = \sum_{k=0}^{y} P_{X|Y=k}\big((-\infty, x]\big) \binom{N}{k} \pi^k (1 - \pi)^{N-k}$$

$$= \sum_{k=0}^{y} \sum_{j=0}^{\min(x,k)} \binom{k}{j} p^j (1-p)^{k-j} \binom{N}{k} \pi^k (1-\pi)^{N-k},$$

for $x = 0, 1, ..., y$, $y = 0, 1, ..., N$. The marginal c.d.f. $F_X(x) = F(x, \infty) = F(x, N)$. The p.d.f. of X w.r.t. counting measure is

$$\begin{aligned} f_X(x) &= \sum_{k=x}^{N} \binom{k}{x} p^x (1-p)^{k-x} \binom{N}{k} \pi^k (1-\pi)^{N-k} \\ &= \binom{N}{x} (\pi p)^x (1-\pi p)^{N-x} \sum_{k=x}^{N} \binom{N-x}{k-x} \left(\frac{\pi - \pi p}{1-\pi p}\right)^{k-x} \left(\frac{1-\pi}{1-\pi p}\right)^{N-k} \\ &= \binom{N}{x} (\pi p)^x (1-\pi p)^{N-x} \end{aligned}$$

for $x = 0, 1, ..., N$. It turns out that the marginal distribution of X is the binomial distribution $Bi(\pi p, N)$. ∎

1.5 Asymptotic Theorems

Asymptotic theory studies limiting behavior of random variables (vectors) and their distributions. It is an important tool for statistical analysis. A more complete coverage of asymptotic theory in statistical analysis can be found in Serfling (1980) and Sen and Singer (1993).

1.5.1 Convergence modes and stochastic orders

There are several convergence modes for random variables/vectors. For any k-vector $c \in \mathcal{R}^k$, $\|c\|$ denotes the usual distance between 0 and c, i.e., $\|c\|^2 = cc^\tau$.

Definition 1.8. Let $X, X_1, X_2, \ldots$ be random k-vectors defined on a probability space.
(i) We say that the sequence $\{X_n\}$ converges to X almost surely (a.s.) and write $X_n \to_{a.s.} X$ if and only if

$$P\left(\lim_{n\to\infty} \|X_n - X\| = 0\right) = 1.$$

(ii) We say that $\{X_n\}$ converges to X in probability and write $X_n \to_p X$ if and only if, for every fixed $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\|X_n - X\| > \epsilon\right) = 0.$$

(iii) We say that $\{X_n\}$ converges to X in L_p (or in pth-moment) and write $X_n \to_{L_p} X$ if and only if

$$\lim_{n\to\infty} E\|X_n - X\|^p = 0,$$

where $p > 0$ is a fixed constant.
(iv) Let F_{X_n} be the c.d.f. of X_n, $n = 1, 2,...$, and F_X be the c.d.f. of X. We say that $\{X_n\}$ converges to X in distribution (or in law) and write $X_n \to_d X$ if and only if, for each continuity point x of F_X,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x). \quad \blacksquare$$

The a.s. convergence in Definition 1.8(i) is almost the same as the pointwise convergence of functions in calculus. The concept of convergence in probability, convergence in L_p, or a.s. convergence represents a sense in which, for n sufficiently large, X_n and X approximate each other as functions on the original probability space. The concept of convergence in distribution in Definition 1.8(iv), however, depends only on the distributions F_{X_n} and F_X and does not necessitate that X_n and X are close in any sense; in fact, Definition 1.8(iv) still makes sense even if X and X_n's are not defined on the same probability space. In Definition 1.8(iv), it is *not* required that $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for every x. However, if $F_X(x)$ is a continuous function, then we have the following stronger result.

Proposition 1.16 (Pólya's theorem). If $X_n \to_d X$ and F_X is continuous, then

$$\lim_{n\to\infty} \sup_x |F_{X_n}(x) - F_X(x)| = 0. \quad \blacksquare$$

The following result describes the relationship among four convergence modes.

Theorem 1.8. Let $X, X_1, X_2, \ldots$ be random k-vectors.
(i) If $X_n \to_{a.s.} X$, then $X_n \to_p X$.
(ii) If $X_n \to_{L_p} X$ for a $p > 0$, then $X_n \to_p X$.
(iii) If $X_n \to_p X$, then $X_n \to_d X$.
(iv) If $X_n \to_d X$, then there are random vectors $Y, Y_1, Y_2, ...$ defined on a probability space such that $P_Y = P_X$, $P_{Y_n} = P_{X_n}$, $n = 1, 2,...$, and $Y_n \to_{a.s.} Y$.
(v) If, for every $\epsilon > 0$,

$$\sum_{n=1}^{\infty} P(\|X_n - X\| \geq \epsilon) < \infty,$$

then $X_n \to_{a.s.} X$.
(vi) If $X_n \to_p X$, then there is a subsequence $\{X_{n_k}, k = 1, 2, ...\}$ such that $X_{n_k} \to_{a.s.} X$ as $k \to \infty$.
(vii) Suppose that $X_n \to_d X$. Then, for any $r > 0$,

$$\lim_{n\to\infty} E\|X_n\|^r = E\|X\|^r < \infty$$

if and only if $\{\|X_n\|^r\}$ is *uniformly integrable* in the sense that

$$\lim_{t\to\infty} \sup_n E\left[\|X_n\|^r I_{(t,\infty)}(\|X_n\|)\right] = 0. \quad \blacksquare \tag{1.45}$$

The converse of Theorem 1.8(i), (ii), or (iii) is generally not true (see Example 1.22 and Exercise 75). Note that part (iv) of Theorem 1.8 is not a converse of part (i), but it is an important result in probability theory. Part (v) of Theorem 1.8 indicates that the converse of part (i) is true under the additional condition that $P(\|X_n - X\| \geq \epsilon)$ tends to 0 fast enough. A consequence of Theorem 1.8(vii) is that if $X_n \to_p X$ and $\{\|X_n - X\|^p\}$ is uniformly integrable, then $X_n \to_{L_p} X$; i.e., the converse of Theorem 1.8(ii) is true under the additional condition of uniform integrability. A useful sufficient condition for uniform integrability of $\{\|X_n\|^r\}$ is that

$$\sup_n E\|X_n\|^{r+\delta} < \infty \tag{1.46}$$

for a $\delta > 0$.

Example 1.22. Let $\theta_n = 1 + n^{-1}$ and X_n be a random variable having the exponential distribution $E(0, \theta_n)$ (Table 1.2), $n = 1, 2, ...$ Let X be a random variable having the exponential distribution $E(0, 1)$. For any $x > 0$,

$$F_{X_n}(x) = 1 - e^{-x/\theta_n} \to 1 - e^{-x} = F_X(x)$$

as $n \to \infty$. Since $F_{X_n}(x) \equiv 0 \equiv F_X(x)$ for $x \leq 0$, we have shown that $X_n \to_d X$.

Is it true that $X_n \to_p X$? This question cannot be answered without any further information about the random variables X and X_n. We consider two cases in which different answers can be obtained. First, suppose that $X_n \equiv \theta_n X$ (then X_n has the given c.d.f.). Note that $X_n - X = (\theta_n - 1)X = n^{-1}X$, which has the c.d.f. $(1 - e^{-nx})I_{[0,\infty)}(x)$. Hence

$$P\left(|X_n - X| \geq \epsilon\right) = e^{-n\epsilon} \to 0$$

for any $\epsilon > 0$. In fact, by Theorem 1.8(iv), $X_n \to_{a.s.} X$; since $E|X_n - X|^p = n^{-p}EX^p < \infty$ for any $p > 0$, $X_n \to_{L_p} X$ for any $p > 0$. Next, suppose

that X_n and X are independent random variables. Using result (1.28) and the fact that the p.d.f.'s for X_n and $-X$ are $\theta_n^{-1}e^{-x/\theta_n}I_{(0,\infty)}(x)$ and $e^x I_{(-\infty,0)}(x)$, respectively, we obtain that

$$P\left(|X_n - X| \leq \epsilon\right) = \int_{-\epsilon}^{\epsilon}\int \theta_n^{-1}e^{-x/\theta_n}e^{y-x}I_{(0,\infty)}(x)I_{(-\infty,x)}(y)dxdy,$$

which converges to (by the dominated convergence theorem)

$$\int_{-\epsilon}^{\epsilon}\int e^{-x}e^{y-x}I_{(0,\infty)}(x)I_{(-\infty,x)}(y)dxdy = 1 - e^{-\epsilon}.$$

Thus, $P\left(|X_n - X| \geq \epsilon\right) \to e^{-\epsilon} > 0$ for any $\epsilon > 0$ and, therefore, $\{X_n\}$ does not converge to X in probability. ∎

The following result gives some useful sufficient and necessary conditions for convergence in distribution.

Theorem 1.9. Let $X, X_1, X_2, \ldots$ be random k-vectors.
(i) $X_n \to_d X$ if and only if $\lim_{n\to\infty} E[h(X_n)] = E[h(X)]$ for every bounded continuous function h from $\mathcal{R}^k$ to $\mathcal{R}$.
(ii) (Lévy-Cramér continuity theorem). Let $\phi_X, \phi_{X_1}, \phi_{X_2}, \ldots$ be the ch.f.'s of $X, X_1, X_2, \ldots$, respectively. $X_n \to_d X$ if and only if $\lim_{n\to\infty}\phi_{X_n}(t) = \phi_X(t)$ for all $t \in \mathcal{R}^k$.
(iii) (Cramér-Wold device). $X_n \to_d X$ if and only if $X_n c^\tau \to_d Xc^\tau$ for every $c \in \mathcal{R}^k$. ∎

Examples of applications of Theorem 1.9 are given as exercises in §1.6. The following result can be used to check whether $X_n \to_d X$ when X has a p.d.f. f and X_n has a p.d.f. f_n.

Proposition 1.17 (Scheffé's theorem). Let $\{f_n\}$ be a sequence of p.d.f.'s on $\mathcal{R}^k$ w.r.t. a measure ν. Suppose that $\lim_{n\to\infty} f_n(x) = f(x)$ a.e. ν and $f(x)$ is a p.d.f. w.r.t. ν. Then $\lim_{n\to\infty}\int |f_n(x) - f(x)|d\nu = 0$.
Proof. Let $g_n(x) = [f(x) - f_n(x)]I_{\{f \geq f_n\}}(x)$, $n = 1, 2, \ldots$. Then

$$\int |f_n(x) - f(x)|d\nu = 2\int g_n(x)d\nu.$$

Since $0 \leq g_n(x) \leq f(x)$ for all x and $g_n \to 0$ a.e. ν, the result follows from the dominated convergence theorem. ∎

As an example, consider the Lebesgue p.d.f. f_n of the t-distribution t_n (Table 1.2), $n = 1, 2, \ldots$. One can show (exercise) that $f_n \to f$, where f is the standard normal p.d.f. This is an important result in statistics.

We now introduce the notion of $O(\cdot)$, $o(\cdot)$, and stochastic $O(\cdot)$ and $o(\cdot)$. In calculus, two sequences of real numbers, $\{a_n\}$ and $\{b_n\}$, satisfy $a_n = O(b_n)$ if and only if $|a_n| \le c|b_n|$ for all n and a constant c; and $a_n = o(b_n)$ if and only if $a_n/b_n \to 0$ as $n \to \infty$.

Definition 1.9. Let $X_1, X_2, ...$ and $Y_1, Y_2, ...$ be random variables defined on a probability space $(\Omega, \mathcal{F}, P)$.
(i) $X_n = O(Y_n)$ a.s. if and only if $X_n(\omega) = O(Y_n(\omega))$ a.s. P.
(ii) $X_n = o(Y_n)$ a.s. if and only if $X_n/Y_n \to_{a.s.} 0$.
(iii) $X_n = O_p(Y_n)$ if and only if, for any $\epsilon > 0$, there is a constant $C_\epsilon > 0$ such that

$$\sup_n P(|X_n| \ge C_\epsilon |Y_n|) < \epsilon.$$

(iv) $X_n = o_p(Y_n)$ if and only if $X_n/Y_n \to_p 0$. ∎

Note that $X_n = o_p(Y_n)$ implies $X_n = O_p(Y_n)$; $X_n = O_p(Y_n)$ and $Y_n = O_p(Z_n)$ implies $X_n = O_p(Z_n)$; but $X_n = O_p(Y_n)$ does not imply $Y_n = O_p(X_n)$. The same conclusion can be obtained if $O_p(\cdot)$ and $o_p(\cdot)$ are replaced by $O(\cdot)$ a.s. and $o(\cdot)$ a.s., respectively. Some results related to O_p are given in Exercise 88. For example, if $X_n \to_d X$ for a random variable X, then $X_n = O_p(1)$. Since $a_n = O(1)$ means that $\{a_n\}$ is bounded, $\{X_n\}$ is said to be bounded in probability if $X_n = O_p(1)$.

1.5.2 Convergence of transformations

Transformation is an important tool in statistics. For random vectors X_n converging to X in some sense, we often want to know whether $g(X_n)$ converges to $g(X)$ in the same sense. The following result provides an answer to most problems.

Theorem 1.10. Let $X, X_1, X_2, ...$ be random k-vectors defined on a probability space and g be a measurable function from $(\mathcal{R}^k, \mathcal{B}^k)$ to $(\mathcal{R}^l, \mathcal{B}^l)$. Suppose that g is continuous a.s. P_X. Then
(i) $X_n \to_{a.s.} X$ implies $g(X_n) \to_{a.s.} g(X)$.
(ii) $X_n \to_p X$ implies $g(X_n) \to_p g(X)$.
(iii) $X_n \to_d X$ implies $g(X_n) \to_d g(X)$. ∎

Example 1.23. (i) Let $X_1, X_2, ...$ be random variables. If $X_n \to_d X$, where X has the $N(0,1)$ distribution, then $X_n^2 \to_d Y$, where Y has the chi-square distribution χ_1^2 (Example 1.13).
(ii) Let (X_n, Y_n) be random 2-vectors satisfying $(X_n, Y_n) \to_d (X, Y)$, where X and Y are independent random variables having the $N(0,1)$ distribution, then $X_n/Y_n \to_d X/Y$, which has the Cauchy distribution $C(0,1)$ (§1.3.1).

(iii) Under the conditions in part (ii), $\max(X_n, Y_n) \to_d \max(X, Y)$, which has the c.d.f. $[\Phi(x)]^2$ ($\Phi(x)$ is the c.d.f. of $N(0,1)$). ∎

In Example 1.23(ii) and (iii), the condition that $(X_n, Y_n) \to_d (X, Y)$ cannot be relaxed to $X_n \to_d X$ and $Y_n \to_d Y$ (exercise); i.e., we need the convergence of the joint c.d.f. of (X_n, Y_n). This is different when $\to_d$ is replaced by $\to_p$ or $\to_{a.s.}$. The following result, which plays an important role in probability and statistics, establishes the convergence in distribution of $X_n + Y_n$ or $X_n Y_n$ when no information regarding the joint c.d.f. of (X_n, Y_n) is provided.

Theorem 1.11 (Slutsky's theorem). Let $X, X_1, X_2, ...,$ $Y_1, Y_2, ...$ be random variables on a probability space. Suppose that $X_n \to_d X$ and $Y_n \to_p c$, where c is a fixed real number. Then
(i) $X_n + Y_n \to_d X + c$;
(ii) $Y_n X_n \to_d cX$;
(iii) $X_n / Y_n \to_d X/c$ if $c \neq 0$.
Proof. We prove (i) only. The proofs of (ii) and (iii) are left as exercises. Let $t \in \mathcal{R}$ and $\epsilon > 0$ be fixed constants. Then

$$\begin{aligned} F_{X_n+Y_n}(t) &= P(X_n + Y_n \leq t) \\ &\leq P(\{X_n + Y_n \leq t\} \cap \{|Y_n - c| < \epsilon\}) + P(|Y_n - c| \geq \epsilon) \\ &\leq P(X_n \leq t - c + \epsilon) + P(|Y_n - c| \geq \epsilon) \end{aligned}$$

and, similarly,

$$F_{X_n+Y_n}(t) \geq P(X_n \leq t - c - \epsilon) - P(|Y_n - c| \geq \epsilon).$$

If $t - c$, $t - c + \epsilon$, and $t - c - \epsilon$ are continuity points of F_X, then it follows from the previous two inequalities and the hypotheses of the theorem that

$$F_X(t - c - \epsilon) \leq \liminf_n F_{X_n+Y_n}(t) \leq \limsup_n F_{X_n+Y_n}(t) \leq F_X(t - c + \epsilon).$$

Since ϵ can be arbitrary (why?),

$$\lim_{n\to\infty} F_{X_n+Y_n}(t) = F_X(t - c).$$

The result follows from $F_{X+c}(t) = F_X(t - c)$. ∎

An application of Theorem 1.11 is given in the proof of the following important result.

Theorem 1.12. Let $X_1, X_2, ...$ and Y be random k-vectors satisfying

$$a_n(X_n - c) \to_d Y, \tag{1.47}$$

where $c \in \mathcal{R}^k$ and $\{a_n\}$ is a sequence of positive numbers with $\lim_{n\to\infty} a_n = \infty$. Let g be a function from $\mathcal{R}^k$ to $\mathcal{R}$.
(i) If g is differentiable at c, then

$$a_n[g(X_n) - g(c)] \to_d \nabla g(c) Y^\tau, \tag{1.48}$$

where $\nabla g(x)$ denotes the k-vector of partial derivatives of g at x.
(ii) Suppose that g has continuous partial derivatives of order $m > 1$ in a neighborhood of c, with all the partial derivatives of order j, $1 \le j \le m-1$, vanishing at c, but with the mth-order partial derivatives not all vanishing at c. Then

$$a_n^m[g(X_n) - g(c)] \to_d \frac{1}{m!} \sum_{i_1=1}^k \cdots \sum_{i_m=1}^k \frac{\partial^m g}{\partial x_{i_1} \cdots \partial x_{i_m}}\bigg|_{x=c} Y_{i_1} \cdots Y_{i_m}, \tag{1.49}$$

where Y_j is the jth component of Y.
Proof. We prove (i) only. The proof of (ii) is similar. Let

$$Z_n = a_n[g(X_n) - g(c)] - a_n \nabla g(c)(X_n - c)^\tau.$$

If we can show that $Z_n = o_p(1)$, then by (1.47), Theorem 1.9(iii), and Theorem 1.11(i), result (1.48) holds.

The differentiability of g at c implies that for any $\epsilon > 0$, there is a $\delta_\epsilon > 0$ such that

$$|g(x) - g(c) - \nabla g(c)(x-c)^\tau| \le \epsilon \|x - c\| \tag{1.50}$$

whenever $\|x - c\| < \delta_\epsilon$. Let $\eta > 0$ be fixed. By (1.50),

$$P(|Z_n| \ge \eta) \le P(\|X_n - c\| \ge \delta_\epsilon) + P(a_n \|X_n - c\| \ge \eta/\epsilon).$$

Since $a_n \to \infty$, (1.47) and Theorem 1.11(ii) imply $X_n \to_p c$. By Theorem 1.10(iii), (1.47) implies $a_n\|X_n - c\| \to_d \|Y\|$. Without loss of generality, we can assume that η/ϵ is a continuity point of $F_{\|Y\|}$. Then

$$\begin{aligned}\limsup_n P(|Z_n| \ge \eta) &\le \lim_{n\to\infty} P(\|X_n - c\| \ge \delta_\epsilon) \\ &\quad + \lim_{n\to\infty} P(a_n\|X_n - c\| \ge \eta/\epsilon) \\ &= P(\|Y\| \ge \eta/\epsilon).\end{aligned}$$

The proof is complete since ϵ can be arbitrary. ∎

In statistics, we often need a nondegenerated limiting distribution of $a_n[g(X_n) - g(c)]$ so that probabilities involving $a_n[g(X_n) - g(c)]$ can be approximated by the c.d.f. of $\nabla g(c) Y^\tau$, if (1.48) holds. Hence, result (1.48) is not useful for this purpose if $\nabla g(c) = 0$, and in such cases result (1.49) may be applied.

A useful method in statistics, called the *delta-method*, is based on the following corollary of Theorem 1.12. Recall that $N(\mu, \sigma^2)$ denotes a random variable having the $N(\mu, \sigma^2)$ distribution.

Corollary 1.1. Assume the conditions of Theorem 1.12. If Y has the $N_k(0, \Sigma)$ distribution, then

$$a_n[g(X_n) - g(c)] \to_d N\left(0, \nabla g(c)\Sigma[\nabla g(c)]^\tau\right). \quad \blacksquare$$

Example 1.24. Let $X_1, X_2, ...$ be random variables such that

$$\sqrt{n}(X_n - c) \to_d N(0, 1).$$

Consider the function $g(x) = x^2$. If $c \neq 0$, then an application of Corollary 1.1 gives that

$$\sqrt{n}(X_n^2 - c^2) \to_d N(0, 4c^2).$$

If $c = 0$, the first-order derivative of g at 0 is 0 but the second-order derivative of $g \equiv 2$. Hence, an application of result (1.49) gives that

$$nX_n^2 \to_d [N(0, 1)]^2,$$

which has the chi-square distribution χ_1^2 (Example 1.13). The last result can also be obtained by applying Theorem 1.10(iii). $\blacksquare$

1.5.3 The law of large numbers

The law of large numbers concerns the limiting behavior of sums of independent random variables. The weak law of large numbers (WLLN) refers to convergence in probability, whereas the strong law of large numbers (SLLN) refers to a.s. convergence. Our first result gives the WLLN and SLLN for a sequence of independent and identically distributed (i.i.d.) random variables.

Theorem 1.13. Let $X_1, X_2, ...$ be i.i.d. random variables having a c.d.f. F.
(i) (The WLLN). The existence of a sequence of real numbers $\{a_n\}$ for which

$$\frac{1}{n}\sum_{i=1}^{n} X_i - a_n \to_p 0$$

if and only if

$$\lim_{x\to\infty} x[1 - F(x) + F(-x)] = 0,$$

in which case we may take $a_n = E[X_1 I_{(-n,n)}(X_1)]$.
(ii) (The SLLN). The existence of a constant c for which

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to_{a.s.} c$$

if and only if $E|X_1| < \infty$, in which case $c = EX_1$ and

$$\frac{1}{n}\sum_{i=1}^{n} c_i(X_i - EX_1) \to_{a.s.} 0$$

for any bounded sequence of real numbers $\{c_i\}$.
(iii) (The Marcinkiewicz SLLN). If $E|X_1|^\delta < \infty$ for a $\delta \in (0,1)$, then

$$\frac{1}{n^{1/\delta}}\sum_{i=1}^{n} |X_i| \to_{a.s.} 0. \quad \blacksquare$$

The proof of this theorem can be found in Billingsley (1986) or Chung (1974). The next result is for sequences of independent but not necessarily identically distributed random variables.

Theorem 1.14. Let $X_1, X_2, ...$ be independent random variables with finite expectations.
(i) (The SLLN). If

$$\sum_{i=1}^{\infty} \frac{\mathrm{Var}(X_i)}{i^2} < \infty,$$

then

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - EX_i) \to_{a.s.} 0.$$

(ii) (The WLLN). If there is a $\delta \in (0,1)$ such that

$$\lim_{n\to\infty} \frac{1}{n^{1+\delta}}\sum_{i=1}^{n} E|X_i|^{1+\delta} = 0,$$

then

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - EX_i) \to_{p} 0. \quad \blacksquare$$

The WLLN and SLLN have many applications in probability and statistics. The following is an example. Other examples can be found in later chapters.

Example 1.25. Let f and g be continuous functions on $[0,1]$ satisfying $0 \le f(x) \le Cg(x)$ for all x, where $C > 0$ is a constant. We now show that

$$\lim_{n\to\infty} \int_0^1 \int_0^1 \cdots \int_0^1 \frac{\sum_{i=1}^n f(x_i)}{\sum_{i=1}^n g(x_i)} dx_1 dx_2 \cdots dx_n = \frac{\int_0^1 f(x)dx}{\int_0^1 g(x)dx} \tag{1.51}$$

(assuming that $\int_0^1 g(x)dx \neq 0$). Let $X_1, X_2, ...$ be i.i.d. random variables having the uniform distribution on $[0,1]$. By Proposition 1.7(i), $E[f(X_1)] = \int_0^1 f(x)dx < \infty$ and $E[g(X_1)] = \int_0^1 g(x)dx < \infty$. By the SLLN (Theorem 1.13(i)),

$$\frac{1}{n}\sum_{i=1}^n f(X_i) \to_{a.s.} E[f(X_1)]$$

and the same result holds when f is replaced by g. By Theorem 1.10(i),

$$\frac{\sum_{i=1}^n f(X_i)}{\sum_{i=1}^n g(X_i)} \to_{a.s.} \frac{E[f(X_1)]}{E[g(X_1)]}. \tag{1.52}$$

Since the random variable on the left-hand side of (1.52) is bounded by C, result (1.51) follows from the dominated convergence theorem and the fact that the left-hand side of (1.51) is the expectation of the random variable on the left-hand side of (1.52). ∎

1.5.4 The central limit theorem

The WLLN and SLLN may not be useful in approximating the distributions of (normalized) sums of independent random variables. We need to use the *Central Limit Theorem* (CLT), which plays a fundamental role in statistical asymptotic theory.

Theorem 1.15 (Lindeberg's CLT). Let $\{X_{nj}, j = 1, ..., k_n\}$ be independent random variables with $0 < \sigma_n^2 = \mathrm{Var}(\sum_{j=1}^{k_n} X_{nj}) < \infty$, $n = 1, 2, ...$, and $k_n \to \infty$ as $n \to \infty$. If

$$\lim_{n\to\infty} \frac{1}{\sigma_n^2} \sum_{j=1}^{k_n} E\left[(X_{nj} - EX_{nj})^2 I_{\{|X_{nj}-EX_{nj}|>\epsilon\sigma_n\}}\right] = 0 \tag{1.53}$$

for any $\epsilon > 0$, then

$$\frac{1}{\sigma_n}\sum_{j=1}^{k_n}(X_{nj} - EX_{nj}) \to_d N(0,1). \quad ∎$$

The proof of this theorem can be found in Billingsley (1986) or Chung (1974). Condition (1.53) with an arbitrary $\epsilon > 0$ is called Lindeberg's condition. It is implied by the following condition, called Liapunov's condition, which is somewhat easier to verify:

$$\sum_{j=1}^{k_n} E|X_{nj} - EX_{nj}|^{2+\delta} = o(\sigma_n^{2+\delta}) \tag{1.54}$$

for some $\delta > 0$.

Example 1.26. Let $X_1, X_2, ...$ be independent random variables. Suppose that X_i has the binomial distribution $Bi(p_i, 1)$, $i = 1, 2,...$, and that $\sigma_n^2 = \sum_{i=1}^n \mathrm{Var}(X_i) = \sum_{i=1}^n p_i(1-p_i) \to \infty$ as $n \to \infty$. For each i, $EX_i = p_i$ and

$$E|X_i - EX_i|^3 = (1-p_i)^3 p_i + p_i^3(1-p_i) \le 2p_i(1-p_i).$$

Hence

$$\sum_{i=1}^n E|X_i - EX_i|^3 \le 2\sigma_n^2,$$

i.e., Liapunov's condition (1.54) holds with $\delta = 1$. Thus, by Theorem 1.15,

$$\frac{1}{\sigma_n}\sum_{i=1}^n (X_i - p_i) \to_d N(0,1). \quad \blacksquare$$

The following are useful corollaries of Theorem 1.15 (and Theorem 1.9(iii)).

Corollary 1.2 (Multivariate CLT). Let $X_1, ..., X_n$ be i.i.d. random k-vectors with a finite $\Sigma = \mathrm{Var}(X_1)$. Then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - EX_1) \to_d N_k(0, \Sigma). \quad \blacksquare$$

Corollary 1.3. Let $X_{ni} \in \mathcal{R}^{m_i}$, $i = 1, ..., k_n$, be independent random vectors with $m_i \le m$ (a fixed integer), $n = 1, 2,...$, $k_n \to \infty$ as $n \to \infty$, and $\inf_{i,n} \lambda_-[\mathrm{Var}(X_{ni})] > 0$, where $\lambda_-[A]$ is the smallest eigenvalue of A. Let $c_{ni} \in \mathcal{R}^{m_i}$ be vectors such that

$$\lim_{n\to\infty}\left(\max_{1\le i\le k_n} \|c_{ni}\|^2 \Big/ \sum_{i=1}^{k_n} \|c_{ni}\|^2\right) = 0.$$

(i) Suppose that $\sup_{i,n} E\|X_{ni}\|^{2+\delta} < \infty$ for some $\delta > 0$. Then

$$\sum_{i=1}^{k_n}(X_{ni} - EX_{ni})c_{ni}^{\tau} \Big/ \left[\sum_{i=1}^{k_n} \mathrm{Var}(X_{ni}c_{ni}^{\tau})\right]^{1/2} \to_d N(0,1). \tag{1.55}$$

(ii) Suppose that X_{ni}, $i = 1, ..., k_n$, $n = 1, 2, ...$, have a common distribution and $E\|X_{11}\|^2 < \infty$. Then (1.55) holds. ∎

Applications of these corollaries can be found in later chapters. More results on the CLT can be found, for example, in Serfling (1980) and Shorack and Wellner (1986).

Let Y_n be a sequence of random variables, $\{\mu_n\}$ and $\{\sigma_n\}$ be sequences of real numbers such that $\sigma_n > 0$ for all n and $(Y_n - \mu_n)/\sigma_n \to_d N(0,1)$. Then, by Proposition 1.16,

$$\lim_{n\to\infty} \sup_x |F_{(Y_n-\mu_n)/\sigma_n}(x) - \Phi(x)| = 0, \tag{1.56}$$

where Φ is the c.d.f. of $N(0,1)$. This implies that for any sequence of real numbers $\{c_n\}$,

$$\lim_{n\to\infty} \left| P\left(Y_n \le c_n\right) - \Phi\left(\frac{c_n - \mu_n}{\sigma_n}\right) \right| = 0,$$

i.e., $P(Y_n \le c_n)$ can be approximated by $\Phi\left(\frac{c_n-\mu_n}{\sigma_n}\right)$, regardless of whether $\{c_n\}$ has a limit. Since $\Phi\left(\frac{t-\mu_n}{\sigma_n}\right)$ is the c.d.f. of $N(\mu_n, \sigma_n^2)$, Y_n is said to be *asymptotically distributed* as $N(\mu_n, \sigma_n^2)$ or simply *asymptotically normal*. For example, $\sum_{i=1}^{k_n} c_{ni}X_{ni}^{\tau}$ in Corollary 1.3 is asymptotically normal. This can be extended to random vectors. For example, $\sum_{i=1}^{n} X_i$ in Corollary 1.2 is asymptotically distributed as $N_k(EX_1, \Sigma/n)$.

1.6 Exercises

1. Let A and B be two nonempty proper subsets of a sample space Ω, $A \neq B$ and $A \cap B \neq \emptyset$. Obtain $\sigma(\{A, B\})$, the smallest σ-field containing A and B.

2. Let $\mathcal{C}$ be a collection of subsets of Ω and let $\Gamma = \{\mathcal{F} : \mathcal{F}$ is a σ-field on Ω and $\mathcal{C} \subset \mathcal{F}\}$. Show that $\Gamma \neq \emptyset$ and $\sigma(\mathcal{C}) = \cap\{\mathcal{F} : \mathcal{F} \in \Gamma\}$.

3. Show that if $\mathcal{C}_1 \subset \mathcal{C}_2$, then $\sigma(\mathcal{C}_1) \subset \sigma(\mathcal{C}_2)$.

4. Let $\mathcal{C}$ be the collection of intervals of the form $(a, b]$, where $-\infty < a < b < \infty$, and let $\mathcal{D}$ be the collection of closed sets on $\mathcal{R}$. Show that $\mathcal{B} = \sigma(\mathcal{C}) = \sigma(\mathcal{D})$, where $\mathcal{B}$ is the Borel σ-field on $\mathcal{R}$.

5. Let $(\Omega, \mathcal{F})$ be a measurable space and $C \in \mathcal{F}$. Show that $\mathcal{F}_C = \{C \cap A : A \in \mathcal{F}\}$ is a σ-field on C.

6. Prove part (ii) and part (iii) of Proposition 1.1.

7. Let ν_i, $i = 1, 2, ...$, be measures on $(\Omega, \mathcal{F})$ and a_i, $i = 1, 2, ...$, be positive numbers. Show that $a_1\nu_1 + a_2\nu_2 + \cdots$ is a measure on $(\Omega, \mathcal{F})$.

8. Let $A_1, A_2, ...$ be a sequence of measurable sets and P be a probability measure. Define

$$\limsup_n A_n = \bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i \quad \text{and} \quad \liminf_n A_n = \bigcup_{n=1}^{\infty} \bigcap_{i=n}^{\infty} A_i.$$

(a) Show that

$$P\left(\liminf_n A_n\right) \leq \liminf_n P(A_n)$$

and

$$\limsup_n P(A_n) \leq P\left(\limsup_n A_n\right).$$

(b) (Borel-Cantelli's lemma). Show that if $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(\limsup_n A_n) = 0$.
(c) (Borel-Cantelli's lemma). Show that if $A_1, A_2, ...$ are independent (Definition 1.7) and $\sum_{n=1}^{\infty} P(A_n) = \infty$, then $P(\limsup_n A_n) = 1$.

9. Prove part (i) of Proposition 1.2.

10. Let $F(x_1, ..., x_k)$ be a c.d.f on $\mathcal{R}^k$. Show that
(a) $F(x_1, ..., x_{k-1}, x_k) \leq F(x_1, ..., x_{k-1}, x'_k)$ if $x_k \leq x'_k$.
(b) $\lim_{x_i \to -\infty} F(x_1, ..., x_k) = 0$ for any $1 \leq i \leq k$.
(c) $F(x_1, ..., x_{k-1}, \infty) = \lim_{x_k \to \infty} F(x_1, ..., x_{k-1}, x_k)$ is a c.d.f on $\mathcal{R}^{k-1}$.

11. Let $(\Omega_i, \mathcal{F}_i) = (\mathcal{R}, \mathcal{B})$, $i = 1, ..., k$. Show that the product σ-field $\sigma(\mathcal{F}_1 \times \cdots \times \mathcal{F}_k)$ is the same as the Borel σ-field generated by $\mathcal{O}$, all open sets in $\mathcal{R}^k$.

12. Let ν and λ be two measures on $(\Omega, \mathcal{F})$ such that $\nu(A) = \lambda(A)$ for any $A \in \mathcal{C}$, where $\mathcal{C} \subset \mathcal{F}$ is a collection having the property that if A and B are in $\mathcal{C}$, then so is $A \cap B$. Assume that there are $A_i \in \mathcal{C}$, $i = 1, 2, ...$, such that $\cup A_i = \Omega$ and $\nu(A_i) < \infty$ for all i. Show that $\nu(A) = \lambda(A)$ for any $A \in \sigma(\mathcal{C})$. This proves the uniqueness part of Proposition 1.3. (Hint: show that $\{A \in \sigma(\mathcal{C}) : \nu(A) = \lambda(A)\}$ is a σ-field.)

13. Show that $f^{-1}(B^c) = (f^{-1}(B))^c$ and $f^{-1}(\cup B_i) = \cup f^{-1}(B_i)$.

14. Show that $f^{-1}(\mathcal{G})$ is a σ-field, if f is a function from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$.

15. Prove parts (i)-(iv) of Proposition 1.4.

16. Show that a monotone function from $\mathcal{R}$ to $\mathcal{R}$ is Borel.

17. Let f be a nonnegative Borel function on $(\Omega, \mathcal{F})$. Show that f is the limit of a sequence of simple functions $\{\varphi_n\}$ satisfying $0 \leq \varphi_1 \leq \varphi_2 \leq \cdots \leq f$.

18. Let f be a function from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ and $A_1, A_2, ...$ be disjoint events in $\mathcal{F}$ such that $\cup A_i = \Omega$. Let f_n be a function from $(A_n, \mathcal{F}_{A_n})$ to $(\Lambda, \mathcal{G})$ such that $f_n(\omega) = f(\omega)$ for any $\omega \in A_n$, $n = 1, 2, ...$ Show that f is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ if and only if f_n is measurable from $(A_n, \mathcal{F}_{A_n})$ to $(\Lambda, \mathcal{G})$ for each n.

19. Let $\prod_{i=1}^k (\Omega_i, \mathcal{F}_i)$ be a product measurable space and π_i be the ith *projection*, i.e., π_i is a function from $\Omega_1 \times \cdots \times \Omega_k$ to Ω_i such that $\pi_i(\omega_1, ..., \omega_k) = \omega_i$, $\omega_i \in \Omega_i$, $i = 1, ..., k$. Show that π_i's are measurable.

20. Let f be a Borel function on $\mathcal{R}^2$. Define a function g from $\mathcal{R}$ to $\mathcal{R}$ as $g(x) = f(x, y)$, where y is a fixed point in $\mathcal{R}$. Show that g is Borel. Is it true that f is Borel from $\mathcal{R}^2$ to $\mathcal{R}$ if $f(x, y)$ with any fixed y or fixed x is Borel from $\mathcal{R}$ to $\mathcal{R}$?

21. Show that the set function defined in (1.8) is a measure.

22. Prove (1.13) in Example 1.5.

23. Prove Proposition 1.5.

24. Prove Proposition 1.6(i).

25. (Chebyshev's inequality). Let X be a random variable and ϕ a strictly positive and increasing function on $(0, \infty)$ satisfying $\phi(-t) = \phi(t)$. Show that for each $t > 0$,

$$P(|X| \geq t) \leq \frac{E\phi(X)}{\phi(t)}.$$

26. Let ν_i, $i = 1, 2$, be measures on $(\Omega, \mathcal{F})$ and f be Borel. Show that

$$\int f d(\nu_1 + \nu_2) = \int f d\nu_1 + \int f d\nu_2,$$

i.e., if either side of the equality is well defined, then so is the other side, and the two sides are equal.

27. Let f be an integrable Borel function on $(\Omega, \mathcal{F}, \nu)$. Show that for each $\epsilon > 0$, there is a δ_ϵ such that $\nu(A) < \delta_\epsilon$ and $A \in \mathcal{F}$ imply $\int_A |f| d\nu < \epsilon$.

28. Consider Example 1.9. Show that (1.16) does not hold for the following function:

$$f(i,j) = \begin{cases} 1 & i = j \\ -1 & i = j-1 \\ 0 & \text{otherwise.} \end{cases}$$

Does this contradict Fubini's theorem?

29. Let f be a nonnegative Borel function on $(\Omega, \mathcal{F}, \nu)$ with a σ-finite ν. Let m be the Lebesgue measure on $(\mathcal{R}, \mathcal{B})$ and

$$A = \{(\omega, x) \in \Omega \times \mathcal{R} : 0 \le x \le f(\omega)\}.$$

Show that $A \in \sigma(\mathcal{F} \times \mathcal{B})$ and $\int_\Omega f d\nu = \nu \times m(A)$.

30. For any c.d.f. F and any $a \ge 0$, show that

$$\int [F(x+a) - F(x)]dx = a.$$

31. (Integration by parts). Let F and G be two c.d.f.'s on $\mathcal{R}$. Show that if F and G have no common points of discontinuity in the interval $[a, b]$, then

$$\int_{(a,b]} G(x)dF(x) = F(b)G(b) - F(a)G(a) - \int_{(a,b]} F(x)dG(x).$$

32. Let f be a Borel function on $\mathcal{R}^2$ such that $f(x,y) = 0$ for each $x \in \mathcal{R}$ and $y \notin C_x$, where $m(C_x) = 0$ for each x and m is the Lebesgue measure. Show that $f(x,y) = 0$ for each $y \notin C$ and $x \notin B_y$, where $m(C) = 0$ and $m(B_y) = 0$ for each $y \notin C$.

33. Show that the set function defined by (1.17) is a measure.

34. Consider Example 1.11. Show that if (1.20) holds, then $P(A) = \int_A f(x)dx$ for any Borel set A. (Hint: $\mathcal{A} = \{A : P(A) = \int_A f(x)dx\}$ is a σ-field containing all sets of the form $(-\infty, x]$.)

35. Prove Proposition 1.7.

36. Let F_i be a c.d.f. having a Lebesgue p.d.f. f_i, $i = 1, 2$. Assume that there is a $c \in \mathcal{R}$ such that $F_1(c) < F_2(c)$. Define

$$F(x) = \begin{cases} F_1(x) & -\infty < x < c \\ F_2(x) & c \le x < \infty. \end{cases}$$

Show that the probability measure P corresponding to F satisfies $P \ll m + \delta_c$, where m is the Lebesgue measure and δ_c is the point mass at c, and find the p.d.f. of F w.r.t. $m + \delta_c$.

37. Let X be a random variable having the Lebesgue p.d.f. $\frac{2x}{\pi^2} I_{(0,\pi)}(x)$. Derive the p.d.f. of $Y = \sin X$.

38. Let X be a random variable having a continuous c.d.f. F. Show that $Y = F(X)$ has the uniform distribution $U(0,1)$ (Table 1.2).

39. Let U be a random variable having the uniform distribution $U(0,1)$ and let F be a c.d.f. Show that the c.d.f. of $Y = F^{-1}(U)$ is F, where $F^{-1}(t) = \inf\{x \in \mathcal{R} : F(x) \geq t\}$.

40. Let X_i, $i = 1, 2, 3$, be independent random variables having the same Lebesgue p.d.f. $f(x) = e^{-x} I_{(0,\infty)}(x)$. Obtain the joint Lebesgue p.d.f. of (Y_1, Y_2, Y_3), where $Y_1 = X_1 + X_2 + X_3$, $Y_2 = X_1/(X_1 + X_2)$, and $Y_3 = (X_1 + X_2)/(X_1 + X_2 + X_3)$. Are the Y_i independent?

41. Let X_1 and X_2 be independent random variables having the standard normal distribution. Obtain the joint Lebesgue p.d.f. of (Y_1, Y_2), where $Y_1 = \sqrt{X_1^2 + X_2^2}$ and $Y_2 = X_1/X_2$. Are the Y_i independent?

42. Let X_1 and X_2 be independent random variables and $Y = X_1 + X_2$. Show that $F_Y(y) = \int F_{X_2}(y - x) dF_{X_1}(x)$.

43. Let X be a random variable and $a > 0$. Show that $E|X|^a < \infty$ if and only if $\sum_{n=1}^{\infty} n^{a-1} P(|X| \geq n) < \infty$.

44. Let X be a random variable with range $\{0, 1, 2, ...\}$. Show that if $EX < \infty$, then
$$EX = \sum_{n=1}^{\infty} P(X \geq n).$$

45. Let X be a random variable having a c.d.f. F_X. Show that if $X \geq 0$ a.s., then
$$EX = \int [1 - F_X(x)] dx;$$
in general, if EX exists, then
$$EX = \int_0^{\infty} [1 - F_X(x)] dx - \int_{-\infty}^{0} F_X(x) dx.$$

46. (Jensen's inequality). Let X be a random variable and f be a convex function on $\mathcal{R}$, i.e., f satisfies
$$f\left(\sum_{i=1}^{k} a_i x_i\right) \leq \sum_{i=1}^{k} a_i f(x_i), \quad x_i \in \mathcal{R}$$

for every set of positive $a_1, ..., a_k$ with $\sum_{i=1}^k a_i = 1$. Suppose that $E|X| < \infty$ and $E|f(X)| < \infty$. Show that $f(EX) \leq Ef(X)$. (Hint: consider nonnegative simple functions first and use the fact that f is continuous.)

47. Show that $EX = \mu$ and $\mathrm{Var}(X) = \Sigma$ for the $N_k(\mu, \Sigma)$ distribution.

48. Let X be a random variable with $EX^2 < \infty$ and let $Y = |X|$. Suppose that X has a p.d.f. symmetric about 0. Show that X and Y are uncorrelated, but they are not independent.

49. Let (X, Y) be a random 2-vector with the following Lebesgue p.d.f.:

$$f(x, y) = \begin{cases} \pi^{-1} & x^2 + y^2 \leq 1 \\ 0 & x^2 + y^2 > 1. \end{cases}$$

Show that X and Y are uncorrelated, but are not independent.

50. Let $X_1, ..., X_k$ be independent random variables and $Y = X_1 + \cdots + X_k$. Prove the following statements, using Proposition 1.10.
(a) If X_i has the binomial distribution $Bi(p, n_i)$, $i = 1, ..., k$, then Y has the binomial distribution $Bi(p, n_1 + \cdots + n_k)$.
(b) If X_i has the Poisson distribution $P(\theta_i)$, $i = 1, ..., k$, then Y has the Poisson distribution $P(\theta_1 + \cdots + \theta_k)$.
(c) If X_i has the negative binomial distribution $NB(p, r_i)$, $i = 1, ..., k$, then Y has the negative binomial distribution $NB(p, r_1 + \cdots + r_k)$.
(d) If X_i has the exponential distribution $E(0, \theta)$, $i = 1, ..., k$, then Y has the gamma distribution $\Gamma(k, \theta)$.
(e) If X_i has the Cauchy distribution $C(0, 1)$, $i = 1, ..., k$, then Y/k has the same distribution as X_1.

51. Show the following properties of the multivariate normal distribution $N_k(\mu, \Sigma)$.
(a) The m.g.f. of $N_k(\mu, \Sigma)$ is $e^{\mu t^\tau + t\Sigma t^\tau/2}$.
(b) Let X be a random k-vector having the $N_k(\mu, \Sigma)$ distribution and $Y = XA + c$, where A is a $k \times l$ matrix of rank $l \leq k$ and $c \in \mathcal{R}^l$. Then Y has the $N_l(\mu A + c, A^\tau \Sigma A)$ distribution.
(c) A random k-vector X has a k-dimensional normal distribution if and only if for any $c \in \mathcal{R}^k$, Xc^τ has a univariate normal distribution.
(d) Let X be a random k-vector having the $N_k(\mu, \Sigma)$ distribution. Let A be a $k \times l$ matrix and B be a $k \times m$ matrix. Then XA and XB are independent if and only if they are uncorrelated.
(e) Let (X_1, X_2) be a random k-vector having the $N_k(\mu, \Sigma)$ distribution with

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where X_1 is a random l-vector and Σ_{11} is an $l \times l$ matrix. Then the conditional p.d.f. of X_2 given $X_1 = x_1$ is

$$N_{k-l}\left(\mu_2 + (x_1 - \mu_1)\Sigma_{11}^{-1}\Sigma_{12}, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right),$$

where $\mu_i = EX_i$, $i = 1, 2$. (Hint: consider $X_2 - \mu_2 - (X_1 - \mu_1)\Sigma_{11}^{-1}\Sigma_{12}$ and $X_1 - \mu_1$.)

52. Let ϕ_n be the ch.f. of a probability measure P_n, $n = 1, 2, ...$ Let $\{a_n\}$ be a sequence of nonnegative numbers with $\sum_{n=1}^{\infty} a_n = 1$. Show that $\sum_{n=1}^{\infty} a_n \phi_n$ is a ch.f. and find its corresponding probability measure.

53. Let Y be a random variable having the noncentral chi-square distribution $\chi_k^2(\delta)$. Show that
(a) Y has the Lebesgue p.d.f. given by (1.31);
(b) the ch.f. of Y is $(1 - 2\sqrt{-1}t)^{k/2} e^{\sqrt{-1}\delta t/(1-2\sqrt{-1}t)}$;
(c) $E(Y) = k + \delta$ and $\mathrm{Var}(Y) = 2k + 4\delta$.

54. Let T be a random variable having the noncentral t-distribution $t_n(\delta)$. Show that
(a) T has the Lebesgue p.d.f. given by (1.32);
(b) $E(T) = \delta\Gamma((n-1)/2)\sqrt{n/2}/\Gamma(n/2)$ when $n > 1$;
(c) $\mathrm{Var}(T) = \frac{n(1+\delta^2)}{n-2} - \frac{\delta^2 n}{2}\left[\frac{\Gamma((n-1)/2)}{\Gamma(n/2)}\right]^2$ when $n > 2$.

55. Let $\mathtt{F}$ be a random variable having the noncentral F-distribution $F_{n_1,n_2}(\delta)$. Show that
(a) $\mathtt{F}$ has the Lebesgue p.d.f. given by (1.33);
(b) $E(\mathtt{F}) = \frac{n_2(n_1+\delta)}{n_1(n_2-2)}$ when $n_2 > 2$;
(c) $\mathrm{Var}(\mathtt{F}) = \frac{2n_2^2[(n_1+\delta)^2+(n_2-2)(n_1+2\delta)]}{n_1^2(n_2-2)^2(n_2-4)}$ when $n_2 > 4$.

56. Let $X = N_n(\mu, I_n)$. Apply Cochran's theorem (Theorem 1.5) to show that if $A^2 = A$, then XAX^τ has the noncentral chi-square distribution $\chi_r^2(\delta)$, where $r =$ rank of A and $\delta = \mu A \mu^\tau$.

57. Let $X_1, ..., X_n$ be independent and $X_i = N(0, \sigma_i^2)$, $i = 1, ..., n$. Let $\tilde{X} = \sum_{i=1}^n \sigma_i^{-2} X_i / \sum_{i=1}^n \sigma_i^{-2}$ and $\tilde{S}^2 = \sum_{i=1}^n \sigma_i^{-2}(X_i - \tilde{X})^2$. Apply Cochran's theorem to show that $\tilde{X}^2$ and $\tilde{S}^2$ are independent and that $\tilde{S}^2$ has the chi-square distribution χ_{n-1}^2.

58. Let $X = N_n(\mu, I_n)$ and $A_i^2 = A_i$, $i = 1, 2$. Show that a necessary and sufficient condition that XA_1X^τ and XA_2X^τ are independent is $A_1A_2 = 0$.

59. Prove Theorem 1.6. (Hint: first consider simple functions.)

60. Prove Proposition 1.12.

61. Let X and Y be random variables on $(\Omega, \mathcal{F}, P)$ and $\mathcal{A} \subset \mathcal{F}$ be a σ-field. Suppose that X is integrable and Y is bounded. Show that $E[YE(X|\mathcal{A})] = E[XE(Y|\mathcal{A})]$.

62. Let (X, Y) be a random 2-vector having a Lebesgue p.d.f. $f(x, y)$. Suppose that $E|X| < \infty$ and $Z = X + Y$. Show that

$$E(X|Z) = \frac{\int xf(x, Z - x)dx}{\int f(x, Z - x)dx}$$

without using Proposition 1.11.

63. (Convergence theorems for conditional expectations). Let $X_1, X_2, ...$ and X be integrable random variables on $(\Omega, \mathcal{F}, P)$ and $\mathcal{A} \subset \mathcal{F}$ be a σ-field. Show that
(a) (Fatou's lemma). If $X_n \geq 0$ for any n, then

$$E\left(\liminf_n X_n | \mathcal{A}\right) \leq \liminf_n E(X_n|\mathcal{A}) \quad \text{a.s.}$$

(b) (Monotone convergence theorem). If $0 \leq X_1 \leq X_2 \leq \cdots$ and $\lim_{n\to\infty} X_n = X$ a.s., then

$$E(X|\mathcal{A}) = \lim_{n\to\infty} E(X_n|\mathcal{A}) \quad \text{a.s.}$$

(c) (Dominated convergence theorem). Suppose that there is an integrable random variable Y such that $|X_n| \leq Y$ for any n and $\lim_{n\to\infty} X_n = X$ a.s. Then the result in (b) holds.

64. Let X be a nonnegative integrable random variable on $(\Omega, \mathcal{F})$ and $\mathcal{A} \subset \mathcal{F}$ be a σ-field. Show that

$$E(X|\mathcal{A}) = \int_0^\infty P(X > t|\mathcal{A})dt.$$

65. Let X be an integrable random variable on $(\Omega, \mathcal{F}, P)$, $\mathcal{A} \subset \mathcal{F}$ be a σ-field, and f be a convex function on $\mathcal{R}$. Show that $f(E(X|\mathcal{A})) \leq E[f(X)|\mathcal{A}]$ a.s.

66. Show that two events A and B are independent if and only if two random variables I_A and I_B are independent.

67. Show that random variables X_i, $i = 1, ..., n$, are independent according to Definition 1.7 if and only if (1.26) holds.

68. Show that a random variable X is independent of itself if and only if X is constant a.s. Can X and $f(X)$ be independent for a Borel f?

69. Let X and Y be independent random variables on a probability space. Show that if $E|X|^a < \infty$ for some $a \geq 1$ and $E|Y| < \infty$, then $E|X+Y|^a \geq E|X+EY|^a$.

70. Let (X,Y) be a random 2-vector with the following Lebesgue p.d.f.:

$$f(x,y) = \begin{cases} 8xy & 0 \leq x \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find the marginal p.d.f.'s of X and Y. Are X and Y independent?

71. Let (X,Y,Z) be a random 3-vector with the following Lebesgue p.d.f.:

$$f(x,y,z) = \begin{cases} \frac{1-\sin x \sin y \sin z}{8\pi^3} & 0 \leq x \leq 2\pi, 0 \leq y \leq 2\pi, 0 \leq z \leq 2\pi \\ 0 & \text{otherwise.} \end{cases}$$

Show that X, Y, and Z are not independent, but are pairwise independent.

72. Let P_Y be a discrete distribution on $\{0,1,2,...\}$ and $P_{X|Y=y}$ be the binomial distribution $Bi(p,y)$. Let (X,Y) be the random vector having the joint c.d.f. given by (1.44). Show that
(a) if Y has the Poisson distribution $P(\theta)$, then the marginal distribution of X is the Poisson distribution $P(p\theta)$;
(b) if $Y+r$ has the negative binomial distribution $NB(\pi,r)$, then the marginal distribution of $X+r$ is the negative binomial distribution $NB(\pi/[1-(1-p)(1-\pi)],r)$.

73. Let $X, X_1, X_2, ...$ be random vectors on a probability space. Show that $X_n \to_{a.s.} X$ if and only if for every $\epsilon > 0$,

$$\lim_{m\to\infty} P(\|X_n - X\| \leq \epsilon \text{ for all } n \geq m) = 1.$$

74. Let $X_1, X_2, ...$ be a sequence of identically distributed random variables and $Y_n = n^{-1}\max_{i\leq n}|X_i|$. Show that $Y_n \to_{a.s.} 0$ and $Y_n \to_{L_1} 0$.

75. Let $X, X_1, X_2, ...$ be random variables. Find an example for each of the following cases:
(a) $X_n \to_p X$, but $\{X_n\}$ does not converge to X a.s.
(b) $X_n \to_p X$, but $\{X_n\}$ does not converge to X in L_p for any $p > 0$.
(c) $X_n \to_d X$, but $\{X_n\}$ does not converge to X in probability (do not use Example 1.22).
(d) $X_n \to_p X$, but $\{g(X_n)\}$ does not converge to $g(X)$ in probability for some function g.

76. Show that (1.46) implies (1.45).

77. Let $X, X_1, X_2, ...$ be random k-vectors satisfying $P(\|X_n\| \geq c) \leq P(\|X\| \geq c)$ for all n and $c > 0$. Show that if $E\|X\| < \infty$, then $\{\|X_n\|\}$ is uniformly integrable.

78. Let $X_1, X_2, ...$ and $Y_1, Y_2, ...$ be random variables. Show that
(a) if $\{|X_n|\}$ and $\{|Y_n|\}$ are uniformly integrable, then $\{|X_n + Y_n|\}$ is uniformly integrable;
(b) if $\{|X_n|\}$ is uniformly integrable, then $\{|n^{-1}\sum_{i=1}^n X_i|\}$ is uniformly integrable.

79. Let $X, Y, X_1, X_2, ...$ be random k-vectors satisfying $X_n \to_p X$ and $P(\|X_n\| \leq \|Y\|) = 1$ for all n. Show that if $E\|Y\|^p < \infty$, then $X_n \to_{L_p} X$.

80. Show that if $X_n \to_d X$ and $X = c$ a.s., where $c \in \mathcal{R}^k$, then $X_n \to_p X$.

81. Show that if $X_n \to_d X$, then for every $\epsilon > 0$, there exists $M_\epsilon > 0$ such that $P(\|X_n\| > M_\epsilon) < \epsilon$.

82. Let X_n, Y_n, $n = 1, 2, ...$ be random k-vectors such that

$$\lim_{n\to\infty} P\left(\|X_n - Y_n\| \geq \epsilon\right) = 0$$

for any $\epsilon > 0$. Show that if $X_n \to_d X$ for a random vector X, then $Y_n \to_d X$.

83. Let $X_1, X_2, ..., X, Y$ be random k-vectors. Show that $X_n \to_p X$ and $X_n \to_p Y$ implies that $P(X = Y) = 1$.

84. Let $X, X_1, X_2, ...$ be random k-vectors and $Y, Y_1, Y_2, ...$ be random l-vectors. Suppose that $X_n \to_d X$, $Y_n \to_d Y$, and X_n and Y_n are independent for each n. Show that (X_n, Y_n) converges in distribution to a random $(k + l)$-vector.

85. Let X_n be a random variable having the $N(\mu_n, \sigma_n^2)$ distribution, $n = 1, 2,...$, and X be a random variable having the $N(\mu, \sigma^2)$ distribution. Show that $X_n \to_d X$ if and only if $\mu_n \to \mu$ and $\sigma_n \to \sigma$.

86. Suppose that X_n is a random variable having the binomial distribution $Bi(p_n, n)$. Show that if $np_n \to \theta > 0$, then $X_n \to_d X$, where X has the Poisson distribution $P(\theta)$.

87. Let f_n be the Lebesgue p.d.f. of the t-distribution t_n, $n = 1, 2,....$ Show that $f_n(x) \to f(x)$ for any $x \in \mathcal{R}$, where f is the Lebesgue p.d.f. of the standard normal distribution.

88. Let $X_1, X_2, ...,\ Y_1, Y_2, ...,\ Z_1, Z_2, ...$ be random variables. Prove the following statements.
(a) If $X_n \to_d X$ for a random variable X, then $X_n = O_p(1)$.
(b) If $X_n = O_p(Z_n)$ and $P(Y_n = 0) = 0$ for all n, then $X_n Y_n = O_p(Y_n Z_n)$.
(c) If $X_n = O_p(Z_n)$ and $Y_n = O_p(Z_n)$, then $X_n + Y_n = O_p(Z_n)$.
(d) If $E|X_n| = O(a_n)$ for a sequence of positive numbers $a_1, a_2, ...$, then $X_n = O_p(a_n)$.

89. Let $X, X_1, X_2, ...$ be random variables such that $X_n \to_{a.s.} X$. Show that $\sup_n |X_n| = O_p(1)$.

90. Prove Theorem 1.10.

91. Show by example that $X_n \to_d X$ and $Y_n \to_d Y$ does not necessarily imply that $g(X_n, Y_n) \to_d g(X, Y)$, where g is a continuous function on $\mathcal{R}^2$.

92. Prove Theorem 1.11(ii)-(iii) and Theorem 1.12(ii).

93. Let $U_1, U_2, ...$ be i.i.d. random variables having the uniform distribution on $[0, 1]$ and $Y_n = \left(\prod_{i=1}^n U_i\right)^{-1/n}$. Show that $\sqrt{n}(Y_n - e) \to_d N(0, e^2)$.

94. Let X_n be a random variable having the Poisson distribution $P(n\theta)$, where $\theta > 0$ and $n = 1, 2, ...$ Show that $X_n/n \to_{a.s.} \theta$. Show that the same conclusion can be drawn if X_n has the binomial distribution $Bi(\theta, n)$.

95. Let $X_1, ..., X_n$ be i.i.d. random variables with

$$P(X_1 = \pm x) = \left(\sum_{x=3}^{\infty} \frac{1}{x^2 \log x}\right)^{-1} \frac{1}{2x^2 \log x}, \quad x = 3, 4, ...$$

Show that $E|X_1| = \infty$ but $n^{-1} \sum_{i=1}^n X_i \to_p 0$, using Theorem 1.13(i).

96. Let $X_1, ..., X_n$ be i.i.d. random variables with $\mathrm{Var}(X_1) < \infty$. Show that

$$\frac{2}{n(n+1)} \sum_{j=1}^n j X_j \to_p EX_1.$$

97. Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables such that

$$P(X_n \le t, Y_n \ge t + \epsilon) + P(X_n \ge t + \epsilon, Y_n \le t) = o(1)$$

for any fixed $t \in \mathcal{R}$ and $\epsilon > 0$. Show that $X_n - Y_n = o_p(1)$.

98. Show that Liapunov's condition (1.54) implies Lindeberg's condition, i.e., condition (1.53) with arbitrary $\epsilon > 0$.

99. Prove Corollaries 1.2 and 1.3.

100. Let X_n be a random variable having the Poisson distribution $P(n\theta)$, where $\theta > 0$, $n = 1, 2, ...$ Show that $(X_n - n\theta)/\sqrt{n\theta} \to_d N(0, 1)$.

101. Let $X_1, ..., X_n$ be random variables and $\{\mu_n\}$, $\{\sigma_n\}$, $\{a_n\}$, and $\{b_n\}$ be sequences of real numbers with $\sigma_n \geq 0$ and $a_n \geq 0$. Suppose that X_n is asymptotically distributed as $N(\mu_n, \sigma_n^2)$. Show that $a_n X_n + b_n$ is asymptotically distributed as $N(\mu_n, \sigma_n^2)$ if and only if $a_n \to 1$ and $[\mu_n(a_n - 1) + b_n]/\sigma_n \to 0$.

102. Let $X_1, X_2, ...$ be independent random variables such that X_j has the uniform distribution on $[-j, j]$, $j = 1, 2, ...$ Show that Lindeberg's condition is satisfied and state the resulting CLT.

103. Let $X_1, X_2, ...$ be independent random variables with $P(X_j = \pm\sqrt{j}\,)$ $= 0.5$, $j = 1, 2, ...$ Can we apply Theorem 1.15 to $\{X_j\}$ by checking Liapunov's condition (1.54)?

104. Let $X_1, X_2, ...$ be independent random variables with $P(X_j = -j^a) = P(X_j = j^a) = P(X_j = 0) = 1/3$, where $a > 0$, $j = 1, 2, ...$ Can we apply Theorem 1.15 to $\{X_j\}$ by checking Liapunov's condition (1.54)?

105. Let $X_1, X_2, ...$ be independent random variables such that for $j = 1, 2, ...,$

$$P(X_j = \pm j^a) = \frac{1}{6j^{2(a-1)}} \quad \text{and} \quad P(X_j = 0) = 1 - \frac{1}{3j^{2(a-1)}},$$

where $a > 1$ is a constant. Show that Lindeberg's condition is satisfied if and only if $a < 1.5$.

106. Suppose that X_n is a random variable having the binomial distribution $Bi(\theta, n)$, where $0 < \theta < 1$, $n = 1, 2, ...$ Define

$$Y_n = \begin{cases} \log(X_n/n) & X_n \geq 1 \\ 1 & X_n = 0. \end{cases}$$

Show that $Y_n \to_{a.s.} \log\theta$ and $\sqrt{n}(Y_n - \log\theta) \to_d N\left(0, \frac{1-\theta}{\theta}\right)$.

Chapter 2

Fundamentals of Statistics

This chapter discusses some fundamental concepts of mathematical statistics. These concepts are essential for the material in later chapters.

2.1 Populations, Samples, and Models

A typical statistical problem can be described as follows. One or a series of random experiments is performed that results in some data; our task is to extract the information from the data and interpret the results. In this book we do not consider the problem of planning experiments and collecting data, but concentrate on statistical analysis of the data, assuming that the data are given.

A descriptive data analysis can be performed to obtain some summary measures of the data, such as the mean, median, range, standard deviation, etc., and some graphical displays, such as the histogram and box-and-whisker diagram, etc. (see, e.g., Hogg and Tanis (1993)). Although this kind of analysis is simple and requires almost no assumptions, it may not allow us to gain enough insight into the problem. We focus on more sophisticated methods of analyzing data: *statistical inference* and *decision theory*.

2.1.1 Populations and samples

In statistical inference and decision theory, the data set is viewed as a realization or observation of a random element defined on a probability space $(\Omega, \mathcal{F}, P)$ related to the random experiment. The probability measure P is called the *population*. The data set or the random element that produces

the data is called a *sample* from P. The size of the data set is called the *sample size*. A population P is *known* if and only if $P(A)$ is a known value for every event A. In a statistical problem, the population P is at least partially unknown and we would like to deduce some properties of P based on the available sample.

Example 2.1 (Measurement problems). To measure an unknown quantity θ (for example, a distance, weight, or temperature), n measurements, $x_1, ..., x_n$, are taken in an experiment of measuring θ. If θ can be measured without errors, then $x_i = \theta$ for all i; otherwise, each x_i has a possible measurement error. In descriptive data analysis, a few summary measures may be calculated, for example, the *sample mean*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and the *sample variance*

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

However, what is the relationship between $\bar{x}$ and θ? Are they close (if not equal) in some sense? The sample variance s^2 is clearly an average of squared deviations of x_i's from their mean. But, what kind of information does s^2 provide? Finally, is it enough to just look at $\bar{x}$ and s^2 for the purpose of measuring θ? These questions cannot be answered in descriptive data analysis.

In statistical inference and decision theory, the data set, $(x_1, ..., x_n)$, is viewed as an outcome of the experiment whose sample space is $\Omega = \mathcal{R}^n$. We usually assume that the n measurements are obtained in n *independent* trials of the experiment. Hence, we can define a random n-vector $X = (X_1, ..., X_n)$ on $\prod_{i=1}^{n}(\mathcal{R}, \mathcal{B}, P)$ whose realization is $(x_1, ..., x_n)$. The population in this problem is P (note that the product probability measure is determined by P) and is at least partially unknown. The random vector X is a sample and n is the sample size. Define

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.1}$$

and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 . \tag{2.2}$$

Then $\bar{X}$ and S^2 are random variables that produce $\bar{x}$ and s^2, respectively. Questions raised previously can be answered if some assumptions are imposed on the population P, which are discussed later. ∎

When the sample $(X_1, ..., X_n)$ has i.i.d. components, which is often the case in applications, the population is determined by the marginal distribution of X_i.

Example 2.2 (Life-time testing problems). Let $x_1, ..., x_n$ be observed life-times of some electronic components. Again, in statistical inference and decision theory, $x_1, ..., x_n$ are viewed as realizations of independent random variables $X_1, ..., X_n$. Suppose that the components are of the same type so that it is reasonable to assume that $X_1, ..., X_n$ have a common marginal c.d.f. F. Then the population is F, which is often unknown. A quantity of interest in this problem is $1 - F(t)$ with a $t > 0$, which is the probability that a component does not fail at time t. It is possible that all x_i's are smaller (or larger) than t. Conclusions about $1 - F(t)$ can be drawn based on data $x_1, ..., x_n$ when certain assumptions on F are imposed. ∎

Example 2.3 (Survey problems). A survey is often conducted when one is not able to evaluate all elements in a collection $\mathcal{P} = \{y_1, ..., y_N\}$ containing N values in $\mathcal{R}^k$, where k and N are finite positive integers but N may be very large. Suppose that the quantity of interest is the *population total* $Y = \sum_{i=1}^N y_i$. In a survey, a subset $\boldsymbol{s}$ of n elements are selected from $\mathcal{P}$ and values y_i, $i \in \boldsymbol{s}$, are obtained. Can we draw some conclusion about Y based on data y_i, $i \in \boldsymbol{s}$?

How do we define some random variables that produce the survey data? First, we need to specify how $\boldsymbol{s}$ is selected. A commonly used probability sampling plan can be described as follows. Assume that every element in $\mathcal{P}$ can be selected at most once, i.e., we consider *sampling without replacement.* The sample space Ω is the collection of all subsets of n distinct elements from $\mathcal{P}$. Let $\mathcal{F}$ be the collection of all subsets of Ω and p be a probability measure on $(\Omega, \mathcal{F})$. Any $\boldsymbol{s} \in \Omega$ is selected with probability $p(\boldsymbol{s})$. Note that $p(\boldsymbol{s})$ is a known value whenever $\boldsymbol{s}$ is given. Let $X_1, ..., X_n$ be random variables such that

$$P(X_1 = y_{i_1}, ..., X_n = y_{i_n}) = p(\boldsymbol{s}), \qquad \boldsymbol{s} = \{i_1, ..., i_n\} \in \Omega. \tag{2.3}$$

Then $(y_i, i \in \boldsymbol{s})$ can be viewed as a realization of the sample $(X_1, ..., X_n)$. If $p(\boldsymbol{s})$ is constant, then the sampling plan is called the *simple random sampling (without replacement)* and $(X_1, ..., X_n)$ is called a *simple random sample.* Although $X_1, ..., X_n$ are identically distributed, they are *not* necessarily independent. Thus, unlike in the previous two examples, the population in this problem may not be specified by the marginal distributions of X_i's. The population is determined by $\mathcal{P}$ and the known selection probability measure p. For this reason, $\mathcal{P}$ is often treated as the population. Conclusions about Y and other characteristics of $\mathcal{P}$ can be drawn based on data y_i, $i \in \boldsymbol{s}$, which are discussed later. ∎

2.1.2 Parametric and nonparametric models

A *statistical model* (a set of assumptions) on the population P in a given problem is often postulated to make the analysis possible or easy. Although testing the correctness of postulated models is part of statistical inference and decision theory, postulated models are often based on knowledge of the problem under consideration.

Definition 2.1. A set of probability measures P_θ on $(\Omega, \mathcal{F})$ indexed by a *parameter* $\theta \in \Theta$ is said to be a *parametric family* if and only if $\Theta \subset \mathcal{R}^d$ for some fixed positive integer d *and* each P_θ is a *known* probability measure when θ is known. The set Θ is called the *parameter space* and d is called its *dimension*. ∎

A *parametric model* refers to the assumption that the population P is in a parametric family. A parametric family $\{P_\theta : \theta \in \Theta\}$ is said to be *identifiable* if and only if $\theta_1 \neq \theta_2$ and $\theta_i \in \Theta$ imply $P_{\theta_1} \neq P_{\theta_2}$. In most cases an identifiable parametric family can be obtained through reparameterization. Hence, we assume in what follows that every parametric family is identifiable.

Let $\mathcal{P}$ be a family of populations and ν be a σ-finite measure on $(\Omega, \mathcal{F})$. If $P \ll \nu$ for all $P \in \mathcal{P}$, then $\mathcal{P}$ is said to be dominated by ν, in which case $\mathcal{P}$ can be identified by the family of densities $\{\frac{dP}{d\nu} : P \in \mathcal{P}\}$ (or $\{\frac{dP_\theta}{d\nu} : \theta \in \Theta\}$ for a parametric family).

Many examples of parametric families can be obtained from Tables 1.1 and 1.2 in §1.3.1. All parametric families from Tables 1.1 and 1.2 are dominated by the counting measure or the Lebesgue measure on $\mathcal{R}$.

Example 2.4 (The k-dimensional normal family). Consider the normal distribution $N_k(\mu, \Sigma)$ given by (1.25) for a fixed positive integer k. An important parametric family in statistics is the family of normal distributions

$$\mathcal{P} = \{N_k(\mu, \Sigma) : \ \mu \in \mathcal{R}^k, \ \Sigma \in \mathcal{M}_k\},$$

where $\mathcal{M}_k$ is the collection of all $k \times k$ symmetric positive definite matrices. This family is dominated by the Lebesgue measure on $\mathcal{R}^k$.

In the measurement problem described in Example 2.1, X_i's are often i.i.d. from the $N(\mu, \sigma^2)$ distribution. Hence we can impose a parametric model on the population, i.e., $\mathcal{P} = \{N(\mu, \sigma^2) : \ \mu \in \mathcal{R}, \ \sigma^2 > 0\}$.

The normal parametric model is perhaps not a good model for the life-time testing problem described in Example 2.2, since clearly $X_i \geq 0$ for all i. In practice, the normal family $\{N(\mu, \sigma^2) : \ \mu \in \mathcal{R}, \ \sigma^2 > 0\}$ can be used for a life-time testing problem if one puts some restrictions on μ and σ so that $P(X_i < 0)$ is negligible. Common parametric models for

life-time testing problems are the exponential model (containing the exponential distributions $E(0,\theta)$ with an unknown parameter θ; see Table 1.2 in §1.3.1), the gamma model (containing the gamma distributions $\Gamma(\alpha,\gamma)$ with unknown parameters α and γ), the log-normal model (containing the log-normal distributions $LN(\mu,\sigma^2)$ with unknown parameters μ and σ), the Weibull model (containing the Weibull distributions $W(\alpha,\theta)$ with unknown parameters α and θ), and any subfamilies of these parametric families (e.g., a family containing the gamma distributions with one known parameter and one unknown parameter).

The normal family is often not a good choice for the survey problem discussed in Example 2.3. ∎

In a given problem, a parametric model is not useful if the dimension of Θ is very high. For example, the survey problem described in Example 2.3 has a natural parametric model, since the population $\mathcal{P}$ can be indexed by the parameter $\theta = (y_1, ..., y_N)$. If there is no restriction on the y-values, however, the dimension of the parameter space is kN, which is usually much larger than the sample size n. If there are some restrictions on the y-values, for example, y_i's are nonnegative integers no larger than a fixed integer m, then the dimension of the parameter space is at most $m+1$ and the parametric model becomes useful.

A family of probability measures is said to be *nonparametric* if it is not parametric according to Definition 2.1. A *nonparametric model* refers to the assumption that the population P is in a nonparametric family. There may be almost no assumption on a nonparametric family, for example, the family of all probability measures on $(\mathcal{R}^k, \mathcal{B}^k)$. But in many applications we may use one or a combination of the following assumptions for a nonparametric family on $(\mathcal{R}^k, \mathcal{B}^k)$:

(1) The joint c.d.f.'s are continuous.

(2) The joint c.d.f.'s have finite moments of order $\leq$ a fixed integer.

(3) The joint c.d.f.'s have p.d.f.'s (e.g., Lebesgue p.d.f.'s).

(4) $k = 1$ and the c.d.f.'s are symmetric.

For instance, in Example 2.1, we may assume a nonparametric model with symmetric and continuous c.d.f.'s. The symmetry assumption may not be suitable for the population in Example 2.2, but the continuity assumption seems to be reasonable.

In statistical inference and decision theory, methods designed for parametric models are called *parametric methods*, whereas methods designed for nonparametric models are called *nonparametric methods*. However, nonparametric methods are used in a parametric model when parametric methods are not effective, such as when the dimension of the parameter space is

too high (Example 2.3). On the other hand, parametric methods may be applied to a nonparametric model when the quantity of interest is not the entire population but a vector of real-valued characteristics (parameters) of the population. Examples are provided later.

2.1.3 Exponential and location-scale families

In this section we discuss two types of parametric families that are of special importance in statistical inference and decision theory.

Definition 2.2 (Exponential families). A parametric family $\{P_\theta : \theta \in \Theta\}$ dominated by a σ-finite measure ν on $(\Omega, \mathcal{F})$ is called an *exponential family* if and only if

$$\frac{dP_\theta}{d\nu}(\omega) = \exp\bigl\{T(\omega)[\eta(\theta)]^\tau - \xi(\theta)\bigr\}h(\omega), \quad \omega \in \Omega, \tag{2.4}$$

where $\exp\{x\} = e^x$ is the exponential function, T is a random p-vector with a fixed positive integer p, η is a function from Θ to $\mathcal{R}^p$, h is a nonnegative Borel function on $(\Omega, \mathcal{F})$, and $\xi(\theta) = \log\bigl\{\int_\Omega e^{T(\omega)[\eta(\theta)]^\tau} h(\omega)d\nu(\omega)\bigr\}$. ∎

In Definition 2.2, T and h are functions of ω only, whereas η and ξ are functions of θ only. Ω is usually $\mathcal{R}^k$. The representation (2.4) of an exponential family is not unique. In fact, any transformation $\tilde{\eta}(\theta) = \eta(\theta)D$ with a $p \times p$ nonsingular matrix D gives another representation (with T replaced by $\tilde{T} = T(D^\tau)^{-1}$). A change of the measure that dominates the family also changes the representation. For example, if we define $\lambda(A) = \int_A h d\nu$ for any $A \in \mathcal{F}$, then we obtain an exponential family with densities

$$\frac{dP_\theta}{d\lambda}(\omega) = \exp\bigl\{T(\omega)[\eta(\theta)]^\tau - \xi(\theta)\bigr\}. \tag{2.5}$$

In an exponential family, consider the reparameterization $\eta = \eta(\theta)$ and

$$\tilde{f}_\eta(\omega) = \exp\bigl\{T(\omega)\eta^\tau - \zeta(\eta)\bigr\}h(\omega), \quad \omega \in \Omega, \tag{2.6}$$

where $\zeta(\eta) = \log\bigl\{\int_\Omega e^{T(\omega)\eta^\tau} h(\omega)d\nu(\omega)\bigr\}$. This is the *canonical form* for the family, which is not unique for the reasons discussed previously. The new parameter η is called the *natural parameter*. The new parameter space is $\Xi = \{\eta(\theta) : \theta \in \Theta\} \subset \mathcal{R}^p$. The set

$$\bigl\{\eta \in \mathcal{R}^p : \int_\Omega e^{T(\omega)\eta^\tau} h(\omega)d\nu(\omega) < \infty\bigr\}$$

is called the *natural parameter space* and is the largest possible parameter space (in canonical form). An exponential family in canonical form with

the natural parameter space is called a *natural exponential family*. If there is an open set contained in the parameter space of an exponential family, then the family is said to be of *full rank*.

Example 2.5. Let P_θ be the binomial distribution $Bi(\theta, n)$ with parameter θ, where n is a fixed positive integer. Then $\{P_\theta : \theta \in (0,1)\}$ is an exponential family, since the p.d.f. of P_θ w.r.t. the counting measure is

$$f_\theta(x) = \exp\left\{x \log \tfrac{\theta}{1-\theta} + n \log(1-\theta)\right\} \binom{n}{x} I_{\{0,1,\ldots,n\}}(x)$$

($T(x) = x$, $\eta(\theta) = \log \frac{\theta}{1-\theta}$, $\xi(\theta) = n\log(1-\theta)$, and $h(x) = \binom{n}{x} I_{\{0,1,\ldots,n\}}(x)$). If we let $\eta = \log \frac{\theta}{1-\theta}$, then $\Xi = \mathcal{R}$ and the family with p.d.f.'s

$$\tilde{f}_\eta(x) = \exp\{x\eta - n\log(1+e^\eta)\} \binom{n}{x} I_{\{0,1,\ldots,n\}}(x)$$

is a natural exponential family of full rank. ∎

Example 2.6. The normal family $\{N(\mu, \sigma^2) : \mu \in \mathcal{R}, \sigma > 0\}$ is an exponential family, since the Lebesgue p.d.f. of $N(\mu, \sigma^2)$ can be written as

$$\frac{1}{\sqrt{2\pi}} \exp\left\{x\frac{\mu}{\sigma^2} - x^2\frac{1}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \log\sigma\right\}.$$

Hence, $T(x) = (x, -x^2)$, $\eta(\theta) = \left(\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2}\right)$, $\theta = (\mu, \sigma^2)$, $\xi(\theta) = -\frac{\mu^2}{2\sigma^2} - \log\sigma$, and $h(x) = 1/\sqrt{2\pi}$. Let $\eta = (\eta_1, \eta_2) = \left(\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2}\right)$. Then $\Xi = \mathcal{R} \times (0, \infty)$ and we can obtain a natural exponential family of full rank with $\zeta(\eta) = -\eta_1^2/(4\eta_2) - \log(1/\sqrt{2\eta_2})$. ∎

For an exponential family, (2.5) implies that there is a nonzero measure λ such that

$$\frac{dP_\theta}{d\lambda}(\omega) > 0 \qquad \text{for all } \omega \text{ and } \theta. \tag{2.7}$$

We can use this fact to show that a family of distributions is not an exponential family. For example, consider the family of uniform distributions, i.e., P_θ is $U(0, \theta)$ with an unknown $\theta \in (0, \infty)$. If $\{P_\theta : \theta \in (0, \infty)\}$ is an exponential family, then from the previous discussion we have a nonzero measure λ such that (2.7) holds. For any $t > 0$, there is a $\theta < t$ such that $P_\theta(t, \infty) = 0$, which with (2.7) implies that $\lambda(t, \infty) = 0$. Also, for any $t \le 0$, $P_\theta(-\infty, t) = 0$, which with (2.7) implies that $\lambda(-\infty, t) = 0$. Since t is arbitrary, $\lambda \equiv 0$. This contradiction implies that $\{P_\theta : \theta \in (0, \infty)\}$ cannot be an exponential family.

The reader may verify which of the parametric families from Tables 1.1 and 1.2 are exponential families. As another example, we consider an important exponential family containing multivariate discrete distributions.

Example 2.7 (The multinomial family). Consider an experiment having $k+1$ possible outcomes with p_i as the probability for the ith outcome, $i = 0, 1, ..., k$, $\sum_{i=0}^k p_i = 1$. In n independent trials of this experiment, let X_i be the number of trials resulting in the ith outcome, $i = 0, 1, ..., k$. Then the joint p.d.f. (w.r.t. counting measure) of $(X_0, X_1, ..., X_k)$ is

$$f_\theta(x_0, x_1, ..., x_k) = \frac{n!}{x_0! x_1! \cdots x_k!} p_0^{x_0} p_1^{x_1} \cdots p_k^{x_k} I_B(x_0, x_1, ..., x_k),$$

where $B = \{(x_0, x_1, ..., x_k) : \ x_i\text{'s are integers} \geq 0, \sum_{i=0}^k x_i = n\}$ and $\theta = (p_0, p_1, ..., p_k)$. The distribution of $(X_0, X_1, ..., X_k)$ is called the *multinomial* distribution, which is an extension of the binomial distribution. In fact, the marginal c.d.f. of each X_i is the binomial distribution $Bi(p_i, n)$. Let $\Theta = \{\theta \in \mathcal{R}^{k+1} : \ 0 < p_i < 1, \sum_{i=0}^k p_i = 1\}$. The parametric family $\{f_\theta : \ \theta \in \Theta\}$ is called the multinomial family. Let $x = (x_0, x_1, ..., x_k)$, $\eta = (\log p_0, \log p_1, ..., \log p_k)$, and $h(x) = [n!/(x_0! x_1! \cdots x_k!)] I_B(x)$. Then

$$f_\theta(x_0, x_1, ..., x_k) = \exp\{x\eta^\tau\} h(x) \qquad x \in \mathcal{R}^{k+1}. \tag{2.8}$$

Hence, the multinomial family is an exponential family with natural parameter η. However, representation (2.8) does not provide an exponential family of full rank, since there is no open set of $\mathcal{R}^{k+1}$ contained in the parameter space Θ or Ξ. A reparameterization leads to an exponential family with full rank. Using the fact that $\sum_{i=0}^k X_i = n$ and $\sum_{i=0}^k p_i = 1$, we obtain that

$$f_\theta(x_0, x_1, ..., x_k) = \exp\{x_* \eta_*^\tau - \zeta(\eta_*)\} h(x) \qquad x \in \mathcal{R}^{k+1}, \tag{2.9}$$

where $x_* = (x_1, ..., x_k)$, $\eta_* = (\log(p_1/p_0), ..., \log(p_k/p_0))$, and $\zeta(\eta_*) = -n \log p_0$. The η_*-parameter space is $\mathcal{R}^k$. Hence the family of densities given by (2.9) is a natural exponential family of full rank. ∎

An important property of exponential families is that if X_1 and X_2 are independent random vectors with p.d.f.'s in exponential families dominated by σ-finite measures ν_1 and ν_2 on $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$, respectively, then the joint p.d.f. of (X_1, X_2) is again in an exponential family dominated by $\nu_1 \times \nu_2$ on $(\Omega_1 \times \Omega_2, \sigma(\mathcal{F}_1 \times \mathcal{F}_2))$. By induction, the result extends to the joint distribution of any $k \geq 2$ random vectors.

The following result summarizes some other useful properties of exponential families. Its proof can be found in Lehmann (1986).

Theorem 2.1. Let $\mathcal{P}$ be a natural exponential family given by (2.6).
(i) The random vector T has the following p.d.f. in an exponential family dominated by some measure on $(\mathcal{R}^p, \mathcal{B}^p)$:

$$\exp\{t\eta^\tau - \zeta(\eta)\} g(t), \quad t \in \mathcal{R}^p,$$

where g is a nonnegative Borel function.
(ii) If η_0 is an interior point of the natural parameter space, then the m.g.f. ψ_{η_0} of $P_{\eta_0} \circ T^{-1}$ is finite in a neighborhood of 0 and is given by

$$\psi_{\eta_0}(t) = \exp\{\zeta(\eta_0 + t) - \zeta(\eta_0)\}.$$

Furthermore, if f is a Borel function satisfying $\int |f| dP_{\eta_0} < \infty$, then the function

$$\int f(\omega) \exp\{T(\omega)\eta^\tau\} h(\omega) d\nu(\omega)$$

is infinitely often differentiable in a neighborhood of η_0, and the derivatives may be computed by differentiation under the integral sign. ∎

Using Theorem 2.1(ii) and the result in Example 2.5, we obtain that the m.g.f. of the binomial distribution $Bi(p, n)$ is

$$\begin{aligned}\psi_\eta(t) &= \exp\{n \log(1 + e^{\eta+t}) - n \log(1 + e^\eta)\} \\ &= \left(\frac{1 + e^\eta e^t}{1 + e^\eta}\right)^n \\ &= (1 - p + pe^t)^n,\end{aligned}$$

since $p = e^\eta/(1 + e^\eta)$.

Definition 2.3 (Location-scale families). Let P be a known probability measure on $(\mathcal{R}^k, \mathcal{B}^k)$ and $\mathcal{M}_k$ be the collection of all $k \times k$ symmetric positive definite matrices. The family

$$\{P_{(\mu,\Sigma)} : \ \mu \in \mathcal{R}^k, \Sigma \in \mathcal{M}_k\} \tag{2.10}$$

is called a *location-scale family* (on $\mathcal{R}^k$), where

$$P_{(\mu,\Sigma)}(B) = P\left((B - \mu)\Sigma^{-1/2}\right), \quad B \in \mathcal{B}^k,$$

$(B - \mu)\Sigma^{-1/2} = \{(x - \mu)\Sigma^{-1/2} : \ x \in B\} \subset \mathcal{R}^k$, and $\Sigma^{-1/2}$ is the inverse of the "square root" matrix $\Sigma^{1/2}$ satisfying $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$. The parameters μ and Σ are called the location and scale parameters, respectively. ∎

There are a number of important subfamilies of the family given by (2.10). Let I_k be the $k \times k$ identity matrix. Then $\{P_{(\mu,I_k)} : \ \mu \in \mathcal{R}^k\}$ is called a *location family*. The family $\{P_{(0,\Sigma)} : \ \Sigma \in \mathcal{M}_k\}$ is called a *scale family*. In some cases we consider a location-scale family $\{P_{(\mu,\sigma^2 I_k)} : \ \mu \in \mathcal{R}^k, \sigma > 0\}$ or $\{P_{(\mu,\sigma^2 I_k)} : \ \mu \in \mathcal{R}_0^k, \sigma > 0\}$, where $\mathcal{R}_0^k = \{(x, ..., x) \in \mathcal{R}^k : \ x \in \mathcal{R}\}$. If $X_1, ..., X_k$ are i.i.d. random variables whose common distribution is in a location-scale family on $\mathcal{R}$, then the joint distribution of $X_1, ..., X_k$ is in $\{P_{(\mu,\sigma^2 I_k)} : \ \mu \in \mathcal{R}_0^k, \sigma > 0\}$.

A location-scale family can be generated as follows. Let X be a random k-vector having a distribution P. Then the distribution of $X\Sigma^{1/2} + \mu$ is $P_{(\mu,\Sigma)}$. On the other hand, if X is a random k-vector whose distribution is in the location-scale family (2.10), then the distribution $XD + c$, $c \in \mathcal{R}^k$ and $D \in \mathcal{M}_k$, is also in the same family.

Let F be the c.d.f. of P. Then the c.d.f. of $P_{(\mu,\Sigma)}$ is $F\left((x-\mu)\Sigma^{-1/2}\right)$, $x \in \mathcal{R}^k$. If F has a Lebesgue p.d.f. f, then the Lebesgue p.d.f. of $P_{(\mu,\Sigma)}$ is $\mathrm{Det}(\Sigma^{-1/2}) f\left((x-\mu)\Sigma^{-1/2}\right)$, $x \in \mathcal{R}^k$ (Proposition 1.8).

Many families of distributions in Table 1.2 (§1.3.1) are location, scale, or location-scale families. For example, the family of exponential distributions $E(a,\theta)$ is a location-scale family on $\mathcal{R}$ with location parameter a and scale parameter θ; the family of uniform distributions $U(0,\theta)$ is a scale family on $\mathcal{R}$ with a scale parameter θ. The k-dimensional normal family discussed in Example 2.4 is a location-scale family on $\mathcal{R}^k$.

2.2 Statistics and Sufficiency

Let us assume now that our data set is a realization of a sample X (a random vector) from an unknown population P on a probability space.

2.2.1 Statistics and their distributions

A measurable function of X, $T(X)$, is called a *statistic* if $T(X)$ is a known value whenever X is known, i.e., the function T is a known function. Statistical analyses are based on various statistics, for various purposes. Of course, X itself is a statistic, but it is a trivial statistic. The range of a nontrivial statistic $T(X)$ is usually simpler than that of X. For example, X may be a random n-vector and $T(X)$ may be a random p-vector with a p much smaller than n. This is desired since $T(X)$ simplifies the original data.

From a probabilistic point of view, the "information" within the statistic $T(X)$ concerning the unknown distribution of X is contained in the σ-field $\sigma(T(X))$. To see this, assume that S is any other statistic for which $\sigma(S(X)) = \sigma(T(X))$. Then by Theorem 1.6, S is a measurable function of T and T is a measurable function of S. Thus, once the value of S (or T) is known, so is the value of T (or S). That is, it is not the particular values of a statistic that contain the information, but the generated σ-field of the statistic. Values of a statistic may be important for other reasons.

Note that $\sigma(T(X)) \subset \sigma(X)$ and the two σ-fields are the same if and only if T is one-to-one. Usually $\sigma(T(X))$ simplifies $\sigma(X)$, i.e., a statistic provides a "reduction" of the σ-field.

Any $T(X)$ is a random element. If the distribution of X is unknown, then the distribution of T may also be unknown, although T is a known function. Finding the form of the distribution of T is one of the major problems in statistical inference and decision theory. Since T is a transformation of X, tools we learn in Chapter 1 for transformations may be useful in finding the distribution or an approximation to the distribution of $T(X)$.

Example 2.8. Let $X_1, ..., X_n$ be i.i.d. random variables having a common distribution P and $X = (X_1, ..., X_n)$. The sample mean $\bar{X}$ and sample variance S^2 defined in (2.1) and (2.2), respectively, are two commonly used statistics. Can we find the joint or the marginal distributions of $\bar{X}$ and S^2? It depends on how much we know about P.

First, let us consider the moments of $\bar{X}$ and S^2. Assume that P has a finite mean denoted by μ. Then

$$E\bar{X} = \mu.$$

If P is in a parametric family $\{P_\theta : \theta \in \Theta\}$, then $E\bar{X} = \int x dP_\theta = \mu(\theta)$ for some function $\mu(\cdot)$. Even if the form of μ is known, $\mu(\theta)$ may still be unknown since θ is unknown. Assume now that P has a finite variance denoted by σ^2. Then

$$\mathrm{Var}(\bar{X}) = \sigma^2/n,$$

which equals $\sigma^2(\theta)/n$ for some function $\sigma^2(\cdot)$ if P is in a parametric family. With a finite $\sigma^2 = \mathrm{Var}(X_1)$, we can also obtain that

$$ES^2 = \sigma^2.$$

With a finite $E|X_1|^3$, we can obtain $E(\bar{X})^3$ and $\mathrm{Cov}(\bar{X}, S^2)$, and with a finite $E|X_1|^4$, we can obtain $\mathrm{Var}(S^2)$ (exercise).

Next, consider the distribution of $\bar{X}$. If P is in a parametric family, we can often find the distribution of $\bar{X}$. See Example 1.17 and some exercises in §1.6. For example, $\bar{X}$ is $N(\mu, \sigma^2/n)$ if P is $N(\mu, \sigma^2)$; $n\bar{X}$ has the gamma distribution $\Gamma(n, \theta)$ if P is the exponential distribution $E(0, \theta)$. If P is not in a parametric family, then it is usually hard to find the exact form of the distribution of $\bar{X}$. One can, however, use the CLT (§1.5.4) to obtain an approximation to the distribution of $\bar{X}$. Applying Corollary 1.2 (for the case of $k = 1$), we obtain that

$$\sqrt{n}(\bar{X} - \mu) \to_d N(0, \sigma^2)$$

and, by (1.56), the distribution of $\bar{X}$ can be approximated by $N(\mu, \sigma^2/n)$, where μ and σ^2 are the mean and variance of P, respectively, and are assumed to be finite.

Compared to $\bar{X}$, the distribution of S^2 is harder to obtain. Assuming that P is $N(\mu, \sigma^2)$, one can show that $(n-1)S^2/\sigma^2$ has the chi-square

distribution χ^2_{n-1} (see Example 2.18). An approximate distribution for S^2 can be obtained from the approximate joint distribution of $\bar{X}$ and S^2 discussed next.

Under the assumption that P is $N(\mu, \sigma^2)$, it can be shown that $\bar{X}$ and S^2 are independent (Example 2.18). Hence the joint distribution of $(\bar{X}, S^2)$ is the product of the marginal distributions of $\bar{X}$ and S^2 given in the previous discussion. Without the normality assumption, an approximate joint distribution can be obtained as follows. Assume again that $\mu = EX_1$, $\sigma^2 = \mathrm{Var}(X_1)$, and $E|X_1|^4$ are finite. Let $Y_i = (X_i - \mu, (X_i - \mu)^2)$, $i = 1, ..., n$. Then $Y_1, ..., Y_n$ are i.i.d. random 2-vectors with $EY_1 = (0, \sigma^2)$ and variance-covariance matrix

$$\Sigma = \begin{pmatrix} \sigma^2 & E(X_1 - \mu)^3 \\ E(X_1 - \mu)^3 & E(X_1 - \mu)^4 - \sigma^4 \end{pmatrix}.$$

Note that $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i = (\bar{X} - \mu, \tilde{S}^2)$, where $\tilde{S}^2 = n^{-1} \sum_{i=1}^n (X_i - \mu)^2$. Applying the CLT (Corollary 1.2) to Y_i's, we obtain that

$$\sqrt{n}(\bar{X} - \mu, \tilde{S}^2 - \sigma^2) \to_d N_2(0, \Sigma).$$

Since

$$S^2 = \frac{n}{n-1} \left[\tilde{S}^2 - (\bar{X} - \mu)^2 \right]$$

and $\bar{X} \to_{a.s.} \mu$ (the SLLN, Theorem 1.13), an application of Slutsky's theorem (Theorem 1.11) leads to

$$\sqrt{n}(\bar{X} - \mu, S^2 - \sigma^2) \to_d N_2(0, \Sigma). \quad \blacksquare$$

Example 2.9 (Order statistics). Let $X = (X_1, ..., X_n)$ with i.i.d. random components and let $X_{(i)}$ be the ith ordered value of $X_1, ..., X_n$. The statistics $X_{(1)}, ..., X_{(n)}$ are called the *order statistics*, which is a set of very useful statistics, in addition to the sample mean and variance in the previous example. Suppose that X_i has a c.d.f. F having a Lebesgue p.d.f. f. Then the joint Lebesgue p.d.f. of $X_{(1)}, ..., X_{(n)}$ is

$$g(x_1, x_2, ..., x_n) = \begin{cases} n! f(x_1) f(x_2) \cdots f(x_n) & x_1 < x_2 < \cdots < x_n \\ 0 & \text{otherwise.} \end{cases}$$

The joint Lebesgue p.d.f. of $X_{(i)}$ and $X_{(j)}$, $1 \le i < j \le n$, is

$$g_{i,j}(x, y) = \begin{cases} \frac{n![F(x)]^{i-1}[F(y)-F(x)]^{j-i-1}[1-F(y)]^{n-j} f(x) f(y)}{(i-1)!(j-i-1)!(n-j)!} & x < y \\ 0 & \text{otherwise} \end{cases}$$

and the Lebesgue p.d.f. of $X_{(i)}$ is

$$g_i(x) = \frac{n!}{(i-1)!(n-i)!} [F(x)]^{i-1} [1 - F(x)]^{n-i} f(x). \quad \blacksquare$$

2.2.2 Sufficiency and minimal sufficiency

Having discussed the reduction of the σ-field $\sigma(X)$ by using a statistic $T(X)$, we now ask whether such a reduction results in any loss of information concerning the unknown population. If a statistic $T(X)$ is fully as informative as the original sample X, then statistical analyses can be done using $T(X)$ which is simpler than X. The next concept describes what we mean by fully informative.

Definition 2.4 (Sufficiency). Let X be a sample from an unknown population $P \in \mathcal{P}$, where $\mathcal{P}$ is a family of populations. A statistic $T(X)$ is said to be *sufficient* for $P \in \mathcal{P}$ (or for $\theta \in \Theta$ when $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a parametric family) if and only if the conditional distribution of X given T is *known* (does not depend on P or θ). ∎

Definition 2.4 can be interpreted as follows. Whence we observe X and compute a sufficient statistic $T(X)$, the original data X do not contain any further information concerning the unknown population P (since its conditional distribution is unrelated to P) and can be discarded. A sufficient statistic $T(X)$ contains all information about P contained in X and provides a reduction of the data if T is not one-to-one. Thus, one of the questions raised in Example 2.1 can be answered as follows: it is enough to just look at $\bar{x}$ and s^2 for the problem of measuring θ if $(\bar{X}, S^2)$ is sufficient for P (or θ).

The concept of sufficiency depends on the given family $\mathcal{P}$. If T is sufficient for $P \in \mathcal{P}$, then T is also sufficient for $P \in \mathcal{P}_0 \subset \mathcal{P}$ but not necessarily sufficient for $P \in \mathcal{P}_1 \supset \mathcal{P}$.

Example 2.10. Suppose that $X = (X_1, ..., X_n)$ and $X_1, ..., X_n$ are i.i.d. from the binomial distribution with the p.d.f. (w.r.t. the counting measure)

$$f_\theta(z) = \theta^z (1-\theta)^{1-z} I_{\{0,1\}}(z), \quad z \in \mathcal{R}, \quad \theta \in (0,1).$$

For any realization x of X, x is a sequence of n ones and zeros. Consider the statistic $T(X) = \sum_{i=1}^n X_i$, which is the number of ones in X. Before showing that T is sufficient, we can intuitively argue that T contains all information about θ, since θ is the probability of an occurrence of a one in x. Given $T = t$ (the number of ones in x), what is left in the data set x is the redundant information about the positions of t ones. Since the random variables are discrete, it is not difficult to compute the conditional distribution of X given $T = t$. Note that

$$P(X = x | T = t) = \frac{P(X = x, T = t)}{P(T = t)}$$

and $P(T=t)=\binom{n}{t}\theta^t(1-\theta)^{n-t}I_{\{0,1,...,n\}}(t)$. Let x_i be the ith component of x. If $t\neq\sum_{i=1}^n x_i$, then $P(X=x,T=t)=0$. If $t=\sum_{i=1}^n x_i$, then

$$\begin{aligned}P(X=x,T=t)&=\prod_{i=1}^n P(X_i=x_i)\\&=\prod_{i=1}^n\theta^{x_i}(1-\theta)^{1-x_i}I_{\{0,1\}}(x_i)\\&=\theta^t(1-\theta)^{n-t}\prod_{i=1}^n I_{\{0,1\}}(x_i).\end{aligned}$$

Let $B_t=\{(x_1,...,x_n):\ x_i=0,1,\ \sum_{i=1}^n x_i=t\}$. Then

$$P(X=x|T=t)=\frac{1}{\binom{n}{t}}I_{B_t}(x)$$

is a known p.d.f. This shows that $T(X)$ is sufficient for $\theta\in(0,1)$, according to Definition 2.4 with the family $\{f_\theta:\theta\in(0,1)\}$. ∎

Finding a sufficient statistic by means of the definition is not convenient since it involves guessing a statistic T that might be sufficient and computing the conditional distribution of X given $T=t$. For families of populations having p.d.f.'s, there is a simple way to find a sufficient statistic.

Theorem 2.2 (The factorization theorem). Suppose that X is a sample from $P\in\mathcal{P}$ and $\mathcal{P}$ is a family of probability measures on $(\mathcal{R}^n,\mathcal{B}^n)$ dominated by a σ-finite measure ν. Then $T(X)$ is sufficient for $P\in\mathcal{P}$ if and only if there are nonnegative Borel functions h (which does not depend on P) on $(\mathcal{R}^n,\mathcal{B}^n)$ and g_P (which depends on P) on the range of T such that

$$\frac{dP}{d\nu}(x)=g_P\big(T(x)\big)h(x). \tag{2.11}$$

Proof. (i) We first show that $\mathcal{P}$ is dominated by a probability measure $Q=\sum_{i=1}^\infty c_iP_i$, where c_i's are nonnegative constants with $\sum_{i=1}^\infty c_i=1$ and $P_i\in\mathcal{P}$. Assume that ν is a finite measure (the case of σ-finite ν is left as an exercise). Let $\mathcal{P}_0$ be the family of all measures of the form $\sum_{i=1}^\infty c_iP_i$, where $P_i\in\mathcal{P}$, $c_i\geq 0$, and $\sum_{i=1}^\infty c_i=1$. Then, it suffices to show that there is a $Q\in\mathcal{P}_0$ such that $Q(A)=0$ implies $P(A)=0$ for all $P\in\mathcal{P}_0$. Let $\mathcal{C}$ be the class of events C for which there exists $P\in\mathcal{P}_0$ such that $P(C)>0$ and $dP/d\nu>0$ a.e. ν on C. Then there exists a sequence $\{C_i\}\subset\mathcal{C}$ such that $\nu(C_i)\to\sup_{C\in\mathcal{C}}\nu(C)$. Let C_0 be the union of all C_i's and $Q=\sum_{i=1}^\infty c_iP_i$, where P_i is the probability measure corresponding to C_i. Then $C_0\in\mathcal{C}$ (exercise). Suppose now that $Q(A)=0$. Let $P\in\mathcal{P}_0$

and $B = \{x : dP/d\nu > 0\}$. Since $Q(A \cap C_0) = 0$, $\nu(A \cap C_0) = 0$ and $P(A \cap C_0) = 0$. Then $P(A) = P(A \cap C_0^c \cap B)$. If $P(A \cap C_0^c \cap B) > 0$, then $\nu(C_0 \cup (A \cap C_0^c \cap B)) > \nu(C_0)$, which contradicts $\nu(C_0) = \sup_{C \in \mathcal{C}} \nu(C)$ since $A \cap C_0^c \cap B$ and therefore $C_0 \cup (A \cap C_0^c \cap B)$ is in $\mathcal{C}$. Thus, $P(A) = 0$ for all $P \in \mathcal{P}_0$.

(ii) Suppose that T is sufficient for $P \in \mathcal{P}$. Then, for any $A \in \mathcal{B}^n$, $P(A|T)$ does not depend on P. For any $B \in \sigma(T)$,

$$\int_B P(A|T)dP = P(A \cap B)$$

and, by Fubini's theorem,

$$\int_B P(A|T)dQ = Q(A \cap B),$$

where Q is the probability measure obtained in part (i) of the proof. This shows that $P(A|T) = E_Q(I_A|T)$, the conditional expectation of I_A given T w.r.t. Q. Let $g_P(T)$ be the Radon-Nikodym derivative dP/dQ on the space $(\mathcal{R}^n, \sigma(T), Q)$. Since $P(A|T)$ is measurable w.r.t. $\sigma(T)$, we obtain that (using Propositions 1.7 and 1.12)

$$\begin{aligned}
P(A) &= \int P(A|T)dP \\
&= \int E_Q(I_A|T)dP \\
&= \int E_Q(I_A|T)g_P(T)dQ \\
&= \int E_Q[g_P(T)I_A|T]dQ \\
&= \int g_P(T)I_A dQ \\
&= \int_A g_P(T)\frac{dQ}{d\nu}d\nu
\end{aligned}$$

for any $A \in \mathcal{B}^n$. Hence, (2.11) holds with $h = dQ/d\nu$.

(iii) Suppose that (2.11) holds. Then

$$\begin{aligned}
\frac{dP}{dQ} &= \frac{dP}{d\nu}\bigg/\frac{dQ}{d\nu} \\
&= \frac{dP}{d\nu}\bigg/\sum_{i=1}^{\infty} c_i \frac{dP_i}{d\nu} \\
&= g_P(T)\bigg/\sum_{i=1}^{\infty} g_{P_i}(T) \quad \text{a.s. } Q,
\end{aligned} \tag{2.12}$$

where the second equality can be proved using the same argument in the proof of Proposition 1.7(ii). Let $\bar{g}_P(T)$ denote the right-hand side of (2.12). Then

$$\frac{dP}{dQ} = \bar{g}_P(T) \quad \text{a.s. } Q. \tag{2.13}$$

Let A be a fixed event and $P \in \mathcal{P}$. The sufficiency of T follows from

$$P(A|T) = E_Q(I_A|T) \quad \text{a.s. } P, \tag{2.14}$$

where $E_Q(I_A|T)$ is given in part (ii) of the proof. This is because $E_Q(I_A|T)$ does not vary with $P \in \mathcal{P}$. Since $\bar{g}_P(T) > 0$ a.s. Q and $\mathcal{P}$ is dominated by Q, (2.14) is the same as

$$\bar{g}_P(T)P(A|T) = \bar{g}_P(T)E_Q(I_A|T) \quad \text{a.s. } Q. \tag{2.15}$$

Since all functions in (2.15) are Borel functions of T, (2.15) follows from

$$\int_B \bar{g}_P(T)P(A|T)dQ = \int_B \bar{g}_P(T)E_Q(I_A|T)dQ \tag{2.16}$$

for any $B \in \sigma(T)$. Let $B \in \sigma(T)$. By Proposition 1.12(vi) and the definition of the conditional expectation, the right-hand side of (2.16) is equal to

$$\int_B E_Q[\bar{g}_P(T)I_A|T]dQ = \int_B \bar{g}_P(T)I_A dQ.$$

By (2.13), Proposition 1.7(i), and the definition of the conditional expectation, the left-hand side of (2.16) is equal to

$$\begin{aligned}\int_B P(A|T)\frac{dP}{dQ}dQ &= \int_B P(A|T)dP \\ &= \int_B I_A dP \\ &= \int_B I_A \frac{dP}{dQ}dQ \\ &= \int_B I_A \bar{g}_P(T)dQ.\end{aligned}$$

This proves (2.16) for any $B \in \sigma(T)$ and completes the proof. ∎

If $\mathcal{P}$ is an exponential family with p.d.f.'s given by (2.4) and $X(\omega) = \omega$, then we can apply Theorem 2.2 with $g_\theta(t) = \exp\{t[\eta(\theta)]^\tau - \xi(\theta)\}$ and conclude that T is a sufficient statistic for $\theta \in \Theta$. In Example 2.10 the joint distribution of X is in an exponential family with $T(X) = \sum_{i=1}^n X_i$. Hence, we can conclude that T is sufficient for $\theta \in (0,1)$ without computing the conditional distribution of X given T.

Example 2.11 (Truncation families). Let $\phi(x)$ be a positive Borel function on $(\mathcal{R}, \mathcal{B})$ such that $\int_a^b \phi(x)dx < \infty$ for any a and b, $-\infty < a < b < \infty$. Let $\theta = (a, b)$, $\Theta = \{(a, b) \in \mathcal{R}^2 : \ a < b\}$, and

$$f_\theta(x) = c(\theta)\phi(x)I_{(a,b)}(x),$$

where $c(\theta) = \left[\int_a^b \phi(x)dx\right]^{-1}$. Then $\{f_\theta : \theta \in \Theta\}$, called a truncation family, is a parametric family dominated by the Lebesgue measure on $\mathcal{R}$. Let $X_1, ..., X_n$ be i.i.d. random variables having the p.d.f. f_θ. Then the joint p.d.f. of $X = (X_1, ..., X_n)$ is

$$\prod_{i=1}^n f_\theta(x_i) = [c(\theta)]^n I_{(a,\infty)}(x_{(1)})I_{(-\infty,b)}(x_{(n)}) \prod_{i=1}^n \phi(x_i), \tag{2.17}$$

where $x_{(i)}$ is the ith ordered value of $x_1, ..., x_n$. Let $T(X) = (X_{(1)}, X_{(n)})$, $g_\theta(t_1, t_2) = [c(\theta)]^n I_{(a,\infty)}(t_1)I_{(-\infty,b)}(t_2)$, and $h(x) = \prod_{i=1}^n \phi(x_i)$. By (2.17) and Theorem 2.2, $T(X)$ is sufficient for $\theta \in \Theta$. ∎

Example 2.12 (Order statistics). Let $X = (X_1, ..., X_n)$ and $X_1, ..., X_n$ be i.i.d. random variables having a distribution $P \in \mathcal{P}$, where $\mathcal{P}$ is the family of distributions on $\mathcal{R}$ having Lebesgue p.d.f.'s. Let $X_{(1)}, ..., X_{(n)}$ be the order statistics given in Example 2.9. Note that the joint p.d.f. of X is

$$f(x_1) \cdots f(x_n) = f(x_{(1)}) \cdots f(x_{(n)}).$$

Hence $T(X) = (X_{(1)}, ..., X_{(n)})$ is sufficient for $P \in \mathcal{P}$. The order statistics can be shown to be sufficient even when $\mathcal{P}$ is not dominated by any σ-finite measure, but Theorem 2.2 is not applicable (see Exercise 27 in §2.6). ∎

There are many sufficient statistics for a given family $\mathcal{P}$. In fact, if T is a sufficient statistic and $T = h(S)$, where h is measurable and S is another statistic, then S is sufficient. This is obvious from Theorem 2.2 if the population has a p.d.f., but it can be proved directly from Definition 2.4 (Exercise 22). For instance, in Example 2.10, $(\sum_{i=1}^m X_i, \sum_{i=m+1}^n X_i)$ is sufficient for θ, where m is any fixed integer between 1 and n. If T is sufficient and $T = h(S)$ with a measurable h, then $\sigma(T) \subset \sigma(S)$ and T is more useful than S, since T provides a further reduction of the data (or σ-field) without loss of information. Is there a sufficient statistic that provides "maximal" reduction of the data?

Before introducing the next concept, we need the following notation. If a statement holds except for outcomes in an event A satisfying $P(A) = 0$ for all $P \in \mathcal{P}$, then we say that the statement holds a.s. $\mathcal{P}$.

Definition 2.5 (Minimal sufficiency). Let T be a sufficient statistic for $P \in \mathcal{P}$. T is called a *minimal sufficient* statistic if and only if, for any other

statistic S sufficient for $P \in \mathcal{P}$, there is a measurable function h such that $T = h(S)$ a.s. $\mathcal{P}$. ∎

If both T and S are minimal sufficient statistics, then by definition there is a one-to-one function h such that $T = h(S)$ a.s. $\mathcal{P}$. Hence the minimal sufficient statistic is unique in the sense that two statistics that are one-to-one functions of each other can be treated as one statistic. Minimal sufficient statistics exist under weak assumptions, e.g., the range of X is $\mathcal{R}^k$ and $\mathcal{P}$ is dominated by a σ-finite measure (Bahadur, 1957).

Example 2.13. Let $X_1, ..., X_n$ be i.i.d. random variables from P_θ, the uniform distribution $U(\theta, \theta+1)$, $\theta \in \mathcal{R}$. Suppose that $n > 1$. The joint Lebesgue p.d.f. of $(X_1, ..., X_n)$ is

$$f_\theta(x) = \prod_{i=1}^n I_{(\theta,\theta+1)}(x_i) = I_{(x_{(n)}-1, x_{(1)})}(\theta), \quad x = (x_1, ..., x_n) \in \mathcal{R}^n,$$

where $x_{(i)}$ denotes the ith ordered value of $x_1, ..., x_n$. By Theorem 2.2, $T = (X_{(1)}, X_{(n)})$ is sufficient for θ. Note that

$$x_{(1)} = \sup\{\theta : f_\theta(x) > 0\} \quad \text{and} \quad x_{(n)} = 1 + \inf\{\theta : f_\theta(x) > 0\}.$$

If $S(X)$ is a statistic sufficient for θ, then by Theorem 2.2, there are Borel functions h and g_θ such that $f_\theta(x) = g_\theta(S(x))h(x)$. For x with $h(x) > 0$,

$$x_{(1)} = \sup\{\theta : g_\theta(S(x)) > 0\} \quad \text{and} \quad x_{(n)} = 1 + \inf\{\theta : g_\theta(S(x)) > 0\}.$$

Hence, there is a measurable function ψ such that $T(x) = \psi(S(x))$ when $h(x) > 0$. Since $h > 0$ a.s. $\mathcal{P}$, we conclude that T is minimal sufficient. ∎

The next result provides a useful method to find minimal sufficient statistics.

Theorem 2.3. (i) Let $\mathcal{P}$ be a family of distributions and $\mathcal{P}_0 \subset \mathcal{P}$ such that a.s. $\mathcal{P}_0$ implies a.s. $\mathcal{P}$. If T is sufficient for $P \in \mathcal{P}$ and minimal sufficient for $P \in \mathcal{P}_0$, then T is minimal sufficient for $P \in \mathcal{P}$.
(ii) Let $\mathcal{P}$ be a family of $k+1$ p.d.f.'s, $f_0, f_1, ..., f_k$, w.r.t. a σ-finite measure on the range of X. Suppose that $\{x : f_i(x) > 0\} \subset \{x : f_0(x) > 0\}$ and that $T_i(X)$ is a statistic satisfying $T_i(x) = f_i(x)/f_0(x)$ when $f_0(x) > 0$, $i = 1, ..., k$. Then $T(X) = (T_1, ..., T_k)$ is minimal sufficient for $P \in \mathcal{P}$.
Proof. (i) If S is sufficient for $P \in \mathcal{P}$, then it is also sufficient for $P \in \mathcal{P}_0$ and, therefore, $T = h(S)$ a.s. $\mathcal{P}_0$ holds for a measurable function h. The result follows from the assumption that a.s. $\mathcal{P}_0$ implies a.s. $\mathcal{P}$.
(ii) Note that $f_0 > 0$ a.s. $\mathcal{P}$. Let $g_0(T) = 1$ and $g_i(T) = T_i$, $i = 1, ..., k$. Then $f_i(x) = g_i(T(x))f_0(x)$ a.s. $\mathcal{P}$. By Theorem 2.2, T is sufficient for

$P \in \mathcal{P}$. Suppose that $S(X)$ is another sufficient statistic. By Theorem 2.2, there are Borel functions h and $\tilde{g}_i$ such that $f_i(x) = \tilde{g}_i(S(x))h(x)$, $i = 0, 1, ..., k$. Then $T_i(x) = \tilde{g}_i(S(x))/\tilde{g}_0(S(x))$ for x's satisfying $f_0(x) > 0$. By Definition 2.5, T is minimal sufficient for $P \in \mathcal{P}$. ∎

Example 2.14. Let $\mathcal{P} = \{f_\theta : \theta \in \Theta\}$ be an exponential family with p.d.f.'s f_θ given by (2.4) and $X(\omega) = \omega$. Suppose that there exists $\Theta_0 = \{\theta_0, \theta_1, ..., \theta_p\} \subset \Theta$ such that the vectors $\eta_i = \eta(\theta_i) - \eta(\theta_0)$, $i = 1, ..., p$, are linearly independent in $\mathcal{R}^p$. (This is true if the family is of full rank.) We have shown that $T(X)$ is sufficient for $\theta \in \Theta$. We now show that T is in fact minimal sufficient. Let $\mathcal{P}_0 = \{f_\theta : \theta \in \Theta_0\}$. By Theorem 2.3(ii),

$$S(X) = \big(\exp\{T(x)\eta_1^\tau - \xi_1\}, ..., \exp\{T(x)\eta_p^\tau - \xi_p\}\big)$$

is minimal sufficient for $\theta \in \Theta_0$, where $\xi_i = \xi(\theta_i) - \xi(\theta_0)$. Since η_i's are linearly independent, there is a one-to-one measurable function ψ such that $T(X) = \psi(S(X))$ a.s. $\mathcal{P}_0$. Hence, T is minimal sufficient for $\theta \in \Theta_0$. It is easy to see that a.s. $\mathcal{P}_0$ implies a.s. $\mathcal{P}$. Thus, by Theorem 2.3(i), T is minimal sufficient for $\theta \in \Theta$. ∎

The sufficiency (and minimal sufficiency) depends on the postulated family $\mathcal{P}$ of populations (statistical models). Hence, it may not be a useful concept if the proposed statistical model is wrong or at least one has some doubts about the correctness of the proposed model. From the examples in this section and some exercises in §2.6, one can find that for a wide variety of models, statistics such as $\bar{X}$ in (2.1), S^2 in (2.2), $(X_{(1)}, X_{(n)})$ in Example 2.11, and the order statistics in Example 2.9 are sufficient. Thus, using these statistics for data reduction and summarization does not lose any information when the true model is one of those models but we do not know exactly which model is correct.

2.2.3 Complete statistics

A statistic $V(X)$ is said to be *ancillary* if its distribution does not depend on the population P and *first-order ancillary* if $E[V(X)]$ is independent of P. A trivial ancillary statistic is the constant statistic $V(X) \equiv c \in \mathcal{R}$. If $V(X)$ is a nontrivial ancillary statistic, then $\sigma(V(X)) \subset \sigma(X)$ is a nontrivial σ-field that does not contain any information about P. Hence, if $S(X)$ is a statistic and $V(S(X))$ is a nontrivial ancillary statistic, it indicates that $\sigma(S(X))$ contains a nontrivial σ-field that does not contain any information about P and, hence, the "data" $S(X)$ may be further reduced. A sufficient statistic T appears to be most successful in reducing the data if no nonconstant function of T is ancillary or even first-order ancillary. This leads to the following concept of completeness.

Definition 2.6 (Completeness). A statistic $T(X)$ is said to be *complete* for $P \in \mathcal{P}$ if and only if, for any Borel f, $E[f(T)] = 0$ for all $P \in \mathcal{P}$ implies $f = 0$ a.s. $\mathcal{P}$. T is said to be *boundedly complete* if and only if the previous statement holds for any bounded Borel f. ∎

A complete statistic is boundedly complete. If T is complete and $S = h(T)$, then S is complete. Intuitively, a complete and sufficient statistic should be minimal sufficient, which was shown by Lehmann and Scheffé (1950) and Bahadur (1957) (see Exercise 37). However, a minimal sufficient statistic is not necessarily complete; for example, the minimal sufficient statistic $(X_{(1)}, X_{(n)})$ in Example 2.13 is not complete (Exercise 36).

Proposition 2.1. If P is in an exponential family of full rank with p.d.f.'s given by (2.6), then $T(X)$ is complete and sufficient for $\eta \in \Xi$.
Proof. We have shown that T is sufficient. Suppose that there is a function f such that $E[f(T)] = 0$ for all $\eta \in \Xi$. By Theorem 2.1(i),

$$\int f(t) \exp\{t\eta^\tau - \zeta(\eta)\} d\lambda = 0 \quad \text{for all } \eta \in \Xi,$$

where λ is a measure on $(\mathcal{R}^p, \mathcal{B}^p)$. Let η_0 be an interior point of Ξ. Then

$$\int f_+(t) e^{t\eta^\tau} d\lambda = \int f_-(t) e^{t\eta^\tau} d\lambda \quad \text{for all } \eta \in N(\eta_0), \tag{2.18}$$

where $N(\eta_0) = \{\eta \in \mathcal{R}^p : \|\eta - \eta_0\| < \epsilon\}$ for some $\epsilon > 0$. In particular,

$$\int f_+(t) e^{t\eta_0^\tau} d\lambda = \int f_-(t) e^{t\eta_0^\tau} d\lambda = c.$$

If $c = 0$, then $f = 0$ a.e. λ. If $c > 0$, then $c^{-1} f_+(t) e^{t\eta_0^\tau}$ and $c^{-1} f_-(t) e^{t\eta_0^\tau}$ are p.d.f.'s w.r.t. λ and (2.18) implies that their m.g.f.'s are the same in a neighborhood of 0. By Proposition 1.10(ii), $c^{-1} f_+(t) e^{t\eta_0^\tau} = c^{-1} f_-(t) e^{t\eta_0^\tau}$, i.e., $f = f_+ - f_- = 0$ a.e. λ. Hence T is complete. ∎

Proposition 2.1 is useful for finding a complete and sufficient statistic when the family of distributions is an exponential family of full rank.

Example 2.15. Suppose that $X_1, ..., X_n$ are i.i.d. random variables having the $N(\mu, \sigma^2)$ distribution, $\mu \in \mathcal{R}$, $\sigma > 0$. From Example 2.6, the joint p.d.f. of $X_1, ..., X_n$ is $(2\pi)^{-n/2} \exp\{T_1\eta_1 + T_2\eta_2 - n\zeta(\eta)\}$, where $T_1 = \sum_{i=1}^n X_i$, $T_2 = -\sum_{i=1}^n X_i^2$, and $\eta = (\eta_1, \eta_2) = \left(\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2}\right)$. Hence the family of distributions for $X = (X_1, ..., X_n)$ is a natural exponential family of full rank ($\Xi = \mathcal{R} \times (0, \infty)$). By Proposition 2.1, $T(X) = (T_1, T_2)$ is complete and sufficient for η. Since there is a one-to-one correspondence between η and $\theta = (\mu, \sigma^2)$, T is also complete and sufficient for θ. It is easy to show that

any one-to-one measurable function of a complete and sufficient statistic is also complete and sufficient. Thus, $(\bar{X}, S^2)$ is complete and sufficient for θ, where $\bar{X}$ and S^2 are the sample mean and variance given by (2.1) and (2.2), respectively. ∎

The following examples show how to find a complete statistic for a non-exponential family.

Example 2.16. Let $X_1, ..., X_n$ be i.i.d. random variables from P_θ, the uniform distribution $U(0, \theta)$, $\theta > 0$. The largest order statistic, $X_{(n)}$, is complete and sufficient for $\theta \in (0, \infty)$. The sufficiency of $X_{(n)}$ follows from the fact that the joint Lebesgue p.d.f. of $X_1, ..., X_n$ is $\theta^{-n} I_{(0,\theta)}(x_{(n)})$. From Example 2.9, $X_{(n)}$ has the Lebesgue p.d.f. $(nx^{n-1}/\theta^n) I_{(0,\theta)}(x)$ on $\mathcal{R}$. Let f be a Borel function such that $E[f(X_{(n)})] = 0$ for all $\theta > 0$. Then

$$\int_0^\theta f(x) x^{n-1} dx = 0 \quad \text{for all } \theta > 0,$$

which implies

$$\int_A f(x) x^{n-1} dx = 0 \quad \text{for all } A \in \mathcal{B}_{(0,\infty)}$$

(exercise). This implies that $f(x)x^{n-1} = 0$ a.e. Lebesgue measure and, hence, $f(x) = 0$ a.e. Lebesgue measure. Therefore, $X_{(n)}$ is complete and sufficient for $\theta \in (0, \infty)$. ∎

Example 2.17. In Example 2.12, we showed that the order statistics $T(X) = (X_{(1)}, ..., X_{(n)})$ of i.i.d. random variables $X_1, ..., X_n$ is sufficient for $P \in \mathcal{P}$, where $\mathcal{P}$ is the family of distributions on $\mathcal{R}$ having Lebesgue p.d.f.'s. We now show that $T(X)$ is also complete for $P \in \mathcal{P}$. Let $\mathcal{P}_0$ be the family of Lebesgue p.d.f.'s of the form

$$f(x) = C(\theta_1, ..., \theta_n) \exp\{-x^{2n} + \theta_1 x + \theta_2 x^2 + \cdots + \theta_n x^n\},$$

where $\theta_j \in \mathcal{R}$ and $C(\theta_1, ..., \theta_n)$ is a normalizing constant such that $\int f(x)dx = 1$. Then $\mathcal{P}_0 \subset \mathcal{P}$ and $\mathcal{P}_0$ is an exponential family of full rank. Note that the joint distribution of $X = (X_1, ..., X_n)$ is also in an exponential family of full rank. Thus, by Proposition 2.1, $U = (U_1, ..., U_n)$ is a complete statistic for $P \in \mathcal{P}_0$, where $U_j = \sum_{i=1}^n X_i^j$. Since a.s. $\mathcal{P}_0$ implies a.s. $\mathcal{P}$, $U(X)$ is also complete for $P \in \mathcal{P}$.

The result follows if we can show that there is a one-to-one correspondence between $T(X)$ and $U(X)$. Let $V_1 = \sum_{i=1}^n X_i$, $V_2 = \sum_{i<j} X_i X_j$, $V_3 = \sum_{i<j<k} X_i X_j X_k$,..., $V_n = X_1 \cdots X_n$. From the identities

$$U_k - V_1 U_{k-1} + V_2 U_{k-2} - \cdots + (-1)^{k-1} V_{k-1} U_1 + (-1)^k k V_k, \quad k = 1, ..., n,$$

there is a one-to-one correspondence between $U(X)$ and $V(X) = (V_1, ..., V_n)$. From the identity

$$(t - X_1) \cdots (t - X_n) = t^n - V_1 t^{n-1} + V_2 t^{n-2} - \cdots + (-1)^n V_n,$$

there is a one-to-one correspondence between $V(X)$ and $T(X)$. This completes the proof and, hence, $T(X)$ is sufficient and complete for $P \in \mathcal{P}$. In fact, both $U(X)$ and $V(X)$ are sufficient and complete for $P \in \mathcal{P}$. ∎

The relationship between an ancillary statistic and a complete and sufficient statistic is characterized in the following result.

Theorem 2.4 (Basu's theorem). Let V and T be two statistics of X from a population $P \in \mathcal{P}$. If V is ancillary and T is boundedly complete and sufficient for $P \in \mathcal{P}$, then V and T are independent w.r.t. any $P \in \mathcal{P}$.
Proof. Let B be an event on the range of V. Since V is ancillary, $P(V^{-1}(B))$ is a constant. Since T is sufficient, $E[I_B(V)|T]$ is a function of T (independent of P). Since $E\{E[I_B(V)|T] - P(V^{-1}(B))\} = 0$ for all $P \in \mathcal{P}$, $P(V^{-1}(B)|T) = E[I_B(V)|T] = P(V^{-1}(B))$ a.s. $\mathcal{P}$, by the bounded completeness of T. Let A be an event on the range of T. Then, $P(T^{-1}(A) \cap V^{-1}(B)) = E\{E[I_A(T)I_B(V)|T]\} = E\{I_A(T)E[I_B(V)|T]\} = E\{I_A(T)P(V^{-1}(B))\} = P(T^{-1}(A))P(V^{-1}(B))$. Hence T and V are independent w.r.t. any $P \in \mathcal{P}$. ∎

Basu's theorem is useful in proving the independence of two statistics.

Example 2.18. Suppose that $X_1, ..., X_n$ are i.i.d. random variables having the $N(\mu, \sigma^2)$ distribution, with $\mu \in \mathcal{R}$ and a known $\sigma > 0$. It can be easily shown that the family $\{N(\mu, \sigma^2) : \mu \in \mathcal{R}\}$ is an exponential family of full rank with natural parameter $\eta = \mu/\sigma^2$. By Proposition 2.1, the sample mean $\bar{X}$ in (2.1) is complete and sufficient for η (and μ). Let S^2 be the sample variance given by (2.2). Since $S^2 = (n-1)^{-1} \sum_{i=1}^n (Z_i - \bar{Z})^2$, where $Z_i = X_i - \mu$ is $N(0, \sigma^2)$ and $\bar{Z} = n^{-1} \sum_{i=1}^n Z_i$, S^2 is an ancillary statistic (σ^2 is known). By Basu's theorem, $\bar{X}$ and S^2 are independent w.r.t. $N(\mu, \sigma^2)$ with $\mu \in \mathcal{R}$. Since σ^2 is arbitrary, $\bar{X}$ and S^2 are independent w.r.t. $N(\mu, \sigma^2)$ for any $\mu \in \mathcal{R}$ and $\sigma^2 > 0$.

Using the independence of $\bar{X}$ and S^2, we now show that $(n-1)S^2/\sigma^2$ has the chi-square distribution χ^2_{n-1}. Note that

$$n\left(\frac{\bar{X} - \mu}{\sigma}\right)^2 + \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2.$$

From the properties of the normal distributions, $n(\bar{X} - \mu)^2/\sigma^2$ has the chi-square distribution χ^2_1 with the m.g.f. $(1 - 2t)^{-1/2}$ and $\sum_{i=1}^n (X_i - \mu)^2/\sigma^2$

has the chi-square distribution χ^2_n with the m.g.f. $(1-2t)^{-n/2}$, $t < 1/2$. By the independence of $\bar{X}$ and S^2, the m.g.f. of $(n-1)S^2/\sigma^2$ is

$$(1-2t)^{-n/2}/(1-2t)^{-1/2} = (1-2t)^{-(n-1)/2}$$

for $t < 1/2$. This is the m.g.f. of the chi-square distribution χ^2_{n-1} and, therefore, the result follows. ∎

2.3 Statistical Decision Theory

In this section we describe some basic elements in statistical decision theory. More developments are given in later chapters.

2.3.1 Decision rules, loss functions, and risks

Let X be a sample from a population $P \in \mathcal{P}$. A *statistical decision* is an *action* that we take after we observe X, for example, a conclusion about P or a characteristic of P. Throughout this section we use $\mathbb{A}$ to denote the set of allowable actions. Let $\mathcal{F}_{\mathbb{A}}$ be a σ-field on $\mathbb{A}$. Then the measurable space $(\mathbb{A}, \mathcal{F}_{\mathbb{A}})$ is called the *action space*. Let $\mathfrak{X}$ be the range of X and $\mathcal{F}_{\mathfrak{X}}$ be a σ-field on $\mathfrak{X}$. A *decision rule* is a measurable function (a statistic) T from $(\mathfrak{X}, \mathcal{F}_{\mathfrak{X}})$ to $(\mathbb{A}, \mathcal{F}_{\mathbb{A}})$. If a decision rule T is chosen, then we take the action $T(X) \in \mathbb{A}$ whence X is observed.

The construction or selection of decision rules cannot be done without any criterion about the performance of decision rules. In *statistical decison theory*, we set a criterion using a *loss function* L, which is a function from $\mathcal{P} \times \mathbb{A}$ to $[0, \infty)$ and is Borel on $(\mathbb{A}, \mathcal{F}_{\mathbb{A}})$ for each fixed $P \in \mathcal{P}$. If $X = x$ is observed and our decision rule is T, then our "loss" (in making a decision) is $L(P, T(x))$. The average loss for the decision rule T, which is called the *risk* of using T, is defined to be

$$R_T(P) = E[L(P, T(X))] = \int_{\mathfrak{X}} L(P, T(x))dP(x). \tag{2.19}$$

The loss and risk functions are denoted by $L(\theta, a)$ and $R_T(\theta)$ if $\mathcal{P}$ is a parametric family indexed by θ. A decision rule with small loss is preferred. But it is difficult to compare $L(P, T_1(X))$ and $L(P, T_2(X))$ for two decision rules T_1 and T_2, since both of them are random. For this reason, the risk function (2.19) is introduced and we compare two decision rules by comparing their risks. A rule T_1 is *as good as* another rule T_2 if and only if

$$R_{T_1}(P) \le R_{T_2}(P) \quad \text{for any } P \in \mathcal{P}, \tag{2.20}$$

and is *better* than T_2 if and only if (2.20) holds and $R_{T_1}(P) < R_{T_2}(P)$ for at least one $P \in \mathcal{P}$. Two decision rules T_1 and T_2 are *equivalent* if and only

if $R_{T_1}(P) = R_{T_2}(P)$ for all $P \in \mathcal{P}$. If there is a decision rule T_* which is as good as any other rule in $\Im$, a class of allowable decision rules, then T_* is said to be $\Im$-*optimal* (or optimal if $\Im$ contains all possible rules).

Example 2.19. Consider the measurement problem in Example 2.1. Suppose that we need a decision on the value of $\theta \in \mathcal{R}$, based on the sample $X = (X_1, ..., X_n)$. If Θ is all possible values of θ, then it is reasonable to consider the action space $(\mathbb{A}, \mathcal{F}_{\mathbb{A}}) = (\Theta, \mathcal{B}_\Theta)$. An example of a decision rule is $T(X) = \bar{X}$, the sample mean defined by (2.1). A common loss function in this problem is the *squared error loss* $L(P, a) = (\theta - a)^2$, $a \in \mathbb{A}$. Then the loss for the decision rule $\bar{X}$ is the squared deviation between $\bar{X}$ and θ. Assuming that the population has mean μ and variance $\sigma^2 < \infty$, we obtain the following risk function for $\bar{X}$:

$$\begin{aligned} R_{\bar{X}}(P) &= E(\theta - \bar{X})^2 \\ &= (\theta - E\bar{X})^2 + E(E\bar{X} - \bar{X})^2 \\ &= (\theta - E\bar{X})^2 + \mathrm{Var}(\bar{X}) \qquad (2.21) \\ &= (\mu - \theta)^2 + \tfrac{\sigma^2}{n}, \qquad (2.22) \end{aligned}$$

where result (2.22) follows from the results for the moments of $\bar{X}$ in Example 2.8. If θ is in fact the mean of the population, then the first term on the right-hand side of (2.22) is 0 and the risk is an increasing function of the population variance σ^2 and a decreasing function of the sample size n.

Consider another decision rule $T_1(X) = (X_{(1)} + X_{(n)})/2$. However, $R_{T_1}(P)$ does not have an explicit form if there is no further assumption on the population P. Suppose that $P \in \mathcal{P}$. Then, for some $\mathcal{P}$, $\bar{X}$ (or T_1) is better than T_1 (or $\bar{X}$) (exercise), whereas for some $\mathcal{P}$, neither $\bar{X}$ nor T_1 is better than the other.

A different loss function may also be considered. For example, $L(P, a) = |\theta - a|$, which is called the *absolute error loss*. However, $R_{\bar{X}}(P)$ and $R_{T_1}(P)$ do not have explicit forms unless $\mathcal{P}$ is of some specific form. ∎

The problem in Example 2.19 is a special case of a general problem called *estimation*, in which the action space is the set of all possible values of a population characteristic ϑ to be estimated. In an estimation problem, a decision rule T is called an *estimator* and result (2.21) holds with $\theta = \vartheta$ and $\bar{X}$ replaced by any estimator with a finite variance. The following example describes another type of important problem called *hypothesis testing*.

Example 2.20. Let $\mathcal{P}$ be a family of distributions, $\mathcal{P}_0 \subset \mathcal{P}$, and $\mathcal{P}_1 = \{P \in \mathcal{P} : P \notin \mathcal{P}_0\}$. A hypothesis testing problem can be formulated as that of deciding which of these two statements is true:

$$H_0 : \ P \in \mathcal{P}_0 \qquad \text{versus} \qquad H_1 : \ P \in \mathcal{P}_1. \qquad (2.23)$$

Here, H_0 is called the *null hypothesis* and H_1 is called the *alternative hypothesis*. The action space for this problem contains only two elements, i.e., $\mathbb{A} = \{0, 1\}$, where 0 is the action of accepting H_0 and 1 is the action of rejecting H_0. A decision rule is called a *test*. Since a test $T(X)$ is a function from $\mathcal{X}$ to $\{0, 1\}$, $T(X)$ must have the form $I_C(X)$, where $C \in \mathcal{F}_{\mathcal{X}}$ is called the *rejection region* or *critical region* for testing H_0.

A simple loss function for this problem is the 0-1 loss: $L(P, a) = 0$ if a correct decision is made and 1 if an incorrect decision is made, i.e., $L(P, j) = 0$ for $P \in \mathcal{P}_j$ and $L(P, j) = 1$ otherwise, $j = 0, 1$. Under this loss, the risk is

$$R_T(P) = \begin{cases} P(T(X) = 1) = P(X \in C) & P \in \mathcal{P}_0 \\ P(T(X) = 0) = P(X \notin C) & P \in \mathcal{P}_1. \end{cases}$$

See Figure 2.2 on page 97 for examples of graphs of $R_T(\theta)$ for some T and P in a parametric family.

The 0-1 loss implies that the loss for two types of incorrect decisions (accepting H_0 when $P \in \mathcal{P}_1$ and rejecting H_0 when $P \in \mathcal{P}_0$) are the same. In some cases one might assume unequal losses: $L(P, j) = 0$ for $P \in \mathcal{P}_j$, $L(P, 0) = c_0$ when $P \in \mathcal{P}_1$, and $L(P, 1) = c_1$ when $P \in \mathcal{P}_0$. ∎

In the following example the decision problem is neither an estimation nor a testing problem.

Example 2.21. A hazardous toxic waste site requires clean-up when the true chemical concentration θ in the contaminated soil is higher than a given level $\theta_0 \geq 0$. Because of the limitation in resources, we would like to spend our money and efforts more in those areas that pose high risk to public health. In a particular area where soil samples are obtained, we would like to take one of these three actions: a complete clean-up (a_1), a partial clean-up (a_2), and no clean-up (a_3). Then $\mathbb{A} = \{a_1, a_2, a_3\}$. Suppose that the cost for a complete clean-up is c_1 and for a partial clean-up is $c_2 < c_1$; the risk to public health is $c_3(\theta - \theta_0)$ if $\theta > \theta_0$ and 0 if $\theta \leq \theta_0$; a complete clean-up can reduce the toxic concentration to a amount $\leq \theta_0$, whereas a partial clean-up can only reduce a fixed amount of the toxic concentration, i.e., the chemical concentration becomes $\theta - t$ after a partial clean-up, where t is a known constant. Then the loss function is given by

$L(\theta, a)$	a_1	a_2	a_3
$\theta \leq \theta_0$	c_1	c_2	0
$\theta_0 < \theta \leq \theta_0 + t$	c_1	c_2	$c_3(\theta - \theta_0)$
$\theta > \theta_0 + t$	c_1	$c_2 + c_3(\theta - \theta_0 - t)$	$c_3(\theta - \theta_0)$

The risk function can be calculated once the decision rule is specified. We discuss this example again in Chapter 4. ∎

Sometimes it is useful to use another type of decision rules, called the *randomized decision rules*. A randomized decision rule is a function δ on $\mathcal{X} \times \mathcal{F}_{\mathbb{A}}$ such that, for every $x \in \mathcal{X}$, $\delta(x, \cdot)$ is a probability measure on $(\mathbb{A}, \mathcal{F}_{\mathbb{A}})$ and, for every $A \in \mathcal{F}_{\mathbb{A}}$, $\delta(\cdot, A)$ is a Borel function. A nonrandomized decision rule $T(X)$ previously discussed can be viewed as a special randomized decision rule with $\delta(x, A) = I_A(T(x))$. If a randomized rule δ is used, then we obtain a probability measure $\delta(x, \cdot)$ on the action space when $X = x$ is observed. If one wants to choose an action in $\mathbb{A}$, then one needs to simulate a pseudorandom element of $\mathbb{A}$ according to $\delta(x, \cdot)$. Thus, an alternative way to describe a randomized rule is to specify the method of simulating the action from $\mathbb{A}$ for each $x \in \mathcal{X}$.

The loss function for a randomized rule δ is defined as

$$L(P, \delta, x) = \int_{\mathbb{A}} L(P, a) d\delta(x, a), \tag{2.24}$$

which reduces to the same loss function we discussed when δ is a nonrandomized rule. The risk of a randomized rule δ is then

$$R_\delta(P) = E[L(P, \delta, X)] = \int_{\mathcal{X}} \int_{\mathbb{A}} L(P, a) d\delta(x, a) dP(x).$$

Examples of using randomized rules are given in §2.3.2, Chapters 4 and 6.

2.3.2 Admissibility and optimality

Consider a given decision problem with a given loss $L(P, a)$.

Definition 2.7 (Admissibility). Let $\Im$ be a class of decision rules (randomized or nonrandomized). A decision rule $T \in \Im$ is called $\Im$-*admissible* (or admissible if $\Im$ contains all possible rules) if there does not exist any $S \in \Im$ that is better than T (in terms of the risk). ∎

If a decision rule T is inadmissible, then there exists a rule better than T. Thus, T should not be used in principle. However, an admissible decision rule is not necessarily good. For example, in an estimation problem a silly estimator $T(X) \equiv$ a constant may be admissible (Exercise 58).

The relationship between the admissibility and optimality defined in §2.3.1 can be described as follows. If T_* is $\Im$-optimal, then it is $\Im$-admissible; if T_* is $\Im$-optimal and T_0 is $\Im$-admissible, then T_0 is also $\Im$-optimal and is equivalent to T_*; if there are two $\Im$-admissible rules that are not equivalent, then there does not exist any $\Im$-optimal rule.

Suppose that we have a sufficient statistic $T(X)$ for $P \in \mathcal{P}$. Intuitively, our decision rule should be a function of T, based on the discussion in

§2.2.2. This is not true in general, but the following result indicates that this is true if randomized decision rules are allowed.

Proposition 2.2. Suppose that $\mathbb{A}$ is a subset of $\mathcal{R}^k$. Let $T(X)$ be a sufficient statistic for $P \in \mathcal{P}$ and let δ_0 be a decision rule. Then

$$\delta_1(t, A) = E[\delta_0(X, A)|T = t], \tag{2.25}$$

which is a randomized decision rule depending only on T, is equivalent to δ_0 if $R_{\delta_0}(P) < \infty$ for any $P \in \mathcal{P}$.

Proof. Note that δ_1 defined by (2.25) is a decision rule since δ_1 does not depend on the unknown P by the sufficiency of T. From (2.24) and (2.25),

$$\begin{aligned} R_{\delta_1}(P) &= E[L(P, \delta_1, X)] \\ &= E\left\{\int_{\mathbb{A}} L(P, a) d\delta_1(X, a)\right\} \\ &= E\left\{E\left[\int_{\mathbb{A}} L(P, a) d\delta_0(X, a)\middle| T\right]\right\} \\ &= E\left\{\int_{\mathbb{A}} L(P, a) d\delta_0(X, a)\right\} \\ &= R_{\delta_0}(P). \quad \blacksquare \end{aligned}$$

Note that Proposition 2.2 does not imply that δ_0 is inadmissible. Also, if δ_0 is a nonrandomized rule,

$$\delta_1(t, A) = E[I_A(\delta_0(X))|T = t] = P(\delta_0(X) \in A|T = t)$$

is still a randomized rule. Hence, Proposition 2.2 does not apply to situations where randomized rules are not allowed.

The following result tells us when nonrandomized rules are all we need and when decision rules that are not functions of sufficient statistics are inadmissible. Recall from calculus that a subset A of $\mathcal{R}^k$ is *convex* if and only if $tx + (1-t)y \in A$ for any $x \in A$, $y \in A$, and $t \in [0, 1]$; a function f from a convex $A \subset \mathcal{R}^k$ to $\mathcal{R}$ is *convex* if and only if

$$f(tx + (1-t)y) \le tf(x) + (1-t)f(y), \quad x \in A, y \in A, t \in [0, 1]; \tag{2.26}$$

and f is *strictly convex* if and only if (2.26) holds with $\le$ replaced by the strictly inequality $<$.

Theorem 2.5. Suppose that $\mathbb{A}$ is a convex subset of $\mathcal{R}^k$ and that for any $P \in \mathcal{P}$, $L(P, a)$ is a convex function of a.

(i) Let δ be a randomized rule satisfying $\int_{\mathbb{A}} \|a\| d\delta(x, a) < \infty$ for any $x \in \mathcal{X}$ and let $T_1(x) = \int_{\mathbb{A}} a d\delta(x, a)$. Then $L(P, T_1(x)) \le L(P, \delta, x)$ (or

$L(P, T_1(x)) < L(P, \delta, x)$ if L is strictly convex in a) for any $x \in \mathfrak{X}$ and $P \in \mathcal{P}$.
(ii) (Rao-Blackwell's theorem). Let T be a sufficient statistic for $P \in \mathcal{P}$, $T_0 \in \mathcal{R}^k$ be a nonrandomized rule satisfying $E\|T_0\| < \infty$, and $T_1 = E[T_0(X)|T] = (E[T_{01}(X)|T], ..., E[T_{0k}(X)|T])$, where T_{0i} is the ith component of T_0. Then $R_{T_1}(P) \leq R_{T_0}(P)$ for any $P \in \mathcal{P}$. If L is strictly convex in a and T_0 is not sufficient for P, then T_0 is inadmissible. ∎

The proof of Theorem 2.5 is an application of Jensen's inequality (Exercise 46 in §1.6) and is left to the reader.

The concept of admissibility helps us to eliminate some decision rules. However, usually there are still too many rules left after the elimination of some rules according to admissibility and sufficiency. Although one is typically interested in a $\Im$-optimal rule, frequently it does not exist, if $\Im$ is either too large or too small. The following examples are illustrations.

Example 2.22. Let $X_1, ..., X_n$ be i.i.d. random variables from a population $P \in \mathcal{P}$ which is the family of populations having finite mean μ and variance σ^2. Consider the estimation of μ ($\mathbb{A} = \mathcal{R}$) under the squared error loss. It can be shown that if we let $\Im$ be the class of all possible estimators, then there is no $\Im$-optimal rule (exercise). Next, let $\Im_1$ be the class of all linear functions in $X = (X_1, ..., X_n)$, i.e., $T(X) = \sum_{i=1}^n c_i X_i$ with known $c_i \in \mathcal{R}$, $i = 1, ..., n$. It follows from (2.21) and the discussion after Example 2.19 that

$$R_T(P) = \mu^2 \left(\sum_{i=1}^n c_i - 1 \right)^2 + \sigma^2 \sum_{i=1}^n c_i^2. \tag{2.27}$$

We now show that there does not exist $T_* = \sum_{i=1}^n c_i^* X_i$ such that $R_{T_*}(P) \leq R_T(P)$ for any $P \in \mathcal{P}$ and $T \in \Im_1$. If there is such a T_*, then $(c_1^*, ..., c_n^*)$ is a minimum of the function of $(c_1, ..., c_n)$ on the right-hand side of (2.27). Then $c_1^*, ..., c_n^*$ must be the same and equal to $\mu^2/(\sigma^2 + n\mu^2)$, which depends on P. Hence T_* is not a statistic. This shows that there is no $\Im_1$-optimal rule.

Consider now a subclass $\Im_2 \subset \Im_1$ with c_i's satisfying $\sum_{i=1}^n c_i = 1$. From (2.27), $R_T(P) = \sigma^2 \sum_{i=1}^n c_i^2$ if $T \in \Im_2$. Minimizing $\sigma^2 \sum_{i=1}^n c_i^2$ subject to $\sum_{i=1}^n c_i = 1$ leads to an optimal solution of $c_i = n^{-1}$ for all i. Thus, the sample mean $\bar{X}$ is $\Im_2$-optimal.

There may not be any optimal rule if we consider a small class of decision rules. For example, if $\Im_3$ contains all the rules in $\Im_2$ except $\bar{X}$, then one can show that there is no $\Im_3$-optimal rule. ∎

Example 2.23. Assume that the sample X has the binomial distribution $Bi(\theta, n)$ with an unknown $\theta \in (0, 1)$ and a fixed integer $n > 1$. Consider the

hypothesis testing problem described in Example 2.20 with $H_0 : \theta \in (0, \theta_0]$ versus $H_1 : \theta \in (\theta_0, 1)$, where $\theta_0 \in (0, 1)$ is a fixed value. Suppose that we are only interested in the following class of nonrandomized decision rules: $\Im = \{T_j : j = 0, 1, ..., n-1\}$, where $T_j(X) = I_{\{j+1,...,n\}}(X)$. From Example 2.20, the risk function for T_j under the 0-1 loss is

$$R_{T_j}(\theta) = P(X > j)I_{(0,\theta_0]}(\theta) + P(X \leq j)I_{(\theta_0,1)}(\theta).$$

For any integers k and j, $0 \leq k < j \leq n-1$,

$$R_{T_j}(\theta) - R_{T_k}(\theta) = \begin{cases} -P(k < X \leq j) < 0 & 0 < \theta \leq \theta_0 \\ P(k < X \leq j) > 0 & \theta_0 < \theta < 1. \end{cases}$$

Hence, neither T_j nor T_k is better than the other. This shows that every T_j is $\Im$-admissible and, thus, there is no $\Im$-optimal rule. ∎

In view of the fact that an optimal rule often does not exist, statisticians adopt the following two approaches to choose a decision rule. The first approach is to define a class $\Im$ of decision rules that have some desirable properties (statistical and/or nonstatistical) and then try to find the best rule in $\Im$. In Example 2.22, for instance, any estimator T in $\Im_2$ has the property that T is linear in X and $E[T(X)] = \mu$. In a general estimation problem, we can use the following concept.

Definition 2.8 (Unbiasedness). In an estimation problem, the *bias* of an estimator $T(X)$ of a real-valued parameter ϑ of the unknown population is defined to be $b_T(P) = E[T(X)] - \vartheta$ (which is denoted by $b_T(\theta)$ if P is in a parametric family indexed by θ). An estimator $T(X)$ is said to be *unbiased* for ϑ if and only if $b_T(P) = 0$ for any $P \in \mathcal{P}$. ∎

Thus, $\Im_2$ in Example 2.22 is the class of unbiased estimators linear in X. In Chapter 3, we discuss how to find a $\Im$-optimal estimator when $\Im$ is the class of unbiased estimators or unbiased estimators linear in X.

Another class of decision rules can be defined after we introduce the concept of *invariance*. In a problem where the distribution of X is in a location-scale family $\mathcal{P}$ on $\mathcal{R}^k$, we often consider location-scale transformations of data X of the form $XA + c$, where $c \in \mathcal{C} \subset \mathcal{R}^k$ and $A \in \mathcal{T}$, a class of invertible $k \times k$ matrices. We assume that if $A_i \in \mathcal{T}$, $i = 1, 2$, then $A_i^{-1} \in \mathcal{T}$ and $A_1A_2 \in \mathcal{T}$, and that if $c_i \in \mathcal{C}$, $i = 1, 2$, then $-c_i \in \mathcal{C}$ and $c_1A + c_2 \in \mathcal{C}$ for any $A \in \mathcal{T}$. The location-scale family $\mathcal{P}$ is said to be *invariant* if P_{XA+c}, the distribution of $XA + c$, is in $\mathcal{P}$ for any $c \in \mathcal{C}$ and $A \in \mathcal{T}$.

Definition 2.9 (Location-scale invariance). Let $\mathcal{P}$ be a location-scale family invariant for given $\mathcal{C}$ and $\mathcal{T}$.

(i) A decision problem is said to be *invariant* if and only if the loss $L(P, a)$ is invariant in the sense that, for every $A \in \mathcal{T}$, every $c \in \mathcal{C}$, and every $a \in \mathbb{A}$, there exists a unique $g_{c,A}(a) \in \mathbb{A}$ such that $L(P_X, a) = L\left(P_{XA+c}, g_{c,A}(a)\right)$.
(ii) A decision rule $T(x)$ is said to be *invariant* if and only if, for every $A \in \mathcal{T}$, every $c \in \mathcal{C}$, and every $x \in \mathfrak{X}$, $T(xA + c) = g_{c,A}(T(x))$. ∎

Invariance means that our decision is not affected by location-scale transformations of data. In Chapters 4 and 6, we discuss the problem of finding a $\Im$-optimal rule when $\Im$ is a class of invariant decision rules.

Example 2.24. Let $X = (X_1, ..., X_n)$ with i.i.d. components from a population in a location family $\mathcal{P} = \{P_\mu : \mu \in \mathcal{R}\}$. Consider the location transformation $g(X) = X + cJ_k$, where $c \in \mathcal{R}$ and J_k is the k-vector whose components are all equal to 1. $\mathcal{P}$ is invariant under the transformation g with $\mathcal{T} = \{I_k\}$ and $\mathcal{C} = \{cJ_k : c \in \mathcal{R}\}$. For estimating μ under the loss $L(\mu, a) = L(\mu - a)$, where $L(\cdot)$ is a nonnegative Borel function, the decision problem is invariant with $g_{c,A}(a) = g_c(a) = a + c$. A decision rule T is invariant if and only if $T(x + cJ_k) = T(x) + c$ for every $x \in \mathcal{R}^k$ and $c \in \mathcal{R}$. An example of an invariant decision rule is $T(x) = xl^\tau$ for some $l \in \mathcal{R}^k$ with $J_k l^\tau = 1$. Note that $T(x) = xl^\tau$ with $J_k l^\tau = 1$ is in the class $\Im_2$ defined in Example 2.22. ∎

The second approach to finding a good decision rule is to consider some characteristic R_T of $R_T(P)$, for a given decision rule T, and then minimize R_T over $T \in \Im$. The following are two popular ways to carry out this idea. The first one is to consider an average of $R_T(P)$ over $P \in \mathcal{P}$:

$$r_T(\Pi) = \int_{\mathcal{P}} R_T(P) d\Pi(P),$$

where Π is a known probability measure on $(\mathcal{P}, \mathcal{F}_{\mathcal{P}})$ with an appropriate σ-field $\mathcal{F}_{\mathcal{P}}$. $r_T(\Pi)$ is called the *Bayes risk* of T w.r.t. Π. If $T_* \in \Im$ and $r_{T_*}(\Pi) \leq r_T(\Pi)$ for any $T \in \Im$, then T_* is called a $\Im$-*Bayes rule* w.r.t. Π. The second method is to consider the worst situation: if $T_* \in \Im$ and

$$\sup_{P \in \mathcal{P}} R_{T_*}(P) \leq \sup_{P \in \mathcal{P}} R_T(P)$$

for any $T \in \Im$, then T_* is called a $\Im$-*minimax* rule. Bayes and minimax rules are discussed in Chapter 4.

Example 2.25. We usually try to find a Bayes rule or a minimax rule in a parametric problem where $P = P_\theta$ for a $\theta \in \mathcal{R}^k$. Consider the special case of $k = 1$ and $L(\theta, a) = (\theta - a)^2$, the squared error loss. Note that

$$r_T(\Pi) = \int_{\mathcal{R}} E[\theta - T(X)]^2 d\Pi(\theta),$$

which is equivalent to $E[\boldsymbol{\theta} - T(X)]^2$, where $\boldsymbol{\theta}$ is a random variable having the distribution Π and given $\boldsymbol{\theta} = \theta$, the conditional distribution of X is P_θ. Then, the problem can be viewed as a prediction problem for $\boldsymbol{\theta}$ using functions of X. Using the result in Example 1.19, the best predictor is $E(\boldsymbol{\theta}|X)$, which is the $\Im$-Bayes rule w.r.t. Π with $\Im$ being the class of rules $T(X)$ satisfying $E[T(X)]^2 < \infty$ for any θ.

As a more specific example, let $X = (X_1, ..., X_n)$ with i.i.d. components having the $N(\mu, \sigma^2)$ distribution with an unknown $\mu = \theta \in \mathcal{R}$ and a known σ^2, and let Π be the $N(\mu_0, \sigma_0^2)$ distribution with known μ_0 and σ_0^2. Then the conditional distribution of $\boldsymbol{\theta}$ given $X = x$ is $N(\mu_*(x), c^2)$ with

$$\mu_*(x) = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{x} \qquad \text{and} \qquad c^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2} \tag{2.28}$$

(exercise). Then $E(\boldsymbol{\theta}|X) = \mu_*(X)$ is the Bayes rule w.r.t. $\Pi = N(\mu_0, \sigma_0^2)$.

In this special case we can show that the sample mean $\bar{X}$ is $\Im$-minimax with $\Im$ being the collection of all decision rules. For any decision rule T,

$$\begin{aligned}
\sup_{\theta\in\mathcal{R}} R_T(\theta) &\geq \int_{\mathcal{R}} R_T(\theta)d\Pi(\theta) \\
&\geq \int_{\mathcal{R}} R_{\mu_*}(\theta)d\Pi(\theta) \\
&= E\left\{[\boldsymbol{\theta} - \mu_*(X)]^2\right\} \\
&= E\left\{E\{[\boldsymbol{\theta} - \mu_*(X)]^2|X\}\right\} \\
&= E(c^2) \\
&= c^2,
\end{aligned}$$

where $\mu_*(X)$ is the Bayes rule given in (2.28) and c^2 is also given in (2.28). Since this result is true for any $\sigma_0^2 > 0$ and $c^2 \to \sigma^2/n$ as $\sigma_0^2 \to \infty$,

$$\sup_{\theta\in\mathcal{R}} R_T(\theta) \geq \frac{\sigma^2}{n} = \sup_{\theta\in\mathcal{R}} R_{\bar{X}}(\theta),$$

where the equality holds because the risk of $\bar{X}$ under the squared error loss is, by (2.22), σ^2/n and independent of $\theta = \mu$. Thus, $\bar{X}$ is minimax.

A minimax rule in a general case may be difficult to obtain. It can be seen that if both μ and σ^2 are unknown in the previous discussion, then

$$\sup_{\theta\in\mathcal{R}\times(0,\infty)} R_{\bar{X}}(\theta) = \infty, \tag{2.29}$$

where $\theta = (\mu, \sigma^2)$. Hence $\bar{X}$ cannot be minimax unless (2.29) holds with $\bar{X}$ replaced by any decision rule T, in which case minimaxity becomes meaningless. ∎

2.4 Statistical Inference

The loss function plays a crucial role in statistical decision theory. Loss functions can be obtained from a utility analysis (Berger, 1985), but in many problems they have to be determined subjectively. In *statistical inference*, we make an inference about the unknown population based on the sample X and *inference procedures* without using any loss function, although any inference procedure can be cast in decision-theoretic terms as a decision rule.

There are three main types of inference procedures: *point estimators*, *hypothesis tests*, and *confidence sets*.

2.4.1 Point estimators

The problem of estimating an unknown parameter related to the unknown population is introduced in Example 2.19 and the discussion after Example 2.19 as a special statistical decision problem. In statistical inference, however, estimators of parameters are derived based on some principle (such as the unbiasedness, invariance, sufficiency, substitution principle, likelihood principle, Bayesian principle, etc.), not based on a loss or risk function. Since confidence sets are sometimes also called *interval estimators* or *set estimators*, estimators of parameters are called point estimators.

In Chapters 3 through 5, we consider how to derive a "good" point estimator based on some principle. Here we focus on how to assess performance of point estimators.

Let $\vartheta \in \tilde{\Theta} \subset \mathcal{R}$ be a parameter to be estimated, which is a function of the unknown population P or θ if P is in a parametric family. An estimator is a statistic with range $\tilde{\Theta}$. First, one has to realize that any estimator $T(X)$ of ϑ is subject to an estimation error $T(x) - \vartheta$ when we observe $X = x$. This is not just because $T(X)$ is random. In some problems $T(x)$ never equals ϑ. A trivial example is when $T(X)$ has a continuous c.d.f. so that $P(T(X) = \vartheta) = 0$. As a nontrivial example, let $X_1, ..., X_n$ be i.i.d. binary random variables (also called Bernoulli variables) with $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$. The sample mean $\bar{X}$ is shown to be a good estimator of $\vartheta = p$ in later chapters, but $\bar{x}$ never equals ϑ if ϑ is not one of j/n, $j = 0, 1, ..., n$. Thus, we cannot assess the performance of $T(X)$ by the values of $T(x)$ with particular x's and it is also not worthwhile to do so.

The bias $b_T(P)$ and unbiasedness of a point estimator $T(X)$ is defined in Definition 2.8. Unbiasedness of $T(X)$ means that the mean of $T(X)$ is equal to ϑ. An unbiased estimator $T(X)$ can be viewed as an estimator without "systematic" error, since, on the average, it does not overestimate (i.e., $b_T(P) > 0$) or underestimate (i.e., $b_T(P) < 0$). However, an unbiased

estimator $T(X)$ may have large positive and negative errors $T(x)-\vartheta$, $x \in \mathcal{X}$, although these errors cancel each other in the calculation of the bias, which is the average $\int [T(x) - \vartheta] dP_X(x)$.

Hence, for an unbiased estimator $T(X)$, it is desired that the values of $T(x)$ be highly concentrated around ϑ. The variance of $T(X)$ is commonly used as a measure of the dispersion of $T(X)$. The *mean squared error* (mse) of $T(X)$ as an estimator of ϑ is defined to be

$$\mathrm{mse}_T(P) = E[T(X) - \vartheta]^2 = [b_T(P)]^2 + \mathrm{Var}(T(X)), \tag{2.30}$$

which is denoted by $\mathrm{mse}_T(\theta)$ if P is in a parametric family. $\mathrm{mse}_T(P)$ is equal to the variance $\mathrm{Var}(T(X))$ if and only if $T(X)$ is unbiased. Note that the mse is simply the risk of T in statistical decision theory under the squared error loss.

In addition to the variance and the mse, the following are other measures of dispersion that are often used in point estimation problems. The first one is the *mean absolute error* of an estimator $T(X)$ defined to be $E|T(X) - \vartheta|$. The second one is the probability of falling outside a stated distance of ϑ, i.e., $P(|T(X) - \vartheta| \geq \epsilon)$ with a fixed $\epsilon > 0$. Again, these two measures of dispersion are risk functions in statistical decision theory with loss functions $|\vartheta - a|$ and $I_{(\epsilon,\infty)}(|\vartheta - a|)$, respectively.

For the bias, variance, mse, and mean absolute error, we have implicitly assumed that certain moments of $T(X)$ exist. On the other hand, the dispersion measure $P(|T(X)-\vartheta| \geq \epsilon)$ depends on the choice of ϵ. It is possible that some estimators are good in terms of one measure of dispersion, but not in terms of other measures of dispersion. The mse, which is a function of bias and variance according to (2.30), is mathematically easy to handle and, hence, is used the most often in the literature. In this book, we use the mse to assess and compare point estimators unless otherwise stated.

Examples 2.19 and 2.22 provide some examples of estimators and their biases, variances, and mse's. The following are two more examples.

Example 2.26. Consider the life-time testing problem in Example 2.2. Let $X_1, ..., X_n$ be i.i.d. from an unknown c.d.f. F. Suppose that the parameter of interest is $\vartheta = 1 - F(t)$ for a fixed $t > 0$. If F is not in a parametric family, then a *nonparametric* estimator of $F(t)$ is the *empirical* c.d.f.

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,t]}(X_i) \qquad t \in \mathcal{R}. \tag{2.31}$$

Since $I_{(-\infty,t]}(X_1), ..., I_{(-\infty,t]}(X_n)$ are i.i.d. binary random variables with $P(I_{(-\infty,t]}(X_i) = 1) = F(t)$, the random variable $nF_n(t)$ has the binomial distribution $Bi(F(t), n)$. Consequently, $F_n(t)$ is an unbiased estimator of

$F(t)$ and $\mathrm{Var}(F_n(t)) = \mathrm{mse}_{F_n(t)}(P) = F(t)[1 - F(t)]/n$. Since any linear combination of unbiased estimators is unbiased for the same linear combination of the parameters (by the linearity of expectations), an unbiased estimator of ϑ is $U(X) = 1 - F_n(t)$, which has the same variance and mse as $F_n(t)$.

The estimator $U(X) = 1 - F_n(t)$ can be improved in terms of the mse if there is further information about F. Suppose that F is the c.d.f. of the exponential distribution $E(0, \theta)$ with an unknown $\theta > 0$. Then $\vartheta = e^{-t/\theta}$. From §2.2.2, the sample mean $\bar{X}$ is sufficient for $\theta > 0$. Since the squared error loss is strictly convex, an application of Theorem 2.5(ii) (Rao-Blackwell's theorem) shows that the estimator $T(X) = E[1 - F_n(t)|\bar{X}]$, which is also unbiased, is better than $U(X)$ in terms of the mse. Figure 2.1 shows graphs of the mse's of $U(X)$ and $T(X)$, as functions of θ, in the special case of $n = 10$, $t = 2$, and $F(x) = (1 - e^{-x/\theta})I_{(0,\infty)}(x)$. ∎

Example 2.27. Consider the sample survey problem in Example 2.3 with a constant selection probability $p(\boldsymbol{s})$ and univariate y_i. Let $\vartheta = Y = \sum_{i=1}^N y_i$, the population total. We now show that the estimator $\hat{Y} = \frac{N}{n}\sum_{i\in\boldsymbol{s}} y_i$ is an unbiased estimator of Y. Let $a_i = 1$ if $i \in \boldsymbol{s}$ and $a_i = 0$ otherwise. Since $p(\boldsymbol{s})$ is constant, $E(a_i) = P(a_i = 1) = n/N$ and

$$E(\hat{Y}) = E\left(\frac{N}{n}\sum_{i=1}^N a_i y_i\right) = \frac{N}{n}\sum_{i=1}^N y_i E(a_i) = \sum_{i=1}^N y_i = Y.$$

Note that

$$\mathrm{Var}(a_i) = E(a_i) - [E(a_i)]^2 = \frac{n}{N}\left(1 - \frac{n}{N}\right)$$

and for $i \neq j$,

$$\mathrm{Cov}(a_i, a_j) = P(a_i = 1, a_j = 1) - E(a_i)E(a_j) = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2}.$$

Hence, the variance or the mse of $\hat{Y}$ is

$$\begin{aligned}
\mathrm{Var}(\hat{Y}) &= \frac{N^2}{n^2} V\left(\sum_{i=1}^N a_i y_i\right) \\
&= \frac{N^2}{n^2}\left[\sum_{i=1}^N y_i^2 \mathrm{Var}(a_i) + 2\sum_{1\le i<j\le N} y_i y_j \mathrm{Cov}(a_i, a_j)\right] \\
&= \frac{N}{n}\left(1 - \frac{n}{N}\right)\left(\sum_{i=1}^N y_i^2 - \frac{2}{N-1}\sum_{1\le i<j\le N} y_i y_j\right) \\
&= \frac{N^2}{n(N-1)}\left(1 - \frac{n}{N}\right)\sum_{i=1}^N \left(y_i - \frac{Y}{N}\right)^2. \quad ∎
\end{aligned}$$

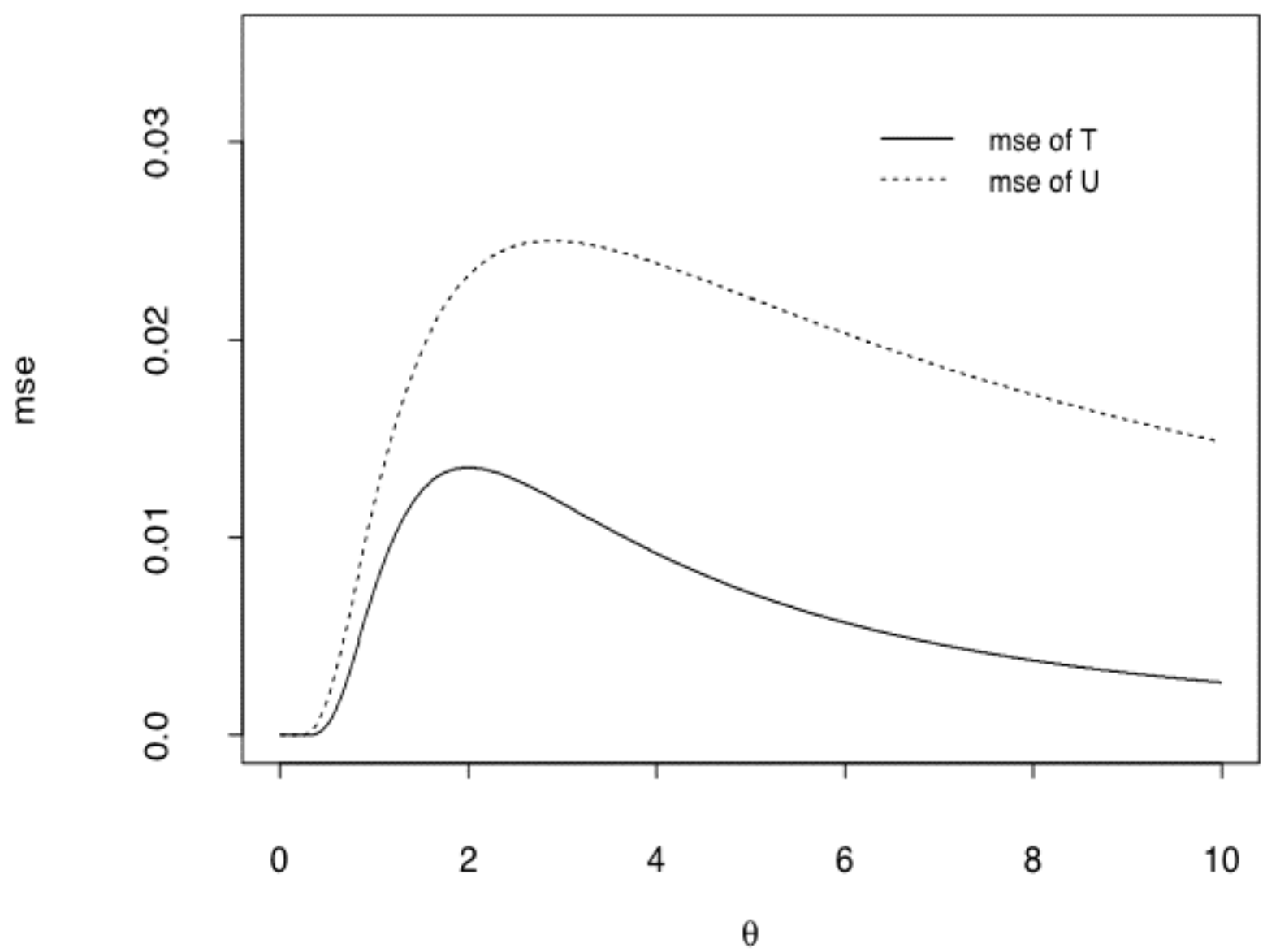

Figure 2.1: mse's of $U(X)$ and $T(X)$ in Example 2.26

2.4.2 Hypothesis tests

The basic elements of a hypothesis testing problem are described in Example 2.20. In statistical inference, tests for a hypothesis are derived based on some principles similar to those given in an estimation problem. Chapter 6 is devoted to deriving tests for various types of hypotheses. Several key ideas are discussed here.

To test the hypotheses H_0 versus H_1 given in (2.23), there are only two types of statistical errors we may commit: rejecting H_0 when H_0 is true (called the *type I error*) and accepting H_0 when H_0 is wrong (called the *type II error*). In statistical inference, a test T, which is a statistic from $\mathcal{X}$ to $\{0, 1\}$, is assessed by the probabilities of making two types of errors:

$$\alpha_T(P) = P(T(X) = 1) \qquad P \in \mathcal{P}_0 \tag{2.32}$$

and

$$1 - \alpha_T(P) = P(T(X) = 0) \qquad P \in \mathcal{P}_1, \tag{2.33}$$

which are denoted by $\alpha_T(\theta)$ and $1 - \alpha_T(\theta)$ if P is in a parametric family indexed by θ. Note that these are risks of T under the 0-1 loss in statistical decision theory. However, an optimal decision rule (test) does not exist even for a very simple problem with a very simple class of tests (Example 2.23).

That is, error probabilities in (2.32) and (2.33) cannot be minimized simultaneously. Furthermore, these two error probabilities cannot be controlled simultaneously when we only have a sample of a fixed size.

Therefore, a common approach to finding an "optimal" test is to assign a small bound α to one of the error probabilities, say $\alpha_T(P)$, $P \in \mathcal{P}_0$, and then to attempt to minimize the other error probability $1-\alpha_T(P)$, $P \in \mathcal{P}_1$, subject to

$$\sup_{P\in\mathcal{P}_0} \alpha_T(P) \leq \alpha. \tag{2.34}$$

The bound α is called the *level of significance*. The left-hand side of (2.34) is called the *size* of the test T. Note that the level of significance should be positive, otherwise no test satisfies (2.34) except the silly test $T(X) \equiv 0$ a.s. $\mathcal{P}$.

Example 2.28. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathcal{R}$ and a known σ^2. Consider the hypotheses

$$H_0 : \mu \leq \mu_0 \qquad \text{versus} \qquad H_1 : \mu > \mu_0,$$

where μ_0 is a fixed constant. Since the sample mean $\bar{X}$ is sufficient for $\mu \in \mathcal{R}$, it is reasonable to consider the following class of tests: $T_c(X) = I_{(c,\infty)}(\bar{X})$, i.e., H_0 is rejected (accepted) if $\bar{X} > c$ ($\bar{X} \leq c$), where $c \in \mathcal{R}$ is a fixed constant. Let Φ be the c.d.f. of $N(0,1)$. Then, by the property of the normal distributions,

$$\alpha_{T_c}(\mu) = P(T_c(X) = 1) = 1 - \Phi\left(\frac{\sqrt{n}(c-\mu)}{\sigma}\right). \tag{2.35}$$

Figure 2.2 provides an example of a graph of two types of error probabilities, with $\mu_0 = 0$. Since $\Phi(t)$ is an increasing function of t,

$$\sup_{P\in\mathcal{P}_0} \alpha_{T_c}(\mu) = 1 - \Phi\left(\frac{\sqrt{n}(c-\mu_0)}{\sigma}\right).$$

In fact, it is also true that

$$\sup_{P\in\mathcal{P}_1} [1 - \alpha_{T_c}(\mu)] = \Phi\left(\frac{\sqrt{n}(c-\mu_0)}{\sigma}\right).$$

If we would like to use an α as the level of significance, then the most effective way is to choose a c_α (a test $T_{c_\alpha}(X)$) such that

$$\alpha = \sup_{P\in\mathcal{P}_0} \alpha_{T_{c_\alpha}}(\mu),$$

in which case c_α must satisfy

$$1 - \Phi\left(\frac{\sqrt{n}(c_\alpha-\mu_0)}{\sigma}\right) = \alpha,$$

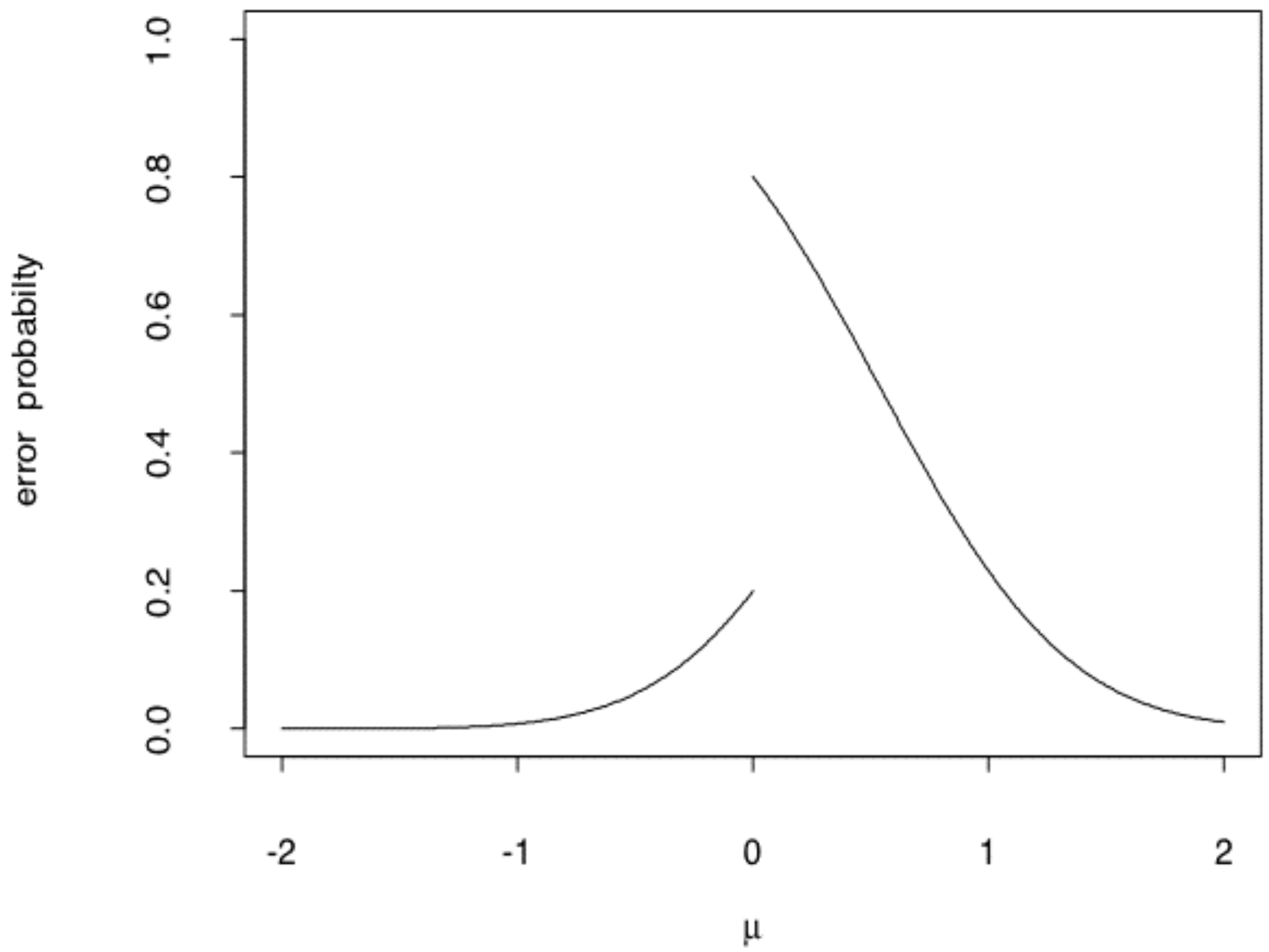

Figure 2.2: Error probabilities in Example 2.28

i.e., $c_\alpha = \sigma\Phi^{-1}(1-\alpha)/\sqrt{n} + \mu_0$. In Chapter 6, it is shown that for any test $T(X)$ satisfying (2.34),

$$1 - \alpha_T(\mu) \geq 1 - \alpha_{T_{c_\alpha}}(\mu), \qquad \mu > \mu_0. \quad \blacksquare$$

The choice of a level of significance α is usually somewhat subjective. In most applications there is no precise limit to the size of T that can be tolerated. Standard values, such as 0.10, 0.05, or 0.01 are often used for convenience.

For most tests satisfying (2.34), a small α leads to a "small" rejection region $\{x : T(x) = 1\}$. It is good practice to determine not only whether H_0 is rejected or accepted for a given α and a given test T, but also the smallest possible level of significance $\hat{\alpha}$ at which H_0 would be rejected for the computed $T(x)$. Such an $\hat{\alpha}$, which depends on x only and is a statistic, is called the *p-value* for T.

Example 2.29. Consider the problem in Example 2.28. Let us calculate the p-value for T_{c_α}. Note that

$$\alpha = 1 - \Phi\left(\frac{\sqrt{n}(c_\alpha - \mu_0)}{\sigma}\right) > 1 - \Phi\left(\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}\right)$$

if and only if $\bar{x} > c_\alpha$ (or $T_{c_\alpha}(x) = 1$). Hence

$$1 - \Phi\left(\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}\right) = \inf\{\alpha \in (0,1) : T_{c_\alpha}(x) = 1\} = \hat{\alpha}(x)$$

is the p-value for T_{c_α}. It turns out that $T_{c_\alpha}(x) = I_{(0,\alpha)}(\hat{\alpha}(x))$. ∎

With the additional information provided by p-values, using p-values is typically more appropriate than using fixed-level tests in a scientific problem. However, a fixed level of significance is unavoidable when acceptance or rejection of H_0 implies an imminent concrete decision. For more discussions about p-values, see Lehmann (1986) and Weerahandi (1995).

In Example 2.28, the equality in (2.34) can always be achieved by a suitable choice of c. This is, however, not true in general. In Example 2.23, for instance, it is possible to find an α such that

$$\sup_{0<\theta\leq\theta_0} P(T_j(X) = 1) \neq \alpha$$

for all T_j's. In such cases we may consider *randomized tests*, which are introduced next.

Recall that a randomized decision rule is a probability measure $\delta(x, \cdot)$ on the action space for any fixed x. Since the action space contains only two points, 0 and 1, for a hypothesis testing problem, any randomized test $\delta(X, A)$ is equivalent to a statistic $T(X) \in [0,1]$ with $T(x) = \delta(x, \{1\})$ and $1 - T(x) = \delta(x, \{0\})$. A nonrandomized test is obviously a special case where $T(x)$ does not take any value in $(0,1)$.

For any randomized test $T(X)$, we define the type I error probability to be $\alpha_T(P) = E[T(X)]$, $P \in \mathcal{P}_0$, and the type II error probability to be $1 - \alpha_T(P) = E[1 - T(X)]$, $P \in \mathcal{P}_1$. For a class of randomized tests, we would like to minimize $1 - \alpha_T(P)$ subject to (2.34).

Example 2.30. Consider Example 2.23 and the following class of randomized tests:

$$T_{j,q}(X) = \begin{cases} 1 & X > j \\ q & X = j \\ 0 & X < j, \end{cases}$$

where $j = 0, 1, ..., n-1$ and $q \in [0,1]$. Then

$$\alpha_{T_{j,q}}(\theta) = P(X > j) + qP(X = j) \qquad 0 < \theta \leq \theta_0$$

and

$$1 - \alpha_{T_{j,q}}(\theta) = P(X < j) + (1-q)P(X = j) \qquad \theta_0 < \theta < 1.$$

It can be shown that for any $\alpha \in (0,1)$, there exist integer j and $q \in (0,1)$ such that the size of $T_{j,q}$ is α (exercise). ∎

2.4.3 Confidence sets

Let ϑ be a real-valued unknown parameter related to the unknown population $P \in \mathcal{P}$ and $C(X) \in \mathcal{B}_{\tilde{\Theta}}$ depending only on the sample X, where $\tilde{\Theta} \in \mathcal{B}$ is the range of ϑ. If

$$\inf_{P \in \mathcal{P}} P(\vartheta \in C(X)) \geq 1 - \alpha, \tag{2.36}$$

where α is a fixed constant in $(0, 1)$, then $C(X)$ is called a *confidence set* for ϑ with *level of significance* $1 - \alpha$. The left-hand side of (2.36) is called the *confidence coefficient* of $C(X)$, which is the highest possible level of significance for $C(X)$. A confidence set is a random element that covers the unknown ϑ with certain probability. If (2.36) holds, then the *coverage probability* of $C(X)$ is at least $1 - \alpha$, although $C(x)$ either covers or does not cover ϑ whence we observe $X = x$. The concept of confidence sets can be extended to the case where ϑ is a vector of k real-valued parameters and $C(X) \in \mathcal{B}^k_{\tilde{\Theta}}$ in an obvious manner.

If $C(X) = [\underline{\vartheta}(X), \overline{\vartheta}(X)]$ for a pair of statistics $\underline{\vartheta}$ and $\overline{\vartheta}$, then $C(X)$ is called a *confidence interval* for ϑ. If $C(X) = (-\infty, \overline{\vartheta}(X)]$ (or $[\underline{\vartheta}(X), \infty)$), then $\overline{\vartheta}$ (or $\underline{\vartheta}$) is called an upper (or a lower) *confidence bound* for ϑ. A confidence interval is also called an interval estimator of ϑ, although it is very different from a point estimator (discussed in §2.4.1). The concepts of level of significance and confidence coefficient are very similar to the level of significance and size in hypothesis testing. In fact, it is shown in Chapter 7 that some confidence sets are closely related to hypothesis tests.

Example 2.31. Consider Example 2.28. Suppose that a confidence interval for $\vartheta = \mu$ is needed. Again, we only need to consider $\underline{\vartheta}(\bar{X})$ and $\overline{\vartheta}(\bar{X})$, since the sample mean $\bar{X}$ is sufficient. Consider confidence intervals of the form $[\bar{X} - c, \bar{X} + c]$, where $c \in (0, \infty)$ is fixed. Note that

$$P\left(\mu \in [\bar{X} - c, \bar{X} + c]\right) = P\left(|\bar{X} - \mu| \leq c\right) = 1 - 2\Phi\left(-\sqrt{n}c/\sigma\right),$$

which is independent of μ. Hence the confidence coefficient of $[\bar{X} - c, \bar{X} + c]$ is $1 - 2\Phi\left(-\sqrt{n}c/\sigma\right)$, which is an increasing function of c and converges to 1 as $c \to \infty$ or 0 as $c \to 0$. Thus, confidence coefficients are positive but less than 1 except for silly confidence intervals $[\bar{X}, \bar{X}]$ and $(-\infty, \infty)$. We can choose a confidence interval with an arbitrarily small confidence coefficient, but the chosen confidence interval may be so wide that it is practically useless.

If σ^2 is also unknown, then $[\bar{X} - c, \bar{X} + c]$ has confidence coefficient 0 and, therefore, is not a good inference procedure. In such a case a different confidence interval for μ with positive confidence coefficient can be derived (Exercise 79 in §2.6). ∎

This example tells us that a reasonable approach is to choose a level of significance $1-\alpha \in (0,1)$ (just like the level of significance in hypothesis testing) and a confidence interval or set satisfying (2.36). In Example 2.31, we may choose a confidence interval whose confidence coefficient is *exactly* $1-\alpha$ for any fixed $\alpha \in (0,1)$, using $c_\alpha = \sigma\Phi^{-1}(1-\alpha/2)/\sqrt{n}$. This is desirable since, for all confidence intervals satisfying (2.36), the one with the shortest interval length is preferred.

For a general confidence interval $[\underline{\vartheta}(X), \overline{\vartheta}(X)]$, its length is $\overline{\vartheta}(X)-\underline{\vartheta}(X)$, which may be random. We may consider the expected (or average) length $E[\overline{\vartheta}(X)-\underline{\vartheta}(X)]$. The confidence coefficient and expected length are a pair of good measures of performance of confidence intervals. Like the two types of error probabilities of a test in hypothesis testing, however, we cannot maximize the confidence coefficient and minimize the length (or expected length) simultaneously. A common approach is to minimize the length (or expected length) subject to (2.36).

For a general confidence set $C(X)$, the length of $C(X)$ may be ∞. Hence we have to define some other measures of performance. For an upper (or a lower) confidence bound, we may consider the distance $\overline{\vartheta}(X)-\vartheta$ (or $\vartheta-\underline{\vartheta}(X)$) or its expectation.

To conclude this section, we discuss an example of a confidence set for a two-dimensional parameter. General discussions about how to construct and assess confidence sets are given in Chapter 7.

Example 2.32. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with both $\mu \in \mathcal{R}$ and $\sigma^2 > 0$ unknown. Let $\theta = (\mu, \sigma^2)$ and $\alpha \in (0,1)$ be given. Let $\bar{X}$ be the sample mean and S^2 be the sample variance. Since $(\bar{X}, S^2)$ is sufficient (Example 2.15), we focus on $C(X)$ which is a function of $(\bar{X}, S^2)$. From Example 2.18, $\bar{X}$ and S^2 are independent and $(n-1)S^2/\sigma^2$ has the chi-square distribution χ^2_{n-1}. Since $\sqrt{n}(\bar{X}-\mu)/\sigma$ has the $N(0,1)$ distribution (Exercise 51 in §1.6),

$$P\left(-\tilde{c}_\alpha \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \tilde{c}_\alpha\right) = \sqrt{1-\alpha},$$

where $\tilde{c}_\alpha = \Phi^{-1}\left(\frac{1+\sqrt{1-\alpha}}{2}\right)$ (verify). Since the chi-square distribution χ^2_{n-1} is a known distribution, we can always find two constants $c_{1\alpha}$ and $c_{2\alpha}$ such that

$$P\left(c_{1\alpha} \le \frac{(n-1)S^2}{\sigma^2} \le c_{2\alpha}\right) = \sqrt{1-\alpha}.$$

Then

$$P\left(-\tilde{c}_\alpha \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \tilde{c}_\alpha, c_{1\alpha} \le \frac{(n-1)S^2}{\sigma^2} \le c_{2\alpha}\right) = 1-\alpha,$$

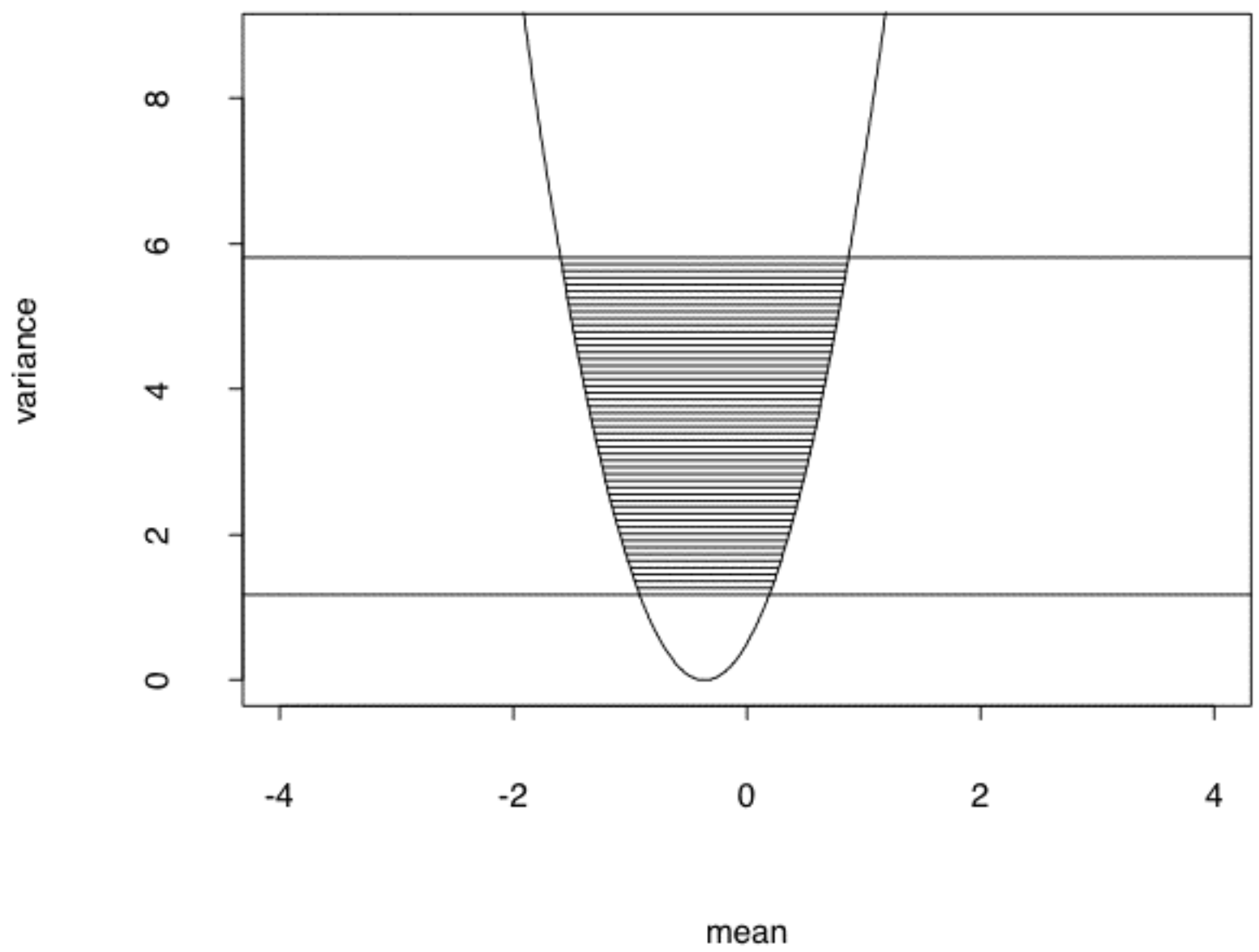

Figure 2.3: A confidence set for θ in Example 2.32

or

$$P\left(\frac{n(\bar{X}-\mu)^2}{\tilde{c}_\alpha^2} \leq \sigma^2, \frac{(n-1)S^2}{c_{2\alpha}} \leq \sigma^2 \leq \frac{(n-1)S^2}{c_{1\alpha}}\right) = 1-\alpha. \qquad (2.37)$$

The left-hand side of (2.37) defines a set in the range of $\theta = (\mu, \sigma^2)$ bounded by two straight lines, $\sigma^2 = (n-1)S^2/c_{i\alpha}$, $i = 1, 2$, and a curve $\sigma^2 = n(\bar{X}-\mu)^2/\tilde{c}_\alpha^2$ (see the shadowed part of Figure 2.3). This set is a confidence set for θ with confidence coefficient $1-\alpha$, since (2.37) holds for any θ. ∎

2.5 Asymptotic Criteria and Inference

We have seen that in statistical decision theory and inference, a key to the success of finding a good decision rule or inference procedure is being able to find some moments and/or distributions of various statistics. Although many examples are presented (including those in the exercises in §2.6), there are more cases in which we are not able to find exactly the moments or distributions of given statistics, especially when the problem is not parametric (see, e.g., the discussions in Example 2.8).

In practice the sample size n is often large, which allows us to approximate the moments and distributions of statistics that are impossible

to derive, using the asymptotic tools discussed in §1.5. In an asymptotic analysis, we consider a sample $X = (X_1, ..., X_n)$ not for fixed n, but as a member of a sequence corresponding to $n = n_0, n_0 + 1, ...$, and obtain the limit of the distribution of an appropriately normalized statistic or variable $T_n(X)$ as $n \to \infty$. The limiting distribution and their moments are used as approximations to the distribution and moments of $T_n(X)$ in the situation with a large but actually finite n. This leads to some asymptotic statistical procedures and asymptotic criteria for assessing their performances, which are introduced in this section.

The asymptotic approach is not only applied to the situation where no exact method is available, but also used to provide an inference procedure simpler (e.g., in terms of computation) than that produced by the exact approach (the approach considering a fixed n). Some examples are given in later chapters.

In addition to providing more theoretical results and/or simpler inference procedures, the asymptotic approach requires less stringent mathematical assumptions than does the exact approach. The mathematical precision of the optimality results obtained in statistical decision theory, for example, tends to obscure the fact that these results are approximations in view of the approximate nature of the assumed models and loss functions. As the sample size increases, the statistical properties become less dependent on the loss functions and models. However, a major weakness of the asymptotic approach is that typically no good estimates are available for the precision of the approximations so that we cannot determine whether a particular n in a problem is large enough to safely apply the asymptotic results. To overcome this difficulty, asymptotic results are frequently used in combination with some numerical/empirical studies for selected values of n to examine the *finite sample* performance of asymptotic procedures.

2.5.1 Consistency

A reasonable point estimator is expected to perform better, at least on the average, if more information about the unknown population is available. With a fixed model assumption and sampling plan, more data (larger sample size n) provide more information about the unknown population. Thus, it is distasteful to use a point estimator T_n which, if sampling were to continue indefinitely, could possibly have a nonzero estimation error, although the estimation error of T_n for a fixed n may never equal 0 (see the discussion in §2.4.1).

Definition 2.10 (Consistency of point estimators). Let $X = (X_1, ..., X_n)$ be a sample from $P \in \mathcal{P}$ and $T_n(X)$ be a point estimator of ϑ for every n.
(i) $T_n(X)$ is called *consistent* for ϑ if and only if $T_n(X) \to_p \vartheta$ w.r.t. any

$P \in \mathcal{P}$.
(ii) Let $\{a_n\}$ be a sequence of positive constants diverging to ∞. $T_n(X)$ is called *a_n-consistent* for ϑ if and only if $a_n[T_n(X) - \vartheta] = O_p(1)$ w.r.t. any $P \in \mathcal{P}$.
(iii) $T_n(X)$ is called *strongly consistent* for ϑ if and only if $T_n(X) \to_{a.s.} \vartheta$ w.r.t. any $P \in \mathcal{P}$.
(iv) $T_n(X)$ is called *L_r-consistent* for ϑ if and only if $T_n(X) \to_{L_r} \vartheta$ w.r.t. any $P \in \mathcal{P}$ for some fixed $r > 0$. ∎

Consistency is actually a concept relating to a sequence of estimators, $\{T_n, n = n_0, n_0 + 1, ...\}$, but we usually just say "consistency of T_n" for simplicity. Each of the four types of consistency in Definition 2.10 describes the convergence of $T_n(X)$ to ϑ in some sense, as $n \to \infty$. In statistics, consistency according to Definition 2.10(i), which is sometimes called *weak consistency* since it is implied by any of the other three types of consistency, is the most useful concept of convergence of T_n to ϑ. L_2-consistency is also called *consistency in mse*, which is the most useful type of L_r-consistency.

Example 2.33. Let $X_1, ..., X_n$ be i.i.d. from $P \in \mathcal{P}$. If $\vartheta = \mu$, which is the mean of P (assumed to be finite), then by the SLLN (Theorem 1.13), the sample mean $\bar{X}$ is strongly consistent for μ and, therefore, is also consistent for μ. If we further assume that the variance of P is finite, then by (2.22), $\bar{X}$ is consistent in mse and is $\sqrt{n}$-consistent. With the finite variance assumption, the sample variance S^2 is strongly consistent for the variance of P, according to the SLLN.

Consider estimators of the form $T_n = \sum_{i=1}^n c_{in} X_i$, where $\{c_{in}\}$ is a double array of constants. If P has a finite variance, then by (2.27), T_n is consistent in mse if and only if $\sum_{i=1}^n c_{in} \to 1$ and $\sum_{i=1}^n c_{in}^2 \to 0$. If we only assume the existence of the mean of P, then T_n with $c_{in} = c_i/n$ satisfying $n^{-1} \sum_{i=1}^n c_i \to 1$ and $\sup_i |c_i| < \infty$ is strongly consistent (Theorem 1.13(ii)). ∎

One or a combination of the law of large numbers, the CLT, Slutsky's theorem (Theorem 1.11), and the continuous mapping theorem (Theorems 1.10 and 1.12) are typically applied to establish consistency of point estimators. In particular, Theorem 1.10 implies that if T_n is (strongly) consistent for ϑ and g is a continuous function of ϑ, then $g(T_n)$ is (strongly) consistent for $g(\vartheta)$. For example, in Example 2.33 the point estimator $\bar{X}^2$ is strongly consistent for μ^2. To show that $\bar{X}^2$ is $\sqrt{n}$-consistent under the assumption that P has a finite variance σ^2, we can use the identity

$$\sqrt{n}(\bar{X}^2 - \mu^2) = \sqrt{n}(\bar{X} - \mu)(\bar{X} + \mu)$$

and the fact that $\bar{X}$ is $\sqrt{n}$-consistent for μ and $\bar{X} + \mu = O_p(1)$. (Note that

$\bar{X}^2$ may not be consistent in mse since we do not assume that P has a finite fourth moment.) Alternatively, we can use the fact that $\sqrt{n}(\bar{X}^2 - \mu^2) \to_d N(0, 4\mu^2\sigma^2)$ (by the CLT and Theorem 1.12) to show the $\sqrt{n}$-consistency of $\bar{X}^2$.

The following example shows another way to establish consistency of some point estimators.

Example 2.34. Let $X_1, ..., X_n$ be i.i.d. from an unknown P with a continuous c.d.f. F satisfying $F(\theta) = 1$ for some $\theta \in \mathcal{R}$ and $F(x) < 1$ for any $x < \theta$. Consider the largest order statistic $X_{(n)}$. For any $\epsilon > 0$, $F(\theta - \epsilon) < 1$ and

$$P(|X_{(n)} - \theta| \geq \epsilon) = P(X_{(n)} \leq \theta - \epsilon) = [F(\theta - \epsilon)]^n ,$$

which imply (according to Theorem 1.8(v)), $X_{(n)} \to_{a.s.} \theta$, i.e., $X_{(n)}$ is strongly consistent for θ. Let $F^{(i)}(\theta-)$ be the ith order left-hand derivative of F at θ. If we assume that $F^{(i)}(\theta-)$, $i = 1, ..., m$, exist and vanish, and that $F^{(m+1)}(\theta-)$ exists and is nonzero, then

$$1 - F(X_{(n)}) = \frac{F^{(m+1)}(\theta-)}{(m+1)!}(\theta - X_{(n)})^{m+1} + o\left(|\theta - X_{(n)}|^{m+1}\right) \quad \text{a.s.}$$

Let

$$h_n(\theta) = \left[\frac{(m+1)!}{nF^{(m+1)}(\theta-)}\right]^{(m+1)^{-1}} .$$

For any $t \leq 0$, by Slutsky's theorem,

$$\begin{aligned}
\lim_{n\to\infty} P\left(\frac{X_{(n)} - \theta}{h_n(\theta)} \leq t\right) &= \lim_{n\to\infty} P\left(\left[\frac{\theta - X_{(n)}}{h_n(\theta)}\right]^{m+1} \geq (-t)^{m+1}\right) \\
&= \lim_{n\to\infty} P\left(n[1 - F(X_{(n)})] \geq (-t)^{m+1}\right) \\
&= \lim_{n\to\infty} \left[1 - (-t)^{m+1}/n\right]^n \\
&= e^{-(-t)^{m+1}} .
\end{aligned}$$

This shows that $(X_{(n)} - \theta)/h_n(\theta) \to_d Y$, where Y is a random variable having the c.d.f. $e^{-(-t)^{m+1}} I_{(-\infty,0)}(t)$. Thus, $X_{(n)}$ is $n^{(m+1)^{-1}}$-consistent. If $m = 0$, then $X_{(n)}$ is n-consistent, which is the most common situation. If $m = 1$, then $X_{(n)}$ is $\sqrt{n}$-consistent. ∎

It can be seen from the previous examples that there are many consistent estimators. Like the admissibility in statistical decision theory, consistency is a very essential requirement in the sense that any inconsistent estimators should not be used, but a consistent estimator is not necessarily good. Thus, consistency should be used together with one or a few more criteria.

We now discuss a situation in which finding a consistent estimator is crucial. Suppose that an estimator T_n of ϑ satisfies

$$c_n[T_n(X) - \vartheta] \to_d \sigma Y, \tag{2.38}$$

where Y is a random variable with a known distribution, $\sigma > 0$ is an unknown parameter, and $\{c_n\}$ is a sequence of constants; for example, in Example 2.33, $\sqrt{n}(\bar{X} - \mu) \to_d N(0, \sigma^2)$; in Example 2.34, (2.38) holds with $c_n = n^{(m+1)^{-1}}$ and $\sigma = [(-1)^m (m+1)!/F^{(m+1)}(\theta-)]^{(m+1)^{-1}}$. If a consistent estimator $\hat{\sigma}_n$ of σ can be found, then, by Slutsky's theorem,

$$c_n[T_n(X) - \vartheta]/\hat{\sigma}_n \to_d Y$$

and, thus, we may approximate the distribution of $c_n[T_n(X) - \vartheta]/\hat{\sigma}_n$ by the known distribution of Y.

2.5.2 Asymptotic bias, variance, and mse

Unbiasedness as a criterion for point estimators is discussed in §2.3.2 and §2.4.1. In some cases, however, there is no unbiased estimator (Exercise 69 in §2.6). Furthermore, having a "slight" bias in some cases may not be a bad idea (see Exercise 52 in §2.6). Let $T_n(X)$ be a point estimator of ϑ for every n and $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to \infty$ or $a_n \to a > 0$. If ET_n exists for every n and $\lim_{n\to\infty} a_n E(T_n - \vartheta) = 0$ for any $P \in \mathcal{P}$, then T_n is called *a_n-approximately unbiased* or *approximately unbiased* if $a_n \equiv 1$.

There are many reasonable point estimators whose expectations are not well defined. For example, consider i.i.d. $(X_1, Y_1), ..., (X_n, Y_n)$ from a bivariate normal distribution with $\mu_x = EX_1$ and $\mu_y = EY_1 \neq 0$. Let $\vartheta = \mu_x/\mu_y$ and $T_n = \bar{X}/\bar{Y}$, the ratio of two sample means. Then ET_n is not defined for any n. It is then desirable to define a concept of *asymptotic bias* for point estimators whose expectations are not well defined.

Definition 2.11. Let $\{\xi_n\}$, $\{\gamma_n\}$, and $\{\varepsilon_n\}$ be sequences of random variables such that $P_{\xi_n} = P_{\gamma_n + \varepsilon_n}$ for any n; $E\gamma_n$ exists for any n; $\varepsilon_n = o_p(1)$; and

$$\lim_{n\to\infty} P(|\varepsilon_n| \geq \epsilon|\gamma_n|, \gamma_n \neq 0) = 0 \qquad \text{for any } \epsilon > 0. \tag{2.39}$$

Then $E\gamma_n$ is called an *asymptotic expectation* of ξ_n. ∎

Note that asymptotic expectations of ξ_n, in most cases, are not unique. The following results can be used to find asymptotic expectations.

Proposition 2.3. Let $\{\xi_n\}$, $\{\gamma_n\}$, and $\{\varepsilon_n\}$ be sequences of random variables given in Definition 2.11.

(i) If $\xi_n = \gamma_n + \varepsilon_n$, then $E\gamma_n$ is an asymptotic expectation of ξ_n.
(ii) Let $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to \infty$ or $a_n \to a > 0$. If $a_n\xi_n \to_d Y$, where Y is a random variable with $E|Y| < \infty$, then EY/a_n is an asymptotic expectation of ξ_n.
Proof. (i) is obvious. We now show (ii). According to Theorem 1.8(iv), $a_n\xi_n \to_d Y$ implies that there are ζ_n and Z such that $P_{\zeta_n} = P_{a_n\xi_n}$, $n = 1, 2, ...$, $P_Z = P_Y$, and $\zeta_n = Z + o_p(1)$. Letting $\gamma_n = Z/a_n$ and $\varepsilon_n = (\zeta_n - Z)/a_n$, the result follows from Definition 2.11 and the fact that $EZ = EY$. ∎

Let $T_n(X)$ be a point estimator of ϑ for every n and $\tilde{b}_{T_n}(P)$ be an asymptotic expectation of $T_n - \vartheta$. Then $\tilde{b}_{T_n}(P)$ is called an *asymptotic bias* of T_n and is denoted by $\tilde{b}_{T_n}(\theta)$ if P is in a parametric family. Note that if the exact bias $b_{T_n}(P)$ exists, then it is an asymptotic bias of T_n. Let $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to \infty$ or $a_n \to a > 0$. If $\lim_{n\to\infty} a_n\tilde{b}_{T_n}(P) = 0$ for any $P \in \mathcal{P}$, then T_n is called *a_n-asymptotically unbiased* or *asymptotically unbiased* if $a_n \equiv 1$.

If T_n is a consistent estimator of ϑ, then $T_n = \vartheta + o_p(1)$ and, by Proposition 2.3, T_n is asymptotically unbiased, although T_n may not be approximately unbiased; in fact, $g(T_n)$ is asymptotically unbiased for $g(\vartheta)$ for any continuous function g. For the example of $T_n = \bar{X}/\bar{Y}$, $T_n \to_{a.s.} \mu_x/\mu_y$ by the SLLN. Hence T_n is asymptotically unbiased, although ET_n may not be defined.

It follows from Proposition 2.3 that in Example 2.34, $X_{(n)}$ has an asymptotic bias $\tilde{b}_{T_n}(P) = h_n(\theta)EY$, which is of order $n^{-(m+1)^{-1}}$; in Example 2.33, $\bar{X}^2$ is $\sqrt{n}$-asymptotically unbiased. A more precise result about the asymptotic bias of $\bar{X}^2$ can be obtained using the following result for functions of unbiased estimators.

Theorem 2.6. Let g be a function on $\mathcal{R}^k$ which is second-order differentiable at $\theta \in \mathcal{R}^k$. Suppose that U_{jn} is an unbiased estimator of the jth component of θ, $\mathrm{Var}(U_{jn}) < \infty$, $j = 1, ..., k$, and $U_n - \theta = o_p(1)$, where $U_n = (U_{1n}, ..., U_{kn})$. Then an asymptotic bias of $T_n = g(U_n)$ as an estimator of $\vartheta = g(\theta)$ is $\mathrm{tr}(\nabla^2 g(\theta)\mathrm{Var}(U_n))/2$, where $\mathrm{tr}(M)$ is the trace of a matrix M and $\nabla^2 g(\theta)$ is the matrix of second-order partial derivatives of g at θ.
Proof. Using Taylor's expansion and the fact that $U_n - \theta = o_p(1)$,

$$T_n - \vartheta = \gamma_n + \varepsilon_n,$$

where

$$\gamma_n = \nabla g(\theta)(U_n - \theta)^\tau + \tfrac{1}{2}(U_n - \theta)\nabla^2 g(\theta)(U_n - \theta)^\tau,$$

$\nabla g(\theta)$ denotes the k-vector of partial derivatives of g at θ, and ε_n satisfies

(2.39). The result follows from $EU_n = \theta$ and

$$E\left[(U_n - \theta)\nabla^2 g(\theta)(U_n - \theta)^\tau\right] = \mathrm{tr}\left(\nabla^2 g(\theta) E(U_n - \theta)^\tau (U_n - \theta)\right). \quad \blacksquare$$

Theorem 2.6 can be applied to the case where U_n is the k-vector of sample means, i.e., $U_n = \bar{X} = n^{-1}\sum_{i=1}^n X_i$ with i.i.d. random k-vectors $X_1, ..., X_n$. A similar result for the exact bias of $g(\bar{X})$ is given in Lehmann (1983, Theorem 2.5.1), which requires a much more stringent condition on the derivatives of g.

Example 2.35. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $P(X_i = 1) = p$, where $p \in (0, 1)$ is unknown. Consider first the estimation of $\vartheta = p(1-p)$. Since the sample mean $\bar{X}$ is unbiased for p with $\mathrm{Var}(\bar{X}) = p(1-p)/n$, an asymptotic bias of $T_n = \bar{X}(1-\bar{X})$ according to Theorem 2.6 with $g(x) = x(1-x)$ is $-p(1-p)/n$. On the other hand, a direct computation shows $E[\bar{X}(1-\bar{X})] = E\bar{X} - E\bar{X}^2 = p - (E\bar{X})^2 - \mathrm{Var}(\bar{X}) = p(1-p) - p(1-p)/n$. Hence the exact bias of T_n is $-p(1-p)/n$, which is the same as the asymptotic bias obtained previously.

Consider next the estimation of $\vartheta = p^{-1}$. In this case there is no unbiased estimator of p^{-1} (Exercise 69 in §2.6). Let $T_n = \bar{X}^{-1}$. Then an asymptotic bias of T_n according to Theorem 2.6 with $g(x) = x^{-1}$ is $(1-p)/(pn)$. On the other hand, $ET_n = \infty$ for every n. $\blacksquare$

Like the bias, the mse of an estimator T_n of ϑ is not well defined if the second moment of T_n does not exist. Thus, an asymptotic expectation of $(T_n - \vartheta)^2$ is defined to be an *asymptotic mean squared error* (amse) of T_n, which is denoted by $\mathrm{amse}_{T_n}(P)$ or $\mathrm{amse}_{T_n}(\theta)$ if P is in a parametric family indexed by θ. An *asymptotic variance* of T_n is defined to be

$$\sigma^2_{T_n}(P) = \mathrm{amse}_{T_n}(P) - [\tilde{b}_{T_n}(P)]^2.$$

In many problems $[\tilde{b}_{T_n}(P)]^2/\mathrm{amse}_{T_n}(P) = o(1)$, in which case asymptotic variances are the same as amse's.

If $a_n(T_n - \vartheta) \to_d Y$ with $0 < EY^2 < \infty$, then by Proposition 2.3, $\mathrm{amse}_{T_n}(P) = EY^2/a_n^2$, and $\sigma^2_{T_n}(P) = \mathrm{Var}(Y)/a_n^2$. For instance, in Example 2.34, $\sigma^2_{X_{(n)}}(P) = [h_n(\theta)]^2\mathrm{Var}(Y)$ and $\mathrm{amse}_{X_{(n)}}(P) = [h_n(\theta)]^2 EY^2$; in Example 2.33, $\mathrm{amse}_{\bar{X}^2}(P) = \sigma^2_{\bar{X}^2}(P) = 4\mu^2\sigma^2/n$.

Since amse's of an estimator are not unique, it is not suitable to use them to assess and compare different estimators. Two estimators T_{1n} and T_{2n} may satisfy $\lim_{n\to\infty}[\mathrm{amse}_{T_{1n}}(P)/\mathrm{amse}_{T_{2n}}(P)] < 1$ for one version of amse's, whereas $\lim_{n\to\infty}[\mathrm{amse}_{T_{1n}}(P)/\mathrm{amse}_{T_{2n}}(P)] > 1$ for another version of amse's (Exercise 97 in §2.6). Thus, we need the following definition.

Definition 2.12. Let T_n be an estimator of ϑ for every n and $E\gamma_n^2$ be an amse of T_n.
(i) Suppose that there is a sequence of positive numbers $\{a_n\}$ such that $a_n \to \infty$ and

$$\lim_{t\to\infty}\lim_{n\to\infty} E[\min(a_n^2\gamma_n^2, t)] = \lim_{n\to\infty} a_n^2 E\gamma_n^2 \in (0,\infty). \tag{2.40}$$

Then $E\gamma_n^2$ is called a *regular* amse of T_n and is denoted by $\underline{\text{amse}}_{T_n}(P)$.
(ii) Let T_n' be another estimator of ϑ. The *asymptotic relative efficiency* of T_n' w.t.r. T_n is defined to be

$$e_{T_n',T_n}(P) = \underline{\text{amse}}_{T_n}(P)/\underline{\text{amse}}_{T_n'}(P).$$

(iii) T_n is said to be *asymptotically more efficient* than T_n' if and only if $\limsup_n e_{T_n',T_n}(P) < 1$ for any P. ∎

The following result shows that the regular amse is unique in the limiting sense so that the concepts of asymptotic relative efficiency in Definition 2.12(ii)-(iii) are well defined. It also shows how to find regular amse's and asymptotic relative efficiencies.

Proposition 2.4. Let T_n be an estimator of ϑ for every n and $r_n = \underline{\text{amse}}_{T_n}(P)$ with $\lim_{n\to\infty} a_n^2 r_n \in (0,\infty)$.
(i) If r_n' is another amse of T_n, then $\liminf_n (r_n'/r_n) \geq 1$.
(ii) If both r_n and r_n' are regular amse's of T_n, then $\lim_{n\to\infty}(r_n/r_n') = 1$.
(iii) If $c_n(T_n - \vartheta) \to_d Y$ for a random variable Y with $0 < EY^2 < \infty$ and a sequence $\{c_n\}$ of positive numbers satisfying $c_n \to \infty$, then $EY^2/c_n^2 = \underline{\text{amse}}_{T_n}(P)$.
Proof. The result in (ii) follows from (i) and the result in (iii) follows directly from Definition 2.12. We only need to prove (i). By definition, there exist $\{\gamma_n\}$ and $\{\gamma_n'\}$ such that $r_n = E\gamma_n^2$, $r_n' = E(\gamma_n')^2$, and $P_{(\gamma_n')^2} = P_{\gamma_n^2[1+o_p(1)]}$. Since

$$\min\{(a_n\gamma_n)^2[1+o_p(1)], t\} - \min\{(a_n\gamma_n)^2, t\} = o_p(1)$$

for any $t > 0$ and $\min[(a_n\gamma_n')^2, t]$ and $\min[(a_n\gamma_n)^2, t]$ are bounded by t,

$$E\{\min[(a_n\gamma_n')^2, t]\} = E\{\min[(a_n\gamma_n)^2, t]\} + o(1),$$

which with (2.40) implies that

$$\lim_{t\to\infty}\lim_{n\to\infty} E\{\min[(a_n\gamma_n')^2, t]\} = \lim_{n\to\infty} a_n^2 E\gamma_n^2.$$

The result follows from

$$\lim_{t\to\infty}\lim_{n\to\infty} E\{\min[(a_n\gamma_n')^2, t]\} \leq \liminf_n E(a_n\gamma_n')^2. \quad ∎$$

It follows from Proposition 2.4 that if $\mathrm{mse}_{T_n}(P)$ exists, then

$$\liminf_n [\mathrm{mse}_{T_n}(P)/\underline{\mathrm{amse}}_{T_n}(P)] \geq 1,$$

since $\mathrm{mse}_{T_n}(P)$ is a particular amse of T_n. It is often true that $\mathrm{mse}_{T_n}(P) = \underline{\mathrm{amse}}_{T_n}(P)$, which is implied by the uniform integrability of $\{a_n^2(T_n - \vartheta)^2\}$ for a_n in Definition 2.12 (exercise).

Example 2.36. Let $X_1, ..., X_n$ be i.i.d. from the Poisson distribution $P(\theta)$ with an unknown $\theta > 0$. Consider the estimation of $\vartheta = P(X_i = 0) = e^{-\theta}$. Let $T_{1n} = F_n(0)$, where F_n is the empirical c.d.f. defined in (2.31). Then T_{1n} is unbiased and has $\mathrm{mse}_{T_{1n}}(\theta) = e^{-\theta}(1 - e^{-\theta})/n$. It can be shown that $n^2 E[T_{1n} - F(0)]^4$ is bounded. Hence $\{n[T_{1n} - F(0)]^2\}$ is uniformly integrable and, therefore, $\underline{\mathrm{amse}}_{T_{1n}}(\theta) = \mathrm{mse}_{T_{1n}}(\theta)$.

Next, consider $T_{2n} = e^{-\bar{X}}$. Note that $ET_{2n} = e^{n\theta(e^{-1/n}-1)}$. Hence $nb_{T_{2n}}(\theta) \to \theta e^{-\theta}/2$. Using Theorem 1.12 and the CLT, we can show that

$$\sqrt{n}(T_{2n} - \vartheta) \to_d N(0, e^{-2\theta}\theta).$$

By Proposition 2.4(iii), $\underline{\mathrm{amse}}_{T_{2n}}(\theta) = e^{-2\theta}\theta/n$. Thus, the asymptotic relative efficiency of T_{1n} w.r.t. T_{2n} is

$$e_{T_{1n},T_{2n}}(\theta) = \theta/(e^{\theta} - 1),$$

which is always less than 1. This shows that T_{2n} is asymptotically more efficient than T_{1n}. ∎

The result for T_{2n} in Example 2.36 is a special case (with $U_n = \bar{X}$) of the following general result.

Theorem 2.7. Let g be a function on $\mathcal{R}^k$ which is differentiable at $\theta \in \mathcal{R}^k$ and let U_n be a k-vector of statistics satisfying that $U_n - \theta = o_p(1)$ and $E\|U_n - \theta\|^2 < \infty$. Let $T_n = g(T_n)$ be an estimator of $\vartheta = g(\theta)$. Then
(i) $\sigma^2_{T_n}(P) = \nabla g(\theta)\mathrm{Var}(U_n)[\nabla g(\theta)]^\tau$ and $\mathrm{amse}_{T_n}(P) = E[\nabla g(\theta)(U_n - \theta)^\tau]^2$;
(ii) $\underline{\mathrm{amse}}_{T_n}(P) = E[\nabla g(\theta)Y^\tau]^2/c_n^2$ if $c_n(U_n - \theta) \to_d Y$ for a random k-vector Y with $0 < E\|Y\|^2 < \infty$. ∎

2.5.3 Asymptotic inference

Statistical inference based on asymptotic criteria and approximations is called *asymptotic statistical inference* or simply *asymptotic inference*. We have previously considered asymptotic estimation. We now focus on asymptotic hypothesis tests and confidence sets.

Definition 2.13. Let $X = (X_1, ..., X_n)$ be a sample from $P \in \mathcal{P}$ and $T_n(X)$ be a test for $H_0 : P \in \mathcal{P}_0$ versus $H_1 : P \in \mathcal{P}_1$.
(i) If $\limsup_n \alpha_{T_n}(P) \leq \alpha$ for any $P \in \mathcal{P}_0$, then α is an *asymptotic significance level* of T_n.
(ii) If $\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} \alpha_{T_n}(P)$ exists, then it is called the *limiting size* of T_n.
(iii) T_n is called *consistent* if and only if the type II error probability converges to 0, i.e., $\lim_{n\to\infty}[1 - \alpha_{T_n}(P)] = 0$, for any $P \in \mathcal{P}_1$.
(iv) T_n is called *Chernoff-consistent* if and only if T_n is consistent *and* the type I error probability converges to 0, i.e., $\lim_{n\to\infty} \alpha_{T_n}(P) = 0$, for any $P \in \mathcal{P}_0$. T_n is called *uniformly Chernoff-consistent* if and only if T_n is consistent and the limiting size of T_n is 0. ∎

Obviously if T_n has size (or significance level) α for all n, then its limiting size (or asymptotic significance level) is α. If the limiting size of T_n is $\alpha \in (0, 1)$, then for any $\epsilon > 0$, T_n has size $\alpha + \epsilon$ for all $n \geq n_0$, where n_0 is independent of P. Hence T_n has level of significance $\alpha + \epsilon$ for any $n \geq n_0$. However, if $\mathcal{P}_0$ is not a parametric family, it is likely that the limiting size of T_n is 1 (see, e.g., Example 2.37). This is the reason why we consider the weaker requirement in Definition 2.13(i). If T_n has asymptotic significance level α, then for any $\epsilon > 0$, $\alpha_{T_n}(P) < \alpha + \epsilon$ for all $n \geq n_0(P)$ but $n_0(P)$ depends on $P \in \mathcal{P}_0$; and there is no guarantee that T_n has significance level $\alpha + \epsilon$ for any n.

The consistency in Definition 2.13(iii) only requires that the type II error probability converge to 0. We may define uniform consistency to be $\lim_{n\to\infty} \sup_{P\in\mathcal{P}_1}[1 - \alpha_{T_n}(P)] = 0$, but it is not satisfied in most problems. If $\alpha \in (0, 1)$ is a pre-assigned level of significance for the problem, then a consistent test T_n having asymptotic significance level α is called *asymptotically correct*, and a consistent test having limiting size α is called *strongly* asymptotically correct.

The Chernoff-consistency (or uniformly Chernoff-consistency) in Definition 2.13(iv) requires that both types of error probabilities converge to 0. Mathematically, Chernoff-consistency (or uniform Chernoff-consistency) is better than asymptotic correctness (or strongly asymptotic correctness). After all, both types of error probabilities should decrease to 0 if sampling can be continued indefinitely. However, if α is chosen to be small enough so that error probabilities smaller than α can be practically treated as 0, then the asymptotic correctness (or strongly asymptotic correctness) is enough, and is probably preferred, since requiring an unnecessarily small type I error probability usually results in an unnecessary increase in the type II error probability as the following example illustrates.

Example 2.37. Consider the testing problem $H_0 : \mu \leq \mu_0$ versus H_1 :

$\mu > \mu_0$ based on i.i.d. $X_1, ..., X_n$ with $EX_1 = \mu \in \mathcal{R}$. If each X_i has the $N(\mu, \sigma^2)$ distribution with a known σ^2, then the test T_{c_α} given in Example 2.28 with $c_\alpha = \sigma^{-1}\Phi^{-1}(1-\alpha)/\sqrt{n} + \mu_0$ and $\alpha \in (0,1)$ has size α (and, therefore, limiting size α). It also follows from (2.35) that for any $\mu > \mu_0$,

$$1 - \alpha_{T_{c_\alpha}}(\mu) = \Phi\left(\Phi^{-1}(1-\alpha) + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right) \to 0 \tag{2.41}$$

as $n \to \infty$. This shows that T_{c_α} is consistent and, hence, is strongly asymptotically correct. Note that the convergence in (2.41) is not uniform in $\mu > \mu_0$, but is uniform in $\mu > \mu_1$ for any fixed $\mu_1 > \mu_0$.

Since the size of T_{c_α} is α for all n, T_{c_α} is not Chernoff-consistent. A uniformly Chernoff-consistent test can be obtained as follows. Let $\{\alpha(n)\} \subset (0,1)$ be a sequence satisfying $\alpha(n) \to 0$ and $\sqrt{n}\alpha(n) \to \infty$. Then $T_{c_{\alpha(n)}}$ has size $\alpha(n)$ for each n and, therefore, its limiting size is 0. On the other hand, (2.41) still holds with α replaced by $\alpha(n)$ (exercise). Hence $T_{c_{\alpha(n)}}$ is uniformly Chernoff-consistent. However, if $\alpha(n) < \alpha$, then, from the left-hand side of (2.41), $1 - \alpha_{T_{c_\alpha}}(\mu) < 1 - \alpha_{T_{c_{\alpha(n)}}}(\mu)$ for any $\mu > \mu_0$.

We now consider the case where the population P is not in a parametric family. We still assume that $\sigma^2 = \mathrm{Var}(X_i)$ is known. Using the CLT, we can show that for $\mu > \mu_0$,

$$\lim_{n\to\infty}[1 - \alpha_{T_{c_\alpha}}(\mu)] = \lim_{n\to\infty} \Phi\left(\Phi^{-1}(1-\alpha) + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right) = 0,$$

i.e., T_{c_α} is still consistent. For $\mu \le \mu_0$,

$$\lim_{n\to\infty} \alpha_{T_{c_\alpha}}(\mu) = 1 - \lim_{n\to\infty} \Phi\left(\Phi^{-1}(1-\alpha) + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right),$$

which equals α if $\mu = \mu_0$ and 0 if $\mu < \mu_0$. Thus, the asymptotic significance level of T_{c_α} is α. Combining these two results, we know that T_{c_α} is asymptotically correct. However, if $\mathcal{P}$ contains all possible populations on $\mathcal{R}$, then one can show that the limiting size of T_{c_α} is 1 (exercise). Using the same argument that is used for the normal case, we can also show that $T_{c_{\alpha(n)}}$ is Chernoff-consistent if $\alpha(n) \to 0$ and $\sqrt{n}\alpha(n) \to \infty$. But $T_{c_{\alpha(n)}}$ is not uniformly Chernoff-consistent if $\mathcal{P}$ contains all possible populations on $\mathcal{R}$. ∎

Definition 2.14. Let ϑ be a k-vector of parameters related to the unknown population $P \in \mathcal{P}$ and $C(X)$ be a confidence set for ϑ.
(i) If $\liminf_n P(\vartheta \in C(X)) \ge 1 - \alpha$ for any $P \in \mathcal{P}$, then $1 - \alpha$ is an *asymptotic significance level* of $C(X)$.
(ii) If $\lim_{n\to\infty} \inf_{P\in\mathcal{P}} P(\vartheta \in C(X))$ exists, then it is called the *limiting confidence coefficient* of $C(X)$. ∎

Note that the asymptotic significance level and limiting confidence coefficient of a confidence set are very similar to the asymptotic significance level and limiting size of a test, respectively. Some conclusions are also similar. For example, in a parametric problem one can often find a confidence set having limiting confidence coefficient $1-\alpha \in (0,1)$, which implies that for any $\epsilon > 0$, the confidence coefficient of $C(X)$ is $1-\alpha-\epsilon$ for all $n \geq n_0$, where n_0 is independent of P; in a nonparametric problem the limiting confidence coefficient of $C(X)$ might be 0, whereas $C(X)$ may have asymptotic significance level $1-\alpha \in (0,1)$, but for any fixed n, the confidence coefficient of $C(X)$ might be 0.

The confidence interval in Example 2.31 with $c = \sigma\Phi^{-1}(1-\alpha/2)/\sqrt{n}$ and the confidence set in Example 2.32 have confidence coefficient $1-\alpha$ for any n and, therefore, have limiting confidence coefficient $1-\alpha$. If we drop the normality assumption and assume $EX_i^4 < \infty$, then these confidence sets have asymptotic significance level $1-\alpha$; their limiting confidence coefficients may be 0 (exercise).

2.6 Exercises

1. Consider Example 2.3. Suppose that $p(\boldsymbol{s})$ is constant. Show that X_i and X_j, $i \neq j$, are not uncorrelated and, hence, $X_1, ..., X_n$ are not independent. Furthermore, if y_i's are either 0 or 1, show that $Z = \sum_{i=1}^n X_i$ has a hypergeometric distribution and compute the mean of Z.

2. Consider Example 2.3. Suppose that we do not require that the elements in $\boldsymbol{s}$ be distinct, i.e., we consider sampling with replacement. Define a suitable sample space Ω, a σ-field on Ω, a probability measure p on $(\Omega, \mathcal{F})$, and a sample $(X_1, ..., X_n)$ such that (2.3) holds. If $p(\boldsymbol{s})$ is constant, are $X_1, ..., X_n$ independent? If $p(\boldsymbol{s})$ is constant and y_i's are either 0 or 1, what are the distribution and mean of $Z = \sum_{i=1}^n X_i$?

3. Show that $\{P_\theta : \theta \in \Theta\}$ is an exponential family and find its canonical form, Ξ, and natural parameter space, when
(a) P_θ is the Poisson distribution $P(\theta)$, $\theta \in \Theta = (0, \infty)$;
(b) P_θ is the negative binomial distribution $NB(\theta, r)$ with a fixed r, $\theta \in \Theta = (0,1)$;
(c) P_θ is the exponential distribution $E(a, \theta)$ with a fixed a, $\theta \in \Theta = (0, \infty)$;
(d) P_θ is the gamma distribution $\Gamma(\alpha, \gamma)$, $\theta = (\alpha, \gamma) \in \Theta = (0, \infty) \times (0, \infty)$;
(e) P_θ is the beta distribution $B(\alpha, \beta)$, $\theta = (\alpha, \beta) \in \Theta = (0,1) \times (0,1)$;

(f) P_θ is the Weibull distribution $W(\alpha, \theta)$ with a fixed $\alpha > 0$, $\theta \in \Theta = (0, \infty)$.

4. Show that the family of exponential distributions $E(a, \theta)$ with two unknown parameters a and θ is not an exponential family.

5. Show that the family of negative binomial distributions $NB(p, r)$ with two unknown parameters p and r is not an exponential family.

6. Show that the family of Cauchy distributions $C(\mu, \sigma)$ with two unknown parameters μ and σ is not an exponential family.

7. Show that the family of Weibull distributions $W(\alpha, \theta)$ with two unknown parameters α and θ is not an exponential family.

8. Is the family of log-normal distributions $LN(\mu, \sigma^2)$ with two unknown parameters μ and σ^2 an exponential family?

9. Show that the family of double exponential distributions $DE(\mu, \theta)$ with two unknown parameters μ and θ is not an exponential family, but the family of double exponential distributions $DE(\mu, \theta)$ with a fixed μ and an unknown parameter θ is an exponential family.

10. Show that the k-dimensional normal family discussed in Example 2.4 is an exponential family. Identify the functions T, η, ξ, and h.

11. Obtain the variance-covariance matrix for $(X_1, ..., X_k)$ in Example 2.7, using (a) Theorem 2.1(ii) and (b) direct computation.

12. Show that the m.g.f. of the gamma distribution $\Gamma(\alpha, \gamma)$ is $(1 - \gamma t)^{-\alpha}$, $t < \gamma^{-1}$, using Theorem 2.1(ii).

13. A discrete random variable X with

$$P(X = x) = \gamma(x)\theta^x / c(\theta), \quad x = 0, 1, 2, ...,$$

where $\gamma(x) \geq 0$, $\theta > 0$, and $c(\theta) = \sum_{x=0}^{\infty} \gamma(x)\theta^x$, is called a random variable with a *power series* distribution. Show that power series distributions with $\theta > 0$ form an exponential family and obtain the m.g.f. of X.

14. Let X be a random variable with a p.d.f. f_θ in an exponential family $\{P_\theta : \theta \in \Theta\}$ and let A be an event. Show that the distribution of X truncated on A has a p.d.f. $f_\theta I_A / P_\theta(A)$ which is in an exponential family.

15. Let $\{P_{(\mu,\Sigma)} : \mu \in \mathcal{R}^k, \Sigma \in \mathcal{M}_k\}$ be a location-scale family on $\mathcal{R}^k$. Suppose that $P_{(0,I_k)}$ has a Lebesgue p.d.f. which is always positive and that the mean and variance-covariance matrix of $P_{(0,I_k)}$ are 0 and I_k, respectively. Show that the mean and variance-covariance matrix of $P_{(\mu,\Sigma)}$ are μ and Σ, respectively.

16. Show that if the distribution of a positive random variable X is in a scale family, then the distribution of $\log X$ is in a location family.

17. Let X be a random variable having the gamma distribution $\Gamma(\alpha, \gamma)$ with a known α and an unknown $\gamma > 0$ and let $Y = \sigma \log X$.
(a) Show that if $\sigma > 0$ is unknown, then the distribution of Y is in a location-scale family.
(b) Show that if $\sigma > 0$ is known, then the distribution of Y is in an exponential family.

18. Let $X_1, ..., X_n$ be i.i.d. random variables having a finite $E|X_1|^4$ and let $\bar{X}$ and S^2 be the sample mean and variance defined by (2.1) and (2.2). Express $E(\bar{X}^3)$, $\mathrm{Cov}(\bar{X}, S^2)$, and $\mathrm{Var}(S^2)$ in terms of $\alpha_k = EX_1^k$, $k = 1, 2, 3, 4$. Find a condition under which $\bar{X}$ and S^2 are uncorrelated.

19. Let $X_i = (Y_i, Z_i)$, $i = 1, ..., n$, be i.i.d. random 2-vectors. The statistic $T(X) = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y})(Z_i - \bar{Z}) / \sqrt{S_Y^2 S_Z^2}$ is called the sample correlation coefficient, where $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$ and $\bar{Z} = n^{-1} \sum_{i=1}^n Z_i$ are two sample means, and $S_Y^2 = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$ and $S_Z^2 = (n-1)^{-1} \sum_{i=1}^n (Z_i - \bar{Z})^2$ are two sample variances.
(a) Assume that $E|Y_i|^4 < \infty$ and $E|Z_i|^4 < \infty$. Show that

$$\sqrt{n}[T(X) - \rho] \to_d N(0, c^2),$$

where ρ is the correlation coefficient between Y_1 and Z_1 and c is a constant.
(b) Assume that Y_i and Z_i are independently distributed as $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. Show that T has the following Lebesgue p.d.f.:

$$f(t) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n-2}{2}\right)} (1 - t^2)^{(n-4)/2} I_{(-1,1)}(t).$$

(c) Assume the conditions in (b). Obtain the result in (a) using Scheffé's theorem.

20. Prove the claims in Example 2.9 for the distributions related to order statistics.

21. Let $X_1, ..., X_n$ be i.i.d. random variables having the exponential distribution $E(a, \theta)$, $a \in \mathcal{R}$, and $\theta > 0$. Show that the smallest order

statistic, $X_{(1)}$, has the exponential distribution $E(a, \theta/n)$ and that $2\sum_{i=1}^{n}(X_i - X_{(1)})/\theta$ has the chi-square distribution χ^2_{2n-2}.

22. Show that if T is a sufficient statistic and $T = h(S)$, where h is measurable and S is another statistic, then S is sufficient.

23. In the proof of Theorem 2.2,
(a) show that $C_0 \in \mathcal{C}$;
(b) show that $\mathcal{P}$ is dominated by Q when ν is σ-finite;
(c) show that (2.12) holds.

24. Let $X_1, ..., X_n$ be i.i.d. random variables from $P_\theta \in \{P_\theta : \theta \in \Theta\}$. Find a sufficient statistic for $\theta \in \Theta$ in the following cases.
(a) P_θ is the Poisson distribution $P(\theta)$, $\theta \in (0, \infty)$.
(b) P_θ is the negative binomial distribution $NB(\theta, r)$ with a known r, $\theta \in (0, 1)$.
(c) P_θ is the exponential distribution $E(0, \theta)$, $\theta \in (0, \infty)$.
(d) P_θ is the gamma distribution $\Gamma(\alpha, \gamma)$, $\theta = (\alpha, \gamma) \in (0, \infty) \times (0, \infty)$.
(e) P_θ is the beta distribution $B(\alpha, \beta)$, $\theta = (\alpha, \beta) \in (0, 1) \times (0, 1)$.
(f) P_θ is the log-normal distribution $LN(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2) \in \mathcal{R} \times (0, \infty)$.
(g) P_θ is the Weibull distribution $W(\alpha, \theta)$ with a known $\alpha > 0$, $\theta \in (0, \infty)$.

25. Let $X_1, ..., X_n$ be i.i.d. random variables from $P_{(a,\theta)}$, where $(a, \theta) \in \mathcal{R}^2$ is a parameter. Find a sufficient statistic for (a, θ) in the following cases.
(a) $P_{(a,\theta)}$ is the exponential distribution $E(a, \theta)$, $a \in \mathcal{R}$, $\theta \in (0, \infty)$.
(b) $P_{(a,\theta)}$ is the Pareto distribution $Pa(a, \theta)$, $a \in (0, \infty)$, $\theta \in (0, \infty)$.

26. In Example 2.11, show that $X_{(1)}$ (or $X_{(n)}$) is sufficient for a (or b) if we consider a subfamily $\{f_{(a,b)} : a < b\}$ with a fixed b (or a).

27. Let $X_1, ..., X_n$ be i.i.d. random variables having a distribution $P \in \mathcal{P}$, where $\mathcal{P}$ is the family of distributions on $\mathcal{R}$ having continuous c.d.f.'s. Let $T = (X_{(1)}, ..., X_{(n)})$ be the vector of order statistics. Show that given T, the conditional distribution of $X = (X_1, ..., X_n)$ is a discrete distribution putting probability $1/n!$ on each of the $n!$ points $(X_{i_1}, ..., X_{i_n}) \in \mathcal{R}^n$, where $\{i_1, ..., i_n\}$ is a permutation of $\{1, ..., n\}$; hence, T is sufficient for $P \in \mathcal{P}$.

28. Let X be a sample from $P \in \mathcal{P}$ containing p.d.f.'s f_P w.r.t. a σ-finite measure. Suppose that there is a statistic $T(X)$ such that, for any two sample points x and y, $f_P(x) = f_P(y)\psi(x, y)$ for all P and some measurable function ψ if and only if $T(x) = T(y)$. Show that if $S(X)$ is a statistic sufficient for $P \in \mathcal{P}$, then $T(X) = h(S(X))$ a.s. $\mathcal{P}$ for some function h.

29. Let $X_1, ..., X_n$ be i.i.d. random variables having the Lebesgue p.d.f.

$$f_\theta(x) = \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^4 - \xi(\theta)\right\},$$

where $\theta = (\mu, \sigma) \in \Theta = \mathcal{R} \times (0, \infty)$. Show that $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is an exponential family, where P_θ is the joint distribution of $X_1, ..., X_n$, and that the statistic $T = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i^3, \sum_{i=1}^n X_i^4\right)$ is minimal sufficient for $\theta \in \Theta$.

30. Let $X_1, ..., X_n$ be i.i.d. random variables having the Lebesgue p.d.f.

$$f_\theta(x) = (2\theta)^{-1}\left[I_{(0,\theta)}(x) + I_{(2\theta,3\theta)}(x)\right].$$

Find a minimal sufficient statistic for $\theta \in (0, \infty)$.

31. Let $X_1, ..., X_n$ be i.i.d. random variables having the Cauchy distribution $C(\mu, \sigma)$ with unknown $\mu \in \mathcal{R}$ and $\sigma > 0$. Show that the vector of order statistics is minimal sufficient for (μ, σ).

32. Let $X_1, ..., X_n$ be i.i.d. random variables having the double exponential distribution $DE(\mu, \theta)$ with unknown $\mu \in \mathcal{R}$ and $\theta > 0$. Show that the vector of order statistics is minimal sufficient for (μ, θ).

33. Let $X_1, ..., X_n$ be i.i.d. random variables having the beta distribution $B(\beta, \beta)$ with an unknown $\beta > 0$. Find a minimal sufficient statistic for β.

34. Let $X_1, ..., X_n$ be i.i.d. random variables having a population P in a parametric family indexed by (θ, j), where $\theta > 0$, $j = 1, 2$. When $j = 1$, P is the $N(0, \theta^2)$ distribution and when $j = 2$, P is the double exponential distribution $DE(0, \theta)$. Show that $T = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n |X_i|)$ is minimal sufficient for (θ, j). Is T complete?

35. Let X be a random variable with a distribution P_θ in $\{P_\theta : \theta \in \Theta\}$, f_θ be the p.d.f. of P_θ w.r.t. a measure ν, A be an event, and $\mathcal{P}_A = \{f_\theta I_A / P_\theta(A) : \theta \in \Theta\}$.
(a) Show that if $T(X)$ is sufficient for $P_\theta \in \mathcal{P}$, then it is sufficient for $P_\theta \in \mathcal{P}_A$.
(b) Show that if T is sufficient and complete for $P_\theta \in \mathcal{P}$, then it is complete for $P_\theta \in \mathcal{P}_A$.

36. Show that $(X_{(1)}, X_{(n)})$ in Example 2.13 is not complete.

37. Let T be a complete (or boundedly complete) and sufficient statistic with $E|T| < \infty$. Suppose that there is a minimal sufficient statistic S. Show that T is minimal sufficient and S is complete (or boundedly complete).

38. Let g be a Borel function on $\mathcal{R}^k$. Show that if, for all rectangles $(a_1, b_1) \times \cdots \times (a_k, b_k)$,

$$\int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} g(x_1, ..., x_n) dx_1 \cdots dx_n = 0,$$

then $g = 0$ a.e. Lebesgue.

39. Find complete and sufficient statistics for the families in Exercises 24 and 25.

40. Show that $(X_{(1)}, X_{(n)})$ in Example 2.11 is complete.

41. Let $(X_1, Y_1), ..., (X_n, Y_n)$ be i.i.d. random 2-vectors having the following Lebesgue p.d.f.

$$f_\theta(x, y) = (2\pi\gamma^2)^{-1} I_{(0,\gamma)} \left(\sqrt{(x-a)^2 + (y-b)^2} \right), \quad (x, y) \in \mathcal{R}^2,$$

where $\theta = (a, b, \gamma) \in \mathcal{R}^2 \times (0, \infty)$.
(a) If $a = 0$ and $b = 0$, find a complete and sufficient statistic for γ.
(b) If all parameters are unknown, show that the convex hull of the sample points is a sufficient statistic for θ.

42. Let X be a discrete random variable with p.d.f.

$$f_\theta(x) = \begin{cases} \theta & x = 0 \\ (1-\theta)^2 \theta^{x-1} & x = 1, 2, ... \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta \in (0, 1)$. Show that X is boundedly complete, but not complete.

43. Show that the sufficient statistic T in Example 2.10 is also complete without using Proposition 2.1.

44. Let $Y_1, ..., Y_n$ be i.i.d. random variables having the Lebesgue p.d.f. $\lambda x^{\lambda-1} I_{(0,1)}(x)$ and let $Z_1, ..., Z_n$ be i.i.d. random variables having the geometric distribution $G(p)$ with an unknown p. Assume that Y_i's and Z_j's are independent. Let $X_i = Y_i + Z_i$, which has a distribution in a parametric family indexed by $\theta = (\lambda, p) \in (0, \infty) \times (0, 1)$, $i = 1, ..., n$. Find a complete and sufficient statistic for θ based on the sample $X = (X_1, ..., X_n)$.

45. Suppose that $(X_1, Y_1), ..., (X_n, Y_n)$ are i.i.d. random 2-vectors and X_i and Y_i are independently distributed as $N(\mu, \sigma_X^2)$ and $N(\mu, \sigma_Y^2)$, respectively, with $\theta = (\mu, \sigma_X^2, \sigma_Y^2) \in \mathcal{R} \times (0, \infty) \times (0, \infty)$. Let $\bar{X}$ and

S_X^2 be the sample mean and variance given by (2.1) and (2.2) for X_i's and $\bar{Y}$ and S_Y^2 be the sample mean and variance for Y_i's. Show that $T = (\bar{X}, \bar{Y}, S_X^2, S_Y^2)$ is minimal sufficient for θ but T is not boundedly complete.

46. Let $X_1, ..., X_n$ be i.i.d. from the $N(\theta, \theta^2)$ distribution, where $\theta > 0$ is a parameter. Find a minimal sufficient statistic for θ and show whether it is complete.

47. Suppose that $(X_1, Y_1), ..., (X_n, Y_n)$ are i.i.d. random 2-vectors having the normal distribution with $EX_1 = EY_1 = 0$, $\mathrm{Var}(X_1) = \mathrm{Var}(Y_1) = 1$, and $\mathrm{Cov}(X_1, Y_1) = \theta \in (-1, 1)$.
(a) Find a minimal sufficient statistic for θ.
(b) Show whether the minimal sufficient statistic in (a) is complete or not.
(c) Prove that $T_1 = \sum_{i=1}^n X_i^2$ and $T_2 = \sum_{i=1}^n Y_i^2$ are both ancillary but that (T_1, T_2) is not ancillary.

48. Let $X_1, ..., X_n$ be i.i.d. random variables having the exponential distribution $E(a, \theta)$.
(a) Show that $\sum_{i=1}^n (X_i - X_{(1)})$ and $X_{(1)}$ are independent for any (a, θ).
(b) Show that $Z_i = (X_{(n)} - X_{(i)})/(X_{(n)} - X_{(n-1)})$, $i = 1, ..., n-2$, are independent of $(X_{(1)}, \sum_{i=1}^n (X_i - X_{(1)}))$.

49. Let $X_1, ..., X_n$ be i.i.d. random variables having the gamma distribution $\Gamma(\alpha, \gamma)$. Show that $\sum_{i=1}^n X_i$ and $\sum_{i=1}^n [\log X_i - \log X_{(1)}]$ are independent for any (α, γ).

50. Let $X_1, ..., X_n$ be i.i.d. random variables having the uniform distribution on the interval (a, b), where $-\infty < a < b < \infty$. Show that $(X_{(i)} - X_{(1)})/(X_{(n)} - X_{(1)})$, $i = 2, ..., n-1$, are independent of $(X_{(1)}, X_{(n)})$ for any a and b.

51. Consider Example 2.19.
(a) Show that $\bar{X}$ is better than T_1 if $P = N(\theta, \sigma^2)$, $\theta \in \mathcal{R}$, $\sigma > 0$.
(b) Show that T_1 is better than $\bar{X}$ if P is the uniform distribution on the interval $(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, $\theta \in \mathcal{R}$.
(c) Find a family $\mathcal{P}$ for which neither $\bar{X}$ nor T_1 is better than the other.

52. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution, where $\mu \in \mathcal{R}$ and $\sigma > 0$. Consider the estimation of σ^2 with the squared error loss. Show that $\frac{n-1}{n} S^2$ is better than S^2, the sample variance. Can you find an estimator of the form cS^2 with a nonrandom c such that it is better than $\frac{n-1}{n} S^2$?

53. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $P(X_i = 1) = \theta \in (0, 1)$. Consider estimating θ with the squared error loss. Calculate the risks of the following estimators:
(a) the nonrandomized estimators $\bar{X}$ (the sample mean) and

$$T_0(X) = \begin{cases} 0 & \text{if more than half of } X_i\text{'s are } 0 \\ 1 & \text{if more than half of } X_i\text{'s are } 1 \\ \frac{1}{2} & \text{if exactly half of } X_i\text{'s are } 0; \end{cases}$$

(b) the randomized estimators

$$T_1(X) = \begin{cases} \bar{X} & \text{with probability } \frac{1}{2} \\ T_0 & \text{with probability } \frac{1}{2} \end{cases}$$

and

$$T_2(X) = \begin{cases} \bar{X} & \text{with probability } \bar{X} \\ \frac{1}{2} & \text{with probability } 1 - \bar{X}. \end{cases}$$

54. Let $X_1, ..., X_n$ be i.i.d. random variables having the exponential distribution $E(0, \theta)$, $\theta \in (0, \infty)$. Consider estimating θ with the squared error loss. Calculate the risks of the sample mean $\bar{X}$ and $cX_{(1)}$, where c is a positive constant. Is $\bar{X}$ better than $cX_{(1)}$ for some c?

55. Let $X_1, ..., X_n$ be i.i.d. random variables having the exponential distribution $E(0, \theta)$, $\theta \in (0, \infty)$. Consider the hypotheses

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0,$$

where $\theta_0 > 0$ is a fixed constant. Obtain the risk function (in terms of θ) of the test rule $T_c(X) = I_{(c,\infty)}(\bar{X})$, under the 0-1 loss.

56. Let $X_1, ..., X_n$ be i.i.d. random variables having the Cauchy distribution $C(\mu, \sigma)$ with unknown $\mu \in \mathcal{R}$ and $\sigma > 0$. Consider the hypotheses

$$H_0 : \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0,$$

where μ_0 is a fixed constant. Obtain the risk function of the test rule $T_c(X) = I_{(c,\infty)}(\bar{X})$, under the 0-1 loss.

57. Consider Example 2.21. Suppose that our decision rule, based on a sample $X = (X_1, ..., X_n)$ with i.i.d. components from the $N(\theta, 1)$ distribution with an unknown $\theta > 0$, is

$$T(X) = \begin{cases} a_1 & b_1 < \bar{X} \\ a_2 & b_0 < \bar{X} \leq b_1 \\ a_3 & \bar{X} \leq b_0. \end{cases}$$

Express the risk of T in terms of θ.

58. Consider an estimation problem with $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ (a parametric family), $\mathbb{A} = \Theta$, and the squared error loss. If $\theta_0 \in \Theta$ satisfies that $P_\theta \ll P_{\theta_0}$ for any $\theta \in \Theta$, show that the estimator $T \equiv \theta_0$ is admissible.

59. Let $\Im$ be a class of decision rules. A subclass $\Im_0 \subset \Im$ is called $\Im$-*complete* if, for any $T \in \Im$ and $T \notin \Im_0$, there is a $T_0 \in \Im_0$ that is better than T, and is called $\Im$-*minimal complete* if $\Im_0$ is $\Im$-complete and no proper subclass of $\Im_0$ is $\Im$-complete. Show that if a $\Im$-minimal complete class exists, then it is exactly the class of $\Im$-admissible rules.

60. Let $X_1, ..., X_n$ be i.i.d. random variables having a distribution $P \in \mathcal{P}$. Assume that $EX_1^2 < \infty$. Consider estimating $\mu = EX_1$ under the squared error loss.
(a) Show that any estimator of the form $a\bar{X} + b$ is inadmissible, where $\bar{X}$ is the sample mean, $a > 1$ and b are constants.
(b) Show that any estimator of the form $\bar{X} + b$ is inadmissible, where $b \neq 0$ is a constant.

61. Consider an estimation problem with $\vartheta \in [a, b] \subset \mathcal{R}$, where a and b are known constants. Suppose that the action space is $\mathbb{A} \supset [a, b]$ and the loss function is $L(|\vartheta - a|)$, where $L(\cdot)$ is an increasing function on $[0, \infty)$. Show that any decision rule T with $P(T(X) \notin [a, b]) > 0$ for some $P \in \mathcal{P}$ is inadmissible.

62. Show that the following functions of x are convex and discuss whether they are strictly convex.
(a) $|x - a|^p$, where $p \geq 1$ and $a \in \mathcal{R}$.
(b) x^{-p}, $x \in (0, \infty)$, where $p > 0$.
(c) e^{cx}, where $c \in \mathcal{R}$.
(d) $-\log x$, $x \in (0, \infty)$.
(e) $\psi(\phi(x))$, $x \in (a, b)$, where $-\infty \leq a < b \leq \infty$, ϕ is convex on (a, b), and ψ is convex and nondecreasing on the range of ϕ.
(f) $\phi(x) = \sum_{i=1}^k c_i \phi_i(x_i)$, $x = (x_1, ..., x_k) \in \prod_{i=1}^k \mathfrak{X}_i$, where c_i is a positive constant and ϕ_i is convex on $\mathfrak{X}_i$, $i = 1, ..., k$.

63. Prove Theorem 2.5.

64. In Exercise 53, use Theorem 2.5 to find decision rules that are better than T_j, $j = 0, 1, 2$.

65. In Exercise 54, use Theorem 2.5 to find a decision rule better than $cX_{(1)}$.

66. Consider Example 2.22.
(a) Show that there is no optimal rule if $\Im$ contains all possible estimators. (Hint: consider constant estimators.)

(b) Find a $\Im_2$-optimal rule if $X_1, ..., X_n$ are independent random variables having a common mean μ and $\mathrm{Var}(X_i) = \sigma^2/a_i$ with known a_i, $i = 1, ..., n$.
(c) Find a $\Im_2$-optimal rule if $X_1, ..., X_n$ are identically distributed but are correlated with a common correlation coefficient ρ.

67. Let $T_0(X)$ be an unbiased estimator of ϑ in an estimation problem. Show that any unbiased estimator of ϑ is of the form $T(X) = T_0(X) - U(X)$, where $U(X)$ is an "unbiased estimator" of 0.

68. Let X be a discrete random variable with

$$P(X = -1) = p, \quad P(X = k) = (1-p)^2 p^k, \quad k = 0, 1, 2, ...,$$

where $p \in (0, 1)$ is unknown.
(a) Show that $U(X)$ is an unbiased estimator of 0 if and only if $U(k) = ak$ for all $k = -1, 0, 1, 2, ...$ and some a.
(b) Show that $T_0(X) = I_{\{0\}}(X)$ is unbiased for $\vartheta = (1-p)^2$ and that, under the squared error loss, T_0 is a $\Im$-optimal rule, where $\Im$ is the class of all unbiased estimators of ϑ.
(c) Show that $T_0(X) = I_{\{-1\}}(X)$ is unbiased for $\vartheta = p$ and that, under the squared error loss, there is no $\Im$-optimal rule, where $\Im$ is the class of all unbiased estimators of ϑ.

69. (Nonexistence of an unbiased estimator). Let X be a random variable having the binomial distribution $Bi(p, n)$ with an unknown $p \in (0, 1)$ and a known n. Consider the problem of estimating $\vartheta = p^{-1}$. Show that there is no unbiased estimator of ϑ.

70. Let $X_1, ..., X_n$ be i.i.d. from the Poisson distribution $P(\theta)$ with an unknown $\theta > 0$. Find the bias and mse of $T_n = (1 - a/n)^{n\bar{X}}$ as an estimator of $\vartheta = e^{-a\theta}$, where $a \neq 0$ is a known constant.

71. Consider a location family $\{P_\mu : \mu \in \mathcal{R}^k\}$ on $\mathcal{R}^k$, where $P_\mu = P_{(\mu, I_k)}$ given in (2.10). Let $\mathcal{T} = \{I_k\}$, $\mathcal{C} = \{cl_0 : c \in \mathcal{R}\}$, where $l_0 \in \mathcal{R}^k$ is fixed, and $L(P, a) = L(\|\mu - a\|)$, where $L(\cdot)$ is a nonnegative Borel function on $[0, \infty)$. Show that the family is invariant and the decision problem is invariant with $g_{c,A}(a) = g_c(a) = a + cl_0$. Find an invariant decision rule.

72. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with unknown $\mu \in \mathcal{R}$ and $\sigma^2 > 0$. Consider the scale transformation aX, $a \in (0, \infty)$.
(a) For estimating σ^2 under the loss function $L(P, a) = (1 - a/\sigma^2)^2$, show that the problem is invariant and that the sample variance S^2 is invariant.

(b) For testing $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$ under the loss

$$L(P, 0) = \frac{\mu}{\sigma} I_{(0,\infty)}(\mu) \quad \text{and} \quad L(P, 1) = \frac{|\mu|}{\sigma} I_{(-\infty,0]}(\mu),$$

show that the problem is invariant and any test that is a function of $\bar{X}/\sqrt{S^2/n}$ is invariant.

73. Let $X_1, ..., X_n$ be i.i.d. random variables having the c.d.f. $F(x - \theta)$, where F is symmetric about 0 and $\theta \in \mathcal{R}$ is unknown.
(a) Show that the c.d.f. of $\sum_{i=1}^n w_i X_{(i)} - \theta$ is symmetric about 0, where $X_{(i)}$ is the ith order statistic and w_i's are constants satisfying $w_i = w_{n-i+1}$ and $\sum_{i=1}^n w_i = 1$.
(b) Show that $\sum_{i=1}^n w_i X_{(i)}$ in (a) is unbiased for θ if the mean of F exists.
(c) Show that $\sum_{i=1}^n w_i X_{(i)}$ is location invariant when $\sum_{i=1}^n w_i = 1$.

74. In Example 2.25, show that the conditional distribution of $\boldsymbol{\theta}$ given $X = x$ is $N(\mu_*(x), c^2)$ with $\mu_*(x)$ and c^2 given by (2.28).

75. A *median* of a random variable Y (or its distribution) is any value m such that $P(Y \leq m) \geq \frac{1}{2}$ and $P(Y \geq m) \geq \frac{1}{2}$.
(a) Show that the set of medians is a closed interval $[m_0, m_1]$.
(b) Suppose that $E|Y| < \infty$. If c is not a median of Y, show that $E|Y - c| \geq E|Y - m|$ for any median m of Y.
(c) Let X be a sample from P_θ, where $\theta \in \Theta \subset \mathcal{R}$. Consider the estimation of θ under the absolute error loss function $|a - \theta|$. Let Π be a given distribution on Θ with finite mean. Find the $\Im$-Bayes rule w.r.t. Π, where $\Im$ is the class of all rules.

76. In Example 2.27, show that $\hat{Y}$ is still unbiased if sampling is with replacement (see Exercise 2), and find the variance of $\hat{Y}$.

77. Let $X_1, ..., X_n$ be i.i.d. random variables having the uniform distribution $U(0, \theta)$, where $\theta > 0$ is unknown. Calculate the bias and mse of $cX_{(n)}$ as an estimator of θ, where c is a positive constant. Find a c such that $cX_{(n)}$ is unbiased for θ.

78. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with unknown $\mu \in \mathcal{R}$ and $\sigma^2 > 0$. Consider estimating $\vartheta = \mu^2$. Calculate the bias and the mse of $\bar{X}^2$ as an estimator of ϑ. Find an unbiased estimator of ϑ based on the complete and sufficient statistic $(\bar{X}, S^2)$ and compare its mse with that of $\bar{X}^2$.

79. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with unknown $\mu \in \mathcal{R}$ and $\sigma^2 > 0$. To test the hypotheses

$$H_0 : \mu \leq \mu_0 \qquad \text{versus} \qquad H_1 : \mu > \mu_0,$$

where μ_0 is a fixed constant, consider a test of the form $T_c(X) = I_{(c,\infty)}(T_{\mu_0})$, where $T_{\mu_0} = (\bar{X} - \mu_0)/\sqrt{S^2/n}$ and c is a fixed constant.
(a) Find the size of T_c. (Hint: $(\bar{X} - \mu_0)/\sqrt{S^2/n}$ has the t-distribution t_{n-1}.)
(b) If α is a given level of significance, find a c_α such that T_{c_α} has size α.
(c) Compute the p-value for T_{c_α} derived in (b).
(d) Find a c_α such that $[\bar{X} - c_\alpha\sqrt{S^2/n}, \bar{X} + c_\alpha\sqrt{S^2/n}]$ is a confidence interval for μ with confidence coefficient $1 - \alpha$. What is the expected interval length?

80. In Exercise 55, calculate the size of $T_c(X)$; find a c_α such that T_{c_α} has size α, a given level of significance; and find the p-value for T_{c_α}.

81. In Exercise 56, assume that σ is known. Calculate the size of $T_c(X)$; find a c_α such that T_{c_α} has size α, a given level of significance; and find the p-value for T_{c_α}.

82. Let $\alpha \in (0, 1)$ be given and $T_{j,q}(X)$ be the test given in Example 2.30. Show that there exist integer j and $q \in (0, 1)$ such that the size of $T_{j,q}$ is α.

83. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(a, \theta)$ with unknown $a \in \mathcal{R}$ and $\theta > 0$. Let $\alpha \in (0, 1)$ be given.
(a) Using $T_1(X) = \sum_{i=1}^n (X_i - X_{(1)})$, construct a confidence interval for θ with confidence coefficient $1 - \alpha$ and find the expected interval length.
(b) Using $T_1(X)$ and $T_2(X) = X_{(1)}$, construct a confidence interval for a with confidence coefficient $1 - \alpha$ and find the expected interval length.
(c) Using the method in Example 2.32, construct a confidence set for the two-dimensional parameter (a, θ) with confidence coefficient $1-\alpha$.

84. Suppose that X is a sample and a statistic $T(X)$ has a distribution in a location family $\{P_\mu : \mu \in \mathcal{R}\}$. Using $T(X)$, derive a confidence interval for μ with level of significance $1 - \alpha$ and obtain the expected interval length. Show that if the c.d.f. of $T(X)$ is continuous, then we can always find a confidence interval for μ with confidence coefficient $1 - \alpha$ for any $\alpha \in (0, 1)$.

85. Let $X = (X_1, ..., X_n)$ be a sample from P_θ, where $\theta \in \{\theta_1, ..., \theta_k\}$ with a fixed integer k. Let $T_n(X)$ be an estimator of θ with range $\{\theta_1, ..., \theta_k\}$.
(a) Show that $T_n(X)$ is consistent if and only if $P_\theta(T_n(X) = \theta) \to 1$.
(b) Show that if $T_n(X)$ is consistent, then it is a_n-consistent for any $\{a_n\}$.

86. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on $(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, where $\theta \in \mathcal{R}$ is unknown. Show that $(X_{(1)} + X_{(n)})/2$ is strongly consistent for θ and also consistent in mse.

87. Let $X_1, ..., X_n$ be i.i.d. from a population with the Lebesgue p.d.f. $f_\theta(x) = 2^{-1}(1 + \theta x)I_{(-1,1)}(x)$, where $\theta \in (-1, 1)$ is an unknown parameter. Find a consistent estimator of θ. Is your estimator $\sqrt{n}$-consistent?

88. Suppose that T_n is an unbiased estimator of ϑ such that for any n, $\mathrm{Var}(T_n) < \infty$ and $\mathrm{Var}(T_n) \leq \mathrm{Var}(U_n)$ for any other unbiased estimator of ϑ. Show that T_n is consistent in mse.

89. Consider the Bayes rule $\mu_*(X)$ in Example 2.25. Show that $\mu_*(X)$ is a strongly consistent, $\sqrt{n}$-consistent, and L_2-consistent estimator of μ. What is the order of the bias of $\mu_*(X)$ as an estimator of μ?

90. Show that the estimator T_0 of θ in Exercise 53 is inconsistent.

91. Let g_1, g_2,... be continuous functions on $(a, b) \subset \mathcal{R}$ such that $g_n(x) \to g(x)$ uniformly for x in any closed subinterval of (a, b). Let T_n be a consistent estimator of $\theta \in (a, b)$. Show that $g_n(T_n)$ is consistent for $\vartheta = g(\theta)$.

92. Let $X_1, ..., X_n$ be i.i.d. from P with unknown mean $\mu \in \mathcal{R}$ and variance $\sigma^2 > 0$, and let $g(\mu) = 0$ if $\mu \neq 0$ and $g(0) = 1$. Find a consistent estimator of $\vartheta = g(\mu)$.

93. Establish results for the smallest order statistic $X_{(1)}$ (based on i.i.d. random variables $X_1, ..., X_n$) similar to those in Example 2.34.

94. (Consistency for finite population). In Example 2.27, show that $\hat{Y} \to_p Y$ as $n \to N$ for any fixed N and population. Is $\hat{Y}$ still consistent if sampling is with replacement?

95. Assume that $X_i = \theta t_i + e_i$, $i = 1, ..., n$, where $\theta \in \Theta$ is an unknown parameter, Θ is a closed subset of $\mathcal{R}$, e_i's are i.i.d. on the interval $[-\tau, \tau]$ with some unknown $\tau > 0$, and t_i's are fixed constants. Let

$$T_n = S_n(\tilde{\theta}_n) = \min_{\gamma \in \Theta} S_n(\gamma),$$

where

$$S_n(\gamma) = 2 \max_{i \leq n} |X_i - \gamma t_i| / \sqrt{1 + \gamma^2}.$$

(a) Assume that $\sup_i |t_i| < \infty$ and $\sup_i t_i - \inf_i t_i > 2\tau$. Show that the sequence $\{\tilde{\theta}_n, n = 1, 2, ...\}$ is bounded.
(b) Let $\theta_n \in \Theta$, $n = 1, 2, ...$ If $\theta_n \to \theta$, show that

$$S_n(\theta_n) - S_n(\theta) = O(|\theta_n - \theta|).$$

(c) Under the conditions in (a), show that T_n is a strongly consistent estimator of $\vartheta = \min_{\gamma\in\Theta} S(\gamma)$, where $S(\gamma) = \lim_{n\to\infty} S_n(\gamma)$ a.s.

96. Let $X_1, ..., X_n$ be i.i.d. from P with $EX_1^4 < \infty$ and unknown mean $\mu \in \mathcal{R}$ and variance $\sigma^2 > 0$. Consider the estimation of $\vartheta = \mu^2$ and the following three estimators: $T_{1n} = \bar{X}^2$, $T_{2n} = \bar{X}^2 - S^2/n$, $T_{3n} = \max(0, T_{2n})$, where $\bar{X}$ and S^2 are the sample mean and variance.
(a) Obtain n^{-1} order asymptotic biases for T_{jn}, $j = 1, 2, 3$.
(b) Show that the regular amse's of T_{jn}, $j = 1, 2, 3$, are the same when $\mu \neq 0$ but may be different when $\mu = 0$. Which estimator has the smallest limiting regular amse when $\mu = 0$?

97. Let T_n be an estimator of ϑ satisfying $\sqrt{n}(T_n - \vartheta) \to_d N(0, \sigma^2)$ for some σ^2 and $\lim_{n\to\infty} n\text{mse}_{T_n}(P) > \sigma^2$. Let $T_{cn} = T_n + \xi_n(c)/\sqrt{n}$, where $\xi_n(c)$ is a random variable independent of T_n, $\xi_n(c) = c$ with probability $1 - n^{-1}$ and $\xi_n(c) = Z$ with probability n^{-1}, c is a fixed constant, and Z is a random variable having a Cauchy distribution.
(a) Show that $\text{amse}_{T_n}(P)/\text{amse}_{T_{cn}}(P) \to \sigma^2/(c^2 + \sigma^2)$.
(b) Show that $\text{mse}_{T_{cn}}(P)$ is not defined for any c and n.
(c) Show that $\text{mse}_{T_n}(P)/\text{amse}_{T_{cn}}(P)$ converges to a constant larger than 1 for some c.

98. Let $X_1, ..., X_n$ be i.i.d. according to $N(\mu, 1)$ with an unknown $\mu \in \mathcal{R}$. Let $\vartheta = P(X_1 \leq c)$ for a fixed constant c. Consider the following estimators of ϑ: $T_{1n} = F_n(c)$, where F_n is the empirical c.d.f. defined in (2.31), and $T_{2n} = \Phi(c - \bar{X})$, where Φ is the c.d.f. of $N(0, 1)$.
(a) Find an n^{-1} order asymptotic bias of T_{2n}.
(b) Find the asymptotic relative efficiency of T_{1n} w.r.t. T_{2n}.

99. Let $X_1, ..., X_n$ be i.i.d. from the $N(0, \sigma^2)$ distribution with an unknown $\sigma > 0$. Consider the estimation of $\vartheta = \sigma$. Find the asymptotic relative efficiency of $\sqrt{\pi/2}\sum_{i=1}^n |X_i|/n$ w.r.t. $(\sum_{i=1}^n X_i^2/n)^{1/2}$.

100. Show that if $E\gamma_n^2 = \text{amse}_{T_n}(P)$, $\lim_{n\to\infty} a_n^2 E\gamma_n^2$ exists, and $\{a_n^2\gamma_n^2\}$ is uniformly integrable, then $E\gamma_n^2 = \underline{\text{amse}}_{T_n}(P)$.

101. Prove Theorem 2.7.

102. Let $X_1, ..., X_n$ be i.i.d. with $EX_i = \mu$, $\text{Var}(X_i) = 1$, and $EX_i^4 < \infty$. Let $T_{1n} = n^{-1}\sum_{i=1}^n X_i^2 - 1$ and $T_{2n} = \bar{X}^2 - n^{-1}$ be estimators of $\vartheta = \mu^2$.
(a) Find the asymptotic relative efficiency of T_{1n} w.r.t. T_{2n}.
(b) Show that $e_{T_{1n},T_{2n}}(P) \leq 1$ if the c.d.f. of $X_i - \mu$ is symmetric about 0.
(c) Find a distribution P for which $e_{T_{1n},T_{2n}}(P) > 1$.

103. Let $X_1, ..., X_n$ be i.i.d. binary random variables with unknown $p = P(X_i = 1) \in (0, 1)$. Consider the estimation of p. Let a and b be two positive constants. Find the asymptotic relative efficiency of the estimator $(a + n\bar{X})/(a + b + n)$ w.r.t. $\bar{X}$.

104. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution $U(0, \theta)$. Consider the following estimators of θ: $T_{1n} = (n + 1)X_{(n)}/n$ and $T_{2n} = X_{(n)}$. Find the biases $b_{T_{jn}}(\theta)$, $j = 1, 2$, and $e_{T_{1n}, T_{2n}}(\theta)$.

105. In Example 2.37, show that
(a) (2.41) holds with α replaced by $\alpha(n)$ satisfying $\alpha(n) \to 0$ and $\sqrt{n}\alpha(n) \to \infty$;
(b) the limiting size of T_{c_α} is 1 if $\mathcal{P}$ contains all possible populations on $\mathcal{R}$.

106. Let $X_1, ..., X_n$ be i.i.d. with unknown mean $\mu \in \mathcal{R}$ and variance $\sigma^2 > 0$. For testing $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$, consider the test T_{c_α} obtained in Exercise 79(b).
(a) Show that T_{c_α} has asymptotic significance level α and is consistent.
(b) Find a test that is Chernoff-consistent.

107. Consider the test T_j in Example 2.23. For each n, find a $j = j_n$ such that T_{j_n} has asymptotic significance level $\alpha \in (0, 1)$.

108. Show that the test T_{c_α} in Exercise 80 is consistent, but T_{c_α} in Exercise 81 is not consistent.

109. In Example 2.31, suppose that we drop the normality assumption but assume that $\mu = EX_i$ and $\sigma^2 = \mathrm{Var}(X_i)$ is known.
(a) Show that the asymptotic significance level of the confidence interval $[\bar{X} - c_\alpha, \bar{X} + c_\alpha]$, $c_\alpha = \sigma\Phi^{-1}(1 - \alpha/2)/\sqrt{n}$ is $1 - \alpha$.
(b) Show that the limiting confidence coefficient of the interval in (a) might be 0 if $\mathcal{P}$ contains all possible populations on $\mathcal{R}$.

110. Let $X_1, ..., X_n$ be i.i.d. with unknown mean $\mu \in \mathcal{R}$ and variance $\sigma^2 > 0$. Show that the confidence interval in Exercise 79(d) has asymptotic significance level $1 - \alpha$.

111. Let $X_1, ..., X_n$ be i.i.d. with unknown mean $\mu \in \mathcal{R}$ and variance $\sigma^2 > 0$. Assume that $EX_1^4 < \infty$. Using the sample variance S^2, construct a confidence interval for σ^2 that has asymptotic significance level $1 - \alpha$.

112. Consider the sample correlation coefficient T defined in Exercise 19. Construct a confidence interval for ρ that has asymptotic significance level $1 - \alpha$, assuming that (Y_i, Z_i) are normally distributed. (Hint: show that the asymptotic variance of T is $(1 - \rho^2)^2$.)

Chapter 3

Unbiased Estimation

Unbiased or asymptotically unbiased estimation plays an important role in point estimation theory. Unbiasedness of point estimators is defined in §2.3. In this chapter we discuss in detail how to derive unbiased estimators and, more importantly, how to find the best unbiased estimators in various situations. Although an unbiased estimator (even the best unbiased estimator if it exists) is not necessarily better than a slightly biased estimator in terms of their mse's (see Exercise 52 in §2.6), unbiased estimators can be used as "building blocks" for the construction of better estimators. Furthermore, one may give up the exact unbiasedness, but cannot give up asymptotic unbiasedness since it is necessary for consistency (see §2.5). Properties and the construction of asymptotically unbiased estimators are studied in the last part of this chapter.

3.1 The UMVUE

Let X be a sample from an unknown population $P \in \mathcal{P}$ and ϑ be a real-valued parameter related to P. Recall that an estimator $T(X)$ of ϑ is unbiased if $E[T(X)] = \vartheta$ for any $P \in \mathcal{P}$. If there exists an unbiased estimator of ϑ, then ϑ is called an *estimable* parameter.

Definition 3.1. An unbiased estimator $T(X)$ of ϑ is called the *uniformly minimum variance unbiased estimator* (UMVUE) if and only if $\mathrm{Var}(T(X)) \leq \mathrm{Var}(U(X))$ for any $P \in \mathcal{P}$ and any other unbiased estimator $U(X)$ of ϑ. ∎

Since the mse of any unbiased estimator is its variance, a UMVUE is $\Im$-optimal in mse with $\Im$ being the class of all unbiased estimators. One

can similarly define the uniformly minimum risk unbiased estimator in statistical decision theory when we use an arbitrary loss instead of the squared error loss that corresponds to the mse.

3.1.1 Sufficient and complete statistics

The derivation of a UMVUE is relatively simple if there exists a sufficient and complete statistic for $P \in \mathcal{P}$.

Theorem 3.1. Suppose that there exists a sufficient and complete statistic $T(X)$ for $P \in \mathcal{P}$. If ϑ is estimable, then there is a unique unbiased estimator of ϑ that is of the form $h(T)$ with a Borel function h. (Two estimators that are equal a.s. $\mathcal{P}$ are treated as one estimator.) Furthermore, $h(T)$ is the unique UMVUE of ϑ. ∎

This theorem is a consequence of Theorem 2.5(ii) (Rao-Blackwell's theorem). One can easily extend this theorem to the case of uniformly minimum risk unbiased estimator under any loss function $L(P, a)$ which is strictly convex in a. The uniqueness of the UMVUE follows from the completeness of $T(X)$.

There are two typical ways to derive a UMVUE when a sufficient and complete statistic T is available. The first one is solving for h when the distribution of T is available. The following are two typical examples.

Example 3.1. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on $(0, \theta)$, $\theta > 0$. Let $\vartheta = g(\theta)$, where g is a differentiable function on $(0, \infty)$. Since the sufficient and complete statistic $X_{(n)}$ has the Lebesgue p.d.f. $n\theta^{-n}x^{n-1}I_{(0,\theta)}(x)$, an unbiased estimator $h(X_{(n)})$ of ϑ must satisfy

$$\theta^n g(\theta) = n\int_0^\theta h(x)x^{n-1}dx \qquad \text{for all } \theta > 0.$$

Assuming that h is continuous and differentiating both sizes of the previous equation lead to

$$n\theta^{n-1}g(\theta) + \theta^n g'(\theta) = nh(\theta)\theta^{n-1}.$$

Hence, the UMVUE of ϑ is $h(X_{(n)}) = g(X_{(n)}) + n^{-1}X_{(n)}g'(X_{(n)})$. In particular, if $\vartheta = \theta$, then the UMVUE of θ is $(1 + n^{-1})X_{(n)}$. ∎

Example 3.2. Let $X_1, ..., X_n$ be i.i.d. from the Poisson distribution $P(\theta)$ with an unknown $\theta > 0$. Then $T(X) = \sum_{i=1}^n X_i$ is sufficient and complete for $\theta > 0$ and has the Poisson distribution $P(n\theta)$. Suppose that $\vartheta = g(\theta)$, where g is a smooth function such that $g(x) = \sum_{j=0}^\infty a_j x^j$, $x > 0$. An

unbiased estimator $h(T)$ of ϑ must satisfy

$$\begin{aligned}\sum_{t=0}^{\infty}\frac{h(t)n^t}{t!}\theta^t &= e^{n\theta}g(\theta)\\ &= \sum_{k=0}^{\infty}\frac{n^k}{k!}\theta^k\sum_{j=0}^{\infty}a_j\theta^j\\ &= \sum_{t=0}^{\infty}\left(\sum_{j,k:j+k=t}\frac{n^k a_j}{k!}\right)\theta^t\end{aligned}$$

for any $\theta > 0$. Thus, a comparison of coefficients in front of θ^t leads to

$$h(t) = \frac{t!}{n^t}\sum_{j,k:j+k=t}\frac{n^k a_j}{k!},$$

i.e., $h(T)$ is the UMVUE of ϑ. In particular, if $\vartheta = \theta^r$ for some fixed integer $r \geq 1$, then $a_r = 1$ and $a_k = 0$ if $k \neq r$ and

$$h(t) = \begin{cases} 0 & t < r \\ \frac{t!}{n^r(t-r)!} & t \geq r. \end{cases} \quad \blacksquare$$

The second method of deriving a UMVUE when there is a sufficient and complete statistic $T(X)$ is conditioning on T, i.e., if $U(X)$ is any unbiased estimator of ϑ, then $E[U(X)|T]$ is the UMVUE of ϑ. To apply this method, we do not need the distribution of T, but need to work out the conditional expectation $E[U(X)|T]$. From the uniqueness of the UMVUE, it does not matter which $U(X)$ is used and, thus, we should choose $U(X)$ so as to make the calculation of $E[U(X)|T]$ as easy as possible.

Example 3.3. Consider the estimation problem in Example 2.26, where $\vartheta = 1 - F_\theta(t)$ and $F_\theta(x) = (1 - e^{-x/\theta})I_{(0,\infty)}(x)$. Since $\bar{X}$ is sufficient and complete for $\theta > 0$ and $U(X) = 1 - F_n(t)$ is unbiased for ϑ, $T(X) = E[U(X)|\bar{X}] = E[1 - F_n(t)|\bar{X}]$ is the UMVUE of ϑ. Since X_i's are i.i.d.,

$$E[1 - F_n(t)|\bar{X}] = \frac{1}{n}\sum_{i=1}^{n}P(X_i > t|\bar{X}) = P(X_1 > t|\bar{X}).$$

If the conditional distribution of X_1 given $\bar{X}$ is available, then we can calculate $P(X_1 > t|\bar{X})$ directly. But the following technique can be applied to avoid the derivation of conditional distributions. By Proposition 1.12(vii),

$$P(X_1 > t|\bar{X} = \bar{x}) = P(X_1/\bar{X} > t/\bar{X}|\bar{X} = \bar{x}) = P(X_1/\bar{X} > t/\bar{x}|\bar{X} = \bar{x}).$$

By Basu's theorem (Theorem 2.4), $X_1/\bar{X}$ and $\bar{X}$ are independent. Hence

$$P(X_1 > t|\bar{X} = \bar{x}) = P(X_1/\bar{X} > t/\bar{x}).$$

To compute this probability, we need the distribution of

$$X_1 \Big/ \sum_{i=1}^n X_i = X_1 \Big/ \left(X_1 + \sum_{i=2}^n X_i \right).$$

Using the transformation technique discussed in §1.3.1 and the fact that $\sum_{i=2}^n X_i$ is independent of X_1 and has a gamma distribution, we obtain that $X_1/\sum_{i=1}^n X_i$ has the Lebesgue p.d.f. $(n-1)(1-x)^{n-2}I_{(0,1)}(x)$. Hence

$$P(X_1 > t|\bar{X} = \bar{x}) = (n-1)\int_{t/(n\bar{x})}^1 (1-x)^{n-2}dx = \left(1 - \frac{t}{n\bar{x}}\right)^{n-1}$$

and the UMVUE of ϑ is $T(X) = E[1 - F_n(t)|\bar{X}] = \left(1 - \frac{t}{n\bar{X}}\right)^{n-1}$. ∎

We now show more examples of applying these two methods to find UMVUE's.

Example 3.4. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with unknown $\mu \in \mathcal{R}$ and $\sigma^2 > 0$. From Example 2.18, $T = (\bar{X}, S^2)$ is sufficient and complete for $\theta = (\mu, \sigma^2)$ and $\bar{X}$ and $(n-1)S^2/\sigma^2$ are independent and have the $N(\mu, \sigma^2/n)$ and chi-square distribution χ^2_{n-1}, respectively. Using the method of solving for h directly, we find that the UMVUE for μ is $\bar{X}$; the UMVUE of μ^2 is $\bar{X}^2 - S^2/n$; the UMVUE for σ^r with $r > 1-n$ is $k_{n-1,r}S^r$, where

$$k_{n,r} = \frac{n^{r/2}\Gamma(n/2)}{2^{r/2}\Gamma\left(\frac{n+r}{2}\right)}$$

(exercise); and the UMVUE of μ/σ is $k_{n-1,-1}\bar{X}/S$, if $n > 2$.

Suppose that ϑ satisfies $P(X_1 \le \vartheta) = p$ with a fixed $p \in (0,1)$. Let Φ be the c.d.f. of the standard normal distribution. Then $\vartheta = \mu + \sigma\Phi^{-1}(p)$ and its UMVUE is $\bar{X} + k_{n-1,1}S\Phi^{-1}(p)$.

Let c be a fixed constant and $\vartheta = P(X_1 \le c) = \Phi\left(\frac{c-\mu}{\sigma}\right)$. We can find the UMVUE of ϑ using the method of conditioning and the technique used in Example 3.3. Since $I_{(-\infty,c)}(X_1)$ is an unbiased estimator of ϑ, the UMVUE of ϑ is $E[I_{(-\infty,c)}(X_1)|T] = P(X_1 \le c|T)$. By Basu's theorem, the ancillary statistic $Z(X) = (X_1 - \bar{X})/S$ is independent of $T = (\bar{X}, S^2)$. Then

$$\begin{aligned} P\left(X_1 \le c|T = (\bar{x}, s^2)\right) &= P\left(\frac{X_1 - \bar{X}}{S} \le \frac{c - \bar{x}}{s}\middle| T = (\bar{x}, s^2)\right) \\ &= P\left(Z \le \frac{c - \bar{x}}{s}\right). \end{aligned}$$

It can be shown that Z has the Lebesgue p.d.f.

$$f(z) = \frac{\sqrt{n}\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}(n-1)\Gamma\left(\frac{n-2}{2}\right)}\left[1 - \frac{nz^2}{(n-1)^2}\right]^{(n/2)-2} I_{(0,(n-1)/\sqrt{n})}(|z|) \quad (3.1)$$

(exercise). Hence the UMVUE of ϑ is

$$P(X_1 \leq c|T) = \int_{-(n-1)/\sqrt{n}}^{(c-\bar{X})/S} f(z)dz \quad (3.2)$$

with f given by (3.1).

Suppose that we would like to estimate $\vartheta = \frac{1}{\sigma}\Phi'\left(\frac{c-\mu}{\sigma}\right)$, the Lebesgue p.d.f. of X_1 evaluated at a fixed c, where Φ' is the first-order derivative of Φ. By (3.2), the conditional p.d.f. of X_1 given $\bar{X} = \bar{x}$ and $S^2 = s^2$ is $s^{-1}f\left(\frac{x-\bar{x}}{s}\right)$. Let f_T be the joint p.d.f. of $T = (\bar{X}, S^2)$. Then

$$\vartheta = \int\int \frac{1}{s}f\left(\frac{c-\bar{x}}{s}\right) f_T(t)dt = E\left[\frac{1}{S}f\left(\frac{c-\bar{X}}{S}\right)\right].$$

Hence the UMVUE of ϑ is

$$\frac{1}{S}f\left(\frac{c-\bar{X}}{S}\right). \quad \blacksquare$$

Example 3.5. Let $X_1, ..., X_n$ be i.i.d. from a power series distribution (see Exercise 13 in §2.6), i.e.,

$$P(X_i = x) = \gamma(x)\theta^x/c(\theta), \qquad x = 0, 1, 2, ...$$

with a known function $\gamma(x) \geq 0$ and an unknown parameter $\theta > 0$. It turns out that the joint distribution of $X = (X_1, ..., X_n)$ is in an exponential family with a sufficient and complete statistic $T(X) = \sum_{i=1}^n X_i$. Furthermore, the distribution of T is also in a power series family, i.e.,

$$P(T = t) = \gamma_n(t)\theta^t/[c(\theta)]^n, \qquad t = 0, 1, 2, ...,$$

where $\gamma_n(t)$ is the coefficient of θ^t in the power series expansion of $[c(\theta)]^n$ (exercise). This result can help us to find the UMVUE of $\vartheta = g(\theta)$. For example, by comparing both sides of

$$\sum_{t=0}^{\infty} h(t)\gamma_n(t)\theta^t = [c(\theta)]^{n-p}\theta^r,$$

we conclude that the UMVUE of $\theta^r/[c(\theta)]^p$ is

$$h(T) = \begin{cases} 0 & T < r \\ \frac{\gamma_{n-p}(T-r)}{\gamma_n(T)} & T \geq r, \end{cases}$$

where r and p are nonnegative integers. In particular, the case of $p = 1$ produces the UMVUE $\gamma(r)h(T)$ of the probability $P(X_1 = r) = \gamma(r)\theta^r/c(\theta)$ for any nonnegative integer r. ∎

Example 3.6. Let $X_1, ..., X_n$ be i.i.d. from an unknown population P in a nonparametric family $\mathcal{P}$. We have discussed in §2.2 that in many cases the vector of order statistics, $T = (X_{(1)}, ..., X_{(n)})$, is sufficient and complete for $P \in \mathcal{P}$. Note that an estimator $\varphi(X_1, ..., X_n)$ is a function of T if and only if the function φ is symmetric in its n arguments. Hence, if T is sufficient and complete, then a symmetric unbiased estimator of any estimable ϑ is the UMVUE. For example, $\bar{X}$ is the UMVUE of $\vartheta = EX_1$; S^2 is the UMVUE of $\mathrm{Var}(X_1)$; $n^{-1}\sum_{i=1}^n X_i^2 - S^2$ is the UMVUE of $(EX_1)^2$; and $F_n(t)$ is the UMVUE of $P(X_1 \le t)$ for any fixed t.

Note that these conclusions are not true if T is *not* sufficient and complete for $P \in \mathcal{P}$. For example, if $\mathcal{P}$ contains all symmetric distributions having Lebesgue p.d.f.'s and finite means, then there is no UMVUE for $\vartheta = EX_1$ (exercise). ∎

More discussions of UMVUE's in nonparametric problems are provided in §3.2.

3.1.2 A necessary and sufficient condition

When a complete and sufficient statistic is not available, it is usually very difficult to derive a UMVUE. In some cases, the following result can be applied, if we have enough knowledge about unbiased estimators of 0.

Theorem 3.2. Let $\mathcal{U}$ be the set of all unbiased estimators of 0 with finite variances and T be an unbiased estimator of ϑ with $E(T^2) < \infty$.
(i) A necessary and sufficient condition for $T(X)$ to be a UMVUE of ϑ is that $E[T(X)U(X)] = 0$ for any $U \in \mathcal{U}$ and any $P \in \mathcal{P}$.
(ii) Suppose that $T = h(\tilde{T})$, where $\tilde{T}$ is a sufficient statistic for $P \in \mathcal{P}$ and h is a Borel function. Let $\mathcal{U}_{\tilde{T}}$ be the subset of $\mathcal{U}$ containing Borel functions of $\tilde{T}$. Then a necessary and sufficient condition for T to be a UMVUE of ϑ is that $E[T(X)U(X)] = 0$ for any $U \in \mathcal{U}_{\tilde{T}}$ and any $P \in \mathcal{P}$.
Proof. (i) Suppose that T is a UMVUE of ϑ. Then $T_c = T + cU$, where $U \in \mathcal{U}$ and c is a fixed constant, is also unbiased for ϑ and, thus,

$$\mathrm{Var}(T_c) \ge \mathrm{Var}(T) \qquad c \in \mathcal{R},\ P \in \mathcal{P},$$

which is the same as

$$c^2\mathrm{Var}(U) + 2c\mathrm{Cov}(T, U) \ge 0 \qquad c \in \mathcal{R},\ P \in \mathcal{P}.$$

This is impossible unless $\mathrm{Cov}(T, U) = E(TU) = 0$ for any $P \in \mathcal{P}$.

Suppose now $E(TU) = 0$ for any $U \in \mathcal{U}$ and $P \in \mathcal{P}$. Let T_0 be another unbiased estimator of ϑ with $\mathrm{Var}(T_0) < \infty$. Then $T - T_0 \in \mathcal{U}$ and, hence,

$$E[T(T - T_0)] = 0 \qquad P \in \mathcal{P},$$

which with the fact that $ET = ET_0$ implies that

$$\mathrm{Var}(T) = \mathrm{Cov}(T, T_0) \qquad P \in \mathcal{P}.$$

By inequality (1.34), $[\mathrm{Cov}(T, T_0)]^2 \leq \mathrm{Var}(T)\mathrm{Var}(T_0)$. Hence $\mathrm{Var}(T) \leq \mathrm{Var}(T_0)$ for any $P \in \mathcal{P}$.
(ii) It suffices to show that $E(TU) = 0$ for any $U \in \mathcal{U}_{\tilde{T}}$ and $P \in \mathcal{P}$ implies that $E(TU) = 0$ for any $U \in \mathcal{U}$ and $P \in \mathcal{P}$. Let $U \in \mathcal{U}$. Then $E(U|\tilde{T}) \in \mathcal{U}_{\tilde{T}}$ and the result follows from the fact that $T = h(\tilde{T})$ and

$$E(TU) = E[E(TU|\tilde{T})] = E[E(h(\tilde{T})U|\tilde{T})] = E[h(\tilde{T})E(U|\tilde{T})]. \quad \blacksquare$$

Theorem 3.2 can be used to find a UMVUE, to check whether a particular estimator is a UMVUE, and to show the nonexistence of any UMVUE. If there is a sufficient statistic, then by Rao-Blackwell's theorem, we only need to focus on functions of the sufficient statistic and, hence, Theorem 3.2(ii) is more convenient to use.

Example 3.7. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on the interval $(0, \theta)$. In Example 3.1, $(1 + n^{-1})X_{(n)}$ is shown to be the UMVUE for θ when the parameter space is $\Theta = (0, \infty)$. Suppose now that $\Theta = [1, \infty)$. Then $X_{(n)}$ is not complete, although it is still sufficient for θ. Thus, Theorem 3.1 does not apply. We now illustrate how to use Theorem 3.2(ii) to find a UMVUE of θ. Let $U(X_{(n)})$ be an unbiased estimator of 0. Since $X_{(n)}$ has the Lebesgue p.d.f. $n\theta^{-n}x^{n-1}I_{(0,\theta)}(x)$,

$$0 = \int_0^1 U(x)x^{n-1}dx + \int_1^\theta U(x)x^{n-1}dx$$

for all $\theta \geq 1$. This implies that $U(x) = 0$ a.e. Lebesgue measure on $[1, \infty)$ and

$$\int_0^1 U(x)x^{n-1}dx = 0.$$

Consider $T = h(X_{(n)})$. To have $E(TU) = 0$, we must have

$$\int_0^1 h(x)U(x)x^{n-1}dx = 0.$$

Thus, we may consider the following function:

$$h(x) = \begin{cases} c & 0 \leq x \leq 1 \\ bx & x > 1, \end{cases}$$

where c and b are some constants. From the previous discussion,

$$E[h(X_{(n)})U(X_{(n)})] = 0, \qquad \theta \geq 1.$$

Since $E[h(X_{(n)})] = \theta$, we obtain that

$$\begin{aligned}\theta &= cP(X_{(n)} \leq 1) + bE[X_{(n)}I_{(1,\infty)}(X_{(n)})] \\ &= c\theta^{-n} + [bn/(n+1)](\theta - \theta^{-n}).\end{aligned}$$

Thus, $c = 1$ and $b = (n+1)/n$. The UMVUE of θ is then

$$T = \begin{cases} 1 & 0 \leq X_{(n)} \leq 1 \\ (1+n^{-1})X_{(n)} & X_{(n)} > 1. \end{cases}$$

This estimator is better than $(1+n^{-1})X_{(n)}$ which is the UMVUE when $\Theta = (0,\infty)$ and does not make use of the information about $\theta \geq 1$. ∎

Example 3.8. Let X be a sample (of size 1) from the uniform distribution $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, $\theta \in \mathcal{R}$. We now apply Theorem 3.2 to show that there is no UMVUE of $\vartheta = g(\theta)$ for any nonconstant continuous g. Note that an unbiased estimator $U(X)$ of 0 must satisfy

$$\int_{\theta-\frac{1}{2}}^{\theta+\frac{1}{2}} U(x)dx = 0 \qquad \theta \in \mathcal{R}$$

and, hence, $U(x) = U(x+1)$ a.e. m, where m is the Lebesgue measure on $\mathcal{R}$. If T is a UMVUE, then $T(X)U(X)$ is unbiased for 0 and, hence, $T(x)U(x) = T(x+1)U(x+1)$ a.e. m, which implies that $T(x) = T(x+1)$ a.e. m. If T is unbiased for $g(\theta)$, then

$$g(\theta) = \int_{\theta-\frac{1}{2}}^{\theta+\frac{1}{2}} T(x)dx \qquad \theta \in \mathcal{R},$$

which implies that

$$g'(\theta) = T\left(\theta + \tfrac{1}{2}\right) - T\left(\theta - \tfrac{1}{2}\right) = 0 \quad \text{a.e. } m \quad ∎$$

As a consequence of Theorem 3.2, we have the following useful result.

Corollary 3.1. (i) Let T_j be a UMVUE of ϑ_j, $j = 1, ..., k$, where k is a fixed positive integer. Then $\sum_{j=1}^k c_j T_j$ is a UMVUE of $\vartheta = \sum_{j=1}^k c_j \vartheta_j$ for any constants $c_1, ..., c_k$.
(ii) Let T_1 and T_2 be two UMVUE's of ϑ. Then $T_1 = T_2$ a.s. P for any $P \in \mathcal{P}$. ∎

3.1.3 Information inequality

Suppose that we have a lower bound for the variances of all unbiased estimators of ϑ and that there is an unbiased estimator T of ϑ whose variance is always the same as the lower bound. Then T is a UMVUE of ϑ. Although this is not an effective way to find UMVUE's (compared with the methods introduced in §3.1.1 and §3.1.2), it provides a way of assessing the performance of UMVUE's. The following result provides such a lower bound in some cases.

Theorem 3.3 (The Cramér-Rao lower bound). Let $X = (X_1, ..., X_n)$ be a sample from $P \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where Θ is an open set in $\mathcal{R}^k$. Suppose that $T(X)$ is an estimator with $E[T(X)] = g(\theta)$ being a differentiable function of θ; the joint distribution of X has a p.d.f. f_θ w.r.t. a measure ν for all $\theta \in \Theta$; and f_θ is differentiable as a function of θ and satisfies

$$\frac{\partial}{\partial\theta}\int h(x)f_\theta(x)d\nu = \int h(x)\frac{\partial}{\partial\theta}f_\theta(x)d\nu, \qquad \theta \in \Theta, \tag{3.3}$$

for $h(x) \equiv 1$ and $h(x) = T(x)$. Then

$$\mathrm{Var}(T(X)) \geq \tfrac{\partial}{\partial\theta}g(\theta)[I(\theta)]^{-1}\left[\tfrac{\partial}{\partial\theta}g(\theta)\right]^\tau, \tag{3.4}$$

where

$$I(\theta) = E\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]^\tau\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right] \tag{3.5}$$

is assumed to be positive definite for any $\theta \in \Theta$.

Proof. We prove the univariate case ($k = 1$) only. The proof for the multivariate case ($k > 1$) is left to the reader. When $k = 1$, (3.4) reduces to

$$\mathrm{Var}(T(X)) \geq \frac{[g'(\theta)]^2}{E\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]^2}. \tag{3.6}$$

From inequality (1.34), we only need to show that

$$E\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]^2 = \mathrm{Var}\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)$$

and

$$g'(\theta) = \mathrm{Cov}\left(T(X), \frac{\partial}{\partial\theta}\log f_\theta(X)\right).$$

These two results are consequences of condition (3.3). ∎

The $k \times k$ matrix $I(\theta)$ in (3.5) is called the *Fisher information matrix*. The greater $I(\theta)$ is, the easier it is to distinguish θ from neighboring values

and, therefore, the more accurately θ can be estimated. In fact, if the equality in (3.6) holds for an unbiased estimator $T(X)$ of $g(\theta)$ (which is then a UMVUE), then the greater $I(\theta)$ is, the smaller $\mathrm{Var}(T(X))$ is. Thus, $I(\theta)$ is the information that X contains about the unknown parameter θ. The inequalities in (3.4) and (3.6) are called *information inequalities*.

The following result is helpful in finding the Fisher information matrix.

Proposition 3.1. (i) Let X and Y be independent with the Fisher information matrices $I_X(\theta)$ and $I_Y(\theta)$, respectively. Then the Fisher information about θ contained in (X, Y) is $I_X(\theta) + I_Y(\theta)$. In particular, if $X_1, ..., X_n$ are i.i.d. and $I(\theta)$ is the Fisher information about θ contained in a single X_i, then the Fisher information about θ contained in $X_1, ..., X_n$ is $nI(\theta)$.
(ii) Suppose that X has the p.d.f. f_θ which is twice differentiable in θ and that (3.3) holds with $h(x) \equiv 1$ and f_θ replaced by $\partial f_\theta / \partial\theta$. Then

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^\tau} \log f_\theta(X)\right]. \tag{3.7}$$

Proof. Result (i) follows from the independence of X and Y and the definition of the Fisher information. Result (ii) follows from the equality

$$\frac{\partial^2}{\partial\theta\partial\theta^\tau} \log f_\theta(X) = \frac{\frac{\partial^2}{\partial\theta\partial\theta^\tau} f_\theta(X)}{f_\theta(X)} - \left[\frac{\partial}{\partial\theta} \log f_\theta(X)\right]^\tau \left[\frac{\partial}{\partial\theta} \log f_\theta(X)\right]. \quad \blacksquare$$

The following example provides a formula for the Fisher information matrix for many parametric families with a two-dimensional parameter θ.

Example 3.9. Let $X_1, ..., X_n$ be i.i.d. with the Lebesgue p.d.f. $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$, where $f(x) > 0$ and $f'(x)$ exists for all $x \in \mathcal{R}$, $\mu \in \mathcal{R}$, and $\sigma > 0$ (a location-scale family). Let $\theta = (\mu, \sigma)$. Then the Fisher information about θ contained in $X_1, ..., X_n$ is (exercise)

$$I(\theta) = \frac{n}{\sigma^2} \begin{pmatrix} \int \frac{[f'(x)]^2}{f(x)} dx & \int x \frac{[f'(x)]^2}{f(x)} dx \\ \int x \frac{[f'(x)]^2}{f(x)} dx & \int \frac{[xf'(x)+f(x)]^2}{f(x)} dx \end{pmatrix}. \quad \blacksquare$$

Note that $I(\theta)$ depends on the particular parameterization. If $\theta = \psi(\eta)$ and ψ is differentiable, then the Fisher information that X contains about η is

$$\tfrac{\partial}{\partial\eta}\psi(\eta) I(\psi(\eta)) \left[\tfrac{\partial}{\partial\eta}\psi(\eta)\right]^\tau.$$

However, it is easy to see that the Cramér-Rao lower bound in (3.4) or (3.6) is not affected by any one-to-one reparameterization.

If we use inequality (3.4) or (3.6) to find a UMVUE $T(X)$, then we obtain a formula for $\mathrm{Var}(T(X))$ at the same time. On the other hand, the Cramér-Rao lower bound in (3.4) or (3.6) is typically not sharp. Under some regularity conditions, the Cramér-Rao lower bound is attained if and only if f_θ is in an exponential family; see Propositions 3.2 and 3.3 and the discussion in Lehmann (1983, p. 123). Some improved information inequalities are available (see, e.g., Lehmann (1983, Sections 2.6 and 2.7)).

Proposition 3.2. Suppose that the distribution of X is from an exponential family $\{f_\theta : \theta \in \Theta\}$, i.e., the p.d.f. of X w.r.t. a measure ν is

$$f_\theta(x) = \exp\big\{T(x)[\eta(\theta)]^\tau - \xi(\theta)\big\}c(x) \tag{3.8}$$

(see §2.1.3), where Θ is an open subset of $\mathcal{R}^k$.
(i) The regularity condition (3.3) is satisfied for any h with $E|h(X)| < \infty$ and (3.7) holds.
(ii) If $\underline{I}(\eta)$ is the Fisher information matrix for the natural parameter η, then the variance-covariance matrix $\mathrm{Var}(T) = \underline{I}(\eta)$.
(iii) If $\overline{I}(\vartheta)$ is the Fisher information matrix for the parameter $\vartheta = E[T(X)]$, then $\mathrm{Var}(T) = [\overline{I}(\vartheta)]^{-1}$.
Proof. (i) This is a direct consequence of Theorem 2.1.
(ii) From (2.6), the p.d.f. under the natural parameter η is

$$\tilde{f}_\eta(x) = \exp\left\{T(x)\eta^\tau - \zeta(\eta)\right\}c(x).$$

From Proposition 1.10 and Theorem 2.1, $E[T(X)] = \frac{\partial}{\partial\eta}\zeta(\eta)$. The result follows from

$$\tfrac{\partial}{\partial\eta}\log\tilde{f}_\eta(x) = T(x) - \tfrac{\partial}{\partial\eta}\zeta(\eta).$$

(iii) Since $\vartheta = E[T(X)] = \frac{\partial}{\partial\eta}\zeta(\eta)$,

$$\underline{I}(\eta) = \tfrac{\partial\vartheta}{\partial\eta}\overline{I}(\vartheta)\tfrac{\partial\vartheta^\tau}{\partial\eta} = \tfrac{\partial^2}{\partial\eta\partial\eta^\tau}\zeta(\eta)\overline{I}(\vartheta)\left[\tfrac{\partial^2}{\partial\eta\partial\eta^\tau}\zeta(\eta)\right]^\tau.$$

By Proposition 1.10, Theorem 2.1, and the result in (ii), $\frac{\partial^2}{\partial\eta\partial\eta^\tau}\zeta(\eta) = \mathrm{Var}(T) = \underline{I}(\eta)$. Hence

$$\overline{I}(\vartheta) = [\underline{I}(\eta)]^{-1}\underline{I}(\eta)[\underline{I}(\eta)]^{-1} = [\underline{I}(\eta)]^{-1} = [\mathrm{Var}(T)]^{-1}. \quad \blacksquare$$

A direct consequence of Proposition 3.2(ii) is that the variance of any linear function of T in (3.8) attains the Cramér-Rao lower bound. The following result gives a necessary condition for $\mathrm{Var}(U(X))$ of an estimator $U(X)$ to attain the Cramér-Rao lower bound.

Proposition 3.3. Let $U(X)$ be an estimator of $g(\theta) = E[U(X)]$. Assume that the conditions in Theorem 3.3 hold for $U(x)$ and that $\Theta \subset \mathcal{R}$.
(i) If $\mathrm{Var}(U(X))$ attains the Cramér-Rao lower bound in (3.6), then

$$a(\theta)[U(X) - g(\theta)] = g'(\theta)\frac{\partial}{\partial\theta}\log f_\theta(X) \quad \text{a.s. } f_\theta$$

for some function $a(\theta)$, $\theta \in \Theta$.
(ii) Let f_θ and T be given by (3.8). If $\mathrm{Var}(U(X))$ attains the Cramér-Rao lower bound, then $U(X)$ is a linear function of $T(X)$ a.s. f_θ, $\theta \in \Theta$. ∎

Example 3.10. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathcal{R}$ and a known σ^2. Let f_μ be the joint distribution of $X = (X_1, ..., X_n)$. Then

$$\tfrac{\partial}{\partial\mu}\log f_\mu(X) = \sum_{i=1}^n (X_i - \mu)/\sigma^2.$$

Thus, $I(\mu) = n/\sigma^2$. It is obvious that $\mathrm{Var}(\bar{X})$ attains the Cramér-Rao lower bound in (3.6). Consider now the estimation of $\vartheta = \mu^2$. Since $E\bar{X}^2 = \mu^2 + \sigma^2/n$, the UMVUE of ϑ is $h(\bar{X}) = \bar{X}^2 - \sigma^2/n$. A straightforward calculation shows that

$$\mathrm{Var}(h(\bar{X})) = \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n^2}.$$

On the other hand, the Cramér-Rao lower bound in this case is $4\mu^2\sigma^2/n$. Hence $\mathrm{Var}(h(\bar{X}))$ does not attain the Cramér-Rao lower bound. The difference is $2\sigma^4/n^2$. ∎

Condition (3.3) is a key regularity condition for the results in Theorem 3.3 and Proposition 3.3. If f_θ is not in an exponential family, then (3.3) has to be checked. Typically, it does not hold if the set $\{x : f_\theta(x) > 0\}$ depends on θ (Exercise 32). More discussions can be found in Pitman (1979).

3.1.4 Asymptotic properties of UMVUE's

UMVUE's are typically consistent (see Exercise 88 in §2.6). If there is an unbiased estimator of ϑ whose mse is of the order a_n^{-2}, where $\{a_n\}$ is a sequence of positive numbers diverging to ∞, then the UMVUE of ϑ (if it exists) has a mse of order a_n^{-2} and is a_n-consistent. For instance, in Example 3.3, the mse of $U(X) = 1 - F_n(t)$ is $F_\theta(t)[1 - F_\theta(t)]/n$; hence the UMVUE $T(X)$ is $\sqrt{n}$-consistent and its mse is of the order n^{-1}.

UMVUE's are exactly unbiased so that there is no need to discuss their asymptotic biases. Their variances (or mse's) are finite, but amse's can

be used to approximate their mse's if the exact forms of these mse's are difficult to obtain. In many cases, although the variance of a UMVUE T_n does not attain the Cramér-Rao lower bound, the limit of the ratio of the amse (or mse) of T_n over the Cramér-Rao lower bound (if it is not 0) is 1. For instance, in Example 3.10,

$$\frac{\mathrm{Var}(\bar{X}^2 - \sigma^2/n)}{\text{the Cramér-Rao lower bound}} = 1 + \frac{\sigma^2}{2\mu^2 n} \to 1$$

if $\mu \neq 0$. In general, under the conditions in Theorem 3.3, if $T_n(X)$ is unbiased for $g(\theta)$ and if for any $\theta \in \Theta$,

$$T_n(X) - g(\theta) = \tfrac{\partial}{\partial\theta} g(\theta)[I(\theta)]^{-1} \left[\tfrac{\partial}{\partial\theta} \log f_\theta(X)\right]^\tau [1 + o_p(1)] \quad \text{a.s. } f_\theta, \tag{3.9}$$

then

$$\mathrm{amse}_{T_n}(\theta) = \text{the Cramér-Rao lower bound} \tag{3.10}$$

whenever the Cramér-Rao lower bound is not 0. Note that the case of zero Cramér-Rao lower bound is not of interest since a zero lower bound does not provide any information on the performance of estimators.

Consider the UMVUE $T_n = \left(1 - \frac{t}{n\bar{X}}\right)^{n-1}$ of $e^{-t/\theta}$ in Example 3.3. Using the fact that

$$\log(1-x) = -\sum_{j=1}^{\infty} \frac{x^j}{j}, \qquad |x| \leq 1,$$

we obtain that

$$T_n - e^{-t/\bar{X}} = O_p\left(n^{-1}\right).$$

Using Taylor's expansion, we obtain that

$$e^{-t/\bar{X}} - e^{-t/\theta} = g'(\theta)(\bar{X} - \theta)[1 + o_p(1)],$$

where $g(\theta) = e^{-t/\theta}$. On the other hand,

$$[I(\theta)]^{-1} \tfrac{\partial}{\partial\theta} \log f_\theta(X) = \bar{X} - \theta.$$

Hence (3.9) and (3.10) hold. Note that the exact variance of T_n is not easy to obtain. In this example, it can be shown that $\{n[T_n - g(\theta)]^2\}$ is uniformly integrable and, therefore,

$$\begin{aligned}
\lim_{n\to\infty} n\mathrm{Var}(T_n) &= \lim_{n\to\infty} n[\underline{\mathrm{amse}}_{T_n}(\theta)] \\
&= \lim_{n\to\infty} n[g'(\theta)]^2[I(\theta)]^{-1} \\
&= \frac{t^2 e^{-2t/\theta}}{\theta^2}.
\end{aligned}$$

It is shown in Chapter 4 that if (3.10) holds, then T_n is asymptotically optimal in some sense. Hence UMVUE's satisfying (3.9), which is often true, are asymptotically optimal, although they may be improved in terms of the exact mse's.

3.2 U-Statistics

Let $X_1, ..., X_n$ be i.i.d. from an unknown population P in a nonparametric family $\mathcal{P}$. In Example 3.6 we argued that if the vector of order statistic is sufficient and complete for $P \in \mathcal{P}$, then a symmetric unbiased estimator of any estimable ϑ is the UMVUE of ϑ. In a large class of problems parameters to be estimated are of the form

$$\vartheta = E[h(X_1, ..., X_m)]$$

with a positive integer m and a Borel function h which is symmetric and satisfies $E|h(X_1, ..., X_m)| < \infty$ for any $P \in \mathcal{P}$. It is easy to see that a symmetric unbiased estimator of ϑ is

$$U_n = \binom{n}{m}^{-1} \sum_c h(X_{i_1}, ..., X_{i_m}), \tag{3.11}$$

where $\sum_c$ denotes the summation over the $\binom{n}{m}$ combinations of m distinct elements $\{i_1, ..., i_m\}$ from $\{1, ..., n\}$.

Definition 3.2. The statistic U_n in (3.11) is called a *U-statistic* with kernel h of order m. ∎

3.2.1 Some examples

The use of U-statistics is an effective way of obtaining unbiased estimators. In nonparametric problems, U-statistics are often UMVUE's, whereas in parametric problems, U-statistics can be used as initial estimators to derive more efficient estimators.

If $m = 1$, U_n in (3.11) is simply a type of sample mean. Examples include the empirical c.d.f. (2.31) evaluated at a particular t and the *sample moments* $n^{-1}\sum_{i=1}^n X_i^k$ for a positive integer k. We now consider some examples with $m > 1$.

Consider the estimation of $\vartheta = \mu^m$, where $\mu = EX_1$ and m is a positive integer. Using $h(x_1, ..., x_m) = x_1 \cdots x_m$, we obtain the following U-statistic unbiased for $\vartheta = \mu^m$:

$$U_n = \binom{n}{m}^{-1} \sum_c X_{i_1} \cdots X_{i_m}. \tag{3.12}$$

Consider next the estimation of $\vartheta = \sigma^2 = \mathrm{Var}(X_1)$. Since

$$\sigma^2 = [\mathrm{Var}(X_1) + \mathrm{Var}(X_2)]/2 = E[(X_1 - X_2)^2/2],$$

we obtain the following U-statistic with kernel $h(x_1, x_2) = (x_1 - x_2)^2/2$:

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \frac{(X_i - X_j)^2}{2} = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right) = S^2,$$

which is the sample variance in (2.2).

In some cases we would like to estimate $\vartheta = E|X_1 - X_2|$, a measure of concentration. Using kernel $h(x_1, x_2) = |x_1 - x_2|$, we obtain the following U-statistic unbiased for $\vartheta = E|X_1 - X_2|$:

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} |X_i - X_j|,$$

which is known as *Gini's mean difference*.

Let $\vartheta = P(X_1 + X_2 \le 0)$. Using kernel $h(x_1, x_2) = I_{(-\infty,0)}(x_1 + x_2)$, we obtain the following U-statistic unbiased for ϑ:

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} I_{(-\infty,0)}(X_i + X_j),$$

which is known as the *one-sample Wilcoxon statistic*.

Let $T_n = T_n(X_1, ..., X_n)$ be a given statistic and let r and d be two positive integers such that $r + d = n$. For any $\boldsymbol{s} = \{i_1, ..., i_r\} \subset \{1, ..., n\}$, define

$$T_{r,\boldsymbol{s}} = T_r(X_{i_1}, ..., X_{i_r}),$$

which is the statistic T_n computed after X_i, $i \notin \boldsymbol{s}$, are deleted from the original sample. Let

$$U_n = \binom{n}{r}^{-1} \sum_c \tfrac{r}{d}(T_{r,\boldsymbol{s}} - T_n)^2. \tag{3.13}$$

Then U_n is a U-statistic with kernel

$$h_n(x_1, ..., x_r) = \tfrac{r}{d}[T_r(x_1, ..., x_r) - T_n(x_1, ..., x_n)].$$

Unlike the kernels in the previous examples, the kernel in this example depends on n. The order of the kernel, r, may also depend on n. The statistic U_n in (3.13) is known as the *delete-d jackknife variance estimator* for T_n (see, e.g., Shao and Tu (1995)), since it is often true that

$$E[h_n(X_1, ..., X_r)] \approx \mathrm{Var}(T_n).$$

It can be shown that if $T_n = \bar{X}$, then nU_n in (3.13) is exactly the same as the sample variance S^2 (exercise).

3.2.2 Variances of U-statistics

If $E[h(X_1, ..., X_m)]^2 < \infty$, then the variance of U_n in (3.11) with kernel h has an explicit form. To derive $\mathrm{Var}(U_n)$, we need some notation. For $k = 1, ..., m$, let

$$\begin{aligned} h_k(x_1, ..., x_k) &= E[h(X_1, ..., X_m)|X_1 = x_1, ..., X_k = x_k] \\ &= E[h(x_1, ..., x_k, X_{k+1}, ..., X_m)]. \end{aligned}$$

It can be shown that

$$h_k(x_1, ..., x_k) = E[h_{k+1}(x_1, ..., x_k, X_{k+1})]. \tag{3.14}$$

Define

$$\tilde{h}_k = h_k - E[h(X_1, ..., X_m)]. \tag{3.15}$$

Then, for any U_n defined by (3.11),

$$U_n - E(U_n) = \binom{n}{m}^{-1} \sum_c \tilde{h}_m(X_{i_1}, ..., X_{i_m}). \tag{3.16}$$

Theorem 3.4 (Hoeffding's theorem). For a U-statistic U_n given by (3.11) with $E[h(X_1, ..., X_m)]^2 < \infty$,

$$\mathrm{Var}(U_n) = \binom{n}{m}^{-1} \sum_{k=1}^{m} \binom{m}{k} \binom{n-m}{m-k} \zeta_k,$$

where

$$\zeta_k = \mathrm{Var}(h_k(X_1, ..., X_k)).$$

Proof. Consider two sets $\{i_1, ..., i_m\}$ and $\{j_1, ..., j_m\}$ of m distinct integers from $\{1, ..., n\}$ with exactly k integers in common. The number of distinct choices of two such sets is $\binom{n}{m}\binom{m}{k}\binom{n-m}{m-k}$. By the symmetry of $\tilde{h}_m$ and independence of $X_1, ..., X_n$,

$$E[\tilde{h}_m(X_{i_1}, ..., X_{i_m})\tilde{h}_m(X_{j_1}, ..., X_{j_m})] = \zeta_k \tag{3.17}$$

for $k = 1, ..., m$ (exercise). Then, by (3.16),

$$\begin{aligned} \mathrm{Var}(U_n) &= \binom{n}{m}^{-2} \sum_c \sum_c E[\tilde{h}_m(X_{i_1}, ..., X_{i_m})\tilde{h}_m(X_{j_1}, ..., X_{j_m})] \\ &= \binom{n}{m}^{-2} \sum_{k=1}^{m} \binom{n}{m} \binom{m}{k} \binom{n-m}{m-k} \zeta_k. \end{aligned}$$

This proves the result. ∎

Corollary 3.2. Under the condition of Theorem 3.4,
(i) $\frac{m^2}{n}\zeta_1 \le \mathrm{Var}(U_n) \le \frac{m}{n}\zeta_m$;
(ii) $(n+1)\mathrm{Var}(U_{n+1}) \le n\mathrm{Var}(U_n)$ for any $n > m$;
(iii) For any fixed m and $k = 1, ..., m$, if $\zeta_j = 0$ for $j < k$ and $\zeta_k > 0$, then

$$\mathrm{Var}(U_n) = \frac{k!\binom{m}{k}^2\zeta_k}{n^k} + O\left(\frac{1}{n^{k+1}}\right). \quad \blacksquare$$

It follows from Corollary 3.2 that a U-statistic U_n as an estimator of its mean is consistent in mse (under the finite second moment assumption on h). In fact, for any fixed m, if $\zeta_j = 0$ for $j < k$ and $\zeta_k > 0$, then the mse of U_n is of the order n^{-k} and, therefore, U_n is $n^{k/2}$-consistent.

Example 3.11. Consider first $h(x_1, x_2) = x_1x_2$ which leads to a U-statistic unbiased for μ^2, $\mu = EX_1$. Note that $h_1(x_1) = \mu x_1$, $\tilde{h}_1(x_1) = \mu(x_1 - \mu)$, $\zeta_1 = E[\tilde{h}_1(X_1)]^2 = \mu^2\mathrm{Var}(X_1) = \mu^2\sigma^2$, $\tilde{h}_2(x_1, x_2) = x_1x_2 - \mu^2$, and $\zeta_2 = \mathrm{Var}(X_1X_2) = E(X_1X_2)^2 - \mu^4 = (\mu^2 + \sigma^2)^2 - \mu^4$. By Theorem 3.4, for $U_n = \binom{n}{2}^{-1}\sum_{1\le i<j\le n} X_iX_j$,

$$\begin{aligned}\mathrm{Var}(U_n) &= \binom{n}{2}^{-1}\left[\binom{2}{1}\binom{n-2}{1}\zeta_1 + \binom{2}{2}\binom{n-2}{0}\zeta_2\right] \\ &= \frac{2}{n(n-1)}\left[2(n-2)\mu^2\sigma^2 + (\mu^2+\sigma^2)^2 - \mu^4\right] \\ &= \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n(n-1)}.\end{aligned}$$

Comparing U_n with $\bar{X}^2 - \sigma^2/n$ in Example 3.10, which is the UMVUE under the normality and known σ^2 assumption, we find that

$$\mathrm{Var}(U_n) - \mathrm{Var}(\bar{X}^2 - \sigma^2/n) = \frac{2\sigma^4}{n^2(n-1)}.$$

Next, consider $h(x_1, x_2) = I_{(-\infty,0)}(x_1 + x_2)$ which leads to the one-sample Wilcoxon statistic. Note that $h_1(x_1) = P(x_1 + X_2 \le 0) = F(-x_1)$, where F is the c.d.f. of P. Then $\zeta_1 = \mathrm{Var}(F(-X_1))$. Let $\vartheta = E[h(X_1, X_2)]$. Then $\zeta_2 = \mathrm{Var}(h(X_1, X_2)) = \vartheta(1-\vartheta)$. Hence, for U_n being the one-sample Wilcoxon statistic,

$$\mathrm{Var}(U_n) = \frac{2}{n(n-1)}\left[2(n-2)\zeta_1 + \vartheta(1-\vartheta)\right].$$

If F is continuous and symmetric about 0, then ζ_1 can be simplified as

$$\zeta_1 = \mathrm{Var}(F(-X_1)) = \mathrm{Var}(1 - F(X_1)) = \mathrm{Var}(F(X_1)) = \tfrac{1}{12},$$

since $F(X_1)$ has the uniform distribution on $[0,1]$.

Finally, consider $h(x_1,x_2)=|x_1-x_2|$, which leads to Gini's mean difference. Note that

$$h_1(x_1)=E|x_1-X_2|=\int|x_1-y|dP(y),$$

and

$$\zeta_1=\mathrm{Var}(h_1(X_1))=\int\left[\int|x-y|dP(y)\right]^2dP(x)-\vartheta^2,$$

where $\vartheta=E|X_1-X_2|$. ∎

3.2.3 The projection method

Since $\mathcal{P}$ is nonparametric, the exact distribution of any U-statistic is hard to derive. In this section we study asymptotic distributions of U-statistics, using the method of *projection*.

Definition 3.3. Let T_n be a given statistic based on $X_1,...,X_n$. The projection of T_n on k_n random elements $Y_1,...,Y_{k_n}$ is defined to be

$$\check{T}_n=E(T_n)+\sum_{i=1}^{k_n}[E(T_n|Y_i)-E(T_n)]. \quad ∎$$

Let $\psi_n(X_i)=E(T_n|X_i)$. If T_n is symmetric (as a function of $X_1,...,X_n$), then $\psi_n(X_1),...,\psi_n(X_n)$ are i.i.d. with mean $E[\psi_n(X_i)]=E[E(T_n|X_i)]=E(T_n)$. If $E(T_n^2)<\infty$ and $\mathrm{Var}(\psi_n(X_i))>0$, then

$$\frac{1}{\sqrt{n\mathrm{Var}(\psi_n(X_1))}}\sum_{i=1}^n[\psi_n(X_i)-E(T_n)]\to_d N(0,1) \tag{3.18}$$

by the CLT. Let $\check{T}_n$ be the projection of T_n on $X_1,...,X_n$. Then

$$T_n-\check{T}_n=T_n-E(T_n)-\sum_{i=1}^n[\psi_n(X_i)-E(T_n)]. \tag{3.19}$$

If we can show that $T_n-\check{T}_n$ has a negligible order of magnitude, then we can derive the asymptotic distribution of T_n by using (3.18)-(3.19) and Slutsky's theorem. The order of magnitude of $T_n-\check{T}_n$ can be obtained with the help of the following lemma.

Lemma 3.1. Let T_n be a symmetric statistic with $\mathrm{Var}(T_n)<\infty$ for every n and $\check{T}_n$ be the projection of T_n on $X_1,...,X_n$. Then $E(T_n)=E(\check{T}_n)$ and

$$E(T_n-\check{T}_n)^2=\mathrm{Var}(T_n)-\mathrm{Var}(\check{T}_n).$$

Proof. Since $E(T_n) = E(\check{T}_n)$,

$$E(T_n - \check{T}_n)^2 = \mathrm{Var}(T_n) + \mathrm{Var}(\check{T}_n) - 2\mathrm{Cov}(T_n, \check{T}_n).$$

From Definition 3.3 with $Y_i = X_i$,

$$\mathrm{Var}(\check{T}_n) = n\mathrm{Var}(E(T_n|X_i)).$$

The result follows from

$$\begin{aligned}
\mathrm{Cov}(T_n, \check{T}_n) &= E(T_n\check{T}_n) - [E(T_n)]^2 \\
&= nE[T_nE(T_n|X_i)] - n[E(T_n)]^2 \\
&= nE\{E[T_nE(T_n|X_i)|X_i]\} - n[E(T_n)]^2 \\
&= nE\{[E(T_n|X_i)]^2\} - n[E(T_n)]^2 \\
&= n\mathrm{Var}(E(T_n|X_i)) \\
&= \mathrm{Var}(\check{T}_n). \quad \blacksquare
\end{aligned}$$

This method of deriving the asymptotic distribution of T_n is known as the method of projection and is particularly effective for U-statistics. For a U-statistic U_n given by (3.11), one can show (exercise) that

$$\check{U}_n = E(U_n) + \frac{m}{n}\sum_{i=1}^{n}\tilde{h}_1(X_i), \tag{3.20}$$

where $\check{U}_n$ is the projection of U_n on $X_1, ..., X_n$ and $\tilde{h}_1$ is defined by (3.15). Hence

$$\mathrm{Var}(\check{U}_n) = m^2\zeta_1/n$$

and, by Corollary 3.2 and Lemma 3.1,

$$E(U_n - \check{U}_n)^2 = O(n^{-2}).$$

If $\zeta_1 > 0$, then (3.18) holds with $\psi_n(X_i) = mh_1(X_i)$, which leads to the result in Theorem 3.5(i) stated later.

If $\zeta_1 = 0$, then $\tilde{h}_1 \equiv 0$ and we have to use another projection of U_n. Suppose that $\zeta_1 = \cdots = \zeta_{k-1} = 0$ and $\zeta_k > 0$ for an integer $k > 1$. Consider the projection $\check{U}_{kn}$ of U_n on $\{X_{i_1}, ..., X_{i_k}\}$, $1 \le i_1 < \cdots < i_k \le n$. We can establish a result similar to that in Lemma 3.1 (exercise) and show that

$$E(U_n - \check{U}_n)^2 = O(n^{-(k+1)}).$$

Also, see Serfling (1980, §5.3.4).

With these results, we obtain the following theorem.

Theorem 3.5. Let U_n be given by (3.11) with $E[h(X_1, ..., X_m)]^2 < \infty$.
(i) If $\zeta_1 > 0$, then

$$\sqrt{n}[U_n - E(U_n)] \to_d N(0, m^2\zeta_1).$$

(ii) If $\zeta_1 = 0$ but $\zeta_2 > 0$, then

$$n[U_n - E(U_n)] \to_d \frac{m(m-1)}{2} \sum_{j=1}^{\infty} \lambda_j(\chi_{1j}^2 - 1), \tag{3.21}$$

where χ_{1j}^2's are i.i.d. random variables having the chi-square distribution χ_1^2 and λ_j's are some constants (which may depend on P) satisfying $\sum_{j=1}^{\infty} \lambda_j^2 = \zeta_2$. ∎

We have actually proved Theorem 3.5(i). A proof for Theorem 3.5(ii) is given in Serfling (1980, §5.5.2). One may derive results for the cases where $\zeta_2 = 0$, but the case of either $\zeta_1 > 0$ or $\zeta_2 > 0$ is the most interesting case in applications.

If $\zeta_1 > 0$, it follows from Theorem 3.5(i) and Corollary 3.2(iii) that $\underline{\text{amse}}_{U_n}(P) = \text{Var}(U_n) = m^2\zeta_1/n$. By Theorem 1.8(vii), $\{n[U_n - E(U_n)]^2\}$ is uniformly integrable.

If $\zeta_1 = 0$ but $\zeta_2 > 0$, it follows from Theorem 3.5(ii) that $\underline{\text{amse}}_{U_n}(P) = EY^2/n^2$, where Y denotes the random variable on the right-hand side of (3.21). The following result provides the value of EY^2.

Lemma 3.2. Let Y be the random variable on the right-hand side of (3.21). Then $EY^2 = \frac{m^2(m-1)^2}{2}\zeta_2$.
Proof. Define

$$Y_k = \frac{m(m-1)}{2} \sum_{j=1}^{k} \lambda_j(\chi_{1j}^2 - 1), \quad k = 1, 2, ...$$

It can be shown (exercise) that $\{Y_k^2\}$ is uniformly integrable. Since $Y_k \to_d Y$ as $k \to \infty$, $\lim_{k\to\infty} EY_k^2 = EY^2$ (Theorem 1.8(vii)). Since χ_{1j}^2's are independent chi-square random variables with $E\chi_{1j}^2 = 1$ and $\text{Var}(\chi_{1j}^2) = 2$, $EY_k = 0$ for any k and

$$\begin{aligned} EY_k^2 &= \frac{m^2(m-1)^2}{4} \sum_{j=1}^{k} \lambda_j^2 \text{Var}(\chi_{1j}^2) \\ &= \frac{m^2(m-1)^2}{4} \left(2 \sum_{j=1}^{k} \lambda_j^2 \right) \\ &\to \frac{m^2(m-1)^2}{2} \zeta_2. \quad ∎ \end{aligned}$$

It follows from Corollary 3.2(iii) and Lemma 3.2 that $\underline{\text{amse}}_{U_n}(P) = \text{Var}(U_n) = \frac{m^2(m-1)^2}{2}\zeta_2/n^2$ if $\zeta_1 = 0$. Again, by Theorem 1.8(vii), the sequence $\{n^2[U_n - E(U_n)]^2\}$ is uniformly integrable.

We now apply Theorem 3.5 to the U-statistics in Example 3.11. For $U_n = \frac{2}{n(n-1)}\sum_{1\le i<j\le n} X_iX_j$, $\zeta_1 = \mu^2\sigma^2$. Thus, if $\mu \neq 0$, the result in Theorem 3.5(i) holds with $\zeta_1 = \mu^2\sigma^2$. If $\mu = 0$, then $\zeta_1 = 0$, $\zeta_2 = \sigma^4 > 0$, and Theorem 3.5(ii) applies. However, it is not convenient to use Theorem 3.5(ii) to find the limiting distribution of U_n. We may derive this limiting distribution using the following technique which is further discussed in §3.5. By the CLT and Theorem 1.10,

$$n\bar{X}^2/\sigma^2 \to_d \chi_1^2$$

when $\mu = 0$, where χ_1^2 is a random variable having the chi-square distribution χ_1^2. Note that

$$\frac{n\bar{X}^2}{\sigma^2} = \frac{1}{\sigma^2 n}\sum_{i=1}^n X_i^2 + \frac{(n-1)U_n}{\sigma^2}.$$

By the SLLN, $\frac{1}{\sigma^2 n}\sum_{i=1}^n X_i^2 \to_{a.s.} 1$. An application of Slutsky's theorem leads to

$$nU_n/\sigma^2 \to_d \chi_1^2 - 1.$$

Since $\mu = 0$, this implies that the right-hand side of (3.21) is $\sigma^2(\chi_1^2 - 1)$, i.e., $\lambda_1 = \sigma^2$ and $\lambda_j = 0$ when $j > 1$.

For the one-sample Wilcoxon statistic, $\zeta_1 = \text{Var}(F(-X_1)) > 0$ unless F is degenerate. Similarly, for Gini's mean difference, $\zeta_1 > 0$ unless F is degenerate. Hence Theorem 3.5(i) applies to these two cases.

Theorem 3.5 does not apply to U_n defined by (3.13), if r, the order of the kernel, depends on n and diverges to ∞ as $n \to \infty$. We consider the simple case where

$$T_n = \frac{1}{n}\sum_{i=1}^n \psi(X_i) + R_n \tag{3.22}$$

for some R_n satisfying $E(R_n^2) = o(n^{-1})$. Note that (3.22) is satisfied for T_n being a U-statistic (exercise). Assume that r/d is bounded. Let $S_\psi^2 = (n-1)^{-1}\sum_{i=1}^n[\psi(X_i) - n^{-1}\sum_{i=1}^n \psi(X_i)]^2$. Then

$$nU_n = S_\psi^2 + o_p(1) \tag{3.23}$$

(exercise). Under (3.22), if $0 < E[\psi(X_i)]^2 < \infty$, then $\underline{\text{amse}}_{T_n}(P) = E[\psi(X_i)]^2/n$. Hence, the jackknife estimator U_n in (3.13) provides a consistent estimator of $\underline{\text{amse}}_{T_n}(P)$, i.e., $U_n/\underline{\text{amse}}_{T_n}(P) \to_p 1$.

3.3 The LSE in Linear Models

One of the most useful statistical models for non-i.i.d. data in applications is the following general linear model

$$X_i = Z_i\beta^\tau + \varepsilon_i, \qquad i = 1, ..., n, \tag{3.24}$$

where X_i is the ith observation and is often called the ith response; β is a p-vector of unknown parameters, $p < n$; Z_i is the ith value of a p-vector of explanatory variables (or covariates); and $\varepsilon_1, ..., \varepsilon_n$ are random errors. Our data in this case are $(X_1, Z_1), ..., (X_n, Z_n)$ (ε_i's are not observed). Throughout this book Z_i's are considered to be nonrandom or given values of a random p-vector, in which case our analysis is conditioned on $Z_1, ..., Z_n$. Each ε_i can be viewed as a random measurement error in measuring the unknown mean of X_i when the covariate vector is equal to Z_i. The main parameter of interest is β. More specific examples of model (3.24) are provided in this section. Other examples and examples of data from model (3.24) can be found in many standard books for linear models, for example, Draper and Smith (1981) and Searle (1971).

3.3.1 The LSE and estimability

Let $X = (X_1, ..., X_n)$, $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$, and Z be the $n \times p$ matrix whose ith row is Z_i, $i = 1, ..., n$. Then a matrix form of model (3.24) is

$$X = \beta Z^\tau + \varepsilon. \tag{3.25}$$

Definition 3.4. Suppose that the range of β in model (3.25) is $B \subset \mathcal{R}^p$. A *least squares estimator* (LSE) of β is defined to be any $\hat{\beta} \in B$ such that

$$\|X - \hat{\beta}Z^\tau\|^2 = \min_{b \in B} \|X - bZ^\tau\|^2. \tag{3.26}$$

For any $l \in \mathcal{R}^p$, $\hat{\beta}l^\tau$ is called an LSE of βl^τ. ∎

Throughout this book we consider $B = \mathcal{R}^p$, unless otherwise stated. Differentiating $\|X - bZ^\tau\|^2$ w.r.t. b, we obtain that any solution of

$$bZ^\tau Z = XZ \tag{3.27}$$

is an LSE of β. If the rank of the matrix Z is p, in which case $(Z^\tau Z)^{-1}$ exists and Z is said to be of full rank, then there is a unique LSE which is

$$\hat{\beta} = XZ(Z^\tau Z)^{-1}. \tag{3.28}$$

If Z is not of full rank, then there are infinitely many LSE's of β. It can be shown (exercise) that any LSE of β is of the form

$$\hat{\beta} = XZ(Z^\tau Z)^-, \tag{3.29}$$

where $(Z^\tau Z)^-$ is called a *generalized inverse* of $Z^\tau Z$ and satisfies

$$Z^\tau Z(Z^\tau Z)^- Z^\tau Z = Z^\tau Z.$$

Generalized inverse matrices are not unique unless Z is of full rank, in which case $(Z^\tau Z)^- = (Z^\tau Z)^{-1}$ and (3.29) reduces to (3.28).

To study properties of LSE's of β, we need some assumptions on the distribution of X. Since Z_i's are nonrandom, assumptions on the distribution of X can be expressed in terms of assumptions on the distribution of ε. Several commonly adopted assumptions are stated as follows.

Assumption A1: ε is distributed as $N_n(0, \sigma^2 I_n)$ with an unknown $\sigma^2 > 0$.
Assumption A2: $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$ with an unknown $\sigma^2 > 0$.
Assumption A3: $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon)$ is an unknown matrix.

Assumption A1 is the strongest and implies a parametric model. We may assume a slightly more general assumption that ε has the $N_n(0, \sigma^2 D)$ distribution with unknown σ^2 but a known positive definite matrix D. Let $D^{-1/2}$ be the inverse of the square root matrix of D. Then model (3.25) with assumption A1 holds if we replace X, Z, and ε by the transformed variables $\tilde{X} = XD^{-1/2}$, $\tilde{Z} = ZD^{-1/2}$, and $\tilde{\varepsilon} = \varepsilon D^{-1/2}$, respectively. A similar conclusion can be made for assumption A2.

Under assumption A1, the distribution of X is $N_n(\beta Z^\tau, \sigma^2 I_n)$, which is in an exponential family $\mathcal{P}$ with parameter $\theta = (\beta, \sigma^2) \in \mathcal{R}^p \times (0, \infty)$. However, if the matrix Z is not of full rank, then $\mathcal{P}$ is not identifiable (see §2.1.2), since $\beta_1 Z^\tau = \beta_2 Z^\tau$ does not imply $\beta_1 = \beta_2$.

Suppose that the rank of Z is $r \le p$. Then there is an $n \times r$ submatrix Z_* of Z such that

$$Z = Z_* Q \tag{3.30}$$

and Z_* is of rank r, where Q is a fixed $r \times p$ matrix. Then

$$\beta Z^\tau = \beta Q^\tau Z_*^\tau$$

and $\mathcal{P}$ is identifiable if we consider the reparameterization $\tilde{\beta} = \beta Q^\tau$. Note that the new parameter $\tilde{\beta}$ is in a subspace of $\mathcal{R}^p$ with dimension r.

In many applications we are interested in estimating some linear functions of β, i.e., $\vartheta = \beta l^\tau$ for some $l \in \mathcal{R}^p$. From the previous discussion, however, estimation of βl^τ is meaningless unless $l = cQ$ for some $c \in \mathcal{R}^r$ so that

$$\beta l^\tau = \beta Q^\tau c^\tau = \tilde{\beta} c^\tau.$$

The following result shows that βl^τ is estimable if $l = cQ$, which is also necessary for βl^τ to be estimable under assumption A1.

Theorem 3.6. Assume model (3.25) with assumption A3.
(i) A necessary and sufficient condition for $l \in \mathcal{R}^p$ being cQ for some $c \in \mathcal{R}^r$ is $l \in \mathcal{R}(Z) = \mathcal{R}(Z^\tau Z)$, where Q is given by (3.30) and $\mathcal{R}(A)$ is the smallest linear subspace of $\mathcal{R}^p$ containing all rows of A.
(ii) If $l \in \mathcal{R}(Z)$, then the LSE $\hat{\beta} l^\tau$ is unique and unbiased for βl^τ.
(iii) If $l \notin \mathcal{R}(Z)$ and assumption A1 holds, then βl^τ is not estimable.
Proof. (i) If $l = cQ$, then

$$l = cQ = c(Z_*^\tau Z_*)^{-1} Z_*^\tau Z_* Q = [c(Z_*^\tau Z_*)^{-1} Z_*^\tau] Z.$$

Hence $l \in \mathcal{R}(Z)$. If $l \in \mathcal{R}(Z)$, then $l = \zeta Z$ for some ζ and

$$l = \zeta Z_* Q = cQ$$

with $c = \zeta Z_*$.
(ii) If $l \in \mathcal{R}(Z)$, then $l = \zeta Z^\tau Z$ for some ζ and by (3.29),

$$\begin{aligned} E(\hat{\beta} l^\tau) &= E[XZ(Z^\tau Z)^- l^\tau] \\ &= \beta Z^\tau Z (Z^\tau Z)^- (Z^\tau Z) \zeta^\tau \\ &= \beta Z^\tau Z \zeta^\tau \\ &= \beta l^\tau. \end{aligned}$$

If $\tilde{\beta}$ is any other LSE of β, then, by (3.27),

$$\hat{\beta} l^\tau - \tilde{\beta} l^\tau = (\hat{\beta} - \tilde{\beta})(Z^\tau Z)\zeta^\tau = (XZ - XZ)\zeta^\tau = 0.$$

(iii) Under assumption A1, if there is an estimator $h(X, Z)$ unbiased for βl^τ, then

$$\beta l^\tau = \int_{\mathcal{R}^n} h(x, Z)(2\pi)^{-n/2} \sigma^{-n} \exp\left\{-\tfrac{1}{2\sigma^2} \|x - \beta Z^\tau\|^2\right\} dx.$$

Differentiating w.r.t. β and applying Theorem 2.1 leads to

$$l^\tau = Z^\tau \int_{\mathcal{R}^n} h(x, Z)(2\pi)^{-n/2} \sigma^{n-2} (x^\tau - Z\beta^\tau) \exp\left\{-\tfrac{1}{2\sigma^2} \|x - \beta Z^\tau\|^2\right\} dx,$$

which implies $l \in \mathcal{R}(Z)$. ∎

Theorem 3.6 shows that LSE's are unbiased for estimable parameters βl^τ. If Z is of full rank, then $\mathcal{R}(Z) = \mathcal{R}^p$ and, therefore, βl^τ is estimable for any $l \in \mathcal{R}^p$.

Example 3.12 (Simple linear regression). Let $\beta = (\beta_0, \beta_1) \in \mathcal{R}^2$ and $Z_i = (1, t_i)$, $t_i \in \mathcal{R}$, $i = 1, ..., n$. Then model (3.24) or (3.25) is called a *simple linear regression* model. It turns out that

$$Z^\tau Z = \begin{pmatrix} n & \sum_{i=1}^n t_i \\ \sum_{i=1}^n t_i & \sum_{i=1}^n t_i^2 \end{pmatrix}.$$

This matrix is invertible if and only if some t_i's are different. Thus, if some t_i's are different, then the unique unbiased LSE of βl^τ for any $l \in \mathcal{R}^2$ is $XZ(Z^\tau Z)^{-1} l^\tau$, which has the normal distribution if assumption A1 holds.

The result can be easily extended to the case of *polynomial regression* of order p in which $\beta = (\beta_0, \beta_1, ..., \beta_{p-1})$ and $Z_i = (1, t_i, ..., t_i^{p-1})$. ∎

Example 3.13 (One-way ANOVA). Suppose that $n = \sum_{j=1}^m n_j$ with m positive integers $n_1, ..., n_m$ and that

$$X_i = \mu_j + \varepsilon_i, \qquad i = n_{j-1} + 1, ..., n_j, j = 1, ..., m,$$

where $n_0 = 0$ and $(\mu_1, ..., \mu_m) = \beta$. Let J_m be the m-vector of ones. Then the matrix Z in this case is a block diagonal matrix with $J_{n_j}^\tau$ as the jth diagonal block. Consequently, $Z^\tau Z$ is an $m \times m$ diagonal matrix whose jth diagonal element is n_j. Thus, $Z^\tau Z$ is invertible and the unique LSE of β is the m-vector whose jth component is $n_j^{-1} \sum_{i=n_{j-1}+1}^{n_j} X_i$, $j = 1, ..., m$.

Sometimes it is more convenient to use the following notation:

$$X_{ij} = X_{n_{i-1}+j}, \ \varepsilon_{ij} = \varepsilon_{n_{i-1}+j}, \qquad j = 1, ..., n_i, i = 1, ..., m,$$

and

$$\mu_i = \mu + \alpha_i, \qquad i = 1, ..., m.$$

Then our model becomes

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad j = 1, ..., n_i, i = 1, ..., m, \tag{3.31}$$

which is called a *one-way analysis of variance* (ANOVA) model. Under model (3.31), $\beta = (\mu, \alpha_1, ..., \alpha_m) \in \mathcal{R}^{m+1}$. The matrix Z under model (3.31) is not of full rank (exercise). The LSE of β under model (3.31) is

$$\hat{\beta} = \left(\bar{X}, \bar{X}_{1\cdot} - \bar{X}, ..., \bar{X}_{m\cdot} - \bar{X}\right),$$

where $\bar{X}$ is still the sample mean of X_{ij}'s and $\bar{X}_{i\cdot}$ is the sample mean of the ith group $\{X_{ij}, j = 1, ..., n_i\}$. The problem of finding the form of $l^\tau \in \mathcal{R}(Z)$ under model (3.31) is left as an exercise. ∎

The notation used in model (3.31) allows us to generalize the one-way ANOVA model to any s-way ANOVA model with a positive integer s under

the so-called factorial experiments. The following example is for the two-way ANOVA model.

Example 3.14 (Two-way balanced ANOVA). Suppose that

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, ..., a, j = 1, ..., b, k = 1, ..., c, \tag{3.32}$$

where a, b, and c are some positive integers. Model (3.32) is called a two-way balanced ANOVA model. If we view model (3.32) as a special case of model (3.25), then the parameter vector β is

$$\beta = (\mu, \alpha_1, ..., \alpha_a, \beta_1, ..., \beta_b, \gamma_{11}, ..., \gamma_{1b}, ..., \gamma_{a1}, ..., \gamma_{ab}). \tag{3.33}$$

One can obtain the matrix Z and show that it is $n \times p$, where $n = abc$ and $p = 1 + a + b + ab$, and is of rank $ab < p$ (exercise). It can also be shown (exercise) that the LSE of β is given by the right-hand side of (3.33) with μ, α_i, β_j, and γ_{ij} replaced by $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$, and $\hat{\gamma}_{ij}$, respectively, where $\hat{\mu} = \bar{X}_{...}$, $\hat{\alpha}_i = \bar{X}_{i..} - \bar{X}_{...}$, $\hat{\beta}_j = \bar{X}_{.j.} - \bar{X}_{...}$, $\hat{\gamma}_{ij} = \bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...}$, and a dot is used to denote averaging over the indicated subscript, e.g.,

$$\bar{X}_{.j.} = \frac{1}{ac} \sum_{i=1}^{a} \sum_{k=1}^{c} X_{ijk}$$

with a fixed j. ∎

3.3.2 The UMVUE and BLUE

We now study UMVUE's in model (3.25) with assumption A1.

Theorem 3.7. Consider model (3.25) with assumption A1.
(i) The LSE $\hat{\beta} l^\tau$ is the UMVUE of βl^τ for any estimable βl^τ.
(ii) The UMVUE of σ^2 is $\hat{\sigma}^2 = (n-r)^{-1} \|X - \hat{\beta} Z^\tau\|^2$, where r is the rank of Z.
(iii) The UMVUE's in (i) and (ii) attain the Cramér-Rao lower bound.
Proof. (i) Let $\hat{\beta}$ be an LSE of β. By (3.27),

$$(X - \hat{\beta} Z^\tau) Z (\hat{\beta} - \beta)^\tau = (XZ - XZ)(\hat{\beta} - \beta)^\tau = 0$$

and, hence,

$$\begin{aligned} \|X - \beta Z^\tau\|^2 &= \|X - \hat{\beta} Z^\tau + \hat{\beta} Z^\tau - \beta Z^\tau\|^2 \\ &= \|X - \hat{\beta} Z^\tau\|^2 + \|\hat{\beta} Z^\tau - \beta Z^\tau\|^2 \\ &= \|X - \hat{\beta} Z^\tau\|^2 - 2XZ\beta^\tau + \|\beta Z^\tau\|^2 + \|\hat{\beta} Z^\tau\|^2. \end{aligned}$$

Using this result and assumption A1, we obtain the following joint Lebesgue p.d.f. of X:

$$(2\pi\sigma^2)^{-n/2}\exp\left\{\frac{xZ\beta^\tau}{\sigma^2}-\frac{\|x-\hat\beta Z^\tau\|^2+\|\hat\beta Z^\tau\|^2}{2\sigma^2}-\frac{\|\beta Z^\tau\|^2}{2\sigma^2}\right\}.$$

By Proposition 2.1 and the fact that $\hat\beta Z^\tau = XZ(Z^\tau Z)^- Z^\tau$ is a function of XZ, $(XZ, \|X-\hat\beta Z^\tau\|^2)$ is complete and sufficient for $\theta=(\beta,\sigma^2)$. Note that $\hat\beta$ is a function of XZ and, hence, a function of the complete sufficient statistic. If βl^τ is estimable, then $\hat\beta l^\tau$ is unbiased for βl^τ (Theorem 3.6) and, hence, $\hat\beta l^\tau$ is the UMVUE of βl^τ.
(ii) Since each column of $Z^\tau \in \mathcal{R}(Z)$, $\hat\beta Z^\tau$ does not depend on the choice of $\hat\beta$ and $E(\hat\beta Z^\tau)=\beta Z^\tau$ (Theorem 3.6). Then

$$\mathrm{Cov}(X-\hat\beta Z^\tau, \hat\beta Z^\tau) = E(X-\hat\beta Z^\tau)Z\hat\beta^\tau = E(XZ-XZ)\hat\beta^\tau = 0 \quad (3.34)$$

and

$$\begin{aligned} E\|X-\hat\beta Z^\tau\|^2 &= E(X-\beta Z^\tau)(X-\beta Z^\tau)^\tau - E[(\beta-\hat\beta)Z^\tau Z(\beta-\hat\beta)^\tau] \\ &= \mathrm{tr}\left(\mathrm{Var}(X)-\mathrm{Var}(\hat\beta Z^\tau)\right) \\ &= \sigma^2[n-\mathrm{tr}\left(Z(Z^\tau Z)^- Z^\tau Z(Z^\tau Z)^- Z^\tau\right)] \\ &= \sigma^2[n-\mathrm{tr}\left((Z^\tau Z)^- Z^\tau Z\right)]. \end{aligned}$$

Since the previous result does not depend on the particular choice of $\hat\beta$ or $(Z^\tau Z)^-$, we can evaluate $\mathrm{tr}((Z^\tau Z)^- Z^\tau Z)$ using a particular $(Z^\tau Z)^-$. From the theory of linear algebra, there exists a $p\times p$ matrix C such that $CC^\tau = I_p$ and

$$C(Z^\tau Z)C^\tau = \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix},$$

where Λ is an $r\times r$ diagonal matrix whose diagonal elements are positive. Then a particular choice of $(Z^\tau Z)^-$ is

$$(Z^\tau Z)^- = C\begin{pmatrix} \Lambda^{-1} & 0 \\ 0 & 0 \end{pmatrix}C^\tau \quad (3.35)$$

and

$$(Z^\tau Z)^- Z^\tau Z = C\begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}C^\tau$$

whose trace is r. Hence $\hat\sigma^2$ is the UMVUE of σ^2, since it is a function of the complete sufficient statistic and

$$E\hat\sigma^2 = (n-r)^{-1}E\|X-\hat\beta Z^\tau\|^2 = \sigma^2.$$

(iii) The result follows from Proposition 3.2. ∎

The vector $X - \hat{\beta}Z^\tau$ is called the *residual vector* and $\|X - \hat{\beta}Z^\tau\|^2$ is called the *sum of squared residuals* and is denoted by SSR. The estimator $\hat{\sigma}^2$ is then equal to $SSR/(n-r)$.

Since $X - \hat{\beta}Z^\tau$ and $\hat{\beta}l^\tau$ are linear in X, they are normally distributed under assumption A1. Then (3.34) and assumption A1 imply that $\hat{\sigma}^2$ and $\hat{\beta}l^\tau$ are independent for any estimable βl^τ. Furthermore, using the generalized inverse matrix in (3.35), we obtain that

$$SSR = X[I_n - Z(Z^\tau Z)^- Z^\tau]X^\tau, \tag{3.36}$$

where $P_n = I_n - Z(Z^\tau Z)^- Z^\tau$ is a projection matrix of rank $n-r$. Then, there exists an $n \times n$ matrix G such that $GG^\tau = I_n$ and

$$P_n G = \left(G_1^\tau, ..., G_{n-r}^\tau, 0, ..., 0\right),$$

where G_j is the jth row of G^τ. This and (3.36) imply that

$$SSR = \sum_{j=1}^{n-r} Y_j^2,$$

where $Y_j = XG_j^\tau$. Let $Y = (Y_1, ..., Y_{n-r})$. Under assumption A1, Y is normal; $\mathrm{Var}(Y) = \sigma^2 I_{n-r}$; and $EY = 0$ since

$$EY_j = E(XG_j^\tau) = \beta Z^\tau P_n G_j^\tau = 0,$$

which follows from the fact that $Z^\tau P_n P_n Z = Z^\tau P_n Z = 0$ by the definition of the generalized inverse. Thus, we have the following result.

Theorem 3.8. Consider model (3.25) with assumption A1. For any estimable parameter βl^τ, the UMVUE's $\hat{\beta}l^\tau$ and $\hat{\sigma}^2$ are independent; the distribution of $\hat{\beta}l^\tau$ is $N(\beta l^\tau, \sigma^2 l(Z^\tau Z)^- l^\tau)$; and $(n-r)\hat{\sigma}^2/\sigma^2$ has the chi-square distribution χ^2_{n-r}. ∎

Example 3.15. In Examples 3.12-3.14, UMVUE's of estimable βl^τ are the LSE's $\hat{\beta}l^\tau$, under assumption A1. In Example 3.13,

$$SSR = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2;$$

in Example 3.14, if $c > 1$,

$$SSR = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk} - \bar{X}_{ij\cdot})^2. \quad ∎$$

We now study properties of $\hat{\beta}l^\tau$ and $\hat{\sigma}^2$ under assumption A2, i.e., without the normality assumption on ε. From Theorem 3.6 and the proof of Theorem 3.7(ii), $\hat{\beta}l^\tau$ (with an $l \in \mathcal{R}(Z)$) and $\hat{\sigma}^2$ are still unbiased without the normality assumption. In what sense are $\hat{\beta}l^\tau$ and $\hat{\sigma}^2$ optimal beyond being unbiased? We have the following result for the LSE $\hat{\beta}l^\tau$. Some discussion about $\hat{\sigma}^2$ can be found, for example, in Rao (1973, p. 228).

Theorem 3.9. Consider model (3.25) with assumption A2.
(i) A necessary and sufficient condition for the existence of a linear function of X that is unbiased for βl^τ is $l \in \mathcal{R}(Z)$.
(ii) (The Gauss-Markov theorem). If $l \in \mathcal{R}(Z)$, then the LSE $\hat{\beta}l^\tau$ is the *best linear unbiased estimator* (BLUE) of βl^τ in the sense that it has the minimum variance in the class of linear unbiased estimators of βl^τ.
Proof. (i) The sufficiency has been established in Theorem 3.6. Suppose now a linear function of X, Xc^τ with $c \in \mathcal{R}^n$, is unbiased for βl^τ. Then

$$\beta l^\tau = E(Xc^\tau) = (EX)c^\tau = \beta Z^\tau c^\tau.$$

Since this equality holds for all β, $l = cZ$, i.e., $l \in \mathcal{R}(Z)$.
(ii) Let $l \in \mathcal{R}(Z) = \mathcal{R}(Z^\tau Z)$. Then $l = \zeta(Z^\tau Z)$ for some ζ and $\hat{\beta}l^\tau = \hat{\beta}(Z^\tau Z)\zeta^\tau = XZ\zeta^\tau$ by (3.27). Let Xc^τ be any linear unbiased estimator of βl^τ. From the proof of (i), $cZ = l$. Then

$$\begin{aligned}\mathrm{Cov}(XZ\zeta^\tau, Xc^\tau - XZ\zeta^\tau) &= E(XZ\zeta^\tau cX^\tau) - E(XZ\zeta^\tau\zeta Z^\tau X^\tau)\\ &= \sigma^2\mathrm{tr}(Z\zeta^\tau c) - \sigma^2\mathrm{tr}(Z\zeta^\tau\zeta Z^\tau)\\ &= \sigma^2[\mathrm{tr}(\zeta^\tau cZ) - \mathrm{tr}(\zeta^\tau l)] = 0.\end{aligned}$$

Hence

$$\begin{aligned}\mathrm{Var}(Xc^\tau) &= \mathrm{Var}(Xc^\tau - XZ\zeta^\tau + XZ\zeta^\tau)\\ &= \mathrm{Var}(Xc^\tau - XZ\zeta^\tau) + \mathrm{Var}(XZ\zeta^\tau)\\ &\quad + 2\mathrm{Cov}(XZ\zeta^\tau, Xc^\tau - XZ\zeta^\tau)\\ &= \mathrm{Var}(Xc^\tau - XZ\zeta^\tau) + \mathrm{Var}(\hat{\beta}l^\tau)\\ &\geq \mathrm{Var}(\hat{\beta}l^\tau). \quad \blacksquare\end{aligned}$$

3.3.3 Robustness of LSE's

Consider now model (3.25) under assumption A3. An interesting question is under what conditions on $\mathrm{Var}(\varepsilon)$, the LSE of βl^τ with $l \in \mathcal{R}(Z)$ is still the BLUE. If $\hat{\beta}l^\tau$ is still the BLUE, then we say that $\hat{\beta}l^\tau$, considered as a BLUE, is *robust* against violation of assumption A2. In general, a statistical procedure having certain properties under an assumption is said

to be robust against violation of the assumption if the statistical procedure still has the same properties if the assumption is (slightly) violated. For example, the LSE of βl^τ with $l \in \mathcal{R}(Z)$, as an unbiased estimator, is robust against violation of assumption A1 or A2, since the LSE is unbiased as long as $E(\varepsilon) = 0$, which can be always assumed without loss of generality. On the other hand, the LSE as a UMVUE may not be robust against violation of assumption A1 (see §3.5).

Theorem 3.10. Consider model (3.25) with assumption A3. The following are equivalent.
(a) $\hat{\beta} l^\tau$ is the BLUE of βl^τ for any $l \in \mathcal{R}(Z)$.
(b) $E(\hat{\beta} l^\tau X \eta^\tau) = 0$ for any $l \in \mathcal{R}(Z)$ and any η such that $EX\eta^\tau = 0$.
(c) $Z^\tau \mathrm{Var}(\varepsilon) U = 0$, where U is a matrix such that $Z^\tau U = 0$ and $\mathcal{R}(U^\tau) + \mathcal{R}(Z^\tau) = \mathcal{R}^n$.
(d) $\mathrm{Var}(\varepsilon) = Z\Lambda_1 Z^\tau + U\Lambda_2 U^\tau$ for some Λ_1 and Λ_2.
(e) The matrix $Z(Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon)$ is symmetric.
Proof. We first show that (a) and (b) are equivalent, which is an analogue of Theorem 3.2(i). Suppose that (b) holds. Let $l \in \mathcal{R}(Z)$. If Xc^τ is another unbiased estimator of βl^τ, then $E(X\eta^\tau) = 0$ with $\eta = c - l(Z^\tau Z)^- Z^\tau$. Hence

$$\begin{aligned}
\mathrm{Var}(Xc^\tau) &= \mathrm{Var}(Xc^\tau - \hat{\beta} l^\tau + \hat{\beta} l^\tau) \\
&= \mathrm{Var}(Xc^\tau - XZ(Z^\tau Z)^- l^\tau + \hat{\beta} l^\tau) \\
&= \mathrm{Var}(X\eta^\tau + \hat{\beta} l^\tau) \\
&= \mathrm{Var}(X\eta^\tau) + \mathrm{Var}(\hat{\beta} l^\tau) + 2\mathrm{Cov}(X\eta^\tau, \hat{\beta} l^\tau) \\
&= \mathrm{Var}(X\eta^\tau) + \mathrm{Var}(\hat{\beta} l^\tau) + 2E(\hat{\beta} l^\tau X\eta^\tau) \\
&= \mathrm{Var}(X\eta^\tau) + \mathrm{Var}(\hat{\beta} l^\tau) \\
&\geq \mathrm{Var}(\hat{\beta} l^\tau).
\end{aligned}$$

Suppose now that there are $l \in \mathcal{R}(Z)$ and η such that $E(X\eta^\tau) = 0$ but $\delta = E(\hat{\beta} l^\tau X\eta^\tau) \neq 0$. Let $c_t = t\eta + l(Z^\tau Z)^- Z^\tau$. From the previous proof we obtain that

$$\mathrm{Var}(Xc_t^\tau) = t^2 \mathrm{Var}(X\eta^\tau) + \mathrm{Var}(\hat{\beta} l^\tau) + 2\delta t.$$

As long as $\delta \neq 0$, there exists a t such that $\mathrm{Var}(Xc_t^\tau) < \mathrm{Var}(\hat{\beta} l^\tau)$. This shows that $\hat{\beta} l^\tau$ cannot be a BLUE and, therefore, (a) implies (b).

We next show that (b) implies (c). Suppose that (b) holds. Since $l \in \mathcal{R}(Z)$, $l = \gamma Z$ for some γ. For any η such that $E(X\eta^\tau) = 0$,

$$0 = E(l\hat{\beta}^\tau X^\tau \eta) = E[\gamma Z(Z^\tau Z)^- Z^\tau X X^\tau \eta] = \gamma Z(Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon)\eta.$$

Since this equality holds for all γ, $Z(Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon)\eta = 0$. Note that $E(X\eta^\tau) = \beta Z\eta^\tau = 0$ for all β. Hence $Z\eta^\tau = 0$, i.e., $\eta \in \mathcal{R}(U)$. Since this

is true for all η,

$$Z(Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon)U = 0,$$

which implies

$$Z^\tau Z(Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon)U = Z^\tau \mathrm{Var}(\varepsilon)U = 0,$$

since $Z^\tau Z(Z^\tau Z)^- Z^\tau = Z^\tau$. Thus, (c) holds.

To show that (c) implies (d), we need to use the following facts from the theory of linear algebra: there exists nonsingular matrix C such that $\mathrm{Var}(\varepsilon) = CC^\tau$ and $C = ZC_1 + UC_2$ for some matrices C_j (since $\mathcal{R}(U^\tau) + \mathcal{R}(Z^\tau) = \mathcal{R}^n$). Let $\Lambda_1 = C_1C_1^\tau$, $\Lambda_2 = C_2C_2^\tau$, and $\Lambda_3 = C_1C_2^\tau$. Then

$$\mathrm{Var}(\varepsilon) = Z\Lambda_1 Z^\tau + U\Lambda_2 U^\tau + Z\Lambda_3 U^\tau + U\Lambda_3^\tau Z^\tau \tag{3.37}$$

and $Z^\tau \mathrm{Var}(\varepsilon)U = Z^\tau Z\Lambda_3 U^\tau U$, which is 0 if (c) holds. Hence, (c) implies

$$0 = Z(Z^\tau Z)^- Z^\tau Z\Lambda_3 U^\tau U(U^\tau U)^- U^\tau = Z\Lambda_3 U^\tau,$$

which with (3.37) implies (d).

If (d) holds, then $Z(Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon) = Z\Lambda_1 Z^\tau$, which is symmetric. Hence (d) implies (e). To complete the proof we need to show that (e) implies (b), which is left as an exercise. ∎

As a corollary of this theorem, the following result shows when the UMVUE's in model (3.25) with assumption A1 is robust against the violation of $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$.

Corollary 3.3. Consider model (3.25) with normally distributed ε and a full rank Z. Then $\hat{\beta}l^\tau$ and $\hat{\sigma}^2$ are still UMVUE's of βl^τ and σ^2, respectively, if and only if one of (b)-(e) in Theorem 3.10 holds. ∎

Example 3.16. Consider model (3.25) with β replaced by a random vector $\boldsymbol{\beta}$ which is independent of ε. Such a model is called a linear model with random coefficients. Suppose that $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$, $E(\boldsymbol{\beta}) = \beta$. Then

$$X = \beta Z^\tau + (\boldsymbol{\beta} - \beta)Z^\tau + \varepsilon = \beta Z^\tau + e, \tag{3.38}$$

where $e = (\boldsymbol{\beta} - \beta)Z^\tau + \varepsilon$ satisfies $E(e) = 0$ and

$$\mathrm{Var}(e) = Z\mathrm{Var}(\boldsymbol{\beta})Z^\tau + \sigma^2 I_n.$$

Since

$$Z(Z^\tau Z)^- Z^\tau \mathrm{Var}(e) = Z\mathrm{Var}(\boldsymbol{\beta})Z^\tau + \sigma^2 Z(Z^\tau Z)^- Z^\tau$$

is symmetric, by Theorem 3.10, the LSE $\hat{\beta}l^\tau$ under model (3.38) is the BLUE for any βl^τ, $l \in \mathcal{R}(Z)$. If Z is of full rank and ε is normal, then, by Corollary 3.3, $\hat{\beta}l^\tau$ is the UMVUE. ∎

Example 3.17 (Random effects models). Suppose that

$$X_{ij} = \mu + A_i + e_{ij}, \quad j = 1, ..., n_i, i = 1, ..., m, \tag{3.39}$$

where $\mu \in \mathcal{R}$ is an unknown parameter, A_i's are i.i.d. random variables having mean 0 and variance σ_a^2, e_{ij}'s are i.i.d. random errors with mean 0 and variance σ^2, and A_i's and e_{ij}'s are independent. Model (3.39) is called a one-way *random effects* model and A_i's are unobserved random effects. Let $\varepsilon_{ij} = A_i + e_{ij}$. Then (3.39) is a special case of the general model (3.25) with

$$\mathrm{Var}(\varepsilon) = \sigma_a^2 \Sigma + \sigma^2 I_n,$$

where Σ is a block diagonal matrix whose ith block is $J_{n_i}^\tau J_{n_i}$ and J_k is the k-vector of ones. Under this model, $Z = J_n^\tau$, $n = \sum_{i=1}^m n_i$, and $Z(Z^\tau Z)^- Z^\tau = n^{-1} J_n^\tau J_n$. Note that

$$J_n^\tau J_n \Sigma = \begin{pmatrix} n_1 J_{n_1}^\tau J_{n_1} & n_2 J_{n_1}^\tau J_{n_2} & \cdots & n_m J_{n_1}^\tau J_{n_m} \\ n_1 J_{n_2}^\tau J_{n_1} & n_2 J_{n_2}^\tau J_{n_2} & \cdots & n_m J_{n_2}^\tau J_{n_m} \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ n_1 J_{n_m}^\tau J_{n_1} & n_2 J_{n_m}^\tau J_{n_2} & \cdots & n_m J_{n_m}^\tau J_{n_m} \end{pmatrix},$$

which is symmetric if and only if $n_1 = n_2 = \cdots = n_m$. Since $J_n^\tau J_n \mathrm{Var}(\varepsilon)$ is symmetric if and only if $J_n^\tau J_n \Sigma$ is symmetric, a necessary and sufficient condition for the LSE of μ to be the BLUE is that all n_i's are the same. This condition is also necessary and sufficient for the LSE of μ to be the UMVUE when ε_{ij}'s are normal. ∎

In some cases we are interested in some (not all) linear functions of β. For example, consider βl^τ with $l \in \mathcal{R}(H)$, where H is an $n \times p$ matrix such that $\mathcal{R}(H) \subset \mathcal{R}(Z)$. We have the following result.

Proposition 3.4. Consider model (3.25) with assumption A3. Suppose that H is a matrix such that $\mathcal{R}(H) \subset \mathcal{R}(Z)$. A necessary and sufficient condition for the LSE $\hat{\beta} l^\tau$ to be the BLUE for any $l \in \mathcal{R}(H)$ is $H(Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon) U = 0$, where U is the same as that in (c) of Theorem 3.10. ∎

Example 3.18. Consider model (3.25) with assumption A3 and $Z = (H_1, H_2)$, where $H_1^\tau H_2 = 0$. Suppose that under the reduced model

$$X = \beta_1 H_1^\tau + \varepsilon,$$

$\hat{\beta}_1 l^\tau$ is the BLUE for any $\beta_1 l^\tau$, $l \in \mathcal{R}(H_1)$, and that under the reduced model

$$X = \beta_2 H_2^\tau + \varepsilon,$$

$\hat{\beta}_2 l^\tau$ is not a BLUE for some $\beta_2 l^\tau$, $l \in \mathcal{R}(H_2)$, where $\beta = (\beta_1, \beta_2)$ and $\hat{\beta}_j$'s are LSE's under the reduced models. Let $H = (H_1, 0)$ be $n \times p$. Note that

$$H(Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon) U = H_1 (H_1^\tau H_1)^- H_1^\tau \mathrm{Var}(\varepsilon) U,$$

which is 0 by Theorem 3.10 for the U given in (c) of Theorem 3.10, and

$$Z(Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon) U = H_2 (H_2^\tau H_2)^- H_2^\tau \mathrm{Var}(\varepsilon) U,$$

which is not 0 by Theorem 3.10. This implies that some LSE $\hat{\beta} l^\tau$ is not a BLUE but $\hat{\beta} l^\tau$ is the BLUE if $l \in \mathcal{R}(H)$. ∎

Finally, we consider model (3.25) with $\mathrm{Var}(\varepsilon)$ being a diagonal matrix whose ith diagonal element is σ_i^2, i.e., ε_i's are uncorrelated but have unequal variances. A straightforward calculation shows that condition (e) in Theorem 3.10 holds if and only if, for all $i \neq j$, $\sigma_i^2 \neq \sigma_j^2$ only when $h_{ij} = 0$, where h_{ij} is the (i, j)th element of the projection matrix $Z(Z^\tau Z)^- Z^\tau$. Thus, the LSE's are not BLUE's in general.

Suppose that the unequal variances of ε_i's are caused by some small perturbations, i.e., $\varepsilon_i = e_i + u_i$, where $\mathrm{Var}(e_i) = \sigma^2$, $\mathrm{Var}(u_i) = \delta_i$, and e_i and u_i are independent so that $\sigma_i^2 = \sigma^2 + \delta_i$. If $\delta_i = 0$ for all i (no perturbations), then assumption A2 holds and any LSE $\hat{\beta} l^\tau$ is the BLUE with variance

$$\mathrm{Var}(\hat{\beta} l^\tau) = \sigma^2 l (Z^\tau Z)^- l^\tau.$$

When $\delta_i > 0$, $\hat{\beta} l^\tau$ is still unbiased for βl^τ, $l \in \mathcal{R}(Z)$, and

$$\mathrm{Var}(\hat{\beta} l^\tau) = l (Z^\tau Z)^- \sum_{i=1}^n \sigma_i^2 Z_i^\tau Z_i (Z^\tau Z)^- l^\tau.$$

Suppose that $\delta_i \leq \sigma^2 \delta$. Then

$$\mathrm{Var}(\hat{\beta} l^\tau) \leq (1 + \delta) \sigma^2 l (Z^\tau Z)^- l^\tau.$$

This indicates that the LSE is robust in the sense that its variance increases slightly when there is a slight violation of the equal variance assumption (small δ).

3.3.4 Asymptotic properties of LSE's

We consider first the consistency of the LSE $\hat{\beta} l^\tau$ with $l \in \mathcal{R}(Z)$ for every n.

Theorem 3.11. Consider model (3.25) with assumption A3. Suppose that $\sup_n \lambda_+[\mathrm{Var}(\varepsilon)] < \infty$, where $\lambda_+[A]$ is the largest eigenvalue of the matrix

A, and that $\lim_{n\to\infty}\lambda_+[(Z^\tau Z)^-] = 0$. Then $\hat{\beta}l^\tau$ is consistent in mse for any $l \in \mathcal{R}(Z)$.

Proof. The result follows from the fact that $\hat{\beta}l^\tau$ is unbiased and

$$\begin{aligned}\mathrm{Var}(\hat{\beta}l^\tau) &= l(Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon) Z(Z^\tau Z)^- l^\tau \\ &\le \lambda_+[\mathrm{Var}(\varepsilon)] l(Z^\tau Z)^- l^\tau. \quad \blacksquare\end{aligned}$$

Without the normality assumption on ε, the exact distribution of $\hat{\beta}l^\tau$ is very hard to obtain. The asymptotic distribution of $\hat{\beta}l^\tau$ is derived in the following result.

Theorem 3.12. Consider model (3.25) with assumption A3. Suppose that $0 < \inf_n \lambda_-[\mathrm{Var}(\varepsilon)]$, where $\lambda_-[A]$ is the smallest eigenvalue of the matrix A, and that

$$\lim_{n\to\infty} \max_{1\le i\le n} Z_i(Z^\tau Z)^- Z_i^\tau = 0. \tag{3.40}$$

Suppose further that $n = \sum_{j=1}^k m_j$ for some integers k, m_j, $j = 1, ..., k$, with m_j's bounded by a fixed integer m, $\varepsilon = (\xi_1, ..., \xi_k)$, $\xi_j \in \mathcal{R}^{m_j}$, and ξ_j's are independent.

(i) If $\sup_i E|\varepsilon_i|^{2+\delta} < \infty$, then for any $l \in \mathcal{R}(Z)$,

$$(\hat{\beta} - \beta)l^\tau \Big/ \sqrt{\mathrm{Var}(\hat{\beta}l^\tau)} \to_d N(0,1). \tag{3.41}$$

(ii) If ξ_i's are i.i.d., then result (3.41) holds.

Proof. Let $l \in \mathcal{R}(Z)$. Then

$$\beta Z^\tau Z(Z^\tau Z)^- l^\tau - \beta l^\tau = 0$$

and

$$(\hat{\beta} - \beta)l^\tau = \varepsilon Z(Z^\tau Z)^- l^\tau = \sum_{j=1}^k \xi_j c_{nj}^\tau,$$

where c_{nj} is the m_j-vector whose components are $l(Z^\tau Z)^- Z_i^\tau$, $i = m_{j-1} + 1, ..., m_j$, $m_0 = 0$. Note that

$$\sum_{j=1}^k \|c_{nj}\|^2 = l(Z^\tau Z)^- Z^\tau Z(Z^\tau Z)^- l^\tau = l(Z^\tau Z)^- l^\tau. \tag{3.42}$$

Also,

$$\begin{aligned}\max_{1\le j\le k} \|c_{nj}\|^2 &\le m \max_{1\le i\le n} [l(Z^\tau Z)^- Z_i^\tau]^2 \\ &\le m l(Z^\tau Z)^- l^\tau \max_{1\le i\le n} Z_i(Z^\tau Z)^- Z_i^\tau,\end{aligned}$$

which, together with (3.42) and condition (3.40), implies that

$$\lim_{n\to\infty}\left(\max_{1\le j\le k}\|c_{nj}\|^2\Big/\sum_{j=1}^{k}\|c_{nj}\|^2\right)=0.$$

The results then follow from Corollary 1.3. ∎

Under the conditions of Theorem 3.12, $\mathrm{Var}(\varepsilon)$ is a diagonal block matrix with $\mathrm{Var}(\xi_j)$ as the jth diagonal block, which includes the case of independent ε_i's as a special case.

The following lemma tells us how to check condition (3.40).

Lemma 3.3. The following are sufficient conditions for (3.40).
(a) $\lambda_+[(Z^\tau Z)^-]\to 0$ and $Z_n(Z^\tau Z)^- Z_n^\tau\to 0$, as $n\to\infty$.
(b) There is an increasing sequence $\{a_n\}$ such that $a_n\to\infty$ and $Z^\tau Z/a_n$ converges to a positive definite matrix. ∎

If $n^{-1}\sum_{i=1}^n t_i^2\to c$ in the simple linear regression model (Example 3.12), where c is a positive constant, then condition (b) in Lemma 3.3 is satisfied with $a_n=n$ and, therefore, Theorem 3.12 applies. In the one-way ANOVA model (Example 3.13),

$$\max_{1\le i\le n} Z_i(Z^\tau Z)^- Z_i^\tau=\lambda_+[(Z^\tau Z)^-]=\max_{1\le j\le m} n_j^{-1}.$$

Hence conditions related to Z in Theorem 3.12 are satisfied if and only if $\min_j n_j\to\infty$. Some similar conclusions can be drawn in the two-way ANOVA model (Example 3.14).

3.4 Unbiased Estimators in Survey Problems

In this section we consider unbiased estimation for another type of non-i.i.d. data often encountered in applications: survey data from finite populations. A description of the problem is given in Example 2.3 of §2.1.1. Examples and a fuller account of theoretical aspects of survey sampling can be found in, for example, Cochran (1977) and Särndal, Swensson, and Wretman (1992).

3.4.1 UMVUE's of population totals

We use the same notation as in Example 2.3. Let $X=(X_1,...,X_n)$ be a sample from a finite population $\mathcal{P}=\{y_1,...,y_N\}$ with

$$P(X_1=y_{i_1},...,X_n=y_{i_n})=p(\boldsymbol{s}),$$

where $\boldsymbol{s} = \{i_1, ..., i_n\}$ is a subset of distinct elements of $\{1, ..., N\}$ and p is a selection probability measure. We consider univariate y_i, although most of our conclusions are valid for the case of multivariate y_i. In many survey problems the parameter to be estimated is $Y = \sum_{i=1}^N y_i$, the population total.

In Example 2.27, it is shown that $\hat{Y} = N\bar{X} = \frac{N}{n}\sum_{i\in\boldsymbol{s}} y_i$ is unbiased for Y if $p(\boldsymbol{s})$ is constant (simple random sampling); a formula of $\mathrm{Var}(\hat{Y})$ is also given. We now show that $\hat{Y}$ is in fact the UMVUE of Y under simple random sampling. Let $\mathcal{Y}$ be the range of y_i, $\theta = (y_1, ..., y_N)$ and $\Theta = \prod_{i=1}^N \mathcal{Y}$. Under simple random sampling, the population under consideration is a parametric family indexed by $\theta \in \Theta$.

Theorem 3.13. (i) (Watson and Royall). If $p(\boldsymbol{s}) > 0$ for all $\boldsymbol{s}$, then the vector of order statistics $X_{(1)} \le \cdots \le X_{(n)}$ is complete for $\theta \in \Theta$.
(ii) Under simple random sampling, the vector of order statistics is sufficient for $\theta \in \Theta$.
(iii) Under simple random sampling, for any estimable function of θ, its unique UMVUE is the unbiased estimator $h(X_1, ..., X_n)$, where h is symmetric in its n arguments.
Proof. (i) Let $h(X)$ be a function of the order statistics. Then h is symmetric in its n arguments. We need to show that if

$$E[h(X)] = \sum_{\boldsymbol{s}=\{i_1,...,i_n\}\subset\{1,...,N\}} p(\boldsymbol{s})h\left(y_{i_1}, ..., y_{i_n}\right) = 0 \tag{3.43}$$

for all $\theta \in \Theta$, then $h(y_{i_1}, ..., y_{i_n}) = 0$ for all $y_{i_1}, ..., y_{i_n}$. First, suppose that all N elements of θ are equal to $a \in \mathcal{Y}$. Then (3.43) implies $h(a, ..., a) = 0$. Next, suppose that $N-1$ elements in θ are equal to a and one is $b > a$. Then (3.43) reduces to

$$q_1 h(a, ..., a) + q_2 h(a, ..., a, b),$$

where q_1 and q_2 are some known numbers in $(0, 1)$. Since $h(a, ..., a) = 0$ and $q_2 \neq 0$, $h(a, ..., a, b) = 0$. Using the same argument, we can show that $h(a, ..., a, b, ..., b) = 0$ for any k a's and $n-k$ b's. Suppose next that elements of θ are equal to a, b, or c, $a < b < c$. Then we can show that $h(a, ..., a, b, ..., b, c, ..., c) = 0$ for any k a's, l b's, and $n-k-l$ c's. Continuing inductively, we see that $h(y_1, ..., y_n) = 0$ for all possible $y_1, ..., y_n$. This completes the proof of (i).
(ii) The result follows from the factorization theorem (Theorem 2.2), the fact that $p(\boldsymbol{s})$ is constant under simple random sampling, and

$$P(X_1 = y_{i_1}, ..., X_n = y_{i_n}) = P(X_{(1)} = y_{(i_1)}, ..., X_{(n)} = y_{(i_n)})/n!,$$

where $y_{(i_1)} \le \cdots \le y_{(i_n)}$ are the ordered values of $y_{i_1}, ..., y_{i_n}$.
(iii) The result follows directly from (i) and (ii). ∎

It is interesting to note the following two issues. (1) Although we have a parametric problem under simple random sampling, the sufficient and complete statistic is the same as that in a nonparametric problem (Example 2.17). (2) For the completeness of the order statistics, we do not need the assumption of simple random sampling.

Example 3.19. From Example 2.27, $\hat{Y} = N\bar{X}$ is unbiased for Y. Since $\hat{Y}$ is symmetric in its arguments, it is the UMVUE of Y. We now derive the UMVUE for $\mathrm{Var}(\hat{Y})$. From Example 2.27,

$$\mathrm{Var}(\hat{Y}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\sigma^2, \qquad \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \frac{Y}{N}\right)^2. \tag{3.44}$$

It can be shown (exercise) that $E(S^2) = \sigma^2$, where S^2 is the usual sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i\in \boldsymbol{s}}\left(y_i - \frac{\hat{Y}}{N}\right)^2.$$

Since S^2 is symmetric in its arguments, $\frac{N^2}{n}\left(1 - \frac{n}{N}\right)S^2$ is the UMVUE of $\mathrm{Var}(\hat{Y})$. ∎

Simple random sampling is rarely used in practice, since it is inefficient unless the population is fairly homogeneous w.r.t. the y_i's. A sampling plan often used in practice is the *stratified sampling* plan which can be described as follows. The population $\mathcal{P}$ is divided into nonoverlapping subpopulations $\mathcal{P}_1, ..., \mathcal{P}_H$ called strata; a sample is drawn from each stratum $\mathcal{P}_h$, independently across the strata. There are many reasons for stratification: (1) it may produce a gain in precision in parameter estimation when a heterogeneous population is divided into strata, 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 stratified sampling, if a simple random sample (without replacement), $X_h = (X_{h1}, ..., X_{hn_h})$, is drawn from each stratum, where n_h is the sample size in stratum h, then the joint distribution of $X = (X_1, ..., X_H)$ is in a parametric family indexed by $\theta = (h, \theta_1, ..., \theta_H)$, where $h = 1, ..., H$ and $\theta_h = (y_i, i \in \mathcal{P}_h)$. Let $\mathcal{Y}_h$ be the range of y_i's in stratum h and $\Theta_h = \prod_{i=1}^{N_h}\mathcal{Y}_h$, where N_h is the size of $\mathcal{P}_h$. We assume that the parameter space is $\Theta = \{1, ..., H\} \times \prod_{i=1}^{H}\Theta_h$. The following result is similar to Theorem 3.13.

Theorem 3.14. Let X be a sample obtained using the stratified simple random sampling plan described previously.
(i) For each h, let Z_h be the vector of the ordered values of the sample in stratum h. Then $(Z_1, ..., Z_H)$ is sufficient and complete for $\theta \in \Theta$.
(ii) For any estimable function of θ, its unique UMVUE is the unbiased estimator $h(X)$ which is symmetric in its first n_1 arguments, symmetric in its second n_2 arguments,..., and symmetric in its last n_H arguments. ∎

Example 3.20. Consider the estimation of the population total Y based on a sample $X = (X_{hi}, i = 1, ..., n_h, h = 1, ..., H)$ obtained by stratified simple random sampling. Let Y_h be the population total of the hth stratum and let $\hat{Y}_h = N_h \bar{X}_{h\cdot}$, where $\bar{X}_{h\cdot}$ is the sample mean of the sample from stratum h, $h = 1, ..., H$. From Example 2.27, each $\hat{Y}_h$ is an unbiased estimator of Y_h. Let

$$\hat{Y}_{st} = \sum_{h=1}^{H} \hat{Y}_h = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{N_h}{n_h} X_{hi}.$$

Then, by Theorem 3.14, $\hat{Y}_{st}$ is the UMVUE of Y. Since $\hat{Y}_1, ..., \hat{Y}_H$ are independent, it follows from (3.44) that

$$\mathrm{Var}(\hat{Y}_{st}) = \sum_{h=1}^{H} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) \sigma_h^2, \tag{3.45}$$

where $\sigma_h^2 = (N_h - 1)^{-1} \sum_{i \in \mathcal{P}_h} (y_i - Y_h/N_h)^2$. A similar argument to that in Example 3.19 shows that the UMVUE of $\mathrm{Var}(\hat{Y}_{st})$ is

$$S_{st}^2 = \sum_{h=1}^{H} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_h^2, \tag{3.46}$$

where S_h^2 is the usual sample variance based on $X_{h1}, ..., X_{hn_h}$.

It is interesting to compare the mse of the UMVUE $\hat{Y}_{st}$ with the mse of the UMVUE $\hat{Y}$ under simple random sampling. Let σ^2 be given by (3.44). Then

$$(N-1)\sigma^2 = \sum_{h=1}^{H} (N_h - 1)\sigma_h^2 + \sum_{h=1}^{H} N_h (\mu_h - \mu)^2,$$

where $\mu_h = Y_h/N_h$ is the population mean of the hth stratum and $\mu = Y/N$ is the overall population mean. By (3.44), (3.45), and (3.46), $\mathrm{Var}(\hat{Y}) \geq \mathrm{Var}(\hat{Y}_{st})$ if and only if

$$\sum_{h=1}^{H} \tfrac{N^2 N_h}{n(N-1)} \left(1 - \tfrac{n}{N}\right)(\mu_h - \mu)^2 \geq \sum_{h=1}^{H} \left[\tfrac{N_h^2}{n_h} \left(1 - \tfrac{n_h}{N_h}\right) - \tfrac{N^2(N_h - 1)}{n(N-1)} \left(1 - \tfrac{n}{N}\right)\right] \sigma_h^2.$$

This means that stratified simple random sampling is better than simple random sampling if the deviations $\mu_h - \mu$ are sufficiently large. If $\frac{n_h}{N_h} \equiv \frac{n}{N}$ (proportional allocation), then this condition simplifies to

$$\sum_{h=1}^{H} N_h(\mu_h - \mu)^2 \geq \sum_{h=1}^{H} \left(1 - \frac{N_h}{N}\right) \sigma_h^2, \tag{3.47}$$

which is usually true when μ_h's are different and some N_h's are large. ∎

3.4.2 Horvitz-Thompson estimators

If some elements of the finite population $\mathcal{P}$ are groups (called clusters) of subunits, then sampling from $\mathcal{P}$ is *cluster sampling*. Cluster sampling is used often because of administrative convenience or economic considerations. 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 the maps of the region, however, it can be divided into units such as cities or blocks in the cities.

In cluster sampling, one may greatly increase the precision of estimation by using sampling with probability proportional to cluster size. Thus, unequal probability sampling is often used.

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. A sample of the subunits in any chosen cluster may be selected. This 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.

When the sampling plan is complex, so is the structure of the observations. We now introduce a general method of deriving unbiased estimators of population totals, which are called *Horvitz-Thompson estimators*.

Theorem 3.15. Let $X = \{y_i, i \in \boldsymbol{s}\}$ denote a sample from $\mathcal{P} = \{y_1, ..., y_N\}$ which is selected, without replacement, by some method. Define

$$\pi_i = \text{probability that } i \in \boldsymbol{s}, \quad i = 1, ..., N.$$

(i) (Horvitz-Thompson). If $\pi_i > 0$ for $i = 1, ..., N$ and π_i is known when $i \in \boldsymbol{s}$, then $\hat{Y}_{ht} = \sum_{i \in \boldsymbol{s}} y_i/\pi_i$ is an unbiased estimator of the population total Y.
(ii) Define

$$\pi_{ij} = \text{probability that } i \in \boldsymbol{s} \text{ and } j \in \boldsymbol{s}, \quad i = 1, ..., N,\ j = 1, ..., N.$$

Then

$$\mathrm{Var}(\hat{Y}_{ht}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 + 2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j \tag{3.48}$$

$$= \sum_{i=1}^{N} \sum_{j=i+1}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 . \tag{3.49}$$

Proof. (i) Let $a_i = 1$ if $i \in \boldsymbol{s}$ and $a_i = 0$ if $i \notin \boldsymbol{s}$, $i = 1, ..., N$. Then $E(a_i) = \pi_i$ and

$$E(\hat{Y}_{ht}) = E\left(\sum_{i=1}^{N} \frac{a_i y_i}{\pi_i} \right) = \sum_{i=1}^{N} y_i = Y.$$

(ii) Since $a_i^2 = a_i$,

$$\mathrm{Var}(a_i) = E(a_i) - [E(a_i)]^2 = \pi_i(1 - \pi_i).$$

For $i \neq j$,

$$\mathrm{Cov}(a_i, a_j) = E(a_i a_j) - E(a_i)E(a_j) = \pi_{ij} - \pi_i \pi_j.$$

Then

$$\begin{aligned}
\mathrm{Var}(\hat{Y}_{ht}) &= \mathrm{Var}\left(\sum_{i=1}^{N} \frac{a_i y_i}{\pi_i} \right) \\
&= \sum_{i=1}^{N} \frac{y_i^2}{\pi_i^2} \mathrm{Var}(a_i) + 2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{y_i y_j}{\pi_i \pi_j} \mathrm{Cov}(a_i, a_j) \\
&= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 + 2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} y_i y_j .
\end{aligned}$$

Hence (3.48) follows. To show (3.49), note that

$$\sum_{i=1}^{N} \pi_i = n \qquad \text{and} \qquad \sum_{j=1,...,N, j \neq i} \pi_{ij} = (n-1)\pi_i,$$

which implies

$$\sum_{j=1,...,N, j \neq i} (\pi_{ij} - \pi_i \pi_j) = (n-1)\pi_i - \pi_i(n - \pi_i) = -\pi_i(1 - \pi_i).$$

Hence

$$\sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} y_i^2 = \sum_{i=1}^{N} \sum_{j=1,\ldots,N, j\neq i} (\pi_i\pi_j - \pi_{ij}) \frac{y_i^2}{\pi_i^2}$$
$$= \sum_{i=1}^{N} \sum_{j=i+1}^{N} (\pi_i\pi_j - \pi_{ij}) \left(\frac{y_i^2}{\pi_i^2} + \frac{y_j^2}{\pi_j^2} \right)$$

and, by (3.48),

$$\mathrm{Var}(\hat{Y}_{ht}) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (\pi_{ij} - \pi_i\pi_j) \left(\frac{y_i^2}{\pi_i^2} + \frac{y_j^2}{\pi_j^2} - \frac{2y_iy_j}{\pi_i\pi_j} \right)$$
$$= \sum_{i=1}^{N} \sum_{j=i+1}^{N} (\pi_i\pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2. \quad \blacksquare$$

Using the same idea, we can obtain unbiased estimators of $\mathrm{Var}(\hat{Y}_{ht})$. Suppose that $\pi_{ij} > 0$ for all i and j and π_{ij} is known when $i \in \boldsymbol{s}$ and $j \in \boldsymbol{s}$. By (3.48), an unbiased estimator of $\mathrm{Var}(\hat{Y}_{ht})$ is

$$v_1 = \sum_{i\in\boldsymbol{s}} \frac{1-\pi_i}{\pi_i^2} y_i^2 + 2 \sum_{i\in\boldsymbol{s}} \sum_{j\in\boldsymbol{s}, j>i} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j\pi_{ij}} y_iy_j. \tag{3.50}$$

By (3.49), an unbiased estimator of $\mathrm{Var}(\hat{Y}_{ht})$ is

$$v_2 = \sum_{i\in\boldsymbol{s}} \sum_{j\in\boldsymbol{s}, j>i} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2. \tag{3.51}$$

Variance estimators v_1 and v_2 may not be the same in general, but they are the same in some special cases (Exercise 84). A more serious problem is that they may take negative values. Some discussions about deriving better estimators of $\mathrm{Var}(\hat{Y}_{st})$ are provided in Cochran (1977, Chapter 9A).

Some special cases of Theorem 3.15 are considered as follows.

Under simple random sampling, $\pi_i = n/N$. Thus, $\hat{Y}$ in Example 3.19 is the Horvitz-Thompson estimator.

Under stratified simple random sampling, $\pi_i = n_h/N_h$ if unit i is in stratum h. Hence, the estimator $\hat{Y}_{st}$ in Example 3.20 is the Horvitz-Thompson estimator.

Suppose now each $y_i \in \mathcal{P}$ is a cluster, i.e., $y_i = (y_{i1}, ..., y_{iM_i})$, where M_i is the size of the ith cluster, $i = 1, ..., N$. The total number of units in $\mathcal{P}$ is then $M = \sum_{i=1}^{N} M_i$. Consider a single-stage sampling plan, i.e., if y_i is selected, then every y_{ij} is observed. If simple random sampling is used,

then $\pi_i = k/N$, where k is the first-stage sample size (the total sample size is $n = \sum_{i=1}^{k} M_i$), and the Horvitz-Thompson estimator is

$$\hat{Y}_s = \frac{N}{k} \sum_{i \in \boldsymbol{s}_1} \sum_{j=1}^{M_i} y_{ij} = \frac{N}{k} \sum_{i \in \boldsymbol{s}_1} Y_i,$$

where $\boldsymbol{s}_1$ is the index set of first-stage sampled clusters and Y_i is the total of the ith cluster. In this case,

$$\mathrm{Var}(\hat{Y}_s) = \frac{N^2}{k(N-1)} \left(1 - \frac{k}{N}\right) \sum_{i=1}^{N} \left(Y_i - \frac{Y}{N}\right)^2.$$

If the selection probability is proportional to the cluster size, then $\pi_i = kM_i/M$ and the Horvitz-Thompson estimator is

$$\hat{Y}_{pps} = \frac{M}{k} \sum_{i \in \boldsymbol{s}_1} \frac{1}{M_i} \sum_{j=1}^{M_i} y_{ij} = \frac{M}{k} \sum_{i \in \boldsymbol{s}_1} \frac{Y_i}{M_i}$$

whose variance is given by (3.48) or (3.49). Usually $\mathrm{Var}(\hat{Y}_{pps})$ is smaller than $\mathrm{Var}(\hat{Y}_s)$; see the discussions in Cochran (1977, Chapter 9A).

Consider next a two-stage sampling in which k first-stage clusters are selected and a simple random sample of size m_i is selected from each sampled cluster y_i, where sampling is independent across clusters. If the first-stage sampling plan is simple random sampling, then $\pi_i = km_i/(NM_i)$ and the Horvitz-Thompson estimator is

$$\hat{Y}_s = \frac{N}{k} \sum_{i \in \boldsymbol{s}_1} \frac{M_i}{m_i} \sum_{j \in \boldsymbol{s}_{2i}} y_{ij},$$

where $\boldsymbol{s}_{2i}$ denotes the second-stage sample from cluster i. If the first-stage selection probability is proportional to the cluster size, then $\pi_i = km_i/M$ and the Horvitz-Thompson estimator is

$$\hat{Y}_{pps} = \frac{M}{k} \sum_{i \in \boldsymbol{s}_1} \frac{1}{m_i} \sum_{j \in \boldsymbol{s}_{2i}} y_{ij}.$$

Finally, let us consider another popular sampling method called *systematic sampling*. Suppose that $\mathcal{P} = \{y_1, ..., y_N\}$ and the population size $N = nk$ for two integers n and k. To select a sample of size n, we first draw a j randomly from $\{1, ..., k\}$. Our sample is then

$$\{y_j, y_{j+k}, y_{j+2k}, \cdots, y_{j+(n-1)k}\}.$$

Systematic sampling is used mainly because it is easier to draw a systematic sample and often easier to execute without mistakes. It is also likely that systematic sampling provides more efficient point estimators than simple random sampling or even stratified sampling, since the sample units are spread more evenly over the population. Under systematic sampling, $\pi_i = k^{-1}$ for every i and the Horvitz-Thompson estimator of the population total is

$$\hat{Y}_{sy} = k \sum_{t=1}^{n} y_{j+(t-1)k}.$$

The unbiasedness of this estimator is a direct consequence of Theorem 3.15, but it can be easily shown as follows. Since j takes value $i \in \{1, ..., k\}$ with probability k^{-1},

$$E(\hat{Y}_{sy}) = k \left(\frac{1}{k} \sum_{i=1}^{k} \sum_{t=1}^{n} y_{i+(t-1)k} \right) = \sum_{i=1}^{N} y_i = Y.$$

The variance of $\hat{Y}_{sy}$ is simply

$$\mathrm{Var}(\hat{Y}_{sy}) = \frac{N^2}{k} \sum_{i=1}^{k} (\mu_i - \mu)^2,$$

where $\mu_i = n^{-1} \sum_{t=1}^{n} y_{i+(t-1)k}$ and $\mu = k^{-1} \sum_{i=1}^{k} \mu_i = Y/N$. Let σ^2 be given by (3.44) and

$$\sigma_{sy}^2 = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{t=1}^{n} (y_{i+(t-1)k} - \mu_i)^2.$$

Then

$$(N-1)\sigma^2 = n \sum_{i=1}^{k} (\mu_i - \mu)^2 + \sum_{i=1}^{k} \sum_{t=1}^{n} (y_{i+(t-1)k} - \mu_i)^2.$$

Thus,

$$(N-1)\sigma^2 = N^{-1}\mathrm{Var}(\hat{Y}_{sy}) + k(n-1)\sigma_{sy}^2$$

and

$$\mathrm{Var}(\hat{Y}_{sy}) = N(N-1)\sigma^2 - N(N-k)\sigma_{sy}^2.$$

Since the variance of the Horvitz-Thompson estimator of the population total under simple random sampling is, by (3.44),

$$\frac{N^2}{n} \left(1 - \frac{n}{N}\right) \sigma^2 = N(k-1)\sigma^2,$$

the Horvitz-Thompson estimator under systematic sampling has a smaller variance if and only if $\sigma_{sy}^2 > \sigma^2$.

3.5 Asymptotically Unbiased Estimators

As we discussed in §2.5, we often need to consider biased but asymptotically unbiased estimators. A large and useful class of such estimators are smooth functions of some exactly unbiased estimators such as UMVUE's, U-statistics, and LSE's. Some other methods of constructing asymptotically unbiased estimators are also introduced in this section.

3.5.1 Functions of unbiased estimators

If the parameter to be estimated is $\vartheta = g(\theta)$ with a vector-valued parameter θ and U_n is a vector of unbiased estimators of components of θ (i.e., $EU_n = \theta$), then $T_n = g(U_n)$ is asymptotically unbiased for ϑ. Assume that g is second-order differentiable and $\|U_n - \theta\| = o_p(1)$. Then

$$\tilde{b}_{T_n}(P) = \mathrm{tr}(\nabla^2 g(\theta)\mathrm{Var}(U_n))/2$$

and

$$\mathrm{amse}_{T_n}(P) = \nabla g(\theta)\mathrm{Var}(U_n)[\nabla g(\theta)]^\tau$$

(Theorem 2.6). Hence, T_n has a good performance in terms of amse if U_n is optimal in terms of mse (such as the UMVUE).

The following are some examples.

Example 3.21 (Ratio estimators). Let $(X_1, Y_1), ..., (X_n, Y_n)$ be i.i.d. random 2-vectors. Consider the estimation of the ratio of two population means: $\vartheta = \mu_y/\mu_x$ ($\mu_x \neq 0$). Note that $(\bar{Y}, \bar{X})$, the vector of sample means, is unbiased for (μ_y, μ_x). The sample means are UMVUE's under some statistical models (§3.1 and §3.2) and are BLUE's in general (Example 2.22). The *ratio* estimator is $T_n = \bar{Y}/\bar{X}$. Assume that $\sigma_x^2 = \mathrm{Var}(X_1)$, $\sigma_y^2 = \mathrm{Var}(Y_1)$, and $\sigma_{xy} = \mathrm{Cov}(X_1, Y_1)$ exist. A direct calculation shows that

$$\tilde{b}_{T_n}(P) = \frac{\vartheta\sigma_x^2 - \sigma_{xy}}{\mu_x^2 n}, \tag{3.52}$$

and

$$\sqrt{n}(T_n - \vartheta) \to_d N\left(0, \frac{\sigma_y^2 - 2\vartheta\sigma_{xy} + \vartheta^2\sigma_x^2}{\mu_x^2}\right), \tag{3.53}$$

which implies

$$\underline{\mathrm{amse}}_{T_n}(P) = \frac{\sigma_y^2 - 2\vartheta\sigma_{xy} + \vartheta^2\sigma_x^2}{\mu_x^2 n}. \tag{3.54}$$

Results (3.52) and (3.54) still hold when $(X_1, Y_1), ..., (X_n, Y_n)$ is a sample from a finite bivariate population of size N (exercise). In some problems we

are not interested in the ratio, but the use of a ratio estimator to improve an estimator of a marginal mean. For example, suppose that μ_x is known and we are interested in estimating μ_y. Consider the following estimator

$$\hat{\mu}_y = (\bar{Y}/\bar{X})\mu_x.$$

Note that $\hat{\mu}_y$ is not unbiased; its n^{-1} order asymptotic bias is

$$\tilde{b}_{\hat{\mu}_y}(P) = \frac{\vartheta\sigma_x^2 - \sigma_{xy}}{\mu_x n};$$

and

$$\underline{\text{amse}}_{\hat{\mu}_y}(P) = \frac{\sigma_y^2 - 2\vartheta\sigma_{xy} + \vartheta^2\sigma_x^2}{n}.$$

Comparing $\hat{\mu}_y$ with the unbiased estimator $\bar{Y}$, we find that $\hat{\mu}_y$ is asymptotically more efficient if and only if

$$2\vartheta\sigma_{xy} > \vartheta^2\sigma_x^2,$$

which means that $\hat{\mu}_y$ is a better estimator if and only if the correlation between X_1 and Y_1 is large enough to pay off the extra variability caused by using $\mu_x/\bar{X}$. ∎

Another example related to a bivariate sample is the sample correlation coefficient defined in Exercise 19 in §2.6.

Example 3.22. Consider a polynomial regression of order p:

$$X_i = \beta Z_i^\tau + \varepsilon_i, \qquad i = 1, ..., n,$$

where $\beta = (\beta_0, \beta_1, ..., \beta_{p-1})$, $Z_i = (1, t_i, ..., t_i^{p-1})$, and ε_i's are i.i.d. with mean 0 and variance $\sigma^2 > 0$. Suppose that the parameter to be estimated is $t_\beta \in \mathcal{R}$ such that

$$\sum_{j=0}^{p-1} \beta_j t_\beta^j = \max_{t\in\mathcal{R}} \sum_{j=0}^{p-1} \beta_j t^j.$$

Note that $t_\beta = g(\beta)$ for some function g. Let $\hat{\beta}$ be the LSE of β. Then the estimator $\hat{t}_\beta = g(\hat{\beta})$ is asymptotically unbiased and its amse can be derived under some conditions (exercise).

Example 3.23. In the study of the reliability of a system component, we assume that

$$X_{ij} = z(t_j)\boldsymbol{\theta}_i^\tau + \varepsilon_{ij}, \quad i = 1, ..., k,\ j = 1, ..., m.$$

Here X_{ij} is the measurement of the ith sample component at time t_j; $z(t)$ is a q-vector whose components are known functions of the time t; $\boldsymbol{\theta}_i$'s are unobservable random q-vectors that are i.i.d. from $N_q(\theta, \Sigma)$, where θ and Σ are unknown; ε_{ij}'s are i.i.d. measurement errors with mean zero and variance σ^2; and $\boldsymbol{\theta}_i$'s and ε_{ij}'s are independent. As a function of t, $z(t)\boldsymbol{\theta}^\tau$ is the degradation curve for a particular component and $z(t)\theta^\tau$ is the mean degradation curve. Suppose that a component will fail to work if $z(t)\boldsymbol{\theta}^\tau < \eta$, a given critical value. Assume that $z(t)'\boldsymbol{\theta}$ is always a decreasing function of t. Then the reliability function of a component is

$$R(t) = P(z(t)\boldsymbol{\theta}^\tau > \eta) = \Phi\left(\frac{z(t)\theta^\tau - \eta}{s(t)}\right),$$

where $s(t) = \sqrt{z(t)\Sigma[z(t)]^\tau}$ and Φ is the standard normal distribution function. For a fixed t, estimators of $R(t)$ can be obtained by estimating θ and Σ, since Φ is a known function. It can be shown (exercise) that the BLUE of θ is the LSE

$$\hat{\theta} = \bar{X}Z(Z^\tau Z)^{-1},$$

where $Z = ([z(t_1)]^\tau, ..., [z(t_k)]^\tau)^\tau$, $X_i = (X_{i1}, ..., X_{im})$, and $\bar{X}$ is the sample mean of X_i's. The estimation of Σ is more difficult. An asymptotically unbiased (as $k \to \infty$) estimator of Σ is

$$\hat{\Sigma} = \frac{1}{k}\sum_{i=1}^{k}(Z^\tau Z)^{-1}Z^\tau(X_i - \bar{X})^\tau(X_i - \bar{X})Z(Z^\tau Z)^{-1} - \hat{\sigma}^2(Z^\tau Z)^{-1},$$

where

$$\hat{\sigma}^2 = \frac{1}{k(m-q)}\sum_{i=1}^{k}[X_iX_i^\tau - X_iZ(Z^\tau Z)^{-1}Z^\tau X_i^\tau].$$

Hence an estimator of $R(t)$ is

$$\hat{R}(t) = \Phi\left(\frac{z(t)\hat{\theta}^\tau - \eta}{\hat{s}(t)}\right),$$

where

$$\hat{s}(t) = \left\{z(t)\hat{\Sigma}[z(t)]^\tau\right\}^{1/2}.$$

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

$$g(y_1, y_2, y_3) = \Phi\left(\frac{y_1 - \eta}{\sqrt{y_2 - y_1^2 - y_3z(t)(Z^\tau Z)^{-1}[z(t)]^\tau}}\right).$$

Suppose that ε_{ij} has a finite fourth moment, which implies the existence of $\mathrm{Var}(Y_i)$. The amse of $\hat{R}(t)$ can be derived (exercise). ∎

3.5.2 The method of moments

The method of moments is the oldest method of deriving point estimators. It almost always produces some asymptotically unbiased estimators, although they may not be the best estimators.

Consider a parametric problem where $X_1, ..., X_n$ are i.i.d. random variables from P_θ, $\theta \in \Theta \subset \mathcal{R}^k$, and $E|X_1|^k < \infty$. Let $\mu_j = EX_1^j$ be the jth moment of P and let

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^n X_i^j$$

be the jth *sample moment*, which is an unbiased estimator of μ_j, $j = 1, ..., k$. Typically,

$$\mu_j = h_j(\theta), \qquad j = 1, ..., k, \tag{3.55}$$

for some functions h_j on $\mathcal{R}^k$. By substituting μ_j's on the left-hand side of (3.55) by the sample moments $\hat{\mu}_j$, we obtain a *moment estimator* $\hat{\theta}$, i.e., $\hat{\theta}$ satisfies

$$\hat{\mu}_j = h_j(\hat{\theta}), \qquad j = 1, ..., k,$$

which is a sample analogue of (3.55). This method of deriving estimators is called the *method of moments*. Note that an important statistical principle, the *substitution principle*, is applied in this method.

Let $\hat{\mu} = (\hat{\mu}_1, ..., \hat{\mu}_k)$ and $h = (h_1, ..., h_k)$. Then $\hat{\mu} = h(\hat{\theta})$. If h^{-1} exists, then the unique moment estimator of θ is $\hat{\theta} = h^{-1}(\hat{\mu})$. When h^{-1} does not exist (i.e., h is not one-to-one), any solution of $\hat{\mu} = h(\hat{\theta})$, denoted by $\hat{\theta} = g(\hat{\mu})$, is a moment estimator of θ.

By the SLLN, $\hat{\mu}_j \to_{a.s.} \mu_j$. Assume that h is one-to-one and let $g = h^{-1}$. Typically, the function g in $\hat{\theta} = g(\hat{\mu})$ is continuous and, therefore, $\hat{\theta}$ is strongly consistent for θ. If g is differentiable and $E|X_1|^{2k} < \infty$, then $\hat{\theta}$ is asymptotically normal, by the CLT and Theorem 2.11, and

$$\underline{\text{amse}}_{\hat{\theta}}(\theta) = n^{-1}\nabla g(\mu) V_\mu [\nabla g(\mu)]^\tau, \tag{3.56}$$

where $\mu = (\mu_1, ..., \mu_k)$ and V_μ is a $k \times k$ matrix whose (i, j)th element is $\mu_{i+j} - \mu_i\mu_j$.

Example 3.24. Let $X_1, ..., X_n$ be i.i.d. from a population P_θ indexed by the parameter $\theta = (\mu, \sigma^2)$, where $\mu = EX_1 \in \mathcal{R}$ and $\sigma^2 = \text{Var}(X_1) \in (0, \infty)$. This includes cases like the family of normal distributions, double exponential distributions, or logistic distributions (Table 1.2, page 20). Since $EX_1 = \mu$ and $EX_1^2 = \text{Var}(X_1) + (EX_1)^2 = \sigma^2 + \mu^2$, setting $\hat{\mu}_1 = \mu$ and $\hat{\mu}_2 = \sigma^2 + \mu^2$ we obtain the moment estimators

$$\hat{\theta} = \left(\bar{X},\ \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right) = \left(\bar{X},\ \frac{n-1}{n}S^2\right).$$

Note that $\bar{X}$ is unbiased, but $\frac{n-1}{n}S^2$ is not. If X_i is normal, then $\hat{\theta}$ is sufficient and is nearly the same as an optimal estimator such as the UMVUE. On the other hand, if X_i is from a double exponential or logistic distribution, then $\hat{\theta}$ is not sufficient and can often be improved.

Consider now the estimation of σ^2 when we know that $\mu = 0$. Obviously we cannot use the equation $\hat{\mu}_1 = \mu$ to solve the problem. Using $\hat{\mu}_2 = \mu_2 = \sigma^2$, we obtain the moment estimator $\hat{\sigma}^2 = \hat{\mu}_2 = n^{-1}\sum_{i=1}^n X_i^2$. This is still a good estimator when X_i is normal, but is not a function of sufficient statistic when X_i is from a double exponential distribution. For the double exponential case one can argue that we should first make a transformation $Y_i = |X_i|$ and then obtain the moment estimator based on the transformed data. The moment estimator of σ^2 based on the transformed data is $\bar{Y} = n^{-1}\sum_{i=1}^n |X_i|$, which is sufficient for σ^2. Note that this estimator can also be obtained based on absolute moment equations. ∎

Example 3.25. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on (θ_1, θ_2), $-\infty < \theta_1 < \theta_2 < \infty$. Note that

$$EX_1 = (\theta_1 + \theta_2)/2$$

and

$$EX_1^2 = (\theta_1^2 + \theta_2^2 + \theta_1\theta_2)/3.$$

Setting $\hat{\mu}_1 = EX_1$ and $\hat{\mu}_2 = EX_1^2$ and substituting θ_1 in the second equation by $2\hat{\mu}_1 - \theta_2$ (the first equation), we obtain that

$$(2\hat{\mu}_1 - \theta_2)^2 + \theta_2^2 + (2\hat{\mu}_1 - \theta_2)\theta_2 = 3\hat{\mu}_2,$$

which is the same as

$$(\theta_2 - \hat{\mu}_1)^2 = 3(\hat{\mu}_2 - \hat{\mu}_1^2).$$

Since $\theta_2 \geq \hat{\mu}_1$, we obtain that

$$\hat{\theta}_2 = \hat{\mu}_1 + \sqrt{3(\hat{\mu}_2 - \hat{\mu}_1^2)} = \bar{X} + \sqrt{\tfrac{3(n-1)}{n}S^2}$$

and

$$\hat{\theta}_1 = \hat{\mu}_1 - \sqrt{3(\hat{\mu}_2 - \hat{\mu}_1^2)} = \bar{X} - \sqrt{\tfrac{3(n-1)}{n}S^2}.$$

These estimators are not functions of the sufficient and complete statistic $(X_{(1)}, X_{(n)})$. ∎

Example 3.26. Let $X_1, ..., X_n$ be i.i.d. from the binomial distribution $Bi(p, k)$ with unknown parameters $k \in \{1, 2, ...\}$ and $p \in (0, 1)$. Since

$$EX_1 = kp$$

and

$$EX_1^2 = kp(1-p) + k^2p^2,$$

we obtain the moment estimators

$$\hat{p} = (\hat{\mu}_1 + \hat{\mu}_1^2 - \hat{\mu}_2)/\hat{\mu}_1 = 1 - \tfrac{n-1}{n} S^2/\bar{X}$$

and

$$\hat{k} = \hat{\mu}_1^2/(\hat{\mu}_1 + \hat{\mu}_1^2 - \hat{\mu}_2) = \bar{X}/(1 - \tfrac{n-1}{n} S^2/\bar{X}).$$

The estimator $\hat{p}$ is in the range of $(0,1)$. But $\hat{k}$ may not be an integer. It can be improved by an estimator which is $\hat{k}$ rounded to the nearest positive integer. ∎

Example 3.27. Suppose that $X_1, ..., X_n$ are i.i.d. from the Pareto distribution $Pa(a, \theta)$ with unknown $a > 0$ and $\theta > 2$ (Table 1.2, page 20). Note that

$$EX_1 = \theta a/(\theta - 1)$$

and

$$EX_1^2 = \theta a^2/(\theta - 2).$$

From the moment equation,

$$\tfrac{(\theta-1)^2}{\theta(\theta-2)} = \hat{\mu}_2/\hat{\mu}_1^2.$$

Note that $\frac{(\theta-1)^2}{\theta(\theta-2)} - 1 = \frac{1}{\theta(\theta-2)}$. Hence

$$\theta(\theta - 2) = \hat{\mu}_1^2/(\hat{\mu}_2 - \hat{\mu}_1^2).$$

Since $\theta > 2$, there is a unique solution

$$\hat{\theta} = 1 + \sqrt{\hat{\mu}_2/(\hat{\mu}_2 - \hat{\mu}_1^2)} = 1 + \sqrt{1 + \tfrac{n}{n-1}\bar{X}^2/S^2}$$

and

$$\begin{aligned}\hat{a} &= \frac{\hat{\mu}_1(\hat{\theta} - 1)}{\hat{\theta}} \\ &= \bar{X}\sqrt{1 + \tfrac{n}{n-1}\bar{X}^2/S^2} \Big/ \left(1 + \sqrt{1 + \tfrac{n}{n-1}\bar{X}^2/S^2}\right). \quad ∎\end{aligned}$$

The method of moments can also be applied to nonparametric problems. Consider, for example, the estimation of the central moments

$$c_j = E(X_1 - \mu)^j, \qquad j = 2, ..., k.$$

Since

$$c_j = \sum_{t=0}^{j} \binom{j}{t} (-\mu)^t \mu_{j-t},$$

the moment estimator of c_j is

$$\hat{c}_j = \sum_{t=0}^{j} \binom{j}{t} (-\bar{X})^t \hat{\mu}_{j-t},$$

where $\hat{\mu}_0 = 1$. It can be shown (exercise) that

$$\hat{c}_j = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^j, \qquad j = 2, ..., k, \tag{3.57}$$

which are sample central moments. From the SLLN, $\hat{c}_j$'s are strongly consistent. If $E|X_1|^{2k} < \infty$, then

$$\sqrt{n}\,(\hat{c}_2 - c_2, ..., \hat{c}_k - c_k) \to_d N_{k-1}(0, D) \tag{3.58}$$

(exercise), where the (i, j)th element of the $(k-1) \times (k-1)$ matrix D is

$$c_{i+j+2} - c_{i+1}c_{j+1} - (i+1)c_i c_{j+2} - (j+1)c_{i+2}c_j + (i+1)(j+1)c_i c_j c_2.$$

3.5.3 V-statistics

Let $X_1, ..., X_n$ be i.i.d. from P. For every U-statistic defined in (3.11) as an estimator of $\vartheta = E[h(X_1, ..., X_m)]$, there is a closely related *V-statistic* defined by

$$V_n = \frac{1}{n^m} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} h(X_{i_1}, ..., X_{i_m}). \tag{3.59}$$

As an estimator of ϑ, V_n is biased; but the bias is small asymptotically as the following results show. For a fixed sample size n, V_n may be better than U_n in terms of their mse's. Consider, for example, the kernel $h(x_1, x_2) = (x_1 - x_2)^2/2$ in §3.2.1, which leads to $\vartheta = \sigma^2 = \mathrm{Var}(X_1)$ and $U_n = S^2$, the sample variance. The corresponding V-statistic is

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{(X_i - X_j)^2}{2} = \frac{1}{n^2} \sum_{1 \le i < j \le n} (X_i - X_j)^2 = \frac{n-1}{n} S^2,$$

which is the moment estimator of σ^2 discussed in Example 3.24. In Exercise 52 in §2.6, $\frac{n-1}{n} S^2$ is shown to have a smaller mse than S^2 in some cases. Of course, there are situations where U-statistics are better than their corresponding V-statistics.

The following result provides orders of magnitude of the bias and variance of a V-statistic as an estimator of ϑ.

Proposition 3.5. Let V_n be defined by (3.59).
(i) Assume that $E|h(X_{i_1}, ..., X_{i_m})| < \infty$ for all $1 \le i_1 \le \cdots \le i_m \le m$. Then the bias of V_n satisfies

$$b_{V_n}(P) = O(n^{-1}).$$

(ii) Assume that $E[h(X_{i_1}, ..., X_{i_m})]^2 < \infty$ for all $1 \le i_1 \le \cdots \le i_m \le m$. Then the variance of V_n satisfies

$$\mathrm{Var}(V_n) = \mathrm{Var}(U_n) + O(n^{-2}),$$

where U_n is given by (3.11).
Proof. (i) Note that

$$U_n - V_n = \left[1 - \tfrac{n!}{n^m(n-m)!}\right](U_n - W_n), \tag{3.60}$$

where W_n is the average of all terms $h(X_{i_1}, ..., X_{i_m})$ with at least one equality $i_m = i_l$, $m \neq l$. The result follows from $E(U_n - W_n) = O(1)$.
(ii) The result follows from $E(U_n - W_n)^2 = O(1)$, $E[W_n(U_n - \vartheta)] = O(n^{-1})$ (exercise), and (3.60). ∎

To study the asymptotic behavior of a V-statistic, we consider the following representation of V_n in (3.59):

$$V_n = \sum_{j=1}^{m} \binom{m}{j} V_{nj},$$

where

$$V_{nj} = \vartheta + \frac{1}{n^j} \sum_{i_1=1}^{n} \cdots \sum_{i_j=1}^{n} g_j(X_{i_1}, ..., X_{i_j})$$

is a "V-statistic" with

$$\begin{aligned} g_j(x_1, ..., x_j) = h_j(x_1, ..., x_j) &- \sum_{i=1}^{j} \int h_j(x_1, ..., x_j) dP(x_i) \\ &+ \sum_{1 \le i_1 < i_2 \le j} \int \int h_j(x_1, ..., x_j) dP(x_{i_1}) dP(x_{i_2}) - \cdots \\ &+ (-1)^j \int \cdots \int h_j(x_1, ..., x_j) dP(x_1) \cdots dP(x_j) \end{aligned}$$

and $h_j(x_1, ..., x_j) = E[h(x_1, ..., x_j, X_{j+1}, ..., X_m)]$. Using a similar argument to the proof of Theorem 3.4, we can show (exercise) that

$$E(V_{nj})^2 = O(n^{-j}), \qquad j = 1, ..., m, \tag{3.61}$$

provided that $E[h(X_{i_1},...,X_{i_m})]^2 < \infty$ for all $1 \le i_1 \le \cdots \le i_m \le m$. Thus,

$$V_n - \vartheta = mV_{n1} + \tfrac{m(m-1)}{2}V_{n2} + o_p(n^{-1}),$$

which leads to the following result similar to Theorem 3.5.

Theorem 3.16. Let V_n be given by (3.59) with $E[h(X_{i_1},...,X_{i_m})]^2 < \infty$ for all $1 \le i_1 \le \cdots \le i_m \le m$.
(i) If $\zeta_1 = \mathrm{Var}(h_1(X_1)) > 0$, then

$$\sqrt{n}(V_n - \vartheta) \to_d N(0, m^2\zeta_1).$$

(ii) If $\zeta_1 = 0$ but $\zeta_2 = \mathrm{Var}(h_2(X_1, X_2)) > 0$, then

$$n(V_n - \vartheta) \to_d \frac{m(m-1)}{2}\sum_{j=1}^{\infty}\lambda_j\chi_{1j}^2,$$

where χ_{1j}^2's and λ_j's are the same as those in (3.21). ∎

Theorem 3.16 indicates that if $\zeta_1 > 0$, then the asymptotic biases and amse's of U_n and V_n are the same. If $\zeta_1 = 0$ but $\zeta_2 > 0$, then a similar argument to that in the proof of Lemma 3.2 leads to

$$\begin{aligned}\underline{\mathrm{amse}}_{V_n}(P) &= \frac{m^2(m-1)^2\zeta_2}{2n^2} + \frac{m^2(m-1)^2}{2n^2}\left(\sum_{j=1}^{\infty}\lambda_j\right)^2 \\ &= \underline{\mathrm{amse}}_{U_n}(P) + \frac{m^2(m-1)^2}{2n^2}\left(\sum_{j=1}^{\infty}\lambda_j\right)^2\end{aligned}$$

(see Lemma 3.2). Hence U_n is asymptotically more efficient than V_n, unless $\sum_{j=1}^{\infty}\lambda_j = 0$. Technically, the proof of the asymptotic results for V_n also requires moment conditions stronger than those for U_n.

Example 3.28. Consider the estimation of μ^2, where $\mu = EX_1$. From the results in §3.2, the U-statistic $U_n = \frac{1}{n(n-1)}\sum_{1\le i<j\le n}X_iX_j$ is unbiased for μ^2. The corresponding V-statistic is simply $V_n = \bar{X}^2$. If $\mu \ne 0$, then $\zeta_1 \ne 0$ and the asymptotic relative efficiency of V_n w.r.t. U_n is 1. If $\mu = 0$, then

$$nV_n \to_d \sigma^2\chi_1^2 \qquad \text{and} \qquad nU_n \to_d \sigma^2(\chi_1^2 - 1),$$

where χ_1^2 is a random variable having the chi-square distribution χ_1^2. Hence the asymptotic relative efficiency of V_n w.r.t. U_n is

$$E(\chi_1^2 - 1)^2/E(\chi_1^2)^2 = 2/3. \quad ∎$$

3.5.4 The weighted LSE

In linear model (3.25), the unbiased LSE of βl^τ may be improved by a slightly biased estimator when $\mathrm{Var}(\varepsilon)$ is not $\sigma^2 I_n$ and the LSE is not BLUE.

Assume that Z in (3.25) is of full rank so that every βl^τ is estimable. For simplicity, let us denote $\mathrm{Var}(\varepsilon)$ by V. If V is known, then the BLUE of βl^τ is $\breve{\beta} l^\tau$, where

$$\breve{\beta} = XV^{-1}Z(Z^\tau V^{-1}Z)^{-1} \tag{3.62}$$

(see the discussion after the statement of assumption A3 in §3.3.1). If V is unknown and $\hat{V}$ is an estimator of V, then an application of the substitution principle leads to a *weighted least squares estimator*

$$\hat{\beta}_w = X\hat{V}^{-1}Z(Z^\tau \hat{V}^{-1}Z)^{-1}. \tag{3.63}$$

The weighted LSE is not linear in X and not necessarily unbiased for β. It is unbiased if $-\varepsilon$ and ε have the same distribution, $E[\lambda_+(\hat{V})]^2 < \infty$, and $\hat{V} = u(\varepsilon)$ for some function u satisfying $u(-\varepsilon) = u(\varepsilon)$. In such a case the LSE $\hat{\beta} l^\tau$ may not be a UMVUE, since $\hat{\beta}_w l^\tau$ may be better than $\hat{\beta} l^\tau$.

Asymptotic properties of the weighted LSE depend on the asymptotic behavior of $\hat{V}$. We say that $\hat{V}$ is consistent for V if and only if

$$\|\hat{V}^{-1}V - I_n\| \to_p 0, \tag{3.64}$$

where $\|A\| = [\mathrm{tr}(A^\tau A)]^{1/2}$ for a matrix A.

Theorem 3.17. Consider model (3.25) with a full rank Z. Let $\breve{\beta}$ and $\hat{\beta}_w$ be defined by (3.62) and (3.63), respectively, with a $\hat{V}$ consistent in the sense of (3.64). Assume the conditions in Theorem 3.12. Then

$$(\hat{\beta}_w l^\tau - \beta l^\tau)/a_n \to_d N(0,1),$$

where $l \in \mathcal{R}^p$, $l \neq 0$, and

$$a_n^2 = \mathrm{Var}(\breve{\beta} l^\tau) = l(Z^\tau V^{-1}Z)^{-1} l^\tau.$$

Proof. Using the same argument as in the proof of Theorem 3.12, we obtain that

$$(\breve{\beta} l^\tau - \beta l^\tau)/a_n \to_d N(0,1).$$

By Slutsky's theorem, the result follows from

$$\hat{\beta}_w l^\tau - \breve{\beta} l^\tau = o_p(a_n).$$

Note that

$$\begin{aligned}
\hat{\beta}_w l^\tau - \breve{\beta} l^\tau &= \varepsilon \hat{V}^{-1}Z(Z^\tau \hat{V}^{-1}Z)^{-1} l^\tau - \varepsilon V^{-1}Z(Z^\tau V^{-1}Z)^{-1} l^\tau \\
&= \varepsilon(\hat{V}^{-1} - V^{-1})Z(Z^\tau \hat{V}^{-1}Z)^{-1} l^\tau \qquad (3.65)\\
&\quad + \varepsilon V^{-1}Z[(Z^\tau \hat{V}^{-1}Z)^{-1} - (Z^\tau V^{-1}Z)^{-1}] l^\tau. \qquad (3.66)
\end{aligned}$$

Let ξ_n be the term in (3.65) and $A_n = V\hat{V}^{-1} - I_n$. Using inequality (1.34), we obtain that

$$\begin{aligned}\xi_n^2 &= [\varepsilon V^{-1/2}V^{-1/2}A_n Z(Z^\tau \hat{V}^{-1}Z)^{-1}l^\tau]^2\\ &\le \varepsilon V^{-1}\varepsilon^\tau l(Z^\tau\hat{V}^{-1}Z)^{-1}Z^\tau A_n^\tau V^{-1}A_n Z(Z^\tau\hat{V}^{-1}Z)^{-1}l^\tau\\ &\le O_p(1)o_p(a_n^2),\end{aligned}$$

since $\|A_n\| = o_p(1)$ by condition (3.64). This proves that $\xi_n = o_p(a_n)$.

Let ζ_n be the term in (3.66), $B_n = Z^\tau V^{-1}Z(Z^\tau\hat{V}^{-1}Z)^{-1} - I_p$, and $C_n = \hat{V}V^{-1} - I_n$. By (3.64), $\|C_n\| = o_p(1)$. Then

$$\begin{aligned}\|B_n\|^2 &= \|Z^\tau\hat{V}^{-1}C_n Z(Z^\tau\hat{V}^{-1}Z)^{-1}\|^2\\ &= \mathrm{tr}\left((Z^\tau\hat{V}^{-1}Z)^{-1}Z^\tau C_n^\tau\hat{V}^{-1}ZZ^\tau\hat{V}^{-1}C_n Z(Z^\tau\hat{V}^{-1}Z)^{-1}\right)\\ &\le \|C_n\|\,\mathrm{tr}\left(\hat{V}^{-1}ZZ^\tau\hat{V}^{-1}C_n Z(Z^\tau\hat{V}^{-1}Z)^{-1}(Z^\tau\hat{V}^{-1}Z)^{-1}Z^\tau\right)\\ &\le \|C_n\|^2\,\mathrm{tr}\left(Z(Z^\tau\hat{V}^{-1}Z)^{-1}(Z^\tau\hat{V}^{-1}Z)^{-1}Z^\tau\hat{V}^{-1}ZZ^\tau\hat{V}^{-1}\right)\\ &= o_p(1)\mathrm{tr}(I_p)\\ &= o_p(1).\end{aligned}$$

Note that (exercise)

$$\varepsilon V^{-1}Z(Z^\tau V^{-1}Z)^{-1}l^\tau = O_p(a_n). \tag{3.67}$$

Then

$$\begin{aligned}\zeta_n^2 &= [\varepsilon V^{-1}Z(Z^\tau V^{-1}Z)^{-1}B_n l^\tau]^2\\ &\le \|\varepsilon V^{-1}Z(Z^\tau V^{-1}Z)^{-1}\|^2\|B_n l^\tau\|^2\\ &= O_p(a_n^2)o_p(1).\end{aligned}$$

This shows that $\zeta_n = o_p(a_n)$ and thus completes the proof. ∎

Theorem 3.17 shows that as long as $\hat{V}$ is consistent in the sense of (3.64), the weighted LSE $\hat{\beta}_w$ is asymptotically as efficient as $\breve{\beta}$, which is the BLUE if V is known. If V is known and ε is normal, then $\mathrm{Var}(\breve{\beta}l^\tau)$ attains the Cramér-Rao lower bound (Theorem 3.7(iii)) and, thus, (3.10) holds with $T_n = \hat{\beta}_w l^\tau$.

By Theorems 3.12 and 3.17, the asymptotic relative efficiency of the LSE $\hat{\beta}l^\tau$ w.r.t. the weighted LSE $\hat{\beta}_w l^\tau$ is

$$\frac{l(Z^\tau V^{-1}Z)^{-1}l^\tau}{l(Z^\tau Z)^{-1}Z^\tau VZ(Z^\tau Z)^{-1}l^\tau},$$

which is always less than 1 and equals 1 if $\hat{\beta}l^\tau$ is a BLUE (in which case $\hat{\beta} = \breve{\beta}$).

Finding a consistent $\hat{V}$ is possible only when V has certain structure. We consider two examples.

Example 3.29. Suppose that V is a block diagonal matrix with the ith diagonal block

$$\sigma^2 I_{m_i} + U_i \Sigma U_i^\tau, \qquad i = 1, ..., k, \tag{3.68}$$

where m_i's are integers bounded by a fixed integer m, $\sigma^2 > 0$ is an unknown parameter, Σ is a $q \times q$ unknown nonnegative definite matrix, U_i is an $m_i \times q$ full rank matrix whose columns are in $\mathcal{R}(W_i^\tau)$, $q < \inf_i m_i$, and W_i is the ith block of $Z = (W_1^\tau, ..., W_k^\tau)^\tau$. Under (3.68), a consistent $\hat{V}$ can be obtained if we can obtain consistent estimators of σ^2 and Σ.

Let $X = (Y_1, ..., Y_k)$, where Y_i is $m_i \times 1$, and let R_i be the matrix containing linearly independent columns of W_i. Then

$$\hat{\sigma}^2 = \frac{1}{n - kq} \sum_{i=1}^{k} Y_i [I_{m_i} - R_i (R_i^\tau R_i)^{-1} R_i^\tau] Y_i^\tau \tag{3.69}$$

is an unbiased estimator of σ^2. Assume that Y_i's are independent and that $\sup_i E|\varepsilon_i|^{2+\delta} < \infty$ for some $\delta > 0$. Then $\hat{\sigma}^2$ is consistent for σ^2 (exercise). A consistent estimator of Σ is then (exercise)

$$\hat{\Sigma} = \frac{1}{k} \sum_{i=1}^{k} \left[(U_i^\tau U_i)^{-1} U_i^\tau r_i^\tau r_i U_i (U_i^\tau U_i)^{-1} - \hat{\sigma}^2 (U_i^\tau U_i)^{-1} \right], \tag{3.70}$$

where $r_i = Y_i - \hat{\beta} W_i^\tau$. ∎

Example 3.30. Suppose that V is diagonal with the ith diagonal element $\sigma_i^2 = \psi(Z_i)$, where ψ is an unknown function. The simplest case is $\psi(t) = \theta_0 + \theta_1 v(Z_i)$ for a known function v and some unknown θ_0 and θ_1. One can then obtain a consistent estimator $\hat{V}$ by using the LSE of θ_0 and θ_1 under the "model"

$$r_i^2 = \theta_0 + \theta_1 v(Z_i), \qquad i = 1, ..., n, \tag{3.71}$$

where $r_i = X_i - \hat{\beta} Z_i^\tau$ (exercise). If ψ is nonlinear or nonparametric, some results are given in Carroll (1982) and Müller and Stadrmüller (1987). ∎

Finally, if $\hat{V}$ is not consistent (i.e., (3.64) does not hold), then the weighted LSE $\hat{\beta}_w l^\tau$ can still be consistent and asymptotically normal, but its asymptotic variance is not $l(Z^\tau V^{-1} Z)^{-1} l^\tau$; in fact, $\hat{\beta}_w l^\tau$ may not be asymptotically as efficient as the LSE $\hat{\beta} l^\tau$ (Carroll and Cline, 1988; Chen and Shao 1993).

3.6 Exercises

1. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $P(X_i = 1) = p \in (0, 1)$.
(a) Find the UMVUE of p^m, $m \leq n$.
(b) Find the UMVUE of $P(X_1 + \cdots + X_m = k)$, where m and k are positive integers $\leq n$.
(c) Find the UMVUE of $P(X_1 + \cdots + X_{n-1} > X_n)$.

2. Let $X_1, ..., X_n$ be i.i.d. having the Poisson distribution $P(\theta)$ with $\theta > 0$. Find the UMVUE of $e^{-t\theta}$ with a fixed $t > 0$.

3. Let $X_1, ..., X_n$ be i.i.d. having the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathcal{R}$ and a known $\sigma^2 > 0$.
(a) Find the UMVUE's of μ^3 and μ^4.
(b) Find the UMVUE's of $P(X_1 \leq t)$ and $\frac{d}{dt}P(X_1 \leq t)$ with a fixed $t \in \mathcal{R}$.

4. In Example 3.4,
(a) show that the UMVUE of σ^r is $k_{n-1,r}S^r$, where $r > 1 - n$;
(b) prove that $(X_1 - \bar{X})/S$ has the p.d.f. given by (3.1);
(c) show that $(X_1 - \bar{X})/S \to_d N(0, 1)$ by using (i) the SLLN and (ii) Scheffé's theorem (Proposition 1.17).

5. Let $X_1, ..., X_m$ be i.i.d. having the $N(\mu_x, \sigma_x^2)$ distribution and let $Y_1, ..., Y_n$ be i.i.d. having the $N(\mu_y, \sigma_y^2)$ distribution. Assume that X_i's and Y_j's are independent.
(a) Assume that $\mu_x \in \mathcal{R}$, $\mu_y \in \mathcal{R}$, $\sigma_x^2 > 0$, and $\sigma_y^2 > 0$. Find the UMVUE's of $\mu_x - \mu_y$ and $(\sigma_x/\sigma_y)^r$, $r > 0$.
(b) Assume that $\mu_x \in \mathcal{R}$, $\mu_y \in \mathcal{R}$, and $\sigma_x^2 = \sigma_y^2 > 0$. Find the UMVUE's of σ_x^2 and $(\mu_x - \mu_y)/\sigma_x$.
(c) Assume that $\mu_x = \mu_y \in \mathcal{R}$, $\sigma_x^2 > 0$, $\sigma_y^2 > 0$, and $\sigma_x^2/\sigma_y^2 = \gamma$ is known. Find the UMVUE of μ_x.
(d) Assume that $\mu_x = \mu_y \in \mathcal{R}$, $\sigma_x^2 > 0$, and $\sigma_y^2 > 0$. Show that a UMVUE of μ_x does not exist.
(e) Assume that $\mu_x = \mu_y \in \mathcal{R}$, $\sigma_x^2 > 0$, and $\sigma_y^2 > 0$. Find the UMVUE of $P(X_1 \leq Y_1)$.
(f) Repeat (e) under the assumption that $\sigma_x = \sigma_y$.

6. Let $X_1, ..., X_n$ be i.i.d. having the uniform distribution on the interval $(\theta_1 - \theta_2, \theta_1 + \theta_2)$, where $\theta_j \in \mathcal{R}$, $j = 1, 2$. Find the UMVUE's of θ_j, $j = 1, 2$, and θ_1/θ_2.

7. Let $X_1, ..., X_n$ be i.i.d. having the exponential distribution $E(a, \theta)$ with parameters $\theta > 0$ and $a \in \mathcal{R}$.
(a) Find the UMVUE of a when θ is known.

(b) Find the UMVUE of θ when a is known.
(c) Find the UMVUE's of θ and a.
(d) Assume that θ is known. Find the UMVUE of $P(X_1 \geq t)$ and $\frac{d}{dt}P(X_1 \geq t)$ for a fixed $t > 0$.
(e) Find the UMVUE of $P(X_1 \geq t)$ for a fixed $t > 0$.

8. Let $X_1, ..., X_n$ be i.i.d. having the Pareto distribution $Pa(a, \theta)$ with $\theta > 0$ and $a > 0$.
(a) Find the UMVUE of θ when a is known.
(b) Find the UMVUE of a when θ is known.
(c) Find the UMVUE's of a and θ.

9. Consider Exercise 41(a) of §2.6. Find the UMVUE of γ.

10. Let $X_1, ..., X_m$ be i.i.d. having the exponential distribution $E(a_x, \theta_x)$ with $\theta_x > 0$ and $a_x \in \mathcal{R}$ and $Y_1, ..., Y_n$ be i.i.d. having the exponential distribution $E(a_y, \theta_y)$ with $\theta_y > 0$ and $a_y \in \mathcal{R}$. Assume that X_i's and Y_j's are independent.
(a) Find the UMVUE's of $a_x - a_y$ and θ_x/θ_y.
(b) Suppose that $\theta_x = \theta_y$ but it is unknown. Find the UMVUE's of θ_x and $(a_x - a_y)/\theta_x$.
(c) Suppose that $a_x = a_y$ but it is unknown. Show that a UMVUE of a_x does not exist.
(d) Suppose that $n = m$ and $a_x = a_y = 0$ and that our sample is $(Z_1, \Delta_1), ..., (Z_n, \Delta_n)$, where $Z_i = \min(X_i, Y_i)$ and $\Delta_i = 1$ if $X_i \geq Y_i$ and 0 otherwise, $i = 1, ..., n$. Find the UMVUE of $\theta_x - \theta_y$.

11. Let $X_1, ..., X_m$ be i.i.d. having the uniform distribution $U(0, \theta_x)$ and $Y_1, ..., Y_n$ be i.i.d. having the uniform distribution $U(0, \theta_y)$. Suppose that X_i's and Y_j's are independent and that $\theta_x > 0$ and $\theta_y > 0$. Find the UMVUE of θ_x/θ_y when $n > 1$.

12. Let X be a random variable having the negative binomial distribution $NB(p, r)$ with an unknown $p \in (0, 1)$ and a known r.
(a) Find the UMVUE of p^t, $t < r$.
(b) Find the UMVUE of $\mathrm{Var}(X)$.
(c) Find the UMVUE of $\log p$.

13. Let $X_1, ..., X_n$ be i.i.d. random variables having the Poisson distribution $P(\theta)$ truncated at 0, i.e., $P(X_i = x) = (e^\theta - 1)^{-1}\theta^x/x!$, $x = 1, 2, ...$, $\theta > 0$. Find the UMVUE of θ when $n = 1, 2$.

14. Let X be a random variable having the negative binomial distribution $NB(p, r)$ truncated at r, where r is known and $p \in (0, 1)$ is unknown. Let k be a fixed positive integer $> r$.
(a) For $r = 1, 2, 3$, find the UMVUE of p^k.
(b) For $r = 1, 2, 3$, find the UMVUE of $P(X = k)$.

15. Let $X_1, ..., X_n$ be i.i.d. having the log-distribution $L(p)$ with an unknown $p \in (0,1)$. Let k be a fixed positive integer.
(a) For $n = 1, 2, 3$, find the UMVUE of p^k.
(b) For $n = 1, 2, 3$, find the UMVUE of $P(X = k)$.

16. Suppose that $(X_0, X_1, ..., X_k)$ has the multinomial distribution in Example 2.7 with $p_i \in (0,1)$, $\sum_{j=0}^k p_j = 1$. Find the UMVUE of $p_0^{r_0} \cdots p_k^{r_k}$, where r_j's are nonnegative integers with $r_0 + \cdots + r_k \leq n$.

17. Let $X_1, ..., X_n$ be i.i.d. from $P \in \mathcal{P}$ containing all symmetric c.d.f.'s with finite means and with Lebesgue p.d.f.'s on $\mathcal{R}$. Show that there is no UMVUE of $\mu = EX_1$.

18. Let $(X_1, Y_1), ..., (X_n, Y_n)$ be i.i.d. random 2-vectors from a population $P \in \mathcal{P}$ which is the family of all bivariate populations with Lebesgue p.d.f.'s.
(a) Show that the set of n pairs (X_i, Y_i) ordered according to the value of their first coordinate constitute a sufficient and complete statistic for $P \in \mathcal{P}$.
(b) A statistic T is a function of the complete and sufficient statistic if and only if T is invariant under permutation of the n pairs.
(c) Show that $(n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$ is the UMVUE of $\mathrm{Cov}(X_1, Y_1)$.
(d) Find the UMVUE's of $P(X_i \leq Y_i)$ and $P(X_i \leq X_j$ and $Y_i \leq Y_j)$, $i \neq j$.

19. Prove Corollary 3.1.

20. Consider the problem in Exercise 68 of §2.6. Use Theorem 3.2 to show that $I_{\{0\}}(X)$ is a UMVUE of $(1-p)^2$ and that there is no UMVUE of p.

21. Let $X_1, ..., X_n$ be i.i.d. from a discrete distribution with
$$P(X_i = \theta - 1) = P(X_i = \theta) = P(X_i = \theta + 1) = \tfrac{1}{3},$$
where θ is an unknown integer. Show that no nonconstant function of θ has a UMVUE.

22. Let X be a random variable having the Lebesgue p.d.f.
$$[(1-\theta) + \theta/(2\sqrt{x})] I_{(0,1)}(x),$$
where $\theta \in [0,1]$. Show that there is no UMVUE of θ.

23. Let X be a discrete random variable with $P(X = -1) = 2p(1-p)$ and $P(X = k) = p^k(1-p)^{3-k}$, $k = 0, 1, 2, 3$, where $p \in (0,1)$.
(a) Determine whether there is a UMVUE of p.
(b) Determine whether there is a UMVUE of $p(1-p)$.

24. Let $X_1, ..., X_n$ be i.i.d. having the exponential distribution $E(a, \theta)$ with a known θ and an unknown $a \leq 0$. Obtain a UMVUE of a.

25. Let $X_1, ..., X_n$ be i.i.d. having the Pareto distribution $Pa(a, \theta)$ with a known $\theta > 1$ and an unknown $a \in (0, 1]$. Obtain a UMVUE of a.

26. Prove Theorem 3.3 for the multivariate case ($k > 1$).

27. Let X be a single sample from P_θ. Find the Fisher information $I(\theta)$ in the following cases.
(a) P_θ is the $N(\mu, \sigma^2)$ distribution with $\theta = \mu \in \mathcal{R}$.
(b) P_θ is the $N(\mu, \sigma^2)$ distribution with $\theta = \sigma^2 > 0$.
(c) P_θ is the $N(\mu, \sigma^2)$ distribution with $\theta = \sigma > 0$.
(d) P_θ is the $N(\sigma, \sigma^2)$ distribution with $\theta = \sigma > 0$.
(e) P_θ is the $N(\mu, \sigma^2)$ distribution with $\theta = (\mu, \sigma^2) \in \mathcal{R} \times (0, \infty)$.
(f) P_θ is negative binomial distribution $NB(\theta, r)$ with $\theta \in (0, 1)$.
(g) P_θ is the gamma distribution $\Gamma(\alpha, \gamma)$ with $\theta = (\alpha, \gamma) \in (0, \infty) \times (0, \infty)$;
(h) P_θ is the beta distribution $B(\alpha, \beta)$ with $\theta = (\alpha, \beta) \in (0, 1) \times (0, 1)$.

28. Find a function of θ for which the amount of information is independent of θ, when P_θ is
(a) the Poisson distribution $P(\theta)$ with $\theta > 0$;
(b) the binomial distribution $Bi(\theta, r)$ with $\theta \in (0, 1)$;
(c) the gamma distribution $\Gamma(\alpha, \theta)$ with $\theta > 0$.

29. Prove the result in Example 3.9. Show that if μ (or σ) is known, then $I_1(\mu)$ (or $I_2(\sigma)$) is the first (or second) diagonal element of $I(\theta)$.

30. Obtain the Fisher information matrix for
(a) the Cauchy distribution $C(\mu, \sigma)$, $\mu \in \mathcal{R}$, $\sigma > 0$;
(b) the double exponential distribution $DE(\mu, \theta)$, $\mu \in \mathcal{R}$, $\theta > 0$;
(c) the logistic distribution $LG(\mu, \sigma)$, $\mu \in \mathcal{R}$, $\sigma > 0$;
(d) $F_r\left(\frac{x-\mu}{\sigma}\right)$, where F_r is the c.d.f. of the t-distribution t_r with a known r, $\mu \in \mathcal{R}$, $\sigma > 0$.

31. Let ϕ be the standard normal p.d.f. Find the Fisher information contained in X which has the Lebesgue p.d.f.

$$f_\theta(x) = (1 - \epsilon)\phi(x - \mu) + \frac{\epsilon}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right),$$

$\theta = (\mu, \sigma, \epsilon) \in \mathcal{R} \times (0, \infty) \times (0, 1)$.

32. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution $U(0, \theta)$ with $\theta > 0$.
(a) Show that condition (3.3) does not hold for $h(X) = X_{(n)}$.
(b) Show that the inequality (3.6) does not apply to the UMVUE of θ.

33. Prove Proposition 3.3.

34. Let X be a single sample from the double exponential distribution $DE(\mu, \theta)$ with $\mu = 0$ and $\theta > 0$. Find the UMVUE's of the following parameters and, in each case, determine whether the variance of the UMVUE attains the Cramér-Rao lower bound.
(a) $\vartheta = \theta$;
(b) $\vartheta = \theta^r$, where $r > 1$;
(c) $\vartheta = (1 + \theta)^{-1}$.

35. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $P(X_i = 1) = p \in (0, 1)$.
(a) Show that the UMVUE of $p(1 - p)$ is $T_n = n\bar{X}(1 - \bar{X})/(n - 1)$.
(b) Show that $\mathrm{Var}(T_n)$ does not attain the Cramér-Rao lower bound.
(c) Show that (3.10) holds.

36. Let $X_1, ..., X_n$ be i.i.d. having the Poisson distribution $P(\theta)$ with $\theta > 0$. Find the amse of the UMVUE of $e^{-t\theta}$ with a fixed $t > 0$ and show that (3.10) holds.

37. Let $X_1, ..., X_n$ be i.i.d. having the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathcal{R}$ and a known $\sigma^2 > 0$.
(a) Find the UMVUE of $\vartheta = e^{t\mu}$ with a fixed $t \neq 0$.
(b) Determine whether the variance of the UMVUE in (a) attains the Cramér-Rao lower bound.
(c) Show that (3.10) holds.

38. Show that if $X_1, ..., X_n$ are i.i.d. binary random variables, U_n in (3.12) equals $T(T - 1) \cdots (T - m + 1)/[n(n - 1) \cdots (n - m + 1)]$, where $T = \sum_{i=1}^n X_i$.

39. Show that if $T_n = \bar{X}$, then U_n in (3.13) is the same as the sample variance S^2 in (2.2). Show that (3.23) holds for T_n given by (3.22) with $E(R_n^2) = o(n^{-1})$.

40. Prove (3.14) and (3.17).

41. Let ζ_k be given in Theorem 3.4. Show that $\zeta_1 \leq \zeta_2 \leq \cdots \leq \zeta_m$.

42. Prove Corollary 3.2.

43. Prove (3.20) and show that $U_n - \check{U}_n$ is also a U-statistic.

44. Let T_n be a symmetric statistic with $\mathrm{Var}(T_n) < \infty$ for every n and $\check{T}_n$ be the projection of T_n on $\binom{n}{k}$ random vectors $\{X_{i_1}, ..., X_{i_k}\}$, $1 \leq i_1 < \cdots < i_k \leq n$. Show that $E(T_n) = E(\check{T}_n)$ and calculate $E(T_n - \check{T}_n)^2$.

45. Let Y_k be defined in Lemma 3.2. Show that $\{Y_k^2\}$ is uniformly integrable.

46. Show that (3.22) with $E(R_n^2) = o(n^{-1})$ is satisfied for T_n being a U-statistic with $E[h(X_1, ..., X_m)]^2 < \infty$.

47. Let S^2 be the sample variance given by (2.2), which is also a U-statistic (§3.2.1). Find the corresponding h_1, h_2, ζ_1, and ζ_2. Discuss how to apply Theorem 3.5 to this case.

48. Let $h(x_1, x_2, x_3) = I_{(-\infty,0)}(x_1 + x_2 + x_3)$. Define the U-statistic with this kernel and find h_k and ζ_k, $k = 1, 2, 3$.

49. Show that any $\hat{\beta}$ given by (3.29) is an LSE of β.

50. Obtain explicit forms for the LSE's of β_j, $j = 0, 1$, and SSR, under the simple linear regression model in Example 3.11, assuming that some t_i's are different.

51. Consider the polynomial model

$$X_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \varepsilon_i, \quad i = 1, ..., n.$$

Find explicit forms for the LSE's of β_j, $j = 0, 1, 2$, and SSR, assuming that some t_i's are different.

52. Suppose that

$$X_{ij} = \alpha_i + \beta t_{ij} + \varepsilon_{ij}, \quad i = 1, ..., a, \ j = 1, ..., b.$$

Find explicit forms for the LSE's of β, α_i, $i = 1, ..., a$, and SSR.

53. Find the matrix Z, $Z^\tau Z$, and the form of $l \in \mathcal{R}(Z)$ under the one-way ANOVA model (3.31).

54. Obtain the matrix Z under the two-way balanced ANOVA model (3.32). Show that the rank of Z is ab. Verify the form of the LSE of β given in Example 3.14. Find the form of $l \in \mathcal{R}(Z)$.

55. Consider the following model as a special case of model (3.25):

$$X_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}, \quad i = 1, ..., a, j = 1, ..., b, k = 1, ..., c.$$

Obtain the matrix Z, the parameter vector β, and the form of LSE's of β. Discuss conditions under which $l \in \mathcal{R}(Z)$.

56. Under model (3.25) and assumption A1, find the UMVUE's of $(\beta l^\tau)^2$, $\beta l^\tau/\sigma$, and $(\beta l^\tau/\sigma)^2$ for an estimable βl^τ.

57. Verify the formulas for SSR's in Example 3.15.

58. Consider model (3.25) with assumption A2. Show that $\mathrm{Var}(\hat{\beta}l^\tau) = \sigma^2 l(Z^\tau Z)^- l^\tau$ for $l \in \mathcal{R}(Z)$.

59. Consider the one-way random effects model in Example 3.17. Assume that $n_i = n$ for all i and that A_i's and e_{ij}'s are normally distributed. Show that the family of populations is an exponential family with sufficient and complete statistics $\bar{X}_{\cdot\cdot}$, $S_A = n\sum_{i=1}^m (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2$, and $S_E = \sum_{i=1}^m \sum_{j=1}^n (X_{ij} - \bar{X}_{i\cdot})^2$. Find the UMVUE's of μ, σ_a^2, and σ^2.

60. Consider model (3.25). Suppose that ε_i's are i.i.d. with a Lebesgue p.d.f. $\sigma^{-1}f(x/\sigma)$, where f is a known Lebesgue p.d.f. and $\sigma > 0$ is unknown.
(a) Show that X is from a subfamily of the location-scale family given by (2.10).
(b) Find the Fisher information about (β, σ) contained in X_i.
(c) Find the Fisher information about (β, σ) contained in X.

61. Consider model (3.25) with assumption A2. Let $c \in \mathcal{R}^p$. Show that if the equation $c = yZ^\tau$ has a solution, then there is a unique solution $y_0 \in \mathcal{R}(Z)$ such that $\mathrm{Var}(Xy_0^\tau) \le \mathrm{Var}(Xy^\tau)$ for any other solution of $c = yZ^\tau$.

62. Consider model (3.25). Show that the number of independent linear functions of X with mean 0 is $n - r$, where r is the rank of Z.

63. Consider model (3.25) with assumption A2. Let $\hat{X}_i = \hat{\beta}Z_i^\tau$, which is called the least squares *prediction* of X_i. Let h_{ij} be the (i,j)th element of $Z(Z^\tau Z)^- Z^\tau$. Show that
(a) $\mathrm{Var}(\hat{X}_i) = \sigma^2 h_{ii}$;
(b) $\mathrm{Var}(X_i - \hat{X}_i) = \sigma^2(1 - h_{ii})$;
(c) $\mathrm{Cov}(\hat{X}_i, \hat{X}_j) = \sigma^2 h_{ij}$;
(d) $\mathrm{Cov}(X_i - \hat{X}_i, X_j - \hat{X}_j) = -\sigma^2 h_{ij}$, $i \ne j$;
(e) $\mathrm{Cov}(\hat{X}_i, X_j - \hat{X}_j) = 0$.

64. Prove that (e) implies (b) in Theorem 3.10.

65. Show that (a) in Theorem 3.10 is equivalent to either
(f) $\mathrm{Var}(\varepsilon)Z = ZB$ for some matrix B, or
(g) $\mathcal{R}(Z)$ is generated by r eigenvectors of $\mathrm{Var}(\varepsilon)$, where r is the rank of Z.

66. Prove Corollary 3.3.

67. Suppose that
$$X = \mu J_n + H\xi + e,$$

where $\mu \in \mathcal{R}$ is an unknown parameter, H is an $n \times p$ known matrix of full rank, ξ is a random p-vector with $E(\xi) = 0$ and $\mathrm{Var}(\xi) = \sigma_\xi^2 I_p$, e is a random n-vector with $E(e) = 0$ and $\mathrm{Var}(e) = \sigma^2 I_n$, and ξ and e are independent. Show that the LSE of μ is the BLUE if and only if the row totals of HH^τ are the same.

68. Consider a special case of model (3.25):

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad i = 1, ..., a, j = 1, ..., b,$$

where μ, α_i's and β_j's are unknown parameters, $E(\varepsilon_{ij}) = 0$, $\mathrm{Var}(\varepsilon_{ij}) = \sigma^2$, $\mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0$ if $i \neq i'$, and $\mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{ij'}) = \sigma^2\rho$ if $j \neq j'$. Show that the LSE of βl^τ is the BLUE for any $l \in \mathcal{R}(Z)$.

69. Consider model (3.25) under assumption A3 with $\mathrm{Var}(\varepsilon)$ = a block diagonal matrix whose ith block diagonal V_i is $n_i \times n_i$ and has a single eigenvalue λ_i with eigenvector J_{n_i} and a repeated eigenvalue ρ_i with multiplicity $n_i - 1$, $i = 1, ..., k$, $\sum_{i=1}^k n_i = n$. Let $U = (U_1^\tau, ..., U_k^\tau)$, where $U_1 = (J_{n_1}, 0, ..., 0)$, $U_2 = (0, J_{n_2}, ..., 0)$,..., $U_k = (0, 0, ..., J_{n_k})$.
(a) If $\mathcal{R}(Z^\tau) \subset \mathcal{R}(U^\tau)$ and $\lambda_i \equiv \lambda$, show that $\hat{\beta} l^\tau$ is the BLUE for any $l \in \mathcal{R}(Z)$.
(b) If $Z^\tau U_i = 0$ for all i and $\rho_i \equiv \rho$, show that $\hat{\beta} l^\tau$ is the BLUE for any $l \in \mathcal{R}(Z)$.

70. Prove Proposition 3.4.

71. Show that the condition $\sup_n \lambda_+[\mathrm{Var}(\varepsilon)] < \infty$ is equivalent to the condition $\sup_i \mathrm{Var}(\varepsilon_i) < \infty$.

72. Find a condition under which the mse of $\hat{\beta} l^\tau$ is of the order n^{-1}. Apply it to problems in Exercises 50-53.

73. Consider model (3.25) with i.i.d. $\varepsilon_1, ..., \varepsilon_n$ having $E(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Let $\hat{X}_i = \hat{\beta} Z_i^\tau$ and $h_{ii} = Z_i (Z^\tau Z)^- Z_i^\tau$.
(a) Show that for any $\epsilon > 0$,

$$P(|\hat{X}_i - E\hat{X}_i| \geq \epsilon) \geq \min[P(\varepsilon_i \geq \epsilon/h_i), P(\varepsilon_i \leq -\epsilon/h_i)].$$

(Hint: for independent random variables X and Y, $P(|X+Y| \geq \epsilon) \geq P(X \geq \epsilon)P(Y \geq 0) + P(X \leq -\epsilon)P(Y < 0)$.)
(b) Show that $\hat{X}_i - E\hat{X}_i \to_p 0$ if and only if $h_{ii} \to 0$.

74. Prove Lemma 3.3 and show that condition (a) is implied by $\{\|Z_i\|\}$ is bounded and $\lambda_+(Z^\tau Z)^- \to 0$.

75. Consider the problem in Exercise 52. Suppose that $\{t_{ij}\}$ is bounded. Find a condition under which (3.40) holds.

76. Consider the one-way random effects model in Example 3.17. Assume that $\{n_i\}$ is bounded and $E|e_{ij}|^{2+\delta} < \infty$ for some $\delta > 0$. Show that the LSE $\hat{\mu}$ of μ is asymptotically normal and derive an explicit form of $\mathrm{Var}(\hat{\mu})$.

77. Suppose that
$$X_i = \rho t_i + \varepsilon_i, \quad i = 1, ..., n,$$
where $\rho \in \mathcal{R}$ is an unknown parameter, $t_i \in (a, b)$, $i = 1, ..., n$, a and b are known positive constants, and ε_i's are independent random variables satisfying $E(\varepsilon_i) = 0$, $E|\varepsilon_i|^{2+\delta} < \infty$ for some $\delta > 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2 t_i$ with an unknown $\sigma^2 > 0$.
(a) Obtain the LSE of ρ.
(b) Obtain the BLUE of ρ.
(c) Show that both the LSE and BLUE are asymptotically normal and obtain the asymptotic relative efficiency of the BLUE w.r.t. the LSE.

78. In Example 3.19, show that $E(S^2) = \sigma^2$ given by (3.44).

79. Suppose that $X = (X_1, ..., X_n)$ is a simple random sample (without replacement) from a finite population $\mathcal{P} = \{y_1, ..., y_N\}$ with univariate y_i.
(a) Show that a necessary condition for $h(\theta)$ to be estimable is that h is symmetric in its N arguments.
(b) Find the UMVUE of Y^m, where m is a fixed positive integer $< n$ and Y is the population total.
(c) Find the UMVUE of $P(X_i \leq X_j)$, $i \neq j$.
(d) Find the UMVUE of $\mathrm{Cov}(X_i, X_j)$, $i \neq j$.

80. Prove Theorem 3.14.

81. Under stratified simple random sampling described in §3.4.1, show that the vector of ordered values of all X_{hi}'s is neither sufficient nor complete for $\theta \in \Theta$.

82. Let $\mathcal{P} = \{y_1, ..., y_N\}$ be a population with univariate y_i. Define the population c.d.f. by
$$F(t) = \frac{1}{N} \sum_{i=1}^{N} I_{(-\infty, t)}(y_i).$$
Find the UMVUE of $F(t)$ under (a) simple random sampling and (b) stratified simple random sampling.

83. Consider the estimation of $F(t)$ in the previous exercise. Suppose that a sample of size n is selected with $\pi_i > 0$. Find the Horvitz-Thompson estimator of $F(t)$. Is it a c.d.f.?

84. Show that v_1 in (3.50) and v_2 in (3.51) are unbiased estimators of $\mathrm{Var}(\hat{Y}_{ht})$. Prove that $v_1 = v_2$ under (a) simple random sampling and (b) stratified simple random sampling.

85. Consider the following two-stage stratified sampling plan. In the first stage, the population is stratified into H strata and k_h clusters are selected from stratum h with probability proportional to cluster size, where sampling is independent across strata. In the second stage, a sample of m_{hi} units are selected from sampled cluster i in stratum h, and sampling is independent across clusters. Find π_i and the Horvitz-Thompson estimator $\hat{Y}_{ht}$ of the population total.

86. In the previous exercise, prove the unbiasedness of $\hat{Y}_{ht}$ directly (without using Theorem 3.15).

87. Under systematic sampling, show that $\mathrm{Var}(\hat{Y}_{sy})$ is equal to

$$\left(1-\frac{1}{N}\right)\frac{\sigma^2}{n}+\frac{2}{nN}\sum_{i=1}^{k}\sum_{1\le t<u\le n}\left(y_{i+(t-1)k}-\frac{Y}{N}\right)\left(y_{i+(u-1)k}-\frac{Y}{N}\right).$$

88. Prove (3.52)-(3.54) in Example 3.21. Show that (3.52) and (3.54) still hold if $(X_1, Y_1), ..., (X_n, Y_n)$ is a simple random sample from a finite bivariate population of size N, as $n \to N$.

89. Derive the n^{-1} order asymptotic bias of the sample correlation coefficient defined in Exercise 19 in §2.6.

90. Derive the n^{-1} order asymptotic bias and amse of $\hat{t}_\beta$ in Example 3.22, assuming that $\sum_{j=0}^{p-1}\beta_j t^j$ is convex in t.

91. Consider Example 3.23.
(a) Show that $\hat{\theta}$ is the BLUE of θ.
(b) Show that $\hat{\sigma}^2$ is unbiased for σ^2.
(c) Show that $\hat{\Sigma}$ is consistent for Σ as $k \to \infty$.
(d) Derive an amse of $\hat{R}(t)$.

92. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$, where $\mu \in \mathcal{R}$ and $\sigma^2 > 0$. Consider the estimation of $\vartheta = E\Phi(a+bX_1)$, where Φ is the standard normal c.d.f. and a and b are known constants. Obtain an explicit form of a function $g(\mu, \sigma^2) = \vartheta$ and an amse of $\hat{\vartheta} = g(\bar{X}, S^2)$.

93. Let $X_1, ..., X_n$ be i.i.d. with mean μ, variance σ^2, and finite $\mu_j = EX_1^j$, $j = 2, 3, 4$. The sample coefficient of variation is defined to be $S/\bar{X}$, where S is the squared root of the sample variance S^2.
(a) If $\mu \neq 0$, show that $\sqrt{n}(S/\bar{X} - \sigma/\mu) \to_d N(0, \tau)$ and obtain an explicit formula of τ in terms of μ, σ^2, and μ_j.
(b) If $\mu = 0$, show that $n^{-1/2}S/\bar{X} \to_d [N(0, 1)]^{-1}$.

94. Prove (3.56).

95. In Exercise 83, discuss how to obtain a consistent (as $n \to N$) estimator $\hat{F}(t)$ of $F(t)$ such that $\hat{F}$ is a c.d.f.

96. Let $X_1, ..., X_n$ be i.i.d. from P in a parametric family. Obtain moment estimators of parameters in the following cases.
(a) P is the gamma distribution $\Gamma(\alpha, \gamma)$, $\alpha > 0$, $\gamma > 0$.
(b) P is the exponential distribution $E(a, \theta)$, $a \in \mathcal{R}$, $\theta > 0$.
(c) P is the beta distribution $B(\alpha, \beta)$, $\alpha > 0$, $\beta > 0$.
(d) P is the log-normal distribution $LN(\mu, \sigma^2)$, $\mu \in \mathcal{R}$, $\sigma > 0$.
(e) P is the uniform distribution $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, $\theta \in \mathcal{R}$.
(f) P is the negative binomial distribution $NB(p, r)$, $p \in (0, 1)$, $r = 1, 2, ...$
(g) P is the log-distribution $L(p)$, $p \in (0, 1)$.
(h) P is the chi-square distribution χ^2_k with an unknown $k = 1, 2, ...$

97. In part (b) of the previous exercise, obtain the asymptotic relative efficiencies of moment estimators w.r.t. UMVUE's.

98. Prove (3.57) and (3.58).

99. In the proof of Proposition 3.5, show that $E[W_n(U_n - \vartheta)] = O(n^{-1})$.

100. Prove (3.61).

101. Let $X_1, ..., X_n$ be i.i.d. with a c.d.f. F and U_n and V_n be the U- and V-statistics with kernel $\int [I_{(-\infty,y)}(x_1) - F_0(y)][I_{(-\infty,y)}(x_2) - F_0(y)]dF_0$, where F_0 is a known c.d.f.
(a) Obtain the asymptotic distributions of U_n and V_n when $F \neq F_0$.
(b) Obtain the asymptotic relative efficiency of U_n w.r.t. V_n when $F = F_0$.

102. Let $X_1, ..., X_n$ be i.i.d. with a c.d.f. F having a finite 6th moment. Consider the estimation of μ^3, where $\mu = EX_1$. When $\mu = 0$, find $\text{amse}_{\bar{X}^3}(P)/\text{amse}_{U_n}(P)$, where $U_n = \binom{n}{3}^{-1} \sum_{1 \le i < j < k \le n} X_i X_j X_k$.

103. Prove (3.67).

104. Prove that $\hat{\sigma}^2$ in (3.69) is unbiased and consistent for σ^2 under model (3.25) with (3.68) and $\sup_i E|\varepsilon_i|^{2+\delta} < \infty$ for some $\delta > 0$. Under the same conditions, show that $\hat{\Sigma}$ in (3.70) is consistent for Σ.

105. Show how to use equation (3.71) to obtain consistent estimators of θ_0 and θ_1.

Chapter 4

Estimation in Parametric Models

In this chapter we consider point estimation methods in parametric models. One such method, the moment method, has been introduced in §3.5.2. It is assumed in this chapter that the sample X is from a population in a parametric family $\{P_\theta : \theta \in \Theta\}$, where $\Theta \subset \mathcal{R}^k$ for a fixed integer $k \geq 1$.

4.1 Bayes Decisions and Estimators

Bayes rules are introduced in §2.3.2 as decision rules minimizing the average risk w.r.t. a given probability measure Π on Θ. Bayes rules, however, are optimal rules in the *Bayesian approach* which is fundamentally different from the classical frequentist approach that we have been adopting.

4.1.1 Bayes actions

In the Bayesian approach, θ is viewed as a realization of a random vector $\boldsymbol{\theta}$ whose *prior* distribution is Π. The prior distribution is based on past experience, past data, or statistician's belief and, thus, can be very subjective. A sample X is drawn from $P_\theta = P_{x|\theta}$, which is viewed as the conditional distribution of X given $\boldsymbol{\theta} = \theta$. The sample $X = x$ is then used to obtain an updated prior distribution, which is called the *posterior* distribution and can be derived as follows. By Proposition 1.15, the joint distribution of X and $\boldsymbol{\theta}$ is a probability measure on $\mathcal{X} \times \Theta$ determined by

$$P(A \times B) = \int_B P_{x|\theta}(A) d\Pi(\theta), \qquad A \in \mathcal{B}_{\mathcal{X}},\ B \in \mathcal{B}_\Theta,$$

where $\mathcal{X}$ is the range of X. The posterior distribution of $\boldsymbol{\theta}$, given $X = x$, is the conditional distribution $P_{\theta|x}$ whose existence is guaranteed by Theorem 1.7 for almost all $x \in \mathcal{X}$. When $P_{x|\theta}$ has a p.d.f., the following result provides a formula for the p.d.f. of the posterior distribution $P_{\theta|x}$.

Theorem 4.1 (The Bayes formula). Assume that $\mathcal{P} = \{P_{x|\theta} : \theta \in \Theta\}$ is dominated by a σ-finite measure ν and $f_\theta(x) = \frac{dP_{x|\theta}}{d\nu}(x)$ is a Borel function on $(\mathcal{X} \times \Theta, \sigma(\mathcal{B}_\mathcal{X} \times \mathcal{B}_\Theta))$. Let Π be a prior distribution on Θ. Suppose that $m(x) = \int_\Theta f_\theta(x) d\Pi > 0$.
(i) The posterior distribution $P_{\theta|x} \ll \Pi$ and

$$\frac{dP_{\theta|x}}{d\Pi} = \frac{f_\theta(x)}{m(x)}.$$

(ii) If $\Pi \ll \lambda$ and $\frac{d\Pi}{d\lambda} = \pi(\theta)$ for a σ-finite measure λ, then

$$\frac{dP_{\theta|x}}{d\lambda} = \frac{f_\theta(x)\pi(\theta)}{m(x)}. \tag{4.1}$$

Proof. Result (ii) follows from result (i) and Proposition 1.7(iii). To show (i), we first show that $m(x) < \infty$ a.e. ν. Note that

$$\int_\mathcal{X} m(x) d\nu = \int_\mathcal{X} \int_\Theta f_\theta(x) d\Pi d\nu = \int_\Theta \int_\mathcal{X} f_\theta(x) d\nu d\Pi = 1, \tag{4.2}$$

where the second equality follows from Fubini's theorem. Thus, $m(x)$ is integrable w.r.t. ν and $m(x) < \infty$ a.e. ν.

For $x \in \mathcal{X}$ with $m(x) < \infty$, define

$$P(B, x) = \frac{1}{m(x)} \int_B f_\theta(x) d\Pi \qquad B \in \mathcal{B}_\Theta.$$

Then $P(\cdot, x)$ is a probability measure on Θ a.e. ν. By Theorem 1.7, it remains to show that

$$P(B, x) = P(\boldsymbol{\theta} \in B | X = x).$$

Note that $P(B, \cdot)$ is a measurable function of x. Let $P_{x,\theta}$ denote the "joint" distribution of $(X, \boldsymbol{\theta})$. For any $A \in \sigma(X)$,

$$\begin{aligned}
\int_{A\times\Theta} I_B(\theta) dP_{x,\theta} &= \int_A \int_B f_\theta(x) d\Pi d\nu \\
&= \int_A \left[\int_B \frac{f_\theta(x)}{m(x)} d\Pi\right] \left[\int_\Theta f_\theta(x) d\Pi\right] d\nu \\
&= \int_\Theta \int_A \left[\int_B \frac{f_\theta(x)}{m(x)} d\Pi\right] f_\theta(x) d\nu d\Pi \\
&= \int_{A\times\Theta} P(B, x) dP_{x,\theta},
\end{aligned}$$

where the third equality follows from Fubini's theorem. This completes the proof. ∎

Because of (4.2), $m(x)$ is called the marginal p.d.f. of X w.r.t. ν. If $m(x) = 0$ for an $x \in \mathfrak{X}$, then $f_\theta(x) = 0$ a.s. Π. Thus, either x should be eliminated from $\mathfrak{X}$ or the prior Π is incorrect and a new prior should be specified. Therefore, without loss of generality we may assume that the assumption of $m(x) > 0$ in Theorem 4.1 is always satisfied.

If both X and $\boldsymbol{\theta}$ are discrete and ν and λ are the counting measures, then (4.1) becomes

$$P(\boldsymbol{\theta} = \theta | X = x) = \frac{P(X = x|\boldsymbol{\theta} = \theta)P(\boldsymbol{\theta} = \theta)}{\sum_{\theta\in\Theta} P(X = x|\boldsymbol{\theta} = \theta)P(\boldsymbol{\theta} = \theta)},$$

which is the Bayes formula that appears in elementary probability.

In the Bayesian approach, the posterior distribution $P_{\theta|x}$ contains all the information we have about θ and, therefore, statistical decisions and inference should be made based on $P_{\theta|x}$, conditional on the observed $X = x$. In the problem of estimating θ, $P_{\theta|x}$ can be viewed as a randomized decision rule under the approach discussed in §2.3.

Definition 4.1. Let $\mathbb{A}$ be an action space in a decision problem and $L(\theta, a) \geq 0$ be a loss function. For any $x \in \mathfrak{X}$, a *Bayes action* w.r.t. Π is any $\delta(x) \in \mathbb{A}$ such that

$$E[L(\boldsymbol{\theta}, \delta(x))|X = x] = \min_{a\in\mathbb{A}} E[L(\boldsymbol{\theta}, a)|X = x], \tag{4.3}$$

where the expectation is w.r.t. the posterior distribution $P_{\theta|x}$. ∎

The existence and uniqueness of Bayes actions can be discussed under some conditions on the loss function and the action space.

Proposition 4.1. Assume that the conditions in Theorem 4.1 hold; $L(\theta, a)$ is convex in a for each fixed θ; and for each $x \in \mathfrak{X}$, $E[L(\boldsymbol{\theta}, a)|X = x] < \infty$ for some a.
(i) If $\mathbb{A}$ is a compact subset of $\mathcal{R}^p$ for some integer $p \geq 1$, then a Bayes action $\delta(x)$ exists for each $x \in \mathfrak{X}$.
(ii) If $\mathbb{A} = \mathcal{R}^p$ and $L(\theta, a)$ tends to ∞ as $a \to \infty$ uniformly in $\theta \in \Theta_0 \subset \Theta$ with $\Pi(\Theta_0) > 0$, then a Bayes action $\delta(x)$ exists for each $x \in \mathfrak{X}$.
(iii) In (i) or (ii), if $L(\theta, a)$ is strictly convex in a for each fixed θ, then the Bayes action is unique.
Proof. The convexity of the loss function implies the convexity and continuity of $E[L(\boldsymbol{\theta}, a)|X = x]$ as a function of a with any fixed x. Then, the result in (i) follows from the fact that any continuous function on a compact

set attains its minimum. The result in (ii) follows from the fact that

$$\lim_{\|a\|\to\infty} E[L(\boldsymbol{\theta}, a)|X = x] \geq \lim_{\|a\|\to\infty} \int_{\Theta_0} L(\theta, a) dP_{\theta|x} = \infty$$

under the assumed condition in (ii). Finally, the result in (iii) follows from the fact that $E[L(\boldsymbol{\theta}, a)|X = x]$ is strictly convex in a for any fixed x under the assumed conditions. ∎

Other conditions on L under which a Bayes action exists can be found, for example, in Lehmann (1983, §1.6 and §4.1).

Example 4.1. Consider the estimation of $\vartheta = g(\theta)$ for some real-valued function g such that $\int_\Theta [g(\theta)]^2 d\Pi < \infty$. Suppose that $\mathbb{A}$ = the range of $g(\theta)$ and $L(\theta, a) = [g(\theta) - a]^2$ (squared error loss). Using the same argument as in Example 1.19, we obtain the Bayes action

$$\delta(x) = \frac{\int_\Theta g(\theta) f_\theta(x) d\Pi}{m(x)} = \frac{\int_\Theta g(\theta) f_\theta(x) d\Pi}{\int_\Theta f_\theta(x) d\Pi}, \tag{4.4}$$

which is the posterior expectation of $g(\boldsymbol{\theta})$, given $X = x$.

More specifically, let us consider the case where $g(\theta) = \theta^j$ for some integer $j \geq 1$, $f_\theta(x) = e^{-\theta}\theta^x I_{\{0,1,2,...\}}(x)/x!$ (the Poisson distribution) with $\theta > 0$, and Π has a Lebesgue p.d.f. $\pi(\theta) = \theta^{\alpha-1}e^{-\theta/\gamma} I_{(0,\infty)}(\theta)/[\Gamma(\alpha)\gamma^\alpha]$ (the gamma distribution $\Gamma(\alpha, \gamma)$ with known $\alpha > 0$ and $\gamma > 0$). Then, for $x = 0, 1, 2, ...$,

$$\frac{f_\theta(x)\pi(\theta)}{m(x)} = c(x)\theta^{x+\alpha-1}e^{-\theta(\gamma+1)/\gamma} I_{(0,\infty)}(\theta), \tag{4.5}$$

where $c(x)$ is some function of x. By using Theorem 4.1 and matching the right-hand side of (4.5) with that of the p.d.f. of the gamma distribution, we know that the posterior is the gamma distribution $\Gamma(x + \alpha, \gamma/(\gamma + 1))$. Hence, without actually working out the integral $m(x)$, we know that $c(x) = (1 + \gamma^{-1})^{x+\alpha}/\Gamma(x + \alpha)$. Then

$$\delta(x) = c(x) \int_0^\infty \theta^{j+x+\alpha-1} e^{-\theta(\gamma+1)/\gamma} d\theta.$$

Note that the integrand is proportional to the p.d.f. of the gamma distribution $\Gamma(j + x + \alpha, \gamma/(\gamma + 1))$. Hence

$$\begin{aligned}\delta(x) &= c(x)\Gamma(j + x + \alpha)/(1 + \gamma^{-1})^{j+x+\alpha} \\ &= (j + x + \alpha - 1) \cdots (x + \alpha)/(1 + \gamma^{-1})^j.\end{aligned}$$

In particular, $\delta(x) = (x + \alpha)\gamma/(\gamma + 1)$ when $j = 1$. ∎

An interesting phenomenon in Example 4.1 is that the prior and the posterior are in the same parametric family of distributions. Such a prior is called a *conjugate* prior. Under a conjugate prior, Bayes actions often have explicit forms (in x) when the loss function is simple. Whether a prior is conjugate involves a pair of families; one is the family $\mathcal{P} = \{f_\theta : \theta \in \Theta\}$ and the other is the family from which Π is chosen. Example 4.1 shows that the Poisson family and the gamma family produce conjugate priors. It can be shown (exercise) that many pairs of families in Table 1.1 (page 18) and Table 1.2 (pages 20-21) produce conjugate priors.

In general, numerical methods have to be used in evaluating the integrals in (4.4) or Bayes actions under general loss functions. Even under a conjugate prior, the integral in (4.4) involving a general g may not have an explicit form. More discussions on the computation of Bayes actions are given in §4.1.4.

As an example of deriving a Bayes action in a general decision problem, we consider Example 2.21.

Example 4.2. Consider the decision problem in Example 2.21. Let $P_{\theta|x}$ be the posterior distribution of $\boldsymbol{\theta}$, given $X = x$. In this problem, $\mathbb{A} = \{a_1, a_2, a_3\}$, which is compact in $\mathcal{R}$. By Proposition 4.1, we know that there is a Bayes action if the mean of $P_{\theta|x}$ is finite. Let $E_{\theta|x}$ be the expectation w.r.t. $P_{\theta|x}$. Since $\mathbb{A}$ contains only three elements, a Bayes action can be obtained by comparing

$$E_{\theta|x}[L(\theta, a_j)] = \begin{cases} c_1 & j = 1 \\ c_2 + c_3 E_{\theta|x}[\psi(\boldsymbol{\theta}, t)] & j = 2 \\ c_3 E_{\theta|x}[\psi(\boldsymbol{\theta}, 0)] & j = 3, \end{cases}$$

where $\psi(\theta, t) = (\theta - \theta_0 - t) I_{(\theta_0 + t, \infty)}(\theta)$. ∎

The minimization problem (4.3) is the same as the minimization problem

$$\int_\Theta L(\theta, \delta(x)) f_\theta(x) d\Pi = \min_{a \in \mathbb{A}} \int_\Theta L(\theta, a) f_\theta(x) d\Pi. \tag{4.6}$$

The minimization problem (4.6) is still defined even if Π is not a probability measure but a σ-finite measure on Θ, in which case $m(x)$ may not be finite. If $\Pi(\Theta) = \infty$, Π is called an *improper prior*. A prior with $\Pi(\Theta) = 1$ is then called a proper prior. An action $\delta(x)$ that satisfies (4.6) with an improper prior is called a *generalized Bayes action*.

The following is a reason why we need to discuss improper priors and generalized Bayes actions. In many cases one has no past information and has to choose a prior subjectively. In such cases one would like to select a *noninformative* prior that tries to treat all parameter values in Θ

equitably. A noninformative prior is often improper. We only provide one example here. For more detailed discussions of the use of improper priors, see Jeffreys (1939, 1948, 1961), Box and Tiao (1973), and Berger (1985).

Example 4.3. Suppose that $X_1, ..., X_n$ are i.i.d. from $N(\mu, \sigma^2)$, where $\mu \in \Theta \subset \mathcal{R}$ is unknown and σ^2 is known. Consider the estimation of $\vartheta = \mu$ under the squared error loss. If $\Theta = [a, b]$ with $-\infty < a < b < \infty$, then a noninformative prior that treats all parameter values equitably is the uniform distribution on $[a, b]$. If $\Theta = \mathcal{R}$, however, the corresponding "uniform distribution" is the Lebesgue measure on $\mathcal{R}$, which is an improper prior. If Π is the Lebesgue measure on $\mathcal{R}$, then

$$(2\pi\sigma^2)^{-n/2} \int_{-\infty}^{\infty} \mu^2 \exp\left\{-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right\} d\mu < \infty.$$

By differentiating a in

$$(2\pi\sigma^2)^{-n/2} \int_{-\infty}^{\infty} (\mu - a)^2 \exp\left\{-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right\} d\mu$$

and using the fact that $\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$, we obtain that

$$\delta(x) = \frac{\int_{-\infty}^{\infty} \mu \exp\left\{-n(\bar{x} - \mu)^2/(2\sigma^2)\right\} d\mu}{\int_{-\infty}^{\infty} \exp\left\{-n(\bar{x} - \mu)^2/(2\sigma^2)\right\} d\mu} = \bar{x},$$

the sample mean of the observations $x_1, ..., x_n$. Thus, the sample mean is a generalized Bayes action under the squared error loss. From Example 2.25 and Exercise 74 in §2.6, if Π is $N(\mu_0, \sigma_0^2)$, then the Bayes action is $\mu_*(x)$ in (2.28). Note that in this case $\bar{x}$ is a limit of $\mu_*(x)$ as $\sigma_0^2 \to \infty$. ∎

4.1.2 Empirical and hierarchical Bayes methods

A Bayes action depends on the chosen prior which may depend on some parameters called *hyperparameters*. In §4.1.1, hyperparameters are assumed to be known. If hyperparameters are unknown, one way to solve the problem is to estimate them using data $x_1, ..., x_n$; the resulting Bayes action is called an *empirical Bayes* action.

The simplest empirical Bayes method is to estimate prior parameters by viewing $x = (x_1, ..., x_n)$ as a "sample" from the marginal distribution

$$P_{x|\xi}(A) = \int_{\Theta} P_{x|\theta}(A) d\Pi_{\theta|\xi}, \qquad A \in \mathcal{B}_{\mathcal{X}},$$

where $\Pi_{\theta|\xi}$ is a prior depending on an unknown vector ξ of hyperparameters, or from the marginal p.d.f. $m(x)$ in (4.2), if $P_{x|\theta}$ has a p.d.f. f_θ. The method of moments introduced in §3.5.3, for example, can be applied to estimate ξ. We consider an example.

Example 4.4. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with an unknown $\mu \in \mathcal{R}$ and a known σ^2. Consider the prior $\Pi_{\mu|\xi} = N(\mu_0, \sigma_0^2)$ with $\xi = (\mu_0, \sigma_0^2)$. To obtain the moment estimates of ξ, we need to calculate

$$\int_{\mathcal{R}^n} x_1 m(x)dx \qquad \text{and} \qquad \int_{\mathcal{R}^n} x_1^2 m(x)dx.$$

These two integrals can be obtained without calculating $m(x)$. Note that

$$\int_{\mathcal{R}^n} x_1 m(x)dx = \int_{\mathcal{R}^n}\int_\Theta x_1 f_\mu(x)dxd\Pi_{\mu|\xi} = \int_{\mathcal{R}} \mu d\Pi_{\mu|\xi} = \mu_0$$

and

$$\int_{\mathcal{R}^n} x_1^2 m(x)dx = \int_{\mathcal{R}^n}\int_\Theta x_1^2 f_\mu(x)dxd\Pi_{\mu|\xi} = \sigma^2 + \int_{\mathcal{R}} \mu^2 d\Pi_{\mu|\xi} = \sigma^2 + \mu_0^2 + \sigma_0^2.$$

Thus, by viewing $x_1, ..., x_n$ as a sample from $m(x)$, we obtain the moment estimates

$$\hat{\mu}_0 = \bar{x} \qquad \text{and} \qquad \hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 - \sigma^2.$$

Replacing μ_0 and σ_0^2 in formula (2.28) (Example 2.25) by $\hat{\mu}_0$ and $\hat{\sigma}_0^2$, respectively, we find that the empirical Bayes action under the squared error loss is simply the sample mean $\bar{x}$ (which is a generalized Bayes action; see Example 4.3). ∎

Note that $\hat{\sigma}_0^2$ in Example 4.4 can be negative. Better empirical Bayes methods can be found, for example, in Berger (1985, §4.5). The following method, called the *hierarchical Bayes* method, is generally better than empirical Bayes methods.

Instead of estimating hyperparameters, in the hierarchical Bayes approach we put a prior on hyperparameters. Let $\Pi_{\theta|\xi}$ be a (first-stage) prior with a hyperparameter vector ξ and let Λ be a prior on Ξ, the range of ξ. Then the "marginal" prior for θ is defined by

$$\Pi(B) = \int_\Xi \Pi_{\theta|\xi}(B)d\Lambda(\xi) \qquad B \in \mathcal{B}_\Theta. \tag{4.7}$$

If the second-stage prior Λ also depends on some unknown hyperparameters, then one can go on to consider a third-stage prior. In most applications,

however, two-stage priors are sufficient, since misspecifying a second-stage prior is much less serious than misspecifying a first-stage prior (Berger, 1985, §4.6). In addition, the second-stage prior can be chosen to be noninformative (improper).

Bayes actions can be obtained in the same way as before, using the prior in (4.7). Thus, the hierarchical Bayes method is simply a Bayes method with a hierarchical prior. Empirical Bayes methods, however, deviate from the Bayes method since $x_1, ..., x_n$ are used to estimate hyperparameters.

Suppose that X has a p.d.f. $f_\theta(x)$ w.r.t. a σ-finite measure ν and $P_{\theta|\xi}$ has a p.d.f. $\pi_{\theta|\xi}(\theta)$ w.r.t. a σ-finite measure κ. Then the prior Π in (4.7) has a p.d.f.

$$\pi(\theta) = \int_\Xi \pi_{\theta|\xi}(\theta) d\Lambda(\xi)$$

w.r.t. κ and

$$m(x) = \int_\Theta \int_\Xi f_\theta(x) \pi_{\theta|\xi}(\theta) d\Lambda d\kappa.$$

Let $P_{\theta|x,\xi}$ be the posterior distribution of $\boldsymbol{\theta}$ given x and ξ (or ξ is assumed known) and

$$m_{x|\xi}(x) = \int_\Theta f_\theta(x) \pi_{\theta|\xi}(\theta) d\kappa,$$

which is the marginal of X given θ and ξ (or ξ is assumed known). Then the posterior distribution $P_{\theta|x}$ has a p.d.f.

$$\begin{aligned}
\frac{dP_{\theta|x}}{d\kappa} &= \frac{f_\theta(x)\pi(\theta)}{m(x)} \\
&= \int_\Xi \frac{f_\theta(x)\pi_{\theta|\xi}(\theta)}{m(x)} d\Lambda(\xi) \\
&= \int_\Xi \frac{f_\theta(x)\pi_{\theta|\xi}(\theta)}{m_{x|\xi}(x)} \frac{m_{x|\xi}(x)}{m(x)} d\Lambda(\xi) \\
&= \int_\Xi \frac{dP_{\theta|x,\xi}}{d\kappa} dP_{\xi|x},
\end{aligned}$$

where $P_{\xi|x}$ is the posterior distribution of ξ given x. Thus, under the estimation problem considered in Example 4.1, the (hierarchical) Bayes action is

$$\delta(x) = \int_\Xi \delta(x, \xi) dP_{\xi|x},$$

where $\delta(x, \xi)$ is the Bayes action when ξ is known.

Example 4.5. Consider Example 4.4 again. Suppose that one of the parameters in the first-stage prior $N(\mu_0, \sigma_0^2)$, μ_0, is unknown and σ_0^2 is

known. Let the second-stage prior for $\xi = \mu_0$ be the Lebesgue measure on $\mathcal{R}$ (improper prior). From Example 2.25,

$$\delta(x, \xi) = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\xi + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{x}.$$

To obtain the Bayes action $\delta(x)$, it suffices to calculate $E_{\xi|x}(\xi)$, where the expectation is w.r.t. $P_{\xi|x}$. Note that the p.d.f. of $P_{\xi|x}$ is proportional to

$$\psi(\xi) = \int_{-\infty}^{\infty} \exp\left\{-\tfrac{n(\bar{x}-\mu)^2}{2\sigma^2} - \tfrac{(\mu-\xi)^2}{2\sigma_0^2}\right\} d\mu.$$

Using the properties of normal distributions, one can show that

$$\begin{aligned}\psi(\xi) &= C_1 \exp\left\{\left(\tfrac{n}{2\sigma^2} + \tfrac{1}{2\sigma_0^2}\right)^{-1}\left(\tfrac{n\bar{x}}{2\sigma^2} + \tfrac{\xi}{2\sigma_0^2}\right)^2 - \tfrac{\xi^2}{2\sigma_0^2}\right\} \\ &= C_2 \exp\left\{-\tfrac{n\xi^2}{2(n\sigma_0^2+\sigma^2)} + \tfrac{n\bar{x}\xi}{n\sigma_0^2+\sigma^2}\right\} \\ &= C_3 \exp\left\{-\tfrac{n(\xi-\bar{x})^2}{2(n\sigma_0^2+\sigma^2)}\right\},\end{aligned}$$

where C_1, C_2, and C_3 are quantities not depending on ξ. Hence $E_{\xi|x}(\xi) = \bar{x}$. The (hierarchical) generalized Bayes action is then

$$\delta(x) = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} E_{\xi|x}(\xi) + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{x} = \bar{x}. \quad \blacksquare$$

4.1.3 Bayes rules and estimators

The discussion in §4.1.1 and §4.1.2 is more general than point estimation and adopts an approach that is different from the frequentist approach used in the rest of this book. In the frequentist approach, a Bayes action $\delta(x)$ viewed as a function of x defines a decision rule. It is easy to see that $\delta(x)$ defined in Definition 4.1 also minimizes the *Bayes risk* (defined in §2.3.2)

$$r_T(\Pi) = \int_\Theta R_T(\theta) d\Pi,$$

where T is any decision rule and $R_T(\theta)$ is the risk function of T defined in (2.19). Thus, $\delta(X)$ is called a Bayes rule (§2.3.2). In an estimation problem, a Bayes rule is called a *Bayes estimator*.

Generalized Bayes risks, generalized Bayes rules (or estimators), and empirical Bayes rules (or estimators) can be defined similarly.

In view of the discussion in §2.3.2, even if we do not adopt the Bayesian approach, the method described in §4.1.1 can be used as a way of generating

decision rules. In this section we study a Bayes rule or estimator in terms of its risk (and bias and consistency for a Bayes estimator).

Bayes rules are typically admissible, since, if there is a rule better than a Bayes rule, then that rule has the same Bayes risk as the Bayes rule and, therefore, is itself a Bayes rule. This actually proves part (i) of the following result. The proof of the other parts of the following result is left as an exercise.

Theorem 4.2. In a decision problem, let $\delta(X)$ be a Bayes rule w.r.t. a prior Π.
(i) If $\delta(X)$ is a unique Bayes rule, then $\delta(X)$ is admissible.
(ii) If Θ is a countable set and Π gives positive probability to each $\theta \in \Theta$, then $\delta(X)$ is admissible.
(iii) If the risk $R_T(\theta)$ is a continuous function of θ for every T (with a finite risk) and Π gives positive probability to any open subset of Θ, then $\delta(X)$ is admissible. ∎

Generalized Bayes rules or estimators are not necessarily admissible. Many generalized Bayes rules are limits of Bayes rules (see Examples 4.3 and 4.7). Limits of Bayes rules are often admissible (Farrell, 1968a,b). The following result shows a technique of proving admissibility using limits of (generalized) Bayes risks.

Theorem 4.3. Suppose that Θ is an open set of $\mathcal{R}^k$. In a decision problem, let $\Im$ be the class of decision rules having continuous risk functions. A decision rule $T \in \Im$ is $\Im$-admissible if there exists a sequence $\{\Pi_j\}$ of (possibly improper) priors such that (a) the Bayes risks $r_T(\Pi_j)$ are finite for all j; (b) for any open neighborhood $O \subset \Theta$, there are $j_0 > 0$ and $c > 0$ such that $\Pi_j(O) \geq c$ for all $j \geq j_0$; and (c) $\lim_{j\to\infty}[r_T(\Pi_j) - r_{\delta_j}(\Pi_j)] = 0$, where δ_j is the Bayes rule w.r.t. Π_j.

Proof. Suppose that T is not $\Im$-admissible. Then there exists $T_0 \in \Im$ such that $R_{T_0}(\theta) \leq R_T(\theta)$ for all θ and $R_{T_0}(\theta_0) < R_T(\theta_0)$ for some $\theta_0 \in \Theta$. From the continuity of the risk functions, we conclude that $R_{T_0}(\theta) < R_T(\theta) - \epsilon$ for all $\theta \in O = \{\theta \in \Theta : \|\theta - \theta_0\| < \eta\}$, where $\epsilon > 0$ and $\eta > 0$ are some constants. From conditions (a) and (b), for sufficiently large j,

$$\begin{aligned} r_T(\Pi_j) - r_{\delta_j}(\Pi_j) &\geq r_T(\Pi_j) - r_{T_0}(\Pi_j) \\ &\geq \int_O [R_T(\theta) - R_{T_0}(\theta)] d\Pi_j(\theta) \\ &\geq \epsilon \Pi_j(O) \\ &\geq \epsilon c, \end{aligned}$$

which contradicts condition (c). Hence, T is $\Im$-admissible. ∎

Example 4.6. Consider Example 4.3 and the estimation of μ under the squared error loss. From Theorem 2.1, the risk function of any decision rule is continuous in μ if the risk is finite. We now apply Theorem 4.3 to show that the sample mean $\bar{X}$ is admissible. Let

$$\pi_j(\mu) = (2\pi)^{-1/2} e^{-\mu^2/(2j)}, \qquad j = 1, 2, ...,$$

and let $\Pi_j(\mu) = \int_{-\infty}^{\mu} \pi_j(t)dt$. Note that Π_j is not a probability measure, but

$$\Pi_j(O) \geq \Pi_1(O) > 0, \quad j = 2, 3, ...$$

for any open interval O, i.e., condition (b) of Theorem 4.3 is satisfied. Note that if we choose $\Pi_j = N(0, j^{-1})$, then condition (b) is not satisfied. A direct calculation shows that

$$r_{\bar{X}}(\Pi_j) = \frac{\sqrt{j}\sigma^2}{n} \qquad \text{and} \qquad r_{\delta_j}(\Pi_j) = \frac{\sqrt{j}j\sigma^2}{nj + \sigma^2}.$$

Hence (a) of Theorem 4.3 is satisfied. Finally,

$$r_{\bar{X}}(\Pi_j) - r_{\delta_j}(\Pi_j) = \frac{\sqrt{j}\sigma^4}{n(nj + \sigma^2)} \to 0$$

as $j \to \infty$. Thus, (c) of Theorem 4.3 is satisfied and, hence, the sample mean $\bar{X}$ is admissible. ∎

From Example 4.6, it can been seen that the choice of Π_j in applying Theorem 4.3 is very elaborate. More results in admissibility can be found in §4.2 and §4.3.

The following result concerns the bias of a Bayes estimator.

Proposition 4.2. Let $\delta(X)$ be a Bayes estimator of $\vartheta = g(\theta)$ under the squared error loss. Then $\delta(X)$ is not unbiased unless the Bayes risk $r_\delta(\Pi) = 0$.

Proof. Suppose that $\delta(X)$ is unbiased. Conditioning on $\boldsymbol{\theta}$ and using Proposition 1.12, we obtain that

$$E[g(\boldsymbol{\theta})\delta(X)] = E\{g(\boldsymbol{\theta})E[\delta(X)|\boldsymbol{\theta}]\} = E[g(\boldsymbol{\theta})]^2.$$

Since $\delta(X) = E[g(\boldsymbol{\theta})|X]$, conditioning on X and using Proposition 1.12, we obtain that

$$E[g(\boldsymbol{\theta})\delta(X)] = E\{\delta(X)E[g(\boldsymbol{\theta})|X]\} = E[\delta(X)]^2.$$

Then

$$r_\delta(\Pi) = E[\delta(X) - g(\boldsymbol{\theta})]^2 = E[\delta(X)]^2 + E[g(\boldsymbol{\theta})]^2 - 2E[g(\boldsymbol{\theta})\delta(X)] = 0. \quad ∎$$

Since $r_\delta = 0$ occurs usually in some trivial cases, a Bayes estimator is typically not unbiased. Hence, Proposition 4.2 can be used to check whether an estimator can be a Bayes estimator w.r.t. some prior under the squared error loss. However, a generalized Bayes estimator may be unbiased; see, for instance, Examples 4.3 and 4.7.

Bayes estimators are usually consistent and asymptotically unbiased. In a particular problem, it is usually easy to check directly whether Bayes estimators are consistent and asymptotically unbiased (Examples 4.7-4.9). Bayes estimators also have some other good asymptotic properties, which are studied in §4.5.3.

Let us consider some examples.

Example 4.7. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(0, \theta)$ with an unknown $\theta > 0$. Let the prior be such that θ^{-1} has the gamma distribution $\Gamma(\alpha, \gamma)$ with known $\alpha > 0$ and $\gamma > 0$. Then the posterior of $\omega = \theta^{-1}$ is the gamma distribution $\Gamma(n + \alpha, (n\bar{x} + \gamma^{-1})^{-1})$ (exercise).

Consider first the estimation of $\theta = \omega^{-1}$. The Bayes estimator of θ under the squared error loss is

$$\delta(x) = \frac{(n\bar{x} + \gamma^{-1})^{n+\alpha}}{\Gamma(n + \alpha)} \int_0^\infty \omega^{n+\alpha-2} e^{-(n\bar{x}+\gamma^{-1})\omega} d\omega = \frac{n\bar{x} + \gamma^{-1}}{n + \alpha - 1}.$$

The bias of $\delta(X)$ is

$$\frac{n\theta + \gamma^{-1}}{n + \alpha - 1} - \theta = \frac{\gamma^{-1} - (\alpha - 1)\theta}{n + \alpha - 1} = O\left(\frac{1}{n}\right).$$

It is also easy to see that $\delta(X)$ is consistent. The UMVUE of θ is $\bar{X}$. Since $\mathrm{Var}(\bar{X}) = \theta^2/n$, $r_{\bar{X}}(\Pi) > 0$ for any Π and, hence, $\bar{X}$ is not a Bayes estimator. In this case, $\bar{X}$ is the generalized Bayes estimator w.r.t. the improper prior $\frac{d\Pi}{d\omega} = I_{(0,\infty)}(\omega)$ and is a limit of Bayes estimators $\delta(X)$ as $\alpha \to 1$ and $\gamma \to \infty$ (exercise).

Consider next the estimation of $e^{-t/\theta} = e^{-t\omega}$ (see Examples 2.26 and 3.3). The Bayes estimator under the squared error loss is

$$\begin{aligned}\delta(x) &= \frac{(n\bar{x} + \gamma^{-1})^{n+\alpha}}{\Gamma(n + \alpha)} \int_0^\infty \omega^{n+\alpha-1} e^{-(n\bar{x}+\gamma^{-1}+t)\omega} d\omega \\ &= \left(1 + \frac{t}{n\bar{x} + \gamma^{-1}}\right)^{-(n+\alpha)}. \end{aligned} \tag{4.8}$$

Again, this estimator is biased and it is easy to show that $\delta(X)$ is consistent as $n \to \infty$. In this case, the UMVUE given in Example 3.3 is neither a Bayes estimator nor a limit of $\delta(X)$. ∎

Example 4.8. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with unknown $\mu \in \mathcal{R}$ and $\sigma^2 > 0$. Let the prior for $\omega = (2\sigma^2)^{-1}$ be the gamma distribution $\Gamma(\alpha, \gamma)$ with known $\alpha > 0$ and $\gamma > 0$ and let the prior for μ be $N(\mu_0, \sigma_0^2/\omega)$ (conditional on ω). Then the posterior p.d.f. of (μ, ω) is proportional to

$$\omega^{(n-1)/2+\alpha-1} \exp\left\{-\left[\gamma^{-1} + y + n(\bar{x}-\mu)^2 + \tfrac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]\omega\right\},$$

where $y = \sum_{i=1}^n (x_i - \bar{x})^2$. Note that

$$n(\bar{x}-\mu)^2 + \tfrac{(\mu-\mu_0)^2}{2\sigma_0^2} = \left(n + \tfrac{1}{2\sigma_0^2}\right)\mu^2 - 2\left(n\bar{x} + \tfrac{\mu_0}{2\sigma_0^2}\right)\mu + n\bar{x}^2 + \tfrac{\mu_0^2}{2\sigma_0^2}.$$

Hence the posterior p.d.f. of (μ, ω) is proportional to

$$\omega^{(n-1)/2+\alpha-1} \exp\left\{-\left[\gamma^{-1} + y + \left(n + \tfrac{1}{2\sigma_0^2}\right)(\mu - \zeta(x))^2\right]\omega\right\},$$

where

$$\zeta(x) = \frac{n\bar{x} + \frac{\mu_0}{2\sigma_0^2}}{n + \frac{1}{2\sigma_0^2}}.$$

Thus, the posterior of ω is the gamma distribution $\Gamma(n/2+\alpha-1, (\gamma^{-1}+y)^{-1})$ and the posterior of μ (given ω and $X = x$) is $N\big(\zeta(x), [(2n + \sigma_0^{-2})\omega]^{-1}\big)$. Under the squared error loss, the Bayes estimator of μ is $\zeta(x)$ and the Bayes estimator of $\sigma^2 = (2\omega)^{-1}$ is $(\gamma^{-1} + y)/(n + 2\alpha - 4)$, provided that $n + 2\alpha > 4$. Apparently, these Bayes estimators are biased but the biases are of the order n^{-1}; and they are consistent as $n \to \infty$. ∎

To consider the last example, we need the following useful lemma whose proof is left as an exercise.

Lemma 4.1. Suppose that X has a p.d.f. $f_\theta(x)$ w.r.t. a σ-finite measure ν. Suppose that $\theta = (\theta_1, \theta_2)$, $\theta_j \in \Theta_j$, and that the prior has a p.d.f.

$$\pi(\theta) = \pi_{\theta_1|\theta_2}(\theta_1)\pi_{\theta_2}(\theta_2),$$

where $\pi_{\theta_2}(\theta_2)$ is a p.d.f. w.r.t. a σ-finite measure ν_2 on Θ_2 and for any given θ_2, $\pi_{\theta_1|\theta_2}(\theta_1)$ is a p.d.f. w.r.t. a σ-finite measure ν_1 on Θ_1. Suppose further that if θ_2 is given, the Bayes estimator of $h(\theta_1) = g(\theta_1, \theta_2)$ under the squared error loss is $\delta(X, \theta_2)$. Then the Bayes estimator of $g(\theta_1, \theta_2)$ under the squared error loss is $\delta(X)$ with

$$\delta(x) = \int_{\Theta_2} \delta(X, \theta_2) p_{\theta_2|x}(\theta_2) d\nu_2,$$

where $p_{\theta_2|x}(\theta_2)$ is the posterior p.d.f. of $\boldsymbol{\theta}_2$ given $X = x$. ∎

Example 4.9. Consider a linear model

$$X_{ij} = \beta Z_i^\tau + \varepsilon_{ij}, \qquad j = 1, ..., n_i,\ i = 1, ..., m,$$

where $\beta \in \mathcal{R}^p$ is unknown, Z_i's are known vectors, ε_{ij}'s are independent, and ε_{ij} is $N(0, \sigma_i^2)$, $j = 1, ..., n_i$, $i = 1, ..., m$. The parameter vector is then $\theta = (\beta, \omega)$, where $\omega = (\omega_1, ..., \omega_k)$ and $\omega_i = (2\sigma_i^2)^{-1}$. Assume that the prior for θ has the Lebesgue p.d.f.

$$c\,\pi(\beta) \prod_{i=1}^m \omega_i^\alpha e^{-\omega_i/\gamma}, \tag{4.9}$$

where $\alpha > 0$, $\gamma > 0$, and $c > 0$ are known constants and $\pi(\beta)$ is a known Lebesgue p.d.f. on $\mathcal{R}^p$. The joint p.d.f. of $(X, \boldsymbol{\theta})$ is then proportional to

$$h(x, \theta) = \pi(\beta) \prod_{i=1}^m \omega_i^{n_i/2+\alpha} e^{-[\gamma^{-1}+v_i(\beta)]\omega_i},$$

where $v_i(\beta) = \sum_{j=1}^{n_i}(x_{ij} - \beta Z_i^\tau)^2$. Suppose first that β is known. Then the Bayes estimator of σ_i^2 under the squared error loss is

$$\int \frac{1}{2\omega_i} \frac{h(x, \theta)}{\int h(x, \theta) d\omega} d\omega = \frac{\gamma^{-1} + v_i(\beta)}{2\alpha + n_i}.$$

By Lemma 4.1, the Bayes estimator of σ_i^2 is

$$\hat{\sigma}_i^2 = \int \frac{\gamma^{-1} + v_i(\beta)}{2\alpha + n_i} f_{\beta|x}(\beta) d\beta, \tag{4.10}$$

where

$$\begin{aligned} f_{\beta|x}(\beta) &\propto \int h(x, \theta) d\omega \\ &\propto \pi(\beta) \prod_{i=1}^m \int \omega_i^{\alpha+n_i/2} e^{-[\gamma^{-1}+v_i(\beta)]\omega_i} d\omega_i \\ &\propto \pi(\beta) \prod_{i=1}^m \left[\gamma^{-1} + v_i(\beta)\right]^{-(\alpha+n_i/2)} \end{aligned} \tag{4.11}$$

is the posterior p.d.f. of β. The Bayes estimator of βl^τ for any $l \in \mathcal{R}^p$ is then the posterior mean of βl^τ w.r.t. the p.d.f. $f_{\beta|x}(\beta)$.

In this problem, Bayes estimators do not have explicit forms. A numerical method (such as one of those in §4.1.4) has to be used to evaluate Bayes estimators (see Example 4.10).

Let $\bar{X}_{i\cdot}$ and S_i^2 be the sample mean and variance of X_{ij}, $j = 1, ..., n_i$ and let $\sigma_0^2 = (2\alpha\gamma)^{-1}$ (the prior mean of σ_i^2). Then the Bayes estimator in (4.10) can be written as

$$\frac{2\alpha}{2\alpha + n_i}\sigma_0^2 + \frac{n_i - 1}{2\alpha + n_i}S_i^2 + \frac{n_i}{2\alpha + n_i}\int (\bar{X}_{i\cdot} - \beta Z_i^\tau)^2 f_{\beta|x}(\beta)d\beta \qquad (4.12)$$

(S_i^2 is defined to be 0 if $n_i = 1$). The Bayes estimator in (4.12) is a weighted average of prior information, "within group" variation, and an averaged squared "residual".

If $n_i \to \infty$, then the first term in (4.12) converges to 0 and the second term in (4.12) converges to σ_i^2 a.s. Hence, the Bayes estimator is consistent as $n_i \to \infty$, since the mean of the third term in (4.12) is bounded by

$$E\int (\bar{X}_{i\cdot} - \beta Z_i^\tau)^2 f_{\beta|x}(\beta)d\beta = \frac{\sigma_0^2}{n_i}. \quad \blacksquare$$

4.1.4 Markov chain Monte Carlo

As we discussed previously, Bayes actions or estimators have to be computed numerically in many applications. Typically we need to compute an integral of the form

$$E_p(g) = \int_\Theta g(\theta)p(\theta)d\nu$$

with some function g, where $p(\theta)$ is a p.d.f. w.r.t. a σ-finite measure ν on $\Theta \subset \mathcal{R}^k$. For example, if g is an indicator function of A and $p(\theta)$ is the posterior p.d.f. of θ given $X = x$, then $E_p(g)$ is the posterior probability of A; under the squared error loss, $E_p(g)$ is the Bayes action (4.4) if $p(\theta)$ is the posterior p.d.f.

There are many numerical methods for computing integrals $E_p(g)$; see, for example, §4.5.3 and Berger (1985, §4.9). In this section we discuss the *Markov Chain Monte Carlo* (MCMC) methods, which are powerful numerical methods not only for Bayesian computations, but also for general statistical computing (see, e.g., §4.4.1).

We start with the simple Monte Carlo method, which can be viewed as a special case of the MCMC. Suppose that we can generate i.i.d. $\theta^{(1)}, ..., \theta^{(m)}$ from a p.d.f. $h(\theta) > 0$ w.r.t. ν. By the SLLN (Theorem 1.13(ii)), as $m \to \infty$,

$$\hat{E}_p(g) = \frac{1}{m}\sum_{j=1}^m \frac{g(\theta^{(j)})p(\theta^{(j)})}{h(\theta^{(j)})} \to_{a.s.} \int_\Theta \frac{g(\theta)p(\theta)}{h(\theta)}h(\theta)d\nu = E_p(g).$$

Hence $\hat{E}_p(g)$ can be used as a numerical approximation to $E_p(g)$. The process of generating $\theta^{(j)}$ according to h is called *importance sampling* and

$h(\theta)$ is called the *importance function*. More discussions on importance sampling can be found, for example, in Berger (1985), Geweke (1989), Shao (1989), and Tanner (1996). When $p(\theta)$ is intractable or complex, it is often difficult to choose a function h that is simple enough for importance sampling and results in a fast convergence of $\hat{E}_p(g)$ as well.

The simple Monte Carlo method, however, may not work well when k, the dimension of Θ, is large. This is because when k is large, the convergence of $\hat{E}_p(g)$ requires a very large m; generating a random vector from a k-dimensional distribution is usually expensive, if not impossible. More sophisticated MCMC methods are different from the simple Monte Carlo in two aspects: generating random vectors can be done using distributions whose dimensions are much lower than k; and $\theta^{(1)}, ..., \theta^{(m)}$ are not independent, but form a Markov chain, which is described next.

A sequence of random k-vectors $\{Y^{(t)} : t = 0, 1, ...\}$ taking values in $\mathcal{Y}$ is a homogeneous Markov chain if and only if

$$P(Y^{(t+1)} \in A | Y^{(0)}, ..., Y^{(t)}) = P(Y^{(1)} \in A | Y^{(0)})$$

for any t. Let

$$P(y, A) = P(Y^{(1)} \in A | Y^{(0)} = y), \qquad y \in \mathcal{Y},\ A \in \mathcal{B}_{\mathcal{Y}},$$

which is called the transition kernel of the Markov chain. Note that $P(y, \cdot)$ is a probability measure for every $y \in \mathcal{Y}$; $P(\cdot, A)$ is a Borel function for every $A \in \mathcal{B}_{\mathcal{Y}}$; and the distribution of a homogeneous Markov chain is determined by $P(y, A)$ and the distribution of $Y^{(0)}$ (initial distribution). MCMC approximates an integral of the form $\int_{\mathcal{Y}} g(y)p(y)d\nu$ by $m^{-1}\sum_{t=1}^{m} g(Y^{(t)})$ with a Markov chain $\{Y^{(t)} : t = 0, 1, ...\}$. The basic justification of the MCMC approximation is given in the following result.

Theorem 4.4. Let $p(y)$ be a p.d.f. on $\mathcal{Y}$ w.r.t. a σ-finite measure ν and g be a Borel function on $\mathcal{Y}$ with $\int_{\mathcal{Y}} |g(y)|p(y)d\nu < \infty$. Let $\{Y^{(t)} : t = 0, 1, ...\}$ be a homogeneous Markov chain taking values on $\mathcal{Y} \subset \mathcal{R}^k$ with the transition kernel $P(y, A)$. Then

$$\frac{1}{m}\sum_{t=1}^{m} g(Y^{(t)}) \to_{a.s.} \int_{\mathcal{Y}} g(y)p(y)d\nu \tag{4.13}$$

and, as $t \to \infty$,

$$P^t(y, A) = P(Y^{(t)} \in A | Y^{(0)} = y) \to_{a.s.} \int_A p(y)d\nu, \tag{4.14}$$

provided that
(a) the Markov chain is *aperiodic* in the sense that there does not exist $d \geq 2$

nonempty disjoint events $A_0, ..., A_{d-1}$ in $\mathcal{B}_{\mathcal{Y}}$ such that for all $i = 0, ..., d-1$ and all $y \in \mathcal{Y}$, $P(y, A_j) = 1$ for $j = i + 1 \pmod d$;
(b) the Markov chain is *p-invariant* in the sense that $\int P(y, A)p(y)d\nu = \int_A p(y)d\nu$ for all $A \in \mathcal{B}_{\mathcal{Y}}$;
(c) the Markov chain is *p-irreducible* in the sense that for any $y \in \mathcal{Y}$ and any A with $\int_A p(y)d\nu > 0$, there exists a positive integer t such that $P^t(y, A)$ in (4.14) is positive; and
(d) the Markov chain is *Harris recurrent* in the sense that for any A with $\int_A p(y)d\nu > 0$, $P\left(\sum_{t=1}^{\infty} I_A(Y^{(t)}) = \infty | Y^{(0)} = y\right) = 1$ for all y. ∎

The proof of these results is beyond the scope of this book and, hence, is omitted. It can be found in, for example, Nummelin (1984), Chan (1993), and Tierney (1994). A homogeneous Markov chain satisfying conditions (a)-(d) in Theorem 4.4 is called *ergodic* with *equilibrium distribution* p. Result (4.13) means that the MCMC approximation is consistent and result (4.14) indicates that p is the limit p.d.f. of the Markov chain.

One of the key issues in MCMC is the choice of the kernel $P(y, A)$. The first requirement on $P(y, A)$ is that conditions (a)-(d) in Theorem 4.4 are satisfied. Condition (a) is usually easy to check for any given $P(y, A)$. In the following we consider two popular MCMC methods satisfying conditions (a)-(d).

Gibbs sampler

One way to construct a p-invariant homogeneous Markov chain is to use conditioning. Suppose that Y has the p.d.f. $p(y)$. Let Y_i (or y_i) be the ith component of Y (or y) and let Y_{-i} (or y_{-i}) be the $(k-1)$-vector containing all components of Y (or y) except Y_i (or y_i). Then

$$P_i(y, A) = P_i(y_{-i}, A) = P(Y \in A | Y_{-i} = y_{-i})$$

is a transition kernel for any i. The MCMC method using this kernel is called the *single-site Gibbs sampler*. Note that

$$\int P_i(y_{-i}, A)p(y)d\nu = E[P(Y \in A|Y_{-i})] = P(Y \in A) = \int_A p(y)d\nu$$

and, therefore, the chain with kernel $P_i(y_{-i}, A)$ is p-invariant. However, this chain is not p-irreducible since $P(y_{-i}, \cdot)$ puts all its mass on the set $\psi_i^{-1}(y_{-i})$, where $\psi_i(y) = y_{-i}$. Gelfand and Smith (1990) considered a *systematic scan Gibbs sampler* whose kernel $P(y, A)$ is a composite of k kernels $P_i(y_{-i}, A)$, $i = 1, ..., k$. More precisely, the chain is defined as follows. Given $Y^{(t-1)} = y^{(t-1)}$, we generate $y_1^{(t)}$ from $P_1(y_2^{(t-1)}, ..., y_k^{(t-1)}, \cdot)$,..., $y_j^{(t)}$ from $P_j(y_1^{(t)}, ..., y_{j-1}^{(t)}, y_{j+1}^{(t-1)}, ..., y_k^{(t-1)}, \cdot)$,..., $y_k^{(t)}$ from $P_k(y_1^{(t)}, ..., y_{k-1}^{(t)}, \cdot)$. The

initial $Y^{(0)}$ is generated from p. It can be shown that this Markov chain is still p-invariant. We illustrate this with the case of $k = 2$. Note that $Y_1^{(1)}$ is generated from $P_2(y_2^{(0)}, \cdot)$, the conditional distribution of Y given $Y_2 = y_2^{(0)}$. Hence $(Y_1^{(1)}, Y_2^{(0)})$ has p.d.f. p. Similarly, we can show that $Y^{(1)} = (Y_1^{(1)}, Y_2^{(1)})$ has p.d.f. p. Thus,

$$\begin{aligned}\int P(y, A)p(y)d\nu &= \int P(Y^{(1)} \in A|Y^{(0)} = y)p(y)d\nu \\ &= E[P(Y^{(1)} \in A|Y^{(0)})] \\ &= P(Y^{(1)} \in A) \\ &= \int_A p(y)d\nu.\end{aligned}$$

This Markov chain is also p-irreducible and aperiodic if $p(y) > 0$ for all $y \in \mathcal{Y}$; see, for example, Chan (1993). Finally, if $p(y) > 0$ for all $y \in \mathcal{Y}$, then $P(y, A) \ll$ the distribution with p.d.f. p for all y and, by Corollary 1 of Tierney (1994), the Markov chain is Harris recurrent. Thus, Theorem 4.4 applies and (4.13) and (4.14) hold.

The previous Gibbs sampler can obviously be extended to the case where y_i's are subvectors (of possibly different dimensions) of y.

Let us now return to Bayesian computation and consider the following example.

Example 4.10. Consider Example 4.9. Under the given prior for $\theta = (\beta, \omega)$, it is difficult to generate random vectors directly from the posterior p.d.f. $p(\theta)$, given $X = x$ (which does not have a familiar form). To apply a Gibbs sampler with $y = \theta$, $y_1 = \beta$, and $y_2 = \omega$, we need to generate random vectors from the posterior of β, given x and ω, and the posterior of ω, given x and β. From (4.9) and (4.11), the posterior of $\omega = (\omega_1, ..., \omega_k)$, given x and β, is a product of marginals of ω_i's that are the gamma distributions $\Gamma(\alpha + 1 + n_i/2, [\gamma^{-1} + v_i(\beta)]^{-1})$, $i = 1, ..., m$. Assume now that $\pi(\beta) \equiv 1$ (noninformative prior for β). It follows from (4.9) that the posterior p.d.f. of β, given x and ω, is proportional to

$$\prod_{i=1}^{k} e^{-\omega_i v_i(\beta)} \propto e^{-\|\beta Z^\tau W^{1/2} - XW^{1/2}\|^2},$$

where W is the diagonal matrix whose ith diagonal is ω_i. Thus, the posterior of $\beta Z^\tau W^{1/2}$, given x and ω, is $N_p(XW^{1/2}, 2^{-1}I_p)$ and the posterior of β, given x and ω, is $N_p(XWZ(Z^\tau WZ)^{-1}, 2^{-1}(Z^\tau WZ)^{-1})$ ($Z^\tau WZ$ is assumed of full rank for simplicity), since $\beta = \beta Z^\tau W^{1/2}[W^{1/2}Z(Z^\tau WZ)^{-1}]$. Note that random generation using these two posterior distributions is fairly easy. ∎

The Metropolis algorithm

A large class of MCMC methods are obtained using the *Metropolis algorithm* (Metropolis et al., 1953). We introduce Hastings' version of the algorithm. Let $Q(y, A)$ be a transition kernel of a homogeneous Markov chain satisfying

$$Q(y, A) = \int_A q(y, z) d\nu(z)$$

for a measurable function $q(y, z) \geq 0$ on $\mathcal{Y} \times \mathcal{Y}$ and a σ-finite measure ν. Without loss of generality, assume that $\int_{\mathcal{Y}} p(y) d\nu = 1$ and that p is not concentrated on a single point. Define

$$\alpha(y, z) = \begin{cases} \min\left[\frac{p(z)q(z,y)}{p(y)q(y,z)}, 1\right] & p(y)q(y, z) > 0 \\ 1 & p(y)q(y, z) = 0 \end{cases}$$

and

$$p(y, z) = \begin{cases} q(y, z)\alpha(y, z) & y \neq z \\ 0 & y = z. \end{cases}$$

The Metropolis kernel $P(y, A)$ is defined by

$$P(y, A) = \int_A p(y, z) d\nu(z) + r(y)\delta_y(A), \tag{4.15}$$

where δ_y is the point mass at y and $r(y) = 1 - \int p(y, z) d\nu(z)$. The corresponding Markov chain can be described as follows. If the chain is currently at a point $Y^{(t)} = y$, then it generates a candidate value z for the next location $Y^{(t+1)}$ from $Q(y, \cdot)$. With probability $\alpha(y, z)$ the chain moves to $Y^{(t+1)} = z$. Otherwise, the chain remains at $Y^{(t+1)} = y$.

Note that this algorithm only depends on $p(y)$ through $p(y)/p(z)$. Thus, it can be used when $p(y)$ is known up to a normalizing constant, which often occurs in Bayesian analysis.

We now show that a Markov chain with a Metropolis kernel $P(y, A)$ is p-invariant. First, by the definition of $p(y, z)$ and $\alpha(y, z)$,

$$p(y)p(y, z) = p(z)p(z, y)$$

for any y and z. Then, for any $A \in \mathcal{B}_{\mathcal{Y}}$,

$$\begin{aligned} \int P(y, A)p(y)d\nu &= \int \left[\int_A p(y, z)d\nu(z)\right] p(y)d\nu(y) + \int r(y)\delta_y(A)p(y)d\nu(y) \\ &= \int_A \left[\int p(y, z)p(y)d\nu(y)\right] d\nu(z) + \int_A r(y)p(y)d\nu(y) \\ &= \int_A \left[\int p(z, y)p(z)d\nu(y)\right] d\nu(z) + \int_A r(y)p(y)d\nu(y) \end{aligned}$$

$$= \int_A [1 - r(z)]p(z)d\nu(z) + \int_A r(z)p(z)d\nu(z)$$
$$= \int_A p(z)d\nu(z).$$

If a Markov chain with a Metropolis kernel defined by (4.15) is p-irreducible and $\int_{r(y)>0} p(y)d\nu > 0$, then, by the results of Nummelin (1984, §2.4), the chain is aperiodic; by Corollary 2 of Tierney (1994), the chain is Harris recurrent. Hence, to apply Theorem 4.4 to a Markov chain with a Metropolis kernel, it suffices to show that the chain is p-irreducible.

Lemma 4.2. Suppose that $Q(y, A)$ is the transition kernel of a p-irreducible Markov chain and that either $q(y, z) > 0$ for all y and z or $q(y, z) = q(z, y)$ for all y and z. Then the chain with the Metropolis kernel $p(y, A)$ in (4.15) is p-irreducible.

Proof. It can be shown (exercise) that if Q is any transition kernel of a homogeneous Markov chain, then

$$Q^t(y, A) = \int_A \int \cdots \int \prod_{j=1}^{t} q(z_{n-j+1}, z_{n-j})d\nu(z_{n-j}), \tag{4.16}$$

where $z_n = y$, $y \in \mathcal{Y}$, and $A \in \mathcal{B}_\mathcal{Y}$. Let $y \in \mathcal{Y}$, $A \in \mathcal{B}_\mathcal{Y}$ with $\int_A p(z)d\nu > 0$, and $B_y = \{z : \alpha(y, z) = 1\}$. If $\int_{A\cap B_y^c} p(z)d\nu > 0$, then

$$P(y, A) \geq \int_{A\cap B_y^c} q(y, z)\alpha(y, z)d\nu(z) = \int_{A\cap B_y^c} \frac{q(z, y)p(z)}{p(y)} d\nu(z) > 0,$$

which follows from either $q(z, y) > 0$ or $q(z, y) = q(y, z) > 0$ on B_y^c. If $\int_{A\cap B_y^c} p(z)d\nu = 0$, then $\int_{A\cap B_y} p(z)d\nu > 0$. From the irreducibility of $Q(y, A)$, there exists a $t \geq 1$ such that $Q^t(y, A \cap B_y) > 0$. Then, by (4.15) and (4.16),

$$P^t(y, A) \geq P^t(y, A \cap B_y) \geq Q^t(y, A \cap B_y) > 0. \quad \blacksquare$$

Two examples of $q(y, z)$ given by Tierney (1994) are $q(y, z) = f(z - y)$ with a Lebesgue p.d.f. f on $\mathcal{R}^k$, which corresponds to a *random walk chain*, and $q(y, z) = f(z)$ with a p.d.f. f, which corresponds to an *independence chain* and is closely related to the importance sampling discussed earlier.

Although the MCMC methods have been used over the last 40 years, the research on the theory of MCMC is still very active. Important topics include the choice of the transition kernel for MCMC; the rate of the convergence in (4.13); the choice of the Monte Carlo size m; and the estimation of the errors due to Monte Carlo. See more results and discussions in Tierney (1994), Basag et al. (1995), Tanner (1996), and their references.

4.2 Invariance

The concept of invariance is introduced in §2.3.2 (Definition 2.9). In this section, we study the best invariant estimators and their properties in one-parameter location families (§4.2.1), in one-parameter scale families (§4.2.2), and in general location-scale families (§4.2.3).

4.2.1 One-parameter location families

Assume that the sample $X = (X_1, ..., X_n)$ has a joint distribution P_μ with a Lebesgue p.d.f.

$$f(x_1 - \mu, ..., x_n - \mu), \tag{4.17}$$

where f is known and $\mu \in \mathcal{R}$ is an unknown location parameter. The p.d.f. in (4.17) is a special case of the general location-scale family in Definition 2.3. The family $\mathcal{P} = \{P_\mu : \mu \in \mathcal{R}\}$ is called a one-parameter location family and is invariant under the location transformations $g_c(X) = (X_1 + c,..., X_n + c)$, $c \in \mathcal{R}$.

We consider the estimation of μ as a statistical decision problem with action space $\mathbb{A} = \mathcal{R}$ and loss function $L(\mu, a)$. It is natural to consider the same transformation in the action space, i.e., if X_i is transformed to $X_i + c$, then our action a is transformed to $a + c$. Consequently, the decision problem is invariant under location transformation if and only if

$$L(\mu, a) = L(\mu + c, a + c) \qquad \text{for all } c \in \mathcal{R},$$

which is equivalent to

$$L(\mu, a) = L(a - \mu) \tag{4.18}$$

for a Borel function $L(\cdot)$ on $\mathcal{R}$.

According to Definition 2.9 (see also Example 2.24), an estimator T (decision rule) of μ is *location invariant* if and only if

$$T(X_1 + c, ..., X_n + c) = T(X_1, ..., X_n) + c. \tag{4.19}$$

Many estimators of μ, such as the sample mean and weighted average of the order statistics, are location invariant. Let $d_i = x_i - x_n$, $D_i = X_i - X_n$, $d = (d_1, ..., d_{n-1})$, and $D = (D_1, ..., D_{n-1})$. The following result provides a characterization of location invariant estimators.

Proposition 4.3. Let T_0 be a location invariant estimator of μ. A necessary and sufficient condition for an estimator T to be location invariant is that there exists a Borel function u on $\mathcal{R}^{n-1}$ ($u \equiv$ a constant if $n = 1$) such that

$$T(x) = T_0(x) - u(d) \qquad \text{for all } x \in \mathcal{R}^n. \tag{4.20}$$

Proof. It is easy to see that T given by (4.20) satisfies (4.19) and, therefore, is location invariant. Suppose that T is location invariant. Let $\tilde{u}(x) = T(x) - T_0(x)$ for any $x \in \mathcal{R}^n$. Then

$$\begin{aligned}\tilde{u}(x_1 + c, ..., x_n + c) &= T(x_1 + c, ..., x_n + c) - T_0(x_1 + c, ..., x_n + c) \\ &= T(x_1, ..., x_n) - T_0(x_1, ..., x_n) \\ &= \tilde{u}(x_1, ..., x_n)\end{aligned}$$

for all $c \in \mathcal{R}$ and $x_i \in \mathcal{R}$. Putting $c = -x_n$ leads to

$$\begin{aligned}u(d_1, ..., d_{n-1}) &= \tilde{u}(x_1 - x_n, ..., x_{n-1} - x_n, 0) \\ &= \tilde{u}(x_1, ..., x_n) \\ &= T(x) - T_0(x)\end{aligned}$$

for all $x \in \mathcal{R}^n$. This proves the result. ∎

The next result states an important property of location invariant estimators.

Proposition 4.4. Let X be distributed with the p.d.f. given by (4.17) and let T be a location invariant estimator of μ under the loss function given by (4.18). If the bias, variance, and risk of T are well defined, then they are all constant (do not depend on μ).
Proof. The result for the bias follows from

$$\begin{aligned}b_T(\mu) &= \int T(x) f(x_1 - \mu, ..., x_n - \mu) dx - \mu \\ &= \int T(x_1 + \mu, ..., x_n + \mu) f(x) dx - \mu \\ &= \int [T(x) + \mu] f(x) dx - \mu \\ &= \int T(x) f(x) dx.\end{aligned}$$

The proof of the result for variance or risk is left as an exercise. ∎

An important consequence of this result is that the problem of finding the best location invariant estimator reduces to comparing constants instead of risk functions. The following definition can be used not only for location invariant estimators, but also for general invariant estimators.

Definition 4.2. Consider an invariant estimation problem in which all invariant estimators have constant risks. An invariant estimator T is called the *minimum risk invariant estimator* (MRIE) if and only if T has the smallest risk among all invariant estimators. ∎

Theorem 4.5. Let X be distributed with the p.d.f. given by (4.17) and consider the estimation of μ under the loss function given by (4.18). Suppose that there is a location invariant estimator T_0 of μ with finite risk.
(i) Assume that for each d there exists a $u_*(d)$ that minimizes

$$E_0[L(T_0(X) - u(d))|D = d]$$

over all functions u, where the conditional expectation E_0 is calculated under the assumption that X has p.d.f. $f(x_1, ..., x_n)$. Then an MRIE exists and is given by

$$T_*(X) = T_0(X) - u_*(D).$$

(ii) The function u_* in (i) exists if $L(t)$ is convex and not monotone; it is unique if L is strictly convex.
(iii) If T_0 and D are independent, then u_* is a constant that minimizes $E_0[L(T_0(X) - u)]$. If, in addition, the distribution of T_0 is symmetric about μ and L is convex and even, then $u_* = 0$.
Proof. By Propositions 4.3 and 4.4,

$$R_T(\mu) = E\{E_0[L(T_0(x) - u(d))|D = d]\},$$

where $T(X) = T_0(X) - u(D)$. This proves part (i). If L is (strictly) convex and not monotone, then $E_0[L(T_0(x) - a)|D = d]$ is (strictly) convex and not monotone (exercise). Hence $\lim_{|a|\to\infty} E_0[L(T_0(x) - a)|D = d] = \infty$. This proves part (ii). The proof of part (iii) is left as an exercise. ∎

Theorem 4.6. Assume the conditions of Theorem 4.5 and that the loss is the squared error loss.
(i) The unique MRIE of μ is

$$T_*(X) = \frac{\int_{-\infty}^{\infty} tf(X_1 - t, ..., X_n - t)dt}{\int_{-\infty}^{\infty} f(X_1 - t, ..., X_n - t)dt},$$

which is known as the *Pitman estimator* of μ.
(ii) The MRIE is unbiased.
Proof. (i) Under the squared error loss,

$$u_*(d) = E_0[T_0(X)|D = d] \tag{4.21}$$

(exercise). Let $T_0(X) = X_n$ (the nth observation). Then X_n is location invariant. If there exists a location invariant estimator of μ with finite risk, then $E_0(X_n|D = d)$ is finite a.s. $\mathcal{P}$ (exercise). By Proposition 1.8, when $\mu = 0$, the joint Lebesgue p.d.f. of (D, X_n) is $f(d_1 + x_n, ..., d_{n-1} + x_n, x_n)$. The conditional p.d.f. of X_n given $D = d$ is then

$$\frac{f(d_1 + x_n, ..., d_{n-1} + x_n, x_n)}{\int_{-\infty}^{\infty} f(d_1 + t, ..., d_{n-1} + t, t)dt}$$

(see (1.39)). By Proposition 1.11,

$$\begin{aligned} E_0(X_n|D=d) &= \frac{\int_{-\infty}^{\infty} tf(d_1+t,...,d_{n-1}+t,t)dt}{\int_{-\infty}^{\infty} f(d_1+t,...,d_{n-1}+t,t)dt} \\ &= \frac{\int_{-\infty}^{\infty} tf(x_1-x_n+t,...,x_{n-1}-x_n+t,t)dt}{\int_{-\infty}^{\infty} f(x_1-x_n+t,...,x_{n-1}-x_n+t,t)dt} \\ &= x_n - \frac{\int_{-\infty}^{\infty} uf(x_1-u,...,x_n-u)du}{\int_{-\infty}^{\infty} f(x_1-u,...,x_n-u)du} \end{aligned}$$

by letting $u = x_n - t$. The result in (i) follows from $T_*(X) = X_n - E(X_n|D)$ (Theorem 4.5).
(ii) Let b be the constant bias of T_* (Proposition 4.4). Then $T_1(X) = T_*(X) - b$ is a location invariant estimator of μ and

$$R_{T_1} = E[T_*(X) - b - \mu]^2 = \mathrm{Var}(T_*) \leq \mathrm{Var}(T_*) + b^2 = R_{T_*}.$$

Since T_* is the MRIE, $b = 0$, i.e., T_* is unbiased. ∎

Theorem 4.6(ii) indicates that we only need to consider the unbiased location invariant estimator in order to find the MRIE, if the loss is the squared error loss. In particular, a location invariant UMVUE is an MRIE.

Example 4.11. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with an unknown $\mu \in \mathcal{R}$ and a known σ^2. Note that $\bar{X}$ is location invariant. Since $\bar{X}$ is the UMVUE of μ (§2.1), it is the MRIE under the squared error loss. Since the distribution of $\bar{X}$ is symmetric about μ and $\bar{X}$ is independent of D (Basu's theorem), it follows from Theorem 4.5(iii) that $\bar{X}$ is an MRIE if L is convex and even. ∎

Example 4.12. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(\mu, \theta)$, where θ is known and $\mu \in \mathcal{R}$ is unknown. Since $X_{(1)} - \theta/n$ is location invariant and is the UMVUE of μ, it is the MRIE under the squared error loss. Note that $X_{(1)}$ is independent of D (Basu's theorem). By Theorem 4.4(iii), an MRIE is of the form $X_{(1)} - u_*$ with a constant u_*. For the absolute error loss, $X_{(1)} - \theta \log 2/n$ is an MRIE (exercise). ∎

Example 4.13. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on $(\mu - \frac{1}{2}, \mu + \frac{1}{2})$ with an unknown $\mu \in \mathcal{R}$. Consider the squared error loss. Note that

$$f(x_1-\mu,...,x_n-\mu) = \begin{cases} 1 & \mu - \frac{1}{2} \leq x_{(1)} \leq x_{(n)} \leq \mu + \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

By Theorem 4.6(i), the MRIE of μ is

$$T_*(X) = \int_{X_{(n)}-\frac{1}{2}}^{X_{(1)}+\frac{1}{2}} t dt \Big/ \int_{X_{(n)}-\frac{1}{2}}^{X_{(1)}+\frac{1}{2}} dt = \frac{X_{(1)}+X_{(n)}}{2}. \quad \blacksquare$$

We end this section with a brief discussion of the admissibility of MRIE's in a one-parameter location problem. Under the squared error loss, the MRIE (Pitman's estimator) is admissible if there exists a location invariant estimator T_0 with $E|T_0(X)|^3 < \infty$ (Stein, 1959). Under a general loss function, an MRIE is admissible when it is a unique MRIE (under some other minor conditions). See Farrell (1964), Brown (1966), and Brown and Fox (1974) for further discussions.

4.2.2 One-parameter scale families

Assume that the sample $X = (X_1, ..., X_n)$ has a joint distribution P_σ with a Lebesgue p.d.f.

$$\tfrac{1}{\sigma^n} f\left(\tfrac{x_1}{\sigma}, ..., \tfrac{x_n}{\sigma}\right), \tag{4.22}$$

where f is known and $\sigma > 0$ is an unknown scale parameter. The family $\mathcal{P} = \{P_\sigma : \sigma > 0\}$ is called a one-parameter scale family and is a special case of the general location-scale family in Definition 2.3. This family is invariant under the scale transformations $g_r(X) = rX$, $r > 0$.

We consider the estimation of σ^h with $\mathbb{A} = [0, \infty)$, where h is a nonzero constant. The transformation g_r induces the transformation $g_r(\sigma^h) = r^h\sigma^h$. Hence a loss function L is scale invariant if and only if

$$L(r\sigma, r^h a) = L(\sigma, a) \qquad \text{for all } r > 0,$$

which is equivalent to

$$L(\sigma, a) = L\left(\tfrac{a}{\sigma^h}\right) \tag{4.23}$$

for a Borel function $L(\cdot)$ on $[0, \infty)$. An example of a loss function satisfying (4.23) is

$$L(\sigma, a) = \left|\frac{a}{\sigma^h} - 1\right|^p = \frac{|a - \sigma^h|^p}{\sigma^{ph}}, \tag{4.24}$$

where $p \geq 1$ is a constant. However, the squared error loss does not satisfy (4.23).

An estimator T of σ^h is *scale invariant* if and only if

$$T(rX_1, ..., rX_n) = r^h T(X_1, ..., X_n).$$

Examples of scale invariant estimators are the sample variance S^2 (for $h = 2$), the sample standard deviation $S = \sqrt{S^2}$ (for $h = 1$), the sample range

$X_{(n)} - X_{(1)}$ (for $h = 1$), and the sample mean deviation $n^{-1}\sum_{i=1}^{n}|X_i - \bar{X}|$ (for $h = 1$).

The following result is an analogue of Proposition 4.3. Its proof is left as an exercise.

Proposition 4.5. Let T_0 be a scale invariant estimator of σ^h. A necessary and sufficient condition for an estimator T to be scale invariant is that there exists a positive Borel function u on $\mathcal{R}^n$ such that

$$T(x) = T_0(x)/u(z) \qquad \text{for all } x \in \mathcal{R}^n,$$

where $z = (z_1, ..., z_n)$, $z_i = x_i/x_n$, $i = 1, ..., n-1$, and $z_n = x_n/|x_n|$. ∎

The next result is similar to Proposition 4.4. It applies to any invariant problem defined in Definition 2.9. We use the notation in Definition 2.9.

Theorem 4.7. Let $\mathcal{P}$ be a location-scale family invariant for given $\mathcal{C}$ and $\mathcal{T}$. Suppose that the loss function is invariant and T is an invariant decision rule. Then the risk function of T is a constant. ∎

The proof is left as an exercise. Note that a special case of Theorem 4.7 is that any scale invariant estimator of σ^h has a constant risk and, therefore, an MRIE (Definition 4.2) of σ^h usually exists. However, Proposition 4.4 is not a special case of Theorem 4.7, since the bias of T may not be a constant in general. For example, the bias of the sample standard deviation is a function of σ.

The next result and its proof are analogues of those of Theorem 4.5.

Theorem 4.8. Let X be distributed with the p.d.f. given by (4.22) and consider the estimation of σ^h under the loss function given by (4.23). Suppose that there is a scale invariant estimator T_0 of σ^h with finite risk.
(i) Assume that for each z there exists a $u_*(z)$ that minimizes

$$E_1[L(T_0(X)/u(z))|Z = z]$$

over all positive functions u, where the conditional expectation E_1 is calculated under the assumption that X has p.d.f. $f(x_1, ..., x_n)$. Then an MRIE exists and is given by

$$T_*(X) = T_0(X)/u_*(Z).$$

(ii) The function u_* in (i) exists if $\gamma(t) = L(e^t)$ is convex and not monotone; it is unique if $\gamma(t)$ is strictly convex. ∎

The loss function given by (4.24) satisfies the condition in Theorem 4.8(ii). A loss function corresponding to the squared error loss in this

problem is the loss function (4.24) with $p = 2$. We have the following result similar to Theorem 4.6 (its proof is left as an exercise).

Corollary 4.1. Under the conditions of Theorem 4.8 and the loss function (4.24) with $p = 2$, the unique MRIE of σ^h is

$$T_*(X) = \frac{T_0(X)E_1[T_0(X)|Z]}{E_1\{[T_0(X)]^2|Z\}} = \frac{\int_0^\infty t^{n+h-1}f(tX_1, ..., tX_n)dt}{\int_0^\infty t^{n+2h-1}f(tX_1, ..., tX_n)dt},$$

which is known as the Pitman estimator of σ^h. ∎

Example 4.14. Let $X_1, ..., X_n$ be i.i.d. from $N(0, \sigma^2)$ and consider the estimation of σ^2. Then $T_0 = \sum_{i=1}^n X_i^2$ is scale invariant. By Basu's theorem, T_0 is independent of Z. Hence u_* in Theorem 4.8 is a constant minimizing $E_1[L(T_0/u)]$ over $u > 0$. When the loss is given by (4.24) with $p = 2$, by Corollary 4.1, the MRIE (Pitman's estimator) is

$$T_*(X) = \frac{T_0(X)E_1[T_0(X)]}{E_1[T_0(X)]^2} = \frac{1}{n+2}\sum_{i=1}^n X_i^2,$$

since T_0 has a chi-square distribution χ_n^2 when $\sigma = 1$. Note that the UMVUE of σ^2 is T_0/n, which is different from the MRIE. ∎

Example 4.15. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on $(0, \sigma)$ and consider the estimation of σ. By Basu's theorem, the scale invariant estimator $X_{(n)}$ is independent of Z. Hence u_* in Theorem 4.8 is a constant minimizing $E_1[L(X_{(n)}/u)]$ over $u > 0$. When the loss is given by (4.24) with $p = 2$, by Corollary 4.1, the MRIE (Pitman's estimator) is

$$T_*(X) = \frac{X_{(n)}E_1X_{(n)}}{E_1X_{(n)}^2} = \frac{(n+2)X_{(n)}}{n+1}.$$

4.2.3 General location-scale families

Assume that $X = (X_1, ..., X_n)$ has a joint distribution P_θ with a Lebesgue p.d.f.

$$\frac{1}{\sigma^n} f\left(\frac{x_1-\mu}{\sigma}, ..., \frac{x_n-\mu}{\sigma}\right), \tag{4.25}$$

where f is known, $\theta = (\mu, \sigma) \in \Theta$, and $\Theta = \mathcal{R} \times (0, \infty)$. The family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a special case of the location-scale family defined in Definition 2.3 and is invariant under the location-scale transformations $g_{c,r}(X) = (rX_1 + c, ..., rX_n + c)$, $c \in \mathcal{R}$, $r > 0$, which induce similar transformations on Θ: $g_{c,r}(\theta) = (r\mu + c, r\sigma)$, $c \in \mathcal{R}$, $r > 0$.

Consider the estimation of σ^h with a fixed $h \neq 0$ under the loss function (4.23), which is invariant under the location-scale transformations $g_{c,r}$. An estimator T of σ^h is *location-scale* invariant if and only if

$$T(rX_1 + c, ..., rX_n + c) = r^h T(X_1, ..., X_n). \tag{4.26}$$

By Theorem 4.7, any location-scale invariant T has a constant risk. Letting $r = 1$ in (4.26), we obtain that

$$T(X_1 + c, ..., X_n + c) = T(X_1, ..., X_n)$$

for all $c \in \mathcal{R}$. Therefore, T is a function of $D = (D_1, ..., D_{n-1})$, $D_i = X_i - X_n$, $i = 1, ..., n-1$. From (4.25), the joint Lebesgue p.d.f. of D is

$$\tfrac{1}{\sigma^{n-1}} \int_{-\infty}^{\infty} f\left(\tfrac{d_1}{\sigma} + t, ..., \tfrac{d_{n-1}}{\sigma} + t, t\right) dt, \tag{4.27}$$

which is of the form (4.22) with n replaced by $n-1$ and x_i's replaced by d_i's. It follows from Theorem 4.8 that if $T_0(D)$ is any finite risk scale invariant estimator of σ^h based on D, then an MRIE of σ^h is

$$T_*(D) = T_0(D)/u_*(W), \tag{4.28}$$

where $W = (W_1, ..., W_{n-1})$, $W_i = D_i/D_{n-1}$, $i = 1, ..., n-2$, $W_{n-1} = D_{n-1}/|D_{n-1}|$, $u_*(w)$ is any number minimizing $\tilde{E}_1[L(T_0(D)/u(w))|W = w]$ over all positive functions u, and $\tilde{E}_1$ is the conditional expectation calculated under the assumption that D has p.d.f. (4.27) with $\sigma = 1$.

Consider next the estimation of μ. Under the location-scale transformation $g_{c,r}$, it can be shown (exercise) that a loss function is invariant if and only if it is of the form

$$L\left(\tfrac{a-\mu}{\sigma}\right). \tag{4.29}$$

An estimator T of μ is location-scale invariant if and only if

$$T(rX_1 + c, ..., rX_n + c) = rT(X_1, ..., X_n) + c.$$

Again, by Theorem 4.7, the risk of an invariant T is a constant.

The following result is an analogue of Proposition 4.3 or 4.5.

Proposition 4.6. Let T_0 be any estimator of μ invariant under location-scale transformation and let T_1 be any estimator of σ satisfying (4.26) with $h = 1$ and $T_1 > 0$. Then an estimator T of μ is location-scale invariant if and only if there is a Borel function u on $\mathcal{R}^{n-1}$ such that

$$T(X) = T_0(X) - u(W)T_1(X),$$

where W is given in (4.28). ∎

The proofs of Proposition 4.6 and the next result, an analogue of Theorem 4.4 or 4.7, are left as exercises.

Theorem 4.9. Let X be distributed with p.d.f. given by (4.25) and consider the estimation of μ under the loss function given by (4.29). Suppose that there exist finite risk location-scale invariant estimators T_0 of μ and T_1 of σ. Then an MRIE of μ is

$$T_*(X) = T_0(X) - u_*(W)T_1(X),$$

where W is given in (4.28), $u_*(w)$ is any number minimizing

$$E_{0,1}[L(T_0(X) - u(w)T_1(X))|W = w]$$

over all functions u, and $E_{0,1}$ is computed under the assumption that X has the p.d.f. (4.25) with $\mu = 0$ and $\sigma = 1$. ∎

Corollary 4.2. Under the conditions of Theorem 4.9 and the loss function $(a-\mu)^2/\sigma^2$, $u_*(w)$ in Theorem 4.9 is equal to

$$u_*(w) = \frac{E_{0,1}[T_0(X)T_1(X)|W = w]}{E_{0,1}\{[T_1(X)]^2|W = w\}}. \quad ∎$$

Example 4.16. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$, where $\mu \in \mathcal{R}$ and $\sigma^2 > 0$ are unknown. Consider first the estimation of σ^2 under loss function (4.23). The sample variance S^2 is location-scale invariant and is independent of W in (4.28) (Basu's theorem). Thus, by (4.28), S^2/u_* is an MRIE, where u_* is a constant minimizing $\tilde{E}_1[L(S^2/u)]$ over all $u > 0$. If the loss function is given by (4.24) with $p = 2$, then by Corollary 4.1, the MRIE of σ^2 is

$$T_*(X) = \frac{S^2\tilde{E}_1(S^2)}{\tilde{E}_1(S^2)^2} = \frac{S^2}{(n^2-1)/(n-1)^2} = \frac{1}{n+1}\sum_{i=1}^n (X_i - \bar{X})^2,$$

since $(n-1)S^2$ has a chi-square distribution χ^2_{n-1} when $\sigma = 1$.

Next, consider the estimation of μ under the loss function (4.29). Since $\bar{X}$ is a location-scale invariant estimator of μ and is independent of W in (4.28) (Basu's theorem), by Theorem 4.9, an MRIE of μ is

$$T_*(X) = \bar{X} - u_* S^2,$$

where u_* is a constant. If L in (4.29) is convex and even, then $u_* = 0$ (see Theorem 4.5(iii)) and, hence, $\bar{X}$ is an MRIE of μ. ∎

Example 4.17. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on $(\mu - \frac{1}{2}\sigma, \mu + \frac{1}{2}\sigma)$, where $\mu \in \mathcal{R}$ and $\sigma > 0$ are unknown. Consider first the

estimation of σ under the loss function (4.24) with $p = 2$. The sample range $X_{(n)} - X_{(1)}$ is a location-scale invariant estimator of σ and is independent of W in (4.28) (Basu's theorem). By (4.28) and Corollary 4.1, the MRIE of σ is

$$T_*(X) = \frac{(X_{(n)} - X_{(1)})\tilde{E}_1(X_{(n)} - X_{(1)})}{\tilde{E}_1(X_{(n)} - X_{(1)})^2} = \frac{(n+2)(X_{(n)} - X_{(1)})}{n}.$$

Consider now the estimation of μ under the loss function (4.29). Since $(X_{(1)} + X_{(n)})/2$ is a location-scale invariant estimator of μ and is independent of W in (4.28) (Basu's theorem), by Theorem 4.9, an MRIE of μ is

$$T_*(X) = \frac{X_{(1)} + X_{(n)}}{2} - u_*(X_{(n)} - X_{(1)}),$$

where u_* is a constant. If L in (4.29) is convex and even, then $u_* = 0$ (see Theorem 4.5(iii)) and, hence, $(X_{(1)} + X_{(n)})/2$ is an MRIE of μ. ∎

Finding MRIE's in various subfamilies of the location-scale family in Definition 2.3 under transformations $XA + c$, where $A \in \mathcal{T}$ and $c \in \mathcal{C}$ with given $\mathcal{T}$ and $\mathcal{C}$, can be done in a similar way. We only provide some brief discussions for two important cases. The first case is the two-sample location-scale problem in which two samples, $X = (X_1, ..., X_m)$ and $Y = (Y_1, ..., Y_n)$, are taken from a distribution with Lebesgue p.d.f.

$$\frac{1}{\sigma_x^m \sigma_y^n} f\left(\frac{x_1 - \mu_x}{\sigma_x}, ..., \frac{x_m - \mu_x}{\sigma_x}, \frac{y_1 - \mu_y}{\sigma_y}, ..., \frac{y_n - \mu_y}{\sigma_y}\right), \tag{4.30}$$

where f is known, $\mu_x \in \mathcal{R}$ and $\mu_y \in \mathcal{R}$ are unknown location parameters, and $\sigma_x > 0$ and $\sigma_y > 0$ are unknown scale parameters. The family of distributions is invariant under the transformations

$$g(X, Y) = (rX_1 + c, ..., rX_m + c, r'Y_1 + c', ..., r'Y_n + c'), \tag{4.31}$$

where $r > 0$, $r' > 0$, $c \in \mathcal{R}$, and $c' \in \mathcal{R}$. The parameters to be estimated in this problem are usually $\Delta = \mu_y - \mu_x$ and $\eta = (\sigma_y/\sigma_x)^h$ with a fixed $h \neq 0$. If X and Y are from two populations, Δ and η are measures of the difference between the two populations. For estimating η, results similar to those in this section can be established. For estimating Δ, MRIE's can be obtained under some conditions. See Exercises 54-56.

The second case is the general linear model (3.25) under the assumption that ε_i's are i.i.d. with the p.d.f. $\sigma^{-1} f(x/\sigma)$, where f is a known Lebesgue p.d.f. The family of populations is invariant under the transformations

$$g(X) = rX + cZ^\tau, \qquad r \in (0, \infty),\ c \in \mathcal{R}^p \tag{4.32}$$

(exercise). The estimation of βl^τ with $l \in \mathcal{R}(Z)$ is invariant under the loss function $L\left(\frac{a-\beta l^\tau}{\sigma}\right)$ and the LSE $\hat{\beta} l^\tau$ is an invariant estimator of βl^τ (exercise). When f is normal, the following result can be established using an argument similar to that in Example 4.16.

Theorem 4.10. Consider model (3.25) with assumption A1.
(i) Under transformations (4.32) and the loss function $L\left(\frac{a-\beta l^\tau}{\sigma}\right)$, where L is convex and even, the LSE $\hat{\beta} l^\tau$ is an MRIE of βl^τ for any $l \in \mathcal{R}(Z)$.
(ii) Under transformations (4.32) and the loss function $(a-\sigma^2)^2/\sigma^4$, the MRIE of σ^2 is $SSR/(n-q+2)$, where SSR is given by (3.36) and q is the rank of Z. ∎

MRIE's in a parametric family with a multi-dimensional θ are often inadmissible. See Lehmann (1983, p. 285) for more discussions.

4.3 Minimaxity and Admissibility

Consider the estimation of a real-valued $\vartheta = g(\theta)$ based on a sample X from P_θ, $\theta \in \Theta$, under a given loss function. A *minimax estimator* minimizes the maximum risk $\sup_{\theta\in\Theta} R_T(\theta)$ over all estimators T (see §2.3.2).

A unique minimax estimator is admissible, since any estimator better than a minimax estimator is also minimax. This indicates that we should consider minimaxity and admissibility together. The situation is different for a UMVUE (or an MRIE), since if a UMVUE (or an MRIE) is inadmissible, it is dominated by an estimator that is not unbiased (or invariant).

4.3.1 Estimators with constant risks

By minimizing the maximum risk, a minimax estimator tries to do as well as possible in the worst case. Such an estimator can be very unsatisfactory. However, if a minimax estimator has some other good properties (e.g., it is a Bayes estimator), then it is often a reasonable estimator. Here we study when estimators having constant risks (e.g., MRIE's) are minimax.

Theorem 4.11. Let Π be a proper prior on Θ and δ be a Bayes estimator of ϑ w.r.t. Π. Let $\Theta_\Pi = \{\theta : R_\delta(\theta) = \sup_{\theta\in\Theta} R_\delta(\theta)\}$. If $\Pi(\Theta_\Pi) = 1$, then δ is minimax. If, in addition, δ is the unique Bayes estimator w.r.t. Π, then it is the unique minimax estimator.
Proof. Let T be any other estimator of ϑ. Then

$$\sup_{\theta\in\Theta} R_T(\theta) \geq \int_{\Theta_\Pi} R_T(\theta) d\Pi \geq \int_{\Theta_\Pi} R_\delta(\theta) d\Pi = \sup_{\theta\in\Theta} R_\delta(\theta).$$

If δ is the unique Bayes estimator, then the second inequality in the previous expression should be replaced by $>$ and, therefore, δ is the unique minimax estimator. ∎

The condition of Theorem 4.11 essentially means that δ has a constant risk. Thus, a Bayes estimator having constant risk is minimax.

Example 4.18. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $P(X_1 = 1) = p \in (0, 1)$. Consider the estimation of p under the squared error loss. The UMVUE $\bar{X}$ has risk $p(1-p)/n$ which is not constant. In fact, $\bar{X}$ is not minimax (Exercise 58). To find a minimax estimator by applying Theorem 4.11, we consider the Bayes estimator w.r.t. the beta distribution $B(\alpha, \beta)$ with known α and β (Exercise 1):

$$\delta(X) = (\alpha + n\bar{X})/(\alpha + \beta + n).$$

A straightforward calculation shows that

$$R_\delta(p) = [np(1-p) + (\alpha - \alpha p - \beta p)^2]/(\alpha + \beta + n)^2.$$

To apply Theorem 4.11, we need to find values of $\alpha > 0$ and $\beta > 0$ such that $R_\delta(p)$ is constant. It can be shown that $R_\delta(p)$ is constant if and only if $\alpha = \beta = \sqrt{n}/2$, which leads to the unique minimax estimator

$$T(X) = (n\bar{X} + \sqrt{n}/2)/(n + \sqrt{n}).$$

The risk of T is $R_T = 1/[4(1 + \sqrt{n})^2]$.

Note that T is a Bayes estimator and has some good properties. Comparing the risk of T with that of $\bar{X}$, we find that T has smaller risk if and only if

$$p \in \left(\tfrac{1}{2} - \tfrac{1}{2}\sqrt{1 - \tfrac{n}{(1+\sqrt{n})^2}},\ \tfrac{1}{2} + \tfrac{1}{2}\sqrt{1 - \tfrac{n}{(1+\sqrt{n})^2}} \right). \tag{4.33}$$

Thus, for small value of n, T is better (and can be much better) than $\bar{X}$ for most of the range of p (Figure 4.1). When $n \to \infty$, the interval in (4.33) shrinks toward $\frac{1}{2}$. Hence, for large (and even moderate) n, $\bar{X}$ is better than T for most of the range of p (Figure 4.1). The limit of the asymptotic relative efficiency of T w.r.t. $\bar{X}$ is $4p(1-p)$, which is always smaller than 1 when $p \neq \frac{1}{2}$ and equals 1 when $p = \frac{1}{2}$.

The minimax estimator depends strongly on the loss function. To see this, let us consider the loss function $L(p, a) = (a-p)^2/[p(1-p)]$. Under this loss function, $\bar{X}$ has constant risk and is the unique Bayes estimator w.r.t. the uniform prior on $(0, 1)$. By Theorem 4.11, $\bar{X}$ is the unique minimax estimator. On the other hand, the risk of T is equal to $1/[4(1+\sqrt{n})^2 p(1-p)]$, which is unbounded. ∎

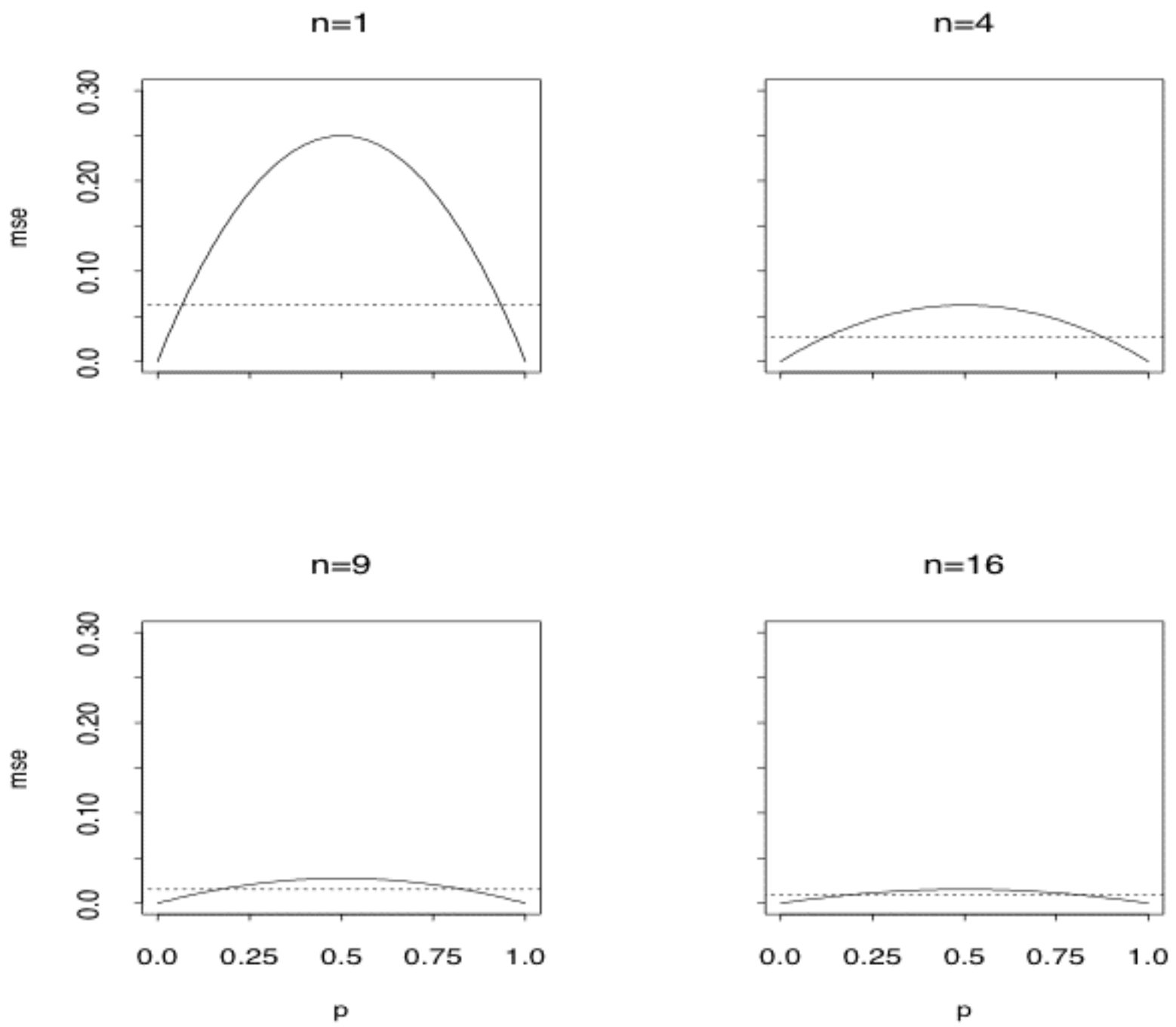

Figure 4.1: mse's of $\bar{X}$ (curve) and $T(X)$ (straightline) in Example 4.18

In many cases a constant risk estimator is not a Bayes estimator (e.g., an unbiased estimator under the squared error loss), but a limit of Bayes estimators w.r.t. a sequence of priors. Then the following result may be used to find a minimax estimator.

Theorem 4.12. Let Π_j, $j = 1, 2, ...$, be a sequence of priors and r_j be the Bayes risk of a Bayes estimator of ϑ w.r.t. Π_j. Let T be a constant risk estimator of ϑ. If $\lim_{j\to\infty} r_j = R_T$, then T is minimax. ∎

The proof of this theorem is similar to that of Theorem 4.11. Although Theorem 4.12 is more general than Theorem 4.11 in finding minimax estimators, it does not provide uniqueness of the minimax estimator even when there is a unique Bayes estimator w.r.t. each Π_j.

In Example 2.25, we actually applied the result in Theorem 4.12 to show the minimaxity of $\bar{X}$ as an estimator of $\mu = EX_1$ when $X_1, ..., X_n$ are i.i.d. from a normal distribution with a known $\sigma^2 = \mathrm{Var}(X_1)$, under the squared error loss. To discuss the minimaxity of $\bar{X}$ in the case where σ^2 is unknown, we need the following lemma.

Lemma 4.3. Let Θ_0 be a subset of Θ and T be a minimax estimator of ϑ when Θ_0 is the parameter space. Then T is a minimax estimator if

$$\sup_{\theta\in\Theta} R_T(\theta) = \sup_{\theta\in\Theta_0} R_T(\theta).$$

Proof. If there is an estimator T_0 with $\sup_{\theta\in\Theta} R_{T_0}(\theta) < \sup_{\theta\in\Theta} R_T(\theta)$, then

$$\sup_{\theta\in\Theta_0} R_{T_0}(\theta) \le \sup_{\theta\in\Theta} R_{T_0}(\theta) < \sup_{\theta\in\Theta} R_T(\theta) = \sup_{\theta\in\Theta_0} R_T(\theta),$$

which contradicts the minimaxity of T when Θ_0 is the parameter space. Hence, T is minimax when Θ is the parameter space. ∎

Example 4.19. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with unknown $\theta = (\mu, \sigma^2)$. Consider the estimation of μ under the squared error loss. Suppose first that $\Theta = \mathcal{R} \times (0, c]$ with a constant $c > 0$. Let $\Theta_0 = \mathcal{R} \times \{c\}$. From Example 2.25, $\bar{X}$ is a minimax estimator of μ when the parameter space is Θ_0. An application of Lemma 4.3 shows that $\bar{X}$ is also minimax when the parameter space is Θ. Although σ^2 is assumed to be bounded by c, the minimax estimator $\bar{X}$ does not depend on c.

Consider next the case where $\Theta = \mathcal{R} \times (0, \infty)$, i.e., σ^2 is unbounded. Let T be any estimator of μ. For any fixed σ^2,

$$\frac{\sigma^2}{n} \le \sup_{\mu\in\mathcal{R}} R_T(\theta),$$

since σ^2/n is the risk of $\bar{X}$ which is minimax when σ^2 is known (Example 2.25). Letting $\sigma^2 \to \infty$, we obtain that $\sup_\theta R_T(\theta) = \infty$ for any estimator T. Thus, minimaxity is meaningless (any estimator is minimax). ∎

Theorem 4.13. Suppose that T as an estimator of ϑ has constant risk and is admissible. Then T is minimax. If the loss function is strictly convex, then T is the unique minimax estimator.

Proof. By the admissibility of T, if there is another estimator T_0 with $\sup_\theta R_{T_0}(\theta) \le R_T$, then $R_{T_0}(\theta) = R_T$ for all θ. This proves that T is minimax. If the loss function is strictly convex and T_0 is another minimax estimator, then

$$R_{(T+T_0)/2}(\theta) < (R_{T_0} + R_T)/2 = R_T$$

for all θ and, therefore, T is inadmissible. This shows that T is unique if the loss is strictly convex. ∎

Combined with Theorem 4.7, Theorem 4.13 tells us that if an MRIE is admissible, then it is minimax. From the discussion at the end of §4.2.1, MRIE's in one-parameter location families (such as Pitman's estimators) are usually minimax.

4.3.2 Results in one-parameter exponential families

The following result provides a sufficient condition for the admissibility of a class of estimators when the population P_θ is in a one-parameter exponential family. Using this result and Theorem 4.13, we can obtain a class of minimax estimators. The proof of this result is an application of the information inequality introduced in §3.1.3.

Theorem 4.14. Suppose that X has the p.d.f. $c(\theta)e^{\theta T(x)}$ w.r.t. a measure ν, where $T(x)$ is real-valued and $\theta \in (\theta_-, \theta_+) \subset \mathcal{R}$. Consider the estimation of $\vartheta = E[T(X)]$ under the squared error loss. Let $\lambda \geq 0$ and γ be known constants and let $T_{\lambda,\gamma}(X) = (T + \gamma\lambda)/(1+\lambda)$. Then a sufficient condition for the admissibility of $T_{\lambda,\gamma}$ is that

$$\int_{\theta_0}^{\theta_+} \frac{e^{\gamma\lambda\theta}}{[c(\theta)]^\lambda} d\theta = \int_{\theta_-}^{\theta_0} \frac{e^{\gamma\lambda\theta}}{[c(\theta)]^\lambda} d\theta = \infty, \tag{4.34}$$

where $\theta_0 \in (\theta_-, \theta_+)$.

Proof. From Theorem 2.1, $\vartheta = E[T(X)] = -c'(\theta)/c(\theta)$ and $\frac{d\vartheta}{d\theta} = \mathrm{Var}(T) = I(\theta)$, the Fisher information defined in (3.5). Suppose that there is an estimator δ such that for all θ,

$$R_\delta(\theta) \leq R_{T_{\lambda,\gamma}}(\theta) = [I(\theta) + \lambda^2(\vartheta - \gamma)^2]/(1+\lambda)^2.$$

From the information inequality (3.6),

$$R_\delta(\theta) \geq [b_\delta(\theta)]^2 + [I(\theta) + b'_\delta(\theta)]^2/I(\theta).$$

Let $h(\theta) = b_\delta(\theta) - \lambda(\gamma - \vartheta)/(1+\lambda)$. Then

$$[h(\theta)]^2 - \frac{2\lambda h(\theta)(\vartheta - \gamma) + 2h'(\theta)}{1+\lambda} + \frac{[h'(\theta)]^2}{I(\theta)} \leq 0,$$

which implies

$$[h(\theta)]^2 - \frac{2\lambda h(\theta)(\vartheta - \gamma) + 2h'(\theta)}{1+\lambda} \leq 0. \tag{4.35}$$

Let $a(\theta) = h(\theta)[c(\theta)]^\lambda e^{\gamma\lambda\theta}$. Differentiation of $a(\theta)$ reduces (4.35) to

$$\frac{[a(\theta)]^2 e^{-\gamma\lambda\theta}}{[c(\theta)]^\lambda} + \frac{2a'(\theta)}{1+\lambda} \leq 0. \tag{4.36}$$

Suppose that $a(\theta_0) < 0$ for some $\theta_0 \in (\theta_-, \theta_+)$. From (4.36), $a'(\theta) \leq 0$ for all θ. Hence $a(\theta) < 0$ for all $\theta \geq \theta_0$ and, for $\theta > \theta_0$, (4.36) can be written as

$$\frac{d}{d\theta}\left[\frac{1}{a(\theta)}\right] \geq \frac{(1+\lambda)e^{-\gamma\lambda\theta}}{2[c(\theta)]^\lambda}.$$

Integrating both sides from θ_0 to θ gives

$$\frac{1+\lambda}{2}\int_{\theta_0}^{\theta}\frac{e^{-\gamma\lambda\theta}}{[c(\theta)]^{\lambda}}d\theta \leq \frac{1}{a(\theta)}-\frac{1}{a(\theta_0)} \leq -\frac{1}{a(\theta_0)}.$$

Letting $\theta \to \theta_+$, the left-hand side of the previous expression diverges to ∞ by condition (4.34), which is impossible. This shows that $a(\theta) \geq 0$ for all θ. Similarly, we can show that $a(\theta) \leq 0$ for all θ. Thus, $a(\theta) = 0$ for all θ. This means that $h(\theta) = 0$ for all θ and the equality in (4.35) holds, which implies $R_\delta(\theta) \equiv R_{T_{\lambda,\gamma}}(\theta)$. This proves the admissibility of $T_{\lambda,\gamma}$. ∎

The reason why $T_{\lambda,\gamma}$ is considered is that it is often a Bayes estimator w.r.t. some prior; see, for example, Examples 2.25, 4.1, 4.7, and 4.8. To find minimax estimators, we may use the following result.

Corollary 4.3. Assume that X has the p.d.f. as described in Theorem 4.14 with $\theta_- = -\infty$ and $\theta_+ = \infty$.
(i) As an estimator of $\vartheta = E(T)$, $T(X)$ is admissible under the squared error loss and the loss $(a-\vartheta)^2/\mathrm{Var}(T)$.
(ii) T is the unique minimax estimator of ϑ under the loss $(a-\vartheta)^2/\mathrm{Var}(T)$.
Proof. (i) With $\lambda = 0$, condition (4.34) is clearly satisfied. Hence, Theorem 4.14 applies under the squared error loss. The admissibility of T under the loss $(a-\vartheta)^2/\mathrm{Var}(T)$ follows from the fact that T is admissible under the squared error loss and $\mathrm{Var}(T) \neq 0$.
(ii) This is a consequence of part (i) and Theorem 4.13. ∎

Example 4.20. Let $X_1, ..., X_n$ be i.i.d. from $N(0, \sigma^2)$ with an unknown $\sigma^2 > 0$. Let $Y = \sum_{i=1}^n X_i^2$. From Example 4.14, $Y/(n+2)$ is the MRIE of σ^2 and has constant risk, under the loss $(a-\sigma^2)^2/\sigma^4$. We now apply Theorem 4.14 to show that $Y/(n+2)$ is admissible. Note that the joint p.d.f. of X_i's is of the form $c(\theta)e^{\theta T(x)}$ with $\theta = -n/(4\sigma^2)$, $c(\theta) = (-2\theta/n)^{n/2}$, $T(X) = 2Y/n$, $\theta_- = -\infty$, and $\theta_+ = 0$. By Theorem 4.14, $T_{\lambda,\gamma}(T+\gamma\lambda)/(1+\lambda)$ is admissible under the squared error loss if

$$\int_{-\infty}^{-c} e^{-\gamma\lambda\theta}\left(\frac{-2\theta}{n}\right)^{-n\lambda/2} d\theta = \int_0^c e^{\gamma\lambda\theta}\theta^{-n\lambda/2}d\theta = \infty$$

for some $c > 0$. This means that $T_{\lambda,\gamma}$ is admissible if $\gamma = 0$ and $\lambda = 2/n$, or if $\gamma > 0$ and $\lambda \geq 2/n$. In particular, $2Y/(n+2)$ is admissible for estimating $E(T) = 2E(Y)/n = 2\sigma^2$, under the squared error loss. It is easy to see that $Y/(n+2)$ is then an admissible estimator of σ^2 under the squared error loss and the loss $(a-\sigma^2)^2/\sigma^4$. Hence $Y/(n+2)$ is minimax under the loss $(a-\sigma^2)^2/\sigma^4$.

Note that we cannot apply Corollary 4.3 directly since $\theta_+ = 0$. ∎

Example 4.21. Let $X_1, ..., X_n$ be i.i.d. from the Poisson distribution $P(\theta)$ with an unknown $\theta > 0$. The joint p.d.f. of X_i's w.r.t. the counting measure is $(x_1! \cdots x_n!)^{-1} e^{-n\theta} e^{n\bar{x}\log\theta}$. For $\eta = n\log\theta$, the conditions of Corollary 4.3 are satisfied with $T(X) = \bar{X}$. Since $E(T) = \theta$ and $\mathrm{Var}(T) = \theta/n$, by Corollary 4.3, $\bar{X}$ is the unique minimax estimator of θ under the loss function $(a-\theta)^2/\theta$. ∎

4.3.3 Simultaneous estimation and shrinkage estimators

In this chapter (and most of Chapter 3) we have been focused on the estimation of a real-valued ϑ. The problem of estimating a vector-valued ϑ under the decision theory approach is called *simultaneous estimation*. Many results for the case of a real-valued ϑ can be extended to simultaneous estimation in a straightforward manner.

Let ϑ be a p-vector of parameters (functions of θ) with range $\tilde{\Theta}$. A vector-valued estimator $T(X)$ can be viewed as a decision rule taking values in the action space $\mathbb{A} = \tilde{\Theta}$. Let $L(\theta, a)$ be a given nonnegative loss function on $\Theta \times \mathbb{A}$. A natural generalization of the squared error loss is

$$L(\theta, a) = \|a - \vartheta\|^2 = \sum_{i=1}^{p} (a_i - \vartheta_i)^2, \tag{4.37}$$

where a_i and ϑ_i are the ith components of a and ϑ, respectively.

A vector-valued estimator T is called unbiased if and only if $E(T) = \vartheta$ for all $\theta \in \Theta$. If there is an unbiased estimator of ϑ, then ϑ is called estimable. It can be seen that the result in Theorem 3.1 extends to the case of vector-valued ϑ with any L strictly convex in a. If the loss function is given by (4.37) and T_i is a UMVUE of ϑ_i for each i, then $T = (T_1, ..., T_p)$ is a UMVUE of ϑ. If there is a sufficient and complete statistic $U(X)$ for θ, then by Theorem 2.5 (Rao-Blackwell's theorem), T must be a function of $U(X)$ and is the unique best unbiased estimator of ϑ.

Example 4.22. Consider the general linear model (3.25) with assumption A1 and a full rank Z. Let $\vartheta = \beta$. An unbiased estimator of β is then the LSE $\hat{\beta}$. From the proof of Theorem 3.7, $\hat{\beta}$ is a function of the sufficient and complete statistic for $\theta = (\beta, \sigma^2)$. Hence, $\hat{\beta}$ is the unique best unbiased estimator of ϑ under any strictly convex loss function. In particular, $\hat{\beta}$ is the UMVUE of β under the loss function (4.37). ∎

Next, we consider Bayes estimators of ϑ, which is still defined to be Bayes actions considered as functions of X. Under the loss function (4.37), the Bayes estimator is still given by (4.4) with vector-valued $g(\theta) = \vartheta$.

Example 4.23. Let $X = (X_0, X_1, ..., X_k)$ have the multinomial distribution given in Example 2.7. Consider the estimation of the vector $\theta = (p_0, p_1, ..., p_k)$ under the loss function (4.37), and the Dirichlet prior for θ which has the Lebesgue p.d.f.

$$\frac{\Gamma(\alpha_0, ..., \alpha_k)}{\Gamma(\alpha_0) \cdots \Gamma(\alpha_k)} p_0^{\alpha_0 - 1} \cdots p_k^{\alpha_k - 1} I_A(\theta), \tag{4.38}$$

where α_j's are known positive constants and $A = \{\theta : 0 \leq p_j, \sum_{j=0}^k p_j = 1\}$. It turns out that the Dirichlet prior is conjugate so that the posterior of θ given $X = x$ is also a Dirichlet distribution having the p.d.f. given by (4.38) with α_j replaced by $\alpha_j + x_j$, $j = 0, 1, ..., k$. Thus, the Bayes estimator of θ is $\delta = (\delta_0, \delta_1, ..., \delta_k)$ with

$$\delta_j(X) = \frac{\alpha_j + X_j}{\alpha_0 + \alpha_1 + \cdots + \alpha_k + n}, \qquad j = 0, 1, ..., k. \quad \blacksquare$$

After a suitable class of transformations is defined, the results in §4.2 for invariant estimators and MRIE's are still valid. This is illustrated by the following example.

Example 4.24. Let X be a sample with the Lebesgue p.d.f. $f(x - \theta)$, where f is a known Lebesgue p.d.f. on $\mathcal{R}^p$ with a finite second moment and $\theta \in \mathcal{R}^p$ is an unknown parameter. Consider the estimation of θ under the loss function (4.37). This problem is invariant under the location transformations $g(X) = X + c$, where $c \in \mathcal{R}^p$. Invariant estimators of θ are of the form $X + l$, $l \in \mathcal{R}^p$. It is easy to show that any invariant estimator has constant bias and risk (a generalization of Proposition 4.4) and the MRIE of θ is the unbiased invariant estimator. In particular, if f is the p.d.f. of $N_p(0, I_p)$, then the MRIE is X. $\blacksquare$

The definition of minimax estimators applies without changes.

Example 4.25. Let X be a sample from $N_p(\theta, I_p)$ with an unknown $\theta \in \mathcal{R}^p$. Consider the estimation of θ under the loss function (4.37). A modification of the proof of Theorem 4.12 with independent priors for θ_i's shows that X is a minimax estimator of θ (exercise). $\blacksquare$

Example 4.26. Consider Example 4.23. If we choose $\alpha_0 = \cdots = \alpha_k = \sqrt{n}/(k+1)$, then the Bayes estimator of θ in Example 4.23 has constant risk. Using the same argument in the proof of Theorem 4.11, we can show that this Bayes estimator is minimax. $\blacksquare$

The previous results for simultaneous estimation are fairly straightforward generalizations of those for the case of real-valued ϑ. Results for

admissibility in simultaneous estimation, however, are quite different. A surprising result, due to Stein (1956), is that in estimating the vector mean $\theta = EX$ of a normally distributed p-vector X (Example 4.25), X is inadmissible under the loss function (4.37) when $p \geq 3$, although X is the UMVUE, MRIE (Example 4.24), and minimax estimator (Example 4.25). Since any estimator better than a minimax estimator is also minimax, there exist many (in fact, infinitely many) minimax estimators in Example 4.25 when $p \geq 3$, which is different from the case of $p = 1$ in which X is the unique admissible minimax estimator (Example 4.6 and Theorem 4.13).

We start with the simple case where X is from $N_p(\theta, I_p)$ with an unknown $\theta \in \mathcal{R}^p$. James and Stein (1961) proposed the following class of estimators of $\vartheta = \theta$ having smaller risks than X when the loss is given by (4.37) and $p \geq 3$:

$$\delta_c = c + \left(1 - \frac{p-2}{\|X-c\|^2}\right)(X - c), \tag{4.39}$$

where $c \in \mathcal{R}^p$ is fixed. The choice of c is discussed next and at the end of this section.

Before we prove that δ_c in (4.39) is better than X, we try to motivate δ_c from two viewpoints. First, suppose that it were thought a priori likely, though not certain, that $\theta = c$. Then we might first test a hypothesis $H_0 : \theta = c$ and estimate θ by c if H_0 is accepted and by X otherwise. The best rejection region has the form $\|X - c\|^2 > t$ for some constant $t > 0$ (see Chapter 6) so that we might estimate θ by

$$I_{(t,\infty)}(\|X-c\|^2)X + [1 - I_{(t,\infty)}(\|X-c\|^2)]c.$$

It can been seen that δ_c in (4.39) is a smoothed version of this estimator, since

$$\delta_c = \psi(\|X-c\|^2)X + [1 - \psi(\|X-c\|^2)]c \tag{4.40}$$

for some function ψ. Any estimator having the form of the right-hand side of (4.40) shrinks the observations toward a given point c and, therefore, is called a *shrinkage estimator*.

Next, δ_c in (4.40) can be viewed as an empirical Bayes estimator (§4.1.2). In view of (2.28) in Example 2.25, a Bayes estimator of θ is of the form

$$\delta = (1 - B)X + Bc,$$

where c is the prior mean of θ and B involves prior variances. If $1 - B$ is "estimated" by $\psi(\|X - c\|^2)$, then δ_c is an empirical Bayes estimator.

Theorem 4.15. Suppose that X is from $N_p(\theta, I_p)$ with $p \geq 3$. Then, under the loss function (4.37), the risks of the following estimators of θ,

$$\delta_{c,r} = c + \left[1 - \frac{r(p-2)}{\|X-c\|^2}\right](X - c), \tag{4.41}$$

are given by

$$R_{\delta_{c,r}}(\theta) = p - (2r - r^2)(p-2)^2 E(\|X - c\|^{-2}), \tag{4.42}$$

where $c \in \mathcal{R}^p$ and $r \in \mathcal{R}$ are known.
Proof. Let $Z = X - c$. Then

$$R_{\delta_{c,r}}(\theta) = E\|\delta_{c,r} - E(X)\|^2 = E\left\| \left(1 - \frac{p-2}{\|Z\|^2}\right) Z - E(Z)\right\|^2.$$

Hence, we only need to show the case of $c = 0$. Let $h(\theta) = R_{\delta_{0,r}}(\theta)$, $g(\theta)$ be the right-hand side of (4.42) with $c = 0$, and $\pi(\theta) = (2\pi\alpha)^{-p/2} e^{-\|\theta\|^2/(2\alpha)}$, which is the p.d.f. of $N_p(0, \alpha I_p)$. Note that the distribution of X can be viewed as the conditional distribution of X given $\boldsymbol{\theta} = \theta$, where $\boldsymbol{\theta}$ has the Lebesgue p.d.f. $\pi(\theta)$. Then

$$\begin{aligned}
\int_{\mathcal{R}^p} g(\theta)\pi(\theta)d\theta &= p - (2r - r^2)(p-2)^2 E[E(\|X\|^{-2}|\boldsymbol{\theta})] \\
&= p - (2r - r^2)(p-2)^2 E(\|X\|^{-2}) \\
&= p - (2r - r^2)(p-2)/(\alpha+1),
\end{aligned}$$

where the expectation in the second line of the previous expression is w.r.t. the joint distribution of $(X, \boldsymbol{\theta})$ and the last equality follows from the fact that the marginal distribution of X is $N_p(0, (\alpha+1)I_p)$, $\|X\|^2/(\alpha+1)$ has the chi-square distribution χ^2_p and, therefore, $E(\|X\|^{-2}) = 1/[(p-2)(\alpha+1)]$. Let $B = 1/(\alpha+1)$ and $\hat{B} = r(p-2)/\|X\|^2$. Then

$$\begin{aligned}
\int_{\mathcal{R}^p} h(\theta)\pi(\theta)d\theta &= E\|(1-\hat{B})X - \boldsymbol{\theta}\|^2 \\
&= E\{E[\|(1-\hat{B})X - \boldsymbol{\theta}\|^2|X]\} \\
&= E\{E[\|\boldsymbol{\theta} - E(\boldsymbol{\theta}|X)\|^2|X] \\
&\quad + \|E(\boldsymbol{\theta}|X) - (1-\hat{B})X\|^2\} \\
&= E\{p(1-B) + (\hat{B} - B)^2\|X\|^2\} \\
&= E\{p(1-B) + B^2\|X\|^2 \\
&\quad - 2Br(p-2) + r^2(p-2)^2\|X\|^{-2}\} \\
&= p - (2r - r^2)(p-2)B,
\end{aligned}$$

where the last equality follows from $E\|X\|^{-2} = B/(p-2)$ and $E\|X\|^2 = p/B$. This proves

$$\int_{\mathcal{R}^p} g(\theta)\pi(\theta)d\theta = \int_{\mathcal{R}^p} h(\theta)\pi(\theta)d\theta. \tag{4.43}$$

Note that $h(\theta)$ and $g(\theta)$ are expectations of functions of $\|X\|^2$, $X\theta^\tau$, and $\|\theta\|^2$. Make an orthogonal transformation from X to Y such that $Y_1 = X\theta^\tau/\|\theta\|$, $EY_j = 0$ for $j > 1$, and $\mathrm{Var}(Y) = I_p$. In terms of Y, $h(\theta)$ and $g(\theta)$ are functions of Y_1, $\sum_{j=2}^p Y_j^2$, and $\|\theta\|^2$. Thus, both h and g are functions of $\|\theta\|^2$.

For the family of p.d.f.'s $\{\pi(\theta) : \alpha > 0\}$, $\|\theta\|^2$ is a complete and sufficient "statistic". Hence, (4.43) and the fact that h and g are functions of $\|\theta\|^2$ imply that $h(\|\theta\|^2) = g(\|\theta\|^2)$ a.e. w.r.t. the Lebesgue measure. From Theorem 2.1, both h and g are continuous functions of $\|\theta\|^2$ and, therefore, $h(\theta) = g(\theta)$ for all $\theta \in \mathcal{R}^p$. This completes the proof. ∎

It follows from Theorem 4.15 that the risk of $\delta_{c,r}$ is smaller than that of X (for every value of θ) when $p \geq 3$ and $0 < r < 2$, since the risk of X is p under the loss function (4.37). From Example 4.6, X is admissible when $p = 1$. When $p = 2$, X is still admissible (Stein, 1956). But we have just shown that X is inadmissible when $p \geq 3$.

The James-Stein estimator δ_c in (4.39), which is a special case of (4.41) with $r = 1$, is better than any $\delta_{c,r}$ in (4.41) with $r \neq 1$, since the factor $2r - r^2$ takes on its maximum value 1 if and only if $r = 1$. To see that δ_c may have a substantial improvement over X in terms of risks, consider the special case where $\theta = c$. Since $\|X - c\|^2$ has a chi-square distribution χ_p^2 when $\theta = c$, $E\|X-c\|^{-2} = (p-2)^{-1}$ and the right-hand side of (4.42) equals 2. Thus, the ratio $R_X(\theta)/R_{\delta_c}(\theta)$ equals $p/2$ when $\theta = c$ and, therefore, can be substantially larger than 1 near $\theta = c$ when p is large.

Since X is minimax (Example 4.25), any shrinkage estimator of the form (4.41) is minimax provided that $p \geq 3$ and $0 < r < 2$.

Unfortunately, the James-Stein estimator with any c is also inadmissible. It is dominated by

$$\delta_c^+ = c + \left[\max\left(1 - \frac{p-2}{\|X-c\|^2},\ 0\right)\right](X - c); \tag{4.44}$$

see, for example, Lehmann (1983, Theorem 4.6.2). This estimator, however, is still inadmissible. An example of an admissible estimator of the form (4.40) is provided by Strawderman (1971); see also Lehmann (1983, p. 304). Although neither the James-Stein estimator δ_c nor δ_c^+ in (4.44) is admissible, it is found that no substantial improvements over δ_c^+ are possible (Efron and Morris, 1973).

To extend Theorem 4.15 to general $\mathrm{Var}(X)$, we consider the case where $\mathrm{Var}(X) = \sigma^2 D$ with an unknown $\sigma^2 > 0$ and a known positive definite matrix D. If σ^2 is known, then an extended James-Stein estimator is

$$\tilde{\delta}_{c,r} = c + \left[1 - \frac{r(p-2)\sigma^2}{\|D^{-1}(X-c)\|^2}\right] D^{-1}(X - c). \tag{4.45}$$

One can show (exercise) that under the loss (4.37), the risk of $\tilde{\delta}_{c,r}$ is

$$\sigma^2 \left[\mathrm{tr}(D) - (2r - r^2)(p-2)^2\sigma^2 E(\|D^{-1}(X-c)\|^{-2})\right]. \tag{4.46}$$

When σ^2 is unknown, we assume that there exists a statistic S_0^2 such that S_0^2 is independent of X and S_0^2/σ^2 has the chi-square distribution χ^2_m (see Example 4.27). Replacing $r\sigma^2$ in (4.45) by $\hat{\sigma}^2 = tS_0^2$ with a constant $t > 0$ leads to the following extended James-Stein estimator

$$\tilde{\delta}_c = c + \left[1 - \frac{(p-2)\hat{\sigma}^2}{\|D^{-1}(X-c)\|^2}\right] D^{-1}(X-c). \tag{4.47}$$

By (4.46) and the independence of $\hat{\sigma}^2$ and X, the risk of $\tilde{\delta}_c$ (as an estimator of $\vartheta = EX$) is

$$\begin{aligned}
R_{\tilde{\delta}_c}(\theta) &= E\left[E(\|\tilde{\delta}_c - \vartheta\|^2|\hat{\sigma}^2)\right] \\
&= E\left[E(\|\tilde{\delta}_{c,(\hat{\sigma}^2/\sigma^2)} - \vartheta\|^2|\hat{\sigma}^2)\right] \\
&= \sigma^2 E\left\{\mathrm{tr}(D) - [2(\hat{\sigma}^2/\sigma^2) - (\hat{\sigma}^2/\sigma^2)^2](p-2)^2\sigma^2\kappa(\theta)\right\} \\
&= \sigma^2\left\{\mathrm{tr}(D) - [2E(\hat{\sigma}^2/\sigma^2) - E(\hat{\sigma}^2/\sigma^2)^2](p-2)^2\sigma^2\kappa(\theta)\right\} \\
&= \sigma^2\left\{\mathrm{tr}(D) - [2tm - t^2m(m+2)](p-2)^2\sigma^2\kappa(\theta)\right\},
\end{aligned}$$

where $\theta = (\vartheta, \sigma^2)$ and $\kappa(\theta) = E(\|D^{-1}(X-c)\|^{-2})$. Since $2tm - t^2m(m+2)$ is maximized at $t = 1/(m+2)$, replacing t by $1/(m+2)$ leads to

$$R_{\tilde{\delta}_c}(\theta) = \sigma^2\left[\mathrm{tr}(D) - m(m+2)^{-1}(p-2)^2\sigma^2 E(\|D^{-1}(X-c)\|^{-2})\right].$$

Hence the risk of the extended James-Stein estimator in (4.47) is smaller than that of X for any fixed θ, when $p \geq 3$.

Example 4.27. Consider the general linear model (3.25) with assumption A1, $p \geq 3$, and a full rank Z, and the estimation of $\vartheta = \beta$ under the loss function (4.37). From Theorem 3.8, the LSE $\hat{\beta}$ is from $N(\beta, \sigma^2 D)$ with a known matrix $D = (Z^\tau Z)^{-1}$; $S_0^2 = SSR$ is independent of $\hat{\beta}$; and S_0^2/σ^2 has the chi-square distribution χ^2_{n-p}. Hence, from the previous discussion, the risk of the shrinkage estimator

$$c + \left[1 - \frac{(p-2)\hat{\sigma}^2}{\|Z^\tau Z(\hat{\beta}-c)\|^2}\right] Z^\tau Z(\hat{\beta}-c)$$

is smaller than that of $\hat{\beta}$ for any β and σ^2, where $c \in \mathcal{R}^p$ is fixed and $\hat{\sigma}^2 = SSR/(n-p+2)$. ∎

From the previous discussion, the James-Stein estimators improve X substantially when we shrink the observations toward a vector c which is

near $\vartheta = EX$. Of course this cannot be done since ϑ is unknown. One may consider shrinking the observations toward the mean of the observations rather than a given point; that is, one may obtain a shrinkage estimator by replacing c in (4.39) or (4.47) by $\bar{X}J_p$, where $\bar{X} = p^{-1}\sum_{i=1}^p X_i$ and J_p is the p-vector of ones. However, we have to replace the factor $p-2$ in (4.39) or (4.47) by $p-3$. This leads to shrinkage estimators

$$\bar{X}J_p + \left(1 - \frac{p-3}{\|X - \bar{X}J_p\|^2}\right)(X - \bar{X}J_p) \tag{4.48}$$

and

$$\bar{X}J_p + \left[1 - \frac{(p-3)\hat{\sigma}^2}{\|D^{-1}(X - \bar{X}J_p)\|^2}\right] D^{-1}(X - \bar{X}J_p). \tag{4.49}$$

These estimators are better than X (and, hence, are minimax) when $p \geq 4$, under the loss function (4.37) (exercise).

The results discussed in this section for the simultaneous estimation of a vector of normal means can be extended to a wide variety of cases where the loss functions are not given by (4.37) (Brown, 1966). The results have also been extended to exponential families and to general location parameter families. For example, Berger (1976) studied the inadmissibility of generalized Bayes estimators of a location vector; Berger (1980) considered simultaneous estimation of gamma scale parameters; and Tsui (1981) investigated simultaneous estimation of several Poisson parameters. See Lehmann (1983, pp. 320-330) for some further references.

4.4 The Method of Maximum Likelihood

So far we have studied estimation methods in parametric families using the decision theory approach. The *maximum likelihood method* introduced next is the most popular method for deriving estimators in statistical inference that does not use any loss function.

4.4.1 The likelihood function and MLE's

To introduce the idea, let us consider an example.

Example 4.28. Let X be a single observation taking values from $\{0, 1, 2\}$ according to P_θ, where $\theta = \theta_0$ or θ_1 and the values of $P_{\theta_j}(\{i\})$ are given by the following table:

	$x = 0$	$x = 1$	$x = 2$
$\theta = \theta_0$	0.8	0.1	0.1
$\theta = \theta_1$	0.2	0.3	0.5

If $X = 0$ is observed, it is more plausible that it came from P_{θ_0}, since $P_{\theta_0}(\{0\})$ is much larger than $P_{\theta_1}(\{0\})$. We then estimate θ by θ_0. On the other hand, if $X = 1$ or 2, it is more plausible that it came from P_{θ_1}, although in this case the difference between the probabilities is not as large as that in the case of $X = 0$. This suggests the following estimator of θ:

$$T(X) = \begin{cases} \theta_0 & X = 0 \\ \theta_1 & X \neq 0. \end{cases} \quad \blacksquare$$

The idea in Example 4.28 can be easily extended to the case where P_θ is a discrete distribution and $\theta \in \Theta \subset \mathcal{R}^k$. If $X = x$ is observed, θ_1 is more plausible than θ_2 if and only if $P_{\theta_1}(\{x\}) > P_{\theta_2}(\{x\})$. We then estimate θ by a $\hat{\theta}$ that maximizes $P_\theta(\{x\})$ over $\theta \in \Theta$, if such a $\hat{\theta}$ exists. The word plausible rather than probable is used because θ is considered to be nonrandom and P_θ is not a distribution of θ. Under the Bayesian approach with a prior that is the discrete uniform distribution on $\{\theta_1, ..., \theta_m\}$, $P_\theta(\{x\})$ is proportional to the posterior probability and we can say that θ_1 is more probable than θ_2 if $P_{\theta_1}(\{x\}) > P_{\theta_2}(\{x\})$.

Note that $P_\theta(\{x\})$ in the previous discussion is the p.d.f. w.r.t. the counting measure. Hence, it is natural to extend the idea to the case of continuous (or arbitrary) X by using the p.d.f. of X w.r.t. some σ-finite measure on the range $\mathfrak{X}$ of X. This leads to the following definition.

Definition 4.3. Let X be a sample from P_θ, $\theta \in \Theta \subset \mathcal{R}^k$. Assume that P_θ's have p.d.f.'s f_θ w.r.t. a σ-finite measure.
(i) For each $x \in \mathfrak{X}$, $f_\theta(x)$ considered as a function of θ is called the *likelihood function* and denoted by $\ell(\theta)$.
(ii) Let $\bar{\Theta}$ be the closure of Θ. A $\hat{\theta} \in \bar{\Theta}$ satisfying $\ell(\hat{\theta}) = \max_{\theta \in \bar{\Theta}} \ell(\theta)$ is called a *maximum likelihood estimate* (MLE) of θ. $\hat{\theta}$ viewed as a function of X is called a *maximum likelihood estimator* (MLE) of θ.
(iii) Let g be a Borel function from Θ to $\mathcal{R}^p$, $p \leq k$. If $\hat{\theta}$ is an MLE of θ, then $\hat{\vartheta} = g(\hat{\theta})$ is defined to be an MLE of $\vartheta = g(\theta)$. $\blacksquare$

Note that $\bar{\Theta}$ instead of Θ is used in the definition of the MLE. This is because a maximum of $\ell(\theta)$ may not exist when Θ is a bounded open set (Examples 4.29 and 4.30). Part (iii) of Definition 4.3 is motivated by a fact given in Exercise 83 of §4.6.

If the parameter space Θ contains finitely many points, then $\bar{\Theta} = \Theta$ and an MLE can always be obtained by comparing finitely many values $\ell(\theta)$, $\theta \in \Theta$. If $\ell(\theta)$ is differentiable on an open set $\Theta^\circ \subset \Theta$, then possible candidates for MLE's are the values of $\theta \in \Theta^\circ$ satisfying

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0, \tag{4.50}$$

which is called the *likelihood equation*. Note that θ's satisfying (4.50) may be local or global minima, local or global maxima, or inflection points. Also, extrema may occur at the boundary of Θ or when $\|\theta\| \to \infty$. Furthermore, if $\ell(\theta)$ is not always differentiable, then extrema may occur at nondifferentiable or discontinuity points of $\ell(\theta)$. Hence, it is important to analyze the entire likelihood function to find its maxima.

Since $\log x$ is a strictly increasing function and $\ell(\theta)$ can be assumed to be positive without loss of generality, $\hat{\theta}$ is an MLE if and only if it maximizes the log-likelihood function $\log \ell(\theta)$. It is often more convenient to work with $\log \ell(\theta)$ and the following analogue of (4.50) (which is called the log-likelihood equation or likelihood equation for simplicity):

$$\frac{\partial \log \ell(\theta)}{\partial \theta} = 0. \tag{4.51}$$

Example 4.29. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $P(X_1 = 1) = p \in \Theta = (0, 1)$. The likelihood function is

$$\ell(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{n\bar{x}}(1-p)^{n(1-\bar{x})}.$$

Note that $\bar{\Theta} = [0, 1]$ and $\Theta^\circ = \Theta$. The likelihood equation (4.51) reduces to

$$\frac{n\bar{x}}{p} - \frac{n(1-\bar{x})}{1-p} = 0.$$

If $0 < \bar{x} < 1$, then this equation has a unique solution $\bar{x}$. The second-order derivative of $\log \ell(p)$ is

$$-\frac{n\bar{x}}{p^2} - \frac{n(1-\bar{x})}{(1-p)^2},$$

which is always negative. Also, when p tends to 0 or 1 (the boundary of Θ), $\ell(p) \to 0$. Thus, $\bar{x}$ is the unique MLE of p.

When $\bar{x} = 0$, $\ell(p) = (1-p)^n$ is a strictly decreasing function of p and, therefore, its unique maximum is 0. Similarly, the MLE is 1 when $\bar{x} = 1$. Combining these results with the previous result, we conclude that the MLE of p is $\bar{x}$.

When $\bar{x} = 0$ or 1, a maximum of $\ell(p)$ does not exist on $\Theta = (0, 1)$, although $\sup_{p \in (0,1)} \ell(p) = 1$; the MLE takes a value outside of Θ and, hence, is not a reasonable estimator. However, if $p \in (0, 1)$, the probability that $\bar{x} = 0$ or 1 tends to 0 quickly as $n \to \infty$. ∎

Example 4.29 indicates that for small n, a maximum of $\ell(\theta)$ may not exist on Θ and an MLE may be an unreasonable estimator; however, this

is unlikely to occur when n is large. A rigorous result of this sort is given in §4.5.2 where we study asymptotic properties of MLE's.

Example 4.30. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with an unknown $\theta = (\mu, \sigma^2)$, where $n \geq 2$. Consider first the case where $\Theta = \mathcal{R} \times (0, \infty)$. The log-likelihood function is

$$\log \ell(\theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{n}{2} \log \sigma^2 - \frac{n}{2} \log(2\pi).$$

The likelihood equation (4.51) becomes

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0 \qquad \text{and} \qquad \frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{n}{2\sigma^2} = 0. \tag{4.52}$$

Solving the first equation in (4.52) for μ we obtain a unique solution $\bar{x}$ and, substituting $\bar{x}$ in the second equation in (4.52), we obtain a unique solution $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$. To show that $\hat{\theta} = (\bar{x}, \hat{\sigma}^2)$ is an MLE, first note that Θ is an open set and $\ell(\theta)$ is differentiable everywhere; as θ tends to the boundary of Θ or $\|\theta\| \to \infty$, $\ell(\theta)$ tends to 0; and

$$\frac{\partial^2 \log \ell(\theta)}{\partial\theta\partial\theta^\tau} = -\begin{pmatrix} \frac{n}{\sigma^2} & \frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu) \\ \frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu) & \frac{1}{\sigma^6} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{n}{\sigma^4} \end{pmatrix}$$

is negative definite when $\mu = \bar{x}$ and $\sigma^2 = \hat{\sigma}^2$. Hence $\hat{\theta}$ is the unique MLE. Sometimes we can avoid the calculation of the second-order derivatives. For instance, in this example we know that $\ell(\theta)$ is bounded and a maximum must be in the interior of Θ. Since (4.52) has a unique solution and a maximum of $\ell(\theta)$ must satisfy (4.52), $\hat{\theta}$ must be the MLE. Another way to show that $\hat{\theta}$ is the MLE is indicated by the following discussion.

Consider next the case where $\Theta = (0, \infty) \times (0, \infty)$, i.e., μ is known to be positive. The likelihood function is differentiable on $\Theta^\circ = \Theta$ and $\bar{\Theta} = [0, \infty) \times [0, \infty)$. If $\bar{x} > 0$, then one can still show that $(\bar{x}, \hat{\sigma}^2)$ is the MLE. If $\bar{x} \leq 0$, then the first equation in (4.52) does not have a solution in Θ. However, the function $\log \ell(\theta) = \log \ell(\mu, \sigma^2)$ is strictly decreasing in μ for any fixed σ^2. Hence a maximum of $\log \ell(\mu, \sigma^2)$ is $\mu = 0$, which does not depend on σ^2. Then, the MLE is $(0, \tilde{\sigma}^2)$, where $\tilde{\sigma}^2$ is the value maximizing $\log \ell(0, \sigma^2)$ over $\sigma^2 \geq 0$. Applying (4.51) to the function $\log \ell(0, \sigma^2)$ leads to $\tilde{\sigma}^2 = n^{-1} \sum_{i=1}^{n} x_i^2$. Thus, the MLE is

$$\hat{\theta} = \begin{cases} (\bar{x}, \hat{\sigma}^2) & \bar{x} > 0 \\ (0, \tilde{\sigma}^2) & \bar{x} \leq 0. \end{cases}$$

Again, the MLE in this case is not in Θ if $\bar{x} \leq 0$. One can show that a maximum of $\ell(\theta)$ does not exist on Θ when $\bar{x} \leq 0$. ∎

Example 4.31. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on an interval $\mathcal{I}_\theta$ with an unknown θ. First, consider the case where $\mathcal{I}_\theta = (0, \theta)$ and $\theta > 0$. The likelihood function is $\ell(\theta) = \theta^{-n} I_{(x_{(n)}, \infty)}(\theta)$, which is not always differentiable. In this case $\Theta^\circ = (0, x_{(n)}) \cup (x_{(n)}, \infty)$. But, on $(0, x_{(n)})$, $\ell \equiv 0$ and on $(x_{(n)}, \infty)$, $\ell'(\theta) = -n\theta^{n-1} < 0$ for all θ. Hence, the method of using the likelihood equation is not applicable to this problem. Since $\ell(\theta)$ is strictly decreasing on $(x_{(n)}, \infty)$ and is 0 on $(0, x_{(n)})$, a unique maximum of $\ell(\theta)$ is $x_{(n)}$, which is a discontinuity point of $\ell(\theta)$. This shows that the MLE of θ is the largest order statistic $X_{(n)}$.

Next, consider the case where $\mathcal{I}_\theta = (\theta - \frac{1}{2}, \theta + \frac{1}{2})$ with $\theta \in \mathcal{R}$. The likelihood function is $\ell(\theta) = I_{(x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2})}(\theta)$. Again, the method of using the likelihood equation is not applicable. However, it follows from Definition 4.3 that any statistic $T(X)$ satisfying $x_{(n)} - \frac{1}{2} \leq T(x) \leq x_{(1)} + \frac{1}{2}$ is an MLE of θ. This example indicates that MLE's may not be unique and can be unreasonable. ∎

Example 4.32. Let X be an observation from the hypergeometric distribution $HG(r, n, \theta - n)$ (Table 1.1, page 18) with known r, n, and an unknown $\theta = n+1, n+2, ...$ In this case, the likelihood function is defined on integers and the method of using the likelihood equation is certainly not applicable. Note that

$$\frac{\ell(\theta)}{\ell(\theta - 1)} = \frac{(\theta - r)(\theta - n)}{\theta(\theta - n - r + x)},$$

which is larger than 1 if and only if $\theta < rn/x$ and is smaller than 1 if and only if $\theta > rn/x$. Thus, $\ell(\theta)$ has a maximum $\theta =$ the integer part of rn/x, which is the MLE of θ. ∎

Example 4.33. Let $X_1, ..., X_n$ be i.i.d. from the gamma distribution $\Gamma(\alpha, \gamma)$ with unknown $\alpha > 0$ and $\gamma > 0$. The log-likelihood function is

$$\log \ell(\theta) = -n\alpha \log \gamma - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^n \log x_i - \frac{1}{\gamma} \sum_{i=1}^n x_i$$

and the likelihood equation (4.51) becomes

$$-n \log \gamma - \frac{n\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^n \log x_i = 0$$

and

$$-\frac{n\alpha}{\gamma} + \frac{1}{\gamma^2} \sum_{i=1}^n x_i = 0.$$

The second equation yields $\gamma = \bar{x}/\alpha$. Substituting $\gamma = \bar{x}/\alpha$ into the first equation we obtain that

$$\log\alpha - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \frac{1}{n}\sum_{i=1}^{n}\log x_i - \log\bar{x} = 0.$$

In this case, the likelihood equation does not have an explicit solution, although it can be shown (exercise) that a solution exists and it is the unique MLE. A numerical method has to be applied to compute the MLE for any given observations $x_1, ..., x_n$. ∎

These examples indicate that we need to use various methods to derive MLE's. In applications MLE's typically do not have analytic forms and some numerical methods have to be used to compute MLE's. A commonly used numerical method is the Newton-Raphson iteration method which repeatedly computes

$$\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} - \left.\frac{\partial\log\ell(\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}^{(t)}}\left[\left.\frac{\partial^2\log\ell(\theta)}{\partial\theta\partial\theta^\tau}\right|_{\theta=\hat{\theta}^{(t)}}\right]^{-1}, \tag{4.53}$$

$t = 0, 1, ...$, where $\hat{\theta}^{(0)}$ is an initial value and $\partial^2\log\ell(\theta)/\partial\theta\partial\theta^\tau$ is assumed of full rank for every $\theta \in \Theta$. If, at each iteration, we replace $\partial^2\log\ell(\theta)/\partial\theta\partial\theta^\tau$ in (4.53) by its expected value $E[\partial^2\log\ell(\theta)/\partial\theta\partial\theta^\tau]$, where the expectation is taken under P_θ, then the method is known as the Fisher-scoring method. If the iteration converges, then $\hat{\theta}^{(\infty)}$ or $\hat{\theta}^{(t)}$ with a sufficiently large t is a numerical approximation to a solution of the likelihood equation (4.51).

The following example shows that the MCMC methods discussed in §4.1.4 can also be useful in computing MLE's.

Example 4.34. Let X be a random k-vector from P_θ with the following p.d.f. w.r.t. a σ-finite measure ν:

$$f_\theta(x) = \int f_\theta(x, y)d\nu(y),$$

where $f_\theta(x, y)$ is a joint p.d.f. w.r.t. $\nu \times \nu$. This type of distribution is called a *mixture* distribution. Thus, the likelihood $\ell(\theta) = f_\theta(x)$ involves a k-dimensional integral. In many cases this integral has to be computed in order to compute an MLE of θ.

Let $\tilde{\ell}_m(\theta)$ be the MCMC approximation to $\ell(\theta)$ based on one of the MCMC methods described in §4.1.4 and a Markov chain of length m. Under the conditions of Theorem 4.4, $\tilde{\ell}_m(\theta) \to_{a.s.} \ell(\theta)$ for every fixed θ and x. Suppose that for each m, there exists $\tilde{\theta}_m$ which maximizes $\tilde{\ell}_m(\theta)$ over $\theta \in \Theta$. Geyer (1994) studies the convergence of $\tilde{\theta}_m$ to an MLE. ∎

In terms of their mse's, MLE's are not necessarily better than UMVUE's or Bayes estimators. Also, MLE's are frequently inadmissible. This is not surprising, since MLE's are not derived under any given loss function. The main theoretical justification for MLE's is provided in the theory of asymptotic efficiency considered in §4.5.

4.4.2 MLE's in generalized linear models

Suppose that X has a distribution from a natural exponential family so that the likelihood function is

$$\ell(\eta) = \exp\{T(x)\eta^\tau - \zeta(\eta)\}h(x),$$

where $\eta \in \Xi$ is a vector of unknown parameters. The likelihood equation (4.51) is then

$$\frac{\partial \log \ell(\eta)}{\partial \eta} = T(x) - \frac{\partial \zeta(\eta)}{\partial \eta} = 0,$$

which has a unique solution $T(x) = \partial\zeta(\eta)/\partial\eta$, assuming that $T(x)$ is in the range of $\partial\zeta(\eta)/\partial\eta$. Note that

$$\frac{\partial^2 \log \ell(\eta)}{\partial \eta \partial \eta^\tau} = -\frac{\partial^2 \zeta(\eta)}{\partial \eta \partial \eta^\tau} = -\mathrm{Var}(T) \tag{4.54}$$

(see the proof of Proposition 3.2). Since $\mathrm{Var}(T)$ is positive definite, the log-likelihood function is convex in η and $T(x)$ is the unique MLE of the parameter $\mu(\eta) = \partial\zeta(\eta)/\partial\eta$. By (4.54) again, the function $\mu(\eta)$ is one-to-one so that μ^{-1} exists. By Definition 4.3, the MLE of η is $\hat{\eta} = \mu^{-1}(T(x))$.

If the distribution of X is in a general exponential family and the likelihood function is

$$\ell(\theta) = \exp\{T(x)[\eta(\theta)]^\tau - \xi(\theta)\}h(x),$$

then the MLE of θ is $\hat{\theta} = \eta^{-1}(\hat{\eta})$, if η^{-1} exists and $\hat{\eta}$ is in the range of $\eta(\theta)$. Of course, $\hat{\theta}$ is also the solution of the likelihood equation

$$\frac{\partial \log \ell(\theta)}{\partial \theta} = \frac{\partial \eta(\theta)}{\partial \theta}[T(x)]^\tau - \frac{\partial \xi(\theta)}{\partial \theta} = 0.$$

The results for exponential families lead to an estimation method in a class of models that have very wide applications. These models are generalizations of the normal linear model (model (3.25) with assumption A1) discussed in §3.3.1-§3.3.2 and, therefore, are named *generalized linear models* (GLM).

A GLM has the following structure. The sample $X = (X_1, ..., X_n) \in \mathcal{R}^n$ has independent components and X_i has the p.d.f.

$$\exp\left\{\frac{\eta_i x_i - \zeta(\eta_i)}{\phi_i}\right\} h(x_i, \phi_i), \qquad i = 1, ..., n, \tag{4.55}$$

w.r.t. a σ-finite measure ν, where η_i and ϕ_i are unknown, $\phi_i > 0$,

$$\eta_i \in \Xi = \left\{\eta : \ 0 < \int h(x, \phi) e^{\eta x/\phi} d\nu(x) < \infty\right\} \subset \mathcal{R}$$

for all i, ζ and h are known functions, and $\zeta''(\eta) > 0$ is assumed for all $\eta \in \Xi^\circ$, the interior of Ξ. Note that the p.d.f. in (4.55) belongs to an exponential family if ϕ_i is known. As a consequence,

$$E(X_i) = \zeta'(\eta_i) \quad \text{and} \quad \text{Var}(X_i) = \phi_i \zeta''(\eta_i), \qquad i = 1, ..., n. \tag{4.56}$$

Define $\mu(\eta) = \zeta'(\eta)$. It is assumed that η_i is related to Z_i, the ith value of a p-vector of covariates (see (3.24)), through

$$g(\mu(\eta_i)) = \beta Z_i^\tau, \qquad i = 1, ..., n, \tag{4.57}$$

where β is a p-vector of unknown parameters and g, called a *link function*, is a known one-to-one, third-order continuously differentiable function on $\{\mu(\eta) : \eta \in \Xi^\circ\}$. If $\mu = g^{-1}$, then $\eta_i = \beta Z_i^\tau$ and g is called the *canonical* or *natural* link function. If g is not canonical, we assume that $\frac{d}{d\eta}(g \circ \mu)(\eta) \neq 0$ for all η.

In a GLM, the parameter of interest is β. We assume that the range of β is $B = \{\beta : (g \circ \mu)^{-1}(\beta z^\tau) \in \Xi^\circ \text{ for all } z \in \mathcal{Z}\}$, where $\mathcal{Z}$ is the range of Z_i's. ϕ_i's are called *dispersion* parameters and are considered to be nuisance parameters. It is often assumed that

$$\phi_i = \phi / t_i, \qquad i = 1, ..., n, \tag{4.58}$$

with an unknown $\phi > 0$ and known positive t_i's.

As we discussed earlier, the linear model (3.24) with $\varepsilon_i = N(0, \phi)$ is a special case of GLM. One can verify this by taking $g(\mu) \equiv \mu$ and $\zeta(\eta) = \eta^2/2$. The usefulness of GLM is that it covers situations where the relationship between $E(X_i)$ and Z_i is nonlinear and/or X_i's are discrete (in which case the linear model (3.24) is clearly not appropriate). The following is an example.

Example 4.35. Let X_i's be independent discrete random variables taking values in $\{0, 1, ..., m\}$, where m is a known positive integer. First, suppose that X_i has the binomial distribution $Bi(p_i, m)$ with an unknown $p_i \in (0, 1)$, $i = 1, ..., n$. Let $\eta_i = \log \frac{p_i}{1-p_i}$ and $\zeta(\eta_i) = m \log(1 + e^{\eta_i})$. Then the p.d.f. of X_i (w.r.t. the counting measure) is given by (4.55) with $\phi_i = 1$,

$h(x_i, \phi_i) = \binom{m}{x_i}$, and $\Xi = \mathcal{R}$. Under (4.57) and the *logit* link (canonical link) $g(t) = \log \frac{t}{1-t}$,

$$E(X_i) = mp_i = \frac{me^{\eta_i}}{1+e^{\eta_i}} = \frac{me^{\beta Z_i^\tau}}{1+e^{\beta Z_i^\tau}}.$$

Another popular link in this problem is the *probit* link $g(t) = \Phi^{-1}(t)$, where Φ is the c.d.f. of the standard normal. Under the probit link, $E(X_i) = m\Phi(\beta Z_i^\tau)$.

The variance of X_i is $mp_i(1-p_i)$ under the binomial distribution assumption. This assumption is often violated in applications, which results in an *over-dispersion*, i.e., the variance of X_i exceeds the nominal variance $mp_i(1-p_i)$. Over-dispersion can arise in a number of ways, but the most common one is clustering in the population. Families, households, and litters are common instances of clustering. For example, suppose that $X_i = \sum_{j=1}^m X_{ij}$, where X_{ij} are binary random variables having a common distribution. If X_{ij}'s are independent, then X_i has a binomial distribution. However, if X_{ij}'s are from the same cluster (family or household), then they are often positively related. Suppose that the correlation coefficient (§1.3.2) between X_{ij} and X_{il}, $j \neq l$, is $\rho_i > 0$. Then

$$\mathrm{Var}(X_i) = mp_i(1-p_i) + m(m-1)\rho_i p_i(1-p_i) = \phi_i mp_i(1-p_i),$$

where $\phi_i = 1 + (m-1)\rho_i$ is the dispersion parameter. Of course, over-dispersion can occur only if $m > 1$ in this case.

This motivates the consideration of GLM (4.55)-(4.57) with dispersion parameters ϕ_i. If X_i has the p.d.f. (4.55) with $\zeta(\eta_i) = m\log(1+e^{\eta_i})$, then

$$E(X_i) = \frac{me^{\eta_i}}{1+e^{\eta_i}} \qquad \text{and} \qquad \mathrm{Var}(X_i) = \phi_i \frac{me^{\eta_i}}{(1+e^{\eta_i})^2},$$

which is exactly (4.56). Of course, the distribution of X_i is not binomial unless $\phi_i = 1$. ∎

We now derive an MLE of β in a GLM under assumption (4.58). Let $\theta = (\beta, \phi)$ and $\psi = (g \circ \mu)^{-1}$. Then the log-likelihood function is

$$\log \ell(\theta) = \sum_{i=1}^n \left[\log h(x_i, \phi/t_i) + \frac{\psi(\beta Z_i^\tau)x_i - \zeta(\psi(\beta Z_i^\tau))}{\phi/t_i} \right]$$

and the likelihood equation is

$$\frac{\partial \log \ell(\theta)}{\partial \beta} = \frac{1}{\phi} \sum_{i=1}^n \{[x_i - \mu(\psi(\beta Z_i^\tau))]\psi'(\beta Z_i^\tau) t_i Z_i\} = 0 \tag{4.59}$$

and

$$\frac{\partial \log \ell(\theta)}{\partial \phi} = \sum_{i=1}^{n} \left\{ \frac{\partial \log h(x_i, \phi/t_i)}{\partial \phi} - \frac{t_i[\psi(\beta Z_i^\tau)x_i - \zeta(\psi(\beta Z_i^\tau))]}{\phi^2} \right\} = 0.$$

From the first equation, an MLE of β, if it exists, can be obtained without estimating ϕ. The second equation, however, is usually difficult to solve. Some other estimators of ϕ are suggested by various researchers; see, for example, McCullagh and Nelder (1989).

Suppose that there is a solution $\hat{\beta} \in B$ to equation (4.59). (The existence of $\hat{\beta}$ is studied in §4.5.2.) We now study whether $\hat{\beta}$ is an MLE of β. Let

$$M_n(\beta) = \sum_{i=1}^{n} [\psi'(\beta Z_i^\tau)]^2 \zeta''(\psi(\beta Z_i^\tau)) t_i Z_i^\tau Z_i \tag{4.60}$$

and

$$R_n(\beta) = \sum_{i=1}^{n} [x_i - \mu(\psi(\beta Z_i^\tau))] \psi''(\beta Z_i^\tau) t_i Z_i^\tau Z_i. \tag{4.61}$$

Then

$$\mathrm{Var}\left(\frac{\partial \log \ell(\theta)}{\partial \beta} \right) = M_n(\beta)/\phi \tag{4.62}$$

and

$$\frac{\partial^2 \log \ell(\theta)}{\partial \beta \partial \beta^\tau} = [R_n(\beta) - M_n(\beta)]/\phi. \tag{4.63}$$

Consider first the simple case of canonical g. Then $\psi'' \equiv 0$ and $R_n \equiv 0$. If $M_n(\beta)$ is positive definite for all β, then $\log \ell(\theta)$ is strictly convex in β for any fixed ϕ and, therefore, $\hat{\beta}$ is the unique MLE of β. For the case of noncanonical g, $R_n(\beta) \neq 0$ and $\hat{\beta}$ is not necessarily an MLE. If $R_n(\beta)$ is dominated by $M_n(\beta)$ (i.e., $[M_n(\beta)]^{-1/2} R_n(\beta) [M_n(\beta)]^{-1/2} \to 0$ in some sense), then $\log \ell(\theta)$ is convex and $\hat{\beta}$ is an MLE for large n; see more details in the proof of Theorem 4.18 in §4.5.2.

Example 4.36. Consider the GLM (4.55) with $\zeta(\eta) = \eta^2/2$, $\eta \in \mathcal{R}$. If g in (4.57) is the canonical link, then the model is the same as (3.24) with independent ε_i's distributed as $N(0, \phi_i)$. If (4.58) holds with $t_i \equiv 1$, then (4.59) is exactly the same as equation (3.27). If Z is of full rank, then $M_n(\beta) = Z^\tau Z$ is positive definite. Thus, we have shown that the LSE $\hat{\beta}$ given by (3.28) is actually the unique MLE of β.

Suppose now that g is noncanonical but (4.58) still holds with $t_i \equiv 1$. Then the model reduces to that X_i's are independent and

$$X_i = N\left(g^{-1}(\beta Z_i^\tau), \phi\right), \qquad i = 1, ..., n. \tag{4.64}$$

This type of model is called a *nonlinear regression model* (with normal errors) and an MLE of β under this model is also called a nonlinear LSE, since maximizing the log-likelihood is equivalent to minimizing the sum of squares $\sum_{i=1}^n [X_i - g^{-1}(\beta Z_i^\tau)]^2$. Under certain conditions the matrix $R_n(\beta)$ is dominated by $M_n(\beta)$ and an MLE of β exists; see more details in §4.5.2. ∎

Example 4.37 (The Poisson model). Consider the GLM (4.55) with $\zeta(\eta) = e^\eta$, $\eta \in \mathcal{R}$. If $\phi_i \equiv 1$, then X_i has the Poisson distribution with mean e^{η_i}. Assume that (4.58) holds. Under the canonical link $g(t) = \log t$,

$$M_n(\beta) = \sum_{i=1}^n e^{\beta Z_i^\tau} t_i Z_i^\tau Z_i,$$

which is positive definite if $\inf_i e^{\beta Z_i^\tau} > 0$ and the matrix $(\sqrt{t}_1 Z_1^\tau, ..., \sqrt{t}_n Z_n^\tau)$ is of full rank.

There is one noncanonical link that deserves attention. Suppose that we choose a link function so that $[\psi'(t)]^2 \zeta''(\psi(t)) \equiv 1$. Then $M_n(\beta) \equiv \sum_{i=1}^n t_i Z_i^\tau Z_i$ does not depend on β. In §4.5.2 it is shown that the asymptotic variance of the MLE $\hat\beta$ is $\phi[M_n(\beta)]^{-1}$. The fact that $M_n(\beta)$ does not depend on β makes the estimation of the asymptotic variance (and, thus, statistical inference) easy. Under the Poisson model, $\zeta''(t) = e^t$ and, therefore, we need to solve the differential equation $[\psi'(t)]^2 e^{\psi(t)} = 1$. A solution is $\psi(t) = 2\log(t/2)$, which gives the link function $g(\mu) = 2\sqrt{\mu}$. ∎

In a GLM, an MLE $\hat\beta$ usually does not have an analytic form. A numerical method such as the Newton-Raphson or the Fisher-scoring method has to be applied. Using the Newton-Raphson method, we have the following iteration procedure:

$$\hat\beta^{(t+1)} = \hat\beta^{(t)} - s_n(\hat\beta^{(t)})[R_n(\hat\beta^{(t)}) - M_n(\hat\beta^{(t)})]^{-1}, \qquad t = 0, 1, ...,$$

where $s_n(\beta) = \phi \partial \log \ell(\theta)/\partial\beta$. Note that $E[R_n(\beta)] = 0$ if β is the true parameter value and x_i is replaced by X_i. This means that the Fisher-scoring method uses the following iteration procedure:

$$\hat\beta^{(t+1)} = \hat\beta^{(t)} + s_n(\hat\beta^{(t)})[M_n(\hat\beta^{(t)})]^{-1}, \qquad t = 0, 1, ...$$

If the canonical link is used, then the two methods are identical.

4.4.3 Quasi-likelihoods and conditional likelihoods

We now introduce two variations of the method of using likelihoods.

Consider a GLM (4.55)-(4.57). Assumption (4.58) is often unrealistic in applications. If there is no restriction on ϕ_i's, however, there are too many parameters and an MLE of β may not exist. (Note that assumption (4.58) reduces n nuisance parameters to one.) One way to solve this problem is to assume that $\phi_i = h(Z_i, \xi)$ for some known function h and unknown parameter vector ξ (which may include β as a subvector). Let $\theta = (\beta, \xi)$. Then we can try to solve the likelihood equation $\partial \log \ell(\theta)/\partial\theta = 0$ to obtain an MLE of β and/or ξ. We omit the details which can be found, for example, in Smyth (1989).

Suppose that we do not impose any assumptions on ϕ_i's but still estimate β by solving

$$\tilde{s}_n(\beta) = \sum_{i=1}^{n} \{[x_i - \mu(\psi(\beta Z_i^\tau))]\psi'(\beta Z_i^\tau) t_i Z_i\} = 0. \tag{4.65}$$

Note that (4.65) is not a likelihood equation unless (4.58) holds. In the special case of Example 4.36 where $X_i = N(\beta Z_i^\tau, \phi_i)$, $i = 1, ..., n$, a solution to (4.65) is simply an LSE of β whose properties are discussed at the end of §3.3.3. Estimating β by solving equation (4.65) is motivated by the following facts. First, if (4.58) does hold, then our estimate is an MLE. Second, if (4.58) is slightly violated, the performance of our estimate is still nearly the same as that of an MLE under assumption (4.58) (see the discussion of robustness at the end of §3.3.3). Finally, estimators obtained by solving (4.65) usually have good asymptotic properties. As a special case of a general result in §5.4, a solution to (4.65) is asymptotically normal under some regularity conditions.

In general, an equation such as (4.65) is called a *quasi-likelihood equation* if it is a likelihood equation when certain assumptions hold. The "likelihood" corresponding to a quasi-likelihood equation is called *quasi-likelihood* and a maximum of the quasi-likelihood is then called a *maximum quasi-likelihood estimate* (MQLE). Thus, a solution to (4.65) is an MQLE.

Note that (4.65) is a likelihood equation if and only if both (4.55) and (4.58) hold. The LSE (§3.3) without normality assumption on X_i's is a simple example of an MQLE without (4.55). Without assumption (4.55), the model under consideration is usually nonparametric and, therefore, the MQLE's are studied in §5.4.

While the quasi-likelihoods are used to relax some assumptions in our models, the *conditional likelihoods* discussed next are used mainly in cases where MLE's are difficult to compute. We consider two cases. In the first case, $\theta = (\theta_1, \theta_2)$, θ_1 is the main parameter vector of interest, and θ_2 is a nuisance parameter vector. Suppose that there is a statistic $T_2(X)$ that is sufficient for θ_2 for each fixed θ_1. By the sufficiency, the conditional distribution of X given T_2 does not depend on θ_2. The likelihood function

corresponding to the conditional p.d.f. of X given T_2 is called the conditional likelihood function. A conditional MLE of θ_1 can then be obtained by maximizing the conditional likelihood function. This method can be applied to the case where the dimension of θ is considerably larger than the dimension of θ_1 so that computing the unconditional MLE of θ is much more difficult than computing the conditional MLE of θ_1. Note that the conditional MLE's are usually different from the unconditional MLE's.

As a more specific example, suppose that X has a p.d.f. in an exponential family:

$$f_\theta(x) = \exp\{T_1(x)\theta_1^\tau + T_2(x)\theta_2^\tau - \zeta(\theta)\}h(x).$$

Then T_2 is sufficient for θ_2 for any given θ_1. Problems of this type are from comparisons of two binomial probabilities or two Poisson distributions (Exercises 104-105).

The second case is when our sample $X = (X_1, ..., X_n)$ follows a first-order autoregressive time series model:

$$X_t - \mu = \rho(X_{t-1} - \mu) + \varepsilon_t, \qquad t = 2, ..., n,$$

where $\mu \in \mathcal{R}$ and $\rho \in (-1, 1)$ are unknown and ε_i's are i.i.d. from $N(0, \sigma^2)$ with an unknown $\sigma^2 > 0$. This model is often a satisfactory representation of the error time series in economic models, and is one of the simplest and most heavily used models in time series analysis (Fuller, 1996). Let $\theta = (\mu, \rho, \sigma^2)$. The log-likelihood function is

$$\begin{aligned}\log \ell(\theta) = &-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 + \frac{1}{2}\log(1-\rho^2) \\ &- \frac{1}{2\sigma^2}\left\{(x_1-\mu)^2(1-\rho^2) + \sum_{t=2}^n [x_t - \mu - \rho(x_{t-1}-\mu)]^2\right\}.\end{aligned}$$

The computation of the MLE is greatly simplified if we consider the conditional likelihood given $X_1 = x_1$:

$$\log \ell(\theta|x_1) = -\frac{n-1}{2}\log(2\pi) - \frac{n-1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=2}^n [x_t - \mu - \rho(x_{t-1}-\mu)]^2.$$

Let $(\bar{x}_{-1}, \bar{x}_0) = (n-1)^{-1}\sum_{t=2}^n (x_{t-1}, x_t)$. If

$$\hat{\rho} = \sum_{t=2}^n (x_t - \bar{x}_0)(x_{t-1} - \bar{x}_{-1}) \Big/ \sum_{t=2}^n (x_{t-1} - \bar{x}_{-1})^2$$

is between -1 and 1, then it is the conditional MLE of ρ and the conditional MLE's of μ and σ^2 are, respectively,

$$\hat{\mu} = (\bar{x}_0 - \hat{\rho}\bar{x}_{-1})/(1-\hat{\rho})$$

and

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{t=2}^{n}[x_t - \bar{x}_0 - \hat{\rho}(x_{t-1} - \bar{x}_{-1})]^2.$$

Obviously, the result can be extended to the case where X follows a pth-order autoregressive time series model:

$$X_t - \mu = \rho_1(X_{t-1} - \mu) + \cdots + \rho_p(X_{t-p} - \mu) + \varepsilon_t, \qquad t = p+1, ..., n, \quad (4.66)$$

where ρ_j's are unknown parameters satisfying the constraint that the roots (which may be complex) of the polynomial $x^p - \rho_1 x^{p-1} - \cdots - \rho_p = 0$ are less than one in absolute value (exercise).

4.5 Asymptotically Efficient Estimation

In this section, we consider asymptotic optimality of point estimators in parametric models. We use the asymptotic mean squared error (amse, see §2.5.2) or its multivariate generalization to assess the performance of an estimator. Reasons for considering asymptotics have been discussed in §2.5.

We focus on estimators that are asymptotically normal, since this covers the majority of cases. Some cases of asymptotically nonnormal estimators are studied in Exercises 97-100 in §4.6.

4.5.1 Asymptotic optimality

Let $\{\hat{\theta}_n\}$ be a sequence of estimators of θ in a parametric model, i.e., $\hat{\theta}_n$ is a statistic based on $X = (X_1, ..., X_n)$ whose distribution is known for all n when θ is known. Suppose that as $n \to \infty$,

$$(\hat{\theta}_n - \theta)[V_n(\theta)]^{-1/2} \to_d N_k(0, I_k), \quad (4.67)$$

where, for each n, $V_n(\theta)$ is a $k \times k$ positive definite matrix depending on θ. If θ is one-dimensional ($k = 1$), then $V_n(\theta)$ is the asymptotic variance as well as the amse of $\hat{\theta}_n$ (§2.5.2). When $k > 1$, $V_n(\theta)$ is called the *asymptotic covariance matrix* of $\hat{\theta}_n$ and can be used as a measure of asymptotic performance of estimators. If $\hat{\theta}_{jn}$ satisfies (4.67) with asymptotic covariance matrix $V_{jn}(\theta)$, $j = 1, 2$, and $V_{1n}(\theta) \le V_{2n}(\theta)$ (in the sense that $V_{2n}(\theta) - V_{1n}(\theta)$ is nonnegative definite) for all $\theta \in \Theta$, then $\hat{\theta}_{1n}$ is said to be asymptotically more efficient than $\hat{\theta}_{2n}$. Of course, some sequences of estimators are not comparable under this criterion. Also, since the asymptotic covariance matrices are unique only in the limiting sense, we have to make our comparison based on their limits. When X_i's are i.i.d., $V_n(\theta)$ is usually of the

form $n^{-\delta}V(\theta)$ for some $\delta > 0$ ($= 1$ in the majority of cases) and a positive definite matrix $V(\theta)$ that does not depend on n.

Note that (4.67) implies that $\hat{\theta}_n$ is an asymptotically unbiased estimator of θ. If $V_n(\theta) = \mathrm{Var}(\hat{\theta}_n)$, then, under some regularity conditions, it follows from Theorem 3.3 that

$$V_n(\theta) \geq [I_n(\theta)]^{-1}, \tag{4.68}$$

where, for every n, $I_n(\theta)$ is the Fisher information matrix (see (3.5)) for X of size n. (Note that (4.68) holds if and only if $lV_n(\theta)l^\tau \geq l[I_n(\theta)]^{-1}l^\tau$ for every $l \in \mathcal{R}^k$.) Unfortunately, when $V_n(\theta)$ is an asymptotic covariance matrix, (4.68) may not hold (even in the limiting sense), even if the regularity conditions in Theorem 3.3 are satisfied.

Example 4.38 (Hodges). Let $X_1, ..., X_n$ be i.i.d. from $N(\theta, 1)$, $\theta \in \mathcal{R}$. Then $I_n(\theta) = n$. Define

$$\hat{\theta}_n = \begin{cases} \bar{X} & |\bar{X}| \geq n^{-1/4} \\ t\bar{X} & |\bar{X}| < n^{-1/4}, \end{cases}$$

where t is a fixed constant. By Proposition 3.2, all conditions in Theorem 3.3 are satisfied. It can be shown (exercise) that (4.67) holds with $V_n(\theta) = V(\theta)/n$, where $V(\theta) = 1$ if $\theta \neq 0$ and $V(\theta) = t^2$ if $\theta = 0$. If $t^2 < 1$, (4.68) does not hold when $\theta = 0$. ∎

However, the following result, due to Le Cam (1953), shows that (4.68) holds for i.i.d. X_i's except for θ in a set of Lebesgue measure 0.

Theorem 4.16. Let $X_1, ..., X_n$ be i.i.d. from a p.d.f. f_θ w.r.t. a σ-finite measure ν on $(\mathcal{R}, \mathcal{B}_\mathcal{R})$, where $\theta \in \Theta$ and Θ is an open set in $\mathcal{R}^k$. Suppose that for every x in the range of X_1, $f_\theta(x)$ is twice continuously differentiable in θ and satisfies

$$\frac{\partial}{\partial\theta}\int \psi_\theta(x)d\nu = \int \frac{\partial}{\partial\theta}\psi_\theta(x)d\nu$$

for $\psi_\theta(x) = f_\theta(x)$ or $= \partial f_\theta(x)/\partial\theta$; the Fisher information matrix

$$\imath(\theta) = E\left[\frac{\partial}{\partial\theta}\log f_\theta(X_1)\right]^\tau \left[\frac{\partial}{\partial\theta}\log f_\theta(X_1)\right]$$

is positive definite; and for any given $\theta \in \Theta$, there exists a positive number c_θ and a positive function h_θ such that $E[h_\theta(X_1)] < \infty$ and

$$\sup_{\gamma:\|\gamma-\theta\|<c_\theta} \left\| \frac{\partial^2 \log f_\gamma(x)}{\partial\gamma\partial\gamma^\tau} \right\| \leq h_\theta(x) \tag{4.69}$$

for all x in the range of X_1. If $\hat{\theta}_n$ is an estimator of θ (based on $X_1, ..., X_n$) and satisfies (4.67) with $V_n(\theta) = V(\theta)/n$, then there is a $\Theta_0 \subset \Theta$ with

Lebesgue measure 0 such that (4.68) holds if $\theta \notin \Theta_0$.

Proof. We adopt the proof given by Bahadur (1964) and prove the case of univariate θ. The proof for multivariate θ is similar and can be found in Bahadur (1964). Let $x = (x_1, ..., x_n)$, $\theta_n = \theta + n^{-1/2} \in \Theta$, and

$$K_n(x, \theta) = [\log \ell(\theta_n) - \log \ell(\theta) + \imath(\theta)/2]/[\imath(\theta)]^{1/2}.$$

Under the assumed conditions, it can be shown (exercise) that

$$K_n(X, \theta) \to_d N(0, 1). \tag{4.70}$$

Let P_{θ_n} (or P_θ) be the distribution of X under the assumption that X_1 has the p.d.f. f_{θ_n} (or f_θ). Define $g_n(\theta) = |P_\theta(\hat{\theta}_n \le \theta) - \frac{1}{2}|$. Let Φ denote the standard normal c.d.f. or its probability measure. By the dominated convergence theorem (Theorem 1.1(i)), as $n \to \infty$,

$$\int g_n(\theta_n) d\Phi(\theta) = \int g_n(\theta) e^{n^{-1/2}\theta - (2n)^{-1}} d\Phi(\theta) \to 0,$$

since $g_n(\theta) \to 0$ under (4.67). By Theorem 1.8(ii) and (vi), there exists a sequence $\{n_k\}$ such that $g_{n_k}(\theta_{n_k}) \to_{a.s.} 0$ w.r.t. Φ. Since Φ is equivalent to the Lebesgue measure, we conclude that there is a $\Theta_0 \subset \Theta$ with Lebesgue measure 0 such that

$$\liminf_n g_n(\theta_n) = 0, \qquad \theta \notin \Theta_0. \tag{4.71}$$

Assume that $\theta \notin \Theta_0$. Then, for any $t > [\imath(\theta)]^{1/2}$,

$$\begin{aligned}
P_{\theta_n}(K_n(X, \theta) \le t) &= \int_{K_n(x,\theta) \le t} \ell(\theta_n) d\nu \times \cdots \times d\nu \\
&= \int_{K_n(x,\theta) \le t} \frac{\ell(\theta_n)}{\ell(\theta)} dP_\theta(x) \\
&= e^{-\imath(\theta)/2} \int_{K_n(x,\theta) \le t} e^{[\imath(\theta)]^{1/2} K_n(x,\theta)} dP_\theta(x) \\
&= e^{-\imath(\theta)/2} \int_{-\infty}^{t} e^{[\imath(\theta)]^{1/2} z} dH_n(z) \\
&= e^{-\imath(\theta)/2} \int_{-\infty}^{t} e^{[\imath(\theta)]^{1/2} z} d\Phi(z) + o(1) \\
&= \Phi\left(t - [\imath(\theta)]^{1/2}\right) + o(1),
\end{aligned}$$

where H_n denotes the distribution of $K_n(X, \theta)$ and the next to last equality follows from (4.70) and the dominated convergence theorem. This result and the fact that

$$\limsup_n P_{(\theta_n)}(\hat{\theta}_n \le \theta_n) \le \tfrac{1}{2}$$

(by (4.71)) imply that there is a sequence $\{n_j\}$ such that for $j = 1, 2, ...,$

$$P_{\theta_{n_j}}(\hat{\theta}_{n_j} \leq \theta_{n_j}) < P_{\theta_{n_j}}(K_{n_j}(X, \theta) \leq t). \tag{4.72}$$

By the Neyman-Pearson lemma (Theorem 6.1 in §6.1.1), we conclude that (4.72) implies that for $j = 1, 2, ...,$

$$P_\theta(\hat{\theta}_{n_j} \leq \theta_{n_j}) < P_\theta(K_{n_j}(X, \theta) \leq t). \tag{4.73}$$

(The reader should come back to this after reading §6.1.1.) From (4.70) and (4.67) with $V_n(\theta) = V(\theta)/n$, (4.73) implies

$$\Phi\big([V(\theta)]^{-1/2}\big) \leq \Phi(t).$$

Hence $[V(\theta)]^{-1/2} \leq t$. Since $I_n(\theta)/n = \imath(\theta)$ (Proposition 3.1(i)) and t is arbitrary, we conclude that (4.68) holds. ∎

Points at which (4.68) do not hold are called points of superefficiency. Motivated by the fact that the set of superefficiency points is of Lebesgue measure 0 under some regularity conditions, we have the following definition.

Definition 4.4. Assume that the Fisher information matrix $I_n(\theta)$ is well defined and positive definite for every n. A sequence of estimators $\{\hat{\theta}_n\}$ satisfying (4.67) is said to be *asymptotically efficient* or *asymptotically optimal* if and only if $V_n(\theta) = [I_n(\theta)]^{-1}$. ∎

Suppose that we are interested in estimating $\vartheta = g(\theta)$, where g is a differentiable function from Θ to $\mathcal{R}^p$, $1 \leq p \leq k$. If $\hat{\theta}_n$ satisfies (4.67), then, by Theorem 1.12(i), $\hat{\vartheta}_n = g(\hat{\theta}_n)$ is asymptotically distributed as $N_p(\vartheta, \nabla g(\theta) V_n(\theta) [\nabla g(\theta)]^\tau)$. Thus, inequality (4.68) becomes

$$\nabla g(\theta) V_n(\theta) [\nabla g(\theta)]^\tau \geq [\tilde{I}_n(\vartheta)]^{-1},$$

where $\tilde{I}_n(\vartheta)$ is the Fisher information matrix about ϑ. If $p = k$ and g is one-to-one, then

$$[\tilde{I}_n(\vartheta)]^{-1} = \nabla g(\theta) [I_n(\theta)]^{-1} [\nabla g(\theta)]^\tau$$

and, therefore, $\hat{\vartheta}_n$ is asymptotically efficient if and only if $\hat{\theta}_n$ is asymptotically efficient. For this reason, in the case of $p < k$, $\hat{\vartheta}_n$ is called asymptotically efficient if $\hat{\theta}_n$ is asymptotically efficient, and we can focus on the estimation of θ only.

4.5.2 Asymptotic efficiency of MLE's and RLE's

We now show that under some regularity conditions, a root of the likelihood equation (RLE), which is a candidate for an MLE, is asymptotically efficient.

Theorem 4.17. Assume the conditions of Theorem 4.16.
(i) There is a sequence of estimators $\{\hat{\theta}_n\}$ such that

$$P\big(s_n(\hat{\theta}_n) = 0\big) \to 1 \qquad \text{and} \qquad \hat{\theta}_n \to_p \theta, \tag{4.74}$$

where $s_n(\gamma) = \partial \log \ell(\gamma)/\partial\gamma$.
(ii) Any consistent sequence $\tilde{\theta}_n$ of RLE's is asymptotically efficient.
Proof. (i) Let $B_n(c) = \{\gamma : \|(\gamma - \theta)[I_n(\theta)]^{1/2}\| \le c\}$ for $c > 0$. Since Θ is open, for each $c > 0$, $B_n(c) \subset \Theta$ for sufficiently large n. Since $B_n(c)$ shrinks to $\{\theta\}$ as $n \to \infty$, the existence of $\hat{\theta}_n$ satisfying (4.74) is implied by that for any $\epsilon > 0$, there exists $c > 0$ and $n_0 > 1$ such that

$$P\big(\log \ell(\gamma) - \log \ell(\theta) < 0 \quad \text{for all } \gamma \in \partial B_n(c)\big) \ge 1 - \epsilon, \qquad n \ge n_0, \tag{4.75}$$

where $\partial B_n(c)$ is the boundary of $B_n(c)$. For $\gamma \in \partial B_n(c)$, the Taylor expansion gives

$$\begin{aligned}\log \ell(\gamma) - \log \ell(\theta) = {} & c s_n(\theta)[I_n(\theta)]^{-1/2}\lambda^\tau \\ & + (c^2/2)\lambda[I_n(\theta)]^{-1/2}\nabla s_n(\gamma^*)[I_n(\theta)]^{-1/2}\lambda^\tau,\end{aligned} \tag{4.76}$$

where $\lambda = (\gamma - \theta)[I_n(\theta)]^{1/2}/c$ satisfying $\|\lambda\| = 1$, $\nabla s_n(\gamma) = \partial s_n(\gamma)/\partial\gamma$, and γ^* lies between γ and θ. Note that

$$\begin{aligned} E\frac{\|\nabla s_n(\gamma^*) - \nabla s_n(\theta)\|}{n} &\le E \max_{\gamma \in B_n(c)} \frac{\|\nabla s_n(\gamma) - \nabla s_n(\theta)\|}{n} \\ &\le E \max_{\gamma \in B_n(c)} \left\| \frac{\partial^2 \log f_\gamma(X_1)}{\partial\gamma\partial\gamma^\tau} - \frac{\partial^2 \log f_\theta(X_1)}{\partial\theta\partial\theta^\tau} \right\| \\ &\to 0, \end{aligned} \tag{4.77}$$

which follows from (a) $\partial^2 \log f_\gamma(x)/\partial\gamma\partial\gamma^\tau$ is continuous in a neighborhood of θ for any fixed x; (b) $B_n(c)$ shrinks to $\{\theta\}$; and (c) for sufficiently large n,

$$\max_{\gamma \in B_n(c)} \left\| \frac{\partial^2 \log f_\gamma(X_1)}{\partial\gamma\partial\gamma^\tau} - \frac{\partial^2 \log f_\theta(X_1)}{\partial\theta\partial\theta^\tau} \right\| \le 2h_\theta(X_1)$$

under condition (4.69). By the SLLN (Theorem 1.13) and Proposition 3.1, $n^{-1}\nabla s_n(\theta) \to_{a.s.} -\imath(\theta)$. These results, together with (4.76), imply that

$$\log \ell(\gamma) - \log \ell(\theta) = c s_n(\theta)[I_n(\theta)]^{-1/2}\lambda^\tau - c^2/2 + o_p(1). \tag{4.78}$$

Note that $\max_\lambda\{s_n(\theta)[I_n(\theta)]^{-1/2}\lambda^\tau\} = \|s_n(\theta)[I_n(\theta)]^{-1/2}\|$. Hence, (4.75) follows from (4.78) and

$$\begin{aligned} P\big(\|s_n(\theta)[I_n(\theta)]^{-1/2}\| < c/4\big) &\geq 1 - (4/c)^2 E\|s_n(\theta)[I_n(\theta)]^{-1/2}\|^2 \\ &= 1 - k(4/c)^2 \\ &\geq 1 - \epsilon \end{aligned}$$

by choosing c sufficiently large. This completes the proof of (i).
(ii) Let $A_\epsilon = \{\gamma : \|\gamma - \theta\| \leq \epsilon\}$ for $\epsilon > 0$. Since Θ is open, $A_\epsilon \subset \Theta$ for sufficiently small ϵ. Let $\{\tilde\theta_n\}$ be a sequence of consistent RLE's, i.e., $P(s_n(\tilde\theta_n) = 0$ and $\tilde\theta_n \in A_\epsilon) \to 1$ for any $\epsilon > 0$. Hence, we can focus on the set on which $s_n(\tilde\theta_n) = 0$ and $\tilde\theta_n \in A_\epsilon$. Using the mean-value theorem for vector-valued functions, we obtain that

$$-s_n(\theta) = (\tilde\theta_n - \theta)\int_0^1 \nabla s_n\big(\theta + t(\tilde\theta_n - \theta)\big)dt.$$

Note that

$$\frac{1}{n}\left\|\int_0^1 \nabla s_n\big(\theta + t(\tilde\theta_n - \theta)\big)dt - \nabla s_n(\theta)\right\| \leq \max_{\gamma\in A_\epsilon}\frac{\|\nabla s_n(\gamma) - \nabla s_n(\theta)\|}{n}.$$

Using the argument in proving (4.77) and the fact that $P(\tilde\theta_n \in A_\epsilon) \to 1$ for arbitrary $\epsilon > 0$, we obtain that

$$\frac{1}{n}\left\|\int_0^1 \nabla s_n\big(\theta + t(\tilde\theta_n - \theta)\big)dt - \nabla s_n(\theta)\right\| \to_p 0.$$

Since $n^{-1}\nabla s_n(\theta) \to_{a.s.} -\imath(\theta)$ and $I_n(\theta)/n = \imath(\theta)$,

$$-s_n(\theta) = -(\tilde\theta_n - \theta)I_n(\theta) + o_p\big(\|(\tilde\theta_n - \theta)I_n(\theta)\|\big).$$

This and Slutsky's theorem (Theorem 1.11) imply that $\sqrt{n}(\tilde\theta_n - \theta)$ has the same asymptotic distribution as

$$\sqrt{n}s_n(\theta)[I_n(\theta)]^{-1} = n^{-1/2}s_n(\theta)[\imath(\theta)]^{-1} \to_d N_k\big(0, [\imath(\theta)]^{-1}\big)$$

by the CLT (Corollary 1.2), since $\mathrm{Var}(s_n(\theta)) = I_n(\theta)$. ∎

Theorem 4.17(i) shows the asymptotic existence of a sequence of consistent RLE's, and Theorem 4.17(ii) shows the asymptotic efficiency of *any* sequence of consistent RLE's. However, for a given sequence of RLE's, its consistency has to be checked, unless the RLE's are unique for sufficiently large n, in which case the consistency of the RLE's is guaranteed by Theorem 4.17(i).

RLE's are not necessarily MLE's. We still have to use the techniques discussed in §4.4 to check whether an RLE is an MLE. However, according to Theorem 4.17, when a sequence of RLE's is consistent, then it is asymptotically efficient and, therefore, we may not need to search for MLE's, if asymptotic efficiency is the only criterion to select estimators. The method of estimating θ by solving $s_n(\gamma) = 0$ over $\gamma \in \Theta$ is called *scoring* and the function $s_n(\gamma)$ is called the *score* function.

Example 4.39. Suppose that X_i has a distribution in a natural exponential family, i.e., the p.d.f. of X_i is

$$f_\eta(x_i) = \exp\{T(x_i)\eta^\tau - \zeta(\eta)\}h(x_i). \tag{4.79}$$

Since $\partial^2 \log f_\eta(x_i)/\partial\eta\partial\eta^\tau = \partial^2\zeta(\eta)/\partial\eta\partial\eta^\tau$, condition (4.69) is satisfied. From Proposition 3.2, other conditions in Theorem 4.16 are also satisfied. For i.i.d. X_i's,

$$s_n(\gamma) = \sum_{i=1}^n \left[T(X_i) - \frac{\partial\zeta(\eta)}{\partial\eta}\right].$$

If $\hat{\theta}_n = n^{-1}\sum_{i=1}^n T(X_i) \in \Theta$, the range of $\theta = g(\eta) = \partial\zeta(\eta)/\partial\eta$, then $\hat{\theta}_n$ is a unique RLE of θ, which is also a unique MLE of θ since $\partial^2\zeta(\eta)/\partial\eta\partial\eta^\tau = \mathrm{Var}(T(X_i))$ is positive definite. Also, $\eta = g^{-1}(\theta)$ exists and a unique RLE (MLE) of η is $\hat{\eta}_n = g^{-1}(\hat{\theta}_n)$.

However, $\hat{\theta}_n$ may not be in Θ and the previous argument fails (e.g., Example 4.29). What Theorem 4.17 tells us in this case is that as $n \to \infty$, $P(\hat{\theta}_n \in \Theta) \to 1$ and, therefore, $\hat{\theta}_n$ (or $\hat{\eta}_n$) is the unique asymptotically efficient RLE (MLE) of θ (or η) in the limiting sense.

In an example like this we can directly show that $P(\hat{\theta}_n \in \Theta) \to 1$, using the fact that $\hat{\theta}_n \to_{a.s.} E[T(X_1)] = g(\eta)$ (the SLLN). ∎

The next theorem provides a similar result for the MLE or RLE in the GLM (§4.4.2).

Theorem 4.18. Consider the GLM (4.55)-(4.58) with t_i's in a fixed interval (t_0, t_∞), $0 < t_0 \le t_\infty < \infty$. Assume that the range of the unknown parameter β in (4.57) is an open subset of $\mathcal{R}^p$; at the true parameter value β, $0 < \inf_i \varphi(\beta Z_i^\tau) \le \sup_i \varphi(\beta Z_i^\tau) < \infty$, where $\varphi(t) = [\psi'(t)]^2\zeta''(\psi(t))$; as $n \to \infty$, $\max_{i\le n} Z_i(Z^\tau Z)^{-1}Z_i^\tau = 0$ and $\lambda_-[Z^\tau Z] \to \infty$, where Z is the $n \times p$ matrix whose ith row is Z_i and $\Lambda_-[A]$ is the smallest eigenvalue of the matrix A.
(i) There is a unique sequence $\{\hat{\beta}_n\}$ such that

$$P(s_n(\hat{\beta}_n) = 0) \to 1 \qquad \text{and} \qquad \hat{\beta}_n \to_p \beta, \tag{4.80}$$

where $s_n(\gamma)$ is the score function defined to be the left-hand side of (4.59) with $\gamma = \beta$.
(ii) Let $I_n(\beta) = \mathrm{Var}(s_n(\beta))$. Then

$$(\hat{\beta}_n - \beta)[I_n(\beta)]^{1/2} \to_d N_p(0, I_p). \tag{4.81}$$

(iii) If ϕ in (4.58) is known or the p.d.f. in (4.55) indexed by $\theta = (\beta, \phi)$ satisfies the conditions for f_θ in Theorem 4.16, then $\hat{\beta}_n$ is asymptotically efficient.
Proof. (i) The proof of the existence of $\hat{\beta}_n$ satisfying (4.80) is the same as that of Theorem 4.17(i) with $\theta = \beta$, except that we need to show

$$\max_{\gamma \in B_n(c)} \left\| [I_n(\beta)]^{-1/2} \nabla s_n(\gamma) [I_n(\beta)]^{-1/2} - I_p \right\| \to_p 0,$$

where $B_n(c) = \{\gamma : \|(\gamma - \beta)[I_n(\beta)]^{1/2}\| \le c\}$. From (4.62) and (4.63), $I_n(\beta) = M_n(\beta)/\phi$ and $\nabla s_n(\gamma) = [R_n(\gamma) - M_n(\gamma)]/\phi$, where $M_n(\gamma)$ and $R_n(\gamma)$ are defined by (4.60)-(4.61) with $\gamma = \beta$. Hence, it suffices to show that for any $c > 0$,

$$\max_{\gamma \in B_n(c)} \left\| [M_n(\beta)]^{-1/2} [M_n(\gamma) - M_n(\beta)] [M_n(\beta)]^{-1/2} \right\| \to 0 \tag{4.82}$$

and

$$\max_{\gamma \in B_n(c)} \left\| [M_n(\beta)]^{-1/2} R_n(\gamma) [M_n(\beta)]^{-1/2} \right\| \to 0. \tag{4.83}$$

The left-hand side of (4.82) is bounded by

$$\sqrt{p} \max_{\gamma \in B_n(c), i \le n} \left| 1 - \varphi(\gamma Z_i^\tau)/\varphi(\beta Z_i^\tau) \right|,$$

which converges to 0 since φ is continuous and for $\gamma \in B_n(c)$,

$$\begin{aligned}
|\gamma Z_i^\tau - \beta Z_i^\tau|^2 &= |(\gamma - \beta)[I_n(\beta)]^{1/2} [I_n(\beta)]^{-1/2} Z_i^\tau|^2 \\
&\le \|(\gamma - \beta)[I_n(\beta)]^{1/2}\|^2 \|[I_n(\beta)]^{-1/2} Z_i^\tau\|^2 \\
&\le c^2 \max_{i \le n} Z_i [I_n(\beta)]^{-1} Z_i^\tau \\
&\le c^2 \phi \big[t_0 \inf_i \varphi(\beta Z_i^\tau) \big]^{-1} \max_{i \le n} Z_i (Z^\tau Z)^{-1} Z_i^\tau \\
&\to 0
\end{aligned}$$

under the assumed conditions. This proves (4.82).

Let $e_i = X_i - \mu(\psi(\beta Z_i^\tau))$,

$$U_n(\gamma) = \sum_{i=1}^n [\mu(\psi(\beta Z_i^\tau)) - \mu(\psi(\gamma Z_i^\tau))] \psi''(\gamma Z_i^\tau) t_i Z_i^\tau Z_i,$$

$$V_n(\gamma) = \sum_{i=1}^{n} e_i[\psi''(\gamma Z_i^\tau) - \psi''(\beta Z_i^\tau)]t_i Z_i^\tau Z_i,$$

and

$$W_n(\beta) = \sum_{i=1}^{n} e_i \psi''(\beta Z_i^\tau) t_i Z_i^\tau Z_i.$$

Then $R_n(\gamma) = U_n(\gamma) + V_n(\gamma) + W_n(\beta)$. Using the same argument as that in proving (4.82), we can show that

$$\max_{\gamma \in B_n(c)} \left\| [M_n(\beta)]^{-1/2} U_n(\gamma) [M_n(\beta)]^{-1/2} \right\| \to 0.$$

Note that $\left\| [M_n(\beta)]^{-1/2} V_n(\gamma) [M_n(\beta)]^{-1/2} \right\|$ is bounded by the product of

$$[M_n(\beta)]^{-1/2} \sum_{i=1}^{n} |e_i| t_i Z_i^\tau Z_i [M_n(\beta)]^{-1/2} = O_p(1)$$

and

$$\max_{\gamma \in B_n(c), i \le n} \left| \psi''(\gamma Z_i^\tau) - \psi''(\beta Z_i^\tau) \right|,$$

which can be shown to be $o(1)$ using the same argument as that in proving (4.82). Hence,

$$\max_{\gamma \in B_n(c)} \left\| [M_n(\beta)]^{-1/2} V_n(\gamma) [M_n(\beta)]^{-1/2} \right\| \to 0$$

and (4.83) follows from

$$\left\| [M_n(\beta)]^{-1/2} W_n(\beta) [M_n(\beta)]^{-1/2} \right\| \to 0.$$

To show this result, we apply Theorem 1.14(ii). Since $E(e_i) = 0$ and e_i's are independent, it suffices to show that

$$\sum_{i=1}^{n} E\left| e_i \psi''(\beta Z_i^\tau) t_i Z_i [M_n(\beta)]^{-1} Z_i^\tau \right|^{1+\delta} \to 0 \tag{4.84}$$

for some $\delta \in (0,1)$. Note that $\sup_i E|e_i|^{1+\delta} < \infty$. Hence, there is a constant $C > 0$ such that the left-hand side of (4.84) is bounded by

$$C \sum_{i=1}^{n} \left| Z_i (Z^\tau Z)^{-1} Z_i^\tau \right|^{1+\delta} \le pC \max_{i \le n} |Z_i (Z^\tau Z)^{-1} Z_i^\tau|^\delta \to 0.$$

Hence, (4.84) follows from Theorem 1.14(ii). This proves (4.80). The uniqueness of $\hat{\beta}_n$ follows from (4.83) and the fact that $M_n(\gamma)$ is positive definite in a neighborhood of β. This completes the proof of (i).

(ii) The proof of (ii) is very similar to that of Theorem 4.17(ii). Using the results in the proof of (i) and Taylor's expansion, we can establish (exercise) that

$$(\hat{\beta}_n - \beta)[I_n(\beta)]^{1/2} = s_n(\beta)[I_n(\beta)]^{-1/2} + o_p(1). \tag{4.85}$$

Using the CLT (e.g., Corollary 1.3) and Theorem 1.9(iii), we can show (exercise) that

$$s_n(\beta)[I_n(\beta)]^{-1/2} \to_d N_p(0, I_p). \tag{4.86}$$

Result (4.81) follows from (4.85)-(4.86) and Slutsky's theorem.
(iii) The result is obvious if ϕ is known. When ϕ is unknown, it follows from (4.59) that

$$\frac{\partial}{\partial\phi}\left[\frac{\partial \log \ell(\theta)}{\partial\beta}\right] = -\frac{s_n(\beta)}{\phi}.$$

Since $E[s_n(\beta)] = 0$, the Fisher information about $\theta = (\beta, \phi)$ is

$$I_n(\beta, \phi) = -E\left[\frac{\partial^2 \log \ell(\theta)}{\partial\theta\partial\theta^\tau}\right] = \begin{pmatrix} I_n(\beta) & 0 \\ 0 & \tilde{I}_n(\phi) \end{pmatrix},$$

where $\tilde{I}_n(\phi)$ is the Fisher information about ϕ. The result then follows from (4.81) and the discussion in the end of §4.5.1. ∎

4.5.3 Other asymptotically efficient estimators

To study other asymptotically efficient estimators, we start with MRIE's in location-scale families. Since MLE's and RLE's are invariant (see Exercise 95 in §4.6), MRIE's are often asymptotically efficient; see, for example, Stone (1974).

Assume the conditions in Theorem 4.16 and let $s_n(\gamma)$ be the score function. Let $\hat{\theta}_n^{(0)}$ be an estimator of θ that may not be asymptotically efficient. The estimator

$$\hat{\theta}_n^{(1)} = \hat{\theta}_n^{(0)} - s_n(\hat{\theta}_n^{(0)})[\nabla s_n(\hat{\theta}_n^{(0)})]^{-1} \tag{4.87}$$

is the first iteration in computing an MLE (or RLE) using the Newton-Raphson iteration method with $\hat{\theta}_n^{(0)}$ as the initial value (see (4.53)) and, therefore, is called the *one-step* MLE. Without any further iteration, $\hat{\theta}_n^{(1)}$ can be used as a numerical approximation to an MLE or RLE; and $\hat{\theta}_n^{(1)}$ is asymptotically efficient under some conditions, as the following result shows.

Theorem 4.19. Assume that the conditions in Theorem 4.16 hold and that $\hat{\theta}_n^{(0)}$ is $\sqrt{n}$-consistent for θ (Definition 2.10).
(i) The one-step MLE $\hat{\theta}_n^{(1)}$ is asymptotically efficient.
(ii) The one-step MLE obtained by replacing $\nabla s_n(\gamma)$ in (4.87) with its

expected value, $-I_n(\gamma)$ (the Fisher-scoring method), is asymptotically efficient.

Proof. Since $\hat{\theta}_n^{(0)}$ is $\sqrt{n}$-consistent, we can focus on the event $\hat{\theta}_n^{(0)} \in A_\epsilon = \{\gamma : \|\gamma - \theta\| \le \epsilon\}$ for a sufficiently small ϵ such that $A_\epsilon \subset \Theta$. From the mean-value theorem,

$$s_n(\hat{\theta}_n^{(0)}) = s_n(\theta) + (\hat{\theta}_n^{(0)} - \theta) \int_0^1 \nabla s_n\big(\theta + t(\hat{\theta}_n^{(0)} - \theta)\big) dt.$$

Substituting this into (4.87) we obtain that

$$\hat{\theta}_n^{(1)} - \theta = -s_n(\theta)[\nabla s_n(\hat{\theta}_n^{(0)})]^{-1} + (\hat{\theta}_n^{(0)} - \theta)[I_k - G_n(\hat{\theta}_n^{(0)})],$$

where

$$G_n(\hat{\theta}_n^{(0)}) = \int_0^1 \nabla s_n\big(\theta + t(\hat{\theta}_n^{(0)} - \theta)\big) dt [\nabla s_n(\hat{\theta}_n^{(0)})]^{-1}.$$

From (4.77), $[I_n(\theta)]^{1/2}[\nabla s_n(\hat{\theta}_n^{(0)})]^{-1}[I_n(\theta)]^{1/2} \to_p -I_k$. Using an argument similar to the proofs of (4.77) and (4.82), we can show that $G_n(\hat{\theta}_n^{(0)}) \to_p I_k$. These results and the fact that $\sqrt{n}(\hat{\theta}_n^{(0)} - \theta) = O_p(1)$ implies

$$\sqrt{n}(\hat{\theta}_n^{(1)} - \theta) = \sqrt{n} s_n(\theta)[I_n(\theta)]^{-1} + o_p(1).$$

This proves (i). The proof for (ii) is similar. ∎

Example 4.40. Let $X_1, ..., X_n$ be i.i.d. from the Weibull distribution $W(\theta, 1)$, where $\theta > 0$ is unknown. Note that

$$s_n(\theta) = \frac{n}{\theta} + \sum_{i=1}^n \log X_i - \sum_{i=1}^n X_i^\theta \log X_i$$

and

$$\nabla s_n(\theta) = -\frac{n}{\theta^2} - \sum_{i=1}^n X_i^\theta (\log X_i)^2.$$

Hence, the one-step MLE of θ is

$$\hat{\theta}_n^{(1)} = \hat{\theta}_n^{(0)} \left[1 + \frac{n + \hat{\theta}_n^{(0)} (\sum_{i=1}^n \log X_i - \sum_{i=1}^n X_i^{\hat{\theta}_n^{(0)}} \log X_i)}{n + (\hat{\theta}_n^{(0)})^2 \sum_{i=1}^n X_i^{\hat{\theta}_n^{(0)}} (\log X_i)^2} \right].$$

Usually one can use a moment estimator (§3.5.2) as the initial estimator $\hat{\theta}_n^{(0)}$. In this example, a moment estimator of θ is the solution of $\bar{X} = \Gamma(\theta^{-1} + 1)$. ∎

Results similar to that in Theorem 4.19 can be obtained in non-i.i.d. cases, for example, the GLM discussed in §4.4.2 (exercise); see also §5.4.

As we discussed in §4.1.3, Bayes estimators are usually consistent. The next result, due to Bickel and Yahav (1969) and Ibragimov and Has'minskii (1972), states that Bayes estimators are asymptotically efficient when X_i's are i.i.d.

Theorem 4.20. Assume the conditions of Theorem 4.16. Let $\pi(\gamma)$ be a prior p.d.f. (which may be improper) w.r.t. the Lebesgue measure on Θ and $p_n(\gamma)$ be the posterior p.d.f., given $X_1, ..., X_n$, $n = 1, 2, ...$ Assume that there exists an n_0 such that $p_{n_0}(\gamma)$ is continuous and positive for all $\gamma \in \Theta$, $\int p_{n_0}(\gamma)d\gamma = 1$ and $\int \|\gamma\| p_{n_0}(\gamma)d\gamma < \infty$. Suppose further that for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$\lim_{n\to\infty} P\left(\sup_{\|\gamma-\theta\|\geq\epsilon} \frac{\log \ell(\gamma) - \log \ell(\theta)}{n} > -\delta\right) = 0 \tag{4.88}$$

and

$$\lim_{n\to\infty} P\left(\sup_{\|\gamma-\theta\|\leq\delta} \frac{\|\nabla s_n(\gamma) - \nabla s_n(\theta)\|}{n} \geq \epsilon\right) = 0, \tag{4.89}$$

where $\ell(\gamma)$ is the likelihood function and $s_n(\gamma)$ is the score function.
(i) Let $p_n^*(\gamma)$ be the posterior p.d.f. of $\sqrt{n}(\gamma - T_n)$, where $T_n = \theta + s_n(\theta)[I_n(\theta)]^{-1}$ and θ is the true parameter value, and let $\psi_n(\gamma)$ be the p.d.f. of $N_k(0, [I_n(\theta)]^{-1})$. Then

$$\int (1 + \|\gamma\|)\big|p_n^*(\gamma) - \psi_n(\gamma)\big|d\gamma \to_p 0. \tag{4.90}$$

(ii) The Bayes estimator of θ under the squared error loss is asymptotically efficient. ∎

The proof of Theorem 4.20 is lengthy and is omitted; see Lehmann (1983, §6.7) for a proof of the case of univariate θ.

A number of conclusions can be drawn from Theorem 4.20. First, result (4.90) shows that the posterior p.d.f. is approximately normal with mean $\theta + s_n(\theta)[I_n(\theta)]^{-1}$ and covariance matrix $[I_n(\theta)]^{-1}$. This result is useful in Bayesian computation; see Berger (1985, §4.9.3). Second, (4.90) shows that the posterior distribution and its first-order moments converge to the degenerate distribution at θ and its first-order moments, which implies the consistency and asymptotic unbiasedness of Bayes estimators such as the posterior means. Third, the Bayes estimator under the squared error loss is asymptotically efficient, which provides an additional support for the early suggestion that the Bayesian approach is a useful method for generating estimators. Finally, the results hold regardless of the prior being used, indicating that the effect of the prior declines as n increases.

In addition to the regularity conditions in Theorem 4.16, Theorem 4.20 requires two more nontrivial regularity conditions, (4.88) and (4.89). Let us verify these conditions for natural exponential families (Example 4.39), i.e., X_i's are i.i.d. with p.d.f. (4.79). Since $\nabla s_n(\eta) = -n\partial^2\zeta(\eta)/\partial\eta\partial\eta^\tau$, (4.89) follows from the continuity of the second-order derivatives of ζ. To show (4.88), consider first the case of univariate η. Without loss of generality we assume that $\gamma > \eta$. Note that

$$\frac{\log \ell(\gamma) - \log \ell(\eta)}{n} = (\gamma - \eta)\left[\bar{T} - \zeta'(\eta) + \zeta'(\eta) - \frac{\zeta(\gamma) - \zeta(\eta)}{\gamma - \eta}\right], \quad (4.91)$$

where $\bar{T}$ is the average of $T(X_i)$'s. Since $\zeta(\gamma)$ is strictly convex, $\gamma > \eta$ implies $\zeta'(\eta) < [\zeta(\gamma) - \zeta(\eta)]/(\gamma - \eta)$. Also, $T \to_{a.s.} \zeta'(\eta)$. Hence, with probability tending to 1, the factor of $(\gamma - \eta)$ on the right-hand side of (4.91) is negative. Then (4.88) holds with

$$\delta = \frac{\epsilon}{2} \inf_{\gamma > \theta} \left[\frac{\zeta(\gamma) - \zeta(\eta)}{\gamma - \eta} - \zeta'(\theta)\right].$$

To show how to extend this to multivariate η, consider the case of bivariate η. Let η_j, γ_j, and ξ_j be the jth components of η, γ, and $\bar{T} - \nabla\zeta(\eta)$, respectively. Assume $\gamma_1 > \eta_1$ and $\gamma_2 > \eta_2$. Let ζ_j' be the derivative of ζ w.r.t. the jth component of η. Then the left-hand side of (4.91) is the sum of

$$(\gamma_1 - \theta_1)\xi_1 - [\zeta(\eta_1, \gamma_2) - \zeta(\eta_1, \eta_2) - (\gamma_2 - \eta_2)\zeta_2'(\eta_1, \eta_2)]$$

and

$$\begin{aligned}&(\gamma_2 - \theta_2)\xi_2 - [\zeta(\gamma_1, \gamma_2) - \zeta(\eta_1, \gamma_2) - (\gamma_1 - \eta_1)\zeta_1'(\eta_1, \eta_2)]\\ &\le (\gamma_2 - \theta_2)\xi_2 - [\zeta(\gamma_1, \gamma_2) - \zeta(\eta_1, \gamma_2) - (\gamma_1 - \eta_1)\zeta_1'(\eta_1, \gamma_2)],\end{aligned}$$

since $\zeta_1'(\eta_1, \eta_2) \le \zeta_1'(\eta_1, \gamma_2)$. The rest of the proof is the same as the case of univariate η.

When Bayes estimators have explicit forms under a specific prior, it is usually easy to prove the asymptotic efficiency of the Bayes estimators directly. For instance, in Example 4.7, the Bayes estimator of θ is

$$\frac{n\bar{X} + \gamma^{-1}}{n + \alpha - 1} = \bar{X} + \frac{(\alpha - 1) + \gamma^{-1}}{n + \alpha - 1} = \bar{X} + O_p(n^{-1}),$$

where $\bar{X}$ is the MLE of θ. Hence the Bayes estimator is asymptotically efficient by Slutsky's theorem. A similar result can be obtained for the Bayes estimator (4.8) in Example 4.7. Theorem 4.20, however, is useful in cases where Bayes estimators do not have explicit forms and/or the prior is not specified clearly. One such example is the problem in Example 4.40 (Exercises 129 and 130).

4.6 Exercises

1. Show that the priors in the following cases are conjugate priors:
(a) $X_1, ..., X_n$ are i.i.d. from $N_k(\theta, I_k)$, $\theta \in \mathcal{R}^k$, and $\Pi = N_k(\mu_0, \Sigma_0)$ (Normal family);
(b) $X_1, ..., X_n$ are i.i.d. from the binomial distribution $Bi(\theta, k)$, $\theta \in (0, 1)$, and $\Pi = B(\alpha, \beta)$ (Beta family);
(c) $X_1, ..., X_n$ are i.i.d. from the uniform distribution $U(0, \theta)$, $\theta > 0$, and $\Pi = Pa(a, b)$ (Pareto family);
(d) $X_1, ..., X_n$ are i.i.d. from the exponential distribution $E(0, \theta)$, $\theta > 0$, $\Pi =$ the inverse gamma distribution $\Gamma^{-1}(\alpha, \gamma)$ (a random variable Y has the inverse gamma distribution $\Gamma^{-1}(\alpha, \gamma)$ if and only if Y^{-1} has the gamma distribution $\Gamma(\alpha, \gamma)$);
(e) X_1 is from the binomial distribution $Bi(p, \theta)$ with a known p, $\theta = 1, 2, ...$, and $\Pi = P(\lambda)$ (Poisson family).

2. In (a)-(e) of Exercise 1, find the posterior mean for each case.

3. Show that if $T(X)$ is a sufficient statistic for $\theta \in \Theta$, then the Bayes action $\delta(x)$ in (4.3) is a function of $T(x)$.

4. Let $X_1, ..., X_n$ be i.i.d. from the $N(\theta, 1)$ distribution and let Π be the double exponential distribution $DE(0, 1)$. Obtain the Bayes action under the squared error loss.

5. Let X be a single observation from $N(\mu, \sigma^2)$ with a known σ^2 and an unknown $\mu > 0$. Consider the estimation of μ under the squared error loss and the noninformative prior $\Pi =$ the Lebesgue measure on $(0, \infty)$. Show that the Bayes action when $X = x$ is $\delta(x) = x + \Phi'(x/\sigma)/[1 - \Phi(-x/\sigma)]$, where Φ is the c.d.f. of the standard normal distribution and Φ' is its derivative.

6. Consider the estimation problem in Example 4.1 with the loss function $L(\theta, a) = w(\theta)[g(\theta) - a]^2$, where $w(\theta) \geq 0$ and $\int_\Theta w(\theta)[g(\theta)]^2 d\Pi < \infty$. Show that the Bayes action is
$$\delta(x) = \frac{\int_\Theta w(\theta) g(\theta) f_\theta(x) d\Pi}{\int_\Theta w(\theta) f_\theta(x) d\Pi}.$$

7. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $P(X_1 = 1) = p \in (0, 1)$. Consider the estimation of p under the loss function $L(p, a) = (p - a)^2/[p(1 - p)]$. Find the Bayes action w.r.t. the uniform prior on $[0, 1]$.

8. Consider Example 4.1 with $g(\theta) = \theta \in \mathcal{R}$. Under the loss function $L(\theta, a) = |\theta - a|$, show that a median of the posterior distribution is a Bayes action (see Exercise 75 in §2.6).

9. Let X be a sample of size 1 from the geometric distribution $G(p)$ with an unknown $p \in (0,1]$. Consider the estimation of p with $\mathbb{A} = [0,1]$ and the loss function $L(p,a) = (p-a)^2/p$.
(a) Show that δ is a Bayes action w.r.t. Π if and only if $\delta(x) = 1 - \int (1-p)^x d\Pi(p) / \int (1-p)^{x-1} d\Pi(p)$, $x = 1, 2, ...$
(b) Let δ_0 be a rule such that $\delta_0(1) = 1/2$ and $\delta_0(x) = 0$ for all $x > 1$. Show that δ_0 is a limit of Bayes actions.
(c) Let δ_0 be a rule such that $\delta_0(x) = 0$ for all $x > 1$ and $\delta_0(1)$ is arbitrary. Show that δ_0 is a generalized Bayes action.

10. Let X be a sample from P_θ having the p.d.f. $h(x)\exp\{x\theta^\tau - \zeta(\theta)\}$ w.r.t. ν. Let Π be the Lebesgue measure on $\Theta = \mathcal{R}^p$. Show that the generalized Bayes action under the loss $L(\theta, a) = \|E(X) - a\|^2$ is $\delta(x) = x$.

11. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$, where μ and σ^2 are unknown. Let the prior for (μ, σ^2) have the improper Lebesgue density $\pi(\mu, \sigma^2) = \sigma^{-2} I_{(0,\infty)}(\sigma^2)$.
(a) Show that the posterior p.d.f. of (μ, σ^2) given $x = (x_1, ..., x_n)$ is $\pi(\mu, \sigma^2|x) = \pi_1(\mu|\sigma^2, x)\pi_2(\sigma^2|x)$, where $\pi_1(\mu|\sigma^2, x)$ is the p.d.f. of $N(\bar{x}, \sigma^2/n)$ and $\pi_2(\sigma^2|x)$ is the p.d.f. of the inverse gamma distribution $\Gamma^{-1}((n-1)/2, [\sum_{i=1}^n (x_i - \bar{x})^2/2]^{-1})$ (see Exercise 1(d)).
(b) Show that the marginal posterior p.d.f. of σ^2 given x is the p.d.f. of the inverse gamma distribution $\Gamma^{-1}((n-1)/2, [\sum_{i=1}^n (x_i - \bar{x})^2/2]^{-1})$.
(c) Show that the marginal posterior p.d.f. of μ given x is $f\left(\frac{\mu - \bar{x}}{\tau}\right)$, where $\tau^2 = \sum_{i=1}^n (x_i - \bar{x})^2/[n(n-1)]$ and f is the p.d.f. of the t-distribution t_{n-1}.

12. Let X be a sample from P_θ, $\theta \in \Theta \subset \mathcal{R}$. Consider the estimation of θ under the loss $L(|\theta - a|)$, where L is an increasing function on $[0, \infty)$. Let $\pi(\theta|x)$ be the posterior p.d.f. of θ given $X = x$. Suppose that $\pi(\theta|x)$ is symmetric and unimodal. Show that δ satisfying $\pi(\delta|x) = \sup_{\theta \in \Theta} \pi(\theta|x)$ is a Bayes action, assuming that all integrals involved are finite.

13. (Bayesian hypothesis testing). Let X be a sample from P_θ, where $\theta \in \Theta$. Let $\Theta_0 \subset \Theta$ and $\Theta_1 = \Theta_0^c$, the complement of Θ_0. Consider the problem of testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ under the loss

$$L(\theta, a_i) = \begin{cases} 0 & \theta \in \Theta_i \\ C_i & \theta \notin \Theta_i, \end{cases}$$

where $C_i > 0$ are known constants and $\{a_0, a_1\}$ is the action space. Let $\Pi_{\theta|x}$ be the posterior distribution of θ w.r.t. a prior distribution Π, given $X = x$. Show that the Bayes action $\delta(x) = a_1$ if and only if $\Pi_{\theta|x}(\Theta_1) > C_1/(C_0 + C_1)$.

14. Let X be a single observation from the Lebesgue p.d.f. $e^{-x+\theta}I_{(\theta,\infty)}(x)$, where $\theta > 0$ is an unknown parameter. Consider the estimation of

$$\vartheta = \begin{cases} j & \theta \in (j-1, j],\ j = 1, 2, 3, \\ 4 & \theta > 3 \end{cases}$$

under the loss $L(i, j)$, $1 \le i, j \le 4$, given by the following matrix

$$\begin{pmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 2 \\ 1 & 2 & 0 & 2 \\ 3 & 3 & 3 & 0 \end{pmatrix}.$$

When $X = 4$, find the Bayes action w.r.t. the prior with the Lebesgue p.d.f. $e^{-\theta}I_{(0,\infty)}(\theta)$.

15. In (b)-(d) of Exercise 1, assume that the parameters in priors are unknown. Using the method of moments, find empirical Bayes actions under the squared error loss.

16. In Example 4.5, assume that both μ_0 and σ_0^2 in the prior for μ are unknown. Let the second-stage joint prior for (μ_0, σ_0^2) be the product of $N(a, v^2)$ and the Lebesgue measure on $(0, \infty)$, where a and v are known. Under the squared error loss, obtain a formula for the hierarchical Bayes action in terms of a one-dimensional integral.

17. Let δ_i be a Bayes estimator of ϑ_i under the squared error loss, $i = 1, ..., p$. Show that $\sum_{j=1}^p c_j\delta_j$ is a Bayes estimator of $\sum_{j=1}^p c_j\vartheta_j$ under the squared error loss.

18. Prove (ii) and (iii) of Theorem 4.2.

19. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $P(X_1 = 1) = p \in (0, 1)$.
(a) Show that $\bar{X}$ is an admissible estimator of p under the loss function $(a - p)^2/[p(1 - p)]$.
(b) Show that $\bar{X}$ is an admissible estimator of p under the squared error loss.

20. Let X be a sample (of size 1) from $N(\mu, 1)$. Consider the estimation of μ under the loss function $L(\mu, a) = |\mu - a|$. Show that X is an admissible estimator.

21. In Exercise 2, consider the posterior mean to be the Bayes estimator of the corresponding parameter in each case.
(a) Show that the bias of the Bayes estimator converges to 0 if $n \to \infty$.
(b) Show that the Bayes estimator is consistent.
(c) Show that the Bayes estimator is admissible.

22. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $P(X_1 = 1) = p \in (0, 1)$.
(a) Obtain the Bayes estimator of $p(1-p)$ w.r.t. $\Pi =$ the beta distribution $B(\alpha, \beta)$ with known α and β, under the squared error loss.
(b) Compare the Bayes estimator in part (a) with the UMVUE of $p(1-p)$.
(c) Discuss the bias, consistency, and admissibility of the Bayes estimator in (a).
(d) If p has the improper prior density $\pi(p) = [p(1-p)]^{-1} I_{(0,1)}(p)$, show that the posterior p.d.f. of p given X is proper provided that $0 < \bar{X} < 1$.
(e) Under the squared error loss, find the generalized Bayes estimator of $p(1-p)$ w.r.t. the improper prior in (d).

23. Let X be an observation from the negative binomial distribution $NB(p, r)$ with a known r and an unknown $p \in (0, 1)$.
(a) Under the squared error loss, find the Bayes estimators of p and p^{-1} w.r.t. $\Pi =$ the beta distribution $B(\alpha, \beta)$ with known α and β.
(b) Show that the Bayes estimators in (a) are consistent.

24. In Example 4.7,
(a) show that the posterior distribution of ω is the gamma distribution $\Gamma(n + \alpha, (n\bar{x} + \gamma^{-1})^{-1})$;
(b) show that $\bar{X}$ is the generalized Bayes estimator of θ w.r.t. the improper prior $\frac{d\Pi}{d\omega} = \omega^{-1} I_{(0,\infty)}(\omega)$ and is a limit of Bayes estimators (as $\alpha \to 1$ and $\gamma \to \infty$).

25. Consider Example 4.8 with $\alpha = \gamma = 0$, which leads to an improper prior. Show that the posterior distribution of $\sqrt{n}(\mu - \bar{x})/\sqrt{y/(n-1)}$, given x, is the t-distribution t_{n-1}.

26. Prove Lemma 4.1.

27. Let X_1 and X_2 be independently distributed as P_{θ_1} and P_{θ_2}, respectively. Suppose that θ_1 and θ_2 are real-valued and independent according to some prior distributions Π_1 and Π_2. Let δ_j be the Bayes estimator of θ_j on the basis of X_j, $j = 1, 2$.
(a) Show that $\delta_1 - \delta_2$ is the Bayes estimator of $\theta_1 - \theta_2$ on the basis of (X_1, X_2).
(b) Show that $\delta_1 \delta_2$ is the Bayes estimator of $\theta_1 \theta_2$ on the basis of (X_1, X_2).

28. In Example 4.9, suppose that ε_{ij} has the Lebesgue p.d.f.

$$\kappa(\delta)\sigma_i^{-1} \exp\left\{-c(\delta)|x/\sigma_i|^{2/(1+\delta)}\right\},$$

where

$$c(\delta) = \left[\frac{\Gamma\left(\frac{3(1+\delta)}{2}\right)}{\Gamma\left(\frac{1+\delta}{2}\right)}\right]^{\frac{1}{1+\delta}}, \qquad \kappa(\delta) = \frac{\left[\Gamma\left(\frac{3(1+\delta)}{2}\right)\right]^{1/2}}{(1+\delta)\left[\Gamma\left(\frac{1+\delta}{2}\right)\right]^{3/2}},$$

$-1 < \delta \leq 1$ and $\sigma_i > 0$.
(a) Assume that δ is known. Let $\omega_i = c(\delta)\sigma_i^{-2/(1+\delta)}$. Under the squared error loss and the same prior in Example 4.9, show that the Bayes estimator of σ_i^2 is

$$q_i(\delta)\int\left[\gamma^{-1}\sum_{j=1}^{n_i}|x_{ij}-\beta Z_i^\tau|^{2/(1+\delta)}\right]^{1+\delta} f(\beta|x,\delta)d\beta,$$

where $q_i(\delta) = [c(\delta)]^{1+\delta}\Gamma\left(\frac{1+\delta}{2}n_i+\alpha-\delta\right)/\Gamma\left(\frac{1+\delta}{2}n_i+\alpha+1\right)$ and

$$f(\beta|x,\delta) \propto \pi(\beta)\prod_{i=1}^{k}\left[\gamma^{-1}+\sum_{j=1}^{n_i}|x_{ij}-\beta Z_i^\tau|^{2/(1+\delta)}\right]^{1+\delta}.$$

(b) Assume that δ has a prior p.d.f. $f(\delta)$ and that given δ, ω_i still has the same prior in (a). Derive a formula (similar to that in (a)) for the Bayes estimator of σ_i^2.

29. Suppose that we have observations

$$X_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, ..., k,\ j = 1, ..., m,$$

where ε_{ij}'s are i.i.d. from $N(0, \sigma_\varepsilon^2)$, μ_i's are i.i.d. from $N(\mu, \sigma_\mu^2)$, and ε_{ij}'s and μ_i's are independent. Suppose that the distribution for σ_ε^2 is the inverse gamma distribution $\Gamma^{-1}(\alpha_1, \beta_1)$ (see Exercise 1(d)); the distribution for σ_μ^2 is the inverse gamma distribution $\Gamma^{-1}(\alpha_2, \beta_2)$; the distribution for μ is $N(\mu_0, \sigma_0^2)$; and σ_ε, σ_μ, and μ are independent. Describe a Gibbs sampler and obtain explicit forms of
(a) the distribution of μ, given X_{ij}'s, μ_i's, σ_ε^2, and σ_μ^2;
(b) the distribution of μ_i, given X_{ij}'s, μ, σ_ε^2, and σ_μ^2;
(c) the distribution of σ_ε^2, given X_{ij}'s, μ_i's, μ, and σ_μ^2;
(d) the distribution of σ_μ^2, given X_{ij}'s, μ_i's, μ, and σ_ε^2.

30. Prove (4.16).

31. Consider a Lebesgue p.d.f. $p(y) \propto (2+y)^{125}(1-y)^{38}y^{34}I_{(0,1)}(y)$. Generate Markov chains of length 10,000 and compute approximations to $\int yp(y)dy$, using the Metropolis kernel with $q(y,z)$ being the p.d.f. of $N(y, r^2)$, given y, where (a) $r = 0.001$; (b) $r = 0.05$; (c) $r = 0.12$.

32. Prove Proposition 4.4 for the cases of variance and risk.

33. In the proof of Theorem 4.5, show that if L is (strictly) convex and not monotone, then $E[L(T_0(x) - a)|D = d]$ is (strictly) convex and not monotone.

34. Prove part (iii) of Theorem 4.5.

35. Under the conditions of Theorem 4.5 and the loss function $L(\mu, a) = |\mu - a|$, show that $u_*(d)$ in Theorem 4.5 is any median (Exercise 75 in §2.6) of $T_0(X)$ under the conditional distribution of X given $D = d$ when $\mu = 0$.

36. Show that if there is a location invariant estimator T_0 of μ with finite mean, then $E_0[T(X)|D = d]$ is finite a.s. $\mathcal{P}$ for any location invariant estimator.

37. Show (4.21) under the squared error loss.

38. Let $X_1, ..., X_n$ be i.i.d. with the Lebesgue p.d.f.

$$f_\theta(x) = \sqrt{\tfrac{2}{\pi}} e^{-(x-\theta)^2/2} I_{(\theta,\infty)}(x).$$

Find the MRIE of θ under the squared error loss.

39. In Example 4.12,
(a) show that $X_{(1)} - \theta \log 2/n$ is an MRIE of μ under the absolute error loss $L(\mu - a) = |\mu - a|$;
(b) show that $X_{(1)} - t$ is an MRIE under the loss function $L(\mu - a) = I_{(t,\infty)}(|\mu - a|)$.

40. In Example 4.13, show that T_* is also an MRIE of μ if the loss function is convex and even. (Hint: the distribution of $T_*(X)$ given D depends only on $X_{(n)} - X_{(1)}$ and is symmetric about 0 when $\mu = 0$.)

41. Let $X_1, ..., X_n$ be i.i.d. from the double exponential distribution $DE(\mu, 1)$ with an unknown $\mu \in \mathcal{R}$. Under the squared error loss, find the MRIE of μ. (Hint: for $x_1 < \cdots < x_n$ and $x_k < t < x_{k+1}$, $\sum_{i=1}^n |x_i - t| = \sum_{i=k+1}^n x_i - \sum_{i=1}^k x_i + (2k - n)t$.)

42. In Example 4.11, find the MRIE of μ under the loss function

$$L(\mu - a) = \begin{cases} -\alpha(\mu - a) & \mu < a \\ \beta(\mu - a) & \mu \geq a, \end{cases}$$

where α and β are positive constants. (Hint: show that if Y is a random variable with c.d.f. F, then $E[L(Y - u)]$ is minimized for any u satisfying $F(u) = \beta/(\alpha + \beta)$.)

43. Let T be a location invariant estimator of μ in a one-parameter location problem. Show that T is an MRIE under the squared error loss if and only if T is unbiased and $E[T(X)U(X)] = 0$ for any $U(X)$ satisfying $U(x_1 + c, ..., x_n + c) = U(x)$ for any c and $E[U(X)] = 0$ for any μ.

44. Assume the conditions in Theorem 4.6. Let T be a sufficient statistic for μ. Show that Pitman's estimator is a function of T.

45. Prove Proposition 4.5, Theorems 4.7 and 4.8, and Corollary 4.1.

46. Under the conditions of Theorem 4.8 and the loss function (4.24) with $p = 1$, show that $u_*(z)$ is any constant $c > 0$ satisfying
$$\int_0^c x dP_{x|z} = \int_c^\infty x dP_{x|z},$$
where $P_{x|z}$ is the conditional distribution of X given Z when $\sigma = 1$.

47. In Example 4.15, show that the MRIE is $2^{(n+1)^{-1}} X_{(n)}$ when the loss is given by (4.24) with $p = 1$.

48. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(0, \theta)$ with an unknown $\theta > 0$.
(a) Find the MRIE of θ under the loss (4.24) with $p = 2$.
(b) Find the MRIE of θ under the loss (4.24) with $p = 1$.
(c) Find the MRIE of θ^2 under the loss (4.24) with $p = 2$.

49. Let $X_1, ..., X_n$ be i.i.d. with a Lebesgue p.d.f. $(2/\sigma)[1-(x/\sigma)]I_{(0,\sigma)}(x)$, where $\sigma > 0$ is an unknown scale parameter. Find Pitman's estimator of σ^h for $n = 2, 3$, and 4.

50. Let $X_1, ..., X_n$ be i.i.d. from the Pareto distribution $Pa(\alpha, \sigma)$, where $\sigma > 0$ is an unknown parameter and $\alpha > 2$ is known. Find the MRIE of σ under the loss function (4.24) with $p = 2$.

51. Assume that the sample X has a joint Lebesgue p.d.f. given by (4.25). Show that a loss function for the estimation of μ is invariant under the location-scale transformations $g_{c,r}(X) = (rX_1 + c, ..., rX_n + c)$, $r > 0$, $c \in \mathcal{R}$, if and only if it is of the form $L\left(\frac{a-\mu}{\sigma}\right)$.

52. Prove Proposition 4.6, Theorem 4.10, and Corollary 4.2.

53. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(\mu, \sigma)$, where $\mu \in \mathcal{R}$ and $\sigma > 0$ are unknown.
(a) Find the MRIE of σ under the loss (4.24) with $p = 1$ or 2.
(b) Under the loss function $(a - \mu)^2/\sigma^2$, find the MRIE of μ.
(c) Compute the bias of the MRIE of μ in (b).

54. Suppose that X and Y are two samples with p.d.f. given by (4.30).
(a) Suppose that $\mu_x = \mu_y = 0$ and consider the estimation of $\eta = (\sigma_y/\sigma_x)^h$ with a fixed $h \neq 0$ under the loss $L(a/\eta)$. Show that the problem is invariant under the transformations $g(X, Y) = (rX, r'Y)$, $r > 0$, $r' > 0$. Generalize Proposition 4.5, Theorem 4.8, and Corollary 4.1 to the present problem.
(b) Generalize the result in (a) to the case of unknown μ_x and μ_y under the transformations in (4.31).

55. Under the conditions of part (a) of the previous exercise and the loss function $(a-\eta)^2/\eta^2$, determine the MRIE of η in the following cases:
(a) $m = n = 1$, X and Y are independent, X has the gamma distribution $\Gamma(\alpha_x, \gamma)$ with a known α_x and an unknown $\gamma = \sigma_x > 0$, and Y has the gamma distribution $\Gamma(\alpha_y, \gamma)$ with a known α_y and an unknown $\gamma = \sigma_y > 0$;
(b) X is $N_m(0, \sigma_x^2 I_m)$, Y is $N_n(0, \sigma_y^2 I_n)$, and X and Y are independent;
(c) X and Y are independent, X_i's are i.i.d. from the uniform distribution $U(0, \sigma_x)$, and Y_i's are i.i.d. from the uniform distribution $U(0, \sigma_y)$.

56. Let $X_1, ..., X_m$ and $Y_1, ..., Y_n$ be two independent samples, where X_i's are i.i.d. having the p.d.f. $\sigma_x^{-1} f\left(\frac{x-\mu_x}{\sigma_x}\right)$ with $\mu_x \in \mathcal{R}$ and $\sigma_x > 0$, and Y_i's are i.i.d. having the p.d.f. $\sigma_y^{-1} f\left(\frac{x-\mu_y}{\sigma_y}\right)$ with $\mu_y \in \mathcal{R}$ and $\sigma_y > 0$. Under the loss function $(a-\eta)^2/\eta^2$ and the transformations in (4.31), obtain the MRIE of $\eta = \sigma_y/\sigma_x$ when
(a) f is the p.d.f. of $N(0, 1)$;
(b) f is the p.d.f. of the exponential distribution $E(0, 1)$;
(c) f is the p.d.f. of the uniform distribution $U\left(-\frac{1}{2}, \frac{1}{2}\right)$;
(d) In (a)-(c), find the MRIE of $\Delta = \mu_y - \mu_x$ under the assumption that $\sigma_x = \sigma_y = \sigma$ and under the loss function $(a - \Delta)^2/\sigma^2$.

57. Consider the general linear model (3.25) under the assumption that ε_i's are i.i.d. with the p.d.f. $\sigma^{-1} f(x/\sigma)$, where f is a known Lebesgue p.d.f.
(a) Show that the family of populations is invariant under the transformations in (4.32).
(b) Show that the estimation of βl^τ with $l \in \mathcal{R}(Z)$ is invariant under the loss function $L\left(\frac{a-\beta l^\tau}{\sigma}\right)$.
(c) Show that the LSE $\hat{\beta} l^\tau$ is an invariant estimator of βl^τ, $l \in \mathcal{R}(Z)$.
(d) Prove Theorem 4.10.

58. In Example 4.18, let T be a randomized estimator of p with probability $n/(n+1)$ being $\bar{X}$ and probability $1/(n+1)$ being $\frac{1}{2}$. Show that

T has a constant risk that is smaller than the maximum risk of $\bar{X}$.

59. Let X be a single sample from the geometric distribution $G(p)$ with an unknown $p \in (0,1)$. Show that $I_{\{1\}}(X)$ is a minimax estimator of p under the loss function $(a-p)^2/[p(1-p)]$.

60. In Example 4.19, show that $\bar{X}$ is a minimax estimator of μ under the loss function $(a-\mu)^2/\sigma^2$ when $\Theta = \mathcal{R} \times (0,\infty)$.

61. Let T be a minimax (or admissible) estimator of ϑ under the squared error loss. Show that $c_1 T + c_0$ is a minimax (or admissible) estimator of $c_1\vartheta + c_0$ under the squared error loss, where c_1 and c_0 are constants.

62. Let X be a sample from P_θ with an unknown $\theta = (\theta_1, \theta_2)$, where $\theta_j \in \Theta_j$, $j = 1, 2$, and let Π_2 be a probability measure on Θ_2. Suppose that an estimator T_0 minimizes $\sup_{\theta_1 \in \Theta_1} \int R_T(\theta) d\Pi_2(\theta_2)$ over all estimators T and that $\sup_{\theta_1 \in \Theta_1} \int R_{T_0}(\theta) d\Pi_2(\theta_2) = \sup_{\theta_1 \in \Theta_1, \theta_2 \in \Theta_2} R_{T_0}(\theta)$. Show that T_0 is a minimax estimator.

63. Let $X_1, ..., X_m$ be i.i.d. from $N(\mu_x, \sigma_x^2)$ and $Y_1, ..., Y_n$ be i.i.d. from $N(\mu_y, \sigma_y^2)$. Assume that X_i's and Y_j's are independent. Consider the estimation of $\Delta = \mu_y - \mu_x$ under the squared error loss.
(a) Show that $\bar{Y} - \bar{X}$ is a minimax estimator of Δ when σ_x and σ_y are known.
(b) Show that $\bar{Y} - \bar{X}$ is a minimax estimator of Δ when $\sigma_x \in (0, c_x]$ and $\sigma_y \in (0, c_y]$, where c_x and c_y are constants.

64. Consider the general linear model (3.25) with assumption A1 and the estimation of βl^τ under the squared error loss, where $l \in \mathcal{R}(Z)$. Show that the LSE $\hat{\beta} l^\tau$ is minimax if $\sigma^2 \in (0, c]$ with a constant c.

65. Let X be a random variable having the hypergeometric distribution $HG(r, \theta, N-\theta)$ (Table 1.1, page 18) with known N and r but an unknown θ. Consider the estimation of θ/N under the squared error loss.
(a) Show that the risk function of $T(X) = \alpha X/r + \beta$ is constant, where $\alpha = \{1 + \sqrt{(N-r)/[r(N-1)]}\}^{-1}$ and $\beta = (1-\alpha)/2$.
(b) Show that T in (a) is the minimax estimator of θ/N and the Bayes estimator w.r.t. the prior

$$\Pi(\{\theta\}) = \frac{\Gamma(2c)}{[\Gamma(c)]^2} \int_0^1 \binom{N}{\theta} t^{\theta+c-1}(1-t)^{N-\theta+c-1} dt, \quad \theta = 1, 2, ...,$$

where $c = \beta/(\alpha/r - 1/N)$.

66. Let X be a single observation from $N(\mu, 1)$ and let μ have the improper prior density $\pi(\mu) = e^\mu$ w.r.t. the Lebesgue measure on $\mathcal{R}$.

Under the squared error loss, show that the generalized Bayes estimator of μ is $X+1$, which is neither minimax nor admissible.

67. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(a, \theta)$ with a known θ and an unknown $a \in \mathcal{R}$. Under the squared error loss, show that $X_{(1)} - \theta/n$ is the unique minimax estimator of a.

68. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution $U(\mu - \frac{1}{2}, \mu + \frac{1}{2})$ with an unknown $\mu \in \mathcal{R}$. Under the squared error loss, show that $(X_{(1)} + X_{(n)})/2$ is the unique minimax estimator of μ.

69. Let $X_1, ..., X_n$ be i.i.d. from the double exponential distribution $DE(\mu, 1)$ with an unknown $\mu \in \mathcal{R}$. Under the squared error loss, find a minimax estimator of μ.

70. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $P(X_1 = 1) = p \in (0, 1)$. Consider the estimation of p under the squared error loss. Using Theorem 4.14, show that $\bar{X}$ and $(\bar{X} + \gamma\lambda)/(1 + \lambda)$ with $\lambda > 0$ and $0 \le \gamma \le 1$ are admissible.

71. Let X be a single observation. Using Theorem 4.14, find values of α and β such that $\alpha X + \beta$ are admissible for estimating EX under the squared error loss, when
(a) X has the Poisson distribution $P(\theta)$ with an unknown $\theta > 0$;
(b) X has the negative binomial distribution $NB(p, r)$ with a known r and an unknown $p \in (0, 1)$.

72. Let X be a single observation having the Lebesgue p.d.f. $\frac{1}{2}c(\theta)e^{\theta x - |x|}$, $|\theta| < 1$.
(a) Show that $c(\theta) = 1 - \theta^2$.
(b) Show that if $0 \le \alpha \le \frac{1}{2}$, then $\alpha X + \beta$ is admissible for estimating $E(X)$ under the squared error loss.

73. In Example 4.23, find the UMVUE of $\theta = (p_1, ..., p_k)$ under the loss function (4.37).

74. Let X be a sample from P_θ, $\theta \in \Theta \subset \mathcal{R}^p$. Consider the estimation of θ under the loss $(\theta - a)Q(\theta - a)^\tau$, where $a \in \mathbb{A} = \Theta$ and Q is a known positive definite matrix. Show that the Bayes action is the posterior mean $E(\theta|X = x)$, assuming that all integrals involved are finite.

75. In Example 4.24, show that X is the MRIE of θ under the squared error loss, if
(a) $f(x - \theta) = \prod_{j=1}^p f_j(x_j - \theta_j)$, where each f_j is a known Lebesgue p.d.f. with mean 0;
(b) $f(x - \theta) = f(\|x - \theta\|)$ with $\int x f(\|x\|)dx = 0$.

76. Prove that X in Example 4.25 is a minimax estimator of θ under the loss function (4.37).

77. Let $X_1, ..., X_k$ be independent random variables, where X_i has the binomial distribution $Bi(p_i, n_i)$ with an unknown $p_i \in (0, 1)$ and a known n_i. For estimating $\theta = (p_1, ..., p_k)$ under the loss (4.37), find a minimax estimator of θ and determine whether it is admissible.

78. Show that the risk function in (4.42) tends to p as $\|\theta\| \to \infty$.

79. Suppose that X is $N_p(\theta, I_p)$. Consider the estimation of θ under the loss $(a-\theta)Q(a-\theta)^\tau$ with a positive definite $p \times p$ matrix Q. Show that the risk of the estimator

$$\delta^Q_{c,r} = c + \left[1 - \frac{r(p-2)}{\|Q^{-1/2}(X-c)\|^2}\right] Q^{-1}(X-c)$$

is equal to

$$\mathrm{tr}(Q) - (2r - r^2)(p-2)^2 E(\|Q^{-1/2}(X-c)\|^{-2}).$$

80. Show that under the loss (4.37), the risk of $\tilde{\delta}_{c,r}$ in (4.45) is given by (4.46).

81. Suppose that X is $N_p(\theta, V)$ with $p \geq 4$. Consider the estimation of θ under the loss function (4.37).
(a) When $V = I_p$, show that the risk of the estimator in (4.48) is

$$p - (p-3)^2 E(\|X - \bar{X} J_p\|^{-2}).$$

(b) When $V = \sigma^2 D$ with an unknown $\sigma^2 > 0$ and a known matrix D, show that the risk function of the estimator in (4.49) is smaller than that of X for any θ and σ^2.

82. Let X be a sample from a p.d.f. f_θ and $T(X)$ be a sufficient statistic for θ. Show that if an MLE exists, it is a function of T but it may not be sufficient for θ.

83. Let $\{f_\theta : \theta \in \Theta\}$ be a family of p.d.f.'s w.r.t. a σ-finite measure, where $\Theta \subset \mathcal{R}^k$; h be a Borel function from Θ onto $\Lambda \subset \mathcal{R}^p$, $1 \leq p \leq k$; and let $\tilde{\ell}(\lambda) = \sup_{\theta: h(\theta)=\lambda} \ell(\theta)$ be the induced likelihood function for the transformed parameter λ. Show that if $\hat{\theta} \in \Theta$ is an MLE of θ, then $\hat{\lambda} = h(\hat{\theta})$ is an MLE of $\lambda = h(\theta)$.

84. Let $X_1, ..., X_n$ be i.i.d. with a p.d.f. f_θ. Find an MLE of θ in each of the following cases.
(a) $f_\theta(x) = \theta^{-1} I_{\{1,...,\theta\}}(x)$, θ is an integer between 1 and θ_0.

(b) $f_\theta(x) = e^{-(x-\theta)} I_{(\theta,\infty)}(x)$, $\theta > 0$.
(c) $f_\theta(x) = \theta(1-x)^{\theta-1} I_{(0,1)}(x)$, $\theta > 1$.
(d) $f_\theta(x) = \frac{\theta}{1-\theta} x^{(2\theta-1)/(1-\theta)} I_{(0,1)}(x)$, $\theta \in (\frac{1}{2}, 1)$.
(e) $f_\theta(x) = 2^{-1} e^{-|x-\theta|}$.
(f) $f_\theta(x) = \theta x^{-2} I_{(\theta,\infty)}(x)$, $\theta > 0$.
(g) $f_\theta(x) = \theta^x (1-\theta)^{1-x} I_{\{0,1\}}(x)$, $\theta \in [\frac{1}{2}, \frac{3}{4}]$.
(h) $f_\theta(x)$ is the p.d.f. of $N(\theta, \theta^2)$, $\theta \in \mathcal{R}$.
(i) $f_\theta(x)$ is the p.d.f. of the exponential distribution $E(\mu, \sigma)$, $\theta = (\mu, \sigma) \in \mathcal{R} \times (0, \infty)$.
(j) $f_\theta(x)$ is the p.d.f. of the log-normal distribution $LN(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2) \in \mathcal{R} \times (0, \infty)$.
(k) $f_\theta(x) = I_{(0,1)}(x)$ if $\theta = 0$ and $f_\theta(x) = (2\sqrt{x})^{-1} I_{(0,1)}(x)$ if $\theta = 1$.
(l) $f_\theta(x) = \beta^{-\alpha} \alpha x^{\alpha-1} I_{(0,\beta)}(x)$, $\alpha > 0$, $\beta > 0$.
(m) $f_\theta(x) = \binom{\theta}{x} p^x (1-p)^{\theta-x} I_{\{0,1,...,\theta\}}(x)$, $\theta = 1, 2, ...$, where $p \in (0, 1)$ is known.

85. Suppose that n observations are taken from $N(\mu, 1)$ with an unknown μ. Instead of recording all the observations, one records only whether the observation is less than 0. Find an MLE of μ.

86. Find MLE's of θ and $e^{-t/\theta}$ in Example 4.7.

87. Let $(Y_1, Z_1), ..., (Y_n, Z_n)$ be i.i.d. with the Lebesgue p.d.f.

$$\lambda^{-1} \mu^{-1} e^{-y/\lambda} e^{-z/\mu} I_{(0,\infty)}(y) I_{(0,\infty)}(z),$$

where $\lambda > 0$ and $\mu > 0$.
(a) Find the MLE of (λ, μ).
(b) Suppose that we only observe $X_i = \min(Y_i, Z_i)$ and $\Delta_i = 1$ if $X_i = Y_i$ and $\Delta_i = 0$ if $X_i = Z_i$. Find the MLE of (λ, μ).

88. In Example 4.33, show that the likelihood equation has a unique solution that is the MLE of $\theta = (\alpha, \gamma)$. Obtain iteration equation (4.53) for this example. Discuss how to apply the Fisher-scoring method in this example.

89. Let $X_1, ..., X_n$ be i.i.d. from the discrete p.d.f.

$$f_\theta(x) = [x!(1 - e^{-\theta})]^{-1} \theta^x e^{-\theta} I_{\{1,2,...\}}(x),$$

where $\theta > 0$. Show that the likelihood equation has a unique root when $\bar{x} > 1$. Discuss whether this root is an MLE of θ.

90. Let $X_1, ..., X_n$ be i.i.d. from the logistic distribution $LG(\mu, \sigma)$ (Table 1.2, page 20).
(a) Show how to find an MLE of μ when $\mu \in \mathcal{R}$ and σ is known.
(b) Show how to find an MLE of σ when $\sigma > 0$ and μ is known.

91. Let $(X_1, Y_1), ..., (X_n, Y_n)$ be i.i.d. from a two-dimensional normal distribution with $E(X_1) = E(Y_1) = 0$, $\mathrm{Var}(X_1) = \mathrm{Var}(Y_1) = 1$, and an unknown correlation coefficient $\rho \in (-1, 1)$. Show that the likelihood equation is a cubic in ρ and the probability that it has a unique root tends to 1 as $n \to \infty$.

92. Let $X_1, ..., X_n$ be i.i.d. from the Weibull distribution $W(\alpha, \theta)$ (Table 1.2, page 20) with unknown $\alpha > 0$ and $\theta > 0$. Show that the likelihood equation is equivalent to $h(\alpha) = n^{-1} \sum_{i=1}^n \log x_i$ and $\theta = n^{-1} \sum_{i=1}^n x_i^\alpha$, where $h(\alpha) = (\sum_{i=1}^n x_i^\alpha)^{-1} \sum_{i=1}^n x_i^\alpha \log x_i - \alpha^{-1}$, and that the likelihood equation has a unique solution.

93. Consider the random effects model in Example 3.17. Assume that $\mu = 0$ and $n_i = n_0$ for all i. Provide a condition on X_{ij}'s under which a unique MLE of (σ_a^2, σ^2) exists and find this MLE.

94. Let $X_1, ..., X_n$ be i.i.d. with the p.d.f. $\theta f(\theta x)$, where f is a Lebesgue p.d.f. on $(0, \infty)$ or symmetric about 0, and $\theta > 0$ is an unknown parameter. Show that the likelihood equation has a unique root if $x f'(x)/f(x)$ is strictly decreasing for $x > 0$. Verify that this condition is satisfied if f is the p.d.f. of the Cauchy distribution $C(0, 1)$.

95. Consider the location family in §4.2.1 and the scale family in §4.2.2. In each case,
(a) show that an MLE of the parameter, if it exists, is invariant;
(b) show that an RLE (root of the likelihood equation), if it exists, is invariant.

96. Let X be a sample from P_θ, $\theta \in \mathcal{R}$. Suppose that P_θ's have p.d.f.'s f_θ w.r.t. a common σ-finite measure and that $\{x : f_\theta(x) > 0\}$ does not depend on θ. Assume further that an estimator $\hat{\theta}$ of θ attains the Cramér-Rao lower bound and that the conditions in Theorem 3.3 hold for $\hat{\theta}$. Show that $\hat{\theta}$ is a unique MLE of θ.

97. Let X_{ij}, $j = 1, ..., r > 1$, $i = 1, ..., n$, be independently distributed as $N(\mu_i, \sigma^2)$. Find the MLE of $(\mu_1, ..., \mu_n, \sigma^2)$. Show that the MLE of σ^2 is not a consistent estimator (as $n \to \infty$).

98. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution $U(0, \theta)$, where $\theta > 0$ is unknown. Let $\hat{\theta}$ be the MLE of θ and T be the UMVUE.
(a) Show that $n\,\mathrm{mse}_T(\theta) \to \theta^2$ and $n\,\mathrm{mse}_{\hat{\theta}}(\theta) \to 2\theta^2$; hence, the MLE is inadmissible when n is large enough.
(b) Let $Z_{a,\theta}$ be a random variable having the exponential distribution $E(a, \theta)$. Prove $n^2(\theta - \hat{\theta}) \to_d Z_{0,\theta}$ and $n^2(\theta - T) \to_d Z_{-\theta,\theta}$. Obtain the asymptotic relative efficiency of $\hat{\theta}$ w.r.t. T.

99. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(a, \theta)$ with unknown a and θ. Obtain the asymptotic relative efficiency of the MLE of a (or θ) w.r.t. the UMVUE of a (or θ).

100. Let $X_1, ..., X_n$ be i.i.d. from the Pareto distribution $Pa(a, \theta)$ with unknown a and θ.
(a) Find the MLE of (a, θ).
(b) Find the asymptotic relative efficiency of the MLE of a w.r.t. the UMVUE of a.

101. In Example 4.36, obtain the MLE of β under the canonical link and assumption (4.58) but $t_i \not\equiv 1$.

102. Consider the GLM in Example 4.35 with $\phi_i \equiv 1$ and the canonical link. Assume that $\sum_{i=1}^n Z_i^\tau Z_i$ is positive definite for $n \geq n_0$. Show that the likelihood equation has at most one solution when $n \geq n_0$ and a solution exists with probability tending to 1.

103. Consider the linear model (3.25) with $\varepsilon = N_n(0, V)$, where V is an unknown positive definite matrix. Show that the LSE $\hat{\beta}$ defined by (3.29) is an MQLE and that $\hat{\beta}$ is an MLE if and only if one of (a)-(e) in Theorem 3.10 holds.

104. Let X_j be a random variable having the binomial distribution $Bi(p_j, n_j)$ with a known n_j and an unknown $p_j \in (0, 1)$, $j = 1, 2$. Assume that X_j's are independent. Obtain a conditional likelihood function of the odds ratio $\theta = \frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2}$, given $X_1 + X_2$.

105. Let X_1 and X_2 be independent from Poisson distributions $P(\mu_1)$ and $P(\mu_2)$, respectively. Suppose that we are interested in $\theta_1 = \mu_1/\mu_2$. Derive a conditional likelihood function of θ_1, using (a) $\theta_2 = \mu_1$; (b) $\theta_2 = \mu_1 + \mu_2$; and (c) $\theta_2 = \mu_1\mu_2$.

106. Assume model (4.66) with $p = 2$ and normally distributed i.i.d. ε_t's. Obtained the conditional likelihood given $(X_1, X_2) = (x_1, x_2)$.

107. Prove the claim in Example 4.38.

108. Prove (4.70). (Hint: Show, using the argument in proving (4.77), that $n^{-1}|\frac{\partial^2}{\partial\theta^2} \log \ell(\xi_n) - \frac{\partial^2}{\partial\theta^2} \log \ell(\theta)| = o_p(1)$ for any random variable $\xi_n \in (\theta, \theta_n)$.)

109. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, 1)$ truncated at two known points $\alpha < \beta$, i.e., the Lebesgue p.d.f. of X_i is

$$\{\sqrt{2\pi}[\Phi(\beta - \mu) - \Phi(\alpha - \mu)]\}^{-1} e^{-(x-\mu)^2/2} I_{(\alpha,\beta)}(x).$$

(a) Using Theorem 4.17, show that $\bar{X}$ is asymptotically efficient for estimating $\theta = E(X_1)$.
(b) Show directly (without using Theorem 4.17) that the likelihood equation for θ has a solution when $\alpha < \bar{x} < \beta$.

110. In Exercise 84, check whether the regularity conditions of Theorem 4.16 are satisfied for cases (b), (c), (d), (e), (g), (h), and (j). Obtain asymptotic distributions of RLE's for cases in which Theorem 4.17 can be applied.

111. In Example 4.30, show that the MLE (or RLE) of θ is asymptotically efficient by (a) applying Theorem 4.17; and (b) directly deriving the asymptotic distribution of the MLE.

112. In Example 4.23, show that there is a unique asymptotically efficient RLE of $\theta = (p_1, ..., p_k)$. Discuss whether this RLE is the MLE.

113. Let $X_1, ..., X_n$ be i.i.d. with $P(X_1 = 0) = 6\theta^2 - 4\theta + 1$, $P(X_1 = 1) = \theta - 2\theta^2$, and $P(X_1 = 2) = 3\theta - 4\theta^2$, where $\theta \in (0, \frac{1}{2})$ is unknown. Apply Theorem 4.17 to obtain the asymptotic distribution of an RLE of θ.

114. In Exercise 91, show that the RLE of ρ is asymptotically distributed as $N\big(\rho, (1-\rho^2)^2/[n(1+\rho^2)]\big)$.

115. In Exercise 94, obtain the asymptotic distribution of the RLE of θ when f is the p.d.f. of the Cauchy distribution $C(0, 1)$.

116. Let $X_1, ..., X_n$ be i.i.d. from the logistic distribution $LG(\mu, \sigma)$ with unknown $\mu \in \mathcal{R}$ and $\sigma > 0$. Obtain the asymptotic distribution of the RLE of (μ, σ).

117. In Exercise 92, show that the conditions of Theorem 4.16 are satisfied.

118. Assume the conditions in Theorem 4.16. Suppose that $\theta = (\theta_1, ..., \theta_k)$ and there is a positive integer $p < k$ such that $\partial \log \ell(\theta)/\partial\theta_i$ and $\partial \log \ell(\theta)/\partial\theta_j$ are uncorrelated whenever $i \le p < j$. Show that the asymptotic distribution of the RLE of $(\theta_1, ..., \theta_p)$ is unaffected by whether $\theta_{p+1}, ..., \theta_k$ are known.

119. Let $X_1, ..., X_n$ be i.i.d. random p-vectors from $N_p(\mu, \Sigma)$ with unknown μ and Σ. Find the MLE's of μ and Σ and derive their asymptotic distributions.

120. Let $X_1, ..., X_n$ be i.i.d. bivariate normal random vectors with mean 0 and an unknown covariance matrix whose diagonal elements are σ_1^2 and σ_2^2 and off-diagonal element is $\sigma_1\sigma_2\rho$. Let $\theta = (\sigma_1^2, \sigma_2^2, \rho)$.

Obtain $I_n(\theta)$ and $[I_n(\theta)]^{-1}$ and derive the asymptotic distribution of the MLE of θ.

121. Let $X_1, ..., X_n$ be i.i.d. each with probability p as $N(\mu, \sigma^2)$ and probability $1-p$ as $N(\eta, \tau^2)$, where $\theta = (\mu, \eta, \sigma^2, \tau^2, p)$ is unknown.
(a) Show that the conditions in Theorem 4.16 are satisfied.
(b) Show that the likelihood function is unbounded and therefore, an MLE does not exist.

122. Let $X_1, ..., X_n$ and $Y_1, ..., Y_n$ be independently distributed as $N(\mu, \sigma^2)$ and $N(\mu, \tau^2)$, respectively, with unknown $\theta = (\mu, \sigma^2, \tau^2)$. Find the MLE of θ and show that it is asymptotically efficient.

123. In Exercise 93, find the asymptotic distribution of the MLE of(σ_a^2, σ^2).

124. Under the conditions in Theorem 4.18, prove (4.85) and (4.86).

125. Assume linear model (3.25) with $\varepsilon = N_n(0, \sigma^2 I_n)$ and a full rank Z. Apply Theorem 4.18 to show that the LSE $\hat{\beta}$ is asymptotically efficient. Compare this result with that in Theorem 3.12.

126. Apply Theorem 4.18 to obtain the asymptotic distribution of the RLE of β in (a) Example 4.35; and (b) Example 4.37.

127. Let $X_1, ..., X_n$ be i.i.d. from the logistic distribution $LG(\mu, \sigma)$, $\mu \in \mathcal{R}$, $\sigma > 0$. Using Newton-Raphson and Fisher-scoring methods, find
(a) one-step MLE's of μ when σ is known;
(b) one-step MLE's of σ when μ is known;
(c) one-step MLE's of (μ, σ).
(d) Show how to obtain $\sqrt{n}$-consistent initial estimators in (a)-(c).

128. Under the GLM (4.55)-(4.58),
(a) show how to obtain a one-step MLE of β, if an initial estimator $\hat{\beta}_n^{(0)}$ is available;
(b) show that under the conditions in Theorem 4.18, the one-step MLE satisfies (4.81) if $\|(\hat{\beta}_n^{(0)} - \beta)[I_n(\beta)]^{1/2}\| = O_p(1)$.

129. In Example 4.40, show that the conditions in Theorem 4.20 concerning the likelihood function are satisfied.

130. Let $X_1, ..., X_n$ be i.i.d. from the logistic distribution $LG(\mu, \sigma)$ with unknown $\mu \in \mathcal{R}$ and $\sigma > 0$. Show that the conditions in Theorem 4.20 concerning the likelihood function are satisfied.

Chapter 5

Estimation in Nonparametric Models

Estimation methods studied in this chapter are useful for nonparametric models as well as for parametric models in which the parametric model assumptions might be violated (so that robust estimators are required) or the number of unknown parameters is exceptionally large. Some such methods have been introduced in Chapter 3; for example, the methods that produce UMVUE's in nonparametric models, the U- and V-statistics, the LSE's and BLUE's, the Horvitz-Thompson estimators, and the sample (central) moments.

The theoretical justification for estimators in nonparametric models, however, relies more on asymptotics than that in parametric models. This means that applications of nonparametric methods usually require large sample sizes. Also, estimators derived using parametric methods are asymptotically more efficient than those based on nonparametric methods when the parametric models are correct. Thus, to choose between a parametric method and a nonparametric method, we need to balance the advantage of requiring weaker model assumptions (robustness) against the drawback of losing efficiency which results in requiring a larger sample size.

It is assumed in this chapter that a sample $X = (X_1, ..., X_n)$ is from a population in a nonparametric family, where X_i's are random vectors.

5.1 Distribution Estimators

In many applications the c.d.f.'s of X_i's are determined by a single c.d.f. F on $\mathcal{R}^d$; for example, X_i's are i.i.d. random d-vectors. In this section we

consider the estimation of F or $F(t)$ for several t's, under a nonparametric model in which very little is assumed about F.

5.1.1 Empirical c.d.f.'s in i.i.d. cases

For i.i.d. random variables $X_1, ..., X_n$, the empirical c.d.f. F_n is defined in (2.31). The definition of the empirical c.d.f. in the case of $X_i \in \mathcal{R}^d$ is analogous. For $t \in \mathcal{R}^d$, let $A(t)$ be the set of all $s \in \mathcal{R}^d$ such that all components of $t - s$ are nonnegative. Then the empirical c.d.f. based on X is defined by

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{A(t)}(X_i), \qquad t \in \mathcal{R}^d. \tag{5.1}$$

Similar to the case of $d = 1$ (Example 2.26), $F_n(t)$ as an estimator of $F(t)$ has the following properties. For any $t \in \mathcal{R}^d$, $nF_n(t)$ has the binomial distribution $Bi(F(t), n)$; $F_n(t)$ is unbiased with variance $F(t)[1 - F(t)]/n$; $F_n(t)$ is the UMVUE under some nonparametric models; and $F_n(t)$ is $\sqrt{n}$-consistent for $F(t)$. For any m fixed distinct points $t_1, ..., t_m$ in $\mathcal{R}^d$, it follows from the multivariate CLT (Corollary 1.2) and (5.1) that as $n \to \infty$,

$$\sqrt{n}\big[\big(F_n(t_1), ..., F_n(t_m)\big) - \big(F(t_1), ..., F(t_m)\big)\big] \to_d N_m(0, \Sigma), \tag{5.2}$$

where Σ is the $m \times m$ matrix whose (i, j)th element is

$$P\big(X_1 \in A(t_i) \cap A(t_j)\big) - F(t_i)F(t_j).$$

Note that these results hold without *any* assumption on F.

Considered as a function of t, F_n is a random element taking values in $\mathcal{F}$, the collection of all c.d.f.'s on $\mathcal{R}^d$. As $n \to \infty$, $\sqrt{n}(F_n - F)$ converges in some sense to a random element defined on some probability space. A detailed discussion of such a result is beyond our scope, and can be found, for example, in Shorack and Wellner (1986). To discuss some global properties of F_n as an estimator of $F \in \mathcal{F}$, we need to define a closeness measure between the elements (c.d.f.'s) in $\mathcal{F}$.

Definition 5.1. Let $\mathcal{F}_0$ be a subset of $\mathcal{F}$. A function ϱ from $\mathcal{F}_0 \times \mathcal{F}_0$ to $[0, \infty)$ is called a *distance* or *metric* on $\mathcal{F}_0$ if and only if for any G_j in $\mathcal{F}_0$,
(a) $\varrho(G_1, G_2) = 0$ if and only if $G_1 = G_2$;
(b) $\varrho(G_1, G_2) = \varrho(G_2, G_1)$;
(c) $\varrho(G_1, G_2) \le \varrho(G_1, G_3) + \varrho(G_3, G_2)$. ∎

The most commonly used distance is the sup-norm distance ϱ_∞ defined on $\mathcal{F}$:

$$\varrho_\infty(G_1, G_2) = \sup_{t \in \mathcal{R}^d} |G_1(t) - G_2(t)|, \qquad G_j \in \mathcal{F}. \tag{5.3}$$

The following result concerning the sup-norm distance between F_n and F is due to Dvoretzky, Kiefer, and Wolfowitz (1956).

Lemma 5.1. (DKW's inequality). Let F_n be the empirical c.d.f. based on i.i.d. $X_1, ..., X_n$ from a c.d.f. F on $\mathcal{R}^d$.
(i) When $d = 1$, there exists a positive constant C (not depending on F) such that

$$P\big(\varrho_\infty(F_n, F) > z\big) \le Ce^{-2nz^2}, \quad z > 0,\ n = 1, 2, ...$$

(ii) When $d \ge 2$, for any $\epsilon > 0$, there exists a positive constant $C_{\epsilon,d}$ (not depending on F) such that

$$P\big(\varrho_\infty(F_n, F) > z\big) \le C_{\epsilon,d}e^{-(2-\epsilon)nz^2}, \quad z > 0,\ n = 1, 2, ... \quad \blacksquare$$

The proof of this lemma is omitted. The following results useful in statistics are direct consequences of Lemma 5.1.

Theorem 5.1. Let F_n be the empirical c.d.f. based on i.i.d. $X_1, ..., X_n$ from a c.d.f. F on $\mathcal{R}^d$. Then
(i) $\varrho_\infty(F_n, F) \to_{a.s.} 0$ as $n \to \infty$;
(ii) $E[\sqrt{n}\varrho_\infty(F_n, F)]^s = O(1)$ for any $s > 0$.
Proof. (i) From DKW's inequality,

$$\sum_{n=1}^{\infty} P\big(\varrho_\infty(F_n, F) > z\big) < \infty.$$

Hence, the result follows from Theorem 1.8(v).
(ii) Using DKW's inequality with $z = y^{1/s}/\sqrt{n}$ and the result in Exercise 45 of §1.6, we obtain that

$$\begin{aligned} E[\sqrt{n}\varrho_\infty(F_n, F)]^s &= \int_0^\infty P\big(\sqrt{n}\varrho_\infty(F_n, F) > y^{1/s}\big)dy \\ &\le C_{\epsilon,d}\int_0^\infty e^{-(2-\epsilon)y^{2/s}}dy \\ &= O(1) \end{aligned}$$

as long as $2 - \epsilon > 0$. $\blacksquare$

Theorem 5.1(i) means that $F_n(t) \to_{a.s.} F(t)$ uniformly in $t \in \mathcal{R}^d$, a result stronger than the strong consistency of $F_n(t)$ for every t. Theorem 5.1(ii) implies that $\sqrt{n}\varrho_\infty(F_n, F) = O_p(1)$, a result stronger than the $\sqrt{n}$-consistency of $F_n(t)$. Again, these results hold without any condition on the unknown F.

Let $p \geq 1$ and $\mathcal{F}_p = \{G \in \mathcal{F} : \int \|t\|^p dG < \infty\}$, which is the subset of c.d.f.'s in $\mathcal{F}$ having finite pth moments. Mallows' distance between G_1 and G_2 in $\mathcal{F}_p$ is defined to be

$$\varrho_{M_p}(G_1, G_2) = \inf(E\|Y_1 - Y_2\|^p)^{1/p}, \tag{5.4}$$

where the infimum is taken over all pairs of Y_1 and Y_2 having c.d.f.'s G_1 and G_2, respectively. Let $\{G_j : j = 0, 1, 2, ...\} \subset \mathcal{F}_p$. Then $\varrho_{M_p}(G_j, G_0) \to 0$ as $j \to \infty$ if and only if $\int \|t\|^p dG_j \to \int \|t\|^p dG_0$ and $G_j(t) \to G_0(t)$ for every $t \in \mathcal{R}^d$ at which G_0 is continuous. It follows from Theorem 5.1 and the SLLN (Theorem 1.13) that $\varrho_{M_p}(F_n, F) \to_{a.s.} 0$ if $F \in \mathcal{F}_p$.

When $d = 1$, another useful distance for measuring the closeness between F_n and F is the L_p distance ($p \geq 1$):

$$\varrho_{L_p}(G_1, G_2) = \left[\int |G_1(t) - G_2(t)|^p dt\right]^{1/p}, \qquad G_j \in \mathcal{F}_1. \tag{5.5}$$

A result similar to Theorem 5.1 is given as follows.

Theorem 5.2. Let F_n be the empirical c.d.f. based on i.i.d. random variables $X_1, ..., X_n$ from a c.d.f. $F \in \mathcal{F}_1$. Then
(i) $\varrho_{L_p}(F_n, F) \to_{a.s.} 0$;
(ii) $E[\sqrt{n}\varrho_{L_p}(F_n, F)] = O(1)$ if $1 \leq p < 2$ and $\int \{F(t)[1 - F(t)]\}^{p/2} dt < \infty$, or $p \geq 2$.
Proof. (i) Since $[\varrho_{L_p}(F_n, F)]^p \leq [\varrho_\infty(F_n, F)]^{p-1}[\varrho_{L_1}(F_n, F)]$ and, by Theorem 5.1, $\varrho_\infty(F_n, F) \to_{a.s.} 0$, it suffices to show the result for $p = 1$. Let $Y_i = \int_{-\infty}^0 [I_{(-\infty,t]}(X_i) - F(t)]dt$. Then $Y_1, ..., Y_n$ are i.i.d. and

$$E|Y_i| \leq \int E|I_{(-\infty,t]}(X_i) - F(t)|dt = 2\int F(t)[1 - F(t)]dt,$$

which is finite under the condition that $F \in \mathcal{F}_1$. By the SLLN,

$$\int_{-\infty}^0 [F_n(t) - F(t)]dt = \frac{1}{n}\sum_{i=1}^n Y_i \to_{a.s.} E(Y_1) = 0. \tag{5.6}$$

Since $[F_n(t) - F(t)]_- \leq F(t)$ and $\int_{-\infty}^0 F(t)dt < \infty$ (Exercise 45 in §1.6), it follows from Theorem 5.1 and the dominated convergence theorem that

$$\int_{-\infty}^0 [F_n(t) - F(t)]_- dt \to_{a.s.} 0,$$

which with (5.6) implies

$$\int_{-\infty}^0 |F_n(t) - F(t)|dt \to_{a.s.} 0. \tag{5.7}$$

The result follows since we can similarly show that (5.7) holds with $\int_{-\infty}^{0}$ replaced by $\int_0^\infty$.
(ii) When $1 \le p < 2$, the result follows from

$$\begin{aligned} E[\varrho_{L_p}(F_n, F)] &\le \left\{ \int E|F_n(t) - F(t)|^p dt \right\}^{1/p} \\ &\le \left\{ \int [E|F_n(t) - F(t)|^2]^{p/2} dt \right\}^{1/p} \\ &= n^{-1/2} \left\{ \int \{F(t)[1 - F(t)]\}^{p/2} dt \right\}^{1/p} \\ &= O(n^{-1/2}), \end{aligned}$$

where the two inequalities follow from Jensen's inequality. When $p \ge 2$,

$$\begin{aligned} E[\varrho_{L_p}(F_n, F)] &\le E\left\{ [\varrho_\infty(F_n, F)]^{1-2/p} [\varrho_{L_2}(F_n, F)]^{2/p} \right\} \\ &\le \left\{ E[\varrho_\infty(F_n, F)]^{(1-2/p)q} \right\}^{1/q} \left\{ E[\varrho_{L_2}(F_n, F)]^2 \right\}^{1/p} \\ &= \left\{ O(n^{-(1-2/p)q/2}) \right\}^{1/q} \left\{ E \int |F_n(t) - F(t)|^2 dt \right\}^{1/p} \\ &= O(n^{-(1-2/p)/2}) \left\{ \frac{1}{n} \int F(t)[1 - F(t)] dt \right\}^{1/p} \\ &= O(n^{-1/2}), \end{aligned}$$

where $\frac{1}{q} + \frac{1}{p} = 1$; the second inequality follows from Hölder's inequality (see, e.g., Serfling (1980, p. 352)); and the first equality follows from Theorem 5.1(ii). ∎

5.1.2 Empirical likelihoods

In §4.4 and §4.5, we have shown that the method of using likelihoods provides some asymptotically efficient estimators. We now introduce some likelihoods in nonparametric models. This not only provides another justification for the use of the empirical c.d.f. in (5.1), but also leads to a useful method of deriving estimators in various (possibly non-i.i.d.) cases, some of which are discussed later in this chapter.

Let $X_1, ..., X_n$ be i.i.d. with $F \in \mathcal{F}$ and P_G be the probability measure corresponding to $G \in \mathcal{F}$. Given $X_1 = x_1, ..., X_n = x_n$, the *nonparametric likelihood* function is defined to be the following functional from $\mathcal{F}$ to $[0, \infty)$:

$$\ell(G) = \prod_{i=1}^{n} P_G(\{x_i\}), \qquad G \in \mathcal{F}. \tag{5.8}$$

Apparently, $\ell(G) = 0$ if $P_G(\{x_i\}) = 0$ for at least one i. The following result, due to Kiefer and Wolfowitz (1956), shows that the empirical c.d.f. F_n is a nonparametric maximum likelihood estimator of F.

Theorem 5.3. Let $X_1, ..., X_n$ be i.i.d. with $F \in \mathcal{F}$ and $\ell(G)$ be defined by (5.8). Then F_n maximizes $\ell(G)$ over $G \in \mathcal{F}$.
Proof. We only need to consider $G \in \mathcal{F}$ such that $\ell(G) > 0$. Let $c \in (0, 1]$ and $\mathcal{F}(c)$ be the subset of $\mathcal{F}$ containing G's satisfying $p_i = P_G(\{x_i\}) > 0$, $i = 1, ..., n$, and $\sum_{i=1}^n p_i = c$. We now apply the Lagrange multiplier method to solve the problem of maximizing $\ell(G)$ over $G \in \mathcal{F}(c)$. Define

$$H(p_1, ..., p_n, \lambda) = \prod_{i=1}^n p_i + \lambda \left(\sum_{i=1}^n p_i - c \right),$$

where λ is the Lagrange multiplier. Set

$$\frac{\partial H}{\partial \lambda} = \sum_{i=1}^n p_i - c = 0, \qquad \frac{\partial H}{\partial p_j} = p_j^{-1} \prod_{i=1}^n p_i + \lambda = 0, \qquad j = 1, ..., n.$$

The solution is $p_i = c/n$, $i = 1, ..., n$, $\lambda = -(c/n)^{n-1}$. It can be shown (exercise) that this solution is a maximum of $H(p_1, ..., p_n, \lambda)$ over $p_i > 0$, $i = 1, ..., n$, $\sum_{i=1}^n p_i = c$. This shows that

$$\max_{G \in \mathcal{F}(c)} \ell(G) = (c/n)^n,$$

which is maximized at $c = 1$ for any fixed n. The result follows from $P_{F_n}(\{x_i\}) = n^{-1}$ for given $X_i = x_i$, $i = 1, ..., n$. ∎

From the proof of Theorem 5.3, F_n maximizes the likelihood $\ell(G)$ in (5.8) over $p_i > 0$, $i = 1, ..., n$, and $\sum_{i=1}^n p_i = 1$, where $p_i = P_G(\{x_i\})$. This method of deriving an estimator of F can be extended to various situations with some modifications of (5.8) and/or constraints on p_i's. Modifications of the likelihood in (5.8) are called *empirical likelihoods* (Owen, 1988, 1990; Qin and Lawless, 1994). An estimator obtained by maximizing an empirical likelihood is then called a *maximum empirical likelihood estimator* (MELE). We now discuss several applications of the method of empirical likelihoods.

Consider first the estimation of F with auxiliary information about F (and i.i.d. $X_1, ..., X_n$). For instance, suppose that there is a known function u from $\mathcal{R}^d$ to $\mathcal{R}^s$ such that

$$\int u(x) dF = 0 \tag{5.9}$$

(e.g., some components of the mean of F are 0). It is then reasonable to expect that any estimate $\hat{F}$ of F has property (5.9), i.e., $\int u(x) d\hat{F} = 0$,

which is not true for the empirical c.d.f. F_n in (5.1), since

$$\int u(x)dF_n = \frac{1}{n}\sum_{i=1}^n u(X_i) \neq 0$$

even if $E[u(X_1)] = 0$. Using the method of empirical likelihood, a natural solution is to put another constraint in the process of maximizing the likelihood. That is, we maximize $\ell(G)$ in (5.8) subject to

$$p_i > 0, \quad i = 1, ..., n, \quad \sum_{i=1}^n p_i = 1, \quad \text{and} \quad \sum_{i=1}^n p_i u(x_i) = 0, \tag{5.10}$$

where $p_i = P_G(\{x_i\})$. Using the Lagrange multiplier method and a similar argument to the proof of Theorem 5.3, it can be shown (exercise) that an MELE of F is

$$\hat{F}(t) = \sum_{i=1}^n \hat{p}_i I_{A(t)}(X_i), \tag{5.11}$$

where

$$\hat{p}_i = \{n[1 + u(X_i)\lambda_n^\tau]\}^{-1}, \qquad i = 1, ..., n, \tag{5.12}$$

and $\lambda_n \in \mathcal{R}^s$ is the Lagrange multiplier satisfying

$$\sum_{i=1}^n \hat{p}_i u(X_i) = \sum_{i=1}^n \frac{u(X_i)}{n[1 + u(X_i)\lambda_n^\tau]} = 0. \tag{5.13}$$

Note that $\hat{F}$ reduces to F_n if $u \equiv 0$.

To see that (5.13) has a solution asymptotically, note that

$$\frac{\partial}{\partial\lambda}\left[\frac{1}{n}\sum_{i=1}^n \log\big(1 + u(X_i)\lambda^\tau\big)\right] = \frac{1}{n}\sum_{i=1}^n \frac{u(X_i)}{1 + u(X_i)\lambda^\tau}$$

and

$$\frac{\partial^2}{\partial\lambda\partial\lambda^\tau}\left[\frac{1}{n}\sum_{i=1}^n \log\big(1 + u(X_i)\lambda^\tau\big)\right] = -\frac{1}{n}\sum_{i=1}^n \frac{[u(X_i)]^\tau u(X_i)}{[1 + u(X_i)\lambda^\tau]^2},$$

which is negative definite if $\mathrm{Var}(u(X_1))$ is positive definite. Also,

$$E\left\{\frac{\partial}{\partial\lambda}\left[\frac{1}{n}\sum_{i=1}^n \log\big(1 + u(X_i)\lambda^\tau\big)\right]\bigg|_{\lambda=0}\right\} = E[u(X_1)] = 0.$$

Hence, using the same argument as in the proof of Theorem 4.18, we can show that there exists a unique sequence $\{\lambda_n(X)\}$ such that as $n \to \infty$,

$$P\left(\sum_{i=1}^n \frac{u(X_i)}{n[1 + u(X_i)\lambda_n^\tau]} = 0\right) \to 1 \qquad \text{and} \qquad \lambda_n \to_p 0. \tag{5.14}$$

Theorem 5.4. Let $X_1, ..., X_n$ be i.i.d. with $F \in \mathcal{F}$, u be a function on $\mathcal{R}^d$ satisfying (5.9), and $\hat{F}$ be given by (5.11)-(5.13). Suppose that $U = \mathrm{Var}(u(X_1))$ is positive definite. Then, for any m fixed distinct $t_1, ..., t_m$ in $\mathcal{R}^d$,

$$\sqrt{n}[\big(\hat{F}(t_1), ..., \hat{F}(t_m)\big) - \big(F(t_1), ..., F(t_m)\big)] \to_d N_m(0, \Sigma_u), \tag{5.15}$$

where

$$\Sigma_u = \Sigma - W^\tau U^{-1} W,$$

$W = \big([W(t_1)]^\tau, ..., [W(t_m)]^\tau\big)$, $W(t_j) = E[u(X_1) I_{A(t_j)}(X_1)]$ ($A(t)$ is given in (5.1)), and Σ is given in (5.2).

Proof. We prove the case of $m = 1$. The case of $m \geq 2$ is left as an exercise. Let $\bar{u} = n^{-1} \sum_{i=1}^n u(X_i)$. It follows from (5.13), (5.14), and Taylor's expansion that

$$\bar{u}^\tau = \frac{1}{n} \sum_{i=1}^n [u(X_i)]^\tau u(X_i) \lambda_n^\tau + o_p(\|\lambda_n\|).$$

By the SLLN and CLT,

$$U^{-1} \bar{u}^\tau = \lambda_n^\tau + o_p(n^{-1/2}).$$

Using Taylor's expansion and the SLLN again, we have

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n I_{A(t)}(X_i)(\hat{p}_i - 1) &= \frac{1}{n} \sum_{i=1}^n I_{A(t)}(X_i) \left[\frac{1}{1 + u(X_i)\lambda_n^\tau} - 1 \right] \\
&= -\frac{1}{n} \sum_{i=1}^n I_{A(t)}(X_i) u(X_i) \lambda_n^\tau + o_p(n^{-1/2}) \\
&= -W(t)\lambda_n^\tau + o_p(n^{-1/2}) \\
&= -W(t) U^{-1} \bar{u}^\tau + o_p(n^{-1/2}).
\end{aligned}$$

Thus,

$$\begin{aligned}
\hat{F}(t) - F(t) &= F_n(t) - F(t) + \frac{1}{n} \sum_{i=1}^n I_{A(t)}(X_i)(\hat{p}_i - 1) \\
&= F_n(t) - F(t) - W(t) U^{-1} \bar{u}^\tau + o_p(n^{-1/2}) \\
&= \frac{1}{n} \sum_{i=1}^n \left\{ I_{A(t)}(X_i) - F(t) - W(t) U^{-1} [u(X_i)]^\tau \right\} + o_p(n^{-1/2}).
\end{aligned}$$

The result follows from the CLT and the fact that

$$\begin{aligned}
\mathrm{Var}\big(W(t) U^{-1} [u(X_i)]^\tau\big) &= W(t) U^{-1} U U^{-1} [W(t)]^\tau \\
&= W(t) U^{-1} [W(t)]^\tau \\
&= E\{W(t) U^{-1} [u(X_i)]^\tau I_{A(t)}(X_i)\} \\
&= \mathrm{Cov}\big(I_{A(t)}(X_i), W(t) U^{-1} [u(X_i)]^\tau\big). \quad \blacksquare
\end{aligned}$$

Comparing (5.15) with (5.2), we conclude that $\hat{F}$ is asymptotically more efficient than F_n.

Example 5.1 (Survey problems). An example of situations in which we have auxiliary information expressed as (5.9) is a survey problem (Example 2.3) where the population $\mathcal{P} = \{y_1, ..., y_N\}$ consists of two-dimensional y_j's, $y_j = (y_{1j}, y_{2j})$, and the population mean $\bar{Y}_2 = N^{-1}\sum_{j=1}^N y_{2j}$ is known. For example, suppose that y_{1j} is the current year's income of unit j in the population and y_{2j} is the last year's income. In many applications the population total or mean of y_{2j}'s is known, for example, from tax return records. Let $X_1, ..., X_n$ be a simple random sample (see Example 2.3) selected from $\mathcal{P}$ *with* replacement. Then X_i's are i.i.d. bivariate random vectors whose c.d.f. is

$$F(t) = \frac{1}{N}\sum_{j=1}^N I_{A(t)}(y_j). \tag{5.16}$$

If $\bar{Y}_2$ is known, then it can be expressed as (5.9) with $u(x_1, x_2) = x_2 - \bar{Y}_2$. In survey problems X_i's are usually sampled *without* replacement so that $X_1, ..., X_n$ are not i.i.d. However, for a simple random sample without replacement, (5.8) can still be treated as an empirical likelihood, given X_i's. Note that F in (5.16) is the c.d.f. of X_i, regardless of whether X_i's are sampled with replacement.

If $X = (X_1, ..., X_n)$ is not a simple random sample, then the likelihood (5.8) has to be modified. Suppose that π_i is the probability that the ith unit is selected (see Theorem 3.15). Given $X = \{y_i, i \in \boldsymbol{s}\}$, an empirical likelihood is

$$\ell(G) = \prod_{i\in\boldsymbol{s}}[P_G(\{y_i\})]^{1/\pi_i} = \prod_{i\in\boldsymbol{s}} p_i^{1/\pi_i}, \tag{5.17}$$

where $p_i = P_G(\{y_i\})$. With the auxiliary information (5.9), an MELE of F in (5.16) can be obtained by maximizing $\ell(G)$ in (5.17) subject to (5.10). In this case F may not be the c.d.f. of X_i, but the c.d.f.'s of X_i's are determined by F and π_i's. It can be shown (exercise) that an MELE is given by (5.11) with

$$\hat{p}_i = \frac{1}{\pi_i[1 + u(y_i)\lambda_n^\tau]} \Big/ \sum_{i\in\boldsymbol{s}} \frac{1}{\pi_i} \tag{5.18}$$

and

$$\sum_{i\in\boldsymbol{s}} \frac{u(y_i)}{\pi_i[1 + u(y_i)\lambda_n^\tau]} = 0. \tag{5.19}$$

If π_i = a constant, then the MELE reduces to that in (5.11)-(5.13). If

$u(x) = 0$ (no auxiliary information), then the MELE is

$$\hat{F}(t) = \sum_{i \in \boldsymbol{s}} \frac{1}{\pi_i} I_{A(t)}(y_i) \Big/ \sum_{i \in \boldsymbol{s}} \frac{1}{\pi_i},$$

which is a ratio of two Horvitz-Thompson estimators (§3.4.2). Some asymptotic properties of the MELE $\hat{F}$ can be found in Chen and Qin (1993). ∎

The second part of Example 5.1 shows how to use empirical likelihoods in a non-i.i.d. problem. Applications of empirical likelihoods in non-i.i.d. problems are usually straightforward extensions of those in i.i.d. cases. The following is another example.

Example 5.2 (Biased sampling). Biased sampling is often used in applications. Suppose that $n = n_1 + \cdots + n_k$, $k \geq 2$; X_i's are independent random variables; $X_1, ..., X_{n_1}$ are i.i.d. with F; and $X_{n_j+1}, ..., X_{n_{j+1}}$ are i.i.d. with the c.d.f.

$$\int_{-\infty}^{t} w_{j+1}(s) dF(s) \Big/ \int_{-\infty}^{\infty} w_{j+1}(s) dF(s),$$

$j = 1, ..., k-1$, where w_j's are some nonnegative functions. A simple example is that $X_1, ..., X_{n_1}$ are sampled from F and $X_{n_1+1}, ..., X_{n_2}$ are sampled from F but conditional on the fact that each sampled value exceeds a given value x_0 (i.e., $w_2(s) = I_{(x_0,\infty)}(s)$). For instance, X_i's are blood pressure measurements; $X_1, ..., X_{n_1}$ are sampled from ordinary people and $X_{n_1+1}, ..., X_{n_2}$ are sampled from patients whose blood pressures are higher than x_0. The name biased sampling comes from the fact that there is a bias in the selection of samples.

For simplicity we consider the case of $k = 2$, since the extension to $k \geq 3$ is straightforward. Denote w_2 by w. An empirical likelihood is

$$\begin{aligned} \ell(G) &= \prod_{i=1}^{n_1} P_G(\{x_i\}) \prod_{i=n_1+1}^{n} \frac{w(x_i) P_G(\{x_i\})}{\int w(s) dG(s)} \\ &= \left[\sum_{i=1}^{n} p_i w(x_i) \right]^{-n_2} \prod_{i=1}^{n} p_i \prod_{i=n_1+1}^{n} w(x_i), \end{aligned} \tag{5.20}$$

where $p_i = P_G(\{x_i\})$. An MELE of F can be obtained by maximizing the empirical likelihood (5.20) subject to $p_i > 0$, $i = 1, ..., n$, and $\sum_{i=1}^{n} p_i = 1$. Using the Lagrange multiplier method we can show (exercise) that an MELE $\hat{F}$ is given by (5.11) with

$$\hat{p}_i = [n_1 + n_2 w(X_i)/\hat{w}]^{-1}, \qquad i = 1, ..., n, \tag{5.21}$$

where $\hat{w}$ satisfies

$$\hat{w} = \sum_{i=1}^{n} \frac{w(X_i)}{n_1 + n_2 w(X_i)/\hat{w}}.$$

An asymptotic result similar to that in Theorem 5.4 can be established (Vardi, 1985; Qin, 1993). ∎

Our last example concerns an important application in survival analysis.

Example 5.3 (Censored data). Let $Z_1, ..., Z_n$ be *survival times* that are i.i.d. nonnegative random variables from a c.d.f. F, and $Y_1, ..., Y_n$ be i.i.d. nonnegative random variables independent of Z_i's. In a variety of applications in biostatistics and life-time testing, we are only able to observe the smaller of Z_i and Y_i and an indicator of which variables is smaller:

$$X_i = \min(Z_i, Y_i), \qquad \delta_i = I_{(0,Y_i)}(Z_i), \qquad i = 1, ..., n.$$

This is called a *random censorship model* and Y_i's are called *censoring times*. We consider the estimation of the survival distribution F; see Kalbfleisch and Prentice (1980) for other problems involving censored data.

An MELE of F can be derived as follows. Let $x_{(1)} \leq \cdots \leq x_{(n)}$ be ordered values of X_i's and $\delta_{(i)}$ be the δ-value associated with $x_{(i)}$. Consider a c.d.f. G that assigns its mass to the points $x_{(1)}, ..., x_{(n)}$ and the interval $(x_{(n)}, \infty)$. Let $p_i = P_G(\{x_{(i)}\})$, $i = 1, ..., n$, and $p_{n+1} = 1 - G(x_{(n)})$. An MELE of F is then obtained by maximizing

$$\ell(G) = \prod_{i=1}^{n} p_i^{\delta_{(i)}} \left(\sum_{j=i+1}^{n+1} p_j \right)^{1-\delta_{(i)}} \tag{5.22}$$

subject to

$$p_i > 0, \quad i = 1, ..., n+1, \qquad \sum_{i=1}^{n+1} p_i = 1. \tag{5.23}$$

It can be shown (exercise) that an MELE is

$$\hat{F}(t) = \sum_{i=1}^{n+1} \hat{p}_i I_{(X_{(i-1)}, X_{(i)})}(t), \tag{5.24}$$

where $X_{(0)} = 0$, $X_{(n+1)} = \infty$, $X_{(1)} \leq \cdots \leq X_{(n)}$ are order statistics, and

$$\hat{p}_1 = \tfrac{1}{n}, \quad \hat{p}_i = \tfrac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \left(1 - \tfrac{\delta_{(j)}}{n-j+1}\right), \quad i = 2, ..., n, \quad \hat{p}_{n+1} = 1 - \sum_{j=1}^{n} \hat{p}_j.$$

The $\hat{F}$ in (5.24) can also be written as (exercise)

$$\hat{F}(t) = 1 - \prod_{X_{(i)} \le t} \left(1 - \tfrac{\delta_{(i)}}{n-i+1}\right), \tag{5.25}$$

which is the well-known Kaplan-Meier (1958) *product-limit* estimator. Some asymptotic results for $\hat{F}$ in (5.25) can be found, for example, in Shorack and Wellner (1986). ∎

5.1.3 Density estimation

Suppose that $X_1, ..., X_n$ are i.i.d. random variables from F and that F is unknown but has a Lebesgue p.d.f. f. Estimation of F can be done by estimating f, which is called *density estimation*. Note that estimators of F derived in §5.1.1 and §5.1.2 do not have Lebesgue p.d.f.'s.

Since $f(t) = F'(t)$ a.e., a simple estimator of $f(t)$ is the difference quotient

$$f_n(t) = \frac{F_n(t+\lambda_n) - F_n(t-\lambda_n)}{2\lambda_n}, \qquad t \in \mathcal{R}, \tag{5.26}$$

where F_n is the empirical c.d.f. given by (2.31) or (5.1) with $d = 1$, and $\{\lambda_n\}$ is a sequence of positive constants. Since $2n\lambda_n f_n(t)$ has the binomial distribution $Bi(F(t+\lambda_n) - F(t-\lambda_n), n)$,

$$E[f_n(t)] \to f(t) \qquad \text{if } \lambda_n \to 0 \text{ as } n \to \infty$$

and

$$\mathrm{Var}\big(f_n(t)\big) \to 0 \qquad \text{if } \lambda_n \to 0 \text{ and } n\lambda_n \to \infty.$$

Thus, we should choose λ_n converging to 0 slower than n^{-1}. If we assume that $\lambda_n \to 0$, $n\lambda_n \to \infty$, and f is continuously differentiable at t, then it can be shown (exercise) that

$$\mathrm{mse}_{f_n(t)}(F) = \frac{f(t)}{2n\lambda_n} + o\left(\frac{1}{n\lambda_n}\right) + O(\lambda_n^2) \tag{5.27}$$

and, under the additional condition that $n\lambda_n^3 \to 0$,

$$\sqrt{n\lambda_n}[f_n(t) - f(t)] \to_d N\big(0, \tfrac{1}{2}f(t)\big). \tag{5.28}$$

A useful class of estimators is the class of *kernel density estimators* of the form

$$\hat{f}(t) = \frac{1}{n\lambda_n} \sum_{i=1}^{n} w\left(\tfrac{t-X_i}{\lambda_n}\right), \tag{5.29}$$

where w is a known Lebesgue p.d.f. on $\mathcal{R}$ and is called the kernel. If we choose $w(t) = \frac{1}{2}I_{[-1,1]}(t)$, then $\hat{f}(t)$ in (5.29) is essentially the same as the so-called histogram. The bias of $\hat{f}(t)$ in (5.29) is

$$\begin{aligned} E[\hat{f}(t)] - f(t) &= \frac{1}{\lambda_n}\int w\left(\tfrac{t-z}{\lambda_n}\right) f(z)dz - f(t) \\ &= \int w(y)[f(t-\lambda_n y) - f(t)]dy. \end{aligned}$$

If f is bounded and is continuous at t, then, by the dominated convergence theorem (Theorem 1.1), the bias of $\hat{f}(t)$ converges to 0 as $\lambda_n \to 0$; if f' is bounded and is continuous at t and $\int |t|w(t)dt < \infty$, then the bias of $\hat{f}(t)$ is $O(\lambda_n)$. The variance of $\hat{f}(t)$ is

$$\begin{aligned} \mathrm{Var}\big(\hat{f}(t)\big) &= \frac{1}{n\lambda_n^2}\mathrm{Var}\left(w\left(\tfrac{t-X_1}{\lambda_n}\right)\right) \\ &= \frac{1}{n\lambda_n^2}\int \left[w\left(\tfrac{t-z}{\lambda_n}\right)\right]^2 f(z)dz \\ &\quad - \frac{1}{n}\left[\frac{1}{\lambda_n}\int w\left(\tfrac{t-z}{\lambda_n}\right) f(z)dz\right]^2 \\ &= \frac{1}{n\lambda_n}\int [w(y)]^2 f(t-\lambda_n y)dy + O\left(\frac{1}{n}\right) \\ &= \frac{w_0 f(t)}{n\lambda_n} + o\left(\frac{1}{n\lambda_n}\right) \end{aligned}$$

if f is bounded and is continuous at t and $w_0 = \int [w(t)]^2 dt < \infty$. Hence, if $\lambda_n \to 0$ and $n\lambda_n \to \infty$ and if f' is bounded and is continuous at t, then

$$\mathrm{mse}_{\hat{f}(t)}(F) = \frac{w_0 f(t)}{n\lambda_n} + O(\lambda_n^2).$$

Using the CLT (Theorem 1.15), one can show (exercise) that if $\lambda_n \to 0$, $n\lambda_n \to \infty$, f is bounded and is continuous at t, and $\int [w(t)]^{2+\delta}dt < \infty$ for some $\delta > 0$, then

$$\sqrt{n\lambda_n}\{\hat{f}(t) - E[\hat{f}(t)]\} \to_d N\big(0, w_0 f(t)\big). \tag{5.30}$$

Furthermore, if f' is bounded and is continuous at t and $n\lambda_n^3 \to 0$, then

$$\sqrt{n\lambda_n}\{E[\hat{f}(t)] - f(t)\} = O\left(\sqrt{n\lambda_n}\lambda_n\right) \to 0$$

and, therefore, (5.30) holds with $E[\hat{f}(t)]$ replaced by $f(t)$.

Similar to the estimation of a c.d.f., we can also study global properties of f_n or $\hat{f}$ as an estimator of the density curve f, using a suitably defined

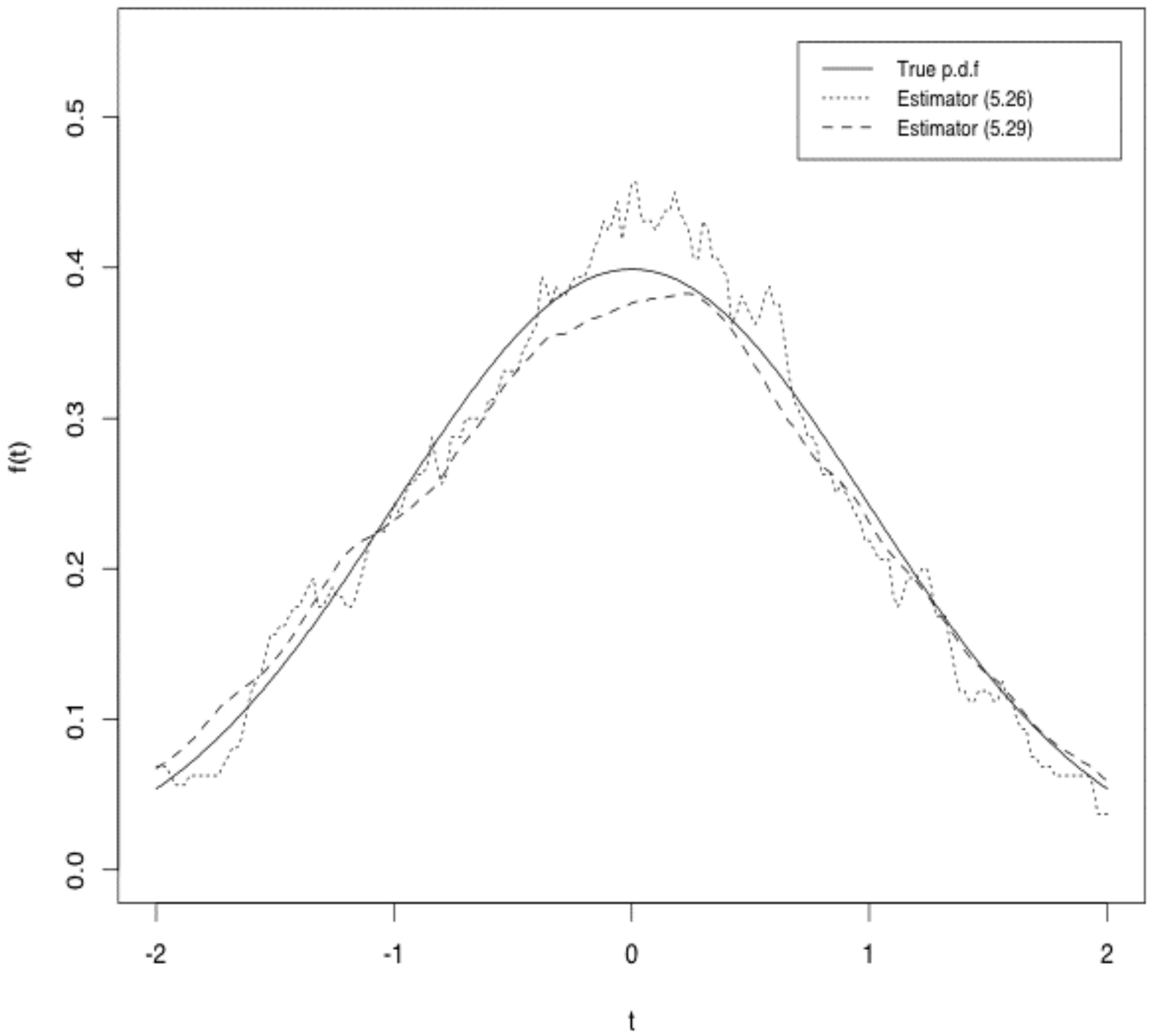

Figure 5.1: Density estimates in Example 5.4

distance between f and its density estimator. For example, we may study the convergence of $\sup_{t\in\mathcal{R}} |\hat{f}(t) - f(t)|$ or $\int |\hat{f}(t) - f(t)|^2 dt$. More details can be found, for example, in Silverman (1986).

Example 5.4. An i.i.d. sample of size $n = 200$ was generated from $N(0, 1)$. Density curve estimates (5.26) and (5.29) are plotted in Figure 5.1 with the curve of the true p.d.f. For the kernel density estimator (5.29), $w(t) = \frac{1}{2}e^{-|t|}$ is used and $\lambda_n = 0.4$. From Figure 5.1, it seems that the kernel estimate (5.29) is much better than the estimate (5.26). ∎

There are many other density estimation methods, for example, the nearest neighbor method (Stone, 1977), the smoothing splines (Wahba, 1990), and the method of empirical likelihoods described in §5.1.2 (see, e.g., Jones (1991)), which produces estimators of the form

$$\hat{f}(t) = \frac{1}{\lambda_n} \sum_{i=1}^{n} \hat{p}_i w\left(\frac{t - X_i}{\lambda_n}\right).$$

5.2 Statistical Functionals

In many nonparametric problems we are interested in estimating some characteristics (parameters) of the unknown population, not the entire population. We assume in this section that X_i's are i.i.d. from an unknown c.d.f. F on $\mathcal{R}^d$. Most characteristics of F can be written as $\mathtt{T}(F)$, where $\mathtt{T}$ is a functional from $\mathcal{F}$ to $\mathcal{R}^s$. If we estimate F by the empirical c.d.f. F_n in (5.1), then a natural estimator of $\mathtt{T}(F)$ is $\mathtt{T}(F_n)$, which is called a *statistical functional*.

Many commonly used statistics can be written as $\mathtt{T}(F_n)$ for some $\mathtt{T}$. Two simple examples are given as follows. Let $\mathtt{T}(F) = \int \psi(x)dF(x)$ with an integrable function ψ, and $\mathtt{T}(F_n) = \int \psi(x)dF_n(x) = n^{-1}\sum_{i=1}^n \psi(X_i)$. The sample moments discussed in §3.5.2 are particular examples of this kind of statistical functionals. For $d = 1$, let $\mathtt{T}(F) = F^{-1}(p) = \inf\{x : F(x) \geq p\}$, where $p \in (0, 1)$ is a fixed constant. $F^{-1}(p)$ is called the pth *quantile* of F. The statistical functional $\mathtt{T}(F_n) = F_n^{-1}(p)$ is called the pth *sample quantile*. More examples of statistical functionals are provided in §5.2.1 and §5.2.2.

In this section we study asymptotic distributions of $\mathtt{T}(F_n)$. We focus on the case of real-valued $\mathtt{T}$ ($s = 1$), since the extension to the case of $s \geq 2$ is straightforward.

5.2.1 Differentiability and asymptotic normality

Note that $\mathtt{T}(F_n)$ is a function of the "statistic" F_n. In Theorem 1.12 (and §3.5.1) we have studied how to use Taylor's expansion to establish asymptotic normality of differentiable functions of statistics that are asymptotically normal. This leads to the approach of establishing asymptotic normality of $\mathtt{T}(F_n)$ by using some generalized Taylor expansions for functionals and using asymptotic properties of F_n given in §5.1.1.

First, we need a suitably defined differential of $\mathtt{T}$. Several versions of differentials are given in the following definition.

Definition 5.2. Let $\mathtt{T}$ be a functional on $\mathcal{F}_0$, a collection of c.d.f.'s on $\mathcal{R}^d$, and let $\mathcal{D} = \{c(G_1 - G_2) : c \in \mathcal{R},\ G_j \in \mathcal{F}_0,\ j = 1, 2\}$.
(i) A functional $\mathtt{T}$ on $\mathcal{F}_0$ is Gâteaux differentiable at $G \in \mathcal{F}_0$ if and only if there is a linear functional $\mathtt{L}_G$ on $\mathcal{D}$ (i.e., $\mathtt{L}_G(c_1\Delta_1 + c_2\Delta_2) = c_1\mathtt{L}_G(\Delta_1) + c_2\mathtt{L}_G(\Delta_2)$ for any $\Delta_j \in \mathcal{D}$ and $c_j \in \mathcal{R}$) such that $\Delta \in \mathcal{D}$ and $G + t\Delta \in \mathcal{F}_0$ imply

$$\lim_{t\to 0}\left[\frac{\mathtt{T}(G + t\Delta) - \mathtt{T}(G)}{t} - \mathtt{L}_G(\Delta)\right] = 0.$$

(ii) Let ϱ be a distance on $\mathcal{F}_0$. Suppose that $\|c(G_1 - G_2)\| = |c|\varrho(G_1, G_2)$, $c \in \mathcal{R}$, $G_j \in \mathcal{F}_0$, defines a norm on $\mathcal{D}$ (i.e., $\|\Delta\| \geq 0$ and $= 0$ if and only

if $\Delta = 0$, $\|c\Delta\| = |c|\|\Delta\|$, and $\|\Delta + \tilde{\Delta}\| \leq \|\Delta\| + \|\tilde{\Delta}\|$, $\Delta \in \mathcal{D}$, $\tilde{\Delta} \in \mathcal{D}$, $c \in \mathcal{R}$). A functional $\mathtt{T}$ on $\mathcal{F}_0$ is ϱ-Hadamard differentiable at $G \in \mathcal{F}_0$ if and only if there is a linear functional $\mathtt{L}_G$ on $\mathcal{D}$ such that for any sequence of numbers $t_j \to 0$ and $\{\Delta, \Delta_j, j = 1, 2, ...\} \subset \mathcal{D}$ satisfying $\|\Delta_j - \Delta\| \to 0$ and $G + t_j\Delta_j \in \mathcal{F}_0$,

$$\lim_{j\to\infty} \left[\frac{\mathtt{T}(G + t_j\Delta_j) - \mathtt{T}(G)}{t_j} - \mathtt{L}_G(\Delta_j)\right] = 0.$$

(iii) Let ϱ be a distance on $\mathcal{F}_0$. A functional $\mathtt{T}$ on $\mathcal{F}_0$ is ϱ-Fréchet differentiable at $G \in \mathcal{F}_0$ if and only if there is a linear functional $\mathtt{L}_G$ on $\mathcal{D}$ such that for any sequence $\{G_j\}$ satisfying $G_j \in \mathcal{F}_0$ and $\varrho(G_j, G) \to 0$,

$$\lim_{j\to\infty} \frac{\mathtt{T}(G_j) - \mathtt{T}(G) - \mathtt{L}_G(G_j - G)}{\varrho(G_j, G)} = 0. \quad \blacksquare$$

The functional $\mathtt{L}_G$ is called the *differential* of $\mathtt{T}$ at G. If we define $h(t) = \mathtt{T}(G + t\Delta)$, then the Gâteaux differentiability is equivalent to the differentiability of the function $h(t)$ at $t = 0$, and $\mathtt{L}_G(\Delta)$ is simply $h'(0)$. Let δ_x denote the d-dimensional c.d.f. degenerated at the point x and $\phi_G(x) = \mathtt{L}_G(\delta_x - G)$. Then $\phi_F(x)$ is called the *influence function* of $\mathtt{T}$ at F, which is an important tool in robust statistics (see Hampel (1974)).

If $\mathtt{T}$ is Gâteaux differentiable at F, then we have the following expansion (taking $t = n^{-1/2}$ and $\Delta = \sqrt{n}(F_n - F)$):

$$\sqrt{n}[\mathtt{T}(F_n) - \mathtt{T}(F)] = \mathtt{L}_F\big(\sqrt{n}(F_n - F)\big) + R_n. \tag{5.31}$$

Since $\mathtt{L}_F$ is linear,

$$\mathtt{L}_F\big(\sqrt{n}(F_n - F)\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \phi_F(X_i) \to_d N(0, \sigma_F^2) \tag{5.32}$$

by the CLT, provided that

$$E[\phi_F(X_1)] = 0 \qquad \text{and} \qquad \sigma_F^2 = E[\phi_F(X_1)]^2 < \infty \tag{5.33}$$

(which is usually true when ϕ_F is bounded or when F has some finite moments). By Slutsky's theorem and (5.32),

$$\sqrt{n}[\mathtt{T}(F_n) - \mathtt{T}(F)] \to_d N(0, \sigma_F^2) \tag{5.34}$$

if R_n in (5.31) is $o_p(1)$.

Unfortunately, Gâteaux differentiability is too weak to be useful in establishing $R_n = o_p(1)$ (or (5.34)). This is why we need other types of

differentiability. Hadamard differentiability, which is also referred to as compact differentiability, is clearly stronger than Gâteaux differentiability but weaker than Fréchet differentiability (exercise). For a given functional $\mathtt{T}$, we can first find $\mathtt{L}_G$ by differentiating $h(t) = \mathtt{T}(G+t\Delta)$ at $t = 0$ and then check whether $\mathtt{T}$ is ϱ-Hadamard (or ϱ-Fréchet) differentiable with a given ϱ. The most commonly used distances on $\mathcal{F}_0$ are the sup-norm distance ϱ_∞ in (5.3) and the L_p distance ϱ_{L_p} in (5.5). Their corresponding norms on $\mathcal{D}$ are $\|\Delta\|_\infty = \sup_x |\Delta(x)|$ and $\|\Delta\|_{L_p} = [\int |\Delta(x)|^p dx]^{1/p}$, $\Delta \in \mathcal{D}$.

Theorem 5.5. Let $X_1, ..., X_n$ be i.i.d. from a c.d.f. F on $\mathcal{R}^d$.
(i) If $\mathtt{T}$ is ϱ_∞-Hadamard differentiable at F, then R_n in (5.31) is $o_p(1)$.
(ii) If $\mathtt{T}$ is ϱ-Fréchet differentiable at F with a distance ϱ satisfying

$$\sqrt{n}\varrho(F_n, F) = O_p(1), \tag{5.35}$$

then R_n in (5.31) is $o_p(1)$.
(iii) In either (i) or (ii), if (5.33) is also satisfied, then (5.34) holds.
Proof. Part (iii) follows directly from (i) or (ii). The proof of (i) involves some high-level mathematics and is omitted; see, for example, Fernholz (1983). We now prove (ii). From Definition 5.2(iii), for any $\epsilon > 0$, there is a $\delta > 0$ such that $|R_n| < \epsilon\sqrt{n}\varrho(F_n, F)$ whenever $\varrho(F_n, F) < \delta$. Then

$$P\left(|R_n| > \eta\right) \leq P\left(\sqrt{n}\varrho(F_n, F) > \eta/\epsilon\right) + P\left(\varrho(F_n, F) \geq \delta\right)$$

for any $\eta > 0$, which implies

$$\limsup_n P\left(|R_n| > \eta\right) \leq \limsup_n P\left(\sqrt{n}\varrho(F_n, F) > \eta/\epsilon\right).$$

The result follows from (5.35) and the fact that ϵ can be arbitrarily small. ∎

Since ϱ-Fréchet differentiability implies ϱ-Hadamard differentiability, Theorem 5.5(ii) is useful when ϱ is not the sup-norm distance. There are functionals that are not ϱ_∞-Hadamard differentiable (and hence not ϱ_∞-Fréchet differentiable). For example, if $d = 1$ and $\mathtt{T}(G) = g(\int x dG)$ with a differentiable function g, then T is not necessarily ϱ_∞-Hadamard differentiable, but is ϱ_{L_1}-Fréchet differentiable (exercise).

From Theorem 5.2, condition (5.35) holds for ϱ_{L_p} under the moment conditions on F given in Theorem 5.2.

Note that if ϱ and $\tilde{\varrho}$ are two distances on $\mathcal{F}_0$ satisfying $\tilde{\varrho}(G_1, G_2) \leq c\varrho(G_1, G_2)$ for a constant c and all $G_j \in \mathcal{F}_0$, then $\tilde{\varrho}$-Hadamard (Fréchet) differentiability implies ϱ-Hadamard (Fréchet) differentiability. This suggests the use of the distance $\varrho_{\infty+p} = \varrho_\infty + \varrho_{L_p}$, which also satisfies (5.35) under the moment conditions in Theorem 5.2. The distance $\varrho_{\infty+p}$ is useful in some cases (Theorem 5.6).

A ϱ_∞-Hadamard differentiable $\mathtt{T}$ having a bounded and continuous influence function ϕ_F is *robust* in Hampel's sense (see, e.g., Huber (1981)). This is motivated by the fact that the asymptotic behavior of $\mathtt{T}(F_n)$ is determined by that of $\mathtt{L}_F(F_n - F)$, and a small change in the sample, i.e., small changes in all x_i's (rounding, grouping) or large changes in a few of x_i's (gross errors, blunders), will result in a small change of $\mathtt{T}(F_n)$ if and only if ϕ_F is bounded and continuous.

We now consider some examples. For the sample moments related to functionals of the form $\mathtt{T}(G) = \int \psi(x)dG(x)$, it is clear that $\mathtt{T}$ is a linear functional. Any linear functional is trivially ϱ-Fréchet differentiable for any ϱ. Next, if F is one-dimensional and $F'(x) > 0$ for all x, then the quantile functional $\mathtt{T}(G) = G^{-1}(p)$ is ϱ_∞-Hadamard differentiable at F (Fernholz, 1983). Hence, Theorem 5.5 applies to these functionals. But the asymptotic normality of sample quantiles can be established under weaker conditions, which are studied in §5.3.1.

Example 5.5 (Convolution functionals). Suppose that F is on $\mathcal{R}$ and for a fixed $z \in \mathcal{R}$,

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

If X_1 and X_2 are i.i.d. with c.d.f. G, then $\mathtt{T}(G)$ is the c.d.f. of $X_1 + X_2$ (Exercise 42 in §1.6), and is also called the convolution of G evaluated at z. For $t_j \to 0$ and $\|\Delta_j - \Delta\|_\infty \to 0$,

$$\mathtt{T}(G + t_j\Delta_j) - \mathtt{T}(G) = 2t_j \int \Delta_j(z-y)dG(y) + t_j^2 \int \Delta_j(z-y)d\Delta_j(y)$$

(for $\Delta = c_1G_1 + c_2G_2$, $G_j \in \mathcal{F}_0$, and $c_j \in \mathcal{R}$, $d\Delta$ denotes $c_1dG_1 + c_2dG_2$). Using Lemma 5.2, one can show (exercise) that

$$\int \Delta_j(z-y)d\Delta_j(y) = O(1). \tag{5.36}$$

Hence $\mathtt{T}$ is ϱ_∞-Hadamard differentiable at any $G \in \mathcal{F}$ with $\mathtt{L}_G(\Delta) = 2\int \Delta(z-y)dG(y)$. The influence function, $\phi_F(x) = 2\int(\delta_x - F)(z-y)dF(y)$, is a bounded function and clearly satisfies (5.33). Thus, (5.34) holds. If F is continuous, then $\mathtt{T}$ is robust in Hampel's sense (exercise). ∎

Three important classes of statistical functionals, i.e., L-estimators, M-estimators, and rank statistics and R-estimators, are considered in §5.2.2.

Lemma 5.2. Let $\Delta \in \mathcal{D}$ and h be a function on $\mathcal{R}$ such that $\int h(x)d\Delta(x)$ is finite. Then

$$\left| \int h(x)d\Delta(x) \right| \le \|h\|_V \|\Delta\|_\infty,$$

where $\|h\|_V$ is the variation norm defined by

$$\|h\|_V = \lim_{a\to-\infty, b\to\infty} \left[\sup \sum_{j=1}^{m} |h(x_j) - h(x_{j-1})|\right]$$

with the supremum being taken over all partitions $a = x_0 < \cdots < x_m = b$ of the interval $[a, b]$. ∎

The proof of Lemma 5.2 can be found, for example, in Natanson (1961, p. 232).

The differentials in Definition 5.2 are first-order differentials. For some functionals, we can also consider their second-order differentials.

Definition 5.3. Let $\mathtt{T}$ be a functional on $\mathcal{F}_0$ and ϱ be a distance on $\mathcal{F}_0$.
(i) $\mathtt{T}$ is second-order ϱ-Hadamard differentiable at $G \in \mathcal{F}_0$ if and only if there is a functional $\mathtt{Q}_G$ on $\mathcal{D}$ such that for any sequence of numbers $t_j \to 0$ and $\{\Delta, \Delta_j, j = 1, 2, ...\} \subset \mathcal{D}$ satisfying $\|\Delta_j - \Delta\| \to 0$ and $G + t_j\Delta_j \in \mathcal{F}_0$,

$$\lim_{j\to\infty} \frac{\mathtt{T}(G + t_j\Delta_j) - \mathtt{T}(G) - \mathtt{Q}_G(t_j\Delta_j)}{t_j^2} = 0,$$

where $\mathtt{Q}_G(\Delta) = \int\int \psi_G(x, y)d(G + t_j\Delta)(x)d(G + t_j\Delta)(y)$ for a function ψ_G satisfying $\psi_G(x, y) = \psi_G(y, x)$, $\int\int \psi_G(x, y)dG(x)dG(y) = 0$, and $\|\cdot\|$ is the same as that in Definition 5.2(ii).
(ii) $\mathtt{T}$ is second-order ϱ-Fréchet differentiable at $G \in \mathcal{F}_0$ if and only if, for any sequence $\{G_j\}$ satisfying $G_j \in \mathcal{F}_0$ and $\varrho(G_j, G) \to 0$,

$$\lim_{j\to\infty} \frac{T(G_j) - T(G) - \mathtt{Q}_G(G_j - G)}{[\varrho(G_j, G)]^2} = 0,$$

where $\mathtt{Q}_G$ is the same as that in (i). ∎

For a second-order differentiable $\mathtt{T}$, we have the following expansion:

$$n[\mathtt{T}(F_n) - \mathtt{T}(F)] = nV_n + R_n, \tag{5.37}$$

where

$$V_n = \mathtt{Q}_G(F_n - F) = \int\int \psi_F(x, y)dF_n(x)dF_n(y) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} \psi_F(X_i, X_j)$$

is a "V-statistic" (§3.5.3) whose asymptotic properties are given by Theorem 3.16. If R_n in (5.37) is $o_p(1)$, then the asymptotic behavior of $\mathtt{T}(F_n) - \mathtt{T}(F)$ is the same as that of V_n.

Proposition 5.1. Let $X_1, ..., X_n$ be i.i.d. from F.
(i) If $\mathtt{T}$ is second-order ϱ_∞-Hadamard differentiable at F, then R_n in (5.37) is $o_p(1)$.
(ii) If $\mathtt{T}$ is second-order ϱ-Fréchet differentiable at F with a distance ϱ satisfying (5.35), then R_n in (5.37) is $o_p(1)$. ∎

Combining Proposition 5.1 with Theorem 3.16, we can summarize the asymptotic behavior of $\mathtt{T}(F_n) - \mathtt{T}(F)$ as follows. If

$$\zeta_1 = \mathrm{Var}\left(\int \psi_F(X_1, y)dF(y)\right)$$

is positive, then (5.34) holds with $\sigma_F^2 = 4\zeta_1$. If $\zeta_1 = 0$, then

$$n[\mathtt{T}(F_n) - \mathtt{T}(F)] \to_d \sum_{j=1}^{\infty} \lambda_j \chi_{1j}^2.$$

If $\mathtt{T}$ is also first-order differentiable, then it can be shown (exercise) that

$$\phi_F(x) = 2\int \psi_F(x, y)dF(y). \tag{5.38}$$

Then $\zeta_1 = 4^{-1}\mathrm{Var}(\phi_F(X_1))$ and $\zeta_1 = 0$ corresponds to the case of $\phi_F(x) \equiv 0$. However, second-order ϱ-Hadamard (Fréchet) differentiability does not in general imply first-order ϱ-Hadamard (Fréchet) differentiability (exercise).

The technique in this section can be applied to non-i.i.d. X_i's when the c.d.f.'s of X_i's are determined by an unknown c.d.f. F, provided that results similar to (5.32) and (5.35) (with F_n replaced by some other estimator $\hat{F}$) can be established.

5.2.2 L-, M-, R-estimators and rank statistics

Three large classes of statistical functionals based on i.i.d. X_i's are studied in this section.

L-estimators

Let F be a c.d.f. on $\mathcal{R}$ and $J(t)$ be a function on $[0, 1]$. An *L-functional* is defined as

$$\mathtt{T}(G) = \int xJ(G(x))dG(x), \qquad G \in \mathcal{F}. \tag{5.39}$$

$\mathtt{T}(F_n)$ is called an L-estimator of $\mathtt{T}(F)$.

Example 5.6. The following are some examples of commonly used L-estimators.
(i) When $J \equiv 1$, $\mathtt{T}(F_n) = \bar{X}$, the sample mean.
(ii) When $J(t) = 4t - 2$, $\mathtt{T}(F_n)$ is called Gini's mean difference.
(iii) When $J(t) = (\beta - \alpha)^{-1} I_{(\alpha,\beta)}(t)$ for some constants $\alpha < \beta$, $\mathtt{T}(F_n)$ is called the trimmed sample mean. ∎

For an L-functional $\mathtt{T}$, it can be shown (exercise) that

$$\mathtt{T}(G) - \mathtt{T}(F) = \int \phi_F(x) d(G - F)(x) + R(G, F), \tag{5.40}$$

where

$$\phi_F(x) = -\int (\delta_x - F)(y) J(F(y)) dy, \tag{5.41}$$

$$R(G, F) = -\int W_G(x)[G(x) - F(x)]dx,$$

and

$$W_G(x) = \begin{cases} [G(x) - F(x)]^{-1} \int_{F(x)}^{G(x)} J(t)dt - J(F(x)) & G(x) \neq F(x) \\ 0 & G(x) = F(x). \end{cases}$$

A sufficient condition for (5.33) in this case is that F has a finite variance (exercise). However, (5.33) is also satisfied if ϕ_F is bounded. The differentiability of $\mathtt{T}$ can be verified under some conditions on J.

Theorem 5.6. Let $\mathtt{T}$ be an L-functional defined by (5.39).
(i) Suppose that J is bounded, $J(t) = 0$ when $t \in [0, \alpha] \cup [\beta, 1]$ for some constants $\alpha < \beta$, and that the set $D = \{x : J \text{ is discontinuous at } F(x)\}$ has Lebesgue measure 0. Then $\mathtt{T}$ is ϱ_∞-Fréchet differentiable at F with the influence function ϕ_F given by (5.41), and ϕ_F is bounded and continuous and satisfies (5.33).
(ii) Suppose that J is bounded, the set D in (i) has Lebesgue measure 0, and J is continuous on $[0, \alpha] \cup [\beta, 1]$ for some constants $\alpha < \beta$. Then $\mathtt{T}$ is $\varrho_{\infty+1}$-Fréchet differentiable at F.
(iii) Suppose that $|J(t) - J(s)| \leq C|t - s|^{p-1}$, where $C > 0$ and $p > 1$ are some constants. Then $\mathtt{T}$ is ϱ_{L_p}-Fréchet differentiable at F.
(iv) If, in addition to the conditions in part (i), J' is continuous on $[\alpha, \beta]$, then $\mathtt{T}$ is second-order ϱ_∞-Fréchet differentiable at F with

$$\psi_F(x, y) = \phi_F(x) + \phi_F(y) - \int (\delta_x - F)(z)(\delta_y - F)(z) J'(F(z)) dz.$$

(v) Suppose that J' is continuous on $[0, 1]$. Then $\mathtt{T}$ is second-order ϱ_{L_2}-Fréchet differentiable at F with the same ψ_F given in (iv).

Proof. We prove (i)-(iii). The proofs for (iv) and (v) are similar and are left to the reader.
(i) Let $G_j \in \mathcal{F}$ and $\varrho_\infty(G_j, F) \to 0$. Let c and d be two constants such that $F(c) > \beta$ and $F(d) < \alpha$. Then, for sufficiently large j, $G_j(x) \in [0, \alpha] \cup [\beta, 1]$ if $x > c$ or $x < d$. Hence, for sufficiently large j,

$$\begin{aligned} |R(G_j, F)| &= \left| \int_d^c W_{G_j}(x)(G_j - F)(x)dx \right| \\ &\leq \varrho_\infty(G_j, F) \int_d^c |W_{G_j}(x)|dx. \end{aligned}$$

Since J is continuous at $F(x)$ when $x \notin D$ and D has Lebesgue measure 0, $W_{G_j}(x) \to 0$ a.e. Lebesgue. By the dominated convergence theorem, $\int_d^c |W_{G_j}(x)|dx \to 0$. This proves that $\mathtt{T}$ is ϱ_∞-Fréchet differentiable. The assertions on ϕ_F can be proved by noting that

$$\phi_F(x) = -\int_d^c (\delta_x - F)(y)J(F(y))dy.$$

(ii) From the proof of (i), we only need to show that

$$\left| \int_A W_{G_j}(x)(G_j - F)(x)dx \right| \Big/ \varrho_{\infty+1}(G_j, F) \to 0, \tag{5.42}$$

where $A = \{x : F(x) \leq \alpha \text{ or } F(x) > \beta\}$. The quantity on the left-hand side of (5.42) is bounded by $\sup_{x \in A} |W_{G_j}(x)|$ which converges to 0 under the continuity assumption of J on $[0, \alpha] \cup [\beta, 1]$. Hence (5.42) follows.
(iii) The result follows from

$$|R(G, F)| \leq C \int |G(x) - F(x)|^p dx = O\left([\varrho_{L_p}(G, F)]^p\right)$$

and the fact that $p > 1$. ∎

An L-estimator with $J(t) = 0$ when $t \in [0, \alpha] \cup [\beta, 1]$ is called a trimmed L-estimator. Theorem 5.6(i) shows that trimmed L-estimators satisfy (5.34) and are robust in Hampel's sense. In case (ii) or (iii) of Theorem 5.6, (5.34) holds if $\mathrm{Var}(X_1) < \infty$, but $\mathtt{T}(F_n)$ may not be robust in Hampel's sense. It can be shown (exercise) that one or several of (i)-(v) of Theorem 5.6 can be applied to each of the L-estimators in Example 5.6.

M-estimators

Let F be a c.d.f. on $\mathcal{R}^d$ and $\rho(x, t)$ be a Borel function on $\mathcal{R}^d \times \mathcal{R}$. An *M-functional* is defined to be a solution of

$$\int \rho(x, \mathtt{T}(G))dG(x) = \min_{t \in \Theta} \int \rho(x, t)dG(x), \qquad G \in \mathcal{F}, \tag{5.43}$$

where Θ is an open subset of $\mathcal{R}$. $\mathtt{T}(F_n)$ is called an M-estimator of $\mathtt{T}(F)$. Assume that $\psi(x,t) = \partial\rho(x,t)/\partial t$ exists a.e. and

$$\lambda_G(t) = \int \psi(x,t)dG(x) = \frac{\partial}{\partial t}\int \rho(x,t)dG(x). \tag{5.44}$$

Then $\lambda_G(\mathtt{T}(G)) = 0$.

Example 5.7. The following are some examples of M-estimators.
(i) If $\rho(x,t) = (x-t)^2/2$, then $\psi(x,t) = t - x$; $\mathtt{T}(G) = \int xdG(x)$ is the mean functional; and $\mathtt{T}(F_n) = \bar{X}$ is the sample mean.
(ii) If $\rho(x,t) = |x-t|^p/p$, where $p \in [1,2)$, then

$$\psi(x,t) = \begin{cases} |x-t|^{p-1} & x < t \\ 0 & x = t \\ -|x-t|^{p-1} & x > t. \end{cases}$$

If $p = 1$, $\mathtt{T}(F_n)$ is the sample median. If $1 < p < 2$, $T(F_n)$ is called the pth least absolute deviations estimator or the minimum L_p distance estimator.
(iii) Let $\mathcal{F}_0 = \{f_\theta : \theta \in \Theta\}$ be a parametric family of p.d.f.'s and $\rho(x,t) = -\log f_t(x)$. Then $\mathtt{T}(F_n)$ is an MLE. This indicates that M-estimators are extensions of MLE's in parametric models.
(iv) Huber (1964) considers

$$\rho(x,t) = \begin{cases} \frac{1}{2}(x-t)^2 & |x-t| \le C \\ C^2 & |x-t| > C \end{cases}$$

with

$$\psi(x,t) = \begin{cases} t - x & |x-t| \le C \\ 0 & |x-t| > C. \end{cases}$$

The corresponding $\mathtt{T}(F_n)$ is a type of trimmed sample mean.
(v) Huber (1964) considers

$$\rho(x,t) = \begin{cases} \frac{1}{2}(x-t)^2 & |x-t| \le C \\ C|x-t| - \frac{1}{2}C^2 & |x-t| > C \end{cases}$$

with

$$\psi(x,t) = \begin{cases} C & t - x > C \\ t - x & |x-t| \le C \\ -C & t - x < -C. \end{cases}$$

The corresponding $\mathtt{T}(F_n)$ is a type of Winsorized sample mean.
(vi) Hampel (1974) considers $\psi(x,t) = \psi_0(t-x)$ with $\psi_0(s) = -\psi_0(-s)$ and

$$\psi_0(s) = \begin{cases} s & 0 \le s \le a \\ a & a < s \le b \\ \frac{a(c-s)}{c-b} & b < s \le c \\ 0 & s > c, \end{cases}$$

where $0 < a < b < c$ are constants. A smoothed version of ψ_0 is

$$\psi_1(s) = \begin{cases} \sin(as) & 0 \le s < \pi/a \\ 0 & s > \pi/a. \end{cases} \qquad \blacksquare$$

For bounded and continuous ψ, the following result shows that $\mathtt{T}$ is ϱ_∞-Hadamard differentiable with a bounded and continuous influence function and, hence, $\mathtt{T}(F_n)$ satisfies (5.34) and is robust in Hampel's sense.

Theorem 5.7. Let $\mathtt{T}$ be an M-functional defined by (5.43). Assume that ψ is a bounded and continuous function on $\mathcal{R}^d \times \mathcal{R}$ and that $\lambda_F(t)$ is continuously differentiable at $\mathtt{T}(F)$ and $\lambda_F'(\mathtt{T}(F)) \neq 0$. Then $\mathtt{T}$ is ϱ_∞-Hadamard differentiable at F with

$$\phi_F(x) = -\psi(x, \mathtt{T}(F))/\lambda_F'(\mathtt{T}(F)).$$

Proof. Let $t_j \to 0$, $\Delta_j \in \mathfrak{D}$, $\|\Delta_j - \Delta\|_\infty \to 0$, and $G_j = F + t_j\Delta_j \in \mathfrak{F}$. Since $\lambda_G(\mathtt{T}(G)) = 0$,

$$|\lambda_F(\mathtt{T}(G_j)) - \lambda_F(\mathtt{T}(F))| = \left| t_j \int \psi(x, \mathtt{T}(G_j)) d\Delta_j(x) \right| \to 0$$

by $\|\Delta_j - \Delta\|_\infty \to 0$ and the boundedness of ψ. Note that $\lambda_F'(\mathtt{T}(F)) \neq 0$. Hence, the inverse of $\lambda_F(t)$ exists and is continuous in a neighborhood of $0 = \lambda_F(\mathtt{T}(F))$. Therefore,

$$\mathtt{T}(G_j) - \mathtt{T}(F) \to 0. \tag{5.45}$$

Let $h_F(\mathtt{T}(F)) = \lambda_F'(\mathtt{T}(F))$, $h_F(t) = [\lambda_F(t) - \lambda_F(\mathtt{T}(F))]/[t - \mathtt{T}(F)]$ if $t \neq \mathtt{T}(F)$,

$$R_{1j} = \int \psi(x, \mathtt{T}(F)) d\Delta_j(x) \left[\frac{1}{\lambda_F'(\mathtt{T}(F))} - \frac{1}{h_F(\mathtt{T}(G_j))} \right],$$

$$R_{2j} = \frac{1}{h_F(\mathtt{T}(G_j))} \int [\psi(x, \mathtt{T}(G_j)) - \psi(x, \mathtt{T}(F))] d\Delta_j(x),$$

and

$$\mathtt{L}_F(\Delta) = -\frac{1}{\lambda_F'(\mathtt{T}(F))} \int \psi(x, \mathtt{T}(F)) d\Delta(x), \qquad \Delta \in \mathfrak{D}.$$

Then

$$\mathtt{T}(G_j) - \mathtt{T}(F) = -\mathtt{L}_F(t_j\Delta_j) + t_j(R_{1j} - R_{2j}).$$

By (5.45), $\|\Delta_j - \Delta\|_\infty \to 0$, and the boundedness of ψ, $R_{j1} \to 0$. The result then follows from $R_{2j} \to 0$, which follows from $\|\Delta_j - \Delta\|_\infty \to 0$ and the boundedness and continuity of ψ (exercise). $\blacksquare$

Some ψ functions in Example 5.7 satisfy the conditions in Theorem 5.7 (exercise). Under more conditions on ψ, it can be shown that an M-functional is ϱ_∞-Fréchet differentiable at F (Clarke, 1986; Shao, 1993). Some M-estimators that satisfy (5.34) but are not differentiable functionals are studied in §5.4.

Rank statistics and R-estimators

Assume that $X_1, ..., X_n$ are i.i.d. from a c.d.f. F on $\mathcal{R}$. The *rank* of X_i among $X_1, ..., X_n$, denoted by R_i, is defined to be the number of X_j's satisfying $X_j \le X_i$, $i = 1, ..., n$. The rank of $|X_i|$ among $|X_1|, ..., |X_n|$ is similarly defined and denoted by $\tilde{R}_i$. A statistic that is a function of R_i's or $\tilde{R}_i$'s is called a *rank statistic*. For $G \in \mathcal{F}$, let

$$\tilde{G}(x) = G(x) - G\big((-x)-\big), \qquad x > 0,$$

where $g(x-)$ denotes the left limit of the function g at x. Define a functional $\mathtt{T}$ by

$$\mathtt{T}(G) = \int_0^\infty J(\tilde{G}(x))dG(x), \qquad G \in \mathcal{F}, \tag{5.46}$$

where J is a function on $[0, 1]$ with a bounded J'. Then

$$\mathtt{T}(F_n) = \int_0^\infty J(\tilde{F}_n(x))dF_n(x) = \frac{1}{n}\sum_{i=1}^n J\left(\frac{\tilde{R}_i}{n}\right) I_{(0,\infty)}(X_i)$$

is a (one-sample) signed rank statistic. If $J(t) = t$, then $\mathtt{T}(F_n)$ is the well-known Wilcoxon signed rank test (§6.5.1).

Statistics based on ranks (or signed ranks) are robust against changes in values of x_i's, but may not provide efficient inference procedures, since the values of x_i's are discarded after ranks (or signed ranks) are determined.

It can be shown (exercise) that $\mathtt{T}$ in (5.46) is ϱ_∞-Hadamard differentiable at F with the differential

$$\mathtt{L}_F(\Delta) = \int_0^\infty J'(\tilde{F}(x))\tilde{\Delta}(x)dF(x) + \int_0^\infty J(\tilde{F}(x))d\Delta(x). \tag{5.47}$$

These results can be extended to the case where $X_1, ..., X_n$ are i.i.d. from a c.d.f. F on $\mathcal{R}^2$. For any c.d.f. G on $\mathcal{R}^2$, let J be a function on $[0, 1]$ with $J(1 - t) = -J(t)$ and a bounded J',

$$\bar{G}(y) = [G(y, \infty) + G(\infty, y)]/2, \qquad y \in \mathcal{R},$$

and

$$\mathtt{T}(G) = \int J(\bar{G}(y))dG(y, \infty). \tag{5.48}$$

Let $X_i = (Y_i, Z_i)$, R_i be the rank of Y_i, and U_i be the number of Z_j's satisfying $Z_j \le Y_i$, $i = 1, ..., n$. Then

$$\mathtt{T}(F_n) = \int J(\bar{F}_n(y))dF_n(y,\infty) = \frac{1}{n}\sum_{i=1}^{n} J\left(\tfrac{R_i+U_i}{2n}\right)$$

is called a two-sample linear rank statistic. It can be shown (exercise) that $\mathtt{T}$ in (5.48) is ϱ_∞-Hadamard differentiable at F with the differential

$$\mathtt{L}_F(\Delta) = \int J'(\bar{F}(y))\bar{\Delta}(y)dF(y,\infty) + \int J(\bar{F}(y))d\Delta(y,\infty). \tag{5.49}$$

Rank statistics (one-sample or two-sample) are asymptotically normal and robust in Hampel's sense (exercise). These results are useful in testing hypotheses (§6.5).

Let F be a continuous c.d.f. on $\mathcal{R}$ symmetric about an unknown parameter $\theta \in \mathcal{R}$. An estimator of θ closely related to a rank statistic can be derived as follows. Let X_i be i.i.d. from F and $W_i = (X_i, 2t - X_i)$ with a fixed $t \in \mathcal{R}$. The functional $\mathtt{T}$ in (5.48) evaluated at the c.d.f. of W_i is equal to

$$\lambda_F(t) = \int J\left(\tfrac{F(x)+1-F(2t-x)}{2}\right) dF(x). \tag{5.50}$$

If J is strictly increasing and F is strictly increasing in a neighborhood of θ, then $\lambda_F(t) = 0$ if and only if $t = \theta$ (exercise). For $G \in \mathcal{F}$, define $\mathtt{T}(G)$ to be a solution of

$$\int J\left(\tfrac{G(x)+1-G(2\mathtt{T}(G)-x)}{2}\right) dG(x) = 0. \tag{5.51}$$

$\mathtt{T}(F_n)$ is called an *R-estimator* of $\mathtt{T}(F) = \theta$. When $J(t) = t - \frac{1}{2}$ (which is related to the Wilcoxon signed rank test), $\mathtt{T}(F_n)$ is the well-known Hodges-Lehmann estimator and is equal to any value between the two middle points of the values $(X_i + X_j)/2$, $i = 1, ..., n$, $j = 1, ..., n$.

Theorem 5.8. Let $\mathtt{T}$ be the functional defined by (5.51). Suppose that F is continuous and symmetric about θ, the derivatives F' and J' exist, and J' is bounded. Then $\mathtt{T}$ is ϱ_∞-Hadamard differentiable at F with the influence function

$$\phi_F(x) = \frac{J(F(x))}{\int J'(F(x))F'(x)dF(x)}.$$

Proof. Since F is symmetric about θ, $F(x) + F(2\theta - x) = 1$. Under the assumed conditions, $\lambda_F(t)$ is continuous and $\int J'(F(x))F'(x)dF(x) = -\lambda_F'(\theta) \neq 0$ (exercise). Hence the inverse of λ_F exists and is continuous

at $0 = \lambda_F(\theta)$. Suppose that $t_j \to 0$, $\Delta_j \in \mathfrak{D}$, $\|\Delta_j - \Delta\|_\infty \to 0$, and $G_j = F + t_j\Delta_j \in \mathfrak{F}$. Then

$$\int [J(G_j(x,t)) - J(F(x,t))]dG_j(x) \to 0$$

uniformly in t, where $G(x,t) = [G(x) + 1 - G(2t - x)]/2$, and

$$\int J(F(x,t))d(G_j - F)(x) = \int (G_j - F)(x)J'(F(x,t))dF(x,t) \to 0$$

uniformly in t. Let $\lambda_G(t)$ be defined by (5.50) with F replaced by G. Then

$$\lambda_{G_j}(t) - \lambda_F(t) \to 0$$

uniformly in t. Thus, $\lambda_F(\mathtt{T}(G_j)) \to 0$, which implies

$$\mathtt{T}(G_j) \to \mathtt{T}(F) = \theta. \tag{5.52}$$

Let $\xi_G(t) = \int J(F(x,t))dG(x)$, $h_F(t) = [\lambda_F(t) - \lambda_F(\theta)]/(t - \theta)$ if $t \neq \theta$, and $h_F(\theta) = \lambda'_F(\theta)$. Then $\mathtt{T}(G_j) - \mathtt{T}(F) - \int \phi_F(x)d(G_j - F)(x)$ is equal to

$$\xi_{G_j}(\theta)\left[\frac{1}{\lambda'_F(\theta)} - \frac{1}{h_F(\mathtt{T}(G_j))}\right] + \frac{\lambda_F(\mathtt{T}(G_j)) - \xi_{G_j}(\theta)}{h_F(\mathtt{T}(G_j))}. \tag{5.53}$$

Note that

$$\xi_{G_j}(\theta) = \int J(F(x))dG_j(x) = t_j \int J(F(x))d\Delta_j(x).$$

By (5.52), Lemma 5.2, and $\|\Delta_j - \Delta\|_\infty \to 0$, the first term in (5.53) is $o(t_j)$. The second term in (5.53) is the sum of

$$-\frac{t_j}{h_F(\mathtt{T}(G_j))}\int [J(F(x,\mathtt{T}(G_j))) - J(F(x))]d\Delta_j(x) \tag{5.54}$$

and

$$\frac{1}{h_F(\mathtt{T}(G_j))}\int [J(F(x,\mathtt{T}(G_j))) - J(G_j(x,\mathtt{T}(G_j)))]dG_j(x). \tag{5.55}$$

From the continuity of J and F, the quantity in (5.54) is $o(t_j)$. Similarly, the quantity in (5.55) is equal to

$$\frac{1}{h_F(\mathtt{T}(G_j))}\int [J(F(x,\mathtt{T}(G_j))) - J(G_j(x,\mathtt{T}(G_j)))]dF(x) + o(t_j). \tag{5.56}$$

Using Taylor's expansion, (5.52), and $\|\Delta_j - \Delta\|_\infty \to 0$, the quantity in (5.56) is equal to

$$\frac{t_j}{h_F(\mathtt{T}(G_j))}\int J'(F(x))\Delta(x,\theta)dF(x) + o(t_j). \tag{5.57}$$

Since $J(1-t) = -J(t)$, the integral in (5.57) is 0. This proves that the second term in (5.53) is $o(t_j)$ and thus the result. ∎

It is clear that the influence function ϕ_F for an R-estimator is bounded and continuous if J and F are continuous. Thus, R-estimators satisfy (5.34) and are robust in Hampel's sense.

Example 5.8. Let $J(t) = t - \frac{1}{2}$. Then $\mathtt{T}(F_n)$ is the Hodges-Lehmann estimator. From Theorem 5.8, $\phi_F(x) = [F(x) - \frac{1}{2}]/\gamma$, where $\gamma = \int F'(x)dF(x)$. Since $F(X_1)$ has a uniform distribution on $[0,1]$, $\phi_F(X_1)$ has mean 0 and variance $(12\gamma^2)^{-1}$. Thus, $\sqrt{n}[\mathtt{T}(F_n) - \mathtt{T}(F)] \to_d N(0, (12\gamma^2)^{-1})$. ∎

5.3 Linear Functions of Order Statistics

In this section we study statistics that are linear functions of order statistics $X_{(1)} \leq \cdots \leq X_{(n)}$, based on independent random variables $X_1, ..., X_n$ (in §5.3.1 and §5.3.2, $X_1, ..., X_n$ are assumed i.i.d.). Order statistics, first introduced in Example 2.9, are usually sufficient and often complete (or minimal sufficient) for nonparametric families (Examples 2.12 and 2.14).

L-estimators defined in §5.2.2 are in fact linear functions of order statistics. If $\mathtt{T}$ is given by (5.39), then

$$\mathtt{T}(F_n) = \int xJ(F_n(x))dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} J\left(\tfrac{i}{n}\right) X_{(i)}, \tag{5.58}$$

since $F_n(X_{(i)}) = i/n$, $i = 1, ..., n$. If J is a smooth function, such as those given in Example 5.6 or those satisfying the conditions in Theorem 5.6, the corresponding L-estimator is often called a smooth L-estimator. Asymptotic properties of smooth L-estimators can be obtained using Theorem 5.6 and the results in §5.2.1. Results on L-estimators that are slightly different from that in (5.58) can be found in Serfling (1980, Chapter 8).

In §5.3.1, we consider another useful class of linear functions of order statistics, the sample quantiles described in the beginning of §5.2. In §5.3.2, we study robust linear functions of order statistics (in Hampel's sense) and their relative efficiencies w.r.t. $\bar{X}$, an efficient but nonrobust estimator. In §5.3.3, extensions to linear models are discussed.

5.3.1 Sample quantiles

Recall that $G^{-1}(p)$ is defined to be $\inf\{x : G(x) \geq p\}$ for any c.d.f. G on $\mathcal{R}$, where $p \in (0,1)$ is a fixed constant. For i.i.d. $X_1, ..., X_n$ from F, let $\theta_p = F^{-1}(p)$ and $\hat{\theta}_p = F_n^{-1}(p)$ denote the pth quantile of F and the pth

sample quantile, respectively. Then

$$\hat{\theta}_p = c_{np} X_{(m_p)} + (1 - c_{np}) X_{(m_p+1)}, \tag{5.59}$$

where m_p is the integer part of np, $c_{np} = 1$ if np is an integer, and $c_{np} = 0$ if np is not an integer. Thus, $\hat{\theta}_p$ is a linear function of order statistics.

Note that $F(\theta_p-) \le p \le F(\theta_p)$. If F is not flat in a neighborhood of θ_p, then $F(\theta_p - \epsilon) < p < F(\theta_p + \epsilon)$ for any $\epsilon > 0$.

Theorem 5.9. Let $X_1, ..., X_n$ be i.i.d. random variables from a c.d.f. F satisfying $F(\theta_p - \epsilon) < p < F(\theta_p + \epsilon)$ for any $\epsilon > 0$. Then, for every $\epsilon > 0$ and $n = 1, 2,...,$

$$P\big(|\hat{\theta}_p - \theta_p| > \epsilon\big) \le 2Ce^{-2n\delta_\epsilon^2}, \tag{5.60}$$

where δ_ϵ is the smaller of $F(\theta_p + \epsilon) - p$ and $p - F(\theta_p - \epsilon)$ and C is the same constant in Lemma 5.1(i).

Proof. Let $\epsilon > 0$ be fixed. Note that $G(x) \ge t$ if and only if $x \ge G^{-1}(t)$ for any c.d.f. G on $\mathcal{R}$ (exercise). Hence

$$\begin{aligned} P\big(\hat{\theta}_p > \theta_p + \epsilon\big) &= P\big(p > F_n(\theta_p + \epsilon)\big) \\ &= P\big(F(\theta_p + \epsilon) - F_n(\theta_p + \epsilon) > F(\theta_p + \epsilon) - p\big) \\ &\le P\big(\varrho_\infty(F_n, F) > \delta_\epsilon\big) \\ &\le Ce^{-2n\delta_\epsilon^2}, \end{aligned}$$

where the last inequality follows from DKW's inequality (Lemma 5.1(i)). Similarly,

$$P\big(\hat{\theta}_p < \theta_p - \epsilon\big) \le Ce^{-2n\delta_\epsilon^2}.$$

This proves (5.60). ∎

Result (5.60) implies that $\hat{\theta}_p$ is strongly consistent for θ_p (exercise) and that $\hat{\theta}_p$ is $\sqrt{n}$-consistent for θ_p if $F'(\theta_p-)$ and $F'(\theta_p+)$ (the left and right derivatives of F at θ_p) exist (exercise).

The exact distribution of $\hat{\theta}_p$ can be obtained as follows. Since $nF_n(t)$ has the binomial distribution $Bi(F(t), n)$ for any $t \in \mathcal{R}$,

$$\begin{aligned} P\big(\hat{\theta}_p \le t\big) &= P\big(F_n(t) \ge p\big) \\ &= \sum_{i=m_p}^{n} \binom{n}{i} [F(t)]^i [1 - F(t)]^{n-i}, \end{aligned} \tag{5.61}$$

where m_p is given in (5.59). If F has a Lebesgue p.d.f. f, then $\hat{\theta}_p$ has the Lebesgue p.d.f.

$$\varphi_n(t) = n\binom{n-1}{m_p - 1} [F(t)]^{m_p - 1} [1 - F(t)]^{n - m_p} f(t). \tag{5.62}$$

The following result provides an asymptotic distribution for $\sqrt{n}(\hat{\theta}_p - \theta_p)$.

Theorem 5.10. Let $X_1, ..., X_n$ be i.i.d. random variables from F.
(i) $P(\sqrt{n}(\hat{\theta}_p - \theta_p) \leq 0) \to \Phi(0) = \frac{1}{2}$, where Φ is the c.d.f. of the standard normal.
(ii) If F is continuous at θ_p and there exists $F'(\theta_p-) > 0$, then

$$P\big(\sqrt{n}(\hat{\theta}_p - \theta_p) \leq t\big) \to \Phi(t/\sigma_F^-), \qquad t < 0,$$

where $\sigma_F^- = \sqrt{p(1-p)}/F'(\theta_p-)$.
(iii) If F is continuous at θ_p and there exists $F'(\theta_p+) > 0$, then

$$P\big(\sqrt{n}(\hat{\theta}_p - \theta_p) \leq t\big) \to \Phi(t/\sigma_F^+), \qquad t > 0,$$

where $\sigma_F^+ = \sqrt{p(1-p)}/F'(\theta_p+)$.
(iv) If $F'(\theta_p)$ exists and is positive, then

$$\sqrt{n}(\hat{\theta}_p - \theta_p) \to_d N(0, \sigma_F^2), \tag{5.63}$$

where $\sigma_F = \sqrt{p(1-p)}/F'(\theta_p)$.
Proof. The proof of (i) is left as an exercise. Part (iv) is a direct consequence of (i)-(iii) and the proofs of (ii) and (iii) are similar. Thus, we only give a proof for (iii).

Let $t > 0$, $p_{nt} = F(\theta_p + t\sigma_F^+ n^{-1/2})$, $c_{nt} = \sqrt{n}(p_{nt} - p)/\sqrt{p_{nt}(1-p_{nt})}$, and $Z_{nt} = [B_n(p_{nt}) - np_{nt}]/\sqrt{np_{nt}(1-p_{nt})}$, where $B_n(q)$ denotes a random variable having the binomial distribution $Bi(q, n)$. Then

$$\begin{aligned} P\big(\hat{\theta}_p \leq \theta_p + t\sigma_F^+ n^{-1/2}\big) &= P\big(p \leq F_n(\theta_p + t\sigma_F^+ n^{-1/2})\big) \\ &= P\big(Z_{nt} \geq -c_{nt}\big). \end{aligned}$$

Under the assumed conditions on F, $p_{nt} \to p$ and $c_{nt} \to t$. Hence, the result follows from

$$P\big(Z_{nt} < -c_{nt}\big) - \Phi(-c_{nt}) \to 0.$$

But this follows from the CLT (Example 1.26) and Pólya's theorem (Proposition 1.16). ∎

If both $F'(\theta_p-)$ and $F'(\theta_p+)$ exist and are positive, but $F'(\theta_p-) \neq F'(\theta_p+)$, then the asymptotic distribution of $\sqrt{n}(\hat{\theta}_p - \theta_p)$ has the c.d.f. $\Phi(t/\sigma_F^-)I_{(-\infty,0)}(t) + \Phi(t/\sigma_F^+)I_{[0,\infty)}(t)$, a mixture of two normal distributions. An example of such a case when $p = \frac{1}{2}$ is

$$F(x) = xI_{[0,\frac{1}{2})}(x) + (2x - \tfrac{1}{2})I_{[\frac{1}{2},\frac{3}{4})}(x) + I_{[\frac{3}{4},\infty)}(x).$$

When $F'(\theta_p-) = F'(\theta_p+) = F'(\theta_p) > 0$, (5.63) shows that the asymptotic distribution of $\sqrt{n}(\hat{\theta}_p - \theta_p)$ is the same as that of $\sqrt{n}[F_n(\theta_p) - F(\theta_p)]/F'(\theta_p)$ (see (5.2)). The following result reveals a stronger relationship between sample quantiles and the empirical c.d.f.

Theorem 5.11 (Bahadur's representation). Let $X_1, ..., X_n$ be i.i.d. random variables from F. Suppose that $F'(\theta_p)$ exists and is positive. Then

$$\hat{\theta}_p = \theta_p + \frac{F(\theta_p) - F_n(\theta_p)}{F'(\theta_p)} + o_p\left(\tfrac{1}{\sqrt{n}}\right). \tag{5.64}$$

Proof. Let $t \in \mathcal{R}$, $\theta_{nt} = \theta_p + tn^{-1/2}$, $Z_n(t) = \sqrt{n}[F(\theta_{nt}) - F_n(\theta_{nt})]/F'(\theta_p)$, and $U_n(t) = \sqrt{n}[F(\theta_{nt}) - F_n(\hat{\theta}_p)]/F'(\theta_p)$. It can be shown (exercise) that

$$Z_n(t) - Z_n(0) = o_p(1). \tag{5.65}$$

Note that $|p - F_n(\hat{\theta}_p)| \le n^{-1}$. Then

$$\begin{aligned} U_n(t) &= \sqrt{n}[F(\theta_{nt}) - p + p - F_n(\hat{\theta}_p)]/F'(\theta_p) \\ &= \sqrt{n}[F(\theta_{nt}) - p]/F'(\theta_p) + O(n^{-1/2}) \\ &\to t. \end{aligned} \tag{5.66}$$

Let $\xi_n = \sqrt{n}(\hat{\theta}_p - \theta_p)$. Then, for any $t \in \mathcal{R}$ and $\epsilon > 0$,

$$\begin{aligned} P\big(\xi_n \le t, Z_n(0) \ge t + \epsilon\big) &= P\big(Z_n(t) \le U_n(t), Z_n(0) \ge t + \epsilon\big) \\ &\le P\big(|Z_n(t) - Z_n(0)| \ge \epsilon/2\big) \\ &\quad + P\big(|U_n(t) - t| \ge \epsilon/2\big) \\ &\to 0 \end{aligned} \tag{5.67}$$

by (5.65) and (5.66). Similarly,

$$P\big(\xi_n \ge t + \epsilon, Z_n(0) \le t\big) \to 0. \tag{5.68}$$

It follows from the result in Exercise 97 of §1.6 that

$$\xi_n - Z_n(0) = o_p(1),$$

which is the same as (5.64). ∎

If F has a positive Lebesgue p.d.f., then $\hat{\theta}_p$ viewed as a statistical functional (§5.2) is ϱ_∞-Hadamard differentiable at F (Fernholz, 1983) with the influence function

$$\phi_F(x) = [F(\theta_p) - I_{(-\infty,\theta_p]}(x)]/F'(\theta_p).$$

This implies result (5.64). Furthermore, $\hat{\theta}_p$ is robust in Hampel's sense; see also §5.3.2.

Corollary 5.1. Let $X_1, ..., X_n$ be i.i.d. random variables from F having positive derivatives at θ_{p_j}, where $0 < p_1 < \cdots < p_m < 1$ are fixed constants. Then

$$\sqrt{n}[(\hat{\theta}_{p_1}, ..., \hat{\theta}_{p_m}) - (\theta_{p_1}, ..., \theta_{p_m})] \to_d N_m(0, D),$$

where D is the $m \times m$ symmetric matrix whose (i, j)th element is

$$p_i(1 - p_j)/[F'(\theta_{p_i})F'(\theta_{p_j})], \qquad i < j. \quad \blacksquare$$

The proof of this corollary is left to the reader.

Example 5.9 (Interquartile range). One application of Corollary 5.1 is the derivation of the asymptotic distribution of the *interquartile range* $\hat{\theta}_{0.75} - \hat{\theta}_{0.25}$. The interquartile range is used as a measure of the variability among X_i's. It can be shown (exercise) that

$$\sqrt{n}[(\hat{\theta}_{0.75} - \hat{\theta}_{0.25}) - (\theta_{0.75} - \theta_{0.25})] \to_d N(0, \sigma_F^2)$$

with

$$\sigma_F^2 = \frac{3}{16[F'(\theta_{0.75})]^2} + \frac{3}{16[F'(\theta_{0.25})]^2} - \frac{1}{8F'(\theta_{0.75})F'(\theta_{0.25})}. \quad \blacksquare$$

There are some applications of using extreme order statistics such as $X_{(1)}$ and $X_{(n)}$. One example is given in Example 2.34. Some other examples and references can be found in Serfling (1980, pp. 89-91).

5.3.2 Robustness and efficiency

Let F be a c.d.f. on $\mathcal{R}$ symmetric about $\theta \in \mathcal{R}$ with $F'(\theta) > 0$. Then $\theta = \theta_{0.5}$ and is called the *median* of F. If F has a finite mean, then θ is also equal to the mean. In this section we consider the estimation of θ based on i.i.d. X_i's from F.

If F is normal, it has been shown in previous chapters that $\bar{X}$ is the UMVUE, MRIE, and MLE of θ, and is asymptotically efficient. On the other hand, if F is the c.d.f. of the Cauchy distribution $C(\theta, 1)$, it follows from Exercise 50 in §1.6 that $\bar{X}$ has the same distribution as X_1, i.e., $\bar{X}$ is as variable as X_1, and is inconsistent as an estimator of θ.

Why does $\bar{X}$ perform so differently? An important difference between the normal and Cauchy p.d.f.'s is that the former tends to 0 at the rate

$e^{-x^2/2}$ as $|x| \to \infty$, whereas the latter tends to 0 at the much slower rate x^{-2}, which results in $\int |x| dF(x) = \infty$. The poor performance of $\bar{X}$ in the Cauchy case is due to the high probability of getting extreme observations and the fact that $\bar{X}$ is sensitive to large changes in a few of the X_i's. (Note that $\bar{X}$ is not robust in Hampel's sense, since the functional $\int x dG(x)$ has an unbounded influence function at F.) This suggests the use of a robust estimator that discards some extreme observations. The *sample median*, which is defined to be the 50%th sample quantile $\hat{\theta}_{0.5}$ described in §5.3.1, is insensitive to the behavior of F as $|x| \to \infty$.

Since both the sample mean and the sample median can be used to estimate θ, a natural question is when one is better than the other, using a criterion such as the amse. Unfortunately, a general answer does not exist, since the asymptotic relative efficiency between these two estimators depends on the unknown distribution F. If F does not have a finite variance, then the sample median is certainly preferred since $\bar{X}$ is inconsistent but $\hat{\theta}_{0.5}$ is consistent and asymptotically normal as long as $F'(\theta) > 0$, and may still have a finite variance (Exercise 54). The following example, which compares the sample mean and median in some cases, shows that the sample median can be better even if $\mathrm{Var}(X_1) < \infty$.

Example 5.10. Suppose that $\mathrm{Var}(X_1) < \infty$. Then, by the CLT,

$$\sqrt{n}(\bar{X} - \theta) \to_d N(0, \mathrm{Var}(X_1)).$$

By Theorem 5.10(iv),

$$\sqrt{n}(\hat{\theta}_{0.5} - \theta) \to_d N(0, [2F'(\theta)]^{-2}).$$

Hence, the asymptotic relative efficiency of $\hat{\theta}_{0.5}$ w.r.t. $\bar{X}$ is

$$e(F) = 4[F'(\theta)]^2 \mathrm{Var}(X_1).$$

(i) If F is the c.d.f. of $N(\theta, \sigma^2)$, then $\mathrm{Var}(X_1) = \sigma^2$, $F'(\theta) = (\sqrt{2\pi}\sigma)^{-1}$, and $e(F) = 2/\pi = 0.637$.
(ii) If F is the c.d.f. of the logistic distribution $LG(\theta, \sigma)$, then $\mathrm{Var}(X_1) = \sigma^2\pi^2/3$, $F'(\theta) = (4\sigma)^{-1}$, and $e(F) = \pi^2/12 = 0.822$.
(iii) If $F(x) = F_0(x - \theta)$ and F_0 is the c.d.f. of the t-distribution t_ν with $\nu \geq 3$, then $\mathrm{Var}(X_1) = \nu/(\nu - 2)$, $F'(\theta) = \Gamma(\frac{\nu+1}{2})/[\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})]$, $e(F) = 1.62$ when $\nu = 3$, $e(F) = 1.12$ when $\nu = 4$, and $e(F) = 0.96$ when $\nu = 5$.
(iv) If F is the c.d.f. of the double exponential distribution $DE(\theta, \sigma)$, then $F'(\theta) = (2\sigma)^{-1}$ and $e(F) = 2$.
(v) Consider the Tukey model

$$F(x) = (1 - \epsilon)\Phi\left(\frac{x-\theta}{\sigma}\right) + \epsilon\Phi\left(\frac{x-\theta}{\tau\sigma}\right), \tag{5.69}$$

where $\sigma > 0$, $\tau > 0$ and $0 < \epsilon < 1$. Then $\mathrm{Var}(X_1) = (1-\epsilon)\sigma^2 + \epsilon\tau^2\sigma^2$, $F'(\theta) = (1-\epsilon+\epsilon/\tau)/(\sqrt{2\pi}\sigma)$, and $e(F) = 2(1-\epsilon+\epsilon\tau^2)(1-\epsilon+\epsilon/\tau)^2/\pi$. Note that $\lim_{\epsilon\to 0} e(F) = 2/\pi$ and $\lim_{\tau\to\infty} e(F) = \infty$. ∎

Since the sample median uses at most two actual values of x_i's, it may go too far in discarding observations, which results in a possible loss of efficiency. The trimmed sample mean introduced in Example 5.6(iii) is a natural compromise between the sample mean and median. Since F is symmetric, we consider $\beta = 1-\alpha$ in the trimmed mean, which results in the following L-estimator

$$\bar{X}_\alpha = \frac{1}{n-2m_\alpha}\sum_{j=m_\alpha+1}^{n-m_\alpha} X_{(j)}, \tag{5.70}$$

where m_α is the integer part of $n\alpha$ and $\alpha \in (0, \frac{1}{2})$. The estimator in (5.70) is called the α-trimmed sample mean. It discards the m_α smallest and m_α largest observations. The sample mean and median can be viewed as two extreme cases of $\bar{X}_\alpha$ as $\alpha \to 0$ and $\frac{1}{2}$, respectively.

It follows from Theorem 5.6 that if $F(x) = F_0(x-\theta)$, where F_0 is symmetric about 0 and has a Lebesgue p.d.f. positive in the range of X_1, then

$$\sqrt{n}(\bar{X}_\alpha - \theta) \to_d N(0, \sigma_\alpha^2), \tag{5.71}$$

where

$$\sigma_\alpha^2 = \frac{2}{(1-2\alpha)^2}\left[\int_0^{\theta_{1-\alpha}} x^2 dF_0(x) + \alpha\theta_{1-\alpha}^2\right].$$

Lehmann (1983, §5.4) provides various values of the asymptotic relative efficiency $e_{\bar{X}_\alpha,\bar{X}}(F) = \mathrm{Var}(X_1)/\sigma_\alpha^2$. For instance, when $F(x) = F_0(x-\theta)$ and F_0 is the c.d.f. of the t-distribution t_3, $e_{\bar{X}_\alpha,\bar{X}}(F) = 1.70$, 1.91, and 1.97 for $\alpha = 0.05$, 0.125, and 0.25, respectively; when F is given by (5.69) with $\tau = 3$ and $\epsilon = 0.05$, $e_{\bar{X}_\alpha,\bar{X}}(F) = 1.20$, 1.19, and 1.09 for $\alpha = 0.05$, 0.125, and 0.25, respectively; when F is given by (5.69) with $\tau = 3$ and $\epsilon = 0.01$, $e_{\bar{X}_\alpha,\bar{X}}(F) = 1.04$, 0.98, and 0.89 for $\alpha = 0.05$, 0.125, and 0.25, respectively.

Robustness and efficiency of other L-estimators can be discussed similarly. For an L-estimator $\mathtt{T}(F_n)$ with $\mathtt{T}$ given by (5.39), if the conditions in one of (i)-(iii) of Theorem 5.6 are satisfied, then (5.34) holds with

$$\sigma_F^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} J(F(x))J(F(y))[F(\min(x,y)) - F(x)F(y)]dxdy, \tag{5.72}$$

provided that $\sigma_F^2 < \infty$ (exercise). If F is symmetric about θ and J is symmetric about $\frac{1}{2}$, then $\mathtt{T}(F) = \theta$ (exercise) and, therefore, the asymptotic relative efficiency of $\mathtt{T}(F_n)$ w.r.t. $\bar{X}$ is $\mathrm{Var}(X_1)/\sigma_F^2$.

5.3.3 L-estimators in linear models

In this section we extend L-estimators to the following linear model:

$$X_i = \beta Z_i^\tau + \varepsilon_i, \qquad i = 1, ..., n, \tag{5.73}$$

with i.i.d. ε_i's having an unknown c.d.f. F_0 and a full rank $Z = (Z_1^\tau, ..., Z_n^\tau)^\tau$. Note that the c.d.f. of X_i is $F_0(x - \beta Z_i^\tau)$. Instead of assuming $E(\varepsilon_i) = 0$ (as we did in Chapter 3), we assume that

$$\int x J(F_0(x)) dF_0(x) = 0, \tag{5.74}$$

where J is a function on $[0, 1]$ (the same as that in (5.39)). Note that (5.74) may hold without any assumption on the existence of $E(\varepsilon_i)$. For instance, (5.74) holds if F_0 is symmetric about 0 and J is symmetric about $\frac{1}{2}$ (Exercise 61).

Since X_i's are not identically distributed, the use of the order statistics and the empirical c.d.f. based on $X_1, ..., X_n$ may not be appropriate. Instead, we consider the ordered values of residuals $r_i = X_i - \hat{\beta} Z_i^\tau$, $i = 1, ..., n$, and some empirical c.d.f.'s based on residuals, where $\hat{\beta} = XZ(Z^\tau Z)^{-1}$ is the LSE of β.

To illustrate the idea, let us start with the case where β and Z_i are univariate. First, assume that $Z_i \geq 0$ for all i. Let $\hat{F}_0$ be the c.d.f. putting mass $Z_i / \sum_{i=1}^n Z_i^2$ at r_i, $i = 1, ..., n$. An L-estimator of β is defined to be

$$\hat{\beta}_L = \hat{\beta} + \int x J(\hat{F}_0(x)) d\hat{F}_0(x) \sum_{i=1}^n Z_i \Big/ \sum_{i=1}^n Z_i^2. \tag{5.75}$$

When $J(t) = (1-2\alpha)^{-1} I_{(\alpha, 1-\alpha)}(t)$ with an $\alpha \in (0, \frac{1}{2})$, which corresponds to the α-trimmed sample mean in the i.i.d. case, $\hat{\beta}_L$ in (5.75) can be computed as follows. Order the residuals and trim off all observations corresponding to residuals $r_{(j)}$ with $w_j = \sum_{i=1}^j Z_{\delta_i} / \sum_{i=1}^n Z_i^2 \in [0, \alpha] \cup [1 - \alpha, 1]$, where $r_{(1)} \leq \cdots \leq r_{(n)}$ are ordered residuals and δ_i satisfies $r_{\delta_i} = r_{(i)}$. Then $\hat{\beta}_L$ is the LSE based on the remaining observations.

If some Z_i's are negative, we can define L-estimators as follows. Let $Z_i^+ = \max(Z_i, 0)$ and $Z_i^- = Z_i^+ - Z_i$. Let $\hat{F}_0^\pm$ be the c.d.f. putting mass $Z_i^\pm / \sum_{i=1}^n Z_i^2$ at r_i, $i = 1, ..., n$. An L-estimator of β is defined to be

$$\begin{aligned}\hat{\beta}_L = \hat{\beta} &+ \int x J(\hat{F}_0^+(x)) d\hat{F}_0^+(x) \sum_{i=1}^n Z_i^+ \Big/ \sum_{i=1}^n Z_i^2 \\ &- \int x J(\hat{F}_0^-(x)) d\hat{F}_0^-(x) \sum_{i=1}^n Z_i^- \Big/ \sum_{i=1}^n Z_i^2.\end{aligned}$$

For general p-vector Z_i, let z_{ij} be the jth component of Z_i, $j = 1, ..., p$. Let $z_{ij}^+ = \max(z_{ij}, 0)$, $z_{ij}^- = z_{ij}^+ - z_{ij}$, and $\hat{F}_{0i}^{\pm}$ be the c.d.f. putting mass $z_{ij}^{\pm}/\sum_{i=1}^n z_{ij}^2$ at r_i, $i = 1, ..., n$. Then an L-estimator of β is defined to be

$$\hat{\beta}_L = \hat{\beta} + (A^+ - A^-)(Z^\tau Z)^{-1}, \tag{5.76}$$

where

$$A^{\pm} = \left(\int xJ(\hat{F}_{01}^{\pm}(x))d\hat{F}_{01}^{\pm}(x) \sum_{i=1}^n z_{i1}^{\pm}, ..., \int xJ(\hat{F}_{0p}^{\pm}(x))d\hat{F}_{0p}^{\pm}(x) \sum_{i=1}^n z_{ip}^{\pm} \right).$$

Obviously, (5.76) reduces to (5.75) if $p = 1$ and $Z_i \geq 0$ for all i.

Theorem 5.12. Assume model (5.73) with i.i.d. ε_i's from a c.d.f. F_0 satisfying (5.74) for a given J. Suppose that F_0 has a uniformly continuous, positive, and bounded derivative on the range of ε_1. Suppose further that the conditions on Z_i's in Theorem 3.12 are satisfied.
(i) If the function J is continuous on (α, β) and equals 0 on $[0, \alpha] \cup [\beta, 1]$, where $0 < \alpha < \beta < 1$ are constants, then

$$\sigma_{F_0}^{-1}(\hat{\beta}_L - \beta)(Z^\tau Z)^{1/2} \to_d N_p(0, I_p), \tag{5.77}$$

where $\sigma_{F_0}^2$ is given by (5.72) with $F = F_0$.
(ii) Result (5.77) also holds if J' is bounded on $[0, 1]$, $E|\varepsilon_1| < \infty$, and $\sigma_{F_0}^2$ is finite. ∎

The proof of this theorem can be found in Bickel (1973). Robustness and efficiency comparisons between the LSE $\hat{\beta}$ and L-estimators $\hat{\beta}_L$ can be made in a similar way to those in §5.3.2.

5.4 Generalized Estimating Equations

The method of *generalized estimating equations* (GEE) is a powerful and general method of deriving point estimators, which includes many previously described methods as special cases. In §5.4.1, we begin with a description of this method and, to motivate the idea, we discuss its relationship with other methods that have been studied. Consistency and asymptotic normality of estimators derived from generalized estimating equations are studied in §5.4.2 and §5.4.3.

Throughout this section we assume that $X_1, ..., X_n$ are independent (not necessarily identically distributed) random vectors, where the dimension of X_i is d_i, $i = 1, ..., n$ ($\sup_i d_i < \infty$), and that we are interested in estimating θ, a k-vector of unknown parameters related to the unknown population.

5.4.1 The GEE method and its relationship with others

The sample mean and, more generally, the LSE in linear models are solutions of equations of the form

$$\sum_{i=1}^{n}(X_i - \gamma Z_i^\tau)Z_i = 0.$$

Also, MLE's (or RLE's) in §4.4 and, more generally, M-estimators in §5.2.2 are solutions of equations of the form

$$\sum_{i=1}^{n}\psi(X_i, \gamma) = 0.$$

This leads to the following general estimation method. Let $\Theta \subset \mathcal{R}^k$ be the range of θ, ψ_i be a Borel function from $\mathcal{R}^{d_i} \times \Theta$ to $\mathcal{R}^k$, $i = 1, ..., n$, and

$$s_n(\gamma) = \sum_{i=1}^{n}\psi_i(X_i, \gamma), \qquad \gamma \in \Theta. \tag{5.78}$$

If θ is estimated by $\hat{\theta} \in \Theta$ satisfying $s_n(\hat{\theta}) = 0$, then $\hat{\theta}$ is called a GEE estimator. The equation $s_n(\gamma) = 0$ is called a GEE. Apparently, the LSE's, RLE's, MQLE's, and M-estimators are special cases of GEE estimators.

Usually GEE's are chosen so that

$$E[s_n(\theta)] = \sum_{i=1}^{n}E[\psi_i(X_i, \theta)] = 0, \tag{5.79}$$

where the expectation E may be replaced by an asymptotic expectation defined in §2.5.2 if the exact expectation does not exist. If this is true, then $\hat{\theta}$ is motivated by the fact that $s_n(\hat{\theta}) = 0$ is a sample analogue of $E[s_n(\theta)] = 0$.

To motivate the idea, let us study the relationship between the GEE method and other methods that have been introduced.

M-estimators

The M-estimators defined in §5.2.2 for unvariate $\theta = \mathtt{T}(F)$ in the i.i.d. case are special cases of GEE estimators. Huber (1981) also considers regression M-estimators in the linear model (5.73). A regression M-estimator of β is defined as a solution to the GEE

$$\sum_{i=1}^{n}\psi(X_i - \gamma Z_i^\tau)Z_i = 0,$$

where ψ is one of the functions given in Example 5.7.

LSE's in linear and nonlinear regression models

Suppose that

$$X_i = f(Z_i, \theta) + \varepsilon_i, \qquad i = 1, ..., n, \tag{5.80}$$

where Z_i's are the same as those in (5.73), θ is an unknown k-vector of parameters, f is a known function, and ε_i's are independent random variables. Model (5.80) is the same as model (5.73) if f is linear in θ and is called a *nonlinear regression model* otherwise. Note that model (4.64) is a special case of model (5.80). The LSE under model (5.80) is any point in Θ minimizing $\sum_{i=1}^n [X_i - f(Z_i, \gamma)]^2$ over $\gamma \in \Theta$. If f is differentiable, then the LSE is a soluation to the GEE

$$\sum_{i=1}^n [X_i - f(Z_i, \gamma)] \frac{\partial f(Z_i, \gamma)}{\partial \gamma} = 0.$$

Quasi-likelihoods

This is a continuation of the discussion of the quasi-likelihoods introduced in §4.4.3. Assume first that X_i's are univariate ($d_i \equiv 1$). If X_i's follow a GLM, i.e., X_i has the p.d.f. in (4.55) and (4.57) holds, and if (4.58) holds, then the likelihood equation (4.59) can be written as

$$\sum_{i=1}^n \frac{x_i - \mu_i(\gamma)}{v_i(\gamma)} G_i(\gamma) = 0, \tag{5.81}$$

where $\mu_i(\gamma) = \mu(\psi(\gamma Z_i^\tau))$, $G_i(\gamma) = \partial \mu_i(\gamma)/\partial \gamma$, $v_i(\gamma) = \mathrm{Var}(X_i)/\phi$, and we have used the following fact:

$$\psi'(t) = (\mu^{-1})'(g^{-1}(t))(g^{-1})'(t) = [\zeta''(\psi(t))]^{-1}(g^{-1})'(t).$$

Equation (5.81) is a quasi-likelihood equation if either X_i does not have the p.d.f. in (4.55) or (4.58) does not hold. Note that this generalizes the discussion in §4.4.3: if X_i does not have the p.d.f. in (4.55), then the problem is often nonparametric. Let $s_n(\gamma)$ be the left-hand side of (5.81). Then $s_n(\gamma) = 0$ is a GEE and $E[s_n(\beta)] = 0$ is satisfied as long as the first condition in (4.56), $E(X_i) = \mu_i(\beta)$, is satisfied.

For general d_i's, let $X_i = (X_{i1}, ..., X_{id_i})$, $i = 1, ..., n$, where each X_{it} satisfies (4.56) and (4.57), i.e.,

$$E(X_{it}) = \mu(\eta_{it}) = g^{-1}(\beta Z_{it}^\tau) \quad \text{and} \quad \mathrm{Var}(X_{it}) = \phi_i \mu'(\eta_{it}),$$

and Z_{it}'s are k-vector values of covariates. In biostatistics and life-time testing problems, components of X_i are *repeated measurements* at different times from subject i and are called *longitudinal data*. Although X_i's are

assumed independent, X_{it}'s are likely to be dependent for each i. Let R_i be the $d_i \times d_i$ correlation matrix whose (t,l)th element is the correlation coefficient between X_{it} and X_{il}. Then

$$\mathrm{Var}(X_i) = \phi_i [D_i(\beta)]^{1/2} R_i [D_i(\beta)]^{1/2}, \tag{5.82}$$

where $D_i(\gamma)$ is the $d_i \times d_i$ diagonal matrix with the tth diagonal element $g^{-1}(\gamma Z_i^\tau)$. If R_i's in (5.82) are known, then an extension of (5.81) to the multivariate x_i's is

$$\sum_{i=1}^n [x_i - \mu_i(\gamma)]\{[D_i(\gamma)]^{1/2} R_i [D_i(\gamma)]^{1/2}\}^{-1} G_i(\gamma) = 0, \tag{5.83}$$

where $\mu_i(\gamma) = (\mu(\psi(\gamma Z_{i1}^\tau)), ..., \mu(\psi(\gamma Z_{in_i}^\tau)))$ and $G_i(\gamma) = \partial \mu_i(\gamma)/\partial \gamma$. In most applications, R_i is unknown and its form is hard to model. Let $\tilde{R}_i$ be a known correlation matrix (called a *working correlation matrix*). Replacing R_i in (5.83) by $\tilde{R}_i$ leads to the quasi-likelihood equation

$$\sum_{i=1}^n [x_i - \mu_i(\gamma)]\{[D_i(\gamma)]^{1/2} \tilde{R}_i [D_i(\gamma)]^{1/2}\}^{-1} G_i(\gamma) = 0. \tag{5.84}$$

For example, we may assume that the components of X_i are independent and take $\tilde{R}_i = I_{d_i}$. Although the working correlation matrix $\tilde{R}_i$ may not be the same as the true unknown correlation matrix R_i, an MQLE obtained from (5.84) is still consistent and asymptotically normal (§5.4.2 and §5.4.3). Of course, MQLE's are asymptotically more efficient if $\tilde{R}_i$ is closer to R_i. Even if $\tilde{R}_i = R_i$ and $\phi_i \equiv \phi$, (5.84) is still a quasi-likelihood equation, since the covariance matrix of X_i cannot determine the distribution of X_i unless X_i is normal.

Since an $\tilde{R}_i$ closer to R_i results in a better MQLE, sometimes it is suggested to replace $\tilde{R}_i$ in (5.84) by $\hat{R}_i$, an estimator of R_i (Liang and Zeger, 1986). The resulting equation is called a *pseudo-likelihood equation*. As long as $\max_{i \le n} \|\hat{R}_i - U_i\| \to_p 0$ as $n \to \infty$, where U_i's are correlation matrices (not necessarily the same as the true correlation matrices), MQLE's are consistent and asymptotically normal.

Profile empirical likelihoods

The previous discussion shows that the GEE method coincides with the method of deriving M-estimators, LSE's, MLE's, or MQLE's. The following discussion indicates that the GEE method is also closely related to the method of empirical likelihoods introduced in §5.1.2.

Assume that X_i's are i.i.d. from a c.d.f. F on $\mathcal{R}^d$ and $\psi_i = \psi$ for all i. Then condition (5.79) reduces to $E[\psi(X_1, \theta)] = 0$. This leads to the

consideration of the empirical likelihood

$$\ell(F) = \prod_{i=1}^{n} p_i \quad \text{subject to} \quad p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \psi(x_i, \theta) = 0,$$

where $p_i = P_F(\{x_i\})$, $i = 1, ..., n$. Maximizing this empirical likelihood is equivalent to maximizing

$$\ell(p_1, ..., p_n, \omega, \lambda, \theta) = \prod_{i=1}^{n} p_i + \omega \left(1 - \sum_{i=1}^{n} p_i\right) + \sum_{i=1}^{n} p_i \psi(x_i, \theta) \lambda^\tau,$$

where ω and λ are Lagrange multipliers.

Suppose that $\ell(\theta, \xi)$ is a likelihood (or empirical likelihood) function, where θ and ξ are not necessarily vector-valued. For example, $\xi = G$, a c.d.f. If maximizing $\ell(\theta, \xi)$ over (θ, ξ) is difficult, sometimes we can apply the method of *profile likelihoods*, which can be described as follows. For each fixed θ, let $\xi(\theta)$ satisfy

$$\ell(\theta, \xi(\theta)) = \sup_{\xi} \ell(\theta, \xi).$$

The function

$$\ell_P(\theta) = \ell(\theta, \xi(\theta))$$

is called a *profile likelihood* function for θ. Suppose that $\hat{\theta}_P$ maximizes $\ell_P(\theta)$. Then $\hat{\theta}_P$ is called a maximum profile likelihood estimator of θ. Note that $\hat{\theta}_P$ may be different from an MLE of θ.

In general, it is difficult to maximize the likelihood $\ell(p_1, ..., p_n, \omega, \lambda, \theta)$. We consider the following profile empirical likelihood. Let θ be fixed. It follows from (5.12) and (5.13) that

$$\omega = n, \qquad \tilde{p}_i = \{n[1 + \psi(x_i, \theta)\lambda_n^\tau]\}^{-1}$$

with a $\lambda_n = \lambda_n(\theta)$ satisfying

$$\sum_{i=1}^{n} \frac{\psi(x_i, \theta)}{n\{1 + \psi(x_i, \theta)[\lambda_n(\theta)]^\tau\}} = 0 \tag{5.85}$$

maximize $\ell(p_1, ...p_n, \omega, \lambda, \theta)$ for any fixed θ. Substituting $\tilde{p}_i$ with $\sum_{i=1}^{n} \tilde{p}_i = 1$ into $\ell(p_1, ...p_n, \omega, \lambda, \theta)$ leads to the following profile empirical likelihood for θ:

$$\ell_P(\theta) = \prod_{i=1}^{n} \frac{1}{n\{1 + \psi(x_i, \theta)[\lambda_n(\theta)]^\tau\}}. \tag{5.86}$$

Let $\hat{\theta}$ be a maximum of $\ell_P(\theta)$ in (5.86). It follows from the proof of Theorem 5.4 that

$$\lambda_n(\hat{\theta}) = O_p(n^{-1/2}) \qquad \text{and} \qquad \hat{\theta} - \theta = O_p(n^{-1/2})$$

(see also Qin and Lawless (1994)), under some conditions on ψ and its derivatives. Using (5.85), we obtain that

$$0 = \sum_{i=1}^{n} \frac{\psi(x_i, \hat{\theta})}{1 + \psi(x_i, \hat{\theta})[\lambda_n(\hat{\theta})]^\tau} = \left[1 + O_p(n^{-1/2})\right] \sum_{i=1}^{n} \psi(x_i, \hat{\theta})$$

and, therefore,

$$P\left(\sum_{i=1}^{n} \psi(X_i, \hat{\theta}) = 0\right) \to 1.$$

That is, $\hat{\theta}$ is a GEE estimator.

5.4.2 Consistency of GEE estimators

We now study under what conditions (besides (5.79)) GEE estimators are consistent. For each n, let $\hat{\theta}_n$ be a GEE estimator, i.e., $s_n(\hat{\theta}_n) = 0$, where $s_n(\gamma)$ is defined by (5.78).

First, Theorem 5.7 and its proof can be extended to multivariate $\mathtt{T}$ in a straightforward manner. Hence, we have the following result.

Proposition 5.2. Suppose that $X_1, ..., X_n$ are i.i.d. from F and $\psi_i \equiv \psi$, a bounded and continuous function from $\mathcal{R}^d \times \Theta$ to $\mathcal{R}^k$. Let $\Psi(t) = \int \psi(x,t)dF(x)$. Suppose that $\Psi(\theta) = 0$ and $\partial\Psi(t)/\partial t$ exists and is of full rank at $t = \theta$. Then $\hat{\theta}_n \to_p \theta$. ∎

For unbounded ψ in the i.i.d. case, the following result and its proof can be found in Qin and Lawless (1994).

Proposition 5.3. Suppose that $X_1, ..., X_n$ are i.i.d. from F and $\psi_i \equiv \psi$. Assume that $\varphi(x,\gamma) = \partial\psi(x,\gamma)/\partial\gamma$ exists in N_θ, a neighborhood of θ, and is continuous at θ; there is a function $h(x)$ such that $\sup_{\gamma \in N_\theta} \|\varphi(x,\gamma)\| \leq h(x)$, $\sup_{\gamma \in N_\theta} \|\psi(x,\gamma)\|^3 \leq h(x)$, and $E[h(X_1)] < \infty$; $E[\varphi(X_1,\theta)]$ is of full rank; $E\{[\psi(X_1,\theta)]^\tau \psi(X_1,\theta)\}$ is positive definite; and (5.79) holds. Then, there exists a sequence of random vectors $\{\hat{\theta}_n\}$ such that

$$P\left(s_n(\hat{\theta}_n) = 0\right) \to 1 \qquad \text{and} \qquad \hat{\theta}_n \to_p \theta. \quad ∎ \tag{5.87}$$

Next, we consider non-i.i.d. X_i's.

Proposition 5.4. Suppose that $X_1, ..., X_n$ are independent and θ is univariate. Assume that $\psi_i(x, \gamma)$ is real-valued and nonincreasing in γ for all i; there is a $\delta > 0$ such that $\sup_i E|\psi_i(X_i, \gamma)|^{1+\delta} < \infty$ for any γ in N_θ, a neighborhood of θ (this condition can be replaced by $E|\psi(X_1, \gamma)| < \infty$ for any γ in N_θ when X_i's are i.i.d. and $\psi_i \equiv \psi$); $\psi_i(x, \gamma)$ are continuous in N_θ; (5.79) holds; and

$$\limsup_n E[\Psi_n(\theta + \epsilon)] < 0 < \liminf_n E[\Psi_n(\theta - \epsilon)] \tag{5.88}$$

for any $\epsilon > 0$, where $\Psi_n(\gamma) = n^{-1}s_n(\gamma)$. Then, there exists a sequence of random variables $\{\hat{\theta}_n\}$ such that (5.87) holds. Furthermore, any sequence $\{\hat{\theta}_n\}$ satisfying $s_n(\hat{\theta}_n) = 0$ satisfies (5.87).

Proof. Since ψ_i's are nonincreasing, the functions $\Psi_n(\gamma)$ and $E[\Psi_n(\gamma)]$ are nonincreasing. Let $\epsilon > 0$ be fixed so that $\theta \pm \epsilon \in N_\theta$. Under the assumed conditions,

$$\Psi_n(\theta \pm \epsilon) - E[\Psi_n(\theta \pm \epsilon)] \to_p 0$$

(Theorem 1.14(ii)). By condition (5.88),

$$P\big(\Psi_n(\theta + \epsilon) < 0 < \Psi_n(\theta - \epsilon)\big) \to 1.$$

The rest of the proof is left as an exercise. ∎

To establish the next result, we need the following lemma. First, we need the following concept. A sequence of functions $\{g_i\}$ from $\mathcal{R}^k$ to $\mathcal{R}^k$ is called *equicontinuous* on an open set $\mathcal{O} \subset \mathcal{R}^k$ if and only if for any $\epsilon > 0$, there is a $\delta_\epsilon > 0$ such that $\sup_i \|g_i(t) - g_i(s)\| < \epsilon$ whenever $t \in \mathcal{O}$, $s \in \mathcal{O}$, and $\|t - s\| < \delta_\epsilon$. Since a continuous function on a compact set is uniformly continuous, functions such as $g_i(\gamma) = g(t_i, \gamma)$ form an equicontinuous sequence on $\mathcal{O}$ if t_i's vary in a compact set containing $\mathcal{O}$ and $g(t, \gamma)$ is a continuous function in (t, γ).

Lemma 5.3. Suppose that Θ is a compact subset of $\mathcal{R}^k$. Let $h_i(X_i) = \sup_{\gamma \in \Theta} \|\psi_i(X_i, \gamma)\|$, $i = 1, 2, ...$ Suppose that $\sup_i E|h_i(X_i)|^{1+\delta} < \infty$ and $\sup_i E\|X_i\|^\delta < \infty$ for some $\delta > 0$ (this condition can be replaced by $E|h(X_1)| < \infty$ when X_i's are i.i.d. and $\psi_i \equiv \psi$). Suppose further that for any $c > 0$ and sequence $\{x_i\}$ satisfying $\|x_i\| \le c$, the sequence of functions $\{g_i(\gamma) = \psi_i(x_i, \gamma)\}$ is equicontinuous on any open subset of Θ. Then

$$\sup_{\gamma \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^n \{\psi_i(X_i, \gamma) - E[\psi_i(X_i, \gamma)]\} \right\| \to_p 0.$$

Proof. Since we only need to consider components of ψ_i's, without loss of generality we can assume that ψ_i's are functions from $\mathcal{R}^{d_i} \times \Theta$ to $\mathcal{R}$. For

any $c > 0$,

$$\sup_n E\left[\frac{1}{n}\sum_{i=1}^n h_i(X_i)I_{(c,\infty)}(\|X_i\|)\right] \le \sup_i E[h_i(X_i)I_{(c,\infty)}(\|X_i\|)].$$

Let $c_0 = \sup_i E|h_i(X_i)|^{1+\delta}$ and $c_1 = \sup_i E\|X_i\|^\delta$. By Hölder's inequality,

$$\begin{aligned} E[h_i(X_i)I_{(c,\infty)}(\|X_i\|)] &\le \left[E|h_i(X_i)|^{1+\delta}\right]^{1/(1+\delta)}\left[P(\|X_i\| > c)\right]^{\delta/(1+\delta)} \\ &\le c_0^{1/(1+\delta)}c_1^{\delta/(1+\delta)}c^{-\delta^2/(1+\delta)} \end{aligned}$$

for all i. For $\epsilon > 0$ and $\tilde{\epsilon} > 0$, choose a c such that $c_0^{1/(1+\delta)}c_1^{\delta/(1+\delta)}c^{-\delta^2/(1+\delta)} < \epsilon\tilde{\epsilon}/2$. Then, for any $\mathcal{O} \subset \Theta$, the probability

$$P\left(\frac{1}{n}\sum_{i=1}^n\left\{\sup_{\gamma\in\mathcal{O}}\psi_i(X_i,\gamma) - \inf_{\gamma\in\mathcal{O}}\psi_i(X_i,\gamma)\right\}I_{(c,\infty)}(\|X_i\|) > \frac{\epsilon}{2}\right) \tag{5.89}$$

is bounded by $\tilde{\epsilon}$ (exercise). From the equicontinuity of $\{\psi_i(x_i,\gamma)\}$, there is a $\delta_\epsilon > 0$ such that

$$\frac{1}{n}\sum_{i=1}^n\left\{\sup_{\gamma\in\mathcal{O}_\epsilon}\psi_i(X_i,\gamma) - \inf_{\gamma\in\mathcal{O}_\epsilon}\psi_i(X_i,\gamma)\right\}I_{[0,c]}(\|X_i\|) < \frac{\epsilon}{2}$$

for sufficiently large n, where $\mathcal{O}_\epsilon$ denotes any open ball in $\mathcal{R}^k$ with radius less than δ_ϵ. These results, together with Theorem 1.14(ii) and the fact that $\|\psi_i(X_i,\gamma)\| \le h_i(X_i)$, imply that

$$P\left(\frac{1}{n}\sum_{i=1}^n\left\{\sup_{\gamma\in\mathcal{O}_\epsilon}\psi_i(X_i,\gamma) - E\left[\inf_{\gamma\in\mathcal{O}_\epsilon}\psi_i(X_i,\gamma)\right]\right\} > \epsilon\right) \to 0. \tag{5.90}$$

Let $H_n(\gamma) = n^{-1}\sum_{i=1}^n\{\psi_i(X_i,\gamma) - E[\psi_i(X_i,\gamma)]\}$. Then

$$\sup_{\gamma\in\mathcal{O}_\epsilon} H_n(\gamma) \le \frac{1}{n}\sum_{i=1}^n\left\{\sup_{\gamma\in\mathcal{O}_\epsilon}\psi_i(X_i,\gamma) - E\left[\inf_{\gamma\in\mathcal{O}_\epsilon}\psi_i(X_i,\gamma)\right]\right\},$$

which with (5.90) implies that

$$P\big(H_n(\gamma) > \epsilon \text{ for all } \gamma \in \mathcal{O}_\epsilon\big) = P\left(\sup_{\gamma\in\mathcal{O}_\epsilon} H_n(\gamma) > \epsilon\right) \to 0.$$

Similarly we can show that

$$P\big(H_n(\gamma) < -\epsilon \text{ for all } \gamma \in \mathcal{O}_\epsilon\big) \to 0.$$

Since Θ is compact, there exists m_ϵ open balls $\mathcal{O}_{\epsilon,j}$ such that $\Theta \subset \cup \mathcal{O}_{\epsilon,j}$. Then, the result follows from

$$P\left(\sup_{\gamma\in\Theta}|H_n(\gamma)| > \epsilon\right) \leq \sum_{j=1}^{m_\epsilon} P\left(\sup_{\gamma\in\mathcal{O}_{\epsilon,j}}|H_n(\gamma)| > \epsilon\right) \to 0. \quad \blacksquare$$

Example 5.11. Consider the quasi-likelihood equation (5.84). Let $\{\tilde{R}_i\}$ be a sequence of working correlation matrices and

$$\psi_i(x_i,\gamma) = [x_i - \mu_i(\gamma)]\{[D_i(\gamma)]^{1/2}\tilde{R}_i[D_i(\gamma)]^{1/2}\}^{-1}G_i(\gamma). \tag{5.91}$$

It can be shown (exercise) that ψ_i's satisfy the conditions of Lemma 5.3 if Θ is compact and $\sup_i \|Z_i\| < \infty$. $\blacksquare$

Proposition 5.5. Assume (5.79) and the conditions in Lemma 5.3. Suppose that the functions $\Delta_n(\gamma) = E[n^{-1}s_n(\gamma)]$ have the property that $\lim_{n\to\infty}\Delta_n(\gamma) = 0$ if and only if $\gamma = \theta$. (If Δ_n converges to a function Δ, then this condition and (5.79) means that Δ has a unique 0 at θ.) Suppose that $\{\hat{\theta}_n\}$ is a sequence of GEE estimators and that $\hat{\theta}_n = O_p(1)$. Then $\hat{\theta}_n \to_p \theta$.

Proof. First, assume that Θ is a compact subset of $\mathcal{R}^k$. From Lemma 5.3 and $s_n(\hat{\theta}_n) = 0$, $\Delta_n(\hat{\theta}_n) \to_p 0$. By Theorem 1.8(vi), there is a subsequence $\{n_i\}$ such that

$$\Delta_{n_i}(\hat{\theta}_{n_i}) \to_{a.s.} 0. \tag{5.92}$$

Let $x_1, x_2, ...$ be a fixed sequence such that (5.92) holds and let θ_0 be a limit point of $\{\hat{\theta}_n\}$. Since Θ is compact, $\theta_0 \in \Theta$ and there is a subsequence $\{m_j\} \subset \{n_i\}$ such that $\hat{\theta}_{m_j} \to \theta_0$. Using the argument in the proof of Lemma 5.3, it can be shown (exercise) that $\{\Delta_n(\gamma)\}$ is equicontinuous on any open subset of Θ. Then

$$\Delta_{m_j}(\hat{\theta}_{m_j}) - \Delta_{m_j}(\theta_0) \to 0,$$

which with (5.92) implies $\Delta_{m_j}(\theta_0) \to 0$. Under the assumed condition, $\theta_0 = \theta$. Since this is true for any limit point of $\{\hat{\theta}_n\}$, $\hat{\theta}_n \to_p \theta$.

Next, consider a general Θ. For any $\epsilon > 0$, there is an $M_\epsilon > 0$ such that $P(\|\hat{\theta}_n\| \leq M_\epsilon) > 1 - \epsilon$. The result follows from the previous proof by considering the closure of $\Theta \cap \{\gamma : \|\gamma\| \leq M_\epsilon\}$ as the parameter space. $\blacksquare$

Condition $\hat{\theta}_n = O_p(1)$ in Proposition 5.5 is obviously necessary for the consistency of $\hat{\theta}_n$. It has to be checked in any particular problem.

If a GEE is a likelihood equation under some conditions, then we can often show, using a similar argument to the proof of Theorem 4.17 or 4.18, that there exists a consistent sequence of GEE estimators.

Proposition 5.6. Suppose that $s_n(\gamma) = \partial \log \ell_n(\gamma)/\partial\gamma$ for some function ℓ_n; $I_n(\theta) = \mathrm{Var}(s_n(\theta)) \to 0$; $\varphi_i(x,\gamma) = \partial\psi_i(x,\gamma)/\partial\gamma$ exists and the sequence of functions $\{\varphi_{ij}, i = 1, 2, ...\}$ satisfies the conditions in Lemma 5.3 with Θ replaced by a compact neighborhood of θ, where φ_{ij} is the jth row of φ_i, $j = 1, ..., k$; and $-\liminf_n [I_n(\theta)]^{1/2} E[\nabla s_n(\theta)][I_n(\theta)]^{1/2}$ is positive definite, where $\nabla s_n(\gamma) = \partial s_n(\gamma)/\partial\gamma$. Then, there exists a sequence of estimators $\{\hat\theta_n\}$ satisfying (5.87). ∎

The proof of Proposition 5.6 is similar to that of Theorem 4.17 or Theorem 4.18 and is left as an exercise.

Example 5.12. Consider the quasi-likelihood equation (5.84) with $\tilde R_i = I_{d_i}$ for all i. Then the GEE is a likelihood equation under a GLM (§4.4.2) assumption. It can be shown (exercise) that the conditions of Proposition 5.6 are satisfied if $\sup_i \|Z_i\| < \infty$. ∎

5.4.3 Asymptotic normality of GEE estimators

Asymptotic normality of a consistent sequence of GEE estimators can be established under some conditions. We first consider the special case where θ is univariate and $X_1, ..., X_n$ are i.i.d.

Theorem 5.13. Let $X_1, ..., X_n$ be i.i.d. from F, $\psi_i \equiv \psi$, and $\theta \in \mathcal{R}$. Suppose that $\Psi(\gamma) = \int \psi(x,\gamma)dF(x) = 0$ if and only if $\gamma = \theta$, $\Psi'(\theta)$ exists and $\Psi'(\theta) \neq 0$.
(i) Assume that $\psi(x,\gamma)$ is nonincreasing in γ and that $\int [\psi(x,\gamma)]^2 dF(x)$ is finite for γ in a neighborhood of θ and is continuous at θ. Then, any sequence of GEE estimators (M-estimators) $\{\hat\theta_n\}$ satisfies

$$\sqrt{n}(\hat\theta_n - \theta) \to_d N(0, \sigma_F^2), \tag{5.93}$$

where

$$\sigma_F^2 = \int [\psi(x,\theta)]^2 dF(x)/[\Psi'(\theta)]^2.$$

(ii) Assume that $\int [\psi(x,\theta)]^2 dF(x) < \infty$, $\psi(x,\gamma)$ is continuous in x, and $\lim_{\gamma\to\theta} \|\psi(\cdot,\gamma) - \psi(\cdot,\theta)\|_V = 0$, where $\|\cdot\|_V$ is the variation norm defined in Lemma 5.2. Then, any consistent sequence of GEE estimators $\{\hat\theta_n\}$ satisfies (5.93).
Proof. (i) Let $\Psi_n(\gamma) = n^{-1} s_n(\gamma)$. Since Ψ_n is nonincreasing,

$$P(\Psi_n(t) < 0) \le P(\hat\theta_n \le t) \le P(\Psi_n(t) \le 0)$$

for any $t \in \mathcal{R}$. Then, (5.93) follows from

$$\lim_{n\to\infty} P\big(\Psi_n(t_n) < 0\big) = \lim_{n\to\infty} P\big(\Psi_n(t_n) \le 0\big) = \Phi(t)$$

for all $t \in \mathcal{R}$, where $t_n = \theta + t\sigma_F n^{-1/2}$. Let $s_{t,n}^2 = \mathrm{Var}(\psi(X_1, t_n))$ and $Y_{ni} = [\psi(X_i, t_n) - \Psi(t_n)]/s_{t,n}$. Then, it suffices to show that

$$\lim_{n\to\infty} P\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_{ni} \le -\frac{\sqrt{n}\Psi(t_n)}{s_{t,n}}\right) = \Phi(t)$$

for all t. Under the assumed conditions, $\sqrt{n}\Psi(t_n) \to \Psi'(\theta)t\sigma_F$ and $s_{t,n} \to -\Psi'(\theta)\sigma_F$. Hence, it suffices to show that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_{ni} \to_d N(0,1).$$

Note that $Y_{n1}, ..., Y_{nn}$ are i.i.d. random variables. Hence we can apply Lindeberg's CLT (Theorem 1.15). In this case, Lindeberg's condition (1.53) is implied by

$$\lim_{n\to\infty} \int_{|\psi(x,t_n)|>\sqrt{n}\epsilon} [\psi(x,t_n)]^2 dF(x) = 0$$

for any $\epsilon > 0$. For any $\eta > 0$, $\psi(x, \theta+\eta) \le \psi(x, t_n) \le \psi(x, \theta-\eta)$ for all x and sufficiently large n. Let $u(x) = \max\{|\psi(x, \theta-\eta)|, |\psi(x, \theta+\eta)|\}$. Then

$$\int_{|\psi(x,t_n)|>\sqrt{n}\epsilon} [\psi(x,t_n)]^2 dF(x) \le \int_{u(x)>\sqrt{n}\epsilon} [u(x)]^2 dF(x),$$

which converges to 0 since $\int [\psi(x,\gamma)]^2 dF(x)$ is finite for γ in a neighborhood of θ. This proves (i).

(ii) Let $\phi_F(x) = -\psi(x,\theta)/\Psi'(\theta)$. Following the proof of Theorem 5.7, we have

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_F(X_i) + R_{1n} - R_{2n},$$

where

$$R_{1n} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(X_i,\theta)\left[\frac{1}{\Psi'(\theta)} - \frac{1}{h_F(\hat{\theta}_n)}\right],$$

$$R_{2n} = \frac{\sqrt{n}}{h_F(\hat{\theta}_n)}\int [\psi(x,\hat{\theta}_n) - \psi(x,\theta)] d(F_n - F)(x),$$

and h_F is defined in the proof of Theorem 5.7 with $\Psi = \lambda_F$. By the CLT and the consistency of $\hat{\theta}_n$, $R_{1n} = o_p(1)$. Hence the result follows if we can show that $R_{2n} = o_p(1)$. By Lemma 5.2,

$$|R_{2n}| \le \sqrt{n}|h_F(\hat{\theta}_n)|^{-1}\varrho_\infty(F_n, F)\|\psi(\cdot,\hat{\theta}_n) - \psi(\cdot,\theta)\|_V.$$

The result follows from the assumed condition on ψ and the fact that $\sqrt{n}\varrho_\infty(F_n, F) = O_p(1)$ (Theorem 5.1). ∎

Note that the result in Theorem 5.13 coincides with the result in Theorem 5.7 and (5.34).

Example 5.13. Consider the M-estimators given in Example 5.7 based on i.i.d. random variables $X_1, ..., X_n$. If ψ is bounded and continuous, then Theorem 5.7 applies and (5.93) holds. For case (ii), $\psi(x, \gamma)$ is not bounded but is nondecreasing in γ ($-\psi(x, \gamma)$ is nonincreasing in γ). Hence Theorem 5.13 can be applied to this case.

Consider Huber's ψ given in (v). Assume that F is differentiable in neighborhoods of $\theta - C$ and $\theta + C$. Then

$$\Psi(\gamma) = \int_{\gamma-C}^{\gamma+C} (\gamma - x)dF(x) - CF(\gamma - C) + C[1 - F(\gamma + C)]$$

is differentiable at θ (exercise); $\Psi(\theta) = 0$ if F is symmetric about θ (exercise); and

$$\int [\psi(x, \gamma)]^2 dF(x) = \int_{\gamma-C}^{\gamma+C} (\gamma - x)^2 dF(x) + C^2 F(\gamma - C) + C^2[1 - F(\gamma + C)]$$

is continuous at θ (exercise). Therefore, (5.93) holds with

$$\sigma_F^2 = \frac{\int_{\theta-C}^{\theta+C} (\theta - x)^2 dF(x) + C^2 F(\theta - C) + C^2[1 - F(\theta + C)]}{[F(\theta + C) - F(\theta - C)]^2}$$

(exercise). Note that Huber's M-estimator is robust in Hampel's sense. Asymptotic relative efficiency of $\hat{\theta}_n$ w.r.t. the sample mean $\bar{X}$ can be obtained (exercise). ∎

The next result is for general θ and independent X_i's.

Theorem 5.14. Suppose that $\varphi_i(x, \gamma) = \partial\psi_i(x, \gamma)/\partial\gamma$ exists and the sequence of functions $\{\varphi_{ij}, i = 1, 2, ...\}$ satisfies the conditions in Lemma 5.3 with Θ replaced by a compact neighborhood of θ, where φ_{ij} is the jth row of φ_i; $\sup_i E\|\psi_i(X_i, \theta)\|^{2+\delta} < \infty$ for some $\delta > 0$ (this condition can be replaced by $E\|\psi(X_1, \theta)\|^2 < \infty$ if X_i's are i.i.d. and $\psi_i \equiv \psi$); $E[\psi_i(X_i, \theta)] = 0$; $\liminf_n \lambda_-[n^{-1}\mathrm{Var}(s_n(\theta))] > 0$ and $\liminf_n \lambda_-[n^{-1}M_n(\theta)] > 0$, where $M_n(\theta) = -E[\nabla s_n(\theta)]$ and $\lambda_-[A]$ is the smallest eigenvalue of the matrix A. If $\{\hat{\theta}_n\}$ is a consistent sequence of GEE estimators, then

$$(\hat{\theta}_n - \theta)V_n^{-1/2} \to_d N_k(0, I_k), \tag{5.94}$$

where

$$V_n = [M_n(\theta)]^{-1}\mathrm{Var}(s_n(\theta))[M_n(\theta)]^{-1}. \tag{5.95}$$

Proof. The proof is similar to that of Theorem 4.17. By the consistency of $\hat{\theta}_n$, we can focus on the event $\{\hat{\theta}_n \in A_\epsilon\}$, where $A_\epsilon = \{\gamma : \|\gamma - \theta\| \le \epsilon\}$ with a given $\epsilon > 0$. For sufficiently small ϵ, it can be shown (exercise) that

$$\max_{\gamma \in A_\epsilon} \frac{\|\nabla s_n(\gamma) - \nabla s_n(\theta)\|}{n} = o_p(1), \tag{5.96}$$

using a similar argument to the proof of Lemma 5.3. From the mean-value theorem and $s_n(\hat{\theta}_n) = 0$,

$$-s_n(\theta) = (\hat{\theta}_n - \theta) \int_0^1 \nabla s_n\big(\theta + t(\hat{\theta}_n - \theta)\big) dt.$$

It follows from (5.96) that

$$\frac{1}{n} \left\| \int_0^1 \nabla s_n\big(\theta + t(\hat{\theta}_n - \theta)\big) dt - \nabla s_n(\theta) \right\| = o_p(1).$$

Also, by Theorem 1.14(ii),

$$n^{-1}[\nabla s_n(\theta) + M_n(\theta)] = o_p(1).$$

This and $\liminf_n \lambda_-[n^{-1} M_n(\theta)] > 0$ imply

$$s_n(\theta)[M_n(\theta)]^{-1} = (\hat{\theta}_n - \theta)[1 + o_p(1)].$$

The result follows if we can show that

$$s_n(\theta)[M_n(\theta)]^{-1} V_n^{-1/2} \to_d N_k(0, I_k). \tag{5.97}$$

For any nonzero $l \in \mathcal{R}^k$,

$$\frac{1}{(l V_n l^\tau)^{1+\delta/2}} \sum_{i=1}^n E|\psi_i(X_i, \theta)[M_n(\theta)]^{-1} l^\tau|^{2+\delta} \to 0, \tag{5.98}$$

since $\liminf_n \lambda_-[n^{-1}\mathrm{Var}(s_n(\theta))] > 0$ and $\sup_i E\|\psi_i(X_i, \theta)\|^{2+\delta} < \infty$ (exercise). Applying the CLT (Theorem 1.15) with Liapunov's condition (5.98), we obtain that

$$s_n(\theta)[M_n(\theta)]^{-1} l^\tau / \sqrt{l V_n l^\tau} \to_d N(0, 1) \tag{5.99}$$

for any l, which implies (5.97) (exercise). ∎

Asymptotic normality of GEE estimators can be established under various other conditions; see, for example, Serfling (1980, Chapter 7) and He and Shao (1996).

If X_i's are i.i.d. and $\psi_i \equiv \psi$, the asymptotic covariance matrix in (5.95) reduces to

$$V_n = n^{-1}\{E[\varphi(X_1,\theta)]\}^{-1}E\{[\psi(X_1,\theta)]^\tau[\psi(X_1,\theta)]\}\{E[\varphi(X_1,\theta)]\}^{-1},$$

where $\varphi(x,\gamma) = \partial\psi(x,\gamma)/\partial\gamma$. When θ is univariate, V_n further reduces to

$$V_n = n^{-1}E[\psi(X_1,\theta)]^2/\{E[\varphi(X_1,\theta)]\}^2.$$

Under the conditions of Theorem 5.14,

$$E[\varphi(X_1,\theta)] = \int \frac{\partial\psi(x,\theta)}{\partial\theta}dF(x) = \frac{\partial}{\partial\theta}\int \psi(x,\theta)dF(x).$$

Hence, the result in Theorem 5.14 coincides with that in Theorem 5.13.

Example 5.14. Consider the quasi-likelihood equation in (5.84) and ψ_i in (5.91). If $\sup_i \|Z_i\| < \infty$, then ψ_i satisfies the conditions in Theorem 5.14 (exercise). Let $\tilde{V}_n(\gamma) = [D_i(\gamma)]^{1/2}\tilde{R}_i[D_i(\gamma)]^{1/2}$. Then

$$\mathrm{Var}(s_n(\theta)) = \sum_{i=1}^n [G_i(\theta)]^\tau[\tilde{V}_n(\theta)]^{-1}\mathrm{Var}(X_i)[\tilde{V}_n(\theta)]^{-1}G_i(\theta)$$

and

$$M_n(\theta) = \sum_{i=1}^n [G_i(\theta)]^\tau[\tilde{V}_n(\theta)]^{-1}G_i(\theta).$$

If $\tilde{R}_i = R_i$ (the true correlation matrix) for all i, then

$$\mathrm{Var}(s_n(\theta)) = \sum_{i=1}^n \phi_i[G_i(\theta)]^\tau[\tilde{V}_n(\theta)]^{-1}G_i(\theta).$$

If, in addition, $\phi_i \equiv \phi$, then

$$V_n = [M_n(\theta)]^{-1}\mathrm{Var}(s_n(\theta))[M_n(\theta)]^{-1} = \phi[M_n(\theta)]^{-1}. \quad \blacksquare$$

5.5 Variance Estimation

In statistical inference the accuracy of a point estimator is usually assessed by its mse or amse. If the bias or asymptotic bias of an estimator is (asymptotically) negligible w.r.t. its mse or amse, then assessing the mse or amse is equivalent to assessing variance or asymptotic variance. Since variances and asymptotic variances usually depend on the unknown population, we have to estimate them in order to report accuracies of point estimators. Variance estimation is an important part of statistical inference, not only for

assessing accuracy, but also for constructing inference procedures studied in Chapters 6 and 7. See also the discussion at the end of §2.5.1.

If the unknown population is in a parametric family indexed by a parameter θ, then the covariance matrix or asymptotic covariance matrix of an estimator of θ is a function of θ, say $V_n(\theta)$. If θ is estimated by $\hat{\theta}$, then it is natural to estimate $V_n(\theta)$ by $V_n(\hat{\theta})$. Thus, variance estimation in parametric problems is usually simple. This idea of substitution can be applied to nonparametric problems in which variance estimation is much more complex.

We introduce three commonly used variance estimation methods in this section, the substitution method, the jackknife, and the bootstrap.

5.5.1 The substitution method

Suppose that we can obtain a formula for the covariance or asymptotic covariance matrix of an estimator $\hat{\theta}_n$. Then a direct method of variance estimation is to substitute unknown quantities in the variance formula by some estimators. To illustrate, consider the simplest case where $X_1, ..., X_n$ are i.i.d. random d-vectors, $\hat{\theta}_n = g(\bar{X})$, and g is a function from $\mathcal{R}^d$ to $\mathcal{R}^k$. Suppose that $E\|X_1\|^2 < \infty$ and g is differentiable at $\mu = E(X_1)$. Then, by the CLT and Theorem 1.12(i),

$$(\hat{\theta}_n - \theta)V_n^{-1/2} \to_d N_k(0, I_k), \tag{5.100}$$

where $\theta = g(\mu)$ and

$$V_n = \nabla g(\mu)\mathrm{Var}(X_1)[\nabla g(\mu)]^\tau/n \tag{5.101}$$

is the asymptotic covariance matrix of $\hat{\theta}_n$ which depends on unknown quantities μ and $\mathrm{Var}(X_1)$. A substitution estimator of V_n is

$$\hat{V}_n = \nabla g(\bar{X})S^2[\nabla g(\bar{X})]^\tau/n, \tag{5.102}$$

where $S^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^\tau(X_i - \bar{X})$ is the sample covariance matrix, an extension of the sample variance to the multivariate X_i's.

An essential asymptotic requirement in variance estimation is the consistency of variance estimators according to the following definition.

Definition 5.4. Let $\{V_n\}$ be a sequence of $k \times k$ positive definite matrices and $\hat{V}_n$ be an estimator of V_n for each n. Then $\{\hat{V}_n\}$ or $\hat{V}_n$ is said to be consistent for V_n if and only if

$$l\hat{V}_n l^\tau / lV_n l \to_p 1 \qquad \text{for any } l \in \mathcal{R}^k. \tag{5.103}$$

$\hat{V}_n$ is strongly consistent if (5.103) holds with $\to_p$ replaced by $\to_{a.s.}$. ∎

If (5.103) holds, then $(\hat{\theta}_n - \theta)\hat{V}_n^{-1/2} \to_d N_k(0, I_k)$, a result useful for asymptotic inference as discussed in Chapters 6 and 7.

By the SLLN, $\hat{V}_n$ in (5.102) is strongly consistent for V_n in (5.101), provided that $\nabla g(\mu) \neq 0$ and ∇g is continuous at μ so that $\nabla g(\bar{X}) \to_{a.s.} \nabla g(\mu)$.

Example 5.15. Let $Y_1, ..., Y_n$ be i.i.d. random variables with finite $\mu_y = EY_1$, $\sigma_y^2 = \mathrm{Var}(Y_1)$, $\gamma_y = EY_1^3$, and $\kappa_y = EY_1^4$. Consider the estimation of $\theta = (\mu_y, \sigma_y^2)$. Let $\hat{\theta}_n = (\bar{X}, \hat{\sigma}_y^2)$, where $\hat{\sigma}_y^2 = n^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$. If $X_i = (Y_i, Y_i^2)$, then $\hat{\theta}_n = g(\bar{X})$ with $g(x) = (x_1, x_2 - x_1^2)$. Hence, (5.100) holds with

$$\mathrm{Var}(X_1) = \begin{pmatrix} \sigma_y^2 & \gamma_y - \mu_y(\sigma_y^2 + \mu_y^2) \\ \gamma_y - \mu_y(\sigma_y^2 + \mu_y^2) & \kappa_y - (\sigma_y^2 + \mu_y^2)^2 \end{pmatrix}$$

and

$$\nabla g(x) = \begin{pmatrix} 1 & 0 \\ -2x_1 & 1 \end{pmatrix}.$$

The estimator $\hat{V}_n$ in (5.102) is strongly consistent, since $\nabla g(x)$ is obviously a continuous function. ∎

Similar results can be obtained for problems in Examples 3.21 and 3.23, and Exercises 92 and 93 in §3.6.

A key step in the previous discussion is the derivation of formula (5.101) for the asymptotic covariance matrix of $\hat{\theta}_n = g(\bar{X})$, via Taylor's expansion (Theorem 1.12) and the CLT. Thus, the idea can be applied to the case where $\hat{\theta}_n = \mathtt{T}(F_n)$, a differentiable statistical functional.

We still consider i.i.d. random d-vectors $X_1, ..., X_n$ from F. Suppose that $\mathtt{T}$ is a vector-valued functional whose components are ϱ-Hadamard differentiable at F, where ϱ is either ϱ_∞ or a distance satisfying (5.35). Let ϕ_F be the vector of influence functions of components of $\mathtt{T}$. If the components of ϕ_F satisfy (5.33), then (5.100) holds with $\theta = \mathtt{T}(F)$, $\hat{\theta}_n = \mathtt{T}(F_n)$, $F_n =$ the empirical c.d.f. in (5.1), and

$$V_n = \frac{\mathrm{Var}(\phi_F(X_1))}{n} = \frac{1}{n}\int [\phi_F(x)]^\tau \phi_F(x) dF(x). \tag{5.104}$$

Formula (5.104) leads to a natural substitution variance estimator

$$\hat{V}_n = \frac{1}{n}\int [\phi_{F_n}(x)]^\tau \phi_{F_n}(x) dF_n(x) = \frac{1}{n^2}\sum_{i=1}^n [\phi_{F_n}(X_i)]^\tau \phi_{F_n}(X_i), \tag{5.105}$$

provided that $\phi_{F_n}(x)$ is well defined, i.e., the components of $\mathtt{T}$ are Gâteaux differentiable at F_n for sufficiently large n. Under some more conditions on ϕ_{F_n} we can establish the consistency of $\hat{V}_n$ in (5.105).

Theorem 5.15. Let $X_1, ..., X_n$ be i.i.d. random d-vectors from F, $\mathtt{T}$ be a vector-valued functional whose components are Gâteaux differentiable at F and F_n, and ϕ_F be the vector of influence functions of components of $\mathtt{T}$. Suppose that $\sup_{\|x\|\le c}\|\phi_{F_n}(x)-\phi_F(x)\| = o_p(1)$ for any $c>0$ and that there exist a constant $c_0>0$ and a function $h(x)\ge 0$ such that $E[h(X_1)]<\infty$ and $P\big(\sup_{\|x\|\ge c_0}\|\phi_{F_n}(x)\|^2\le h(x)\big)\to 1$. Then $\hat{V}_n$ in (5.105) is consistent for V_n in (5.104).

Proof. Let $\zeta(x)=[\phi_F(x)]^\tau[\phi_F(x)]$ and $\zeta_n(x)=[\phi_{F_n}(x)]^\tau[\phi_{F_n}(x)]$. By the SLLN,

$$\frac{1}{n}\sum_{i=1}^n \zeta(X_i)\to_{a.s.}\int \zeta(x)dF(x).$$

Hence the result follows from

$$\frac{1}{n}\sum_{i=1}^n[\zeta_n(X_i)-\zeta(X_i)]=o_p(1).$$

Using the assumed conditions and the argument in the proof of Lemma 5.3, we can show that for any $\epsilon>0$, there is a $c>0$ such that

$$P\left(\frac{1}{n}\sum_{i=1}^n\|\zeta_n(X_i)-\zeta(X_i)\|I_{(c,\infty)}(\|X_i\|)>\frac{\epsilon}{2}\right)\le\epsilon$$

and

$$P\left(\frac{1}{n}\sum_{i=1}^n\|\zeta_n(X_i)-\zeta(X_i)\|I_{[0,c]}(\|X_i\|)>\frac{\epsilon}{2}\right)\le\epsilon$$

for sufficiently large n. This completes the proof. ∎

Example 5.16. Consider the L-functional defined in (5.39) and the L-estimator $\hat{\theta}_n=\mathtt{T}(F_n)$. Theorem 5.6 shows that $\mathtt{T}$ is Hadamard differentiable at F under some conditions on J. It can be shown (exercise) that $\mathtt{T}$ is Gâteaux differentiable at F_n with $\phi_{F_n}(x)$ given by (5.41) (with F replaced by F_n). Then the difference $\phi_{F_n}(x)-\phi_F(x)$ is equal to

$$\int(F_n-F)(y)J(F_n(y))dy+\int(F-\delta_x)(y)[J(F_n(y))-J(F(y))]dy.$$

One can show (exercise) that the conditions in Theorem 5.15 are satisfied if the conditions in Theorem 5.6(i) or (ii) (with $E|X_1|<\infty$) hold. ∎

Substitution variance estimators for M-estimators and U-statistics can also be derived (exercises).

The substitution method can clearly be applied to non-i.i.d. cases. For example, the LSE $\hat{\beta}$ in linear model (3.25) with a full rank Z and i.i.d. ε_i's

has $\mathrm{Var}(\hat{\beta}) = \sigma^2(Z^\tau Z)^{-1}$, where $\sigma^2 = \mathrm{Var}(\varepsilon_1)$. A consistent substitution estimator of $\mathrm{Var}(\hat{\beta})$ can be obtained by replacing σ^2 in the formula of $\mathrm{Var}(\hat{\beta})$ by a consistent estimator of σ^2 such as $SSR/(n-p)$ (see (3.36)).

We now consider variance estimation for the GEE estimators described in §5.4.1. By Theorem 5.14, the asymptotic covariance matrix of the GEE estimator $\hat{\theta}_n$ is given by (5.95), where

$$\mathrm{Var}(s_n(\theta)) = \sum_{i=1}^n E\{[\psi_i(X_i,\theta)]^\tau \psi_i(X_i,\theta)\},$$

$$M_n(\theta) = \sum_{i=1}^n E[\varphi_i(X_i,\theta)],$$

and $\varphi_i(x,\gamma) = \partial\psi_i(x,\gamma)/\partial\gamma$. Substituting θ by $\hat{\theta}_n$ and the expectations by their empirical analogues, we obtain the substitution estimator $\hat{V}_n = \hat{M}_n^{-1}\widehat{\mathrm{Var}}(s_n)\hat{M}_n^{-1}$, where

$$\widehat{\mathrm{Var}}(s_n) = \sum_{i=1}^n [\psi_i(X_i,\hat{\theta}_n)]^\tau \psi_i(X_i,\hat{\theta}_n)$$

and

$$\hat{M}_n = \sum_{i=1}^n \varphi_i(X_i,\hat{\theta}_n).$$

The proof of the following result is left as an exercise.

Theorem 5.16. Let $X_1, ..., X_n$ be independent and $\{\hat{\theta}_n\}$ be a consistent sequence of GEE estimators. Assume the conditions in Theorem 5.14. Suppose further that the sequence of functions $\{h_{ij}, i = 1, 2, ...\}$ satisfies the conditions in Lemma 5.3 with Θ replaced by a compact neighborhood of θ, where $h_{ij}(x,\gamma)$ is the jth row of $[\psi_i(x,\gamma)]^\tau\psi_i(x,\gamma)$. Let V_n be given by (5.95). Then $\hat{V}_n = \hat{M}_n^{-1}\widehat{\mathrm{Var}}(s_n)\hat{M}_n^{-1}$ is consistent for V_n. ∎

5.5.2 The jackknife

Applying the substitution method requires the derivation of a formula for the covariance matrix or asymptotic covariance matrix of a point estimator. There are variance estimation methods that can be used without actually deriving such a formula (only the existence of the covariance matrix or asymptotic covariance matrix is assumed), at the expense of requiring a large number of computations. These methods are called *resampling* methods, *replication* methods, or *data reuse* methods. The *jackknife* method introduced here and the *bootstrap* method in §5.5.3 are the most popular resampling methods.

The jackknife method was proposed by Quenouille (1949) and Tukey (1958). Let $\hat{\theta}_n$ be a vector-valued estimator based on independent X_i's, where each X_i is a random d_i-vector and $\sup_i d_i < \infty$. Let $\hat{\theta}_{-i}$ be the same estimator but based on $X_1, ..., X_{i-1}, X_{i+1}, ..., X_n$, $i = 1, ..., n$. Note that $\hat{\theta}_{-i}$ also depends on n but the subscript n is omitted for simplicity. Since $\hat{\theta}_n$ and $\hat{\theta}_{-1}, ..., \hat{\theta}_{-n}$ are estimators of the same quantity, the "sample covariance matrix"

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\theta}_{-i}-\bar{\theta}_n\right)^{\tau}\left(\hat{\theta}_{-i}-\bar{\theta}_n\right) \tag{5.106}$$

can be used as a measure of the variation of $\hat{\theta}_n$, where $\bar{\theta}_n$ is the average of $\hat{\theta}_{-i}$'s.

There are two major differences between the quantity in (5.106) and the sample covariance matrix S^2 previously discussed. First, $\hat{\theta}_{-i}$'s are not independent. Second, $\hat{\theta}_{-i} - \hat{\theta}_{-j}$ usually converges to 0 at a fast rate (such as n^{-1}). Hence, to estimate the asymptotic covariance matrix of $\hat{\theta}_n$, the quantity in (5.106) should be multiplied by a correction factor c_n. If $\hat{\theta}_n = \bar{X}$ ($d_i \equiv d$), then $\hat{\theta}_{-i} = (n-1)^{-1}(\bar{X} - X_i)$ and the quantity in (5.106) reduces to

$$\frac{1}{(n-1)^3}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^{\tau}\left(X_i-\bar{X}\right) = \frac{1}{(n-1)^2}S^2,$$

where S^2 is the sample covariance matrix. Thus, the correction factor c_n is $(n-1)^2/n$ for the case of $\hat{\theta}_n = \bar{X}$, since, by the SLLN, S^2/n is consistent for $\mathrm{Var}(\bar{X})$.

It turns out that the same correction factor works for many other estimators. This leads to the following *jackknife variance estimator* for $\hat{\theta}_n$:

$$\hat{V}_J = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{-i}-\bar{\theta}_n\right)^{\tau}\left(\hat{\theta}_{-i}-\bar{\theta}_n\right). \tag{5.107}$$

Theorem 5.17. Let $X_1, ..., X_n$ be i.i.d. random d-vectors from F with finite $\mu = E(X_1)$ and $\mathrm{Var}(X_1)$, and let $\hat{\theta}_n = g(\bar{X})$. Suppose that ∇g is continuous at μ and $\nabla g(\mu) \neq 0$. Then the jackknife variance estimator $\hat{V}_J$ in (5.107) is strongly consistent for V_n in (5.101).

Proof. From Definition 5.4, it suffices to show the case where g is real-valued. Let $\bar{X}_{-i}$ be the sample mean based on $X_1, ..., X_{i-1}, X_{i+1}, ..., X_n$. From the mean-value theorem, we have

$$\begin{aligned}\hat{\theta}_{-i} - \hat{\theta}_n &= g(\bar{X}_{-i}) - g(\bar{X}) \\ &= \nabla g(\xi_{n,i})(\bar{X}_{-i} - \bar{X})^{\tau} \\ &= \nabla g(\bar{X})(\bar{X}_{-i} - \bar{X})^{\tau} + R_{n,i},\end{aligned}$$

where $R_{n,i} = \big[\nabla g(\xi_{n,i}) - \nabla g(\bar{X})\big](\bar{X}_{-i} - \bar{X})^\tau$ and $\xi_{n,i}$ is a point on the line segment between $\bar{X}_{-i}$ and $\bar{X}$. From $\bar{X}_{-i} - \bar{X} = (n-1)^{-1}(\bar{X} - X_i)$, it follows that $\sum_{i=1}^n (\bar{X}_{-i} - \bar{X}) = 0$ and

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

From the definition of the jackknife estimator in (5.107),

$$\hat{V}_J = A_n + B_n + 2C_n,$$

where

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

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

and

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

By $\bar{X}_{-i} - \bar{X} = (n-1)^{-1}(\bar{X} - X_i)$, the SLLN, and the continuity of ∇g at μ,

$$A_n/V_n \to_{a.s.} 1.$$

Also,

$$(n-1)\sum_{i=1}^n \|\bar{X}_{-i} - \bar{X}\|^2 = \frac{1}{n-1}\sum_{i=1}^n \|X_i - \bar{X}\|^2 = O(1) \quad \text{a.s.} \tag{5.108}$$

Hence

$$\max_{i\le n} \|\bar{X}_{-i} - \bar{X}\|^2 \to_{a.s.} 0,$$

which, together with the continuity of ∇g at μ and $\|\xi_{n,i} - \bar{X}\| \le \|\bar{X}_{-i} - \bar{X}\|$, implies that

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

From (5.101) and (5.108), $\sum_{i=1}^n \|\bar{X}_{-i} - \bar{X}\|^2/V_n = O(1)$ a.s. Hence

$$\frac{B_n}{V_n} \le \frac{n-1}{V_n n}\sum_{i=1}^n R_{n,i}^2 \le \frac{u_n}{V_n}\sum_{i=1}^n \|\bar{X}_{-i} - \bar{X}\|^2 \to_{a.s.} 0.$$

By the Cauchy-Schwarz inequality, $(C_n/V_n)^2 \le (A_n/V_n)(B_n/V_n) \to_{a.s.} 0$. This proves the result. ∎

A key step in the proof of Theorem 5.17 is that $\hat{\theta}_{-i} - \hat{\theta}_n$ can be approximated by $\nabla g(\bar{X})(\bar{X}_{-i} - \bar{X})^\tau$ and the contributions of the remainders, $R_{n,1}, ..., R_{n,n}$, are sufficiently small, i.e., $B_n/V_n \to_{a.s.} 0$. This indicates that the jackknife estimator (5.107) is consistent for $\hat{\theta}_n$ that can be well approximated by some linear statistic. In fact, the jackknife estimator (5.107) has been shown to be consistent when $\hat{\theta}_n$ is a U-statistic (Arvesen, 1969) or a statistical functional that is Hadamard differentiable and *continuously* Gâteaux differentiable at F (which includes certain types of L-estimators and M-estimators). More details can be found in Shao and Tu (1995, Chapter 2).

The jackknife method can be applied to non-i.i.d. problems. A detailed discussion of the use of the jackknife method in survey problems can be found in Shao and Tu (1995, Chapter 6). We now consider the jackknife variance estimator for the LSE $\hat{\beta}$ in linear model (3.25). For simplicity, assume that Z is of full rank. Assume also that ε_i's are independent with $E(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma_i^2$. Then

$$\mathrm{Var}(\hat{\beta}) = (Z^\tau Z)^{-1} \sum_{i=1}^n \sigma_i^2 Z_i^\tau Z_i (Z^\tau Z)^{-1}.$$

Let $\hat{\beta}_{-i}$ be the LSE of β based on the data with the ith pair (X_i, Z_i) deleted. Using the fact that $(A + c^\tau c)^{-1} = A^{-1} - A^{-1}c^\tau c A^{-1}/(1 + cA^{-1}c^\tau)$ for a matrix A and a vector c, we can show that (exercise)

$$\hat{\beta}_{-i} = \hat{\beta} - r_i Z_i/(1 - h_{ii}), \tag{5.109}$$

where $r_i = X_i - \hat{\beta} Z_i^\tau$ is the ith residual and $h_{ii} = Z_i(Z^\tau Z)^{-1} Z_i^\tau$. Hence

$$\hat{V}_J = \frac{n-1}{n}(Z^\tau Z)^{-1} \left[\sum_{i=1}^n \frac{r_i^2 Z_i^\tau Z_i}{(1-h_{ii})^2} - \frac{1}{n} \sum_{i=1}^n \frac{r_i Z_i^\tau}{1-h_{ii}} \sum_{i=1}^n \frac{r_i Z_i}{1-h_{ii}} \right] (Z^\tau Z)^{-1}.$$

Wu (1986) proposed the following weighted jackknife variance estimator that improves $\hat{V}_J$:

$$\hat{V}_{WJ} = \sum_{i=1}^n (1-h_{ii}) \left(\hat{\beta}_{-i} - \hat{\beta}\right)^\tau \left(\hat{\beta}_{-i} - \hat{\beta}\right) = (Z^\tau Z)^{-1} \sum_{i=1}^n \frac{r_i^2 Z_i^\tau Z_i}{1-h_{ii}} (Z^\tau Z)^{-1}.$$

Theorem 5.18. Assume the conditions in Theorem 3.12 and that ε_i's are independent. Then both $\hat{V}_J$ and $\hat{V}_{WJ}$ are consistent for $\mathrm{Var}(\hat{\beta})$.

Proof. Let $l \in \mathcal{R}^p$ be a fixed nonzero vector and $l_i = l(Z^\tau Z)^{-1} Z_i^\tau$. Since $\max_{i \le n} h_{ii} \to 0$, the result for $\hat{V}_{WJ}$ follows from

$$\sum_{i=1}^n l_i^2 r_i^2 \Big/ \sum_{i=1}^n l_i^2 \sigma_i^2 \to_p 1. \tag{5.110}$$

By the WLLN (Theorem 1.14(ii)),

$$\sum_{i=1}^n l_i^2\varepsilon_i^2 \Big/ \sum_{i=1}^n l_i^2\sigma_i^2 \to_p 1.$$

Note that $r_i = \varepsilon_i + (\beta - \hat\beta)Z_i^\tau$ and

$$\max_{i\le n}[(\beta-\hat\beta)Z_i^\tau]^2 \le \|(\beta-\hat\beta)Z^\tau\|^2 \max_{i\le n} h_{ii} = o_p(1).$$

Hence (5.110) holds.

The consistency of $\hat V_J$ follows from (5.110) and

$$\frac{n-1}{n^2}\left(\sum_{i=1}^n \frac{l_i r_i}{1-h_{ii}}\right)^2 \Big/ \sum_{i=1}^n l_i^2\sigma_i^2 = o_p(1). \tag{5.111}$$

The proof of (5.111) is left as an exercise. ∎

Finally, let us consider the jackknife estimators for GEE estimators in §5.4.1. Under the conditions of Proposition 5.5 or 5.6, it can be shown that

$$\max_{i\le n}\|\hat\theta_{-i} - \hat\theta\| = o_p(1), \tag{5.112}$$

where $\hat\theta_{-i}$ is a root of $s_{ni}(\gamma) = 0$ and

$$s_{ni}(\gamma) = \sum_{j\ne i, j\le n} \psi_j(X_j,\gamma).$$

Using Taylor's expansion and the fact that $s_{ni}(\hat\theta_{-i}) = 0$ and $s_n(\hat\theta_n) = 0$, we obtain that

$$\psi_i(X_i,\hat\theta_{-i}) = (\hat\theta_{-i} - \hat\theta_n)\int_0^1 \nabla s_n\big(\hat\theta_n + t(\hat\theta_{-i} - \hat\theta_n)\big)dt.$$

Following the proof of Theorem 5.14, we obtain that

$$\hat V_J = [M_n(\theta)]^{-1}\sum_{i=1}^n [\psi_i(X_i,\hat\theta_{-i})]^\tau \psi_i(X_i,\hat\theta_{-i})[M_n(\theta)]^{-1} + R_n,$$

where R_n denotes a quantity satisfying $lR_nl^\tau/lV_nl^\tau = o_p(1)$ for V_n in (5.95). Under the conditions of Theorem 5.16, it follows from (5.112) that $\hat V_J$ is consistent.

If $\hat\theta_n$ is computed using an iteration method, then the computation of $\hat V_J$ requires n additional iteration processes. We may use the idea of a one-step MLE to reduce the amount of computation. For each i, let

$$\hat\theta_{-i} = \hat\theta_n - s_{ni}(\hat\theta_n)[\nabla s_{ni}(\hat\theta_n)]^{-1}, \tag{5.113}$$

which is the result from the first iteration when Newton-Raphson's method is applied in computing a root of $s_{ni}(\gamma) = 0$ and $\hat{\theta}_n$ is used as the initial point. Note that $\hat{\theta}_{-i}$'s in (5.113) satisfy (5.112) (exercise). If the jackknife variance estimator is based on $\hat{\theta}_{-i}$'s in (5.113), then

$$\hat{V}_J = [M_n(\theta)]^{-1} \sum_{i=1}^{n} [\psi_i(X_i, \hat{\theta}_n)]^\tau \psi_i(X_i, \hat{\theta}_n) [M_n(\theta)]^{-1} + R_n.$$

These results are summarized in the following theorem.

Theorem 5.19. Assume the conditions in Theorems 5.14 and 5.16. Assume further that $\hat{\theta}_{-i}$'s are given by (5.113) or GEE estimators satisfying (5.112). Then the jackknife variance estimator $\hat{V}_J$ is consistent for V_n given in (5.95). ∎

5.5.3 The bootstrap

The basic idea of the bootstrap method can be described as follows. Suppose that P is a population or model that generates the sample X and that we need to estimate $\mathrm{Var}(\hat{\theta})$, where $\hat{\theta} = \hat{\theta}(X)$ is an estimator, a statistic based on X. Suppose further that the unknown population P is estimated by $\hat{P}$, based on the sample X. Let X^* be a sample (called a *bootstrap sample*) taken from the estimated population $\hat{P}$ using the same or a similar sampling procedure used to obtain X, and let $\hat{\theta}^* = \hat{\theta}(X^*)$, which is the same as $\hat{\theta}$ but with X replaced by X^*. If we believe that $P = \hat{P}$ (i.e., we have a perfect estimate of the population), then $\mathrm{Var}(\hat{\theta}) = \mathrm{Var}_*(\hat{\theta}^*)$, where Var_* is the conditional variance w.r.t. the randomness in generating X^*, given X. In general, $P \neq \hat{P}$ and, therefore, $\mathrm{Var}(\hat{\theta}) \neq \mathrm{Var}_*(\hat{\theta}^*)$. But $\hat{V}_B = \mathrm{Var}_*(\hat{\theta}^*)$ is an empirical analogue of $\mathrm{Var}(\hat{\theta})$ and can be used as an estimate of $\mathrm{Var}(\hat{\theta})$.

In a few cases, an explicit form of $\hat{V}_B = \mathrm{Var}_*(\hat{\theta}^*)$ can be obtained. First, consider i.i.d. $X_1, ..., X_n$ from a c.d.f. F on $\mathcal{R}^d$. The population is determined by F. Suppose that we estimate F by the empirical c.d.f. F_n in (5.1) and that $X_1^*, ..., X_n^*$ are i.i.d. from F_n. For $\hat{\theta} = \bar{X}$, its bootstrap analogue is $\hat{\theta}^* = \bar{X}^*$, the average of X_i^*'s. Then

$$\hat{V}_B = \mathrm{Var}_*(\bar{X}^*) = \frac{1}{n^2} \sum_{i=1}^{n} (X_i - \bar{X})^\tau (X_i - \bar{X}) = \frac{n-1}{n^2} S^2,$$

where S^2 is the sample covariance matrix. In this case $\hat{V}_B = \mathrm{Var}_*(\bar{X}^*)$ is a strongly consistent estimator for $\mathrm{Var}(\bar{X})$. Next, consider i.i.d. random variables $X_1, ..., X_n$ from a c.d.f. F on $\mathcal{R}$ and $\hat{\theta} = F_n^{-1}(\frac{1}{2})$, the sample

median. Suppose that $n = 2l - 1$ for an integer l. Let $X_1^*, ..., X_n^*$ be i.i.d. from F_n and $\hat{\theta}^*$ be the sample median based on $X_1^*, ..., X_n^*$. Then

$$\hat{V}_B = \mathrm{Var}_*(\hat{\theta}^*) = \sum_{j=1}^{n} p_j \left(X_{(j)} - \sum_{i=1}^{n} p_i X_{(i)} \right)^2,$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ are order statistics and $p_j = P(\hat{\theta}^* = X_{(j)}|X)$. It can be shown (exercise) that

$$p_j = \sum_{t=0}^{l-1} \binom{n}{t} \frac{(j-1)^t (n-j+1)^{n-t} - j^t (n-j)^{n-t}}{n^n}. \tag{5.114}$$

However, in most cases $\hat{V}_B$ does not have a simple explicit form. When P is known, the Monte Carlo method described in §4.1.4 can be used to approximate $\mathrm{Var}(\hat{\theta})$. That is, we draw repeatedly new data sets from P and then use the sample covariance matrix based on the values of $\hat{\theta}$ computed from new data sets as a numerical approximation to $\mathrm{Var}(\hat{\theta})$. This idea can be used to approximate $\hat{V}_B$, since $\hat{P}$ is a known population. That is, we can draw m bootstrap data sets $X^{*1}, ..., X^{*m}$ independently from $\hat{P}$ (conditioned on X), compute $\hat{\theta}^{*j} = \hat{\theta}(X^{*j})$, $j = 1, ..., m$, and approximate $\hat{V}_B$ by

$$\hat{V}_B^m = \frac{1}{m} \sum_{j=1}^{m} \left(\hat{\theta}^{*j} - \bar{\theta}^* \right)^\tau \left(\hat{\theta}^{*j} - \bar{\theta}^* \right),$$

where $\bar{\theta}^*$ is the average of $\hat{\theta}^{*j}$'s. Since each X^{*j} is a data set generated from $\hat{P}$, $\hat{V}_B^m$ is a resampling estimator. From the SLLN, as $m \to \infty$, $\hat{V}_B^m \to_{a.s.} \hat{V}_B$, conditioned on X. Both $\hat{V}_B$ and its Monte Carlo approximation $\hat{V}_B^m$ are called *bootstrap variance estimators* for $\hat{\theta}$. $\hat{V}_B^m$ is more useful in practical applications, whereas in theoretical studies, we usually focus on $\hat{V}_B$.

The consistency of the bootstrap variance estimator $\hat{V}_B$ is a much more complicated problem than that of the jackknife variance estimator in §5.5.2. Some examples can be found in Shao and Tu (1995, §3.2.2).

The bootstrap method can also be applied to estimate quantities other than $\mathrm{Var}(\hat{\theta})$. For example, let $K(t) = P(\hat{\theta} \leq t)$ be the c.d.f. of a real-valued estimator $\hat{\theta}$. From the previous discussion, a bootstrap estimator of $K(t)$ is the conditional probability $P(\hat{\theta}^* \leq t|X)$, which can be approximated by the Monte Carlo approximation $m^{-1} \sum_{j=1}^{m} I_{(-\infty,t]}(\hat{\theta}^{*j})$. An important application of bootstrap distribution estimators in problems of constructing confidence sets is studied in §7.4. Here, we study the use of a bootstrap distribution estimator to form a consistent estimator of the asymptotic variance of a real-valued estimator $\hat{\theta}$.

Suppose that

$$\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, v), \tag{5.115}$$

where v is unknown. Let $H_n(t)$ be the c.d.f. of $\sqrt{n}(\hat{\theta} - \theta)$ and

$$\hat{H}_B(t) = P(\sqrt{n}(\hat{\theta}^* - \hat{\theta}) \leq t|X) \tag{5.116}$$

be a bootstrap estimator of $H_n(t)$. If

$$\hat{H}_B(t) - H_n(t) \to_p 0$$

for any t, then, by (5.115),

$$\hat{H}_B(t) - \Phi\left(t/\sqrt{v}\right) \to_p 0,$$

which implies (exercise) that

$$\hat{H}_B^{-1}(\alpha) \to_p \Phi^{-1}\left(\alpha/\sqrt{v}\right) = \sqrt{v}\Phi^{-1}(\alpha)$$

for any $\alpha \in (0, 1)$. Then, for $\alpha \neq \frac{1}{2}$,

$$\hat{H}_B^{-1}(1-\alpha) - \hat{H}_B^{-1}(\alpha) \to_p \sqrt{v}[\Phi^{-1}(1-\alpha) - \Phi^{-1}(\alpha)].$$

Therefore, a consistent estimator of v/n, the asymptotic variance of $\hat{\theta}$, is

$$\tilde{V}_B = \frac{1}{n}\left[\frac{\hat{H}_B^{-1}(1-\alpha) - \hat{H}_B^{-1}(\alpha)}{\Phi^{-1}(1-\alpha) - \Phi^{-1}(\alpha)}\right]^2.$$

The following result gives some conditions under which $\hat{H}_B(t) - H_n(t) \to_p 0$. The proof of part (i) is omitted. The proof of part (ii) is given in Exercises 97-99 in §5.6.

Theorem 5.20. Suppose that $X_1, ..., X_n$ are i.i.d. from a c.d.f. F on $\mathcal{R}^d$. Let $\hat{\theta} = \mathtt{T}(F_n)$, where $\mathtt{T}$ is a real-valued functional, $\hat{\theta}^* = \mathtt{T}(F_n^*)$, where F_n^* is the empirical c.d.f. based on a bootstrap sample $X_1^*, ..., X_n^*$ i.i.d. from F_n, and let $\hat{H}_B$ be given by (5.116).
(i) If $\mathtt{T}$ is ϱ_∞-Hadamard differentiable at F and (5.33) holds, then

$$\varrho_\infty(\hat{H}_B, H_n) \to_p 0. \tag{5.117}$$

(ii) If $d = 1$ and $\mathtt{T}$ is ϱ_{L_p}-Fréchet differentiable at F ($\int\{F(t)[1-F(t)]\}^{p/2}dt < \infty$ if $1 \leq p < 2$) and (5.33) holds, then (5.117) holds. ∎

Applications of the bootstrap method to non-i.i.d. cases can be found, for example, in Efron and Tibshirani (1993), Hall (1992), and Shao and Tu (1995).

5.6 Exercises

1. Let ϱ_∞ be defined by (5.3).
(a) Show ϱ_∞ is a distance on $\mathcal{F}$.
(b) Find an example of a sequence $\{G_j\} \subset \mathcal{F}$ for which $\lim_{j\to\infty} G_j(t) = G_0(t)$ for every t at which G_0 is continuous, but $\varrho_\infty(G_j, G_0)$ does not converge to 0.

2. Let $X_1, ..., X_n$ be i.i.d. random d-vectors with c.d.f. F and F_n be the empirical c.d.f. defined by (5.1). Show that for any $t > 0$ and $\epsilon > 0$, there is a $C_{\epsilon,d}$ such that for all $n = 1, 2, ...,$

$$P\left(\sup_{m\geq n} \varrho_\infty(F_m, F) > t\right) \leq \frac{C_{\epsilon,d} e^{-(2-\epsilon)t^2 n}}{1 - e^{-(2-\epsilon)t^2}}.$$

3. Show that ϱ_{M_p} defined by (5.4) is a distance on $\mathcal{F}_p$, $p \geq 1$.

4. Show that ϱ_{L_p} defined by (5.5) is a distance on $\mathcal{F}_1$ for any $p \geq 1$.

5. Let $\mathcal{F}_1$ be the collection of c.d.f.'s on $\mathcal{R}$ with finite means.
(a) Show that $\varrho_{M_1}(G_1, G_2) = \int_0^1 |G_1^{-1}(z) - G_2^{-1}(z)|dz$, where $G^{-1}(z) = \inf\{t : G(t) \geq z\}$ for any $G \in \mathcal{F}$.
(b) Show that $\varrho_{M_1}(G_1, G_2) = \varrho_{L_1}(G_1, G_2)$.

6. Find an example of a sequence $\{G_j\} \subset \mathcal{F}$ for which
(a) $\lim_{j\to\infty} \varrho_\infty(G_j, G_0) = 0$ but $\varrho_{M_2}(G_j, G_0)$ does not converge to 0;
(b) $\lim_{j\to\infty} \varrho_{M_2}(G_j, G_0) = 0$ but $\varrho_\infty(G_j, G_0)$ does not converge to 0.

7. Repeat the previous exercise with ϱ_{M_2} replaced by ϱ_{L_2}.

8. Let X be a random variable having c.d.f. F. Show that
(a) $E|X|^2 < \infty$ implies $\int \{F(t)[1 - F(t)]\}^{p/2} dt < \infty$ for $p \in (1, 2)$;
(b) $E|X|^{2+\delta} < \infty$ with some $\delta > 0$ implies $\int \{F(t)[1 - F(t)]\}^{1/2} dt < \infty$.

9. For any one-dimensional $G_j \in \mathcal{F}_1$, $j = 1, 2$, show that $\varrho_{L_1}(G_1, G_2) \geq |\int x dG_1 - \int x dG_2|$.

10. In the proof of Theorem 5.3, show that $p_i = c/n$, $i = 1, ..., n$, $\lambda = -(c/n)^{n-1}$ is a maximum of the function $H(p_1, ..., p_n, \lambda)$ over $p_i > 0$, $i = 1, ..., n$, $\sum_{i=1}^n p_i = c$.

11. Show that (5.11)-(5.13) is a solution to the problem of maximizing $\ell(G)$ in (5.8) subject to (5.10).

12. In the proof of Theorem 5.4, prove the case of $m \geq 2$.

13. Show that a maximum of $\ell(G)$ in (5.17) subject to (5.10) is given by (5.11) with $\hat{p}_i$ defined by (5.18) and (5.19).

14. In Example 5.2, show that an MELE is given by (5.11) with $\hat{p}_i$'s given by (5.21).

15. In Example 5.3, show that
(a) maximizing (5.22) subject to (5.23) is equivalent to maximizing

$$\prod_{i=1}^{n} q_i^{\delta_{(i)}}(1-q_i)^{n-i+1-\delta_{(i)}},$$

where $q_i = p_i/\sum_{j=i}^{n+1} p_j$, $i = 1, ..., n$;
(b) $\hat{F}$ given by (5.24) maximizes (5.22) subject to (5.23); (Hint: use part (a) and the fact that $p_i = q_i \prod_{j=1}^{i-1}(1-q_j)$.)
(c) $\hat{F}$ given by (5.25) is the same as that in (5.24).
(d) if $\delta_i = 1$ for all i (no censoring), then $\hat{F}$ in (5.25) is the same as the empirical c.d.f. in (5.1).

16. Let f_n be given by (5.26).
(a) Show that f_n is a Lebesgue p.d.f. on $\mathcal{R}$.
(b) Suppose that f is continuously differentiable at t, $\lambda_n \to 0$, and $n\lambda_n \to \infty$. Show that (5.27) holds.
(c) Under $n\lambda_n^3 \to 0$ and the conditions of (b), show that (5.28) holds.
(d) Suppose that f is continuous on $[a, b]$, $-\infty < a < b < \infty$, $\lambda_n \to 0$, and $n\lambda_n \to \infty$. Show that $\int_a^b f_n(t)dt \to_p \int_a^b f(t)dt$.

17. Let $\hat{f}$ be given by (5.29).
(a) Show that $\hat{f}$ is a Lebesgue p.d.f. on $\mathcal{R}$.
(b) Prove (5.30) under the condition that $\lambda_n \to 0$, $n\lambda_n \to \infty$, and f is bounded and is continuous at t and $\int [w(t)]^{2+\delta}dt < \infty$ for some $\delta > 0$. (Hint: check Liapunov's condition and apply Theorem 1.15.)
(c) Suppose that $\lambda_n \to 0$, $n\lambda_n \to \infty$, and f is bounded and is continuous on $[a, b]$, $-\infty < a < b < \infty$. Show that $\int_a^b \hat{f}(t)dt \to_p \int_a^b f(t)dt$.

18. Show that ϱ-Fréchet differentiability implies ϱ-Hadamard differentiability.

19. Suppose that a functional $\mathtt{T}$ is Gâteaux differentiable at F with a continuous differential $\mathtt{L}_F$ in the sense that $\varrho_\infty(\Delta_j, \Delta) \to 0$ implies $\mathtt{L}_F(\Delta_j) \to \mathtt{L}_F(\Delta)$. Show that ϕ_F is bounded.

20. Suppose that a functional $\mathtt{T}$ is Gâteaux differentiable at F with a bounded and continuous influence function ϕ_F. Show that the differential $\mathtt{L}_F$ is continuous in the sense described in the previous exercise.

21. Let $\mathtt{T}(G) = g(\int x dG)$ be a functional defined on $\mathcal{F}_1$, the collection of one-dimensional c.d.f.'s with finite means.
(a) Find a differentiable function g for which the functional $\mathtt{T}$ is not ϱ_∞-Hadamard differentiable at F.
(b) Show that if g is a differentiable function, then $\mathtt{T}$ is ϱ_{L_1}-Fréchet differentiable at F. (Hint: use the result in Exercise 9.)

22. In Example 5.5, show that (5.36) holds. (Hint: for $\Delta = c(G_1 - G_2)$, show that $\|\Delta\|_V \le |c|(\|G_1\|_V + \|G_2\|_V) = 2|c|$.)

23. In Example 5.5, show that ϕ_F is continuous if F is continuous.

24. In Example 5.5, show that $\mathtt{T}$ is not ϱ_∞-Fréchet differentiable at F.

25. Prove Proposition 5.1(ii).

26. Suppose that $\mathtt{T}$ is first-order and second-order ϱ-Hadamard differentiable at F. Prove (5.38).

27. Find an example of a second-order ϱ-Fréchet differentiable functional $\mathtt{T}$ that is not first-order ϱ-Hadamard differentiable.

28. Prove (5.40) and that (5.33) is satisfied if F has a finite variance.

29. Prove (iv) and (v) of Theorem 5.6.

30. Discuss which of (i)-(v) in Theorem 5.6 can be applied to each of the L-estimators in Example 5.6.

31. Obtain explicit forms of the influence functions for L-estimators in Example 5.6. Discuss which of them are bounded and continuous.

32. Provide an example in which the L-functional $\mathtt{T}$ given by (5.39) is not ϱ_∞-Hadamard differentiable at F. (Hint: consider an untrimmed J.)

33. Discuss which M-functionals defined in (i)-(vi) of Example 5.7 satisfy the conditions of Theorem 5.7.

34. In the proof of Theorem 5.7, show that $R_{2j} \to 0$.

35. Show that the second equality in (5.44) holds when ψ is Borel and bounded.

36. Show that the functional $\mathtt{T}$ in (5.46) is ϱ_∞-Hadamard differentiable at F with the differential given by (5.47). Obtain the influence function ϕ_F and show that it is bounded and continuous if F is continuous.

37. Show that the functional $\mathtt{T}$ in (5.48) is ϱ_∞-Hadamard differentiable at F with the differential given by (5.49). Obtain the influence function ϕ_F and show that it is bounded and continuous if $F(y,\infty)$ and $F(\infty,z)$ are continuous.

38. Let F be a continuous c.d.f. on $\mathcal{R}$. Suppose that F is symmetric about θ and is strictly increasing in a neighborhood of θ. Show that $\lambda_F(t)=0$ if and only if $t=\theta$, where $\lambda_F(t)$ is defined by (5.50) with a strictly increasing J satisfying $J(1-t)=-J(t)$.

39. Show that $\lambda_F(t)$ in (5.50) is differentiable at θ and $\lambda_F'(\theta)$ is equal to $-\int J'(F(x))F'(x)dF(x)$.

40. Let $\mathtt{T}(F_n)$ be an R-estimator satisfying the conditions in Theorem 5.8. Show that (5.34) holds with

$$\sigma_F^2=\int_0^1[J(t)]^2dt\bigg/\left[\int_{-\infty}^{\infty}J'(F(x))F'(x)dF(x)\right]^2.$$

41. Calculate the asymptotic relative efficiency of the Hodges-Lehmann estimator in Example 5.8 w.r.t. the sample mean based on an i.i.d. sample from F when
(a) F is the c.d.f. of $N(\mu,\sigma^2)$;
(b) F is the c.d.f. of the logistic distribution $LG(\mu,\sigma)$;
(c) F is the c.d.f. of the double exponential distribution $DE(\mu,\sigma)$;
(d) $F(x)=F_0(x-\theta)$, where $F_0(x)$ is the c.d.f. of the t-distribution t_ν with $\nu\geq 3$.

42. Let G be a c.d.f. on $\mathcal{R}$. Show that $G(x)\geq t$ if and only if $x\geq G^{-1}(t)$.

43. Show that (5.60) implies that $\hat\theta_p$ is strongly consistent for θ_p and is $\sqrt{n}$-consistent for θ_p if $F'(\theta_p-)$ and $F'(\theta_p+)$ exist.

44. Under the condition of Theorem 5.9, show that for $\rho_\epsilon=e^{-2\delta_\epsilon^2}$,

$$P\left(\sup_{m\geq n}|\hat\theta_p-\theta_p|>\epsilon\right)\leq\frac{2C\rho_\epsilon^n}{1-\rho_\epsilon},\qquad n=1,2,....$$

45. Prove that $\varphi_n(t)$ in (5.62) is the Lebesgue p.d.f. of the pth sample quantile $\hat\theta_p$ when F has the Lebesgue p.d.f. f, by
(a) differentiating the c.d.f. of $\hat\theta_p$ in (5.61);
(b) using result (5.59) and the result in Example 2.9.

46. Let $X_1,...,X_n$ be i.i.d. random variables from F with a finite mean. Show that $\hat\theta_p$ has a finite jth moment for sufficiently large n, $j=1,2,....$

47. Prove Theorem 5.10(i).

48. Suppose that a c.d.f. F has a Lebesgue p.d.f. f. Using the p.d.f. in (5.62) and Scheffé's theorem (Proposition 1.17), prove part (iv) of Theorem 5.10.

49. Let $\{k_n\}$ be a sequence of integers satisfying $k_n/n = p + o(n^{-1/2})$ with $p \in (0,1)$, and let $X_1, ..., X_n$ be i.i.d. random variables from a c.d.f. F with $F'(\theta_p) > 0$. Show that
$$\sqrt{n}(X_{(k_n)} - \theta_p) \to_d N(0, p(1-p)/[F'(\theta_p)]^2).$$

50. In the proof of Theorem 5.11, prove (5.65), (5.68), and inequality (5.67).

51. Prove Corollary 5.1.

52. Prove the claim in Example 5.9.

53. Let $T(G) = G^{-1}(p)$ be the pth quantile functional. Suppose that F has a positive derivative F' in a neighborhood of $\theta = F^{-1}(p)$. Show that T is Gâteaux differentiable at F and obtain the influence function $\phi_F(x)$.

54. Let $X_1, ..., X_n$ be i.i.d. from the Cauchy distribution $C(0,1)$.
(a) Show that $E(X_{(j)})^2 < \infty$ if and only if $3 \le j \le n-2$.
(b) Show that $E(\hat{\theta}_{0.5})^2 < \infty$ for $n \ge 5$.

55. Suppose that F is the c.d.f. of the uniform distribution $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, $\theta \in \mathcal{R}$. Obtain the asymptotic relative efficiency of the sample median w.r.t. the sample mean, based on an i.i.d. sample of size n from F.

56. Suppose that $F(x) = F_0(x - \theta)$ and F_0 is the c.d.f. of the Cauchy distribution $C(0,1)$ truncated at c and $-c$, i.e., F_0 has the Lebesgue p.d.f. $(1+x^2)^{-1} I_{(-c,c)}(x) / \int_{-c}^{c} (1+x^2)^{-1} dt$. Obtain the asymptotic relative efficiency of the sample median w.r.t. the sample mean, based on an i.i.d. sample of size n from F.

57. Show that $\bar{X}_\alpha$ in (5.70) is the L-estimator corresponding to the J function given in Example 5.6(iii) with $\beta = 1 - \alpha$.

58. Let $X_1, ..., X_n$ be i.i.d. random variables from F, where F is symmetric about θ.
(a) Show that $X_{(j)} - \theta$ and $\theta - X_{(n-j+1)}$ have the same distribution.
(b) Show that $\sum_{j=1}^n w_j X_{(j)}$ has a c.d.f. symmetric about θ, if w_i's are constants satisfying $\sum_{i=1}^n w_i = 1$ and $w_j = w_{n-j+1}$ for all j.
(c) Show that the trimmed sample mean $\bar{X}_\alpha$ has a c.d.f. symmetric about θ.

59. Under the conditions in one of (i)-(iii) of Theorem 5.6, show that (5.34) holds for $\mathtt{T}(F_n)$ with σ_F^2 given by (5.72), if $\sigma_F^2 < \infty$.

60. Prove (5.71) under the assumed conditions.

61. For the functional $\mathtt{T}$ given by (5.39), show that $\mathtt{T}(F) = \theta$ if F is symmetric about θ and J is symmetric about $\frac{1}{2}$.

62. Suppose that F is the double exponential distribution $DE(\theta, 1)$, where $\theta \in \mathcal{R}$. Obtain the asymptotic relative efficiency of the trimmed sample mean $\bar{X}_\alpha$ w.r.t. the sample mean, based on an i.i.d. sample of size n from F.

63. Show that the method of moments in §3.5.2 is a special case of the GEE method.

64. Let $\ell(\theta, \xi)$ be a likelihood. Show that a maximum profile likelihood estimator $\hat{\theta}$ of θ is an MLE if $\xi(\theta)$, the maximum of $\sup_\xi \ell(\theta, \xi)$ for a fixed θ, does not depend on θ.

65. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$. Derive the profile likelihood function for μ or σ^2. Discuss in each case whether the maximum profile likelihood estimator is the same as the MLE.

66. Complete the proof of Proposition 5.4.

67. In the proof of Lemma 5.3, show that the probability in (5.89) is bounded by ϵ.

68. In Example 5.11, show that ψ_i's satisfy the conditions of Lemma 5.3 if Θ is compact and $\sup_i \|Z_i\| < \infty$.

69. In the proof of Proposition 5.5, show that $\{\Delta_n(\gamma)\}$ is equicontinuous on any open subset of Θ.

70. Prove Proposition 5.6.

71. Prove the claim in Example 5.12.

72. Prove the claims in Example 5.13.

73. For Huber's M-estimator discussed in Example 5.13, obtain a formula for $e(F)$, the asymptotic relative efficiency of $\hat{\theta}_n$ w.r.t. $\bar{X}$, when F is given by (5.69). Show that $\lim_{\tau\to\infty} e(F) = \infty$. Find the value of $e(F)$ when $\epsilon = 0$, $\sigma = 1$, and $C = 1.5$.

74. Consider the ψ function in Example 5.7(ii). Show that under some conditions on F, ψ satisfies the conditions given in Theorem 5.13(i) or (ii). Obtain σ_F^2 in (5.93) in this case.

75. In the proof of Theorem 5.14, show that
(a) (5.96) holds;
(b) (5.98) holds;
(c) (5.99) implies (5.97). (Hint: use Theorem 1.9(iii).)

76. Prove the claim in Example 5.14, assuming some necessary moment conditions.

77. Derive the asymptotic distribution of the MQLE (the GEE estimator based on (5.84)), assuming that $X_i = (X_{i1}, ..., X_{id_i})$, $E(X_{it}) = me^{\eta_i}/(1+e^{\eta_i})$, $\mathrm{Var}(X_{it}) = m\phi_i e^{\eta_i}/(1+e^{\eta_i})^2$, and (4.57) holds with $g(t) = \log \frac{t}{1-t}$.

78. Repeat the previous exercise under the assumption that $E(X_{it}) = e^{\eta_i}$, $\mathrm{Var}(X_{it}) = \phi_i e^{\eta_i}$, and (4.57) holds with $g(t) = \log t$ or $g(t) = 2\sqrt{t}$.

79. In Theorem 5.14, show that result (5.94) still holds if $\tilde{R}_i$ is replaced by an estimator $\hat{R}_i$ satisfying $\max_{i\le n} \|\hat{R}_i - U_i\| = o_p(1)$, where U_i's are correlation matrices.

80. Suppose that $X_1, ..., X_n$ are independent (not necessarily identically distributed) random d-vectors with $E(X_i) = \mu$ for all i. Suppose also that $\sup_i E\|X_i\|^{2+\delta} < \infty$ for some $\delta > 0$. Let $\mu = E(X_1)$, $\theta = g(\mu)$, and $\hat{\theta}_n = g(\bar{X})$. Show that
(a) (5.100) holds with $V_n = n^{-2}\nabla g(\mu) \sum_{i=1}^n \mathrm{Var}(X_i)[\nabla g(\mu)]^\tau$;
(b) $\hat{V}_n$ in (5.102) is consistent for V_n in part (a).

81. Consider the ratio estimator in Example 3.21. Derive the estimator $\hat{V}_n$ given by (5.102) and show that $\hat{V}_n$ is consistent for the asymptotic variance of the ratio estimator.

82. Derive a consistent variance estimator for $\hat{R}(t)$ in Example 3.23.

83. Prove the claims in Example 5.16.

84. Derive a consistent variance estimator for a U-statistic satisfying the conditions in Theorem 3.5(i).

85. Derive a consistent variance estimator for Huber's M-estimator discussed in Example 5.13.

86. Assume the conditions in Theorem 5.8. Let $r \in (0, \frac{1}{2})$.
(a) Show that $n^r \lambda_F(\mathtt{T}(F_n) + n^{-r}) \to_p \lambda_F(\mathtt{T}(F))$.
(b) Show that $n^r[\lambda_{F_n}(\mathtt{T}(F_n) + n^{-r}) - \lambda_F(\mathtt{T}(F_n) + n^{-r})] \to_p 0$.
(c) Derive a consistent estimator of the asymptotic variance of $\mathtt{T}(F_n)$, using the results in (a) and (b).

87. Prove Theorem 5.16.

88. Let $X_1, ..., X_n$ be random variables and $\hat{\theta} = \bar{X}^2$. Show that the jackknife estimator in (5.107) equals $\frac{4\bar{X}^2\hat{c}_2}{n-1} - \frac{4\bar{X}\hat{c}_3}{(n-1)^2} + \frac{\hat{c}_4-\hat{c}_2^2}{(n-1)^3}$, where $\hat{c}_j$'s are the sample central moments defined by (3.57).

89. Prove (5.109).

90. In the proof of Theorem 5.18, prove (5.111).

91. Show that $\hat{\theta}_{-i}$'s in (5.113) satisfy (5.112), under the conditions of Theorem 5.14.

92. Prove Theorem 5.19.

93. Prove (5.114).

94. Let $X_1, ..., X_n$ be random variables and $\hat{\theta} = \bar{X}^2$. Show that the bootstrap variance estimator based on i.i.d. X_i^*'s from F_n is equal to $\hat{V}_B = \frac{4\bar{X}^2\hat{c}_2}{n} + \frac{4\bar{X}\hat{c}_3}{n^2} + \frac{\hat{c}_4}{n^3}$, where $\hat{c}_j$'s are the sample central moments defined by (3.57).

95. Let $X_1, ..., X_n$ be i.i.d. from a Lebesgue p.d.f. $\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$ on $\mathcal{R}$, where f is known. Let $H_n(t) = P(\sqrt{n}(\bar{X} - \mu)/S \leq t)$ and $\hat{H}_B(t) = P(\sqrt{n}(\bar{X}^* - \bar{X})/S^* \leq t|X)$ be the bootstrap estimator of H_n, where S^2 is the sample variance, X_i^*'s are i.i.d. from $\frac{1}{s}f\left(\frac{x-\bar{x}}{s}\right)$, given $\bar{X} = \bar{x}$ and $S = s$, and S^* is the bootstrap analogue of S. Show that $\hat{H}_B \equiv H_n$.

96. Let G, G_1, G_2,..., be c.d.f.'s on $\mathcal{R}$. Suppose that $\varrho_\infty(G_j, G) \to 0$ as $j \to \infty$ and $G'(x)$ exists and is positive for all $x \in \mathcal{R}$. Show that $G_j^{-1}(p) \to G^{-1}(p)$ for any $p \in (0, 1)$.

97. Let $X_1, ..., X_n$ be i.i.d. from a c.d.f. F on $\mathcal{R}^d$ with a finite $\mathrm{Var}(X_1)$. Let $X_1^*, ..., X_n^*$ be i.i.d. from the empirical c.d.f. F_n. Show that for almost all given sequences $X_1, X_2, ...$, $\sqrt{n}(\bar{X}^* - \bar{X}) \to_d N(0, \mathrm{Var}(X_1))$. (Hint: verify Lindeberg's condition.)

98. Let $X_1, ..., X_n$ be i.i.d. from a c.d.f. F on $\mathcal{R}^d$, $X_1^*, ..., X_n^*$ be i.i.d. from the empirical c.d.f. F_n, and let F_n^* be the empirical c.d.f. based on X_i^*'s. Using DKW's inequality (Lemma 5.1), show that
(a) $\varrho_\infty(F_n^*, F) \to_{a.s.} 0$;
(b) $\varrho_\infty(F_n^*, F) = O_p(n^{-1/2})$;
(c) $\varrho_{L_p}(F_n^*, F) = O_p(n^{-1/2})$, under the condition in Theorem 5.20(ii).

99. Using the results from the previous two exercises, prove Theorem 5.20(ii).

100. Under the conditions in Theorem 5.11, establish a Bahadur's representation for the bootstrap sample quantile $\hat{\theta}_p^*$.

Chapter 6

Hypothesis Tests

A general theory of testing hypotheses is presented in this chapter. Let X be a sample from a population P in $\mathcal{P}$, a family of populations. Based on the observed X, we test a given hypothesis $H_0 : P \in \mathcal{P}_0$ versus $H_1 : P \in \mathcal{P}_1$, where $\mathcal{P}_0$ and $\mathcal{P}_1$ are two disjoint subsets of $\mathcal{P}$ and $\mathcal{P}_0 \cup \mathcal{P}_1 = \mathcal{P}$. Notational conventions and basic concepts (such as two types of errors, significance levels, and sizes) given in Example 2.20 and §2.4.2 are used in this chapter.

6.1 UMP Tests

A test for a hypothesis is a statistic $T(X)$ taking values in $[0, 1]$. When $X = x$ is observed, we reject H_0 with probability $T(x)$ and accept H_0 with probability $1-T(x)$. If $T(X) = 1$ or 0 a.s. $\mathcal{P}$, then $T(X)$ is a nonrandomized test. Otherwise $T(X)$ is a randomized test. For a given test $T(X)$, the *power function* of $T(X)$ is defined to be

$$\beta_T(P) = E[T(X)], \qquad P \in \mathcal{P}, \tag{6.1}$$

which is the type I error probability of $T(X)$ when $P \in \mathcal{P}_0$ and one minus the type II error probability of $T(X)$ when $P \in \mathcal{P}_1$.

As we discussed in §2.4.2, with a sample of a fixed size, we are not able to minimize two error probabilities simultaneously. Our approach involves maximizing the power $\beta_T(P)$ over all $P \in \mathcal{P}_1$ (i.e., minimizing the type II error probability) and over all tests T satisfying

$$\sup_{P \in \mathcal{P}_0} \beta_T(P) \leq \alpha, \tag{6.2}$$

where $\alpha \in [0, 1]$ is a given level of significance.

Definition 6.1. A test T_* of size α is a *uniformly most powerful* (UMP) test if and only if $\beta_{T_*}(P) \geq \beta_T(P)$ for all $P \in \mathcal{P}_1$ and T of level α. ∎

If $U(X)$ is a sufficient statistic for $P \in \mathcal{P}$, then for any test $T(X)$, $E(T|U)$ has the same power function as T and, therefore, to find a UMP test we may consider tests that are functions of U only.

The existence and characteristics of UMP tests are studied in this section.

6.1.1 The Neyman-Pearson lemma

A hypothesis H_0 (or H_1) is said to be *simple* if and only if $\mathcal{P}_0$ (or $\mathcal{P}_1$) contains exactly one population. The following useful result, which has already been used once in the proof of Theorem 4.16, provides the form of UMP tests when *both* H_0 and H_1 are simple.

Theorem 6.1. (The Neyman-Pearson lemma). Suppose that $\mathcal{P}_0 = \{P_0\}$ and $\mathcal{P}_1 = \{P_1\}$. Let f_j be the p.d.f. of P_j w.r.t. a σ-finite measure ν (e.g., $\nu = P_0 + P_1$), $j = 0, 1$.
(i) (Existence of a UMP test). For every α, there exists a UMP test of size α, which is equal to

$$T_*(X) = \begin{cases} 1 & f_1(X) > cf_0(X) \\ \gamma & f_1(X) = cf_0(X) \\ 0 & f_1(X) < cf_0(X), \end{cases} \tag{6.3}$$

where $\gamma \in (0, 1)$ and $c \geq 0$ are some constants chosen so that $E[T_*(X)] = \alpha$ when $P = P_0$ ($c = \infty$ is allowed).
(ii) (Uniqueness). If T_{**} is a UMP test of size α, then

$$T_{**}(X) = \begin{cases} 1 & f_1(X) > cf_0(X) \\ 0 & f_1(X) < cf_0(X) \end{cases} \quad \text{a.s. } \mathcal{P}. \tag{6.4}$$

Proof. The proof for the case of $\alpha = 0$ or 1 is left as an exercise. Assume now that $0 < \alpha < 1$.
(i) We first show that there exist γ and c such that $E_0[T_*(X)] = \alpha$, where E_j is the expectation w.r.t. P_j. Let $\gamma(t) = P_0(f_1(X) > tf_0(X))$. Then $\gamma(t)$ is nonincreasing, $\gamma(-\infty) = 1$, and $\gamma(\infty) = 0$ (why?). Thus, there exists a $c \in (0, \infty)$ such that $\gamma(c) \leq \alpha \leq \gamma(c-)$. Set

$$\gamma = \begin{cases} \frac{\alpha - \gamma(c)}{\gamma(c-) - \gamma(c)} & \gamma(c-) \neq \gamma(c) \\ 0 & \gamma(c-) = \gamma(c). \end{cases}$$

Note that $\gamma(c-) - \gamma(c) = P(f_1(X) = cf_0(X))$. Then

$$E_0[T_*(X)] = P_0\big(f_1(X) > cf_0(X)\big) + \gamma P_0\big(f_1(X) = cf_0(X)\big) = \alpha.$$

Next, we show that T_* in (6.3) is a UMP test. Suppose that $T(X)$ is a test satisfying $E_0[T(X)] \leq \alpha$. If $T_*(x) - T(x) > 0$, then $T_*(x) > 0$ and, therefore, $f_1(x) \geq cf_0(x)$. If $T_*(x) - T(x) < 0$, then $T_*(x) < 1$ and, therefore, $f_1(x) \leq cf_0(x)$. In any case, $[T_*(x) - T(x)][f_1(x) - cf_0(x)] \geq 0$ and, therefore,

$$\int [T_*(x) - T(x)][f_1(x) - cf_0(x)]d\nu \geq 0,$$

i.e.,

$$\int [T_*(x) - T(x)]f_1(x)d\nu \geq c\int [T_*(x) - T(x)]f_0(x)d\nu. \tag{6.5}$$

The left-hand side of (6.5) is $E_1[T_*(X)] - E_1[T(X)]$ and the right-hand side of (6.5) is $c\{E_0[T_*(X)] - E_0[T(X)]\} = c\{\alpha - E_0[T(X)]\} \geq 0$. This proves the result in (i).
(ii) Let $T_{**}(X)$ be a UMP test of size α. Define

$$A = \{x : T_*(x) \neq T_{**}(x), \quad f_1(x) \neq cf_0(x)\}.$$

Then $[T_*(x) - T_{**}(x)][f_1(x) - cf_0(x)] > 0$ when $x \in A$ and $= 0$ when $x \in A^c$, and

$$\int [T_*(x) - T_{**}(x)][f_1(x) - cf_0(x)]d\nu = 0,$$

since both T_* and T_{**} are UMP tests of size α. By Proposition 1.6(ii), $\nu(A) = 0$. This proves (6.4). ∎

Theorem 6.1 shows that when both H_0 and H_1 are simple, there exists a UMP test that can be determined by (6.4) uniquely (a.s. $\mathcal{P}$) except on the set $B = \{x : f_1(x) = cf_0(x)\}$. If $\nu(B) = 0$, then we have a unique nonrandomized UMP test; otherwise UMP tests are randomized on the set B and the randomization is necessary for UMP tests to have the given size α; furthermore, we can always choose a UMP test that is constant on B.

Example 6.1. Suppose that X is a sample of size 1, $\mathcal{P}_0 = \{P_0\}$ and $\mathcal{P}_1 = \{P_1\}$, where P_0 is $N(0,1)$ and P_1 is the double exponential distribution $DE(0,2)$ with the p.d.f. $4^{-1}e^{-|x|/2}$. Since $P(f_1(X) = cf_0(X)) = 0$, there is a unique nonrandomized UMP test. From (6.3), the UMP test $T_*(x) = 1$ if and only if $\frac{\pi}{8}e^{x^2-|x|} > c^2$ for some $c > 0$, which is equivalent to $|x| > t$ or $|x| < 1 - t$ for some $t > \frac{1}{2}$. Suppose that $\alpha < \frac{1}{4}$. To determine t, we use

$$\alpha = E_0[T_*(X)] = P_0(|X| > t) + P_0(|X| < 1 - t). \tag{6.6}$$

If $t \leq 1$, then $P_0(|X| > t) \geq P_0(|X| > 1) = 0.3374 > \alpha$. Hence t should be larger than 1 and (6.6) becomes

$$\alpha = P_0(|X| > t) = \Phi(-t) + 1 - \Phi(t).$$

Thus, $t = \Phi^{-1}(1-\alpha/2)$ and $T_*(X) = I_{(t,\infty)}(|X|)$. Note that it is not necessary to find out what c is.

Intuitively, the reason why the UMP test in this example rejects H_0 when $|X|$ is large is that the probability of getting a large $|X|$ is much higher under H_1 (i.e., P is the double exponential distribution $DE(0,2)$).

The power of T_* when $P \in \mathcal{P}_1$ is

$$E_1[T_*(X)] = P_1(|X| > t) = 1 - \frac{1}{4}\int_{-t}^{t} e^{-|x|/2}dx = e^{-t/2}. \quad \blacksquare$$

Example 6.2. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $p = P(X_1 = 1)$. Suppose that $H_0 : p = p_0$ and $H_1 : p = p_1$, where $0 < p_0 < p_1 < 1$. By Theorem 6.1, a UMP test of size α is

$$T_*(Y) = \begin{cases} 1 & \lambda(Y) > c \\ \gamma & \lambda(Y) = c \\ 0 & \lambda(Y) < c, \end{cases}$$

where $Y = \sum_{i=1}^n X_i$ and

$$\lambda(Y) = \left(\frac{p_1}{p_0}\right)^Y \left(\frac{1-p_1}{1-p_0}\right)^{n-Y}.$$

Since $\lambda(Y)$ is increasing in Y, there is an integer $m > 0$ such that

$$T_*(Y) = \begin{cases} 1 & Y > m \\ \gamma & Y = m \\ 0 & Y < m, \end{cases}$$

where m and γ satisfy $\alpha = E_0[T_*(Y)] = P_0(Y > m) + \gamma P_0(Y = m)$. Since Y has the binomial distribution $Bi(p,n)$, we can determine m and γ from

$$\alpha = \sum_{j=m+1}^{n} \binom{n}{j} p_0^j (1-p_0)^{n-j} + \gamma \binom{n}{m} p_0^m (1-p_0)^{n-m}. \tag{6.7}$$

Unless

$$\alpha = \sum_{j=m+1}^{n} \binom{n}{j} p_0^j (1-p_0)^{n-j}$$

for some integer m, in which case we can choose $\gamma = 0$, the UMP test T_* is a randomized test. $\blacksquare$

An interesting phenomenon in Example 6.2 is that the UMP test T_* does not depend on p_1. In such a case T_* is in fact a UMP test for testing $H_0 : p = p_0$ versus $H_1 : p > p_0$.

Lemma 6.1. Suppose that there is a test T_* of size α such that for every $P_1 \in \mathcal{P}_1$, T_* is UMP for testing H_0 versus the hypothesis $P = P_1$. Then T_* is UMP for testing H_0 versus H_1. ∎

The proof of this lemma is left as an exercise.

We conclude this section with the following generalized Neyman-Pearson lemma. Its proof is left to the reader.

Proposition 6.1. Let $f_1, ..., f_{m+1}$ be real-valued functions on $\mathcal{R}^p$ that are integrable w.r.t. a σ-finite measure ν. For given constants $t_1, ..., t_m$, let $\mathcal{T}$ be the class of Borel functions ϕ (from $\mathcal{R}^p$ to $[0, 1]$) satisfying

$$\int \phi f_i d\nu \leq t_i, \quad i = 1, ..., m, \tag{6.8}$$

and $\mathcal{T}_0$ be the set of ϕ's in $\mathcal{T}$ satisfying (6.8) with all inequalities replaced by equalities. If there are constants $c_1, ..., c_m$ such that

$$\phi_*(x) = \begin{cases} 1 & f_{m+1}(x) > c_1 f_1(x) + \cdots + c_m f_m(x) \\ 0 & f_{m+1}(x) < c_1 f_1(x) + \cdots + c_m f_m(x) \end{cases} \tag{6.9}$$

is a member of $\mathcal{T}_0$, then ϕ_* maximizes $\int \phi f_{m+1} d\nu$ over $\phi \in \mathcal{T}_0$. If $c_i \geq 0$ for all i, then ϕ_* maximizes $\int \phi f_{m+1} d\nu$ over $\phi \in \mathcal{T}$. ∎

The existence of constants c_i's in (6.9) is considered in the following lemma whose proof can be found in Lehmann (1986, pp. 97-99).

Lemma 6.2. Let $f_1, ..., f_m$ and ν be given by Proposition 6.1. Then the set $M = \left\{ (\int \phi f_1 d\nu, ..., \int \phi f_m d\nu) : \ \phi \text{ is from } \mathcal{R}^p \text{ to } [0, 1] \right\}$ is convex and closed. If $(t_1, ..., t_m)$ is an interior point of M, then there exist constants $c_1, ..., c_m$ such that the function defined by (6.9) is in $\mathcal{T}_0$. ∎

6.1.2 Monotone likelihood ratio

The case of both H_0 and H_1 are simple is mainly of theoretical interest. If a hypothesis is not simple, it is called composite. As we discussed in §6.1.1, UMP tests for composite H_1 exist in the problem discussed in Example 6.2. We now extend this result to a class of parametric problems in which the likelihood functions have a special property.

Definition 6.2. Suppose that the distribution of X is in $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, a parametric family indexed by a real-valued θ, and that $\mathcal{P}$ is dominated by a σ-finite measure ν. Let $f_\theta = dP_\theta/d\nu$. The family $\mathcal{P}$ is said to have *monotone likelihood ratio* in $Y(X)$ (a real-valued statistic) if and only if, for

any $\theta_1 < \theta_2$, $f_{\theta_2}(x)/f_{\theta_1}(x)$ is a nondecreasing function of $Y(x)$ for values x at which at least one of $f_{\theta_1}(x)$ and $f_{\theta_2}(x)$ is positive. ∎

The following lemma states a useful result for a family with monotone likelihood ratio.

Lemma 6.3. Suppose that the distribution of X is in a parametric family $\mathcal{P}$ indexed by a real-valued θ and that $\mathcal{P}$ has monotone likelihood ratio in $Y(X)$. If ψ is a nondecreasing function of Y, then $g(\theta) = E[\psi(Y)]$ is a nondecreasing function of θ.

Proof. Let $\theta_1 < \theta_2$, $A = \{x : f_{\theta_1}(x) > f_{\theta_2}(x)\}$, $a = \sup_{x\in A} \psi(Y(x))$, $B = \{x : f_{\theta_1}(x) < f_{\theta_2}(x)\}$, and $b = \inf_{x\in B} \psi(Y(x))$. Since $\mathcal{P}$ has monotone likelihood ratio in $Y(X)$ and ψ is nondecreasing in Y, $b \geq a$. Then the result follows from

$$\begin{aligned} g(\theta_2) - g(\theta_1) &= \int \psi(Y(x))(f_{\theta_2} - f_{\theta_1})(x)d\nu \\ &\geq a \int_A (f_{\theta_2} - f_{\theta_1})(x)d\nu + b \int_B (f_{\theta_2} - f_{\theta_1})(x)d\nu \\ &= (b-a) \int_B (f_{\theta_2} - f_{\theta_1})(x)d\nu \\ &\geq 0. \quad ∎ \end{aligned}$$

Before discussing UMP tests in families with monotone likelihood ratio, let us consider some examples of such families.

Example 6.3. Let θ be real-valued and $\eta(\theta)$ be a nondecreasing function of θ. Then the one-parameter exponential family with

$$f_\theta(x) = \exp\{\eta(\theta)Y(x) - \xi(\theta)\}h(x) \tag{6.10}$$

has monotone likelihood ratio in $Y(X)$. From Tables 1.1-1.2 (§1.3.1), this includes the binomial family $\{Bi(\theta, r)\}$, the Poisson family $\{P(\theta)\}$, the negative binomial family $\{NB(\theta, r)\}$, the log-distribution family $\{L(\theta)\}$, the normal family $\{N(\theta, c^2)\}$ or $\{N(c, \theta)\}$, the exponential family $\{E(c, \theta)\}$, the gamma family $\{\Gamma(\theta, c)\}$ or $\{\Gamma(c, \theta)\}$, the beta family $\{B(\theta, c)\}$ or $\{B(c, \theta)\}$, and the double exponential family $\{DE(c, \theta)\}$, where r or c is known. ∎

Example 6.4. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on $(0, \theta)$, where $\theta > 0$. The Lebesgue p.d.f. of $X = (X_1, ..., X_n)$ is $f_\theta(x) = \theta^{-n} I_{(0,\theta)}(x_{(n)})$, where $x_{(n)}$ is the value of the largest order statistic $X_{(n)}$. For $\theta_1 < \theta_2$,

$$\frac{f_{\theta_2}(x)}{f_{\theta_1}(x)} = \frac{\theta_1^n}{\theta_2^n} \frac{I_{(0,\theta_2)}(x_{(n)})}{I_{(0,\theta_1)}(x_{(n)})},$$

which is a nondecreasing function of $x_{(n)}$ for x's at which at least one of $f_{\theta_1}(x)$ and $f_{\theta_2}(x)$ is positive, i.e., $x_{(n)} < \theta_2$. Hence the family of distributions of X has monotone likelihood ratio in $X_{(n)}$. ∎

Example 6.5. The following families have monotone likelihood ratio:
(a) the double exponential distribution family $\{DE(\theta, c)\}$ with a known c;
(b) the exponential distribution family $\{E(\theta, c)\}$ with a known c;
(c) the logistic distribution family $\{LG(\theta, c)\}$ with a known c;
(d) the uniform distribution family $\{U(\theta, \theta + 1)\}$;
(e) the hypergeometric distribution family $\{HG(r, \theta, N - \theta)\}$ with known r and N (Table 1.1, page 18).

An example of a family that does not have monotone likelihood ratio is the Cauchy distribution family $\{C(\theta, c)\}$ with a known c. ∎

Hypotheses of the form $H_0 : \theta \le \theta_0$ (or $H_0 : \theta \ge \theta_0$) versus $H_1 : \theta > \theta_0$ (or $H_1 : \theta < \theta_0$) are called *one-sided* hypotheses for any given constant θ_0. The following result provides UMP tests for testing one-sided hypotheses when the distribution of X is in a parametric family with monotone likelihood ratio.

Theorem 6.2. Suppose that the distribution of X is in a parametric family $\mathcal{P}$ indexed by a real-valued θ and that $\mathcal{P}$ has monotone likelihood ratio in $Y(X)$. Consider the problem of testing $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$, where θ_0 is a given constant.
(i) There exists a UMP test of size α, which is given by

$$T_*(X) = \begin{cases} 1 & Y(X) > c \\ \gamma & Y(X) = c \\ 0 & Y(X) < c, \end{cases} \tag{6.11}$$

where c and γ are determined by $\beta_{T_*}(\theta_0) = \alpha$, and $\beta_T(\theta)$ is the power function of a test T.
(ii) The power function $\beta_{T_*}(\theta) = E[T_*(X)]$ is strictly increasing for all θ's for which $0 < \beta_{T_*}(\theta) < 1$.
(iii) For any $\theta < \theta_0$, T_* minimizes $\beta_T(\theta)$ (the type I error probability of T) among all tests T satisfying $\beta_T(\theta_0) = \alpha$.
(iv) For any θ_1, T_* is UMP for testing $H_0 : \theta \le \theta_1$ versus $H_1 : \theta > \theta_1$, with size $\beta_{T_*}(\theta_1)$.
Proof. (i) Consider the hypotheses $\theta = \theta_0$ versus $\theta = \theta_1$ with any $\theta_1 > \theta_0$. From Theorem 6.1, a UMP test is given by (6.3) with $f_j = f_{\theta_j}$, $j = 0, 1$. Since $\mathcal{P}$ has monotone likelihood ratio in $Y(X)$, this UMP test is the same as T_* in (6.11) (with a different c), as long as γ and c satisfy $\beta_{T_*}(\theta_0) = \alpha$. Since T_* does not depend on θ_1, it follows from Lemma 6.1 that T_* is UMP for testing the hypotheses $\theta = \theta_0$ versus H_1.

Note that if T_* is UMP for testing $\theta = \theta_0$ versus H_1, then it is UMP for testing H_0 versus H_1 provided that $\beta_{T_*}(\theta) \le \alpha$ for all $\theta \le \theta_0$, i.e., the size of T_* is α. But this follows from Lemma 6.3, i.e., $\beta_{T_*}(\theta)$ is nondecreasing in θ. This proves (i).
(ii) See Exercise 2 in §6.6.
(iii) The result can be proved using Theorem 6.1 with all inequalities reversed.
(iv) The proof for (iv) is similar to that of (i). ∎

By reversing inequalities throughout, we can obtain UMP tests for testing $H_0 : \theta \ge \theta_0$ versus $H_1 : \theta < \theta_0$.

A major application of Theorem 6.2 is to problems with one-parameter exponential families.

Corollary 6.1. Suppose that X has the p.d.f. given by (6.10) w.r.t. a σ-finite measure ν, where η is a strictly monotone function of θ. If η is increasing, then T_* given by (6.11) is UMP for testing $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$, where γ and c are determined by $\beta_{T_*}(\theta_0) = \alpha$. If η is decreasing or $H_0 : \theta \ge \theta_0$ ($H_1 : \theta < \theta_0$), the result is still valid by reversing inequalities in (6.11). ∎

Example 6.6. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathcal{R}$ and a known σ^2. Consider the hypotheses

$$H_0 : \mu \le \mu_0 \qquad \text{versus} \qquad H_1 : \mu > \mu_0,$$

where μ_0 is a fixed constant. The joint distribution of X is of the form (6.10) with $Y(X) = \bar{X}$ and $\eta(\mu) = n\mu/\sigma^2$. By Corollary 6.1 and the fact that $\bar{X}$ is $N(\mu, \sigma^2/n)$, the UMP test is $T_*(X) = I_{(c_\alpha,\infty)}(\bar{X})$, where $c_\alpha = \sigma\Phi^{-1}(1-\alpha)/\sqrt{n} + \mu_0$ (see also Example 2.28). ∎

To derive a UMP test for testing $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$ when X has the p.d.f. (6.10), it is essential to know the distribution of $Y(X)$. Typically, a nonrandomized test can be obtained if the distribution of Y is continuous; otherwise UMP tests are randomized.

Example 6.7. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $p = P(X_1 = 1)$. The distribution of X is of the form (6.10) with $Y(X) = \sum_{i=1}^n X_i$ and $\eta(p) = \log \frac{p}{1-p}$. Note that $\eta(p)$ is a strictly increasing function of p. By Corollary 6.1, a UMP test for $H_0 : p \le p_0$ versus $H_1 : p > p_0$ is given by (6.11), where c and γ are determined by (6.7) with $c = m$. ∎

Example 6.8. Let $X_1, ..., X_n$ be i.i.d. random variables from the Poisson distribution $P(\theta)$ with an unknown $\theta > 0$. The distribution of X is of the

form (6.10) with $Y(X) = \sum_{i=1}^n X_i$ and $\eta(\theta) = \log\theta$. Note that Y has the Poisson distribution $P(n\theta)$. By Corollary 6.1, a UMP test for $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$ is given by (6.11) with c and γ determined by

$$\alpha = \sum_{j=c+1}^{\infty} \frac{e^{n\theta_0}(n\theta_0)^j}{j!} + \gamma \frac{e^{n\theta_0}(n\theta_0)^c}{c!}. \quad \blacksquare$$

Example 6.9. Let $X_1, ..., X_n$ be i.i.d. random variables from the uniform distribution $U(0,\theta)$, $\theta > 0$. Consider the hypotheses $H_0 : \theta \le \theta_0$ and $H_1 : \theta > \theta_0$. Since the distribution of X is in a family with monotone likelihood ratio in $X_{(n)}$ (Example 6.4), by Theorem 6.2, a UMP test is of the form (6.11). Since $X_{(n)}$ has the Lebesgue p.d.f. $n\theta^{-n}x^{n-1}I_{(0,\theta)}(x)$, the UMP test in (6.11) is nonrandomized and c is determined by

$$\alpha = \beta_{T_*}(\theta_0) = \frac{n}{\theta_0^n}\int_c^{\theta_0} x^{n-1}dx = 1 - \frac{c^n}{\theta_0^n}.$$

Hence $c = \theta_0(1-\alpha)^{1/n}$. The power function of T_* when $\theta > \theta_0$ is

$$\beta_{T_*}(\theta) = \frac{n}{\theta^n}\int_c^{\theta} x^{n-1}dx = 1 - \frac{\theta_0^n(1-\alpha)}{\theta^n}.$$

In this problem, however, UMP tests are not unique. It can be shown (exercise) that the following test is also UMP with size α:

$$T(X) = \begin{cases} 1 & X_{(n)} > \theta_0 \\ \alpha & X_{(n)} \le \theta_0. \end{cases} \quad \blacksquare$$

6.1.3 UMP tests for two-sided hypotheses

The following hypotheses are called two-sided hypotheses:

$$H_0 : \ \theta \le \theta_1 \text{ or } \theta \ge \theta_2 \quad \text{versus} \quad H_1 : \ \theta_1 < \theta < \theta_2, \tag{6.12}$$

$$H_0 : \ \theta_1 \le \theta \le \theta_2 \quad \text{versus} \quad H_1 : \ \theta < \theta_1 \text{ or } \theta > \theta_2, \tag{6.13}$$

$$H_0 : \ \theta = \theta_0 \quad \text{versus} \quad H_1 : \ \theta \ne \theta_0, \tag{6.14}$$

where $\theta_1 < \theta_2$ and θ_0 are given constants.

Theorem 6.3. Suppose that X has the p.d.f. given by (6.10) w.r.t. a σ-finite measure ν, where η is a strictly increasing function of θ.
(i) For testing hypotheses (6.12), a UMP test of size α is

$$T_*(X) = \begin{cases} 1 & c_1 < Y(X) < c_2 \\ \gamma_i & Y(X) = c_i, \ i = 1,2 \\ 0 & Y(X) < c_1 \text{ or } Y(X) > c_2, \end{cases} \tag{6.15}$$

where $c_1 < c_2$ and γ_i's are determined by

$$\beta_{T_*}(\theta_1) = \beta_{T_*}(\theta_2) = \alpha. \tag{6.16}$$

(ii) The test defined by (6.15) minimizes $\beta_T(\theta)$ over all $\theta < \theta_1$, $\theta > \theta_2$, and T satisfying $\beta_T(\theta_1) = \beta_T(\theta_2) = \alpha$.
(iii) If T_* and T_{**} are two tests satisfying (6.15) and $\beta_{T_*}(\theta_1) = \beta_{T_{**}}(\theta_1)$ and if the region $\{T_{**} = 1\}$ is to the right of $\{T_* = 1\}$, then $\beta_{T_*}(\theta) < \beta_{T_{**}}(\theta)$ for $\theta > \theta_1$ and $\beta_{T_*}(\theta) > \beta_{T_{**}}(\theta)$ for $\theta < \theta_1$. If both T_* and T_{**} satisfy (6.15) and (6.16), then $T_* = T_{**}$ a.s. $\mathcal{P}$.
Proof. (i) The distribution of Y has a p.d.f.

$$g_\theta(y) = \exp\{\eta(\theta)y - \xi(\theta)\} \tag{6.17}$$

(Theorem 2.1). Since Y is sufficient for θ, we only need to consider tests of the form $T(Y)$. Let $\theta_1 < \theta_3 < \theta_2$. Consider the problem of maximizing $\beta_T(\theta_3)$ subject to $\beta_T(\theta_1) = \beta_T(\theta_2) = \alpha$. Clearly, (α, α) is an interior point of the set of all points $(\beta_T(\theta_1), \beta_T(\theta_2))$ as T ranges over all tests of the form $T(Y)$. By (6.17), Lemma 6.2, and Proposition 6.1, there are constants a_1, a_2, b_1, and b_2 ($b_1 < 0 < b_2$) such that

$$T_*(Y) = \begin{cases} 1 & a_1 e^{b_1 Y} + a_2 e^{b_2 Y} < 1 \\ 0 & a_1 e^{b_1 Y} + a_2 e^{b_2 Y} > 1 \end{cases}$$

maximizes $\beta_T(\theta_0)$ and satisfies (6.16). Clearly a_i's cannot both be ≤ 0. If one of the a_i's is ≤ 0 and the other is > 0, then $a_1 e^{b_1 Y} + a_2 e^{b_2 Y}$ is strictly monotone and T_* is of the form (6.11), which has a strictly monotone power function (Theorem 6.2) and, therefore, cannot satisfy (6.16). Thus, both a_i's are positive and T_* is of the form (6.15). It follows from Proposition 6.1 that T_* is UMP for testing $\theta = \theta_1$ or $\theta = \theta_2$ versus $\theta = \theta_3$. Since T_* does not depend on θ_3, it follows from Lemma 6.1 that T_* is UMP for testing $\theta = \theta_1$ or $\theta = \theta_2$ versus H_1.

To show that T_* is a UMP test of size α for testing H_0 versus H_1, it remains to show that $\beta_{T_*}(\theta) \leq \alpha$ for $\theta \leq \theta_1$ or $\theta \geq \theta_2$. But this follows from part (ii) of the theorem by comparing T_* with the test $T(Y) \equiv \alpha$.
(ii) The proof is similar to that in (i) and is left as an exercise.
(iii) The first claim in (iii) follows from Lemma 6.4, since the function $T_{**} - T_*$ has a single change of sign. The second claim in (iii) follows from the first claim. ∎

Lemma 6.4. Suppose that X has a p.d.f. in $\{f_\theta(x) : \theta \in \Theta\}$, a parametric family of p.d.f.'s w.r.t. a single σ-finite measure ν on $\mathcal{R}$, where $\Theta \subset \mathcal{R}$. Suppose that this family has monotone likelihood ratio in X. Let ψ be a function with a single change of sign.
(i) There exists $\theta_0 \in \Theta$ such that $E_\theta[\psi(X)] \leq 0$ for $\theta < \theta_0$ and $E_\theta[\psi(X)] \geq 0$

for $\theta > \theta_0$, where E_θ is the expectation w.r.t. f_θ.
(ii) Suppose that $f_\theta(x) > 0$ for all x and θ, that $f_{\theta_1}(x)/f_\theta(x)$ is strictly increasing in x for $\theta < \theta_1$, and that $\nu(\{x : \psi(x) \neq 0\}) > 0$. If $E_{\theta_0}[\psi(X)] = 0$, then $E_\theta[\psi(X)] < 0$ for $\theta < \theta_0$ and $E_\theta[\psi(X)] > 0$ for $\theta > \theta_0$.
Proof. (i) Suppose that there is an $x_0 \in \mathcal{R}$ such that $\psi(x) \leq 0$ for $x < x_0$ and $\psi(x) \geq 0$ for $x > x_0$. Let $\theta_1 < \theta_2$. We first show that $E_{\theta_1}[\psi(X)] > 0$ implies $E_{\theta_2}[\psi(X)] \geq 0$. If $f_{\theta_2}(x_0)/f_{\theta_1}(x_0) = \infty$, then $f_{\theta_1}(x) = 0$ for $x \geq x_0$ and, therefore, $E_{\theta_1}[\psi(X)] \leq 0$. Hence $f_{\theta_2}(x_0)/f_{\theta_1}(x_0) = c < \infty$. Then $\psi(x) \geq 0$ on the set $A = \{x : f_{\theta_1}(x) = 0 \text{ and } f_{\theta_2}(x) > 0\}$. Thus,

$$\begin{aligned} E_{\theta_2}[\psi(X)] &\geq \int_{A^c} \psi \frac{f_{\theta_2}}{f_{\theta_1}} f_{\theta_1} d\nu \\ &\geq \int_{x<x_0} c\psi f_{\theta_1} d\nu + \int_{x\geq x_0} c\psi f_{\theta_1} d\nu \qquad (6.18) \\ &= cE_{\theta_1}[\psi(X)]. \end{aligned}$$

The result follows by letting $\theta_0 = \inf\{\theta : E_\theta[\psi(X)] > 0\}$.
(ii) Under the assumed conditions $f_{\theta_2}(x_0)/f_{\theta_1}(x_0) = c < \infty$. The result follows from the proof in (i) with θ_1 replaced by θ_0 and the fact that $\geq$ should be replaced by $>$ in (6.18) under the assumed conditions. ∎

Part (iii) of Theorem 6.3 shows that the c_i's and γ_i's are uniquely determined by (6.15) and (6.16). It also indicates how to select the c_i's and γ_i's. One can start with some trial values $c_1^{(0)}$ and $\gamma_1^{(0)}$, find $c_2^{(0)}$ and $\gamma_2^{(0)}$ such that $\beta_{T_*}(\theta_1) = \alpha$, and compute $\beta_{T_*}(\theta_2)$. If $\beta_{T_*}(\theta_2) < \alpha$, by Theorem 6.3(iii), the correct rejection region $\{T_* = 1\}$ is to the right of the one chosen so that one should try $c_1^{(1)} > c_1^{(0)}$ or $c_1^{(1)} = c_1^{(0)}$ and $\gamma_1^{(1)} < \gamma_1^{(0)}$; the converse holds if $\beta_{T_*}(\theta_2) > \alpha$.

Example 6.10. Let $X_1, ..., X_n$ be i.i.d. from $N(\theta, 1)$. By Theorem 6.3, a UMP test for testing (6.12) is $T_*(X) = I_{(c_1,c_2)}(\bar{X})$, where c_i's are determined by

$$\Phi\big(\sqrt{n}(c_2 - \theta_1)\big) - \Phi\big(\sqrt{n}(c_1 - \theta_1)\big) = \alpha$$

and

$$\Phi\big(\sqrt{n}(c_2 - \theta_2)\big) - \Phi\big(\sqrt{n}(c_1 - \theta_2)\big) = \alpha. \quad ∎$$

When the distribution of X is not given by (6.10), UMP tests for hypotheses (6.12) exist in some cases (see Exercises 15 and 24). Unfortunately, a UMP test does not exist in general for testing hypotheses (6.13) or (6.14) (Exercise 25).

6.2 UMP Unbiased Tests

When a UMP test does not exist, we may use the same approach used in estimation problems, i.e., imposing a reasonable restriction on the tests to be considered and finding optimal tests within the class of tests under the restriction. Two such types of restrictions in estimation problems are unbiasedness and invariance. We consider *unbiased tests* in this section. The class of *invariant tests* is studied in §6.3.

6.2.1 Unbiasedness and similarity

A UMP test T of size α has the property that

$$\beta_T(P) \leq \alpha, \quad P \in \mathcal{P}_0 \qquad \text{and} \qquad \beta_T(P) \geq \alpha, \quad P \in \mathcal{P}_1. \tag{6.19}$$

This means that T is at least as good as the silly test $T \equiv \alpha$. Thus, we have the following definition.

Definition 6.3. A test of $H_0 : P \in \mathcal{P}_0$ versus $H_1 : P \in \mathcal{P}_1$ is said to be unbiased if and only if (6.19) holds for some α. A test of size α is called a *uniformly most powerful unbiased* (UMPU) test if and only if it is UMP within the class of unbiased tests of level α. ∎

In a large class of problems for which a UMP test does not exist, there do exist UMPU tests.

Suppose that U is a sufficient statistic for $P \in \mathcal{P}$. Then, similar to the search for a UMP test, we need to consider functions of U only in order to find a UMPU test, since for any unbiased test $T(X)$, $E(T|U)$ is unbiased and has the same power function as T.

Throughout this section we consider the following hypotheses:

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1, \tag{6.20}$$

where $\theta = \theta(P)$ is a functional from $\mathcal{P}$ onto Θ and Θ_0 and Θ_1 are two disjoint Borel sets with $\Theta_0 \cup \Theta_1 = \Theta$. Note that $\mathcal{P}_j = \{P : \theta \in \Theta_j\}$, $j = 0, 1$. For instance, $X_1, ..., X_n$ are i.i.d. from F but we are interested in testing $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$, where $\theta = EX_1$ or the median of F.

Definition 6.4. Consider the hypotheses specified by (6.20). Let α be a given level of significance and let $\bar{\Theta}_{01}$ be the common boundary of Θ_0 and Θ_1, i.e., the set of points θ that are points or limit points of both Θ_0 and Θ_1. A test T is *similar* on $\bar{\Theta}_{01}$ if and only if

$$\beta_T(P) = \alpha, \qquad \theta \in \bar{\Theta}_{01}. \quad ∎ \tag{6.21}$$

It is more convenient to work with (6.21) than to work with (6.19) when the hypotheses are given by (6.20). Thus, the following lemma is useful. For a given test T, the power function $\beta_T(P)$ is said to be continuous in θ if and only if for any $\{\theta_j : j = 0, 1, 2, ...\} \subset \Theta$, $\theta_j \to \theta_0$ implies $\beta_T(P_j) \to \beta_T(P_0)$, where $P_j \in \mathcal{P}$ satisfying $\theta(P_j) = \theta_j$, $j = 0, 1, ...$. Note that if β_T is a function of θ, then this continuity property is simply the continuity of $\beta_T(\theta)$.

Lemma 6.5. Consider hypotheses (6.20). Suppose that for every T, $\beta_T(P)$ is continuous in θ. If T_* is UMP among all tests satisfying (6.21) and has size α, then T_* is a UMPU test.
Proof. Under the continuity assumption on β_T, the class of tests satisfying (6.21) contains the class of tests satisfying (6.19). Since T_* is uniformly at least as powerful as the test $T \equiv \alpha$, T_* is unbiased. Hence, T_* is a UMPU test. ∎

Using Lemma 6.5, we can derive a UMPU test for testing hypotheses given by (6.13) or (6.14), when X has the p.d.f. (6.10) in a one-parameter exponential family. (Note that a UMP test does not exist in these cases.) We do not provide the details here, since the results for one-parameter exponential families are special cases of those in §6.2.2 for multiparameter exponential families. To prepare for the discussion in §6.2.2, we introduce the following result that simplifies (6.21) when there is a statistic sufficient and complete for $P \in \bar{\mathcal{P}} = \{P : \theta(P) \in \bar{\Theta}_{01}\}$.

Let $U(X)$ be a sufficient statistic for $P \in \bar{\mathcal{P}}$ and let $\bar{\mathcal{P}}_U$ be the family of distributions of U as P ranges over $\bar{\mathcal{P}}$. If T is a test satisfying

$$E[T(X)|U] = \alpha \qquad \text{a.s. } \bar{\mathcal{P}}_U, \tag{6.22}$$

then

$$E[T(X)] = E\{E[T(X)|U]\} = \alpha \qquad P \in \bar{\mathcal{P}},$$

i.e., T is similar on $\bar{\Theta}_{01}$. A test satisfying (6.22) is said to have *Neyman structure* w.r.t. U. If all tests similar on $\bar{\Theta}_{01}$ have Neyman structure w.r.t. U, then working with (6.21) is the same as working with (6.22).

Lemma 6.6. Let $U(X)$ be a sufficient statistic for $P \in \bar{\mathcal{P}}$. Then a necessary and sufficient condition for all tests similar on $\bar{\Theta}_{01}$ to have Neyman structure w.r.t. U is that U is boundedly complete for $P \in \bar{\mathcal{P}}$.
Proof. (i) Suppose first that U is boundedly complete for $P \in \bar{\mathcal{P}}$. Let $T(X)$ be a test similar on $\bar{\Theta}_{01}$. Then $E[T(X) - \alpha] = 0$ for all $P \in \bar{\mathcal{P}}$. From the boundedness of $T(X)$, $E[T(X)|U]$ is bounded (Proposition 1.12). Since $E\{E[T(X)|U] - \alpha\} = E[T(X) - \alpha] = 0$ for all $P \in \bar{\mathcal{P}}$, (6.22) holds.
(ii) Suppose now that U is not boundedly complete for $P \in \bar{\mathcal{P}}$. Then there is a function h such that $|h(u)| \le C$, $E[h(U)] = 0$ for all $P \in \bar{\mathcal{P}}$, and $h(U) \neq 0$ with positive probability for some $P \in \bar{\mathcal{P}}$. Let $T(X) = \alpha + ch(U)$,

where $c = \min(\alpha, 1-\alpha)/C$. The result follows from the fact that T is a test similar on $\bar{\Theta}_{01}$ but does not have Neyman structure w.r.t. U. ∎

6.2.2 UMPU tests in exponential families

Suppose that the distribution of X is in a multiparameter natural exponential family (§2.1.3) with the following p.d.f. w.r.t. a σ-finite measure ν:

$$f_{\theta,\varphi}(x) = \exp\left\{\theta Y(x) + U(x)\varphi^\tau - \zeta(\theta,\varphi)\right\}, \tag{6.23}$$

where θ is a real-valued parameter, φ is a vector-valued parameter, and Y (real-valued) and U (vector-valued) are statistics. It follows from Theorem 2.1 that (Y, U) has the p.d.f. $\exp\{\theta y + u\varphi^\tau - \zeta(\theta,\varphi)\}$ w.r.t. some measure and, given $U = u$, the conditional distribution of Y has the p.d.f. $\exp\{\theta y\}$ w.r.t. some measure ν_u, which is also in a natural exponential family.

Theorem 6.4. Suppose that the distribution of X is in a multiparameter natural exponential family given by (6.23).
(i) For testing $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$, a UMPU test of size α is

$$T_*(Y,U) = \begin{cases} 1 & Y > c(U) \\ \gamma(U) & Y = c(U) \\ 0 & Y < c(U), \end{cases} \tag{6.24}$$

where $c(u)$ and $\gamma(u)$ are Borel functions determined by

$$E_{\theta_0}[T_*(Y,U)|U = u] = \alpha \tag{6.25}$$

for every u, and E_{θ_0} is the expectation w.r.t. $f_{\theta_0,\varphi}$.
(ii) For testing hypotheses (6.12), a UMPU test of size α is

$$T_*(Y,U) = \begin{cases} 1 & c_1(U) < Y < c_2(U) \\ \gamma_i(U) & Y = c_i(U),\ i = 1,2, \\ 0 & Y < c_1(U) \text{ or } Y > c_2(U), \end{cases} \tag{6.26}$$

where $c_i(u)$'s and $\gamma_i(u)$'s are Borel functions determined by

$$E_{\theta_1}[T_*(Y,U)|U = u] = E_{\theta_2}[T_*(Y,U)|U = u] = \alpha \tag{6.27}$$

for every u.
(iii) For testing hypotheses (6.13), a UMPU test of size α is

$$T_*(Y,U) = \begin{cases} 1 & Y < c_1(U) \text{ or } Y > c_2(U) \\ \gamma_i(U) & Y = c_i(U),\ i = 1,2, \\ 0 & c_1(U) < Y < c_2(U), \end{cases} \tag{6.28}$$

where $c_i(u)$'s and $\gamma_i(u)$'s are Borel functions determined by (6.27) for every u.
(iv) For testing hypotheses (6.14), a UMPU test of size α is given by (6.28), where $c_i(u)$'s and $\gamma_i(u)$'s are Borel functions determined by (6.25) and

$$E_{\theta_0}[T_*(Y,U)Y|U=u] = \alpha E_{\theta_0}(Y|U=u) \tag{6.29}$$

for every u.
Proof. Since (Y,U) is sufficient for (θ,φ), we only need to consider tests that are functions of (Y,U). Hypotheses in (i)-(iv) are of the form (6.20) with $\bar{\Theta}_{01} = \{(\theta,\varphi) : \theta = \theta_0\}$ or $= \{(\theta,\varphi) : \theta = \theta_i,\ i = 1,2\}$. In any case, U is sufficient and complete for $P \in \bar{\mathcal{P}}$ and, hence, Lemma 6.6 applies. By Theorem 2.1, the power functions of all tests are continuous and, hence, Lemma 6.5 applies. Thus, for (i)-(iii), we only need to show that T_* is UMP among all tests T satisfying (6.25) (for part (i)) or (6.27) (for part (ii) or (iii)) with T_* replaced by T. For (iv), any unbiased T should satisfy (6.25) with T_* replaced by T and

$$\frac{\partial}{\partial\theta} E_{\theta,\varphi}[T(Y,U)] = 0, \qquad \theta \in \bar{\Theta}_{01}. \tag{6.30}$$

By Theorem 2.1, the differentiation can be carried out under the expectation sign. Hence, one can show (exercise) that (6.30) is equivalent to

$$E_{\theta,\varphi}[T(Y,U)Y - \alpha Y] = 0, \qquad \theta \in \bar{\Theta}_{01}. \tag{6.31}$$

Using the argument in the proof of Lemma 6.6, one can show (exercise) that (6.31) is equivalent to (6.29) with T_* replaced by T. Hence, to prove (iv) we only need to show that T_* is UMP among all tests T satisfying (6.25) and (6.29) with T_* replaced by T.

Note that the power function of any test $T(Y,U)$ is

$$\beta_T(\theta,\varphi) = \int \left[\int T(y,u) dP_{Y|U=u}(y)\right] dP_U(u).$$

Thus, it suffices to show that for every fixed u and $\theta \in \Theta_1$, T_* maximizes

$$\int T(y,u) dP_{Y|U=u}(y)$$

over all T subject to the given side conditions. Since $P_{Y|U=u}$ is in a one-parameter exponential family, the results in (i) and (ii) follow from Corollary 6.1 and Theorem 6.3, respectively. The result in (iii) follows from Theorem 6.3(ii) by considering $1 - T_*$ with T_* given by (6.15). To prove the result in (iv), it suffices to show that if Y has the p.d.f. given by (6.10) and if U is treated as a constant in (6.25), (6.28), and (6.29), T_*

in (6.28) is UMP subject to conditions (6.25) and (6.29). We now omit U in the following proof for (iv), which is very similar to the proof of Theorem 6.3. First, $(\alpha, \alpha E_{\theta_0}(Y))$ is an interior point of the set of points $(E_{\theta_0}[T(Y)], E_{\theta_0}[T(Y)Y])$ as T ranges over all tests of the form $T(Y)$. By Lemma 6.2 and Proposition 6.1, for testing $\theta = \theta_0$ versus $\theta = \theta_1$, the UMP test is equal to 1 when

$$(k_1 + k_2 y)e^{\theta_0 y} < C(\theta_0, \theta_1)e^{\theta_1 y}, \tag{6.32}$$

where k_i's and $C(\theta_0, \theta_1)$ are constants. Note that (6.32) is equivalent to

$$a_1 + a_2 y < e^{by}$$

for some constants a_1, a_2, and b. This region is either one-sided or the outside of an interval. By Theorem 6.2(ii), a one-sided test has a strictly monotone power function and therefore cannot satisfy (6.29). Thus, this test must have the form (6.28). Since T_* in (6.28) does not depend on θ_1, by Lemma 6.1, it is UMP over all tests satisfying (6.25) and (6.29), in particular, the test $\equiv \alpha$. Thus, T_* is UMPU.

Finally, it can be shown that all the c- and γ-functions in (i)-(iv) are Borel functions (see Lehmann (1986, p. 149)). ∎

Example 6.11. A problem arising in many different contexts is the comparison of two treatments. If the observations are integer-valued, the problem often reduces to testing the equality of two Poisson distributions (e.g., a comparison of the radioactivity of two substances or the car accident rate in two cities) or two binomial distributions (when the observation is the number of successes in a sequence of trials for each treatment).

Consider first the Poisson problem in which X_1 and X_2 are independently distributed as the Poisson distributions $P(\lambda_1)$ and $P(\lambda_2)$, respectively. The p.d.f. of $X = (X_1, X_2)$ is

$$\frac{e^{-(\lambda_1+\lambda_2)}}{x_1!x_2!} \exp\{x_2 \log(\lambda_2/\lambda_1) + (x_1 + x_2)\log\lambda_2\} \tag{6.33}$$

w.r.t. the counting measure on $\{(i,j) : i = 0, 1, 2, ..., j = 0, 1, 2, ...\}$. Let $\theta = \log(\lambda_2/\lambda_1)$. Then hypotheses such as $\lambda_1 = \lambda_2$ and $\lambda_1 \geq \lambda_2$ are equivalent to $\theta = 1$ and $\theta \leq 1$, respectively. The p.d.f. in (6.33) is of the form (6.23) with $\varphi = \log\lambda_2$, $Y = X_2$, and $U = X_1 + X_2$. Thus, Theorem 6.4 applies. To obtain various tests in Theorem 6.4, it is enough to derive the conditional distribution of $Y = X_2$ given $U = X_1 + X_2 = u$. Using the fact that $X_1 + X_2$ has the Poisson distribution $P(\lambda_1 + \lambda_2)$, one can show that

$$P(Y = y|U = u) = \binom{u}{y} p^y (1-p)^{u-y} I_{\{0,1,...,u\}}(y), \quad u = 0, 1, 2, ...,$$

where $p = \lambda_2/(\lambda_1 + \lambda_2) = e^\theta/(1 + e^\theta)$. This is the binomial distribution $Bi(p, u)$. On the boundary set $\bar{\Theta}_{01}$, $\theta = \theta_j$ (a known value) and the distribution $P_{Y|U=u}$ is known.

The previous result can obviously be extended to the case where two independent samples, $X_{i1}, ..., X_{in_i}$, $i = 1, 2$, are i.i.d. from the Poisson distributions $P(\lambda_i)$, $i = 1, 2$, respectively.

Consider next the binomial problem in which X_j, $j = 1, 2$, are independently distributed as the binomial distributions $Bi(p_j, n_j)$, $j = 1, 2$, respectively, where n_j's are known but p_j's are unknown. The p.d.f. of $X = (X_1, X_2)$ is

$$\binom{n_1}{x_1}\binom{n_2}{x_2}(1-p_1)^{n_1}(1-p_2)^{n_2}\exp\left\{x_2 \log \tfrac{p_2(1-p_1)}{p_1(1-p_2)} + (x_1 + x_2)\log \tfrac{p_1}{(1-p_1)}\right\}$$

w.r.t. the counting measure on $\{(i, j) : i = 0, 1, ..., n_1,\ j = 0, 1, ..., n_2\}$. This p.d.f. is of the form (6.23) with $\theta = \log \frac{p_2(1-p_1)}{p_1(1-p_2)}$, $Y = X_2$, and $U = X_1 + X_2$. Thus, Theorem 6.4 applies. Note that hypotheses such as $p_1 = p_2$ and $p_1 \geq p_2$ are equivalent to $\theta = 0$ and $\theta \leq 0$, respectively. Using the joint distribution of (X_1, X_2), one can show (exercise) that

$$P(Y = y|U = u) = K_u(\theta)\binom{n_1}{u-y}\binom{n_2}{y}e^{\theta y}I_A(y), \qquad u = 0, 1, ..., n_1 + n_2,$$

where $A = \{y : y = 0, 1, ..., \min(u, n_2), u - y \leq n_1\}$ and

$$K_u(\theta) = \left[\sum_{y\in A}\binom{n_1}{u-y}\binom{n_2}{y}e^{\theta y}\right]^{-1}. \tag{6.34}$$

If $\theta = 0$, this distribution reduces to a known distribution: the hypergeometric distribution $HG(u, n_2, n_1)$ (Table 1.1, page 18). ∎

Example 6.12 (2×2 contingency tables). Let A and B be two different events in a probability space related to a random experiment. Suppose that n independent trials of the experiment are carried out and that we observe the frequencies of the occurrence of the events $A \cap B$, $A \cap B^c$, $A^c \cap B$, and $A^c \cap B^c$. The results can be summarized in the following 2×2 *contingency table*:

	A	A^c	Total
B	X_{11}	X_{12}	n_1
B^c	X_{21}	X_{22}	n_2
Total	m_1	m_2	n

The distribution of $X = (X_{11}, X_{12}, X_{21}, X_{22})$ is multinomial (Example 2.7) with probabilities p_{11}, p_{12}, p_{21}, and p_{22}, where $p_{ij} = E(X_{ij})/n$. Thus, the p.d.f. of X is

$$\frac{n!}{x_{11}!x_{12}!x_{21}!x_{22}!} p_{22}^n \exp\left\{x_{11}\log\frac{p_{11}}{p_{22}} + x_{12}\log\frac{p_{12}}{p_{22}} + x_{21}\log\frac{p_{21}}{p_{22}}\right\}$$

w.r.t. the counting measure on the range of X. This p.d.f. is clearly of the form (6.23). By Theorem 6.4, we can derive UMPU tests for any parameter of the form

$$\theta = a_0 \log\frac{p_{11}}{p_{22}} + a_1\log\frac{p_{12}}{p_{22}} + a_2\log\frac{p_{21}}{p_{22}},$$

where a_i's are given constants. In particular, testing independence of A and B is equivalent to the hypotheses $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ when $a_1 = a_2 = 1$ and $a_0 = -1$ (exercise).

For hypotheses concerning θ with $a_1 = a_2 = 1$ and $a_0 = -1$, the p.d.f. of X can be written as (6.23) with $Y = X_{11}$ and $U = (X_{11} + X_{12}, X_{11} + X_{21})$. A direct calculation shows that $P(Y = y | X_{11} + X_{12} = n_1, X_{11} + X_{21} = u_2)$ is equal to

$$K_{u_2}(\theta)\binom{n_1}{y}\binom{n_2}{u_2 - y} e^{\theta(u_2 - y)} I_A(y),$$

where $A = \{y : y = 0, 1, ..., \min(u_2, n_1), u_2 - y \leq n_2\}$ and $K_u(\theta)$ is given by (6.34). This distribution is known when $\theta = \theta_j$ is known. In particular, for testing independence of A and B, $\theta = 0$ implies that $P_{Y|U=u}$ is the hypergeometric distribution $HG(u_2, n_1, n_2)$, and the UMPU test in Theorem 6.4(iv) is also known as Fisher's exact test.

Suppose that X_{ij}'s in the 2×2 contingency table are from two binomial distributions, i.e., X_{i1} is from the binomial distribution $Bi(p_i, n_i)$, $X_{i2} = n_i - X_{i1}$, $i = 1, 2$, and that X_{i1}'s are independent. Then the UMPU test for independence of A and B previously derived is exactly the same as the UMPU test for $p_1 = p_2$ given in Example 6.11. The only difference is that n_i's are fixed for testing the equality of two binomial distributions whereas n_i's are random for testing independence of A and B. This is also true for the general $r \times c$ contingency tables considered in §6.4.3. ∎

6.2.3 UMPU tests in normal families

An important application of Theorem 6.4 to problems with continuous distributions in exponential families is the derivation of UMPU tests in normal families. The results presented here are the basic justifications for tests in elementary textbooks concerning parameters in normal families.

We start with the following lemma, which is useful especially when X is from a normal family.

Lemma 6.7. Suppose that X has the p.d.f. (6.23) and that $V(Y,U)$ is a statistic independent of U when $\theta=\theta_j$, where θ_j's are known values given in the hypotheses in (i)-(iv) of Theorem 6.4.
(i) If $V(y,u)$ is increasing in y for each u, then the UMPU tests in (i)-(iii) of Theorem 6.4 are equivalent to those given by (6.24)-(6.28) with Y and (Y,U) replaced by V and with $c_i(U)$'s and $\gamma_i(U)$'s replaced by constants c_i's and γ_i's.
(ii) If there are functions $a(u)>0$ and $b(u)$ such that $V(y,u)=a(u)y+b(u)$, then the UMPU test in (iv) of Theorem 6.4 is equivalent to that given by (6.25), (6.28), and (6.29) with Y and (Y,U) replaced by V and with $c_i(U)$'s and $\gamma_i(U)$'s replaced by constants c_i's and γ_i's.
Proof. (i) Since V is increasing in y, $Y>c_i(u)$ is equivalent to $V>d_i(u)$ for some d_i. The result follows from the fact that V is independent of U so that d_i's and γ_i's do not depend on u when Y is replaced by V.
(ii) Since $V=a(U)Y+b(U)$, the UMPU test in Theorem 6.4(iv) is the same as

$$T_*(V,U)=\begin{cases}1 & V<c_1(U) \text{ or } V>c_2(U)\\ \gamma_i(U) & V=c_i(U),\ i=1,2,\\ 0 & c_1(U)<V<c_2(U),\end{cases} \tag{6.35}$$

subject to $E_{\theta_0}[T_*(V,U)|U=u]=\alpha$ and

$$E_{\theta_0}\left[T_*(V,U)\frac{V-b(U)}{a(U)}\middle|U\right]=\alpha E_{\theta_0}\left[\frac{V-b(U)}{a(U)}\middle|U\right]. \tag{6.36}$$

Under $E_{\theta_0}[T_*(V,U)|U=u]=\alpha$, (6.36) is the same as $E_{\theta_0}[T_*(V,U)V|U]=\alpha E_{\theta_0}(V|U)$. Since V and U are independent, $c_i(u)$'s and $\gamma_i(u)$'s do not depend on u and, therefore, T_* in (6.35) does not depend on U. ∎

If the conditions of Lemma 6.7 are satisfied, then UMPU tests can be derived by working with the distribution of V instead of $P_{Y|U=u}$. In exponential families, a $V(Y,U)$ independent of U can often be found by applying Basu's theorem (Theorem 2.4).

When we consider normal families, γ_i's can be chosen to be 0 since the c.d.f. of Y or V is continuous.

One-sample problems

Let $X_1,...,X_n$ be i.i.d. from $N(\mu,\sigma^2)$ with unknown $\mu\in\mathcal{R}$ and $\sigma^2>0$, where $n\geq 2$. The joint p.d.f. of $X=(X_1,...,X_n)$ is

$$\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2+\frac{n\mu}{\sigma^2}\bar{x}-\frac{n\mu^2}{2\sigma^2}\right\}.$$

Consider first hypotheses concerning σ^2. The p.d.f. of X has the form (6.23) with $\theta = -(2\sigma^2)^{-1}$, $\varphi = n\mu/\sigma^2$, $Y = \sum_{i=1}^n X_i^2$, and $U = \bar{X}$. By Basu's theorem, $V = (n-1)S^2$ (S^2 is the sample variance) is independent of $U = \bar{X}$ (Example 2.18). Also,

$$\sum_{i=1}^n X_i^2 = (n-1)S^2 + n\bar{X}^2,$$

i.e., $V = Y - nU^2$. Hence the conditions of Lemma 6.7 are satisfied. Since V/σ^2 has the chi-square distribution χ^2_{n-1} (Example 2.18), values of c_i's for hypotheses in (i)-(iii) of Theorem 6.4 are related to quantiles of χ^2_{n-1}. For testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ (which is equivalent to testing $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 \neq \sigma_0^2$), $d_i = c_i/\sigma_0^2$, $i = 1, 2$, are determined by

$$\int_{d_1}^{d_2} \chi^2_{n-1}(v)dv = 1 - \alpha \qquad \text{and} \qquad \int_{d_1}^{d_2} v\chi^2_{n-1}(v)dv = (n-1)(1-\alpha),$$

where χ^2_m is the p.d.f. of the chi-square distribution χ^2_m. Since $v\chi^2_{n-1}(v) = (n-1)\chi^2_{n+1}(v)$, d_1 and d_2 are determined by

$$\int_{d_1}^{d_2} \chi^2_{n-1}(v)dv = \int_{d_1}^{d_2} \chi^2_{n+1}(v)dv = 1 - \alpha.$$

If $n - 1 \approx n + 1$, then d_1 and d_2 are nearly the $(\alpha/2)$th and $(1 - \alpha/2)$th quantiles of χ^2_{n-1}, respectively, in which case the UMPU test in Theorem 6.4(iv) is the same as the "equal-tailed" chi-square test for H_0 in elementary textbooks.

Consider next hypotheses concerning μ. The p.d.f. of X has the form (6.23) with $\theta = n\mu/\sigma^2$, $\varphi = -(2\sigma^2)^{-1}$, $Y = \bar{X}$, and $U = \sum_{i=1}^n X_i^2$. For testing hypotheses $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$, we take V to be $t(X) = \sqrt{n}(\bar{X} - \mu_0)/S$. By Basu's theorem, $t(X)$ is independent of U when $\mu = \mu_0$. Hence it satisfies the conditions in Lemma 6.7(i). From Examples 1.15 and 2.18, $t(X)$ has the t-distribution t_{n-1} when $\mu = \mu_0$. Thus, $c(U)$ in Theorem 6.4(i) is the $(1 - \alpha)$th quantile of t_{n-1}. For the two-sided hypotheses $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, we consider $V = (\bar{X} - \mu_0)/\sum_{i=1}^n (X_i - \mu_0)^2$, which satisfies the conditions in Lemma 6.7(ii) and has a distribution symmetric about 0 when $\mu = \mu_0$. Then the UMPU test in Theorem 6.4(iv) rejects H_0 when $|V| > d$ with $P_{\mu_0}(|V| > d) = \alpha$. Since

$$t(X) = \sqrt{(n-1)n}V(X)/\sqrt{1 - n[V(X)]^2},$$

the UMPU test rejects H_0 if and only if $|t(X)| > t_{n-1,\alpha/2}$, where $t_{n-1,\alpha}$ is the $(1 - \alpha)$th quantile of the t-distribution t_{n-1}. The UMPU tests derived here are the so-called one-sample t-tests in elementary textbooks.

The power function of a one-sample t-test is related to the noncentral t-distribution introduced in §1.3.1 (see Exercise 32).

Two-sample problems

The problem of comparing the parameters of two normal distributions arises in the comparison of two treatments, products, and so on (see also Example 6.11). Suppose that we have two independent samples, $X_{i1}, ..., X_{in_i}$, $i = 1, 2$, i.i.d. from $N(\mu_i, \sigma_i^2)$, $i = 1, 2$, respectively, where $n_i \geq 2$. The joint p.d.f. of X_{ij}'s is

$$C(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \exp\left\{ -\sum_{i=1}^{2} \frac{1}{2\sigma_i^2} \sum_{j=1}^{n_i} x_{ij}^2 + \sum_{i=1}^{2} \frac{n_i \mu_i}{\sigma_i^2} \bar{x}_i \right\},$$

where $\bar{x}_i$ is the sample mean based on $x_{i1}, ..., x_{in_i}$ and $C(\cdot)$ is a known function.

Consider first the hypothesis $H_0 : \sigma_2^2/\sigma_1^2 \leq \Delta_0$ or $H_0 : \sigma_2^2/\sigma_1^2 = \Delta_0$. The p.d.f. of X_{ij}'s is of the form (6.23) with

$$\theta = \frac{1}{2\Delta_0 \sigma_1^2} - \frac{1}{2\sigma_2^2}, \qquad \varphi = \left(-\frac{1}{2\sigma_1^2}, \frac{n_1 \mu_1}{\sigma_1^2}, \frac{n_2 \mu_2}{\sigma_2^2} \right),$$

$$Y = \sum_{j=1}^{n_2} X_{2j}^2, \qquad U = \left(\sum_{j=1}^{n_1} X_{1j}^2 + \frac{1}{\Delta_0} \sum_{j=1}^{n_2} X_{2j}^2,\ \bar{X}_1,\ \bar{X}_2 \right).$$

To apply Lemma 6.7, consider

$$V = \frac{(n_2 - 1) S_2^2/\Delta_0}{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2/\Delta_0} = \frac{(Y - n_2 U_3)/\Delta_0}{U_1 - n_1 U_2 - n_2 U_3/\Delta_0},$$

where S_i^2 is the sample variance based on $X_{i1}, ..., X_{in_i}$ and U_j is the jth component of U. By Basu's theorem, V and U are independent when $\theta = 0$ ($\sigma_2^2 = \Delta_0 \sigma_1^2$). Since V is increasing and linear in Y, the conditions of Lemma 6.7 are satisfied. Thus, a UMPU test rejects $H_0 : \theta \leq 0$ (which is equivalent to $H_0 : \sigma_2^2/\sigma_1^2 \leq \Delta_0$) when $V > c_0$ with $P_{\theta=0}(V > c_0) = \alpha$; a UMPU test rejects $H_0 : \theta = 0$ (which is equivalent to $H_0 : \sigma_2^2/\sigma_1^2 = \Delta_0$) when $V < c_1$ or $V > c_2$ with $P_{\theta=0}(c_1 < V < c_2) = 1 - \alpha$ and $E_{\theta=0}[V T_*(V)] = \alpha E_{\theta=0}(V)$. Note that

$$V = \frac{(n_2 - 1)\mathtt{F}}{n_1 - 1 + (n_2 - 1)\mathtt{F}} \qquad \text{with} \qquad \mathtt{F} = \frac{S_2^2/\Delta_0}{S_1^2}.$$

It follows from Example 1.15 that $\mathtt{F}$ has the F-distribution F_{n_2-1, n_1-1} (Table 1.2, page 20) when $\theta = 0$. Since V is a strictly increasing function of $\mathtt{F}$, a UMPU test rejects $H_0 : \theta \leq 0$ when $\mathtt{F} > F_{n_2-1, n_1-1, \alpha}$, where $F_{a,b,\alpha}$ is the $(1 - \alpha)$th quantile of the F-distribution $F_{a,b}$. This is the F-test in elementary textbooks.

When $\theta = 0$, V has the beta distribution $B((n_2-1)/2, (n_1-1)/2)$ and $E_{\theta=0}(V) = (n_2-1)/(n_1+n_2-2)$ (Table 1.2). Then condition $E_{\theta=0}[VT_*(V)] = \alpha E_{\theta=0}(V)$ is the same as

$$\frac{(1-\alpha)(n_2-1)}{n_1+n_2-2} = \int_{c_1}^{c_2} v f_{(n_2-1)/2,(n_1-1)/2}(v)dv,$$

where $f_{a,b}$ is the p.d.f. of the beta distribution $B(a,b)$. Using the fact that $vf_{(n_2-1)/2,(n_1-1)/2}(v) = (n_1+n_2-2)^{-1}(n_2-1)f_{(n_2+1)/2,(n_1-1)/2}(v)$, we conclude that a UMPU test rejects $H_0 : \theta = 0$ when $V < c_1$ or $V > c_2$, where c_1 and c_2 are determined by

$$1-\alpha = \int_{c_1}^{c_2} v f_{(n_2-1)/2,(n_1-1)/2}(v)dv = \int_{c_1}^{c_2} f_{(n_2+1)/2,(n_1-1)/2}(v)dv.$$

If $n_2 - 1 \approx n_2 + 1$ (i.e., n_2 is large), then this UMPU test can be approximated by the F-test which rejects $H_0 : \theta = 0$ if and only if $\mathtt{F} < F_{n_2-1,n_1-1,1-\alpha/2}$ or $\mathtt{F} > F_{n_2-1,n_1-1,\alpha/2}$.

Consider next the hypothesis $H_0 : \mu_1 \geq \mu_2$ or $H_0 : \mu_1 = \mu_2$. If $\sigma_1^2 \neq \sigma_2^2$, the problem is the so-called Behrens-Fisher problem and is not accessible by the method introduced in this section. We now assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ but σ^2 is unknown. The p.d.f. of X_{ij}'s is then

$$C(\mu_1, \mu_2, \sigma^2) \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} x_{ij}^2 + \frac{n_1\mu_1}{\sigma^2}\bar{x}_1 + \frac{n_2\mu_2}{\sigma^2}\bar{x}_2 \right\},$$

which is of the form (6.23) with

$$\theta = \frac{\mu_2 - \mu_1}{(n_1^{-1} + n_2^{-1})\sigma^2}, \qquad \varphi = \left(\frac{n_1\mu_1 + n_2\mu_2}{(n_1+n_2)\sigma^2}, \ -\frac{1}{2\sigma^2} \right),$$

$$Y = \bar{X}_2 - \bar{X}_1, \qquad U = \left(n_1\bar{X}_1 + n_2\bar{X}_2, \ \sum_{i=1}^{2}\sum_{j=1}^{n_i} X_{ij}^2 \right).$$

For testing $H_0 : \theta \leq 0$ (i.e., $\mu_1 \geq \mu_2$) versus $H_1 : \theta > 0$, we consider V in Lemma 6.7 to be

$$t(X) = \frac{(\bar{X}_2 - \bar{X}_1)/\sqrt{n_1^{-1} + n_2^{-1}}}{\sqrt{[(n_1-1)S_1^2 + (n_2-1)S_2^2]/(n_1+n_2-2)}}. \tag{6.37}$$

When $\theta = 0$, $t(X)$ is independent of U (Basu's theorem) and satisfies the conditions in Lemma 6.7(i); the numerator and the denominator of $t(X)$ (after division by σ) are independently distributed as $N(0,1)$ and

the chi-square distribution $\chi^2_{n_1+n_2-2}$, respectively. Hence $t(X)$ has the t-distribution $t_{n_1+n_2-2}$ and a UMPU test rejects H_0 when $t(X) > t_{n_1+n_2-2,\alpha}$, where $t_{n_1+n_2-2,\alpha}$ is the $(1-\alpha)$th quantile of the t-distribution $t_{n_1+n_2-2}$. This is the so-called (one-sided) two-sample t-test.

For testing $H_0 : \theta = 0$ (i.e., $\mu_1 = \mu_2$) versus $H_1 : \theta \neq 0$, it follows from a similar argument used in the derivation of the (two-sided) one-sample t-test that a UMPU test rejects H_0 when $|t(X)| > t_{n_1+n_2-2,\alpha/2}$ (exercise). This is the (two-sided) two-sample t-test.

The power function of a two-sample t-test is related to a noncentral t-distribution.

Normal linear models

Consider linear model (3.25) with assumption A1, i.e.,

$$X = (X_1, ..., X_n) \ \text{ is } \ N_n(\beta Z^\tau, \sigma^2 I_n), \tag{6.38}$$

where β is a p-vector of unknown parameters, $Z = (Z_1^\tau, ..., Z_n^\tau)^\tau$, Z_i's are the values of a p-vector of deterministic covariates, and $\sigma^2 > 0$ is an unknown parameter. Assume that $n > p$ and the rank of Z is $r \leq p$. Let $l \in \mathcal{R}(Z)$ (the linear space generated by the rows of Z) and θ_0 be a fixed constant. We consider the hypotheses

$$H_0 : \beta l^\tau \leq \theta_0 \qquad \text{versus} \qquad H_1 : \beta l^\tau > \theta_0 \tag{6.39}$$

or

$$H_0 : \beta l^\tau = \theta_0 \qquad \text{versus} \qquad H_1 : \beta l^\tau \neq \theta_0. \tag{6.40}$$

Since $H = Z(Z^\tau Z)^- Z^\tau$ is a projection matrix of rank r, there exists an $n \times n$ orthogonal matrix Γ such that

$$\Gamma = (\Gamma_1, \Gamma_2) \qquad \text{and} \qquad H\Gamma = (\Gamma_1, 0), \tag{6.41}$$

where Γ_1 is $n \times r$ and Γ_2 is $n \times (n-r)$. Let $Y_j = X\Gamma_j$, $j = 1, 2$. Consider the transformation $(Y_1, Y_2) = X\Gamma$. Since $\Gamma^\tau \Gamma = I_n$ and X is $N_n(\beta Z^\tau, \sigma^2 I_n)$, (Y_1, Y_2) is $N_n(\beta Z^\tau \Gamma, \sigma^2 I_n)$. It follows from (6.41) that

$$E(Y_2) = E(X\Gamma_2) = \beta Z^\tau \Gamma_2 = \beta Z^\tau H \Gamma_2 = 0.$$

Let $\eta = \beta Z^\tau \Gamma_1 = E(Y_1)$. Then the p.d.f. of (Y_1, Y_2) is

$$\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ \frac{Y_1 \eta^\tau}{\sigma^2} - \frac{\|Y_1\|^2 + \|Y_2\|^2}{2\sigma^2} - \frac{\|\eta\|^2}{2\sigma^2} \right\}. \tag{6.42}$$

Since l in (6.39) or (6.40) is in $\mathcal{R}(Z)$, there exists $\lambda \in \mathcal{R}^n$ such that $l = \lambda Z$. Then

$$\hat{\beta} l^\tau = XH\lambda^\tau = XH\Gamma\Gamma^\tau \lambda^\tau = X\Gamma_1 \Gamma_1^\tau \lambda^\tau = Y_1 \Gamma_1^\tau \lambda^\tau, \tag{6.43}$$

where $\hat{\beta}$ is the LSE defined by (3.27). By (6.43) and Theorem 3.6(ii),

$$E(\hat{\beta}l^\tau) = \beta l^\tau = E(Y_1)\Gamma_1^\tau \lambda^\tau = \eta a^\tau,$$

where $a = \lambda\Gamma_1$. Let $\eta = (\eta_1, ..., \eta_r)$ and $a = (a_1, ..., a_r)$. Without loss of generality, we assume that $a_1 \neq 0$. Then the p.d.f. in (6.42) is of the form (6.23) with

$$\theta = \frac{\eta a^\tau}{a_1\sigma^2}, \qquad \varphi = \left(-\frac{1}{2\sigma^2}, \frac{\eta_2}{\sigma^2}, ..., \frac{\eta_r}{\sigma^2}\right),$$

$$Y = Y_{11}, \qquad U = \left(\|Y_1\|^2 + \|Y_2\|^2, Y_{12} - \frac{a_2 Y_{11}}{a_1}, ..., Y_{1r} - \frac{a_r Y_{11}}{a_1}\right),$$

where Y_{1j} is the jth component of Y_1. By Basu's theorem,

$$t(X) = \frac{\sqrt{n-r}(Y_1 a^\tau - \theta_0)}{\|Y_2\|\,\|a\|}$$

is independent of U when $\eta a^\tau = \beta l^\tau = \theta_0$. Note that $\|Y_2\|^2 = SSR$ in (3.36) and $\|a\|^2 = \lambda\Gamma_1\Gamma_1^\tau\lambda^\tau = \lambda H \lambda^\tau = l(Z^\tau Z)^- l^\tau$. Hence, by (6.43),

$$t(X) = \frac{\hat{\beta}l^\tau - \theta_0}{\sqrt{l(Z^\tau Z)^- l^\tau}\sqrt{SSR/(n-r)}},$$

which has the t-distribution t_{n-r} (Theorem 3.8). Using the same arguments in deriving the one-sample or two-sample t-test, we obtain that a UMPU test for the hypotheses in (6.39) rejects H_0 when $t(X) > t_{n-r,\alpha}$, and that a UMPU test for the hypotheses in (6.40) rejects H_0 when $|t(X)| > t_{n-r,\alpha/2}$.

Testing for independence in the bivariate normal family

Suppose that $X_1, ..., X_n$ are i.i.d. from a bivariate normal distribution, i.e., the p.d.f. of $X = (X_1, ..., X_n)$ is

$$\frac{1}{(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2})^n} \exp\left\{-\frac{\|Y_1-\mu_1\|^2}{2\sigma_1^2(1-\rho^2)} + \frac{\rho(Y_1-\mu_1)(Y_2-\mu_2)^\tau}{\sigma_1\sigma_2(1-\rho^2)} - \frac{\|Y_2-\mu_2\|^2}{2\sigma_2^2(1-\rho^2)}\right\}, \quad (6.44)$$

where Y_j is the n-vector containing the jth components of $X_1, ..., X_n$, $j = 1, 2$.

Testing for independence of the two components of X_1 (or Y_1 and Y_2) is equivalent to testing $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$. In some cases one may also be interested in the one-sided hypotheses $H_0 : \rho \leq 0$ versus $H_1 : \rho > 0$. It can be shown (exercise) that the p.d.f. in (6.44) is of the form (6.23) with $\theta = \frac{\rho}{\sigma_1\sigma_2(1-\rho^2)}$ and

$$Y = \sum_{i=1}^n X_{i1}X_{i2}, \qquad U = \left(\sum_{i=1}^n X_{i1}^2, \sum_{i=1}^n X_{i2}^2, \sum_{i=1}^n X_{i1}, \sum_{i=1}^n X_{i2}\right),$$

where X_{ij} is the jth component of X_i, $j = 1, 2$.

The hypothesis $\rho \leq 0$ is equivalent to $\theta \leq 0$. The sample correlation coefficient

$$R = \sum_{i=1}^{n}(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) \Big/ \left[\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)^2 \sum_{i=1}^{n}(X_{i2} - \bar{X}_2)^2\right]^{1/2},$$

where $\bar{X}_j$ is the sample mean of $X_{1j}, ..., X_{nj}$, $j = 1, 2$, is independent of U when $\rho = 0$ (Basu's theorem). To apply Lemma 6.7, we consider

$$V = \sqrt{n-2}R/\sqrt{1-R^2}. \tag{6.45}$$

It can be shown (exercise) that R is linear in Y and that V has the t-distribution t_{n-2} when $\rho = 0$. Hence, a UMPU test for $H_0 : \rho \leq 0$ versus $H_1 : \rho > 0$ rejects H_0 when $V > t_{n-2,\alpha}$ and a UMPU test for $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$ rejects H_0 when $|V| > t_{n-2,\alpha/2}$, where $t_{n-2,\alpha}$ is the $(1-\alpha)$th quantile of the t-distribution t_{n-2}.

6.3 UMP Invariant Tests

In the previous section the unbiasedness principle is considered to derive an optimal test within the class of unbiased tests when a UMP test does not exist. In this section we study the same problem with unbiasedness replaced by invariance under a given group of transformations. The principles of unbiasedness and invariance often complement each other in that each is successful in cases where the other is not.

6.3.1 Invariance and UMPI tests

The invariance principle considered here is similar to that introduced in §2.3.2 (Definition 2.9) and in §4.2, but is more general in the sense that we are not restricted to location-scale families (although most examples in this section are about location-scale families).

Definition 6.5. Let X be a sample from $P \in \mathcal{P}$.
(i) A class $\mathcal{G}$ of one-to-one transformations of X is called a *group* if and only if $g_i \in \mathcal{G}$ implies $g_1 \circ g_2 \in \mathcal{G}$ and $g_i^{-1} \in \mathcal{G}$.
(ii) We say that $\mathcal{P}$ is *invariant* under $\mathcal{G}$ if and only if $\bar{g}(P_X) = P_{g(X)}$ is a one-to-one transformation from $\mathcal{P}$ onto $\mathcal{P}$ for each $g \in \mathcal{G}$.
(iii) We say that the problem of testing $H_0 : P \in \mathcal{P}_0$ versus $H_1 : P \in \mathcal{P}_1$ is *invariant* under $\mathcal{G}$ if and only if both $\mathcal{P}_0$ and $\mathcal{P}_1$ are invariant under $\mathcal{G}$.

(iv) In an invariant testing problem, a test $T(X)$ is said to be *invariant* under $\mathcal{G}$ if and only if

$$T(g(x)) = T(x) \qquad \text{for all } x \text{ and } g. \tag{6.46}$$

(v) A test of size α is said to be a *uniformly most powerful invariant* (UMPI) test if and only if it is UMP within the class of level α tests that are invariant under $\mathcal{G}$.
(vi) A statistic $M(X)$ is said to be *maximal invariant* under $\mathcal{G}$ if and only if (6.46) holds with T replaced by M and

$$M(x_1) = M(x_2) \qquad \text{implies } x_1 = g(x_2) \text{ for some } g \in \mathcal{G}. \quad \blacksquare \tag{6.47}$$

The following result indicates that invariance reduces the data X to a maximal invariant statistic $M(X)$ whose distribution may depend only on a functional of P that shrinks $\mathcal{P}$.

Proposition 6.2. Let $M(X)$ be maximal invariant under $\mathcal{G}$.
(i) A test $T(X)$ is invariant under $\mathcal{G}$ if and only if there is a function h such that $T(x) = h(M(x))$ for all x.
(ii) Suppose that there is a functional $\theta(P)$ on $\mathcal{P}$ satisfying $\theta(\bar{g}(P)) = \theta(P)$ for all $g \in \mathcal{G}$ and $P \in \mathcal{P}$ and

$$\theta(P_1) = \theta(P_2) \qquad \text{implies } P_1 = \bar{g}(P_2) \text{ for some } g \in \mathcal{G}$$

(i.e., $\theta(P)$ is "maximal invariant"). Then the distribution of $M(X)$ depends only on $\theta(P)$.
Proof. (i) If $T(x) = h(M(x))$ for all x, then $T(g(x)) = h(M(g(x))) = h(M(x)) = T(x)$ so that T is invariant. If T is invariant and if $M(x_1) = M(x_2)$, then $x_1 = g(x_2)$ for some g and $T(x_1) = T(g(x_2)) = T(x_2)$. Hence T is a function of M.
(ii) Suppose that $\theta(P_1) = \theta(P_2)$. Then $P_2 = \bar{g}(P_1)$ for some $g \in \mathcal{G}$ and for any event B in the range of $M(X)$,

$$\begin{aligned} P_2\big(M(X) \in B\big) &= \bar{g}(P_1)\big(M(X) \in B\big) \\ &= P_1\big(M(g(X)) \in B\big) \\ &= P_1\big(M(X) \in B\big). \end{aligned}$$

Hence the distribution of $M(X)$ depends only on $\theta(P)$. $\blacksquare$

In applications, maximal invariants $M(X)$ and $\theta = \theta(P)$ are frequently real-valued. If the hypotheses of interest can be expressed in terms of θ, then there may exist a test UMP among those depending only on $M(X)$ (e.g., when the distribution of $M(X)$ is in a parametric family having monotone likelihood ratio). Such a test is then a UMPI test.

Example 6.13 (Location-scale families). Suppose that X has the Lebesgue p.d.f. $f_{i,\mu}(x) = f_i(x_1 - \mu, ..., x_n - \mu)$, where $n \geq 2$, $\mu \in \mathcal{R}$ is unknown, and f_i, $i = 0, 1$, are known Lebesgue p.d.f.'s. We consider the problem of testing

$$H_0 : X \text{ is from } f_{0,\mu} \qquad \text{versus} \qquad H_1 : X \text{ is from } f_{1,\mu}. \tag{6.48}$$

Consider $\mathcal{G} = \{g_c : c \in \mathcal{R}\}$ with $g_c(x) = (x_1 + c, ..., x_n + c)$. For any $g_c \in \mathcal{G}$, it induces a transformation $\bar{g}_c(f_{i,\mu}) = f_{i,\mu+c}$ and the problem of testing H_0 versus H_1 in (6.48) is invariant under $\mathcal{G}$.

We now show that a maximal invariant under $\mathcal{G}$ is $D(X) = (D_1, ..., D_{n-1}) = (X_1 - X_n, ..., X_{n-1} - X_n)$. First, it is easy to see that $D(X)$ is invariant under $\mathcal{G}$. Let $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ be two points in the range of X. Suppose that $x_i - x_n = y_i - y_n$ for $i = 1, ..., n-1$. Putting $c = y_n - x_n$, we have $y_i = x_i + c$ for all i. Hence, $D(X)$ is maximal invariant under $\mathcal{G}$.

By Proposition 1.8, D has the p.d.f. $\int f_i(d_1 + t, ..., d_{n-1} + t, t)dt$ under H_i, $i = 0, 1$, which does not depend on μ. In fact, in this case Proposition 6.2 applies with $M(X) = D(X)$ and $\theta(f_{i,\mu}) = i$. If we consider tests that are functions of $D(X)$, then the problem of testing the hypotheses in (6.48) becomes one of testing a simple hypothesis versus a simple hypothesis. By Theorem 6.1, the test UMP among functions of $D(X)$, which is then the UMPI test, rejects H_0 in (6.48) when

$$\frac{\int f_1(d_1 + t, ..., d_{n-1} + t, t)dt}{\int f_0(d_1 + t, ..., d_{n-1} + t, t)dt} = \frac{\int f_1(x_1 + t, ..., x_n + t)dt}{\int f_0(x_1 + t, ..., x_n + t)dt} > c,$$

where c is determined by the size of the UMPI test.

The previous result can be extended to the case of a location-scale family where the p.d.f. of X is one of $f_{i,\mu,\sigma} = \frac{1}{\sigma^n} f_i\left(\frac{x_1-\mu}{\sigma}, ..., \frac{x_n-\mu}{\sigma}\right)$, $i = 0, 1$, $f_{i,\mu,\sigma}$ is symmetric about μ, the hypotheses of interest are given by (6.48) with $f_{i,\mu}$ replaced by $f_{i,\mu,\sigma}$, and $\mathcal{G} = \{g_{c,r} : c \in \mathcal{R}, r \neq 0\}$ with $g_{c,r}(x) = (rx_1 + c, ..., rx_n + c)$. When $n \geq 3$, it can be shown that a maximal invariant under $\mathcal{G}$ is $W(X) = (W_1, ..., W_{n-2})$, where $W_i = (X_i - X_n)/(X_{n-1} - X_n)$, and that the p.d.f. of W does not depend on (μ, σ). A UMPI test can then be derived (exercise). ∎

The next example considers finding a maximal invariant in a problem that is not a location-scale family problem.

Example 6.14. Let $\mathcal{G}$ be the set of $n!$ permutations of the components of $x \in \mathcal{R}^n$. Then a maximal invariant is the vector of order statistics. This is because a permutation of the components of x does not change the values of these components and two x's with the same set of ordered components can be obtained from each other through a permutation of coordinates.

Suppose that $\mathcal{P}$ contains continuous c.d.f.'s on $\mathcal{R}^n$. Let $\mathcal{G}$ be the class of all transformations of the form $g(x) = (\psi(x_1), ..., \psi(x_n))$, where ψ is continuous and strictly increasing. For $x = (x_1, ..., x_n)$, let $R(x) = (R_1, ..., R_n)$ be the vector of ranks (§5.2.2), i.e., $x_i = x_{(R_i)}$, where $x_{(j)}$ is the jth ordered value of x_i's. Clearly, $R(g(x)) = R(x)$ for any $g \in \mathcal{G}$. For any x and y in $\mathcal{R}^n$ with $R(x) = R(y)$, define $\psi(t) = t + (y_{(1)} - x_{(1)})$ for $t \leq x_{(1)}$, $\psi(t) = t + (y_{(n)} - x_{(n)})$ for $t \geq x_{(n)}$, and to be linear between $x_{(j)}$ and $x_{(j+1)}$, $j = 1, ..., n-1$. Then $\psi(x_i) = \psi(y_i)$, $i = 1, ..., n$. This shows that the vector of rank statistics is maximal invariant. ∎

When there is a sufficient statistic $U(X)$, it is convenient first to reduce the data to $U(X)$ before applying invariance. If there is a test $T(U)$ UMP among all invariant tests depending only on U, one would like to conclude that $T(U)$ is a UMPI test. Unfortunately, this may not be true in general, since it is not clear that for any invariant test based on X there is an equivalent invariant test based only on $U(X)$. The following result provides a sufficient condition under which it is enough to consider invariant tests depending only on $U(X)$. Its proof is omitted and can be found in Lehmann (1986, pp. 297-302).

Proposition 6.3. Let $\mathcal{G}$ be a group of transformations on $\mathfrak{X}$ (the range of X) and $(\mathcal{G}, \mathcal{B}_\mathcal{G}, \lambda)$ be a measure space with a σ-finite λ. Suppose that the testing problem under consideration is invariant under $\mathcal{G}$, that for any set $A \in \mathcal{B}_\mathfrak{X}$, the set of points (x, g) for which $g(x) \in A$ is in $\sigma(\mathcal{B}_\mathfrak{X} \times \mathcal{B}_\mathcal{G})$, and that $\lambda(B) = 0$ implies $\lambda(\{h \circ g : h \in B\}) = 0$ for all $g \in \mathcal{G}$. Suppose further that there is a statistic $U(X)$ sufficient for $P \in \mathcal{P}$ and that $U(x_1) = U(x_2)$ implies $U(g(x_1)) = U(g(x_2))$ for all $g \in \mathcal{G}$ so that $\mathcal{G}$ induces a group $\mathcal{G}_U$ of transformations on the range of U through $g_U(U(x)) = U(g(x))$. Then, for any test $T(X)$ invariant under $\mathcal{G}$, there exists a test based on $U(X)$ that is invariant under $\mathcal{G}$ (and $\mathcal{G}_U$) and has the same power function as $T(X)$. ∎

In many problems $g(x) = \psi(x, g)$, where g ranges over a set $\mathcal{G}$ in $\mathcal{R}^m$ and ψ is a Borel function on $\mathcal{R}^{n+m}$. Then the measurability condition in Proposition 6.3 is satisfied by choosing $\mathcal{B}_\mathcal{G}$ to be the Borel σ-field on $\mathcal{G}$. In such cases it is usually not difficult to find a measure λ satisfying the condition in Proposition 6.3.

Example 6.15. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with unknown $\mu \in \mathcal{R}$ and $\sigma^2 > 0$. The problem of testing $H_0 : \sigma^2 \geq \sigma_0^2$ versus $H_1 : \sigma^2 < \sigma_0^2$ is invariant under $\mathcal{G} = \{g_c : c \in \mathcal{R}\}$ with $g_c(x) = (x_1 + c, ..., x_n + c)$. It can be shown (exercise) that $\mathcal{G}$ and the sufficient statistic $U = (\bar{X}, S^2)$ satisfy the conditions in Proposition 6.3 with $\mathcal{G}_U = \{h_c : c \in \mathcal{R}\}$ and $h_c(u_1, u_2) = (u_1 + c, u_2)$. A maximal invariant under $\mathcal{G}_U$ is S^2. It follows from Proposition 6.3, Corollary 6.1, and the fact that $(n-1)S^2/\sigma_0^2$ has

the chi-square distribution χ^2_{n-1} when $\sigma^2 = \sigma_0^2$ that a UMPI test of size α rejects H_0 when $(n-1)S^2/\sigma_0^2 \le \chi^2_{n-1,1-\alpha}$, where $\chi^2_{n-1,\alpha}$ is the $(1-\alpha)$th quantile of the chi-square distribution χ^2_{n-1}. This test coincides with the UMPU test given in §6.2.3. ∎

Example 6.16. Let $X_{i1}, ..., X_{in_i}$, $i = 1, 2$, be two independent samples i.i.d. from $N(\mu_i, \sigma_i^2)$, $i = 1, 2$, respectively. The problem of testing $H_0 : \sigma_2^2/\sigma_1^2 \le \Delta_0$ versus $H_1 : \sigma_2^2/\sigma_1^2 > \Delta_0$ is invariant under

$$\mathcal{G} = \{g_{c_1,c_2,r} : c_i \in \mathcal{R}, i = 1, 2, r > 0\}$$

with

$$g_{c_1,c_2,r}(x_1, x_2) = (rx_{11} + c_1, ..., rx_{1n_1} + c_1, rx_{21} + c_2, ..., rx_{2n_2} + c_2).$$

It can be shown (exercise) that the sufficient statistic $U = (\bar{X}_1, \bar{X}_2, S_1, S_2)$ and $\mathcal{G}$ satisfy the conditions in Proposition 6.3 with

$$\mathcal{G}_U = \{h_{c_1,c_2,r} : c_i \in \mathcal{R}, i = 1, 2, r > 0\}$$

and

$$h_{c_1,c_2,r}(u_1, u_2, u_3, u_4) = (ru_1 + c_1, ru_2 + c_2, ru_3, ru_4).$$

A maximal invariant under $\mathcal{G}_U$ is S_2/S_1. Let $\Delta = \sigma_2^2/\sigma_1^2$. Then $(S_2^2/S_1^2)/\Delta$ has an F-distribution and, therefore, $V = S_2^2/S_1^2$ has a Lebesgue p.d.f. of the form

$$f_\Delta(v) = C(\Delta)v^{(n_2-3)/2}[\Delta + (n_2-1)v/(n_1-1)]^{-(n_1+n_2-2)/2} I_{(0,\infty)}(v),$$

where $C(\Delta)$ is a known function of Δ. It can be shown (exercise) that the family $\{f_\Delta : \Delta > 0\}$ has monotone likelihood ratio in V so that a UMPI test of size α rejects H_0 when $V > F_{n_2-1,n_1-1,\alpha}$, where $F_{a,b,\alpha}$ is the $(1-\alpha)$th quantile of the F-distribution $F_{a,b}$. Again, this UMPI test coincides with the UMPU test given in §6.2.3. ∎

The following result shows that in Examples 6.15 and 6.16, the fact that UMPI tests are the same as the UMPU tests is not a simple coincidence.

Proposition 6.4. Consider a testing problem invariant under $\mathcal{G}$. If there exists a UMPI test of size α, then it is unbiased. If there also exists a UMPU test of size α that is invariant under $\mathcal{G}$, then the two tests have the same power function on $P \in \mathcal{P}_1$. If either the UMPI test or the UMPU test is unique a.s. $\mathcal{P}$, then the two tests are equal a.s. $\mathcal{P}$.

Proof. We only need to prove that a UMPI test of size α is unbiased. This follows from the fact that the test $T \equiv \alpha$ is invariant under $\mathcal{G}$. ∎

The next example shows an application of invariance in a situation where a UMPU test may not exist.

Example 6.17. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with unknown μ and σ^2. Let $\theta = (\mu - u)/\sigma$, where u is a known constant. Consider the problem of testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Note that H_0 is the same as $P(X_1 \leq u) \geq p_0$ for a known constant $p_0 = \Phi(-\theta_0)$. Without loss of generality, we consider the case of $u = 0$.

The problem is invariant under $\mathcal{G} = \{g_r : r > 0\}$ with $g_r(x) = rx$. By Proposition 6.3, we can consider tests that are functions of the sufficient statistic $(\bar{X}, S^2)$ only. A maximal invariant under $\mathcal{G}$ is $t(X) = \sqrt{n}\bar{X}/S$. To find a UMPI test, it remains to find a test UMP among all tests that are functions of $t(X)$.

From the discussion in §1.3.1, $t(X)$ has the noncentral t-distribution $t_{n-1}(\sqrt{n}\theta)$. Let $f_\theta(t)$ be the Lebesgue p.d.f. of $t(X)$, i.e., f_θ is given by (1.32) with n replaced by $n-1$ and $\delta = \sqrt{n}\theta$. It can be shown (exercise) that the family of p.d.f.'s, $\{f_\theta(t) : \theta \in \mathcal{R}\}$, has monotone likelihood ratio in t. Hence, by Theorem 6.2, a UMPI test of size α rejects H_0 when $t(X) > c$, where c is the $(1-\alpha)$th quantile of $t_{n-1}(\sqrt{n}\theta_0)$.

In some problems we may have to apply both unbiasedness and invariance principles. For instance, suppose that in the current problem we would like to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. The problem is still invariant under $\mathcal{G}$. Following the previous discussion, we only need to consider tests that are functions of $t(X)$. But a test UMP among functions of $t(X)$ does not exist in this case. A test UMP among all unbiased tests that are functions of $t(X)$ rejects H_0 when $t(X) < c_1$ or $t(X) > c_2$, where c_1 and c_2 are determined by

$$\int_{c_1}^{c_2} f_{\theta_0}(t)dt = 1 - \alpha \qquad \text{and} \qquad \frac{d}{d\theta}\left[\int_{c_1}^{c_2} f_\theta(t)dt\right]\Bigg|_{\theta=\theta_0} = 0$$

(see Exercise 24). This test is then UMP among all tests that are unbiased and invariant. Whether it is also UMPU without the restriction to invariant tests is an open problem. ∎

6.3.2 UMPI tests in normal linear models

Consider normal linear model (6.38):

$$X = N_n(\beta Z^\tau, \sigma^2 I_n),$$

where β is a p-vector of unknown parameters, $\sigma^2 > 0$ is unknown, and Z is a fixed $n \times p$ matrix of rank $r \leq p < n$. In §6.2.3, UMPU tests for testing

(6.39) or (6.40) are derived. A frequently encountered problem in practice is to test

$$H_0 : \beta L^\tau = 0 \qquad \text{versus} \qquad H_1 : \beta L^\tau \neq 0, \tag{6.49}$$

where L is an $s \times p$ matrix of rank $s \leq r$ and all rows of L are in $\mathcal{R}(Z)$. However, a UMPU test for (6.49) does not exist if $s > 1$. We now derive a UMPI test for testing (6.49). We use without proof the following result from linear algebra: there exists an orthogonal matrix Γ such that

$$(\eta, 0) = \beta Z^\tau \Gamma,$$

where η is an r-vector and 0 is the $(n-r)$-vector of 0's, and that (6.49) is equivalent to

$$H_0 : \eta_1 = 0 \qquad \text{versus} \qquad H_1 : \eta_1 \neq 0, \tag{6.50}$$

where η_1 is the s-vector containing the first s components of η. Let $Y = (Y_1, Y_2) = X\Gamma$, where Y_1 is $r \times 1$. Then $Y = N_n((\eta, 0), \sigma^2 I_n)$ with the p.d.f. given by (6.42). Let $Y_1 = (Y_{11}, Y_{12})$, where Y_{11} is $s \times 1$, and let

$$\mathcal{G} = \{g_{\Lambda,c,\gamma} : c \in \mathcal{R}^{r-s},\ \gamma > 0,\ \Lambda \text{ is an } s \times s \text{ orthogonal matrix}\}$$

with

$$g_{\Lambda,c,\gamma}(Y) = \gamma(Y_{11}\Lambda, Y_{12} + c, Y_2).$$

Testing (6.50) is invariant under $\mathcal{G}$. By Proposition 6.3, we can restrict our attention to the sufficient statistic $(Y_1, \|Y_2\|^2)$. The statistic

$$M(Y) = \|Y_{11}\|^2 / \|Y_2\|^2 \tag{6.51}$$

is clearly invariant. We now show that $M(Y)$ is maximal invariant. Let $u_i \in \mathcal{R}^s$, $u_i \neq 0$, and $t_i \in (0, \infty)$, $i = 1, 2$. If $\|u_1\|^2/t_1^2 = \|u_2\|^2/t_2^2$, then $t_1 = \gamma t_2$ with $\gamma = \|u_1\|/\|u_2\|$. Since $u_1/\|u_1\|$ and $u_2/\|u_2\|$ are two points having the same distance from the origin, there exists an orthogonal matrix Λ such that $u_1/\|u_1\| = (u_2/\|u_2\|)\Lambda$, i.e., $u_1 = \gamma u_2 \Lambda$. This proves that if $M(y^{(1)}) = M(y^{(2)})$ for two points $y^{(1)}$ and $y^{(2)}$ in $\mathcal{R}^n$, then $y_{11}^{(1)} = \gamma y_{11}^{(2)} \Lambda$ and $\|y_2^{(1)}\| = \gamma \|y_2^{(2)}\|$ for some $\gamma > 0$ and orthogonal matrix Λ and, therefore, $y^{(1)} = g_{\Lambda,c,\gamma}(y^{(2)})$ with $c = \gamma^{-1} y_{12}^{(1)} - y_{12}^{(2)}$. Thus, $M(Y)$ is maximal invariant under $\mathcal{G}$.

It can be shown (exercise) that $W = M(Y)(n-r)/s$ has the noncentral F-distribution $F_{s,n-r}(\theta)$ with $\theta = \|\eta_1\|^2/\sigma^2$ (see §1.3.1). Let $f_\theta(w)$ be the Lebesgue p.d.f. of W, i.e., f_θ is given by (1.33) with $n_1 = s$, $n_2 = n - r$, and $\delta = \theta$. Note that under H_0, $\theta = 0$ and f_θ reduces to the p.d.f. of the central F-distribution $F_{s,n-r}$ (Table 1.2, page 20). Also, it can be shown (exercise) that the ratio $f_{\theta_1}(w)/f_0(w)$ is an increasing function of w for any given $\theta_1 \neq 0$. By Theorem 6.1, a UMPI test of size α for testing $H_0 : \theta = 0$ versus $H_1 : \theta = \theta_1$ rejects H_0 when $W > F_{s,n-r,\alpha}$, where $F_{s,n-r,\alpha}$ is the

$(1-\alpha)$th quantile of the F-distribution $F_{s,n-r}$. Since this test does not depend on θ_1, by Lemma 6.1, it is also a UMPI test of size α for testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$, which is equivalent to testing (6.50).

In applications it is not convenient to carry out the test by finding explicitly the orthogonal matrix Γ. Hence, we now express the statistic W in terms of X. Since $Y = X\Gamma$ and $E(Y) = E(X)\Gamma = \beta Z^\tau \Gamma$,

$$\|Y_1 - \eta\|^2 + \|Y_2\|^2 = \|X - \beta Z^\tau\|^2$$

and, therefore,

$$\min_\eta \|Y_1 - \eta\|^2 + \|Y_2\|^2 = \min_\beta \|X - \beta Z^\tau\|^2,$$

which is the same as

$$\|Y_2\|^2 = \|X - \hat\beta Z^\tau\|^2 = SSR,$$

where $\hat\beta$ is the LSE defined by (3.27). Similarly,

$$\|Y_{11}\|^2 + \|Y_2\|^2 = \min_{\beta:\beta L^\tau = 0} \|X - \beta Z^\tau\|^2.$$

If we define $\hat\beta_{H_0}$ to be a solution of

$$\|X - \hat\beta_{H_0} Z^\tau\|^2 = \min_{\beta:\beta L^\tau = 0} \|X - \beta Z^\tau\|^2,$$

which is called the LSE of β under H_0 or the LSE of β subject to $\beta L^\tau = 0$, then

$$W = \frac{(\|X - \hat\beta_{H_0} Z^\tau\|^2 - \|X - \hat\beta Z^\tau\|^2)/s}{\|X - \hat\beta Z^\tau\|^2/(n-r)}. \tag{6.52}$$

Thus, the UMPI test for (6.49) can be used without finding Γ.

When $s = 1$, the UMPI test derived here is the same as the UMPU test for (6.40) given in §6.2.3.

Example 6.18. Consider the one-way ANOVA model in Example 3.13:

$$X_{ij} = N(\mu_i, \sigma^2), \qquad j = 1, ..., n_i,\ i = 1, ..., m,$$

and X_{ij}'s are independent. A common testing problem in applications is the test for homogeneity of means, i.e.,

$$H_0 : \mu_1 = \cdots = \mu_m \qquad \text{versus} \qquad H_1 : \mu_i \neq \mu_k \text{ for some } i \neq k. \tag{6.53}$$

One can easily find a matrix L for which (6.53) is equivalent to (6.49). But it is not necessary to find such a matrix in order to compute the

statistic W that defines the UMPI test. Note that the LSE of $(\mu_1, ..., \mu_m)$ is $(\bar{X}_{1\cdot}, ..., \bar{X}_{m\cdot})$, where $\bar{X}_{i\cdot}$ is the sample mean based on $X_{i1}, ..., X_{in_i}$, and the LSE under H_0 is simply $\bar{X}$, the sample mean based on all X_{ij}'s. Thus,

$$SSR = \|X - \hat{\beta}Z^\tau\|^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2,$$

$$SST = \|X - \hat{\beta}_{H_0}Z^\tau\|^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2,$$

and

$$SSA = SST - SSR = \sum_{i=1}^{m} n_i(\bar{X}_{i\cdot} - \bar{X})^2.$$

Then

$$W = \frac{SSA/(m-1)}{SSR/(n-m)},$$

where $n = \sum_{i=1}^{m} n_i$. The name ANOVA comes from the fact that the UMPI test is carried out by comparing two sources of variation: the variation within each group of observations (measured by SSR) and the variation among m groups (measured by SSA), and that $SSA + SSR = SST$ is the total variation in the data set.

In this case, the distribution of W can also be derived using Cochran's theorem (Theorem 1.5). See Exercise 64. ∎

Example 6.19. Consider the two-way balanced ANOVA model in Example 3.14:

$$X_{ijk} = N(\mu_{ij}, \sigma^2), \qquad i = 1, ..., a,\ j = 1, ..., b, k = 1, ..., c,$$

where $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$, $\sum_{i=1}^{a}\alpha_i = \sum_{j=1}^{b}\beta_j = \sum_{i=1}^{a}\gamma_{ij} = \sum_{j=1}^{b}\gamma_{ij} = 0$, and X_{ijk}'s are independent. Typically the following hypotheses are of interest:

$$H_0 : \alpha_i = 0 \text{ for all } i \qquad \text{versus} \qquad H_1 : \alpha_i \neq 0 \text{ for some } i, \tag{6.54}$$

$$H_0 : \beta_j = 0 \text{ for all } j \qquad \text{versus} \qquad H_1 : \beta_j \neq 0 \text{ for some } j, \tag{6.55}$$

and

$$H_0 : \gamma_{ij} = 0 \text{ for all } i, j \qquad \text{versus} \qquad H_1 : \gamma_{ij} \neq 0 \text{ for some } i, j. \tag{6.56}$$

In applications, α_i's are effects of a factor A (a variable taking finitely many values), β_j's are effects of a factor B, and γ_{ij}'s are effects of the interaction of factors A and B. Hence, testing hypotheses in (6.54), (6.55), and (6.56)

are the same as testing effects of factor A, of factor B, and of the interaction between A and B, respectively.

The LSE of μ, α_i, β_j, and γ_{ij} are given by (Example 3.14) $\hat{\mu} = \bar{X}_{...}$, $\hat{\alpha}_i = \bar{X}_{i..} - \bar{X}_{...}$, $\hat{\beta}_j = \bar{X}_{.j.} - \bar{X}_{...}$, $\hat{\gamma}_{ij} = \bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...}$, and a dot is used to denote averaging over the indicated subscript. Let

$$SSR = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk} - \bar{X}_{ij.})^2,$$

$$SSA = bc\sum_{i=1}^{a}(X_{i..} - \bar{X}_{...})^2,$$

$$SSB = ac\sum_{j=1}^{b}(X_{.j.} - \bar{X}_{...})^2,$$

and

$$SSC = c\sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij.} - X_{i..} - X_{.j.} + \bar{X}_{...})^2.$$

Then, one can show (exercise) that for testing (6.54), (6.55), and (6.56), the statistics W in (6.52) (for the UMPI tests) are, respectively,

$$\frac{SSA/(a-1)}{SSR/[(c-1)ab]}, \quad \frac{SSB/(b-1)}{SSR/[(c-1)ab]}, \quad \text{and} \quad \frac{SSC/[(a-1)(b-1)]}{SSR/[(c-1)ab]}. \quad \blacksquare$$

We end this section with a discussion of testing for random effects in the following balanced one-way random effects model (Example 3.17):

$$X_{ij} = \mu + A_i + e_{ij}, \qquad i = 1, ..., a, j = 1, ..., b, \tag{6.57}$$

where μ is an unknown parameter, A_i's are i.i.d. random effects from $N(0, \sigma_a^2)$, e_{ij}'s are i.i.d. measurement errors from $N(0, \sigma^2)$, and A_i's and e_{ij}'s are independent. Consider the problem of testing

$$H_0 : \sigma_a^2/\sigma^2 \le \Delta_0 \qquad \text{versus} \qquad H_1 : \sigma_a^2/\sigma^2 > \Delta_0 \tag{6.58}$$

for a given Δ_0. When Δ_0 is small, hypothesis H_0 in (6.58) means that the random effects are negligible relative to the measurement variation.

Let $(Y_{i1}, ..., Y_{ib}) = (X_{i1}, ..., X_{ib})\Gamma$, where Γ is a $b \times b$ orthogonal matrix whose elements in the first column are all equal to $1/\sqrt{b}$. Then

$$Y_{i1} = \sqrt{b}\bar{X}_{i.} = \sqrt{b}(\mu + A_i + \bar{e}_{i.}), \qquad i = 1, ..., a,$$

are i.i.d. from $N(\sqrt{b}\mu, \sigma^2 + b\sigma_a^2)$, Y_{ij}, $i = 1, ..., a, j = 2, ..., b$, are i.i.d. from $N(0, \sigma^2)$, and Y_{ij}'s are independent. The reason why $E(Y_{ij}) = 0$ when $j > 1$ is because column j of Γ is orthogonal to the first column of Γ.

Let Λ be an $a \times a$ orthogonal matrix whose elements in the first column are all equal to $1/\sqrt{a}$ and $(U_{11}, ..., U_{a1}) = (Y_{1\cdot}, ..., Y_{a\cdot})\Lambda$. Then $U_{11} = \sqrt{a}\bar{Y}_{\cdot 1}$ is $N(\sqrt{ab}\mu, \sigma^2 + b\sigma_a^2)$, U_{i1}, $i = 2, ..., a$, are from $N(0, \sigma^2 + b\sigma_a^2)$, and U_{i1}'s are independent. Let $U_{ij} = Y_{ij}$ for $j = 2, ..., b$, $i = 1, ..., a$.

The problem of testing (6.58) is invariant under the group of transformations that transform U_{11} to $rU_{11}+c$ and U_{ij} to rU_{ij}, $(i, j) \neq (1, 1)$, where $r > 0$ and $c \in \mathcal{R}$. It can be shown (exercise) that the maximal invariant under this group of transformations is SSA/SSR, where

$$SSA = \sum_{i=2}^{a} U_{i1}^2 \qquad \text{and} \qquad SSR = \sum_{i=1}^{a}\sum_{j=2}^{b} U_{ij}^2.$$

Note that H_0 in (6.58) is equivalent to $(\sigma^2 + b\sigma_a^2)/\sigma^2 \leq 1 + b\Delta_0$. Also, $SSA/(\sigma^2 + b\sigma_a^2)$ has the chi-square distribution χ^2_{a-1} and SSR/σ^2 has the chi-square distribution $\chi^2_{a(b-1)}$. Hence, the p.d.f. of the statistic

$$W = \frac{1}{1 + b\Delta_0} \frac{SSA/(a-1)}{SSR/[a(b-1)]}$$

is in a family indexed by $(\sigma^2 + b\sigma_a^2)/\sigma^2$ with monotone likelihood ratio in W. Thus, a UMPI test of size α for testing (6.58) rejects H_0 when $W > F_{a-1,a(b-1),\alpha}$, where $F_{a-1,a(b-1),\alpha}$ is the $(1-\alpha)$th quantile of the F-distribution $F_{a-1,a(b-1)}$.

It remains to express W in terms of X_{ij}'s. Note that

$$SSR = \sum_{i=1}^{a}\sum_{j=2}^{b} Y_{ij}^2 = \sum_{i=1}^{a}\left(\sum_{j=1}^{b} e_{ij}^2 - b\bar{e}_{i\cdot}^2\right) = \sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij} - \bar{X}_{i\cdot})^2$$

and

$$SSA = \sum_{i=1}^{a} U_{i1}^2 - U_{11}^2 = \sum_{i=1}^{a} Y_{i1}^2 - a\bar{Y}_{\cdot 1}^2 = b\sum_{i=1}^{a}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2.$$

The SSR and SSA derived here are the same as those in Example 6.18 when $n_i = b$ for all i and $m = a$. It can also be seen that if $\Delta_0 = 0$, then testing (6.58) is equivalent to testing $H_0 : \sigma_a^2 = 0$ versus $H_1 : \sigma_a^2 \neq 0$ and the derived UMPI test is exactly the same as that in Example 6.18, although the testing problems are different in these two cases.

Extensions to balanced two-way random effects models can be found in Lehmann (1986, §7.12).

6.4 Tests in Parametric Models

A UMP, UMPU, or UMPI test often does not exist in a particular problem. In the rest of this chapter we study some methods for constructing tests that have intuitive appeal and frequently coincide with optimal tests (UMP or UMPU tests) when optimal tests do exist. We consider tests in parametric models in this section, whereas tests in nonparametric models are studied in §6.5.

When the hypothesis H_0 is not simple, it is often difficult or even impossible to obtain a test that has exactly a given size α, since this involves finding a population P that maximizes the power function of the test over all $P \in \mathcal{P}_0$. In such cases a common approach is to find tests having asymptotic significance level α (Definition 2.13). This involves finding the limit of the power of a test at $P \in \mathcal{P}_0$, which is studied in this section and §6.5.

Throughout this section we assume that a sample X is from $P \in \mathcal{P} = \{f_\theta : \theta \in \Theta\}$, where f_θ's are p.d.f.'s w.r.t. a common σ-finite measure and $\Theta \subset \mathcal{R}^k$, and that the testing problem is

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1, \tag{6.59}$$

where $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$.

6.4.1 Likelihood ratio tests

When both H_0 and H_1 are simple (i.e., both $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$ are single-point sets), Theorem 6.1 applies and a UMP test rejects H_0 when

$$\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} > c_0 \tag{6.60}$$

for some $c_0 > 0$, which is equivalent to (exercise)

$$\frac{f_{\theta_0}(X)}{\max[f_{\theta_0}(X), f_{\theta_1}(X)]} < c \tag{6.61}$$

for some $c > 0$. The following definition is a natural extension of this idea.

Definition 6.6. Let $\ell(\theta) = f_\theta(X)$ be the likelihood function. For testing (6.59), a *likelihood ratio* (LR) test is any test that rejects H_0 if and only if $\lambda(X) < c$, where $c \in [0, 1]$ and $\lambda(X)$ is the likelihood ratio defined by

$$\lambda(X) = \frac{\sup\limits_{\theta \in \Theta_0} \ell(\theta)}{\sup\limits_{\theta \in \Theta} \ell(\theta)}. \quad \blacksquare$$

Note that $\lambda(X) \leq 1$. The rationale behind LR tests is that when H_0 is true, $\lambda(X)$ tends to be close to 1, whereas when H_1 is true, $\lambda(X)$ tends to be close to 0. LR tests are as widely applicable as MLE's in §4.4 and, in fact, they are closely related to MLE's. If $\hat{\theta}$ is an MLE of θ and $\hat{\theta}_0$ is an MLE of θ subject to $\theta \in \Theta_0$ (i.e., Θ_0 is treated as the parameter space), then

$$\lambda(X) = \ell(\hat{\theta}_0)/\ell(\hat{\theta}).$$

If the c.d.f. of $\lambda(X)$ is continuous, then an LR test of a given size α can be obtained by finding a $c_\alpha \in [0,1]$ such that

$$\sup_{\theta \in \Theta_0} P_\theta(\lambda(X) < c_\alpha) = \alpha.$$

On the other hand, it may not be possible to find an LR test of size α when the c.d.f. of $\lambda(X)$ is not continuous, which is due to the nonrandomized nature of LR tests. Of course, one can define randomized LR tests so that an LR test of size α can always be obtained in principle.

When a UMP, UMPU, or UMPI test exists, an LR test is often the same as this optimal test. For real-valued θ, we have the following result.

Proposition 6.5. Suppose that X has the p.d.f. given by (6.10) w.r.t. a σ-finite measure ν, where η is a strictly increasing function of θ.
(i) For testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, there is an LR test whose rejection region is the same as that of the UMP test T_* given by (6.11).
(ii) For testing the hypotheses in (6.12), there is an LR test whose rejection region is the same as that of the UMP test T_* given by (6.15).
(iii) For testing the hypotheses in (6.13) or (6.14), there is an LR test whose rejection region is equivalent to $Y(X) < c_1$ or $Y(X) > c_2$ for some constants c_1 and c_2.
Proof. (i) The condition on f_θ ensures that the MLE $\hat{\theta}$ of θ is a strictly increasing function of $Y(X)$. Hence, we only need to show that $\lambda(X) < c$ is equivalent to $\hat{\theta} > c_0$. Note that the p.d.f. f_θ has a strictly decreasing derivative for any given X. Thus, $\ell(\theta)$ is increasing when $\theta \leq \hat{\theta}$ and decreasing when $\theta > \hat{\theta}$, and

$$\lambda(X) = \begin{cases} 1 & \hat{\theta} \leq \theta_0 \\ \frac{\ell(\theta_0)}{\ell(\hat{\theta})} & \hat{\theta} > \theta_0. \end{cases}$$

Then $\lambda(X) < c$ is the same as $\hat{\theta} > c_0$ since $\ell(\theta)$ is increasing when $\theta \leq \hat{\theta}$.
(ii) The proof is similar to that in (i). Note that

$$\lambda(X) = \begin{cases} \frac{\ell(\theta_1)}{\ell(\hat{\theta})} & \hat{\theta} < \theta_1 \\ 1 & \theta_1 \leq \hat{\theta} \leq \theta_2 \\ \frac{\ell(\theta_2)}{\ell(\hat{\theta})} & \hat{\theta} > \theta_2. \end{cases}$$

Hence $\lambda(X) < c$ is equivalent to $c_1 < Y < c_2$.
(iii) The proof for (iii) is left as an exercise. ∎

Proposition 6.5 can be applied to problems concerning one-parameter exponential families such as the binomial, Poisson, negative binomial, and normal (with one parameter known) families. The following example shows that the same result holds in a situation where Proposition 6.5 is not applicable.

Example 6.20. Consider the testing problem $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ based on i.i.d. $X_1, ..., X_n$ from the uniform distribution $U(0, \theta)$. We now show that the UMP test with rejection region $X_{(n)} > \theta_0$ or $X_{(n)} \leq \theta_0 \alpha^{1/n}$ given in Exercise 17(c) is an LR test. Note that $\ell(\theta) = \theta^{-n} I_{(X_{(n)}, \infty)}(\theta)$. Hence

$$\lambda(X) = \begin{cases} (X_{(n)}/\theta_0)^n & X_{(n)} \leq \theta_0 \\ 0 & X_{(n)} > \theta_0 \end{cases}$$

and $\lambda(X) < c$ is equivalent to $X_{(n)} > \theta_0$ or $X_{(n)}/\theta_0 < c^{1/n}$. Taking $c = \alpha$ ensures that the LR test has size α. ∎

More examples of this kind can be found in §6.6. The next example considers multivariate θ.

Example 6.21. Consider normal linear model (6.38) and the hypotheses in (6.49). The likelihood function in this problem is

$$\ell(\theta) = \left(\tfrac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\tfrac{1}{2\sigma^2}\|X - \beta Z^\tau\|^2\right\},$$

where $\theta = (\beta, \sigma^2)$. Let $\hat{\beta}$ be the LSE defined by (3.27). Since $\|X - \beta Z^\tau\|^2 \geq \|X - \hat{\beta} Z^\tau\|^2$ for any β,

$$\ell(\theta) \leq \left(\tfrac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\tfrac{1}{2\sigma^2}\|X - \hat{\beta} Z^\tau\|^2\right\}.$$

Treating the right-hand side of the previous expression as a function of σ^2, it is easy to show that it has a maximum at $\sigma^2 = \hat{\sigma}^2 = \|X - \hat{\beta} Z^\tau\|^2/n$ and, therefore,

$$\sup_{\theta \in \Theta} \ell(\theta) = (2\pi\hat{\sigma}^2)^{-n/2} e^{-n/2}.$$

Similarly, let $\hat{\beta}_{H_0}$ be the LSE under H_0 and $\hat{\sigma}^2_{H_0} = \|X - \hat{\beta}_{H_0} Z^\tau\|^2/n$. Then

$$\sup_{\theta \in \Theta_0} \ell(\theta) = (2\pi\hat{\sigma}^2_{H_0})^{-n/2} e^{-n/2}.$$

Thus,

$$\lambda(X) = (\hat{\sigma}^2/\hat{\sigma}^2_{H_0})^{n/2} = \left(\frac{\|X - \hat{\beta} Z^\tau\|^2}{\|X - \hat{\beta}_{H_0} Z^\tau\|^2}\right)^{n/2} = \left(\frac{sW}{n-r} + 1\right)^{-n/2},$$

where W is given in (6.52). This shows that LR tests are the same as the UMPI tests derived in §6.3.2.

The one-sample or two-sample two-sided t-tests derived in §6.2.3 are special cases of LR tests. For a one-sample problem, we define $\beta = \mu$ and $Z^\tau = J_n$, the n-vector of ones. Note that $\hat{\beta} = \bar{X}$, $\hat{\sigma}^2 = (n-1)S^2/n$, $\hat{\beta}^2_{H_0} = 0$ ($H_0 : \beta = 0$), and $\hat{\sigma}^2_{H_0} = \|X\|^2/n = (n-1)S^2/n + \bar{X}^2$. Hence

$$\lambda(X) = \left[1 + \frac{n\bar{X}^2}{(n-1)S^2}\right]^{-n/2} = \left(1 + \frac{[t(X)]^2}{n-1}\right)^{-n/2},$$

where $t(X) = \sqrt{n}\bar{X}/S$ has the t-distribution t_{n-1} under H_0. Thus, $\lambda(X) < c$ is equivalent to $|t(X)| > c_0$, which is the rejection region of a one-sample two-sided t-test.

For a two-sample problem, we let $n = n_1 + n_2$, $\beta = (\mu_1, \mu_2)$, and

$$Z^\tau = \begin{pmatrix} J_{n_1} & 0 \\ 0 & J_{n_2} \end{pmatrix}.$$

Testing $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ is the same as testing (6.49) with $L = (1, -1)$. Since $\hat{\beta}_{H_0} = \bar{X}$ and $\hat{\beta} = (\bar{X}_1, \bar{X}_2)$, where $\bar{X}_1$ and $\bar{X}_2$ are the sample means based on $X_1, ..., X_{n_1}$ and $X_{n_1+1}, ..., X_n$, respectively, we have

$$n\hat{\sigma}^2 = \sum_{i=1}^{n_1}(X_i - \bar{X}_1)^2 + \sum_{i=n_1+1}^{n}(X_i - \bar{X}_2)^2 = (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2$$

and

$$n\hat{\sigma}^2_{H_0} = (n-1)S^2 = n^{-1}n_1n_2(\bar{X}_1 - \bar{X}_2)^2 + (n_1 - 1)S_1^2 + (n_2 - 1)S_2^2.$$

Therefore, $\lambda(X) < c$ is equivalent to $|t(X)| > c_0$, where $t(X)$ is given by (6.37), and LR tests are the same as the two-sample two-sided t-tests in §6.2.3. ∎

6.4.2 Asymptotic tests based on likelihoods

As we can see from Proposition 6.5 and the previous examples, an LR test is often equivalent to a test based on a statistic $Y(X)$ whose distribution under H_0 can be used to determine the rejection region of the LR test with size α. When this technique fails, it is difficult or even impossible to find an LR test with size α, even if the c.d.f. of $\lambda(X)$ is continuous. The following result shows that in the i.i.d. case we can obtain the asymptotic distribution (under H_0) of the likelihood ratio $\lambda(X)$ so that an LR test

having asymptotic significance level α can be obtained. Assume that Θ_0 is determined by

$$H_0 : \theta = g(\vartheta), \tag{6.62}$$

where ϑ is a $(k-r)$-vector of unknown parameters and g is a continuously differentiable function from $\mathcal{R}^{k-r}$ to $\mathcal{R}^k$. For example, if $\Theta = \mathcal{R}^2$ and $\Theta_0 = \{(\theta_1, \theta_2) \in \Theta : \theta_1 = 0\}$, then $\vartheta = \theta_2$, $g_1(\vartheta) = 0$, and $g_2(\vartheta) = \vartheta$.

Theorem 6.5. Assume the conditions in Theorem 4.16. Suppose that H_0 is determined by (6.62).
(i) Under H_0, $-2\log\lambda_n \to_d \chi^2_r$, where $\lambda_n = \lambda(X)$ and χ^2_r is a random variable having the chi-square distribution χ^2_r.
(ii) The LR test with rejection region $\lambda_n < e^{-\chi^2_{r,\alpha}/2}$ has asymptotic significance level α, where $\chi^2_{r,\alpha}$ is the $(1-\alpha)$th quantile of the chi-square distribution χ^2_r.
Proof. Part (ii) is a direct consequence of part (i). Thus, we only need to prove part (i). Without loss of generality, we assume that there exist an MLE $\hat{\theta}$ and an MLE $\hat{\vartheta}$ under H_0 such that

$$\sup_{\theta\in\Theta} \ell(\theta) = \ell(\hat{\theta})$$

and

$$\sup_{\theta\in\Theta_0} \ell(\theta) = \sup_{\vartheta} \ell(g(\vartheta)) = \ell(g(\hat{\vartheta})).$$

Following the proof of Theorem 4.17 in §4.5.2, we can obtain that

$$\sqrt{n}(\hat{\theta} - \theta)\imath(\theta) = n^{-1/2} s_n(\theta) + o_p(1),$$

where $s_n(\theta) = \partial \log \ell(\theta)/\partial\theta$ and $\imath(\theta)$ is the Fisher information about θ contained in a single observation X_i, and that

$$2[\log \ell(\hat{\theta}) - \log \ell(\theta)] = n(\hat{\theta} - \theta)\imath(\theta)(\hat{\theta} - \theta)^\tau + o_p(1).$$

Then

$$2[\log \ell(\hat{\theta}) - \log \ell(\theta)] = n^{-1} s_n(\theta)[\imath(\theta)]^{-1}[s_n(\theta)]^\tau + o_p(1).$$

Similarly, under H_0,

$$2[\log \ell(g(\hat{\vartheta})) - \log \ell(g(\vartheta))] = n^{-1}\tilde{s}_n(\vartheta)[\jmath(\vartheta)]^{-1}[\tilde{s}_n(\vartheta)]^\tau + o_p(1),$$

where

$$\tilde{s}_n(\vartheta) = \frac{\partial \log \ell(g(\vartheta))}{\partial\vartheta} = \frac{\partial \log \ell(\theta)}{\partial\theta}\frac{\partial g(\vartheta)}{\partial\vartheta}$$

and $\jmath(\vartheta)$ is the Fisher information about ϑ (under H_0) contained in X_i. Combining these results, we obtain that

$$\begin{aligned}-2\log\lambda_n &= 2[\log\ell(\hat\theta) - \log\ell(g(\hat\vartheta))]\\ &= n^{-1}s_n(g(\vartheta))B(\vartheta)[s_n(g(\vartheta))]^\tau + o_p(1)\end{aligned}$$

under H_0, where

$$B(\vartheta) = [\imath(g(\vartheta))]^{-1} - D(\vartheta)[\jmath(\vartheta)]^{-1}[D(\vartheta)]^\tau$$

and $D(\vartheta) = \partial g(\vartheta)/\partial\vartheta$. By the CLT, $n^{-1/2}s_n(\theta) \to_d N_k\big(0, \imath(\theta)\big)$. Then, it follows from Theorem 1.10(iii) that, under H_0,

$$-2\log\lambda_n \to_d Z[\imath(g(\vartheta))]^{1/2}B(\vartheta)[\imath(g(\vartheta))]^{1/2}Z^\tau,$$

where $Z = N_k(0, I_k)$. Let $D = D(\vartheta)$, $B = B(\vartheta)$, $A = \imath(g(\vartheta))$, and $C = \jmath(\vartheta)$. Then

$$\begin{aligned}(A^{1/2}BA^{1/2})^2 &= A^{1/2}BABA^{1/2}\\ &= A^{1/2}(A^{-1} - DC^{-1}D^\tau)A(A^{-1} - DC^{-1}D^\tau)A^{1/2}\\ &= (I_k - A^{1/2}DC^{-1}D^\tau A^{1/2})(I_k - A^{1/2}DC^{-1}D^\tau A^{1/2})\\ &= I_k - 2A^{1/2}DC^{-1}D^\tau A^{1/2} + A^{1/2}DC^{-1}D^\tau ADC^{-1}D^\tau A^{1/2}\\ &= I_k - A^{1/2}DC^{-1}D^\tau A^{1/2}\\ &= A^{1/2}BA^{1/2},\end{aligned}$$

where the next to last equality follows from the fact that $C = D^\tau AD$. This shows that $A^{1/2}BA^{1/2}$ is a projection matrix. The rank of $A^{1/2}BA^{1/2}$ is

$$\begin{aligned}\mathrm{tr}(A^{1/2}BA^{1/2}) &= \mathrm{tr}(I_k - DC^{-1}D^\tau A)\\ &= k - \mathrm{tr}(C^{-1}D^\tau AD)\\ &= k - \mathrm{tr}(C^{-1}C)\\ &= k - (k - r)\\ &= r.\end{aligned}$$

Thus, there is an orthogonal matrix Γ such that

$$A^{1/2}BA^{1/2} = \Gamma\begin{pmatrix} I_r & 0\\ 0 & 0\end{pmatrix}\Gamma^\tau.$$

Since $Z\Gamma = N_k(0, I_k)$, $Z[\imath(g(\vartheta))]^{1/2}B(\vartheta)[\imath(g(\vartheta))]^{1/2}Z^\tau = \chi^2_r$. ∎

As an example, Theorem 6.5 can be applied to testing problems in Example 4.33 where the exact rejection region of the LR test of size α is difficult to obtain but the likelihood ratio λ_n can be calculated numerically.

Tests whose rejection regions are constructed using asymptotic theory (so that these tests have asymptotic significance level α) are called *asymptotic tests*, which are useful when a test of exact size α is difficult to find. There are two popular asymptotic tests based on likelihoods that are asymptotically equivalent to LR tests. Note that the hypothesis in (6.62) is equivalent to a set of $r \leq k$ equations:

$$H_0 : R(\theta) = 0, \tag{6.63}$$

where $R(\theta)$ is a continuously differentiable function from $\mathcal{R}^k$ to $\mathcal{R}^r$. Wald (1943) introduced a test that rejects H_0 if and only if $W_n > c$, where

$$W_n = R(\hat{\theta})\{C(\hat{\theta})[I_n(\hat{\theta})]^{-1}[C(\hat{\theta})]^\tau\}^{-1}[R(\hat{\theta})]^\tau,$$

$I_n(\theta)$ is the Fisher information matrix based on $X_1, ..., X_n$, $\hat{\theta}$ is an MLE or RLE of θ, and $C(\theta) = \partial R(\theta)/\partial\theta$.

Rao (1947) introduced a *score* test that rejects H_0 if and only if $R_n > c$, where

$$R_n = s_n(g(\hat{\vartheta}))[I_n(g(\hat{\vartheta}))]^{-1}[s_n(g(\hat{\vartheta}))]^\tau,$$

g is given in (6.62), $s_n(\theta) = \partial \log \ell(\theta)/\partial\theta$ is the score function, and $\hat{\vartheta}$ is an MLE or RLE of ϑ under H_0 in (6.62).

Theorem 6.6. Assume the conditions in Theorem 4.16.
(i) Suppose that H_0 is determined by (6.63). Under H_0, $W_n \to_d \chi^2_r$ and, therefore, the test rejects H_0 if and only if $W_n > \chi^2_{r,\alpha}$ has asymptotic significance level α, where $\chi^2_{r,\alpha}$ is the $(1-\alpha)$th quantile of the chi-square distribution χ^2_r.
(ii) Suppose that H_0 is determined by (6.62). Under H_0, $R_n \to_d \chi^2_r$ and, therefore, the test rejects H_0 if and only if $R_n > \chi^2_{r,\alpha}$ has asymptotic significance level α.
Proof. (i) Using Theorems 1.12 and 4.17,

$$\sqrt{n}[R(\hat{\theta}) - R(\theta)] \to_d N_r(0, C(\theta)[\imath(\theta)]^{-1}[C(\theta)]^\tau),$$

where $\imath(\theta) = n^{-1}I_n(\theta)$. Under H_0, $R(\theta) = 0$ and, therefore,

$$nR(\hat{\theta})\{C(\theta)[\imath(\theta)]^{-1}[C(\theta)]^\tau\}^{-1}[R(\hat{\theta})]^\tau \to_d \chi^2_r$$

(Theorem 1.10). The result follows from

$$C(\hat{\theta})[\imath(\hat{\theta})]^{-1}[C(\hat{\theta})]^\tau \to_p C(\theta)[\imath(\theta)]^{-1}[C(\theta)]^\tau,$$

since both $\imath(\theta)$ and $C(\theta)$ are continuous at θ.
(ii) The proof of a special case is given as an exercise. For the proof of the general case, see Sen and Singer (1993, pp. 242-244). ∎

Thus, Wald's tests, Rao's score tests, and LR tests are asymptotically equivalent. Note that Wald's test requires computing $\hat{\theta}$, not $\hat{\vartheta}$, whereas Rao's score test requires computing $\hat{\vartheta}$, not $\hat{\theta}$. On the other hand, an LR test requires computing both $\hat{\theta}$ and $\hat{\vartheta}$ (or solving two maximization problems). Hence, one may choose one of these tests which is easy to compute in a particular application.

The results in Theorems 6.5 and 6.6 can be extended to non-i.i.d. situations (e.g., the GLM in §4.4.2). We state without proof the following result.

Theorem 6.7. Assume the conditions in Theorem 4.18. Consider the problem of testing H_0 in (6.63) (or equivalently, (6.62)) with $\theta = (\beta, \phi)$. Then the results in Theorems 6.5 and 6.6 still hold. ∎

Example 6.22. Consider the GLM (4.55)-(4.58) with t_i's in a fixed interval (t_0, t_∞), $0 < t_0 \leq t_\infty < \infty$. Then the Fisher information matrix

$$I_n(\theta) = \begin{pmatrix} \phi^{-1} M_n(\beta) & 0 \\ 0 & \tilde{I}_n(\beta, \phi) \end{pmatrix},$$

where $M_n(\beta)$ is given by (4.60) and $\tilde{I}_n(\beta, \phi)$ is the Fisher information about ϕ.

Consider the problem of testing $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$, where β_0 is a fixed vector. Let $(\hat{\beta}, \hat{\phi})$ be the MLE (or RLE) of (β, ϕ). Then, Wald's test is based on

$$W_n = (\hat{\beta} - \beta_0) M_n(\hat{\beta}) (\hat{\beta} - \beta_0)^\tau / \hat{\phi}$$

and Rao's score test is based on

$$R_n = \tilde{s}_n(\beta_0) [M_n(\beta_0)]^{-1} [\tilde{s}_n(\beta_0)]^\tau / \tilde{\phi},$$

where $\tilde{s}_n(\beta)$ is given by (4.65) and $\tilde{\phi}$ is a solution of $\partial \log \ell(\beta_0, \phi)/\partial \phi = 0$. It follows from Theorem 4.18 that both W_n and R_n are asymptotically distributed as χ^2_r under H_0. ∎

Wald's tests, Rao's score tests, and LR tests are typically consistent according to Definition 2.13(iii). They are also Chernoff-consistent (Definition 2.13(iv)) if α is chosen to be $\alpha_n \to 0$ and $\chi^2_{r,\alpha_n} = o(n)$ as $n \to \infty$ (exercise). Other asymptotic optimality properties of these tests are discussed in Wald (1943); see also Serfling (1980, Chapter 10).

6.4.3 χ^2-tests

A test that is related to the asymptotic tests described in §6.4.2 is the so-called χ^2-test for testing cell probabilities in a multinomial distribu-

tion. Consider a sequence of n independent trials with k possible outcomes for each trial. Let $p_j > 0$ be the cell probability of occurrence of the jth outcome in any given trial and X_j be the number of occurrences of the jth outcome in n trials. Then $X = (X_1, ..., X_k)$ has the multinomial distribution (Example 2.7) with the parameter $\boldsymbol{P} = (p_1, ..., p_k)$. Let $\xi_i = (0, ..., 0, 1, 0, ..., 0)$, where the single nonzero component 1 is located in the jth position if the ith trial yields the jth outcome. Then $\xi_1, ..., \xi_n$ are i.i.d. and $X/n = \bar{\xi} = \sum_{i=1}^n \xi_i/n$. By the CLT,

$$Z_n(\boldsymbol{P}) = \sqrt{n}\left(\frac{X}{n} - \boldsymbol{P}\right) = \sqrt{n}(\bar{\xi} - \boldsymbol{P}) \to_d N_k(0, \Sigma), \tag{6.64}$$

where $\Sigma = \mathrm{Var}(X/\sqrt{n})$ is a symmetric $k \times k$ matrix whose ith diagonal element is $p_i(1 - p_i)$ and (i, j)th off-diagonal element is $-p_i p_j$.

Consider the problem of testing

$$H_0 : \boldsymbol{P} = \boldsymbol{P}_0 \qquad \text{versus} \qquad H_1 : \boldsymbol{P} \neq \boldsymbol{P}_0, \tag{6.65}$$

where $\boldsymbol{P}_0 = (p_{01}, ..., p_{0k})$ is a known vector of cell probabilities. A popular test for (6.65) is based on the following χ^2-statistic:

$$\chi^2 = \sum_{j=1}^k \frac{(X_j - np_{0j})^2}{np_{0j}} = \|Z_n(\boldsymbol{P}_0)D(\boldsymbol{P}_0)\|^2, \tag{6.66}$$

where $Z_n(\boldsymbol{P})$ is given by (6.64) and $D(\boldsymbol{P})$ is the $k \times k$ diagonal matrix whose jth diagonal element is $p_i^{-1/2}$. Another popular test is based on the following modified χ^2-statistic:

$$\tilde{\chi}^2 = \sum_{j=1}^k \frac{(X_j - np_{0j})^2}{X_j} = \|Z_n(\boldsymbol{P}_0)D(X/n)\|^2. \tag{6.67}$$

Note that X/n is an unbiased estimator of $\boldsymbol{P}$.

Theorem 6.8. Let $\phi = (\sqrt{p_1}, ..., \sqrt{p_k})$ and Λ be a $k \times k$ projection matrix.
(i) If $\phi\Lambda = a\phi$, then

$$Z_n(\boldsymbol{P})D(\boldsymbol{P})\Lambda[Z_n(\boldsymbol{P})D(\boldsymbol{P})]^\tau \to_d \chi_r^2,$$

where χ_r^2 has the chi-square distribution χ_r^2 with $r = \mathrm{tr}(\Lambda) - a$.
(ii) The same result holds if $D(\boldsymbol{P})$ in (i) is replaced by $D(X/n)$.
Proof. (i) Let $D = D(\boldsymbol{P})$, $Z_n = Z_n(\boldsymbol{P})$, and $Z = N_k(0, I_k)$. From (6.64) and Theorem 1.10,

$$Z_n D\Lambda(Z_n D)^\tau \to_d ZAZ^\tau \qquad \text{with} \quad A = \Sigma^{1/2} D\Lambda D\Sigma^{1/2}.$$

Following the proof of Theorem 6.5, the result in (i) follows if we can show that $A^2 = A$ (i.e., A is a projection matrix) and $r = \mathrm{tr}(A)$. Since Λ is a projection matrix and $\phi\Lambda = a\phi$, a must be either 0 or 1. Note that

$$D\Sigma D = I_k - \phi^\tau \phi.$$

Then

$$\begin{aligned} A^3 &= \Sigma^{1/2} D\Lambda D\Sigma D\Lambda D\Sigma D\Lambda D\Sigma^{1/2} \\ &= \Sigma^{1/2} D(\Lambda - a\phi^\tau\phi)(\Lambda - a\phi^\tau\phi)\Lambda D\Sigma^{1/2} \\ &= \Sigma^{1/2} D(\Lambda - 2a\phi^\tau\phi + a^2\phi^\tau\phi)\Lambda D\Sigma^{1/2} \\ &= \Sigma^{1/2} D(\Lambda - a\phi^\tau\phi)\Lambda D\Sigma^{1/2} \\ &= \Sigma^{1/2} D\Lambda D\Sigma D\Lambda D\Sigma^{1/2} \\ &= A^2, \end{aligned}$$

which implies that the eigenvalues of A must be 0 or 1. Therefore, $A^2 = A$. Also,

$$\mathrm{tr}(A) = \mathrm{tr}[\Lambda(D\Sigma D)] = \mathrm{tr}(\Lambda - a\phi^\tau\phi) = \mathrm{tr}(\Lambda) - a.$$

(ii) The result in (ii) follows from the result in (i) and the fact that $X/n \to_p \boldsymbol{P}$. ∎

Note that the χ^2-statistic in (6.66) and the modified χ^2-statistic in (6.67) are special cases of the statistics in Theorem 6.8(i) and (ii), respectively, with $\Lambda = I_k$ satisfying $\phi\Lambda = \phi$. Hence, a test of asymptotic significance level α for testing (6.65) rejects H_0 when $\chi^2 > \chi^2_{k-1,\alpha}$ (or $\tilde{\chi}^2 > \chi^2_{k-1,\alpha}$), where $\chi^2_{k-1,\alpha}$ is the $(1-\alpha)$th quantile of χ^2_{k-1}. These tests are called (asymptotic) χ^2-tests.

Example 6.23 (Goodness of fit tests). Let $Y_1, ..., Y_n$ be i.i.d. from F. Consider the problem of testing

$$H_0 : F = F_0 \qquad \text{versus} \qquad H_1 : F \neq F_0, \tag{6.68}$$

where F_0 is a known c.d.f. For instance, $F_0 = N(0,1)$. One way to test (6.68) is to partition the range of Y_1 into k disjoint events $A_1, ..., A_k$ and test (6.65) with $p_j = P_F(A_j)$ and $p_{0j} = P_{F_0}(A_j)$, $j = 1, ..., k$. Let X_j be the number of Y_i's in A_j, $j = 1, ..., k$. Based on X_j's, the χ^2-tests discussed previously can be applied to this problem and they are called *goodness of fit* tests. ∎

In the goodness of fit tests discussed in Example 6.23, F_0 in H_0 is known so that p_{0j}'s can be computed. In some cases we need to test the following hypotheses that are slightly different from those in (6.68):

$$H_0 : F = F_\theta \qquad \text{versus} \qquad H_1 : F \neq F_\theta, \tag{6.69}$$

where θ is an unknown parameter in $\Theta \subset \mathcal{R}^s$. For example, $F_\theta = N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$. If we still try to test (6.65) with $p_j = P_{F_\theta}(A_j)$, $j = 1, ..., k$, the result in Example 6.23 is not applicable since $\boldsymbol{P}$ is unknown under H_0. A generalized χ^2-test for (6.69) can be obtained using the following result. Let $\boldsymbol{P}(\theta) = (p_1(\theta), ..., p_k(\theta))$ be a k-vector of known functions of $\theta \in \Theta \subset \mathcal{R}^s$, where $s < k$. Consider the testing problem

$$H_0 : \boldsymbol{P} = \boldsymbol{P}(\theta) \qquad \text{versus} \qquad H_1 : \boldsymbol{P} \neq \boldsymbol{P}(\theta). \tag{6.70}$$

Note that (6.65) is the special case of (6.70) with $s = 0$, i.e., θ is known. Let $\hat{\theta}$ be an MLE of θ under H_0. Then, by Theorem 6.5, the LR test that rejects H_0 when $-2\log\lambda_n > \chi^2_{k-s-1,\alpha}$ has asymptotic significance level α, where $\chi^2_{k-s-1,\alpha}$ is the $(1-\alpha)$th quantile of χ^2_{k-s-1} and

$$\lambda_n = \prod_{j=1}^k [p_j(\hat{\theta})]^{X_j} \Big/ (X_j/n)^{X_j}.$$

Using the fact that $p_j(\hat{\theta})/(X_j/n) \to_p 1$ under H_0 and

$$\log(1+x) = x - x^2/2 + o(|x|^2) \qquad \text{as } |x| \to 0,$$

we obtain that

$$\begin{aligned}
-2\log\lambda_n &= -2\sum_{j=1}^k X_j \log\left(1 + \frac{p_j(\hat{\theta})}{X_j/n} - 1\right) \\
&= -2\sum_{j=1}^k X_j \left(\frac{p_j(\hat{\theta})}{X_j/n} - 1\right) + \sum_{j=1}^k X_j \left(\frac{p_j(\hat{\theta})}{X_j/n} - 1\right)^2 + o_p(1) \\
&= \sum_{j=1}^k \frac{[X_j - np_j(\hat{\theta})]^2}{X_j} + o_p(1) \\
&= \sum_{j=1}^k \frac{[X_j - np_j(\hat{\theta})]^2}{np_j(\hat{\theta})} + o_p(1),
\end{aligned}$$

where the third equality follows from $\sum_{j=1}^k p_j(\hat{\theta}) = \sum_{j=1}^k X_j/n = 1$. Define the generalized χ^2-statistics χ^2 and $\tilde{\chi}^2$ to be the χ^2 and $\tilde{\chi}^2$ in (6.66) and (6.67), respectively, with p_{0j}'s replaced by $p_j(\hat{\theta})$'s. We then have the following result.

Theorem 6.9. Under H_0 given by (6.70), the generalized χ^2-statistics converge in distribution to χ^2_{k-s-1}. The χ^2-test with rejection region $\chi^2 > \chi^2_{k-s-1,\alpha}$ (or $\tilde{\chi}^2 > \chi^2_{k-s-1,\alpha}$) has asymptotic significance level α, where $\chi^2_{k-s-1,\alpha}$ is the $(1-\alpha)$th quantile of χ^2_{k-s-1}. ∎

Theorem 6.9 can be applied to derive a goodness of fit test for hypotheses (6.69). However, one has to formulate (6.70) and compute an MLE of θ under $H_0 : \boldsymbol{P} = \boldsymbol{P}(\theta)$, which is different from an MLE under $H_0 : F = F_\theta$ unless (6.69) and (6.70) are the same; see Moore and Spruill (1975). The next example is the main application of Theorem 6.9.

Example 6.24 ($r \times c$ contingency tables). In Example 6.12 we considered the 2×2 contingency table. The following $r \times c$ contingency table is a natural extension:

	A_1	A_2	$\cdots$	A_c	Total
B_1	X_{11}	X_{12}	$\cdots$	X_{1c}	n_1
B_2	X_{21}	X_{22}	$\cdots$	X_{2c}	n_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
B_r	X_{r1}	X_{r2}	$\cdots$	X_{rc}	n_r
Total	m_1	m_2	$\cdots$	m_c	n

where A_i's are disjoint events with $A_1 \cup \cdots \cup A_c = \Omega$ (the sample space of a random experiment), B_i's are disjoint events with $B_1 \cup \cdots \cup B_r = \Omega$, and X_{ij} is the observed frequency of the outcomes in $A_j \cap B_i$. Similar to the case of the 2×2 contingency table discussed in Example 6.12, there are two important applications in this problem. We first consider testing independence of $\{A_j : j = 1, ..., c\}$ and $\{B_i : i = 1, ..., r\}$ with hypotheses

$$H_0 : p_{ij} = p_i q_j \ \text{ for all } i, j \qquad \text{versus} \qquad H_1 : p_{ij} \neq p_i q_j \ \text{ for some } i, j,$$

where $p_{ij} = P(A_j \cap B_i) = E(X_{ij})/n$, $p_i = P(B_i)$, and $q_j = P(A_j)$, $i = 1, ..., r$, $j = 1, ..., c$. In this case, $X = (X_{ij}, i = 1, ..., r, j = 1, ..., c)$ has the multinomial distribution with parameters p_{ij}, $i = 1, ..., r$, $j = 1, ..., c$. Under H_0, MLE's of p_i's and q_j's are $\bar{X}_{i\cdot} = n_i/n$ and $\bar{X}_{\cdot j} = m_j/n$, $i = 1, ..., r$, $j = 1, ..., c$ (exercise). By Theorem 6.9, the χ^2-test rejects H_0 when $\chi^2 > \chi^2_{(r-1)(c-1),\alpha}$, where

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(X_{ij} - n\bar{X}_{i\cdot}\bar{X}_{\cdot j})^2}{n\bar{X}_{i\cdot}\bar{X}_{\cdot j}} \tag{6.71}$$

and $\chi^2_{(r-1)(c-1),\alpha}$ is the $(1-\alpha)$th quantile of the chi-square distribution $\chi^2_{(r-1)(c-1)}$ (exercise). One can also obtain the modified χ^2-test by replacing $n\bar{X}_{i\cdot}\bar{X}_{\cdot j}$ by X_{ij} in the denominator of each term of the sum in (6.71).

Next, suppose that $(X_{1j}, ..., X_{rj})$, $j = 1, ..., c$, are c independent random vectors having the multinomial distributions with parameters $(p_{1j}, ..., p_{rj})$, $j = 1, ..., c$, respectively. Consider the problem of testing whether c multinomial distributions are the same, i.e.,

$$H_0 : p_{ij} = p_{i1} \ \text{ for all } i, j \qquad \text{versus} \qquad H_1 : p_{ij} \neq p_{i1} \ \text{ for some } i, j.$$

It turns out that the rejection region of the χ^2-test given in Theorem 6.9 is still $\chi^2 > \chi^2_{(r-1)(c-1),\alpha}$ with χ^2 given by (6.71) (exercise).

One can also obtain the LR test in this problem. When $r = c = 2$, the LR test is equivalent to Fisher's test given in Example 6.12, which is a UMPU test. When $r > 2$ or $c > 2$, however, a UMPU test does not exist in this problem. ∎

6.4.4 Bayes tests

An LR test actually compares $\sup_{\theta\in\Theta_0} \ell(\theta)$ with $\sup_{\theta\in\Theta_1} \ell(\theta)$ for testing (6.59). Instead of comparing two maximum values, one may compare two averages such as $\hat{\pi}_j = \int_{\Theta_j} \ell(\theta)d\Pi(\theta)/\int_\Theta \ell(\theta)d\Pi(\theta)$, $j = 0, 1$, where $\Pi(\theta)$ is a c.d.f. on Θ, and reject H_0 when $\hat{\pi}_1 > \hat{\pi}_0$. If Π is treated as a prior c.d.f., then $\hat{\pi}_j$ is the posterior probability of Θ_j, and this test is a particular Bayes action (see Exercise 13 in §4.6) and is called a Bayes test.

In Bayesian analysis one often considers the *Bayes factor* defined to be

$$\beta = \frac{\text{posterior odds ratio}}{\text{prior odds ratio}} = \frac{\hat{\pi}_0/\hat{\pi}_1}{\pi_0/\pi_1},$$

where $\pi_j = \Pi(\Theta_j)$ is the prior probability of Θ_j.

Clearly, if there is a statistic sufficient for θ, then the Bayes test and Bayes factor depend only on the sufficient statistic.

Consider the special case where $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$ are simple hypotheses. Then $\Pi(\Theta_j) = \pi_j$ and, for given $X = x$,

$$\hat{\pi}_j = \frac{\pi_j f_{\theta_j}(x)}{\pi_0 f_{\theta_0}(x) + \pi_1 f_{\theta_1}(x)}.$$

Rejecting H_0 when $\hat{\pi}_1 > \hat{\pi}_0$ is the same as rejecting H_0 when

$$\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} > \frac{\pi_0}{\pi_1}. \tag{6.72}$$

This is equivalent to the UMP test T_* in (6.3) (Theorem 6.1) with $c = \pi_0/\pi_1$ and $\gamma = 0$. The Bayes factor in this case is

$$\beta = \frac{\hat{\pi}_0\pi_1}{\hat{\pi}_1\pi_0} = \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)}.$$

Thus, the UMP test T_* in (6.3) is equivalent to the test using the Bayes factor. Note that the rejection region given by (6.72) depends on prior probabilities whereas the Bayes factor does not.

When either Θ_0 or Θ_1 is not simple, however, Bayes factors also depend on the prior Π.

If Π is an improper prior, the Bayes test is still defined as long as the posterior probabilities $\hat{\pi}_j$ are finite. However, the Bayes factor is not well defined when Π is improper.

Example 6.25. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with an unknown $\mu \in \mathcal{R}$ and a known $\sigma^2 > 0$. Let the prior of μ be $N(\xi, \tau^2)$. Then the posterior of μ is $N(\mu_*(x), c^2)$, where

$$\mu_*(x) = \frac{\sigma^2}{n\tau^2 + \sigma^2}\xi + \frac{n\tau^2}{n\tau^2 + \sigma^2}\bar{x} \qquad \text{and} \qquad c^2 = \frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2}$$

(see Example 2.25). Consider first the problem of testing $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$. Let Φ be the c.d.f. of the standard normal. Then the posterior probability of Θ_0 and the Bayes factor are, respectively,

$$\hat{\pi}_0 = \Phi\left(\tfrac{\mu_0 - \mu_*(x)}{c}\right) \qquad \text{and} \qquad \beta = \frac{\Phi\left(\frac{\mu_0 - \mu_*(x)}{c}\right)\Phi\left(\frac{\mu_0 - \xi}{\tau}\right)}{\Phi\left(\frac{\mu_*(x) - \mu_0}{c}\right)\Phi\left(\frac{\xi - \mu_0}{\tau}\right)}.$$

It is interesting to see that if we let $\tau \to \infty$, which is the same as considering the improper prior $\Pi =$ the Lebesgue measure on $\mathcal{R}$, then

$$\hat{\pi}_0 \to \Phi\left(\tfrac{\mu_0 - \bar{x}}{\sigma/\sqrt{n}}\right),$$

which is exactly the p-value $\hat{\alpha}(x)$ derived in Example 2.29.

Consider next the problem of testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. In this case the prior c.d.f. can not be continuous at μ_0. We consider $\Pi(\mu) = \pi_0 I_{[\mu_0, \infty)}(\mu) + (1 - \pi_0)\Phi\left(\frac{\mu - \xi}{\tau}\right)$. Let $\ell(\mu)$ be the likelihood function based on $\bar{x}$. Then

$$m_1(x) = \int_{\mu \neq \mu_0} \ell(\mu) d\Phi\left(\tfrac{\mu - \xi}{\tau}\right) = \frac{1}{\sqrt{\tau^2 + \sigma^2/n}}\Phi'\left(\frac{\bar{x} - \xi}{\sqrt{\tau^2 + \sigma^2/n}}\right),$$

where $\Phi'(t)$ is p.d.f. of the standard normal distribution, and

$$\hat{\pi}_0 = \frac{\pi_0 \ell(\mu_0)}{\pi_0 \ell(\mu_0) + (1 - \pi_0) m_1(x)} = \left(1 + \frac{1 - \pi_0}{\pi_0 \beta}\right)^{-1},$$

where

$$\beta = \frac{\ell(\mu_0)}{m_1(x)} = \frac{\sqrt{n\tau^2 + \sigma^2}\Phi'\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right)}{\sigma\Phi'\left(\frac{\bar{x} - \xi}{\sqrt{\tau^2 + \sigma^2/n}}\right)}$$

is the Bayes factor. ∎

More discussions about Bayesian hypothesis tests can be found in Berger (1985, §4.3.3).

6.5 Tests in Nonparametric Models

In a nonparametric problem, a UMP, UMPU, or UMPI test usually does not exist. In this section we study some nonparametric tests that have size α, limiting size α, or asymptotic significance level α. Consistency (Definition 2.13) of these nonparametric tests is also discussed.

Nonparametric tests are derived using some intuitively appealing ideas. They are commonly referred to as *distribution-free* tests, since almost no assumption is imposed on the population under consideration. But a nonparametric test may not be as good as a parametric test (in terms of its power) when the parametric model is correct. This is very similar to the case where we consider parametric estimation methods versus nonparametric estimation methods.

6.5.1 Sign, permutation, and rank tests

Three popular classes of nonparametric tests are introduced here. The first one is the class of *sign tests*. Let $X_1, ..., X_n$ be i.i.d. random variables from F, u be a fixed constant, and $p = F(u)$. Consider the problem of testing $H_0 : p \leq p_0$ versus $H_1 : p > p_0$, or testing $H_0 : p = p_0$ versus $H_1 : p \neq p_0$, where p_0 is a fixed constant in $(0, 1)$. Let

$$\Delta_i = \begin{cases} 1 & X_i - u \leq 0 \\ 0 & X_i - u > 0, \end{cases} \qquad i = 1, ..., n.$$

Then $\Delta_1, ..., \Delta_n$ are i.i.d. binary random variables with $p = P(\Delta_i = 1)$. For testing $H_0 : p \leq p_0$ versus $H_1 : p > p_0$, it follows from Corollary 6.1 that the test

$$T_*(Y) = \begin{cases} 1 & Y > m \\ \gamma & Y = m \\ 0 & Y < m \end{cases} \tag{6.73}$$

is of size α and UMP among tests based on Δ_i's, where $Y = \sum_{i=1}^n \Delta_i$ and m and γ satisfy (6.7). Although T_* is of size α, we cannot conclude immediately that T_* is a UMP test, since $\Delta_1, ..., \Delta_n$ may not be sufficient for F. However, it can be shown that T_* is in fact a UMP test (Lehmann, 1986, pp. 106-107) in this particular case. Note that no assumption is imposed on F.

For testing $H_0 : p = p_0$ versus $H_1 : p \neq p_0$, it follows from Theorem 6.4 that the test

$$T_*(Y) = \begin{cases} 1 & Y < c_1 \text{ or } Y > c_2 \\ \gamma_i & Y = c_i, \ i = 1, 2, \\ 0 & c_1 < Y < c_2 \end{cases} \tag{6.74}$$

is of size α and UMP among unbiased tests based on Δ_i's, where γ and c_i's are chosen so that $E_{p=p_0}(T_*) = \alpha$ and $E_{p=p_0}(T_*Y) = \alpha n p_0$. This test is in fact a UMPU test (Lehmann, 1986, p. 166).

Since Y is equal to the number of nonnegative signs of $(u - X_i)$'s, tests based on T_* in (6.73) or (6.74) are called sign tests. One can easily extend the sign tests to the case where $p = P(X_1 \in B)$ with any fixed event B. Another extension is to the case where we observe i.i.d. $(X_1, Y_1), ..., (X_n, Y_n)$ (matched pairs). By using $\Delta_i = X_i - Y_i - u$, one can obtain sign tests for hypotheses concerning $P(X_1 - Y_1 \leq u)$.

Next, we introduce the class of *permutation tests*. Let $X_{i1}, ..., X_{in_i}$, $i = 1, 2$, be two independent samples i.i.d. from F_i, $i = 1, 2$, respectively, where F_i's are c.d.f.'s on $\mathcal{R}$. In §6.2.3, we showed that the two-sample t-tests are UMPU tests for testing hypotheses concerning the means of F_i's, under the assumption that F_i's are normal with the same variance. Such types of testing problems arise from the comparison of two treatments. Suppose now we remove the normality assumption and replace it by a much weaker assumption that F_i's are in the nonparametric family $\mathcal{F}$ containing all continuous c.d.f.'s on $\mathcal{R}$. Consider the problem of testing

$$H_0 : F_1 = F_2 \qquad \text{versus} \qquad H_1 : F_1 \neq F_2, \tag{6.75}$$

which is the same as testing the equality of the means of F_i's when F_i's are normal with the same variance.

Let $X = (X_{ij}, j = 1, ..., n_i, i = 1, 2)$, $n = n_1 + n_2$, and α be a given significance level. A test $T(X)$ satisfying

$$\frac{1}{n!} \sum_{z \in \pi(x)} T(z) = \alpha \tag{6.76}$$

is called a permutation test, where $\pi(x)$ is the set of $n!$ points obtained from $x \in \mathcal{R}^n$ by permuting the components of x. Permutation tests are of size α (exercise). Under the assumption that $F_1(x) = F_2(x - \theta)$ and $F_1 \in \mathcal{F}$ containing all c.d.f.'s having Lebesgue p.d.f.'s that are continuous a.e., which is still much weaker than the assumption that F_i's are normal with the same variance, the class of permutation tests of size α is exactly the same as the class of unbiased tests of size α; see, for example, Lehmann (1986, p. 231).

Unfortunately, a test UMP among all permutation tests of size α does not exist. In applications, we usually choose a Lebesgue p.d.f. h and define a permutation test

$$T(X) = \begin{cases} 1 & h(X) > h_m \\ \gamma & h(X) = h_m \\ 0 & h(X) < h_m, \end{cases} \tag{6.77}$$

where h_m is the $(m+1)$th largest value of the set $\{h(z) : z \in \pi(x)\}$, m is the integer part of $\alpha n!$, and $\gamma = \alpha n! - m$. This permutation test is optimal in some sense (Lehmann, 1986, §5.11).

While the class of permutation tests is motivated by the unbiasedness principle, the third class of tests introduced here is motivated by the invariance principle.

Consider first the one-sample problem in which $X_1, ..., X_n$ are i.i.d. random variables from a continuous c.d.f. F and we would like to test

$$H_0 : F \text{ is symmetric about } 0 \quad \text{versus} \quad H_1 : F \text{ is not symmetric about } 0.$$

Let $\mathcal{G}$ be the class of transformations $g(x) = (\psi(x_1), ..., \psi(x_n))$, where ψ is continuous, odd, and strictly increasing. Let $\tilde{R}(X)$ be the vector of ranks of $|X_i|$'s and $R_+(X)$ (or $R_-(X)$) be the subvector of $\tilde{R}(X)$ containing ranks corresponding to positive (or negative) X_i's. It can be shown (exercise) that (R_+, R_-) is maximal invariant under $\mathcal{G}$. Furthermore, sufficiency permits a reduction from R_+ and R_- to R_+^o, the vector of ordered components of R_+. A test based on R_+^o is called a (one-sample) *signed rank test*.

Similar to the case of permutation tests, there is no UMP test within the class of signed rank tests. A common choice is the signed rank test that rejects H_0 when $W(R_+^o)$ is too large or too small, where

$$W(R_+^o) = J(R_{+1}^o/n) + \cdots + J(R_{+n_*}^o/n), \tag{6.78}$$

J is a continuous and strictly increasing function on $[0, 1]$, R_{+i}^o is the ith component of R_+^o, and n_* is the number of positive X_i's. This is motivated by the fact that H_0 is unlikely to be true if W in (6.78) is too large or too small. Note that W/n is equal to $\mathtt{T}(F_n)$ with $\mathtt{T}$ given by (5.46) and $J(t) = t$, and the test based on W in (6.78) is the well-known one-sample Wilcoxon signed rank test.

Under H_0, $P(R_+^o = y) = 2^{-n}$ for each $y \in \mathcal{Y}$ containing 2^n n_*-tuples $y = (y_1, ..., y_{n_*})$ satisfying $1 \le y_1 < \cdots < y_{n_*} \le n$. Then, the following signed rank test is of size α:

$$T(X) = \begin{cases} 1 & W(R_+^o) < c_1 \text{ or } W(R_+^o) > c_2 \\ \gamma & W(R_+^o) = c_i, i = 1, 2 \\ 0 & c_1 < W(R_+^o) < c_2, \end{cases} \tag{6.79}$$

where c_1 and c_2 are the $(m+1)$th smallest and largest values of the set $\{W(y) : y \in \mathcal{Y}\}$, m is the integer part of $\alpha 2^n/2$, and $\gamma = \alpha 2^n/2 - m$.

Consider next the two-sample problem of testing (6.75) based on two independent samples, $X_{i1}, ..., X_{in_i}$, $i = 1, 2$, i.i.d. from F_i, $i = 1, 2$, respectively. Let $\mathcal{G}$ be the class of transformations $g(x) = (\psi(x_{ij}), j = 1, ..., n_i, i =$

$1, 2)$, where ψ is continuous and strictly increasing. Let $R(X)$ be the vector of ranks of all X_{ij}'s. In Example 6.14, we showed that R is maximal invariant under $\mathcal{G}$. Again, sufficiency permits a reduction from R to R_1^o, the vector of ordered values of the ranks of $X_{11}, ..., X_{1n_1}$. A test for (6.75) based on R_1^o is called a two-sample *rank test*. Under H_0, $P(R_1^o = y) = \binom{n}{n_1}^{-1}$ for each $y \in \mathcal{Y}$ containing $\binom{n}{n_1}$ n_1-tuples $y = (y_1, ..., y_{n_1})$ satisfying $1 \le y_1 < \cdots < y_{n_1} \le n$. Let $R_1^o = (R_{11}^o, ..., R_{1n_1}^o)$. Then a commonly used two-sample rank test is given by (6.78)-(6.79) with R_{+i}^o, n_*, and 2^n replaced by R_{1i}^o, n_1, and $\binom{n}{n_1}$, respectively. When $n_1 = n_2$, the statistic W/n is equal to $\mathtt{T}(F_n)$ with $\mathtt{T}$ given by (5.48). When $J(t) = t - \frac{1}{2}$, this reduces to the well-known two-sample Wilcoxon rank test.

A common feature of the permutation and rank tests previously introduced is that tests of size α can be obtained for each fixed sample size n, but the computation involved in determining the rejection regions $\{T(X) = 1\}$ may be cumbersome if n is large. Thus, one may consider approximations to permutation and rank tests when n is large. Permutation tests can often be approximated by the two-sample t-tests derived in §6.2.3 (Lehmann, 1986, §5.13). Using the results in §5.2.2, we now derive one-sample signed rank tests having limiting size α (Definition 2.13(ii)), which can be viewed as signed rank tests of size approximately α when n is large.

From the discussion in §5.2.2, $W/n = \mathtt{T}(F_n)$ with a ϱ_∞-Hadamard differentiable functional $\mathtt{T}$ given by (5.46) and, by Theorem 5.5,

$$\sqrt{n}[W/n - \mathtt{T}(F)] \to_d N(0, \sigma_F^2),$$

where $\sigma_F^2 = E[\phi_F(X_1)]^2$ and

$$\phi_F(x) = \int_0^\infty J'(\tilde{F}(y))(\tilde{\delta}_x - \tilde{F})(y)dF(y) + J(\tilde{F}(x)) - \mathtt{T}(F)$$

(see (5.47)). Since F is continuous, $\tilde{F}(x) = F(x) - F(-x)$. Under H_0, $F(x) = 1 - F(-x)$. Hence, σ_F^2 under H_0 is equal to $v_1 + v_2 + 2v_{12}$, where

$$v_1 = \mathrm{Var}\big(J(\tilde{F}(X_1))\big) = \frac{1}{2}\int_0^\infty [J(\tilde{F}(x))]^2 d\tilde{F}(x),$$

$$\begin{aligned}
v_2 &= \mathrm{Var}\left(\int_0^\infty J'(\tilde{F}(y))(\tilde{\delta}_{X_1} - \tilde{F})(y)dF(y)\right) \\
&= E\int_0^\infty\int_0^\infty J'(\tilde{F}(y))J'(\tilde{F}(z))(\tilde{\delta}_{X_1} - \tilde{F})(y)(\tilde{\delta}_{X_1} - \tilde{F})(z)dF(y)dF(z) \\
&= \frac{1}{4}\int_0^\infty\int_0^\infty J'(\tilde{F}(y))J'(\tilde{F}(z))[\tilde{F}(\min(y,z)) - \tilde{F}(y)\tilde{F}(z)]d\tilde{F}(y)d\tilde{F}(z) \\
&= \frac{1}{2}\int_{0<z<y<\infty} J'(\tilde{F}(y))J'(\tilde{F}(z))\tilde{F}(z)[1 - \tilde{F}(y)]d\tilde{F}(y)d\tilde{F}(z),
\end{aligned}$$

and

$$\begin{aligned} v_{12} &= \mathrm{Cov}\left(J(\tilde{F}(X_1)), \int_0^\infty J'(\tilde{F}(y))(\tilde{\delta}_{X_1} - \tilde{F})(y)dF(y)\right) \\ &= E\int_0^\infty J(\tilde{F}(X_1))J'(\tilde{F}(y))(\tilde{\delta}_{X_1} - \tilde{F})(y)dF(y) \\ &= \int_{-\infty}^\infty \int_0^\infty J(\tilde{F}(x))J'(\tilde{F}(y))(\delta_{|x|} - \tilde{F})(y)dF(y)dF(x) \\ &= \frac{1}{2}\int_0^\infty \int_0^\infty J(\tilde{F}(x))J'(\tilde{F}(y))(\delta_x - \tilde{F})(y)d\tilde{F}(y)d\tilde{F}(x). \end{aligned}$$

Note that under H_0, the distribution of W is completely known. Indeed, letting $s = \tilde{F}(y)$ and $t = \tilde{F}(z)$, we conclude that $\sigma_F^2 = v_1 + v_2 + 2v_{12}$ and

$$\mathtt{T}(F) = \int_0^\infty J(\tilde{F}(x))dF(x) = \frac{1}{2}\int_0^1 J(s)ds$$

do not depend on F. Hence, a signed rank test T that rejects H_0 when

$$\sqrt{n}|W/n - t_0| > \sigma_0 z_{1-\alpha/2}, \tag{6.80}$$

where $z_a = \Phi^{-1}(a)$ and $t_0 = \mathtt{T}(F)$ and $\sigma_0^2 = \sigma_F^2$ under H_0 are known constants, has the property that

$$\begin{aligned} \sup_{P\in\mathcal{P}_0} \beta_T(P) &= \sup_{P\in\mathcal{P}_0} P\left(\sqrt{n}|W/n - t_0| > \sigma_0 z_{1-\alpha/2}\right) \\ &= P_W\left(\sqrt{n}|W/n - t_0| > \sigma_0 z_{1-\alpha/2}\right) \\ &\to \alpha, \end{aligned}$$

i.e., T has limiting size α.

Two-sample rank tests having limiting size α can be similarly derived (exercise).

6.5.2 Kolmogorov-Smirnov and Cramér-von Mises tests

In this section we introduce two types of tests for hypotheses concerning continuous c.d.f.'s on $\mathcal{R}$. Let $X_1, ..., X_n$ be i.i.d. random variables from a continuous c.d.f. F. Suppose that we would like to test hypotheses (6.68), i.e., $H_0 : F = F_0$ versus $H_1 : F \neq F_0$ with a fixed F_0. Let F_n be the empirical c.d.f. and

$$D_n(F) = \sup_{x\in\mathcal{R}} |F_n(x) - F(x)|, \tag{6.81}$$

which is in fact the distance $\varrho_\infty(F_n, F)$. Intuitively, $D_n(F_0)$ should be small if H_0 is true. From the results in §5.1.1 we know that $D_n(F_0) \to_{a.s.} 0$ if and

only if H_0 is true. The statistic $D_n(F_0)$ is called the Kolmogorov-Smirnov statistic. Tests with rejection region $D_n(F_0) > c$ are called Kolmogorov-Smirnov tests.

In some cases we would like to test "one-sided" hypotheses $H_0 : F = F_0$ versus $H_1 : F \geq F_0,\ F \neq F_0$, or $H_0 : F = F_0$ versus $H_1 : F \leq F_0,\ F \neq F_0$. The corresponding Kolmogorov-Smirnov statistic is $D_n^+(F_0)$ or $D_n^-(F_0)$, where

$$D_n^+(F) = \sup_{x \in \mathcal{R}} [F_n(x) - F(x)] \tag{6.82}$$

and

$$D_n^-(F) = \sup_{x \in \mathcal{R}} [F(x) - F_n(x)].$$

The rejection regions of one-sided Kolmogorov-Smirnov tests are, respectively, $D_n^+(F_0) > c$ and $D_n^-(F_0) > c$.

Let $X_{(1)} < \cdots < X_{(n)}$ be the order statistics and define $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$. Since $F_n(x) = i/n$ when $X_{(i)} \leq x < X_{(i+1)}$, $i = 0, 1, ..., n$,

$$\begin{aligned} D_n^+(F) &= \max_{0 \leq i \leq n} \sup_{X_{(i)} \leq x < X_{(i+1)}} \left[\frac{i}{n} - F(x) \right] \\ &= \max_{0 \leq i \leq n} \left[\frac{i}{n} - \inf_{X_{(i)} \leq x < X_{(i+1)}} F(x) \right] \\ &= \max_{0 \leq i \leq n} \left[\frac{i}{n} - F(X_{(i)}) \right]. \end{aligned}$$

When F is continuous, $F(X_{(i)})$ is the ith order statistic of a sample of size n from the uniform distribution $U(0, 1)$ irrespective of what F is. Therefore, the distribution of $D_n^+(F)$ does not depend on F, if we restrict our attention to continuous c.d.f.'s on $\mathcal{R}$. The distribution of $D_n^-(F)$ is the same as that of $D_n^+(F)$ because of symmetry. Since

$$D_n(F) = \max[D_n^+(F), D_n^-(F)],$$

the distribution of $D_n(F)$ does not depend on F. This means that the distributions of Kolmogorov-Smirnov statistics are known under H_0.

Theorem 6.10. Let $D_n(F)$ and $D_n^+(F)$ be defined by (6.81) and (6.82).
(i) For any fixed n,

$$P\big(D_n^+(F) \leq t\big) = \begin{cases} 0 & t \leq 0 \\ n! \displaystyle\prod_{i=1}^{n} \int_{\frac{i}{n}-t}^{u_{i+1}} du & 0 < t < 1 \\ 1 & t \geq 1 \end{cases}$$

and

$$P\big(D_n(F) \le t\big) = \begin{cases} 0 & t \le (2n)^{-1} \\ n!\prod_{i=1}^{n} \int_{2i-t}^{2i-n^{-1}+t} du & (2n)^{-1} < t < 1 \\ 1 & t \ge 1, \end{cases}$$

where $du = du_1 \cdots du_n$ $(u_1 < \cdots < u_n)$ and $u_{n+1} = 1$.
(ii) For $t > 0$,

$$\lim_{n\to\infty} P\big(\sqrt{n}D_n^+(F) \le t\big) = 1 - e^{-2t^2}$$

and

$$\lim_{n\to\infty} P\big(\sqrt{n}D_n(F) \le t\big) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2t^2}. \quad \blacksquare$$

The proof of Theorem 6.10(i) is left as an exercise. The proof of Theorem 6.10(ii) can be found in Kolmogorov (1933) and Smirnov (1944).

When n is not large, Kolmogorov-Smirnov tests of size α can be obtained using the results in Theorem 6.10(i). When n is large, using the results in Theorem 6.10(i) is not convenient. We can obtain Kolmogorov-Smirnov tests of limiting size α using the results in Theorem 6.10(ii).

Another test for $H_0 : F = F_0$ versus $H_1 : F \ne F_0$ is the Cramér-von Mises test which rejects H_0 when $C_n(F_0) > c$, where

$$C_n(F) = \int [F_n(x) - F(x)]^2 dF(x) \tag{6.83}$$

is another measure of disparity between F_n and F. Similar to $D_n(F)$, the distribution of $C_n(F)$ does not depend on F (exercise). Hence, a Cramér-von Mises test of size α can be obtained. When n is large, it is more convenient to use a Cramér-von Mises test of limiting size α. Note that $C_n(F_0)$ is actually a V-statistic (§3.5.3) with kernel

$$h(x_1, x_2) = \int [\delta_{x_1}(y) - F_0(y)][\delta_{x_2}(y) - F_0(y)]dF_0(y)$$

and

$$h_1(x_1) = E[h(x_1, X_2)] = \int [\delta_{x_1}(y) - F_0(y)][F(y) - F_0(y)]dF_0(y).$$

It follows from Theorem 3.16 that if H_1 is true, $C_n(F_0)$ is asymptotically normal whereas if H_0 is true, $h_1(x_1) \equiv 0$ and

$$nC_n(F_0) \to_d \sum_{j=1}^{\infty} \lambda_j \chi_{1j}^2,$$

where χ^2_{1j}'s are i.i.d. from the chi-square distribution χ^2_1 and λ_j's are constants. In this case, Durbin (1973) showed that $\lambda_j = j^{-2}\pi^{-2}$.

For testing (6.68), it is worthwhile to compare the goodness of fit test introduced in Example 6.23 with the Kolmogorov-Smirnov test (or Cramér-von Mises test). The former requires a partition of the range of observations and may lose information through partitioning, whereas the latter requires that F be continuous and univariate; the latter is of size α (or limiting size α), whereas the former is only of asymptotic significance level α; the former can be modified to allow estimation of unknown parameters under H_0 (i.e., hypotheses (6.69)), whereas the latter does not have this flexibility. Note that goodness of fit tests are nonparametric in nature, although χ^2-tests are derived from a parametric model.

Kolmogorov-Smirnov tests can be extended to two-sample problems to test hypotheses in (6.75). Let $X_{i1}, ..., X_{in_i}$, $i = 1, 2$, be two independent samples i.i.d. from F_i on $\mathcal{R}$, $i = 1, 2$, and let F_{in_i} be the empirical c.d.f. based on $X_{i1}, ..., X_{in_i}$. A Kolmogorov-Smirnov test rejects H_0 when $D_{n_1,n_2} > c$, where

$$D_{n_1,n_2} = \sup_{x\in\mathcal{R}} |F_{1n_1}(x) - F_{2n_2}(x)|.$$

A Kolmogorov-Smirnov test of limiting size α can be obtained using

$$\lim_{n_1,n_2\to\infty} P\big(\sqrt{n_1n_2/(n_1+n_2)}D_{n_1,n_2} \le t\big) = \sum_{j=-\infty}^{\infty} (-1)^{j-1}e^{-2j^2t^2}, \quad t > 0.$$

6.5.3 Empirical likelihood ratio tests

The method of likelihood ratio is useful in deriving tests under parametric models. In nonparametric problems, we now introduce a similar method based on the empirical likelihoods introduced in §5.1.2.

Suppose that a sample X is from a population determined by a c.d.f. $F \in \mathcal{F}$, where $\mathcal{F}$ is a class of c.d.f.'s on $\mathcal{R}^d$. Consider the problem of testing

$$H_0 : \mathtt{T}(F) = t_0 \qquad \text{versus} \qquad H_1 : \mathtt{T}(F) \neq t_0, \tag{6.84}$$

where $\mathtt{T}$ is a functional from $\mathcal{F}$ to $\mathcal{R}^k$ and t_0 is a fixed vector in $\mathcal{R}^k$. Let $\ell(G)$, $G \in \mathcal{F}$, be a given empirical likelihood, $\hat{F}$ be an MELE of F, and $\hat{F}_{H_0}$ be an MELE of F under H_0, i.e., $\hat{F}_{H_0}$ is an MELE of F subject to $\mathtt{T}(F) = t_0$. Then the empirical likelihood ratio is defined as

$$\lambda_n(X) = \ell(\hat{F}_{H_0})/\ell(\hat{F}).$$

A test with rejection region $\lambda_n(X) < c$ is called an *empirical likelihood ratio* test.

As a specific example, consider the following empirical likelihood (or nonparametric likelihood) when $X = (X_1, ..., X_n)$ with i.i.d. X_i's:

$$\ell(G) = \prod_{i=1}^{n} p_i \qquad \text{subject to} \quad p_i \geq 0, \ \sum_{i=1}^{n} p_i = 1,$$

where $p_i = P_G(\{x_i\})$, $i = 1, ..., n$. Suppose that $\mathtt{T}(G) = \int u(x)dG(x)$ with a known function $u(x)$ from $\mathcal{R}^d$ to $\mathcal{R}^r$. Then $\hat{F} = F_n$; H_0 in (6.84) with $t_0 = 0$ is the same as that assumption (5.9) holds; $\hat{F}_{H_0}$ is the MELE given by (5.11); and the empirical likelihood ratio is

$$\lambda_n(X) = n^n \prod_{i=1}^{n} \hat{p}_i, \tag{6.85}$$

where $\hat{p}_i$ is given by (5.12). An empirical likelihood ratio test with asymptotic significance level α can be obtained using the following result.

Theorem 6.11. Assume the conditions in Theorem 5.4. Under the hypothesis H_0 in (6.84) with $t_0 = 0$ (i.e., (5.9) holds),

$$-2 \log \lambda_n(X) \to_d \chi^2_r,$$

where $\lambda_n(X)$ is given by (6.85) and χ^2_r has the chi-square distribution χ^2_r. ∎

The proof of this result can be found in Owen (1988, 1990). In fact, the result in Theorem 6.11 holds for some other functionals $\mathtt{T}$ such as the median functional.

We can also derive tests based on the profile empirical likelihoods discussed in §5.4.1. Consider an empirical likelihood

$$\ell(G) = \prod_{i=1}^{n} p_i \quad \text{subject to} \quad p_i \geq 0, \ \sum_{i=1}^{n} p_i = 1, \ \sum_{i=1}^{n} p_i \psi(x_i, \theta) = 0,$$

where θ is a k-vector of unknown parameters and ψ is a known function. Let $\theta = (\vartheta, \varphi)$, where ϑ is $r \times 1$ and φ is $(k-r) \times 1$. Suppose that we would like to test

$$H_0 : \vartheta = \vartheta_0 \qquad \text{versus} \qquad H_1 : \vartheta \neq \vartheta_0,$$

where ϑ_0 is a fixed r-vector. Let $\hat{\theta}$ be a maximum of the profile empirical likelihood $\ell_P(\theta)$ given by (5.86) and let $\hat{\varphi}$ be a maximum of $\ell_P(\varphi) = \ell_P(\vartheta_0, \varphi)$. Then a profile empirical likelihood ratio test rejects H_0 when $\lambda_n(X) < c$, where

$$\lambda_n(X) = \prod_{i=1}^{n} \frac{1 + \psi(x_i, \hat{\theta})[\xi_n(\hat{\theta})]^\tau}{1 + \psi(x_i, \vartheta_0, \hat{\varphi})[\zeta_n(\vartheta_0, \hat{\varphi})]^\tau}, \tag{6.86}$$

$\xi_n(\hat{\theta})$ satisfies

$$\sum_{i=1}^{n} \frac{\psi(x_i, \hat{\theta})}{1 + \psi(x_i, \hat{\theta})[\xi_n(\hat{\theta})]^\tau} = 0$$

and $\zeta_n(\vartheta_0, \hat{\varphi})$ satisfies

$$\sum_{i=1}^{n} \frac{\psi(x_i, \vartheta_0, \hat{\varphi})}{1 + \psi(x_i, \vartheta_0, \hat{\varphi})[\zeta_n(\vartheta_0, \hat{\varphi})]^\tau} = 0.$$

From the discussion in §5.4.1, $\hat{\theta}$ can be approximated by a solution of the GEE $\sum_{i=1}^n \psi(X_i, \theta) = 0$ and $\hat{\varphi}$ can be approximated by a solution of the reduced GEE $\sum_{i=1}^n \psi(X_i, \vartheta_0, \varphi) = 0$. Under some regularity conditions (e.g., the conditions in Proposition 5.3), Qin and Lawless (1994) showed that the result in Theorem 6.11 holds with $\lambda_n(X)$ given by (6.86). Thus, a profile empirical likelihood ratio test with asymptotic significance level α can be obtained.

Example 6.26. Let $Y_1, ..., Y_n$ be i.i.d. random 2-vectors from F. Consider the problem of testing $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$, where $(\mu_1, \mu_2) = E(Y_1)$. Let $Y_i = (Y_{i1}, Y_{i2})$, $X_{i1} = Y_{i1} - Y_{i2}$, $X_{i2} = Y_{i1} + Y_{i2}$, and $X_i = (X_{i1}, X_{i2})$, $i = 1, ..., n$. Then $X_1, ..., X_n$ are i.i.d. with $E(X_1) = \theta = (\vartheta, \varphi)$, where $\vartheta = \mu_1 - \mu_2$ and $\varphi = \mu_1 + \mu_2$. The hypotheses of interest becomes $H_0 : \vartheta = 0$ versus $H_1 : \vartheta \neq 0$.

To apply the profile empirical likelihood method, we define $\psi(x, \theta) = x - \theta$, $x \in \mathcal{R}^2$. Note that a solution of the GEE $\sum_{i=1}^n (X_i - \theta) = 0$ is $\hat{\theta} = (\bar{X}_{1\cdot}, \bar{X}_{2\cdot}) = \bar{X}$, and a solution of the reduced GEE $\sum_{i=1}^n (X_{2i} - \varphi) = 0$ is $\hat{\varphi} = \bar{X}_{2\cdot}$. The profile empirical likelihood ratio is then given by

$$\lambda_n(X) = \prod_{i=1}^{n} \frac{1 + (X_i - \bar{X})[\xi_n(\bar{X})]^\tau}{1 + [X_i - (0, \bar{X}_{2\cdot})][\zeta_n(0, \bar{X}_{2\cdot})]^\tau},$$

where $\xi_n(\bar{X})$ satisfies

$$\sum_{i=1}^{n} \frac{X_i - \bar{X}}{1 + (X_i - \bar{X})[\xi_n(\bar{X})]^\tau} = 0$$

and $\zeta_n(0, \bar{X}_{2\cdot})$ satisfies

$$\sum_{i=1}^{n} \frac{X_i - (0, \bar{X}_{2\cdot})}{1 + [X_i - (0, \bar{X}_{2\cdot})][\zeta_n(0, \bar{X}_{2\cdot})]^\tau} = 0. \quad \blacksquare$$

Empirical likelihood ratio tests or profile empirical likelihood ratio tests in various other problems can be found, for example, in Owen (1988, 1990, 1991), Chen and Qin (1993), Qin (1993), and Qin and Lawless (1994).

6.5.4 Asymptotic tests

We now introduce a simple method of constructing asymptotic tests (i.e., tests with asymptotic significance level α). This method works for almost all problems (parametric or nonparametric) in which the hypotheses being tested are $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, where θ is a vector of parameters, and an asymptotically normally distributed estimator of θ can be found. However, this simple method may not provide the best or even nearly best solution to the problem, especially when there are different asymptotically normally distributed estimators of θ.

Let X be a sample of size n from a population P and $\hat{\theta}_n$ be an estimator of θ, a k-vector of parameters related to P. Suppose that under H_0,

$$(\hat{\theta}_n - \theta)V_n^{-1/2} \to_d N_k(0, I_k), \tag{6.87}$$

where V_n is the asymptotic covariance matrix of $\hat{\theta}_n$. If V_n is known when $\theta = \theta_0$, then a test with rejection region

$$(\hat{\theta}_n - \theta_0)V_n^{-1}(\hat{\theta}_n - \theta_0)^\tau > \chi^2_{k,\alpha} \tag{6.88}$$

has asymptotic significance level α, where $\chi^2_{k,\alpha}$ is the $(1-\alpha)$th quantile of the chi-squared distribution χ^2_k. If the distribution of $\hat{\theta}_n$ does not depend on the unknown population P under H_0 and (6.87) holds, then a test with rejection region (6.88) has limiting size α.

If V_n in (6.88) depends on the unknown population P even if H_0 is true ($\theta = \theta_0$), then we have to replace V_n in (6.88) by an estimator $\hat{V}_n$. If, under H_0, $\hat{V}_n$ is consistent according to Definition 5.4, then the test having rejection region (6.88) with V_n replaced by $\hat{V}_n$ has asymptotic significance level α. Variance estimation methods introduced in §5.5 can be used to construct a consistent estimator $\hat{V}_n$.

In some cases result (6.87) holds for any P. Then, the following result shows that the test having rejection region (6.88) is asymptotically correct (§2.5.3), i.e., it is a consistent asymptotic test (Definition 2.13).

Theorem 6.12. Assume that (6.87) holds for any P and that $\lambda_+[V_n] \to 0$, where $\lambda_+[V_n]$ is the largest eigenvalue of V_n.
(i) The test having rejection region (6.88) (with a known V_n or V_n replaced by an estimator $\hat{V}_n$ that is consistent for any P) is consistent.
(ii) If we choose $\alpha = \alpha_n \to 0$ as $n \to \infty$ and $\chi^2_{k,1-\alpha_n}\lambda_+[V_n] = o(1)$, then the test in (i) is Chernoff-consistent.
Proof. The proof of (ii) is left as an exercise. We only prove (i) for the case where V_n is known. Let $Z_n = (\hat{\theta}_n - \theta)V_n^{-1/2}$ and $l_n = (\theta - \theta_0)V_n^{-1/2}$. Then $\|Z_n\| = O_p(1)$ and $\|l_n\| = \|(\theta - \theta_0)V_n^{-1/2}\| \to \infty$ when $\theta \neq \theta_0$. The

result follows from the fact that when $\theta \neq \theta_0$,

$$\begin{aligned}(\hat{\theta}_n - \theta_0)V_n^{-1}(\hat{\theta}_n - \theta_0)^\tau &= \|Z_n\|^2 + \|l_n\|^2 + 2l_n Z_n^\tau \\ &\geq \|Z_n\|^2 + \|l_n\|^2 - 2\|l_n\|\|Z_n\| \\ &= O_p(1) + \|l_n\|^2[1 - o_p(1)]\end{aligned}$$

and, therefore,

$$P\left((\hat{\theta}_n - \theta_0)V_n^{-1}(\hat{\theta}_n - \theta_0)^\tau > \chi^2_{k,\alpha}\right) \to 1. \quad \blacksquare$$

Example 6.27. Let $X_1, ..., X_n$ be i.i.d. random variables from a symmetric c.d.f. F having finite variance and positive F'. Consider the problem of testing $H_0 : F$ is symmetric about 0 versus $H_1 : F$ is not symmetric about 0. Under H_0, there are many estimators satisfying (6.87). We consider the following five estimators:
(1) $\hat{\theta}_n = \bar{X}$ and $\theta = E(X_1)$;
(2) $\hat{\theta}_n = \hat{\theta}_{0.5}$ (the sample median) and $\theta = F^{-1}(\frac{1}{2})$ (the median of F);
(3) $\hat{\theta}_n = \bar{X}_a$ (the a-trimmed sample mean defined by (5.70)) and $\theta = \mathtt{T}(F)$, where $\mathtt{T}$ is given by (5.39) with $J(t) = (1-2a)^{-1}I_{(a,1-a)}(t)$, $a \in (0, \frac{1}{2})$;
(4) $\hat{\theta}_n =$ the Hodges-Lehmann estimator (Example 5.8) and $\theta = F^{-1}(\frac{1}{2})$;
(5) $\hat{\theta}_n = W/n - \frac{1}{2}$, where W is given by (6.78) with $J(t) = t$, and $\theta = \mathtt{T}(F) - \frac{1}{2}$ with $\mathtt{T}$ given by (5.46).

Although θ in (1)-(5) are different in general, in all cases $\theta = 0$ is equivalent to that H_0 holds.

For $\bar{X}$, it follows from the CLT that (6.87) holds with $V_n = \sigma^2/n$ for any F, where $\sigma^2 = \mathrm{Var}(X_1)$. From the SLLN, S^2/n is a consistent estimator of V_n for any F. Thus, the test having rejection region (6.88) with $\hat{\theta}_n = \bar{X}$ and V_n replaced by S^2/n is asymptotically correct. This test is asymptotically equivalent to the one-sample t-test derived in §6.2.3.

From Theorem 5.10, $\hat{\theta}_{0.5}$ satisfies (6.87) with $V_n = 4^{-1}[F'(\theta)]^{-2}n^{-1}$ for any F. A consistent estimator can be obtained using the bootstrap method considered in §5.5.3. Another consistent estimator of V_n can be obtained using the Woodruff's interval introduced in §7.4 (see Exercise 70 in §7.6). The test having rejection region (6.88) with $\hat{\theta}_n = \hat{\theta}_{0.5}$ and V_n replaced by a consistent estimator is asymptotically correct.

It follows from the discussion in §5.3.2 that $\bar{X}_a$ satisfies (6.87) for any F. A consistent estimator of V_n can be obtained using formula (5.105) or the jackknife method in §5.5.2. The test having rejection region (6.88) with $\hat{\theta}_n = \bar{X}_a$ and V_n replaced by a consistent estimator is asymptotically correct.

From Example 5.8, the Hodges-Lehmann estimator satisfies (6.87) for any F and $V_n = 12^{-1}\gamma^{-2}n^{-1}$ under H_0, where $\gamma = \int F'(x)dF(x)$. A

consistent estimator of V_n under H_0 can be obtained using the result in Exercise 86 in §5.6. The test having rejection region (6.88) with $\hat{\theta}_n =$ the Hodges-Lehmann estimator and V_n replaced by a consistent estimator is asymptotically correct.

Note that all tests discussed so far are not of limiting size α, since the distributions of $\hat{\theta}_n$ are still unknown under H_0.

The test having rejection region (6.88) with $\hat{\theta}_n = W/n - \frac{1}{2}$ and $V_n = (12n)^{-1}$ is equivalent to the one-sample Wilcoxon signed rank test and is shown to have limiting size α (§6.5.1). Also, (6.87) is satisfied for any F (§5.2.2). Although Theorem 6.12 is not applicable, a modified proof of Theorem 6.12 can be used to show the consistency of this test (exercise).

It is not clear which one of the five tests discussed here is to be preferred in general.

The results for $\hat{\theta}_n$ in (1)-(3) and (5) still hold for testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ without the assumption that F is symmetric. ∎

An example of asymptotic tests for one-sided hypotheses is given in Exercise 107. Most tests in §6.1-§6.4 derived under parametric models are asymptotically correct even when the parametric model assumptions are removed. Some examples are given in Exercises 105-107.

Finally, a study of asymptotic efficiencies of various tests can be found, for example, in Serfling (1980, Chapter 10).

6.6 Exercises

1. Prove Theorem 6.1 for the case of $\alpha = 0$ or 1.

2. Assume the conditions in Theorem 6.1. Let $\beta(P)$ be the power function of a UMP test of size $\alpha \in (0, 1)$. Show that $\alpha < \beta(P_1)$ unless $P_0 = P_1$.

3. Let T_* be given by (6.3) with $c = c(\alpha)$ for an $\alpha > 0$.
(a) Show that if $\alpha_1 < \alpha_2$, then $c(\alpha_1) \geq c(\alpha_2)$.
(b) Show that if $\alpha_1 < \alpha_2$, then the type II error probability of T_* of size α_1 is larger than that of T_* of size α_2.

4. Let H_0 and H_1 be simple and let $\alpha \in (0, 1)$. Suppose that T_* is a UMP test of size α for testing H_0 versus H_1 and that $\beta < 1$, where β is the power of T_* when H_1 is true. Show that $1 - T_*$ is a UMP test of size $1 - \beta$ for testing H_1 versus H_0.

5. Let X be a sample of size 1 from a Lebesgue p.d.f. f_θ. Find a UMP test of size $\alpha \in (0, \frac{1}{2})$ for $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ in the

following cases:
(a) $f_\theta(x) = 2\theta^{-2}(\theta - x)I_{(0,\theta)}(x)$, $\theta_0 < \theta_1$;
(b) $f_\theta(x) = 2[\theta x + (1-\theta)(1-x)]I_{(0,1)}(x)$, $0 \le \theta_1 < \theta_0 \le 1$;
(c) f_{θ_0} is the p.d.f. of $N(0,1)$ and f_{θ_1} is the p.d.f. of the Cauchy distribution $C(0,1)$;
(d) $f_{\theta_0}(x) = 4xI_{(0,\frac{1}{2})}(x) + 4(1-x)I_{(\frac{1}{2},1)}(x)$ and $f_{\theta_1}(x) = I_{(0,1)}(x)$;
(e) f_θ is the p.d.f. of the Cauchy distribution $C(\theta,1)$, $\theta_0 = 0$, and $\theta_1 = 1$.

6. Let $X_1, ..., X_n$ be i.i.d. from a Lebesgue p.d.f. f_θ. Find a UMP test of size α for $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ in the following cases:
(a) $f_\theta(x) = e^{-(x-\theta)}I_{(\theta,\infty)}(x)$, $\theta_0 < \theta_1$;
(b) $f_\theta(x) = \theta x^{-2}I_{(\theta,\infty)}(x)$, $\theta_0 \neq \theta_1$.

7. Prove Lemma 6.1.

8. Let f_0 and f_1 be Lebesgue integrable functions on $\mathcal{R}$ and

$$\phi_*(x) = \begin{cases} 1 & f_0(x) < 0 \text{ or } f_0(x) = 0,\ f_1(x) \ge 0 \\ 0 & f_0(x) > 0 \text{ or } f_0(x) = 0,\ f_1(x) < 0. \end{cases}$$

Show that ϕ_* maximizes $\int \phi(x)f_1(x)dx$ over all functions ϕ from $\mathcal{R}$ to $[0,1]$ such that $\int \phi(x)f_0(x)dx = \int \phi_*(x)f_0(x)dx$.

9. Prove Proposition 6.1.

10. Prove the claims in Example 6.5.

11. Show that the family $\{f_\theta : \theta \in \mathcal{R}\}$ with $f_\theta(x) = c(\theta)h(x)I_{(a(\theta),b(\theta))}(x)$ has monotone likelihood ratio, where $h(x)$ is a positive Lebesgue integrable function, and $a(\theta)$ and $b(\theta)$ are nondecreasing functions of θ.

12. Let $X_1, ..., X_n$ be i.i.d. from a p.d.f. f_θ, $\theta \in \Theta \subset \mathcal{R}$. Find a UMP test of size α for testing $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$ in the following cases:
(a) $f_\theta(x) = \theta^{-1}e^{-x/\theta}I_{(0,\infty)}(x)$, $\theta > 0$;
(b) $f_\theta(x) = \theta^{-1}x^{\theta-1}I_{(0,1)}(x)$, $\theta > 0$;
(c) $f_\theta(x)$ is the p.d.f. of $N(1,\theta)$;
(d) $f_\theta(x) = \theta^{-c}cx^{c-1}e^{-(x/\theta)^c}I_{(0,\infty)}(x)$, $\theta > 0$, where $c > 0$ is known.

13. Suppose that the distribution of X is in a family with monotone likelihood ratio in $Y(X)$, where $Y(X)$ has a continuous distribution. Consider the hypotheses $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$. Show that the p-value (§2.4.2) of the UMP test is given by $P_{\theta_0}(Y \ge y)$, where y is the observed value of Y.

14. Let $X_1, ..., X_m$ be i.i.d. from $N(\mu_x, \sigma_x^2)$ and $Y_1, ..., Y_n$ be i.i.d. from $N(\mu_y, \sigma_y^2)$. Suppose that X_i's and Y_j's are independent.
(a) When $\sigma_x = \sigma_y = 1$, find a UMP test of size α for testing $H_0 : \mu_x \leq \mu_y$ versus $H_1 : \mu_x > \mu_y$. (Hint: see Lehmann (1986, §3.9).)
(b) When μ_x and μ_y are known, find a UMP test of size α for testing $H_0 : \sigma_x \leq \sigma_y$ versus $H_1 : \sigma_x > \sigma_y$. (Hint: see Lehmann (1986, §3.9).)

15. Let f and g be two known p.d.f.'s w.r.t. ν. Suppose that X has the p.d.f. $\theta f(x) + (1-\theta)g(x)$, $\theta \in \mathcal{R}$. Show that the test $T_*(X) \equiv \alpha$ is a UMP test of size α for testing $H_0 : \theta \leq \theta_1$ or $\theta \geq \theta_2$ versus $H_1 : \theta_1 < \theta < \theta_2$.

16. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution $U(\theta, \theta+1)$, $\theta \in \mathcal{R}$. Suppose that $n \geq 2$.
(a) Find the joint distribution of $X_{(1)}$ and $X_{(n)}$.
(b) Show that a UMP test of size α for testing $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$ is of the form

$$T_*(X_{(1)}, X_{(n)}) = \begin{cases} 0 & X_{(1)} < 1 - \alpha^{1/n},\ X_{(n)} < 1 \\ 1 & \text{otherwise.} \end{cases}$$

(c) Does the family of all possible distributions of $(X_{(1)}, X_{(n)})$ have monotone likelihood ratio? (Hint: see Lehmann (1986, p. 115).)

17. Suppose that $X_1, ..., X_n$ are i.i.d. from the discrete uniform distribution $DU(1, ..., \theta)$ (Table 1.1, page 18) with an unknown $\theta = 1, 2, ...$
(a) Consider $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Show that

$$T_*(X) = \begin{cases} 1 & X_{(n)} > \theta_0 \\ \alpha & X_{(n)} \leq \theta_0 \end{cases}$$

is a UMP test of size α.
(b) Consider $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Show that

$$T_*(X) = \begin{cases} 1 & X_{(n)} > \theta_0 \text{ or } X_{(n)} \leq \theta_0 \alpha^{1/n} \\ 0 & \text{otherwise} \end{cases}$$

is a UMP test of size α.
(c) Show that the results in (a) and (b) still hold if the discrete uniform distribution is replaced by the uniform distribution $U(0, \theta)$, $\theta > 0$.

18. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(\theta, b)$, $\theta \in \mathcal{R}$, $b > 0$.
(a) Derive a UMP test of size α for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, when b is known.
(b) For testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1 < \theta_0$, show that any

UMP test T_* of size α satisfies $\beta_{T_*}(\theta_1) = 1 - (1-\alpha)e^{-n(\theta_0-\theta_1)/b}$.
(c) For testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1 < \theta_0$, show that the power of any size α test that rejects H_0 when $Y \le c_1$ or $Y \ge c_2$ is the same as that in part (b), where $Y = (X_{(1)} - \theta_0)/\sum_{i=1}^n (X_i - X_{(1)})$.
(d) Show that the test in part (c) is a UMP test of size α for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \ne \theta_0$.
(e) Derive a UMP test of size α for testing $H_0 : \theta = \theta_0, b = b_0$ versus $H_1 : \theta < \theta_0, b < b_0$.

19. Let $X_1, ..., X_n$ be i.i.d. from the Pareto distribution $Pa(\theta, c)$, $\theta > 0$, $c > 0$.
(a) Derive a UMP test of size α for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \ne \theta_0$ when c is known.
(b) Derive a UMP test of size α for testing $H_0 : \theta = \theta_0, c = c_0$ versus $H_1 : \theta < \theta_0, c > c_0$.

20. Prove part (ii) of Theorem 6.3.

21. Let F_1 and F_2 be two c.d.f.'s on $\mathcal{R}$. Show that $F_1(x) \le F_2(x)$ for all x if and only if $\int g(x)dF_2(x) \le \int g(x)dF_1(x)$ for any nondecreasing function g.

22. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $p = P(X_1 = 1)$. Determine the c_i's and γ_i's in (6.15) and (6.16) for testing $H_0 : p \le 0.2$ or $p \ge 0.7$ when $\alpha = 0.1$ and $n = 15$. Find the power of the UMP test (6.15) when $p = 0.4$.

23. Consider Example 6.10. Suppose that $\theta_2 = -\theta_1$. Show that $c_2 = -c_1$ and discuss how to find a c_2 such that the test $T_* = I_{(-c_2, c_2)}(\bar{X})$ has size α.

24. Suppose that the distribution of X is in a family of p.d.f.'s indexed by a real-valued parameter θ; there is a real-valued sufficient statistic $U(X)$ such that the $f_{\theta_2}(u)/f_{\theta_1}(u)$ is strictly increasing in u for $\theta_1 < \theta_2$, where $f_\theta(u)$ is the Lebesgue p.d.f. of $U(X)$ and is continuous in u for each θ; and that for all $\theta_1 < \theta_2 < \theta_3$ and $u_1 < u_2 < u_3$,

$$\begin{vmatrix} f_{\theta_1}(u_1) & f_{\theta_1}(u_2) & f_{\theta_1}(u_3) \\ f_{\theta_2}(u_1) & f_{\theta_2}(u_2) & f_{\theta_2}(u_3) \\ f_{\theta_3}(u_1) & f_{\theta_3}(u_2) & f_{\theta_3}(u_3) \end{vmatrix} > 0.$$

Show that the conclusions of Theorem 6.3 remain valid.

25. Suppose that X has the p.d.f. (6.10). Consider hypotheses (6.13) or (6.14). Show that a UMP test does not exist. (Hint: this follows from a consideration of the UMP tests for the one-sided hypotheses $H_0 : \theta \ge \theta_1$ and $H_0 : \theta \le \theta_2$.)

26. (p-values). Suppose that X has a distribution from a parametric family indexed by θ. Consider a family of nonrandomized level α tests for $H_0 : \theta = \theta_0$ (or $\theta \le \theta_0$) with rejection region C_α such that $P_{\theta=\theta_0}(X \in C_\alpha) = \alpha$ for all $0 < \alpha < 1$ and $C_{\alpha_1} = \cap_{\alpha>\alpha_1} C_\alpha$ for all $0 < \alpha_1 < 1$.
(a) Show that the p-value is $\hat{\alpha}(x) = \inf\{\alpha : x \in C_\alpha\}$.
(b) Show that when $\theta = \theta_0$, $\hat{\alpha}(X)$ has the uniform distribution $U(0,1)$.
(c) If the tests with rejection regions C_α are unbiased, show that under H_1, $P_\theta(\hat{\alpha}(X) \le \alpha) \ge \alpha$.

27. In the proof of Theorem 6.4, show that
(a) (6.30) is equivalent to (6.31);
(b) (6.31) is equivalent to (6.29) with T_* replaced by T.

28. Show that the UMPU tests for hypotheses in Theorem 6.4 are unique a.s. $\mathcal{P}$ if attention is restricted to tests depending on Y and U.

29. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $p = P(X_1 = 1)$. Derive a UMPU test of size α for $H_0 : p = p_0$ versus $H_1 : p \ne p_0$ when
(a) $n = 10$, $\alpha = 0.1$, and $p_0 = 0.2$;
(b) $n = 10$, $\alpha = 0.05$, and $p_0 = 0.4$.

30. Suppose that X has the Poisson distribution $P(\theta)$ with an unknown $\theta > 0$. Show that (6.29) reduces to

$$\sum_{x=c_1+1}^{c_2-1} \frac{\theta_0^{x-1} e^{-\theta_0}}{(x-1)!} + \sum_{i=1}^{2} (1-\gamma_i) \frac{\theta_0^{c_i-1} e^{-\theta_0}}{(c_i-1)!} = 1 - \alpha,$$

provided that $c_1 > 1$.

31. Let X be a random variable from the geometric distribution $G(p)$. Find a UMPU test of size α for $H_0 : p = p_0$ versus $H_1 : p \ne p_0$.

32. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with unknown μ and σ^2.
(a) Show how the power of the one-sample t-test relates a noncentral t-distribution.
(b) Show that the power of the one-sample t-test is an increasing function of μ/σ in the one-sided case ($H_0 : \mu \le \mu_0$ versus $H_1 : \mu > \mu_0$), and of $|\mu|/\sigma$ in the two-sided case ($H_0 : \mu = \mu_0$ versus $H_1 : \mu \ne \mu_0$).

33. Let $X_1, ..., X_n$ be i.i.d. from the gamma distribution $\Gamma(\theta, \gamma)$ with unknown θ and γ.
(a) For testing $H_0 : \gamma \le \gamma_0$ versus $H_1 : \gamma > \gamma_0$ and $H_0 : \gamma = \gamma_0$ versus $H_1 : \gamma \ne \gamma_0$, show that there exist UMPU tests whose rejection regions are based on $V = \prod_{i=1}^n (X_i/\bar{X})$.

(b) For testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, show that a UMPU test rejects H_0 when $\sum_{i=1}^n X_i > C(\prod_{i=1}^n X_i)$ for some function C.

34. Let X_1 and X_2 be independently distributed as the Poisson distributions $P(\lambda_1)$ and $P(\lambda_2)$, respectively.
(a) Find a UMPU test of size α for testing $H_0 : \lambda_1 \geq \lambda_2$ versus $H_1 : \lambda_1 < \lambda_2$.
(b) Calculate the power of the UMPU test in (a) when $\alpha = 0.1$, $(\lambda_1, \lambda_2) = (0.1, 0.2)$, (1,2), (10,20), and (0.1,0.4).

35. Consider the binomial problem in Example 6.11.
(a) Prove the claim about $P(Y = y|U = u)$.
(b) Find a UMPU test of size α for testing $H_0 : p_1 \geq p_2$ versus $H_1 : p_1 < p_2$.
(c) Repeat (b) for $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$.

36. Let X_1 and X_2 be independently distributed as the negative binomial distributions $NB(p_1, n_1)$ and $NB(p_2, n_2)$, respectively, where n_i's are known and p_i's are unknown.
(a) Show that there exists a UMPU test of size α for testing $H_0 : p_1 \leq p_2$ versus $H_1 : p_1 > p_2$.
(b) Determine the conditional distribution $P_{Y|U=u}$ in Theorem 6.4 when $n_1 = n_2 = 1$.

37. In Example 6.12, show that A and B are independent if and only if $\log \frac{p_{11}}{p_{22}} = \log \frac{p_{12}}{p_{22}} + \log \frac{p_{21}}{p_{22}}$.

38. Let X_1 and X_2 be independently distributed according to p.d.f.'s given by (6.10) with ξ, η, θ, Y, and h replaced by ξ_i, η_i, θ_i, Y_i, and h_i, $i = 1, 2$, respectively. Show that there exists a UMPU test of size α for testing
(a) $H_0 : \eta_2(\theta_2) - \eta_1(\theta_1) \leq \eta_0$ versus $H_1 : \eta_2(\theta_2) - \eta_1(\theta_1) > \eta_0$;
(b) $H_0 : \eta_2(\theta_2) + \eta_1(\theta_1) \leq \eta_0$ versus $H_1 : \eta_2(\theta_2) + \eta_1(\theta_1) > \eta_0$.

39. Let X_j, $j = 1, 2, 3$, be independent from the Poisson distributions $P(\lambda_j)$, $j = 1, 2, 3$, respectively. Show that there exists a UMPU test of size α for testing $H_0 : \lambda_1\lambda_2 \leq \lambda_3^2$ versus $H_1 : \lambda_1\lambda_2 > \lambda_3^2$.

40. Let $X_{11}, ..., X_{1n_1}$ and $X_{21}, ..., X_{2n_2}$ be two independent samples i.i.d. from the gamma distributions $\Gamma(\theta_1, \gamma_1)$ and $\Gamma(\theta_2, \gamma_2)$, respectively.
(a) Assume that γ_1 and γ_2 are known. For testing $H_0 : \theta_1 \leq \theta_2$ versus $H_1 : \theta_1 > \theta_2$ and $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_1 \neq \theta_2$, show that there exist UMPU tests and that the rejection regions can be determined by using beta distributions.
(b) If γ_i's are unknown in (a), show that there exist UMPU tests and describe their general forms.

(c) Assume that $\theta_1 = \theta_2$ (unknown). For testing $H_0 : \gamma_1 \leq \gamma_2$ versus $H_1 : \gamma_1 > \gamma_2$ and $H_0 : \gamma_1 = \gamma_2$ versus $H_1 : \gamma_1 \neq \gamma_2$, show that there exist UMPU tests and describe their general forms.

41. Let N be a random variable with the following discrete p.d.f.:

$$P(N = n) = C(\lambda)a(n)\lambda^n I_{\{0,1,2,...\}}(n),$$

where $\lambda > 0$ is unknown and a and C are known functions. Suppose that given $N = n$, $X_1, ..., X_n$ are i.i.d. from the p.d.f. given in (6.10). Show that, based on $(N, X_1, ..., X_N)$, there exists a UMPU test of size α for $H_0 : \eta(\theta) \leq \eta_0$ versus $H_1 : \eta(\theta) > \eta_0$.

42. Let $X_{i1}, ..., X_{in_i}$, $i = 1, 2$, be two independent samples i.i.d. from $N(\mu_i, \sigma^2)$, respectively, $n_i \geq 2$. Show that a UMPU test of size α for $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ rejects H_0 when $|t(X)| > t_{n_1+n_2-1,\alpha/2}$, where $t(X)$ is given by (6.37) and $t_{n_1+n_2-1,\alpha}$ is the $(1-\alpha)$th quantile of the t-distribution $t_{n_1+n_2-1}$. Derive the power function of this test.

43. In the two-sample problem discussed in §6.2.3, show that when $n_1 = n_2$, a UMPU test of size α for testing $H_0 : \sigma_2^2 = \Delta_0\sigma_1^2$ versus $H_1 : \sigma_2^2 \neq \Delta_0\sigma_1^2$ rejects H_0 when

$$\max\left(\frac{S_2^2}{\Delta_0 S_1^2}, \frac{\Delta_0 S_1^2}{S_2^2}\right) > \frac{1-c}{c},$$

where $\int_0^c f_{n_1-1,n_1-1}(v)dv = \alpha/2$ and $f_{a,b}$ is the p.d.f. of the beta distribution $B(a, b)$.

44. Suppose that $X_i = \beta_0 + \beta_1 t_i + \varepsilon_i$, where t_i's are fixed constants that are not all the same, ε_i's are i.i.d. from $N(0, \sigma^2)$, and β_0, β_1, and σ^2 are unknown parameters. Derive a UMPU test of size α for testing
(a) $H_0 : \beta_0 \leq \theta_0$ versus $H_1 : \beta_0 > \theta_0$;
(b) $H_0 : \beta_0 = \theta_0$ versus $H_1 : \beta_0 \neq \theta_0$;
(c) $H_0 : \beta_1 \leq \theta_0$ versus $H_1 : \beta_1 > \theta_0$;
(d) $H_0 : \beta_1 = \theta_0$ versus $H_1 : \beta_1 \neq \theta_0$.

45. Consider the normal linear model in §6.2.3 (i.e., model (3.25) with $\varepsilon = N_n(0, \sigma^2 I_n)$). For testing $H_0 : \sigma^2 \leq \sigma_0^2$ versus $H_1 : \sigma^2 > \sigma_0^2$ and $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 \neq \sigma_0^2$, show that UMPU tests of size α are functions of SSR and their rejection regions can be determined using chi-square distributions.

46. In the problem of testing for independence in the bivariate normal family, show that

(a) the p.d.f. in (6.44) is of the form (6.23) and identify φ;
(b) the sample correlation coefficient R is independent of U when $\rho = 0$;
(c) R is linear in Y, and V in (6.45) has the t-distribution t_{n-2} when $\rho = 0$.

47. Let $X_1, ..., X_n$ be i.i.d. bivariate normal with the p.d.f. in (6.44) and let $S_j^2 = \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2$ and $S_{12} = \sum_{j=1}^n (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)$.
(a) Show that a UMPU test for testing $H_0 : \sigma_2/\sigma_1 = \Delta_0$ versus $H_1 : \sigma_2/\sigma_1 \neq \Delta_0$ rejects H_0 when

$$R = |\Delta_0^2 S_1^2 - S_2^2| / \sqrt{(\Delta_0^2 S_1^2 + S_2^2)^2 - 4\Delta_0^2 S_{12}^2} > c.$$

(b) Find the p.d.f. of R in (a) when $\sigma_2/\sigma_1 = \Delta_0$.
(c) Assume that $\sigma_1 = \sigma_2$. Show that a UMPU test for $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ rejects H_0 when

$$V = |\bar{X}_2 - \bar{X}_1| / \sqrt{S_1^2 + S_2^2 - 2S_{12}} > c.$$

(d) Find the p.d.f. of V in (c) when $\mu_1 = \mu_2$.

48. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(a, \theta)$ with unknown a and θ.
(a) For testing $H_0 : \theta = 1$ versus $H_1 : \theta \neq 1$, show that a UMPU test of size α rejects H_0 when $V < c_1$ or $V > c_2$, where $V = 2\sum_{i=1}^n (X_i - X_{(1)})$, c_i's are determined by

$$\int_{c_1}^{c_2} \chi_{2n-2}^2(v)dv = \int_{c_1}^{c_2} \chi_{2n}^2(v)dv = 1 - \alpha,$$

and $\chi_m^2(v)$ is the p.d.f. of the chi-square distribution χ_m^2.
(b) For testing $H_0 : a = 0$ versus $H_1 : a \neq 0$, show that a UMPU test of size α rejects H_0 when $X_{(1)} < 0$ or $2nX_{(1)}/V > c$, where c is determined by

$$(n-1)\int_0^c (1+v)^{-n} dv = 1 - \alpha.$$

49. Let $X_1, ..., X_n$ be i.i.d. random variables from the uniform distribution $U(\theta, \vartheta)$, $-\infty < \theta < \vartheta < \infty$.
(a) Show that the conditional distribution of $X_{(1)}$ given $X_{(n)} = x$ is the distribution of the minimum of a sample of size $n-1$ from the uniform distribution $U(\theta, x)$.
(b) Find a UMPU test of size α for testing $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$.

50. Let $\mathcal{X} = \{x \in \mathcal{R}^n$: all components of x are nonzero$\}$ and $\mathcal{G}$ be the group of transformations $g(x) = (cx_1, ..., cx_n)$, $c > 0$. Show that a maximal invariant under $\mathcal{G}$ is $(\text{sgn}x_n, x_1/x_n, ..., x_{n-1}/x_n)$, where sgn$x$ is 1 or -1 as x is positive or negative.

51. Let $X_1, ..., X_n$ be i.i.d. with a Lebesgue p.d.f. $\sigma^{-1}f(x/\sigma)$ and f_i, $i = 0, 1$, be two known Lebesgue p.d.f.'s on $\mathcal{R}$ that are either 0 for $x < 0$ or symmetric about 0. Consider $H_0 : f = f_0$ versus $H_1 : f = f_1$ and $\mathcal{G} = \{g_r : r > 0\}$ with $g_r(x) = rx$.
(a) Show that a UMPI test rejects H_0 when
$$\frac{\int_0^\infty v^{n-1} f_1(vX_1) \cdots f_1(vX_n) dv}{\int_0^\infty v^{n-1} f_0(vX_1) \cdots f_0(vX_n) dv} > c.$$
(b) Show that if $f_0 = N(0, 1)$ and $f_1(x) = e^{-|x|}/2$, then the UMPI test in (a) rejects H_0 when $(\sum_{i=1}^n X_i^2)^{1/2} / \sum_{i=1}^n |X_i| > c$.
(c) Show that if $f_0(x) = I_{(0,1)}(x)$ and $f_1(x) = 2xI_{(0,1)}(x)$, then the UMPI test in (a) rejects H_0 when $X_{(n)}/(\prod_{i=1}^n X_i)^{1/n} < c$.
(d) Find the value of c in part (c) when the UMPI test is of size α.

52. Consider the location-scale family problem (with unknown parameters μ and σ) in Example 6.13.
(a) Show that W is maximal invariant under the given $\mathcal{G}$.
(b) Show that Proposition 6.2 applies and find the form of the functional $\theta(f_{i,\mu,\sigma})$.
(c) Derive the p.d.f. of $W(X)$ under H_i, $i = 0, 1$.
(d) Obtain a UMPI test.

53. In Example 6.13, find the rejection region of the UMPI test when $X_1, ..., X_n$ are i.i.d. and
(a) $f_{0,\mu,\sigma}$ is $N(\mu, \sigma^2)$ and $f_{1,\mu,\sigma}$ is the p.d.f. of the uniform distribution $U(\mu - \frac{1}{2}\sigma, \mu + \frac{1}{2}\sigma)$;
(b) $f_{0,\mu,\sigma}$ is $N(\mu, \sigma^2)$ and $f_{1,\mu,\sigma}$ is the p.d.f. of the exponential distribution $E(\mu, \sigma)$;
(c) $f_{0,\mu,\sigma}$ is the p.d.f. of $U(\mu - \frac{1}{2}\sigma, \mu + \frac{1}{2}\sigma)$ and $f_{1,\mu,\sigma}$ is the p.d.f. of $E(\mu, \sigma)$;
(d) $f_{0,\mu}$ is $N(\mu, 1)$ and $f_{1,\mu}(x) = \exp\{-e^{x-\mu} + x - \mu\}$.

54. Prove the claims in Example 6.15.

55. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with unknown μ and σ^2. Consider the problem of testing $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ and the group of transformations $g_c(X_i) = cX_i$, $c \neq 0$.
(a) Show that the testing problem is invariant under $\mathcal{G}$.
(b) Show that the one-sample two-sided t-test in §6.2.3 is a UMPI test.

56. Prove the claims in Example 6.16.

57. Consider Example 6.16 with H_0 and H_1 replaced by $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$, and with $\mathcal{G}$ changed to $\{g_{c_1,c_2,r} : c_1 = c_2 \in \mathcal{R}, r \neq 0\}$.
(a) Show that the testing problem is invariant under $\mathcal{G}$.
(b) Show that the two-sample two-sided t-test in §6.2.3 is a UMPI test.

58. Show that the UMPU tests in Exercise 33(a) and Exercise 40(a) are also UMPI tests under $\mathcal{G} = \{g_r : r > 0\}$ with $g_r(x) = rx$.

59. In Example 6.17, show that $t(X)$ has the noncentral t-distribution $t_{n-1}(\sqrt{n}\theta)$; the family $\{f_\theta(t) : \theta \in \mathcal{R}\}$ has monotone likelihood ratio in t; and that for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, a test that is UMP among all unbiased tests based on $t(X)$ rejects H_0 when $t(X) < c_1$ or $t(X) > c_2$. (Hint: consider Exercise 24.)

60. Let X_1 and X_2 be independently distributed as the exponential distributions $E(0, \theta_i)$, $i = 1, 2$, respectively. Define $\theta = \theta_1/\theta_2$.
(a) For testing $H_0 : \theta \leq 1$ versus $\theta > 1$, show that the problem is invariant under the group of transformations $g_c(x_1, x_2) = (cx_1, cx_2)$, $c > 0$, and that a UMPI test of size α rejects H_0 when $X_2/X_1 > (1-\alpha)/\alpha$.
(b) For testing $H_0 : \theta = 1$ versus $\theta \neq 1$, show that the problem is invariant under the group of transformations in (a) and $g(x_1, x_2) = (x_2, x_1)$, and that a UMPI test of size α rejects H_0 when $X_1/X_2 > (2-\alpha)/\alpha$ and $X_2/X_1 > (2-\alpha)/\alpha$.

61. Let $X_1, ..., X_m$ and $Y_1, ..., Y_n$ be two independent samples i.i.d. from the exponential distributions $E(a_1, \theta_1)$ and $E(a_2, \theta_2)$, respectively. Let $g_{r,c,d}(x, y) = (rx_1 + c, ..., rx_m + c, ry_1 + d, ..., ry_n + d)$ and let $\mathcal{G} = \{g_{r,c,d} : r > 0, c \in \mathcal{R}, d \in \mathcal{R}\}$.
(a) Show that a UMPI test of size α for testing $H_0 : \theta_1/\theta_2 \geq \Delta_0$ versus $H_1 : \theta_1/\theta_2 < \Delta_0$ rejects H_0 when $\sum_{i=1}^n (Y_i - Y_{(1)}) > c \sum_{i=1}^m (X_i - X_{(1)})$ for some constant c.
(b) Find the value of c in (a).
(c) Show that the UMPI test in (a) is also a UMPU test.

62. Let $M(Y)$ be given by (6.51) and $W = M(Y)(n-r)/s$.
(a) Show that W has the noncentral F-distribution $F_{s,n-r}(\theta)$.
(b) Show that $f_{\theta_1}(w)/f_0(w)$ is an increasing function of w for any given $\theta_1 \neq 0$.

63. Consider normal linear model (6.38). Show that
(a) the UMPI test derived in §6.3.2 for testing (6.49) is the same as the UMPU test for (6.40) given in §6.2.3 when $s = 1$ and $\theta_0 = 0$;

(b) the test with the rejection region $W > F_{s,n-r,\alpha}$ is a UMPI test of size α for testing $H_0 : \beta L^\tau = \theta_0$ versus $H_1 : \beta L^\tau \neq \theta_0$, where W is given by (6.52), θ_0 is a fixed constant, L is the same as that in (6.49), and $F_{s,n-r,\alpha}$ is the $(1-\alpha)$th quantile of the F-distribution $F_{s,n-r}$.

64. In Examples 6.18-6.19,
(a) prove the claim in Example 6.19;
(b) derive the distribution of W by applying Cochran's theorem (Theorem 1.5).

65. (Two-way additive model). Suppose that

$$X_{ij} = N(\mu_{ij}, \sigma^2), \qquad i = 1, ..., a,\ j = 1, ..., b,$$

where $\mu_{ij} = \mu + \alpha_i + \beta_j$, $\sum_{i=1}^a \alpha_i = \sum_{j=1}^b \beta_j = 0$, and X_{ij}'s are independent. Derive the forms of the UMPI tests in §6.3.2 for testing (6.54) and (6.55).

66. Let X_{ijk}, $i = 1, ..., a$, $j = 1, ..., b$, $k = 1, ..., c$, be independently normally distributed with common variance σ^2 and means

$$E(X_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k$$

$(\sum_{i=1}^a \alpha_i = \sum_{j=1}^b \beta_j = \sum_{k=1}^c \gamma_k = 0)$. Derive the UMPI test based on the W in (6.52) for testing $H_0 : \alpha_i = 0$ for all i versus $H_1 : \alpha_i \neq 0$ for some i.

67. Let $X_1, ..., X_m$ and $Y_1, ..., Y_n$ be independently normally distributed with a common variance σ^2 and means

$$E(X_i) = \mu_x + \beta_x(u_i - \bar{u}) \qquad E(Y_j) = \mu_y + \beta_y(v_j - \bar{v}),$$

where u_i's and v_j's are some constants and μ_x, μ_y, β_x, and β_y are unknown. Derive the UMPI test based on the W in (6.52) for testing
(a) $H_0 : \beta_x = \beta_y$ versus $H_1 : \beta_x \neq \beta_y$;
(b) $H_0 : \beta_x = \beta_y$ and $\mu_x = \mu_y$ versus $H_1 : \beta_x \neq \beta_y$ or $\mu_x \neq \mu_y$.

68. Let $(X_1, Y_1), ..., (X_n, Y_n)$ be i.i.d. from a bivariate normal distribution with unknown means, variances, and correlation coefficient ρ.
(a) Show that the problem of testing $H_0 : \rho \leq \rho_0$ versus $H_1 : \rho > \rho_0$ is invariant under $\mathcal{G}$ containing transformations $rX_i + c$, $sY_i + d$, $i = 1, ..., n$, where $r > 0$, $s > 0$, $c \in \mathcal{R}$, and $d \in \mathcal{R}$. Show that a UMPI test rejects H_0 when $R > c$, where R is the sample correlation coefficient given in (6.45). (Hint: see Lehmann (1986, p. 340).)
(b) Show that the problem of testing $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$ is invariant in addition (to the transformations in (a)) under the transformation $g(X_i, Y_i) = (X_i, -Y_i)$, $i = 1, ..., n$. Show that a UMPI test rejects H_0 when $|R| > c$.

69. Under the random effects model (6.57), show that
(a) SSA/SSR is maximal invariant under the group of transformations described in §6.3.2;
(b) the UMPI test for (6.58) derived in §6.3.2 is also a UMPU test.

70. Show that (6.60) is equivalent to (6.61).

71. Prove part (iii) of Proposition 6.5.

72. Let $X_1, ..., X_n$ be i.i.d. from the discrete uniform distribution on $\{1, ..., \theta\}$, where θ is an integer ≥ 2. Find a level α LR test for
(a) $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, where θ_0 is a known integer ≥ 2;
(b) $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

73. Let X be a sample of size 1 from the p.d.f. $2\theta^{-2}(\theta - x)I_{(0,\theta)}(x)$, where $\theta > 0$ is unknown. Find an LR test of size α for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

74. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(a, \theta)$.
(a) Suppose that θ is known. Find an LR test of size α for testing $H_0 : a \leq a_0$ versus $H_1 : a > a_0$.
(b) Suppose that θ is known. Find an LR test of size α for testing $H_0 : a = a_0$ versus $H_1 : a \neq a_0$.
(c) Repeat part (a) for the case where θ is also unknown.

75. Let $X_1, ..., X_n$ be i.i.d. from the Pareto distribution $Pa(\gamma, \theta)$, where $\theta > 0$ and $\gamma > 0$ are unknown.
(a) Show that an LR test for $H_0 : \theta = 1$ versus $H_1 : \theta \neq 1$ rejects H_0 when $Y < c_1$ or $Y > c_2$, where c_1 and c_2 are positive constants and $Y = \log\left(\prod_{i=1}^n X_i / X_{(1)}^n\right)$.
(b) Find values of c_1 and c_2 so that the LR test in (a) has size α.

76. Let $X_{i1}, ..., X_{in_i}$, $i = 1, 2$, be two independent samples i.i.d. from $N(\mu_i, \sigma_i^2)$, $i = 1, 2$, respectively, where μ_i's and σ_i^2's are unknown. For testing $H_0 : \sigma_2^2/\sigma_1^2 = \Delta_0$ versus $H_1 : \sigma_2^2/\sigma_1^2 \neq \Delta_0$, derive an LR test of size α and compare it with the UMPU test derived in §6.2.3.

77. Let $(X_{11}, X_{12}), ..., (X_{n1}, X_{n2})$ be i.i.d. from a bivariate normal distribution with unknown mean and covariance matrix. For testing $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$, where ρ is the correlation coefficient, show that the test rejecting H_0 when $|W| > 0$ is an LR test, where

$$W = \sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) \Big/ \left[\sum_{i=1}^n (X_{i1} - \bar{X}_1)^2 + \sum_{i=1}^n (X_{i2} - \bar{X}_2)^2\right].$$

Find the distribution of W.

78. Let X_1 and X_2 be independently distributed as the Poisson distributions $P(\lambda_1)$ and $P(\lambda_2)$, respectively. Find an LR test of significance level α for testing
(a) $H_0 : \lambda_1 = \lambda_2$ versus $H_1 : \lambda_1 \neq \lambda_2$;
(b) $H_0 : \lambda_1 \geq \lambda_2$ versus $H_1 : \lambda_1 < \lambda_2$. (Is this test a UMPU test?)

79. Let X_1 and X_2 be independently distributed as the binomial distributions $Bi(p_1, n_1)$ and $Bi(p_2, n_2)$, respectively, where n_i's are known and p_i's are unknown. Find an LR test of significance level α for testing
(a) $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$;
(b) $H_0 : p_1 \geq p_2$ versus $H_1 : p_1 < p_2$. (Is this test a UMPU test?)

80. Let X_1 and X_2 be independently distributed as the negative binomial distributions $NB(p_1, n_1)$ and $NB(p_2, n_2)$, respectively, where n_i's are known and p_i's are unknown. Find an LR test of significance level α for testing
(a) $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$;
(b) $H_0 : p_1 \leq p_2$ versus $H_1 : p_1 > p_2$.

81. Let X_1 and X_2 be independently distributed as the exponential distributions $E(0, \theta_i)$, $i = 1, 2$, respectively. Define $\theta = \theta_1/\theta_2$. Find an LR test of size α for testing
(a) $H_0 : \theta = 1$ versus $H_1 : \theta \neq 1$;
(b) $H_0 : \theta \leq 1$ versus $H_1 : \theta > 1$.

82. Let $X_{i1}, ..., X_{in_i}$, $i = 1, 2$, be independently distributed as the beta distributions with p.d.f.'s $\theta_i x^{\theta_i - 1} I_{(0,1)}(x)$, $i = 1, 2$, respectively. For testing $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_1 \neq \theta_2$, find the forms of LR test, Wald's test, and Rao's score test.

83. Prove Theorem 6.6(ii) for the special case where $\Theta_0 = \{\theta_0\}$.

84. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$.
(a) Suppose that $\sigma^2 = \gamma\mu^2$ with unknown $\gamma > 0$ and $\mu \in \mathcal{R}$. Find an LR test for testing $H_0 : \gamma = 1$ versus $H_1 : \gamma \neq 1$.
(b) In the testing problem in (a), find the forms of W_n for Wald's test and R_n for Rao's score test, and discuss whether Theorems 6.5 and 6.6 can be applied.
(c) Repeat (a) and (b) when $\sigma^2 = \gamma\mu$ with unknown $\gamma > 0$ and $\mu > 0$.

85. Suppose that $X_1, ..., X_n$ are i.i.d. from the Weibull distribution with p.d.f. $\theta^{-1}\gamma x^{\gamma-1} e^{-x^\gamma/\theta} I_{(0,\infty)}(x)$, where $\gamma > 0$ and $\theta > 0$ are unknown. Consider the problem of testing $H_0 : \gamma = 1$ versus $H_1 : \gamma \neq 1$.
(a) Find an LR test and discuss whether Theorem 6.5 can be applied to this case.
(b) Find the forms of W_n for Wald's test and R_n for Rao's score test.

86. Suppose that $X = (X_1, ..., X_k)$ has the multinomial distribution with the parameter $\boldsymbol{P} = (p_1, ..., p_k)$. Consider the problem of testing (6.65). Find the forms of W_n for Wald's test and R_n for Rao's score test.

87. Prove the claims in Example 6.24.

88. Consider testing independence in the $r \times c$ contingency table problem in Example 6.24. Find the forms of W_n for Wald's test and R_n for Rao's score test.

89. Under the conditions of Theorems 6.5 and 6.6, show that Wald's tests are Chernoff-consistent (Definition 2.13) if α is chosen to be $\alpha_n \to 0$ and $\chi^2_{r,\alpha_n} = o(n)$ as $n \to \infty$, where $\chi^2_{r,\alpha}$ is the $(1-\alpha)$th quantile of χ^2_r.

90. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $\theta = P(X_1 = 1)$.
(a) Let the prior $\Pi(\theta)$ be the c.d.f. of the beta distribution $B(a, b)$. Find the Bayes factor and the Bayes test for $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$.
(b) Let the prior c.d.f. be $\pi_0 I_{[\theta_0,\infty)}(\theta) + (1-\pi_0)\Pi(\theta)$, where Π is the same as that in (a). Find the Bayes factor and the Bayes test for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

91. Let $X_1, ..., X_n$ be i.i.d. from the Poisson distribution $P(\theta)$.
(a) Let the prior $\Pi(\theta) = (1 - e^{-\theta})I_{(0,\infty)}(\theta)$. Find the Bayes factor and the Bayes test for $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$.
(b) Let the prior be $\pi_0 I_{[\theta_0,\infty)}(\theta) + (1-\pi_0)\Pi(\theta)$, where Π is the same as that in (a). Find the Bayes factor and the Bayes test for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

92. Find a condition under which the UMPI test given in Example 6.17 is better than the sign test given by (6.73) in terms of their power functions under H_1.

93. For testing (6.75), show that a test T satisfying (6.76) is of size α and that the test in (6.77) satisfies (6.76).

94. Let $\mathcal{G}$ be the class of transformations $g(x) = (\psi(x_1), ..., \psi(x_n))$, where ψ is continuous, odd, and strictly increasing. Show that (R_+, R_-) is maximal invariant under $\mathcal{G}$. (Hint: see Example 6.14).

95. Under H_0, obtain the distribution of W in (6.78) for the one-sample Wilcoxon signed rank test, when $n = 3$ or 4.

96. For the one-sample Wilcoxon signed rank test, show that t_0 and σ_0^2 in (6.80) are equal to $\frac{1}{4}$ and $\frac{1}{12}$, respectively.

97. Using the results in §5.2.2, derive a two-sample rank test for testing (6.75) that has limiting size α.

98. Prove Theorem 6.10(i).

99. Show that the one-sided and two-sided Kolmogorov-Smirnov tests are consistent according to Definition 2.13.

100. Show that the distribution of $C_n(F)$ in (6.83) does depend on F. (Hint: construct i.i.d. random variables $U_1, ..., U_n$ from the uniform distribution $U(0,1)$ such that $P(X_i = F^{-1}(U_i)) = 1$, $i = 1, ..., n$.)

101. Show that the Cramér-von Mises tests are consistent.

102. In Example 6.27, show that the one-sample Wilcoxon signed rank test is consistent.

103. Let $X_1, ..., X_n$ be i.i.d. from a c.d.f. F on $\mathcal{R}^d$ and $\theta = E(X_1)$.
(a) Derive the empirical likelihood ratio for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.
(b) Let $\theta = (\vartheta, \varphi)$. Derive the profile empirical likelihood ratio for testing $H_0 : \vartheta = \vartheta_0$ versus $H_1 : \vartheta \neq \vartheta_0$.

104. Prove Theorem 6.12(ii).

105. Let $X_{i1}, ..., X_{in_i}$, $i = 1, 2$, be two independent samples i.i.d. from F_i on $\mathcal{R}$, $i = 1, 2$, respectively, and let $\mu_i = E(X_i)$.
(a) Show that the two-sample t-test derived in §6.2.3 for testing $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ has asymptotic significance level α and is consistent, if $n_1 \to \infty$, $n_1/n_2 \to c \in (0, 1)$, and $\sigma_1^2 = \sigma_2^2$.
(b) Derive a consistent asymptotic test for testing $H_0 : \mu_1/\mu_2 = \Delta_0$ versus $H_1 : \mu_1/\mu_2 \neq \Delta_0$, assuming that $\mu_2 \neq 0$.

106. Consider the general linear model (3.25) with i.i.d. ε_i's having $E(\varepsilon_i) = 0$ and $E(\varepsilon_i^2) = \sigma^2$.
(a) Under the conditions of Theorem 3.12, derive a consistent asymptotic test based on the LSE $\hat{\beta} l^\tau$ for testing $H_0 : \beta l^\tau = \theta_0$ versus $H_1 : \beta l^\tau \neq \theta_0$, where $l \in \mathcal{R}(Z)$.
(b) Show that the LR test in Example 6.21 has asymptotic significance level α and is consistent.

107. Let $\hat{\theta}_n$ be an estimator of a real-valued parameter θ such that (6.87) holds for any θ and let $\hat{V}_n$ be a consistent estimator of V_n. Suppose that $V_n \to 0$.
(a) Show that the test with rejection region $\hat{V}_n^{-1/2}(\hat{\theta}_n - \theta_0) > z_{1-\alpha}$ is a consistent asymptotic test for testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$.
(b) Apply the result in (a) to show that the one-sample one-sided t-test in §6.2.3 is a consistent asymptotic test.

Chapter 7

Confidence Sets

Various methods of constructing confidence sets are introduced in this chapter, along with studies of properties of confidence sets. Throughout this chapter $X = (X_1, ..., X_n)$ denotes a sample from a population $P \in \mathcal{P}$; $\theta = \theta(P)$ denotes a functional from $\mathcal{P}$ to $\Theta \subset \mathcal{R}^k$ for a fixed integer k; and $C(X)$ denotes a *confidence set* for θ, a set in $\mathcal{B}_\Theta$ (the class of Borel sets on Θ) depending only on X. We adopt the basic concepts of confidence sets introduced in §2.4.3. In particular, $\inf_{P\in\mathcal{P}} P(\theta \in C(X))$ is the confidence coefficient of $C(X)$ and, if the confidence coefficient of $C(X)$ is $\geq 1-\alpha$ for fixed $\alpha \in (0,1)$, then we say that $C(X)$ has significance level $1-\alpha$ or $C(X)$ is a level $1-\alpha$ confidence set.

7.1 Construction of Confidence Sets

In this section, we introduce some basic methods for constructing confidence sets that have a given significance level (or confidence coefficient) for any fixed n. Properties and comparisons of confidence sets are given in §7.2.

7.1.1 Pivotal quantities

Perhaps the most popular method of constructing confidence sets is the use of pivotal quantities defined as follows.

Definition 7.1. A known Borel function of (X, θ), $\Re(X, \theta)$, is a *pivotal quantity* if the distribution of $\Re(X, \theta)$ does not depend on P. ∎

Note that a pivotal quantity depends on P through $\theta = \theta(P)$. A pivotal quantity is usually not a statistic, although its distribution is known.

With a pivotal quantity $\Re(X,\theta)$, a level $1-\alpha$ confidence set for any given $\alpha \in (0,1)$ can be obtained as follows. First, find two constants c_1 and c_2 such that

$$P(c_1 \le \Re(X,\theta) \le c_2) \ge 1-\alpha. \tag{7.1}$$

Next, define

$$C(X) = \{\theta \in \Theta : \ c_1 \le \Re(X,\theta) \le c_2\}. \tag{7.2}$$

Then $C(X)$ is a level $1-\alpha$ confidence set, since

$$\begin{aligned}\inf_{P\in\mathcal{P}} P\big(\theta \in C(X)\big) &= \inf_{P\in\mathcal{P}} P\big(c_1 \le \Re(X,\theta) \le c_2\big)\\ &= P\big(c_1 \le \Re(X,\theta) \le c_2\big)\\ &\ge 1-\alpha.\end{aligned}$$

Note that the confidence coefficient of $C(X)$ may not be $1-\alpha$. If $\Re(X,\theta)$ has a continuous c.d.f., then we can choose c_i's such that the equality in (7.1) holds and, therefore, the confidence set $C(X)$ has confidence coefficient $1-\alpha$.

In a given problem, there may not exist any pivotal quantity, or there may be many different pivotal quantities. When there are many pivotal quantities, one has to choose one based on some principles or criteria, which are discussed in §7.2. For example, pivotal quantities based on sufficient statistics are certainly preferred. In many cases we also have to choose c_i's in (7.1) based on some criteria.

When $\Re(X,\theta)$ and c_i's are chosen, we need to compute the confidence set $C(X)$ in (7.2). This can be done by inverting $c_1 \le \Re(X,\theta) \le c_2$. For example, if θ is real-valued and $\Re(X,\theta)$ is monotone in θ when X is fixed, then $C(X) = \{\theta : \underline{\theta}(X) \le \theta \le \overline{\theta}(X)\}$ for some $\underline{\theta}(X) < \overline{\theta}(X)$, i.e., $C(X)$ is an interval (finite or infinite); if $\Re(X,\theta)$ is not monotone, then $C(X)$ may be a union of several intervals. For real-valued θ, a confidence interval rather than a complex set such as a union of several intervals is generally preferred since it is simple and the result is easy to interpret. When θ is multivariate, inverting $c_1 \le \Re(X,\theta) \le c_2$ may be complicated. In most cases where explicit forms of $C(X)$ do not exist, $C(X)$ can still be obtained numerically.

Example 7.1 (Location-scale families). Suppose that $X_1, ..., X_n$ are i.i.d. with a Lebesgue p.d.f. $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$, where $\mu \in \mathcal{R}$, $\sigma > 0$, and f is a known Lebesgue p.d.f.

Consider first the case where σ is known and $\theta = \mu$. For any fixed i, $X_i - \mu$ is a pivotal quantity. Also, $\bar{X} - \mu$ is a pivotal quantity, since any function of independent pivotal quantites is pivotal. In many cases $\bar{X} - \mu$ is preferred. Let c_1 and c_2 be constants such that $P(c_1 \le \bar{X} - \mu \le c_2) = 1-\alpha$.

Then $C(X)$ in (7.2) is

$$C(X) = \{\mu : c_1 \leq \bar{X} - \mu \leq c_2\} = \{\mu : \bar{X} - c_2 \leq \mu \leq \bar{X} - c_1\},$$

i.e., $C(X)$ is the interval $[\bar{X} - c_2, \bar{X} - c_1] \subset \mathcal{R} = \Theta$. This interval has confidence coefficient $1 - \alpha$. The choice of c_i's is not unique. Some criteria discussed in §7.2 can be applied to choose c_i's. One particular choice (not necessarily the best choice) frequently used by practitioners is $c_1 = -c_2$. The resulting $C(X)$ is symmetric about $\bar{X}$ and is also an *equal-tail* confidence interval (a confidence interval $[\underline{\theta}, \overline{\theta}]$ is equal-tail if $P(\theta < \underline{\theta}) = P(\theta > \overline{\theta})$) if the distribution of $\bar{X}$ is symmetric about μ. Note that the confidence interval in Example 2.31 is a special case of the intervals considered here.

Consider next the case where μ is known and $\theta = \sigma$. The following quantities are pivotal: $(X_i - \mu)/\sigma$, $i = 1, ..., n$, $\prod_{i=1}^n (X_i - \mu)/\sigma$, $(\bar{X} - \mu)/\sigma$, and S/σ, where S^2 is the sample variance. Consider the confidence set (7.2) based on S/σ. Let c_1 and c_2 be chosen such that $P(c_1 \leq S/\sigma \leq c_2) = 1 - \alpha$. If both c_i's are positive, then

$$C(X) = \{\sigma : S/c_2 \leq \sigma \leq S/c_1\} = [S/c_2, S/c_1]$$

is a finite interval. Similarly, if $c_1 = 0$ $(0 < c_2 < \infty)$ or $c_2 = \infty$ $(0 < c_1 < \infty)$, then $C(X) = [S/c_2, \infty)$ or $(0, S/c_1]$.

When $\theta = \sigma$ and μ is also unknown, S/σ is still a pivotal quantity and, hence, confidence intervals of σ based on S are still valid. Note that $(\bar{X} - \mu)/\sigma$ and $\prod_{i=1}^n (X_i - \mu)/\sigma$ are not pivotal when μ is unknown.

Finally, we consider the case where both μ and σ are unknown and $\theta = \mu$. There are still many different pivotal quantities, but the most commonly used pivotal quantity is $t(X) = \sqrt{n}(\bar{X} - \mu)/S$. The distribution of $t(X)$ does not depend on (μ, σ). When f is normal, $t(X)$ has the t-distribution t_{n-1}. The pivotal quantity $t(X)$ is often called a studentized statistic or t-statistic, although $t(X)$ is not a statistic and $t(X)$ does not have a t-distribution when f is not normal. A confidence interval for μ based on $t(X)$ is of the form

$$\{\mu : c_1 \leq \sqrt{n}(\bar{X} - \mu)/S \leq c_2\} = [\bar{X} - c_2 S/\sqrt{n}, \bar{X} - c_1 S/\sqrt{n}],$$

where c_i's are chosen so that $P(c_1 \leq t(X) \leq c_2) = 1 - \alpha$. ∎

Example 7.2. Let $X_1, ..., X_n$ be i.i.d. random variables from the uniform distribution $U(0, \theta)$. Consider the problem of finding a confidence set for θ. Note that the family $\mathcal{P}$ in this case is a scale family so that the results in Example 7.1 can be used. But a better confidence interval can be obtained based on the sufficient and complete statistic $X_{(n)}$ for which $X_{(n)}/\theta$ is a pivotal quantity (Example 7.13). Note that $X_{(n)}/\theta$ has the Lebesgue p.d.f.

$nx^{n-1}I_{(0,1)}(x)$. Hence c_i's in (7.1) should satisfy $c_2^n - c_1^n = 1-\alpha$. The resulting confidence interval for θ is $[c_2^{-1}X_{(n)}, c_1^{-1}X_{(n)}]$. Choices of c_i's are discussed in Example 7.13. ∎

Example 7.3 (Fieller's interval). Let $X_i = (X_{i1}, X_{i2})$, $i = 1, ..., n$, be i.i.d. bivariate normal with unknown $(\mu_1, \mu_2) = E(X_1)$ and $\mathrm{Var}(X_1)$. Let $\theta = \mu_2/\mu_1$ be the parameter of interest ($\mu_1 \neq 0$). Define $Y_i(\theta) = X_{i2} - \theta X_{i1}$. Then $Y_1(\theta), ..., Y_n(\theta)$ are i.i.d. with $N(0, \sigma_2^2 - 2\theta\sigma_{12} + \theta^2\sigma_1^2)$, where $\sigma_j^2 = \mathrm{Var}(X_{1j})$ and $\sigma_{12} = \mathrm{Cov}(X_{11}, X_{12})$. Let

$$S^2(\theta) = \frac{1}{n-1}\sum_{i=1}^n [Y_i(\theta) - \bar{Y}(\theta)]^2 = S_2^2 - 2\theta S_{12} + \theta^2 S_1^2,$$

where $\bar{Y}(\theta)$ is the average of $Y_i(\theta)$'s and S_i^2 and S_{12} are sample variances and covariance based on X_{ij}'s. It follows from Examples 1.15 and 2.18 that $\sqrt{n}\bar{Y}(\theta)/S(\theta)$ has the t-distribution t_{n-1} and, therefore, is a pivotal quantity. Let $t_{n-1,\alpha}$ be the $(1-\alpha)$th quantile of the t-distribution t_{n-1}. Then

$$C(X) = \{\theta : n[\bar{Y}(\theta)]^2/S^2(\theta) \leq t_{n-1,\alpha/2}^2\}$$

is a confidence set for θ with confidence coefficient $1-\alpha$. Note that $n[\bar{Y}(\theta)]^2 = t_{n-1,\alpha/2}^2 S^2(\theta)$ defines a parabola in θ. Depending on the roots of the parabola, $C(X)$ can be a finite interval, the complement of a finite interval, or the whole real line (exercise). ∎

Example 7.4. Consider the normal linear model $X = N_n(\beta Z^\tau, \sigma^2 I_n)$, where $\theta = \beta$ is a p-vector of unknown parameters and Z is a known $n \times p$ matrix of full rank. A pivotal quantity is

$$\Re(X, \beta) = \frac{(\hat{\beta} - \beta)Z^\tau Z(\hat{\beta} - \beta)^\tau/p}{\|X - \hat{\beta}Z^\tau\|^2/(n-p)},$$

where $\hat{\beta}$ is the LSE of β. By Theorem 3.8 and Example 1.15, $\Re(X, \beta)$ has the F-distribution $F_{p,n-p}$. We can then obtain a confidence set

$$C(X) = \{\beta : c_1 \leq \Re(X, \beta) \leq c_2\}.$$

Note that $\{\beta : \Re(X, \beta) < c\}$ is the interior of an ellipsoid in $\mathcal{R}^p$. ∎

The following result indicates that in many problems, there exist pivotal quantities.

Proposition 7.1. Let $T(X) = (T_1(X), ..., T_s(X))$ and $T_1, ..., T_s$ be independent statistics. Suppose that each T_i has a continuous c.d.f. $F_{T_i,\theta}$ indexed by θ. Then $\Re(X, \theta) = \prod_{i=1}^s F_{T_i,\theta}(T_i(X))$ is a pivotal quantity.

Proof. The result follows from the fact that $F_{T_i,\theta}(T_i)$'s are i.i.d. from the uniform distribution $U(0,1)$. ∎

When $X_1, ..., X_n$ are i.i.d. from a parametric family indexed by θ, the simplest way to apply Proposition 7.1 is to take $T(X) = X$. However, the resulting pivotal quantity may not be the best pivotal quantity. For instance, the pivotal quantity in Example 7.2 is a function of the one obtained by applying Proposition 7.1 with $T(X) = X_{(n)}$ $(s = 1)$, which is better than the one obtained by using $T(X) = X$ (Example 7.13).

The result in Proposition 7.1 holds even when P is in a nonparametric family, but in a nonparametric problem, it may be difficult to find a statistic T whose c.d.f. is indexed by θ, the parameter vector of interest.

When θ and T in Proposition 7.1 are real-valued, we can use the following result to construct confidence intervals for θ even when the c.d.f. of T is not continuous.

Theorem 7.1. Suppose that P is in a parametric family indexed by a real-valued θ. Let $T(X)$ be a real-valued statistic with c.d.f. $F_{T,\theta}(t)$ and let α_1 and α_2 be fixed positive constants such that $\alpha_1 + \alpha_2 = \alpha < \frac{1}{2}$.
(i) Suppose that $F_{T,\theta}(t)$ and $F_{T,\theta}(t-)$ are nonincreasing in θ for each fixed t. Define

$$\overline{\theta} = \sup\{\theta : F_{T,\theta}(T) \geq \alpha_1\} \qquad \text{and} \qquad \underline{\theta} = \inf\{\theta : F_{T,\theta}(T-) \leq 1 - \alpha_2\}.$$

Then $[\underline{\theta}(T), \overline{\theta}(T)]$ is a level $1 - \alpha$ confidence interval for θ.
(ii) If $F_{T,\theta}(t)$ and $F_{T,\theta}(t-)$ are nondecreasing in θ for each t, then the same result holds with

$$\underline{\theta} = \inf\{\theta : F_{T,\theta}(T) \geq \alpha_1\} \qquad \text{and} \qquad \overline{\theta} = \sup\{\theta : F_{T,\theta}(T-) \leq 1 - \alpha_2\}.$$

(iii) If $F_{T,\theta}$ is a continuous c.d.f. for any θ, then $F_{T,\theta}(T)$ is a pivotal quantity and the confidence interval in (i) or (ii) has confidence coefficient $1 - \alpha$.
Proof. We only need to prove (i). Under the given condition, $\theta > \overline{\theta}$ implies $F_{T,\theta}(T) < \alpha_1$ and $\theta < \underline{\theta}$ implies $F_{T,\theta}(T-) > 1 - \alpha_2$. Hence,

$$P\big(\underline{\theta} \leq \theta \leq \overline{\theta}\big) \geq 1 - P\big(F_{T,\theta}(T) < \alpha_1\big) - P\big(F_{T,\theta}(T-) > 1 - \alpha_2\big).$$

The result follows from

$$P\big(F_{T,\theta}(T) < \alpha_1\big) \leq \alpha_1 \quad \text{and} \quad P\big(F_{T,\theta}(T-) > 1 - \alpha_2\big) \leq \alpha_2. \tag{7.3}$$

The proof of (7.3) is left as an exercise. ∎

When the parametric family in Theorem 7.1 has monotone likelihood ratio in $T(X)$, then it follows from Lemma 6.3 that the condition in Theorem 7.1(i) holds; in fact, it follows from Exercise 2 in §6.6 that $F_{T,\theta}(t)$ is

strictly decreasing for any t at which $0 < F_{T,\theta}(t) < 1$. If $F_{T,\theta}(t)$ is also continuous in θ, $\lim_{\theta\to\theta_-} F_{T,\theta}(t) > \alpha_1$ and $\lim_{\theta\to\theta_+} F_{T,\theta}(t) < \alpha_1$, where θ_- and θ_+ are the two ends of the parameter space, then $\overline{\theta}$ is the unique solution of $F_{T,\theta}(t) = \alpha_1$. A similar conclusion can be drawn for $\underline{\theta}$.

Theorem 7.1 can be applied to obtain the confidence interval for θ in Example 7.2 (exercise). The following example concerns a discrete $F_{T,\theta}$.

Example 7.5. Let $X_1, ..., X_n$ be i.i.d. random variables from the Poisson distribution $P(\theta)$ with an unknown $\theta > 0$ and $T(X) = \sum_{i=1}^n X_i$. Note that T is sufficient and complete for θ and has the Poisson distribution $P(n\theta)$. Thus,

$$F_{T,\theta}(t) = \sum_{j=0}^{t} \frac{e^{-n\theta}(n\theta)^j}{j!}, \qquad t = 0, 1, 2, ...$$

Since the Poisson family has monotone likelihood ratio in T and $0 < F_{T,\theta}(t) < 1$ for any t, $F_{T,\theta}(t)$ is strictly decreasing in θ. Also, $F_{T,\theta}(t)$ is continuous in θ and $F_{T,\theta}(t)$ tends to 1 and 0 as θ tends to 0 and ∞, respectively. Thus, Theorem 7.1 applies and $\overline{\theta}$ is the unique solution of $F_{T,\theta}(T) = \alpha_1$. Since $F_{T,\theta}(t-) = F_{T,\theta}(t-1)$ for $t > 0$, $\underline{\theta}$ is the unique solution of $F_{T,\theta}(t-1) = 1 - \alpha_2$ when $T = t > 0$ and $\underline{\theta} = 0$ when $T = 0$. In fact, in this case explicit forms of $\underline{\theta}$ and $\overline{\theta}$ can be obtained from

$$\frac{1}{\Gamma(t)} \int_\lambda^\infty x^{t-1} e^{-x} dx = \sum_{j=0}^{t-1} \frac{e^{-\lambda}\lambda^j}{j!}, \qquad t = 1, 2, ...$$

Using this equality, it can be shown (exercise) that

$$\overline{\theta} = (2n)^{-1}\chi^2_{2(T+1),1-\alpha_1} \qquad \text{and} \qquad \underline{\theta} = (2n)^{-1}\chi^2_{2T,\alpha_2}, \tag{7.4}$$

where $\chi^2_{r,\alpha}$ is the $(1-\alpha)$th quantile of the chi-square distribution χ^2_r and $\chi^2_{0,a}$ is defined to be 0. ∎

So far we have considered examples for parametric problems. In a nonparametric problem, a pivotal quantity may not exist and we have to consider approximate pivotal quantities (§7.3 and §7.4). The following is an example of a nonparametric problem in which there exist pivotal quantities.

Example 7.6. Let $X_1, ..., X_n$ be i.i.d. random variables from $F \in \mathcal{F}$ containing all continuous and symmetric distributions on $\mathcal{R}$. Suppose that F is symmetric about θ. Let $\tilde{R}(\theta)$ be the vector of ranks of $|X_i - \theta|$'s and $R_+(\theta)$ be the subvector of $\tilde{R}(\theta)$ containing ranks corresponding to positive $(X_i - \theta)$'s. Then, any real-valued Borel function of $R_+(\theta)$ is a pivotal quantity (see the discussion in §6.5.1). Various confidence sets can be constructed using these pivotal quantities. More details can be found in Example 7.10. ∎

7.1.2 Inverting acceptance regions of tests

Another popular method of constructing confidence sets is to use a close relationship between confidence sets and hypothesis tests. For any test T, the complement of $\{x : T(x) = 1\}$ is called the *acceptance region*. Note that this terminology is not precise when T is a randomized test.

Theorem 7.2. For each $\theta_0 \in \Theta$, let T_{θ_0} be a test for $H_0 : \theta = \theta_0$ (versus some H_1) with significance level α and acceptance region $A(\theta_0)$. For each x in the range of X, define

$$C(x) = \{\theta : x \in A(\theta)\}.$$

Then $C(X)$ is a level $1 - \alpha$ confidence set for θ. If T_{θ_0} is nonrandomized and has size α for every θ_0, then $C(X)$ has confidence coefficient $1 - \alpha$.
Proof. We prove the first assertion only. The proof for the second assertion is similar. Under the given condition,

$$\sup_{\theta=\theta_0} P\big(X \notin A(\theta_0)\big) = \sup_{\theta=\theta_0} P(T_{\theta_0} = 1) \le \alpha,$$

which is the same as

$$1 - \alpha \le \inf_{\theta=\theta_0} P\big(X \in A(\theta_0)\big) = \inf_{\theta=\theta_0} P\big(\theta_0 \in C(X)\big).$$

Since this holds for all θ_0, the result follows from

$$\inf_{P\in\mathcal{P}} P\big(\theta \in C(X)\big) = \inf_{\theta_0\in\Theta} \inf_{\theta=\theta_0} P\big(\theta_0 \in C(X)\big) \ge 1 - \alpha. \quad \blacksquare$$

The converse of Theorem 7.2 is also true, which is stated in the next result whose proof is left as an exercise.

Proposition 7.2. Let $C(X)$ be a confidence set for θ with significance level (or confidence coefficient) $1 - \alpha$. For any $\theta_0 \in \Theta$, define a region $A(\theta_0) = \{x : \theta_0 \in C(x)\}$. Then the test $T(X) = 1 - I_{A(\theta_0)}(X)$ has significance level α for testing $H_0 : \theta = \theta_0$ versus some H_1. $\blacksquare$

In general, $C(X)$ in Theorem 7.2 can be determined numerically, if it does not have an explicit form. Theorem 7.2 can be best illustrated in the case where θ is real-valued and $A(\theta) = \{Y : a(\theta) \le Y \le b(\theta)\}$ for a real-valued statistic $Y(X)$ and some nondecreasing functions $a(\theta)$ and $b(\theta)$. When we observe $Y = y$, $C(X)$ is an interval with limits $\underline{\theta}$ and $\overline{\theta}$, which are the θ-values at which the horizontal line $Y = y$ intersects the curves $Y = b(\theta)$ and $Y = a(\theta)$ (Figure 7.1), respectively. If $y = b(\theta)$ (or $y = a(\theta)$) has no solution or more than one solution, $\underline{\theta} = \inf\{\theta : y \le b(\theta)\}$

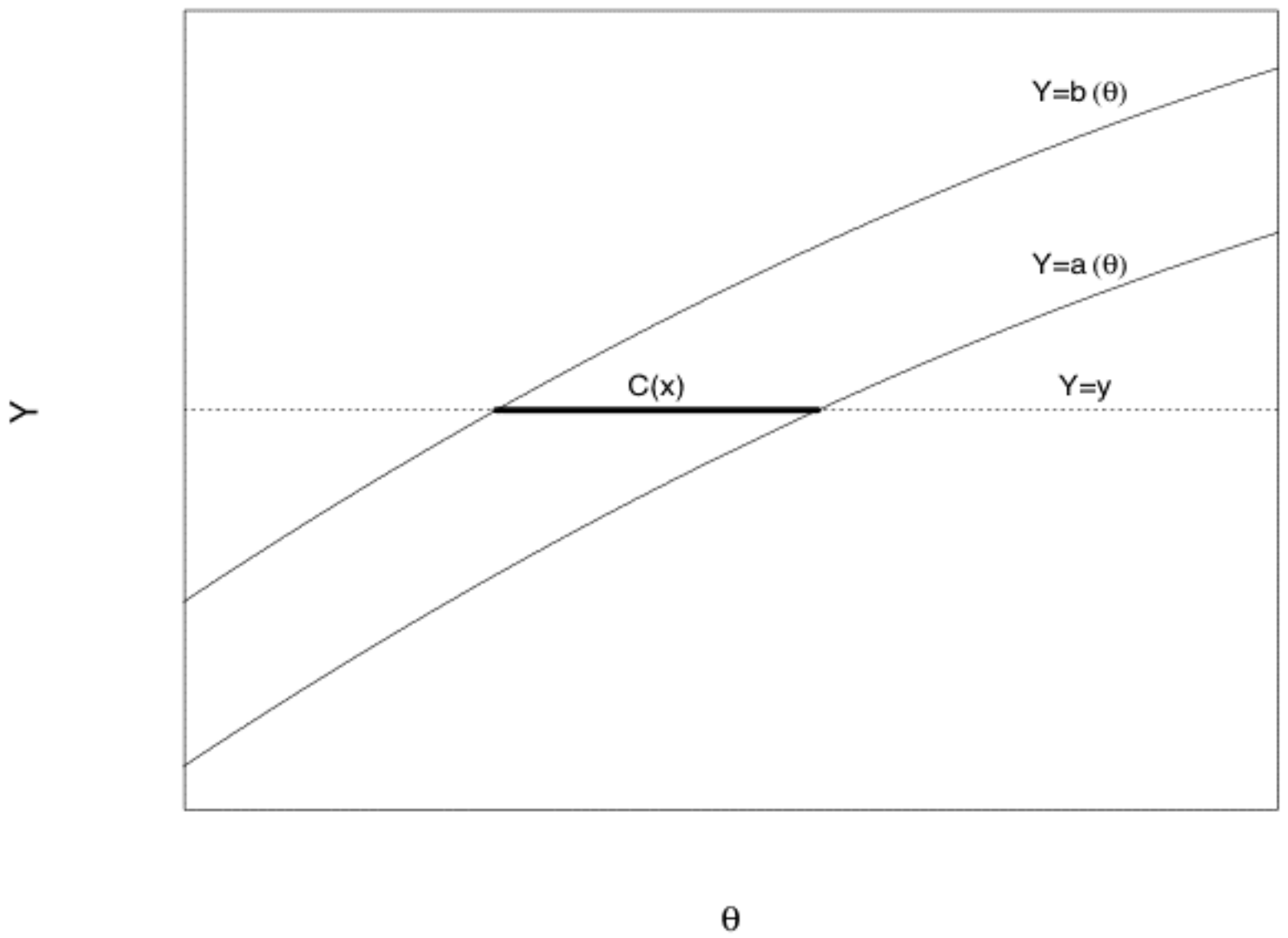

Figure 7.1: A confidence interval obtained by inverting $A(\theta) = [a(\theta), b(\theta)]$

(or $\overline{\theta} = \sup\{\theta : a(\theta) \leq y\}$). $C(X)$ does not include $\underline{\theta}$ (or $\overline{\theta}$) if and only if at $\underline{\theta}$ (or $\overline{\theta}$), $b(\theta)$ (or $a(\theta)$) is only left-continuous (or right-continuous).

Example 7.7. Suppose that X has the following p.d.f. in a one-parameter exponential family: $f_\theta(x) = \exp\{\eta(\theta)Y(x) - \xi(\theta)\}h(x)$, where θ is real-valued and $\eta(\theta)$ is nondecreasing in θ. First, we apply Theorem 7.2 with $H_0 : \theta = \theta_0$ and $H_1 : \theta > \theta_0$. By Theorem 6.2, the acceptance region of the UMP test of size α given by (6.11) is $A(\theta_0) = \{x : Y(x) \leq c(\theta_0)\}$, where $c(\theta_0) = c$ in (6.11). It can be shown (exercise) that $c(\theta)$ is nondecreasing in θ. Inverting $A(\theta)$ according to Figure 7.1 with $b(\theta) = c(\theta)$ and $a(\theta)$ ignored, we obtain $C(X) = [\underline{\theta}(X), \infty)$ or $(\underline{\theta}(X), \infty)$, a one-sided confidence interval for θ with significance level $1 - \alpha$. ($\underline{\theta}(X)$ is a called a lower confidence bound for θ in §2.4.3.) When the c.d.f. of $Y(X)$ is continuous, $C(X)$ has confidence coefficient $1 - \alpha$.

In the previous derivation, if $H_0 : \theta = \theta_0$ and $H_1 : \theta < \theta_0$ are considered, then $C(X) = \{\theta : Y(X) \geq c(\theta)\}$ and is of the form $(-\infty, \overline{\theta}(X)]$ or $(-\infty, \overline{\theta}(X))$. ($\overline{\theta}(X)$ is called an upper confidence bound for θ.)

Consider next $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$. By Theorem 6.4, the acceptance region of the UMPU test of size α defined in (6.28) is given by $A(\theta_0) = \{x : c_1(\theta_0) \leq Y(x) \leq c_2(\theta_0)\}$, where $c_i(\theta)$ are nondecreasing (exercise). A confidence interval can be obtained by inverting $A(\theta)$ according to Figure 7.1 with $a(\theta) = c_1(\theta)$ and $b(\theta) = c_2(\theta)$.

Let us consider a specific example in which $X_1, ..., X_n$ are i.i.d. binary random variables with $p = P(X_i = 1)$. Note that $Y(X) = \sum_{i=1}^n X_i$. Suppose that we need a lower confidence bound for p so that we consider $H_0 : p = p_0$ and $H_1 : p > p_0$. From Example 6.2, the acceptance region of a UMP test of size $\alpha \in (0, 1)$ is $A(p_0) = \{y : y \le m(p_0)\}$, where $m(p_0)$ is an integer between 0 and n such that

$$\sum_{j=m(p_0)+1}^{n} \binom{n}{j} p_0^j (1-p_0)^{n-j} \le \alpha < \sum_{j=m(p_0)}^{n} \binom{n}{j} p_0^j (1-p_0)^{n-j}.$$

Thus, $m(p)$ is an integer-valued, nondecreasing step-function of p. Define

$$\underline{p} = \inf\{p : m(p) \ge y\} = \inf\left\{p : \sum_{j=y}^{n} \binom{n}{j} p^j (1-p)^{n-j} \ge \alpha\right\}. \tag{7.5}$$

Then a level $1-\alpha$ confidence interval for p is $(\underline{p}, 1]$ (exercise). One can compare this confidence interval with the one obtained by applying Theorem 7.1 (exercise). See also Example 7.16. ∎

Example 7.8. Suppose that X has the following p.d.f. in a multiparameter exponential family: $f_{\theta,\varphi}(x) = \exp\{\theta Y(x) + U(x)\varphi^\tau - \zeta(\theta, \varphi)\}$. By Theorem 6.4, the acceptance region of a UMPU test of size α for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ or $H_0 : \theta = \theta_0$ versus $H_1 : \theta \ne \theta_0$ is

$$A(\theta_0) = \{(y, u) : y \le c_2(u, \theta_0)\}$$

or

$$A(\theta_0) = \{(y, u) : c_1(u, \theta_0) \le y \le c_2(u, \theta_0)\},$$

where $c_i(u, \theta)$, $i = 1, 2$, are nondecreasing functions of θ. Confidence intervals for θ can then be obtained by inverting $A(\theta)$ according to Figure 7.1 with $b(\theta) = c_2(u, \theta)$ and $a(\theta) = c_1(u, \theta)$ or $a(\theta) \equiv -\infty$, for any observed u.

Consider more specifically the case where X_1 and X_2 are independently distributed as the Poisson distributions $P(\lambda_1)$ and $P(\lambda_2)$, respectively, and we need a lower confidence bound for the ratio $\rho = \lambda_2/\lambda_1$. From Example 6.11, a UMPU test of size α for testing $H_0 : \rho = \rho_0$ versus $H_1 : \rho > \rho_0$ has the acceptance region $A(\rho_0) = \{(y, u) : y \le c(u, \rho_0)\}$, where $c(u, \rho_0)$ is determined by the conditional distribution of $Y = X_2$ given $U = X_1 + X_2 = u$. Since the conditional distribution of Y given $U = u$ is the binomial distribution $Bi(\rho/(1+\rho), u)$, we can use the result in Example 7.7, i.e., $c(u, \rho)$ is the same as $m(p)$ in Example 7.7 with $n = u$ and $p = \rho/(1+\rho)$. Then a level $1-\alpha$ lower confidence bound for p is $\underline{p}$ given by (7.5) with $n = u$. Since $\rho = p/(1-p)$ is a strictly increasing function of p, a level $1-\alpha$ lower confidence bound for ρ is $\underline{p}/(1-\underline{p})$. ∎

Example 7.9. Consider the normal linear model $X = N_n(\beta Z^\tau, \sigma^2 I_n)$ and the problem of constructing a confidence set for $\theta = \beta L^\tau$, where L is an $s \times p$ matrix of rank s and all rows of L are in $\mathcal{R}(Z)$. It follows from the discussion in §6.3.2 and Exercise 63 in §6.6 that a nonrandomized UMPI test of size α for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ has the acceptance region

$$A(\theta_0) = \{X : W(X, \theta_0) \leq c_\alpha\},$$

where c_α is the $(1-\alpha)$th quantile of the F-distribution $F_{s,n-r}$,

$$W(X, \theta) = \frac{[\|X - \hat{\beta}(\theta)Z^\tau\|^2 - \|X - \hat{\beta}Z^\tau\|^2]/s}{\|X - \hat{\beta}Z^\tau\|^2/(n-r)},$$

r is the rank of Z, $r \geq s$, $\hat{\beta}$ is the LSE of β and, for each fixed θ, $\hat{\beta}(\theta)$ is a solution of

$$\|X - \hat{\beta}(\theta)Z^\tau\|^2 = \min_{\beta:\beta L^\tau = \theta} \|X - \beta Z^\tau\|^2.$$

Inverting $A(\theta)$, we obtain the following confidence set for θ with confidence coefficient $1-\alpha$: $C(X) = \{\theta : W(X, \theta) \leq c_\alpha\}$, which is the interior and the boundary of an ellipsoid in $\mathcal{R}^s$. ∎

The last example concerns inverting the acceptance regions of tests in a nonparametric problem.

Example 7.10. Consider the problem in Example 7.6. We now derive a confidence interval for θ by inverting the acceptance regions of the signed rank tests given by (6.79). Note that testing whether the c.d.f. of X_i is symmetric about θ is equivalent to testing whether the c.d.f. of $X_i - \theta$ is symmetric about 0. Let c_i's be given by (6.79), W be given by (6.78), and, for each θ, let $R_+^o(\theta)$ be the vector of ordered components of $R_+(\theta)$ described in Example 7.6. A level $1-\alpha$ confidence set for θ is

$$C(X) = \{\theta : c_1 \leq W(R_+^o(\theta)) \leq c_2\}.$$

The region $C(X)$ can be computed numerically for any observed X. From the discussion in Example 7.6, $W(R_+^o(\theta))$ is a pivotal quantity and, therefore, $C(X)$ is the same as the confidence set obtained by using a pivotal quantity. ∎

7.1.3 The Bayesian approach

In Bayesian analysis, analogues to confidence sets are called *credible sets*. Consider a sample X from a population in a parametric family indexed by

$\theta \in \Theta \subset \mathcal{R}^k$ and dominated by a σ-finite measure. Let $f_\theta(x)$ be the p.d.f. of X and $\pi(\theta)$ be a prior p.d.f. w.r.t. a σ-finite measure λ on $(\Theta, \mathcal{B}_\Theta)$. Let

$$p_x(\theta) = f_\theta(x)\pi(\theta)/m(x)$$

be the posterior p.d.f. w.r.t. λ, where x is the observed X and $m(x) = \int_\Theta f_\theta(x)\pi(\theta)d\lambda$. For any $\alpha \in (0,1)$, a level $1-\alpha$ credible set for θ is any $C \in \mathcal{B}_\Theta$ with

$$P_{\theta|x}(\theta \in C) = \int_C p_x(\theta)d\lambda \geq 1-\alpha. \tag{7.6}$$

A level $1-\alpha$ *highest posterior density* (HPD) credible set for θ is defined to be the event

$$C(x) = \{\theta : p_x(\theta) \geq c_\alpha\}, \tag{7.7}$$

where c_α is chosen so that $\int_{C(x)} p_x(\theta)d\lambda \geq 1-\alpha$. When $p_x(\theta)$ has a continuous c.d.f., we can replace $\geq$ in (7.6) and (7.7) by $=$. An HPD credible set is often an interval with the shortest length among all credible intervals of the same level (Exercise 30).

The Bayesian credible sets and the confidence sets we have discussed so far are very different in terms of their meanings and interpretations, although sometimes they look similar. In a credible set, x is fixed and θ is considered random and the probability statement in (7.6) is w.r.t. the posterior probability $P_{\theta|x}$. On the other hand, in a confidence set θ is nonrandom (although unknown) but X is considered random, and the significance level is w.r.t. $P(\theta \in C(X))$, the probability related to the distribution of X. The set $C(X)$ in (7.7) is not necessarily a confidence set with significance level $1-\alpha$.

When $\pi(\theta)$ is constant, which is usually an improper prior, the HPD credible set $C(x)$ in (7.7) is related to the idea of maximizing likelihood (a non-Bayesian approach introduced in §4.4; see also §7.3.2), since $p_x(\theta) = f_\theta(x)/m(x)$ is proportional to $f_\theta(x) = \ell(\theta)$, the likelihood function. In such a case $C(X)$ may be a confidence set with significance level $1-\alpha$.

Example 7.11. Let $X_1, ..., X_n$ be i.i.d. as $N(\theta, \sigma^2)$ with an unknown $\theta \in \mathcal{R}$ and a known σ^2. Let $\pi(\theta)$ be the p.d.f. of $N(\mu_0, \sigma_0^2)$ with known μ_0 and σ_0^2. Then, $p_x(\theta)$ is the p.d.f. of $N(\mu_*(x), c^2)$ (Example 2.25), where $\mu_*(x)$ and c^2 are given by (2.28), and the HPD credible set in (7.7) is

$$\begin{aligned} C(x) &= \left\{\theta : e^{-[\theta-\mu_*(x)]^2/(2c^2)} \geq c_\alpha\sqrt{2\pi}c\right\} \\ &= \left\{\theta : |\theta-\mu_*(x)| \leq \sqrt{2}c[-\log(c_\alpha\sqrt{2\pi}c)]^{1/2}\right\}. \end{aligned}$$

Let Φ be the standard normal c.d.f. The quantity $\sqrt{2}c[-\log(c_\alpha\sqrt{2\pi}c)]^{1/2}$ must be $c\Phi^{-1}(1-\alpha/2)$, since it is chosen so that $P_{\theta|x}(C(x)) = 1-\alpha$ and

$P_{\theta|x} = N(\mu_*(x), c^2)$. Therefore,

$$C(x) = [\mu_*(x) - c\Phi^{-1}(1-\alpha/2),\ \mu_*(x) + c\Phi^{-1}(1-\alpha/2)].$$

If we let $\sigma_0^2 = \infty$, which is equivalent to taking the Lebesgue measure as the (improper) prior, then $\mu_*(x) = \bar{x}$, $c^2 = \sigma^2/n$, and

$$C(x) = [\,\bar{x} - \sigma\Phi^{-1}(1-\alpha/2)/\sqrt{n},\ \bar{x} + \sigma\Phi^{-1}(1-\alpha/2)/\sqrt{n}\,],$$

which looks the same as the classical confidence interval for θ with confidence coefficient $1-\alpha$ (Example 2.31). Although the Bayesian credible set coincides with the classical confidence interval, which is frequently the case when a noninformative prior is used, their interpretations are still different. ∎

More details about Bayesian credible sets can be found, for example, in Berger (1985, §4.3).

7.1.4 Prediction sets

In some problems the quantity of interest is the future (or unobserved) value of a random variable ξ. An inference procedure about a random quantity instead of an unknown nonrandom parameter is called *prediction*. If the distribution of ξ is known, then a level $1-\alpha$ *prediction set* for ξ is any event C satisfying $P_\xi(\xi \in C) \geq 1-\alpha$. In applications, however, the distribution of ξ is usually unknown.

Suppose that the distribution of ξ is related to the distribution of a sample X from which prediction will be made. For instance, $X = (X_1, ..., X_n)$ is the observed sample and $\xi = X_{n+1}$ is to be predicted, where $X_1, ..., X_n, X_{n+1}$ are i.i.d. random variables. A set $C(X)$ depending only on the sample X is said to be a level $1-\alpha$ *prediction set* for ξ if

$$\inf_{P\in\mathcal{P}} P\big(\xi \in C(X)\big) \geq 1-\alpha,$$

where P is the joint distribution of ξ and X and $\mathcal{P}$ contains all possible P.

Note that prediction sets are very similar and closely related to confidence sets. Hence, some methods for constructing confidence sets can be applied to obtained prediction sets. For example, if $\Re(X,\xi)$ is a pivotal quantity in the sense that its distribution does not depend on P, then a prediction set can be obtained by inverting $c_1 \leq \Re(X,\xi) \leq c_2$. The following example illustrates this idea.

Example 7.12. Many prediction problems encountered in practice can be formulated as follows. The variable ξ to be predicted is related to a vector-valued covariate ζ (called predictor) according to $E(\xi|\zeta) = \beta\zeta^\tau$, where β

is a p-vector of unknown parameters. Suppose that at $\zeta = Z_i$, we observe $\xi = X_i$, $i = 1, ..., n$, and X_i's are independent. The Z_i's are either fixed or random observations (in the latter case all probabilities given in the following discussion are conditional probabilities, given $Z_1, ..., Z_n$). Based on $(X_1, Z_1), ..., (X_n, Z_n)$, we would like to construct a prediction set for the value of $\xi = X_0$ when $\zeta = Z_0 \in \mathcal{R}(Z)$, where $Z^\tau = (Z_1^\tau, ..., Z_n^\tau)$.

Assume further that $X_0, X_1, ..., X_n$ are independently normal with a common variance, i.e., $X = (X_1, ..., X_n) = N_n(\beta Z^\tau, \sigma^2 I_n)$ follows a normal linear model and is independent of $X_0 = N(\beta Z_0^\tau, \sigma^2)$. Let $\hat{\beta}$ be the LSE of β, $\hat{\sigma}^2 = \|X - \hat{\beta} Z^\tau\|^2/(n-r)$, and $\|Z_0\|_Z^2 = Z_0(Z^\tau Z)^- Z_0^\tau$, where r is the rank of Z. Then

$$\Re(X, X_0) = \frac{X_0 - \hat{\beta} Z_0^\tau}{\hat{\sigma}\sqrt{1 + \|Z_0\|_Z^2}}$$

has the t-distribution t_{n-r} and, therefore, is a pivotal quantity. This is because X_0 and $\hat{\beta} Z_0^\tau$ are independently normal,

$$E(X_0 - \hat{\beta} Z_0^\tau) = 0, \qquad \mathrm{Var}(X_0 - \hat{\beta} Z_0^\tau) = \sigma^2(1 + \|Z_0\|_Z^2),$$

$(n-r)\hat{\sigma}^2$ has the chi-square distribution χ^2_{n-r}, and X_0, $\hat{\beta} Z_0^\tau$, and $\hat{\sigma}^2$ are independent (Theorem 3.8). A level $1-\alpha$ prediction interval for X_0 is then

$$\left[\hat{\beta} Z_0^\tau - t_{n-r,\alpha/2}\hat{\sigma}\sqrt{1 + \|Z_0\|_Z^2},\ \hat{\beta} Z_0^\tau + t_{n-r,\alpha/2}\hat{\sigma}\sqrt{1 + \|Z_0\|_Z^2}\right], \tag{7.8}$$

where $t_{n-r,\alpha}$ is the $(1-\alpha)$th quantile of the t-distribution t_{n-r}.

To compare prediction sets with confidence sets, let us consider a confidence interval for $E(X_0) = \beta Z_0^\tau$. Using the pivotal quantity

$$\Re(X, \beta Z_0^\tau) = \frac{\beta Z_0^\tau - \hat{\beta} Z_0^\tau}{\hat{\sigma}\|Z_0\|_Z},$$

we obtain the following confidence interval for βZ_0^τ with confidence coefficient $1-\alpha$:

$$\left[\hat{\beta} Z_0^\tau - t_{n-r,\alpha/2}\hat{\sigma}\|Z_0\|_Z,\ \hat{\beta} Z_0^\tau + t_{n-r,\alpha/2}\hat{\sigma}\|Z_0\|_Z\right]. \tag{7.9}$$

Since a random variable is more variable than its average (an unknown parameter), the prediction interval (7.8) is much longer than the confidence interval (7.9), although each of them covers the quantity of interest with probability $1-\alpha$. In fact, when $\|Z_0\|_Z^2 \to 0$ as $n \to \infty$, the length of the confidence interval (7.9) tends to 0 a.s., whereas the length of the prediction interval (7.8) tends to a positive constant a.s. ∎

Because of the similarity between confidence sets and prediction sets, in the rest of this chapter we do not discuss prediction sets in detail. Some examples are given in Exercises 21 and 22.

7.2 Properties of Confidence Sets

In this section we study some properties of confidence sets and introduce several criteria for comparing confidence sets.

7.2.1 Lengths of confidence intervals

For confidence intervals of a real-valued θ with the same confidence coefficient, an apparent measure of their performance is the interval length. Shorter confidence intervals are preferred, since they are more informative. In most problems, however, shortest-length confidence intervals do not exist. A common approach is to consider a reasonable class of confidence intervals (with the same confidence coefficient) and then find a confidence interval with the shortest length within the class.

When confidence intervals are constructed by using pivotal quantities or by inverting acceptance regions of tests, choosing a reasonable class of confidence intervals amounts to selecting good pivotal quantities or tests. Functions of sufficient statistics should be used, when sufficient statistics exist. In many problems pivotal quantities or tests are related to some point estimators of θ. For example, in a location family problem (Example 7.1), a confidence interval for $\theta = \mu$ is often of the form $[\hat{\theta} - c, \hat{\theta} + c]$, where $\hat{\theta}$ is an estimator of θ and c is a constant. In such a case a more accurate estimator of θ should intuitively result in a better confidence interval. For instance, when $X_1, ..., X_n$ are i.i.d. $N(\mu, 1)$, it can be shown (exercise) that the interval $[\bar{X} - c_1, \bar{X} + c_1]$ is better than the interval $[X_1 - c_2, X_1 + c_2]$ in terms of their lengths, where c_i's are chosen so that these confidence intervals have confidence coefficient $1 - \alpha$. However, we cannot have the same conclusion when X_i's are from the Cauchy distribution $C(\mu, 1)$ (exercise). The following is another example.

Example 7.13. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution $U(0, \theta)$ with an unknown $\theta > 0$. A confidence interval for θ of the form $[b^{-1}X_{(n)}, a^{-1}X_{(n)}]$ is derived in Example 7.2, where a and b are constants chosen so that this confidence interval has confidence coefficient $1 - \alpha$. Another confidence interval obtained by applying Proposition 7.1 with $T = X$ is of the form $[b_1^{-1}\tilde{X}, a_1^{-1}\tilde{X}]$, where $\tilde{X} = (\prod_{i=1}^n X_i)^{1/n}$. We now argue that when n is large enough, the former has a shorter length than the latter. Note that $\sqrt{n}(\tilde{X} - \theta)/\theta \to_d N(0, 1)$. Thus,

$$P\left(\left(1 + \frac{d}{\sqrt{n}}\right)^{-1}\tilde{X} \leq \theta \leq \left(1 + \frac{c}{\sqrt{n}}\right)^{-1}\tilde{X}\right) = P\left(\frac{c}{\sqrt{n}} \leq \frac{\tilde{X} - \theta}{\theta} \leq \frac{d}{\sqrt{n}}\right) \to 1 - \alpha$$

for some constants c and d. This means that $a_1 \approx 1 + c/\sqrt{n}$, $b_1 \approx 1 + d/\sqrt{n}$, and the length of $[b_1^{-1}\tilde{X}, a_1^{-1}\tilde{X}]$ converges to 0 a.s. at the rate $n^{-1/2}$. On

the other hand,

$$P\left(\left(1+\tfrac{d}{n}\right)^{-1}X_{(n)} \le \theta \le \left(1+\tfrac{c}{n}\right)^{-1}X_{(n)}\right) = P\left(\tfrac{c}{n} \le \tfrac{X_{(n)}-\theta}{\theta} \le \tfrac{d}{n}\right) \to 1-\alpha$$

for some constants c and d, since $n(X_{(n)}-\theta)/\theta$ has a known limiting distribution (Example 2.34). This means that the length of $[b^{-1}X_{(n)}, a^{-1}X_{(n)}]$ converges to 0 a.s. at the rate n^{-1} and, therefore, $[b^{-1}X_{(n)}, a^{-1}X_{(n)}]$ is shorter than $[b_1^{-1}\tilde{X}, a_1^{-1}\tilde{X}]$ for sufficiently large n a.s.

Similarly, one can show that the confidence interval based on the pivotal quantity $\bar{X}/\theta$ is not as good as $[b^{-1}X_{(n)}, a^{-1}X_{(n)}]$ in terms of their lengths.

Thus, it is reasonable to consider the class of confidence intervals of the form $[b^{-1}X_{(n)}, a^{-1}X_{(n)}]$ subject to $P(b^{-1}X_{(n)} \le \theta \le a^{-1}X_{(n)}) = 1-\alpha$. The shortest-length interval within this class can be derived as follows. Note that $X_{(n)}/\theta$ has the Lebesgue p.d.f. $nx^{n-1}I_{(0,1)}(x)$. Hence

$$1-\alpha = P(b^{-1}X_{(n)} \le \theta \le a^{-1}X_{(n)}) = \int_a^b nx^{n-1}dx = b^n - a^n.$$

This implies that $1 \ge b > a \ge 0$ and $\frac{da}{db} = (\frac{b}{a})^{n-1}$. Since the length of the interval $[b^{-1}X_{(n)}, a^{-1}X_{(n)}]$ is $\psi(a,b) = X_{(n)}(a^{-1}-b^{-1})$,

$$\frac{d\psi}{db} = X_{(n)}\left(\frac{1}{b^2} - \frac{1}{a^2}\frac{da}{db}\right) = X_{(n)}\frac{a^{n+1}-b^{n+1}}{b^2a^{n+1}} < 0.$$

Hence the minimum occurs at $b = 1$ $(a = \alpha^{1/n})$. This shows that the shortest-length interval is $[X_{(n)}, \alpha^{-1/n}X_{(n)}]$. ∎

As Example 7.13 indicates, whence a reasonable class of confidence intervals is chosen (using some good estimators, pivotal quantities, or tests), we may find the shortest-length confidence interval within the class by directly analyzing the lengths of the intervals. For a large class of problems, the following result can be used.

Theorem 7.3. Let θ be a real-valued parameter and $T(X)$ be a real-valued statistic.
(i) Let $U(X)$ be a positive statistic. Suppose that $(T-\theta)/U$ is a pivotal quantity having a Lebesgue p.d.f. f that is *unimodal* at $x_0 \in \mathcal{R}$ in the sense that $f(x)$ is nondecreasing for $x \le x_0$ and $f(x)$ is nonincreasing for $x \ge x_0$. Consider the following class of confidence intervals for θ:

$$\mathcal{C} = \left\{[T-bU, T-aU] : \int_a^b f(x)dx = 1-\alpha\right\}. \tag{7.10}$$

If $[T-b_*U, T-a_*U] \in \mathcal{C}$, $f(a_*) = f(b_*) > 0$, and $a_* \le x_0 \le b_*$, then the interval $[T-b_*U, T-a_*U]$ has the shortest length within $\mathcal{C}$.

(ii) Suppose that $T > 0$, $\theta > 0$, T/θ is a pivotal quantity having a Lebesgue p.d.f. f, and that $x^2 f(x)$ is unimodal at x_0. Consider the following class of confidence intervals for θ:

$$\mathcal{C} = \left\{ [b^{-1}T, a^{-1}T] : \int_a^b f(x)dx = 1 - \alpha \right\}. \tag{7.11}$$

If $[b_*^{-1}T, a_*^{-1}T] \in \mathcal{C}$, $a_*^2 f(a_*) = b_*^2 f(b_*) > 0$, and $a_* \le x_0 \le b_*$, then the interval $[b_*^{-1}T, a_*^{-1}T]$ has the shortest length within $\mathcal{C}$.

Proof. We prove (i) only. The proof of (ii) is left as an exercise. Note that the length of an interval in $\mathcal{C}$ is $(b-a)U$. Thus, it suffices to show that if $a < b$ and $b - a < b_* - a_*$, then $\int_a^b f(x)dx < 1 - \alpha$. Assume that $a < b$, $b - a < b_* - a_*$, and $a \le a_*$ (the proof for $a > a_*$ is similar).

If $b \le a_*$, then $a \le b \le a_* \le x_0$ and

$$\int_a^b f(x)dx \le f(a_*)(b-a) < f(a_*)(b_* - a_*) \le \int_{a_*}^{b_*} f(x)dx = 1 - \alpha,$$

where the first inequality follows from the unimodality of f, the strict inequality follows from $b - a < b_* - a_*$ and $f(a_*) > 0$, and the last inequality follows from the unimodality of f and the fact that $f(a_*) = f(b_*)$.

If $b > a_*$, then $a \le a_* < b < b_*$. By the unimodality of f,

$$\int_a^{a_*} f(x)dx \le f(a_*)(a_* - a) \qquad \text{and} \qquad \int_b^{b_*} f(x)dx \ge f(b_*)(b_* - b).$$

Then

$$\begin{aligned}
\int_a^b f(x)dx &= \int_{a_*}^{b_*} f(x)dx + \int_a^{a_*} f(x)dx - \int_b^{b_*} f(x)dx \\
&= 1 - \alpha + \int_a^{a_*} f(x)dx - \int_b^{b_*} f(x)dx \\
&\le 1 - \alpha + f(a_*)(a_* - a) - f(b_*)(b_* - b) \\
&= 1 - \alpha + f(a_*)[(a_* - a) - (b_* - b)] \\
&= 1 - \alpha + f(a_*)[(b - a) - (b_* - a_*)] \\
&< 1 - \alpha. \quad \blacksquare
\end{aligned}$$

Example 7.14. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with unknown μ and σ^2. Confidence intervals for $\theta = \mu$ using the pivotal quantity $\sqrt{n}(\bar{X} - \mu)/S$ form the class $\mathcal{C}$ in (7.10) with f being the p.d.f. of the t-distribution t_{n-1}, which is unimodal at $x_0 = 0$. Hence, we can apply Theorem 7.3(i). Since f is symmetric about 0, $f(a_*) = f(b_*)$ implies $a_* = -b_*$ (exercise). Therefore, the equal-tail confidence interval

$$\left[\bar{X} - t_{n-1,\alpha/2} S/\sqrt{n}, \bar{X} + t_{n-1,\alpha/2} S/\sqrt{n} \right] \tag{7.12}$$

has the shortest length within $\mathcal{C}$.

If $\theta = \mu$ and σ^2 is known, then we can replace S by σ and f by the standard normal p.d.f. (i.e., use the pivotal quantity $\sqrt{n}(\bar{X} - \mu)/\sigma$ instead of $\sqrt{n}(\bar{X} - \mu)/S$). The resulting confidence interval is

$$\left[\, \bar{X} - \Phi^{-1}(1 - \alpha/2)\sigma/\sqrt{n}, \bar{X} + \Phi^{-1}(1 - \alpha/2)\sigma/\sqrt{n} \,\right], \tag{7.13}$$

which is the shortest interval of the form $[\bar{X} - b, \bar{X} - a]$ with confidence coefficient $1-\alpha$. The difference in length of the intervals in (7.12) and (7.13) is a random variable so that we cannot tell which one is better in general. But the expected length of the interval (7.13) is always shorter than that of the interval (7.12) (exercise). This again shows the importance of picking the right pivotal quantity.

Consider next confidence intervals for $\theta = \sigma^2$ using the pivotal quantity $(n-1)S^2/\sigma^2$, which form the class $\mathcal{C}$ in (7.11) with f being the p.d.f. of the chi-square distribution χ^2_{n-1}. Note that $x^2 f(x)$ is unimodal, but not symmetric. By Theorem 7.3(ii), the shortest-length interval within $\mathcal{C}$ is

$$[\, b_*^{-1}(n-1)S^2, a_*^{-1}(n-1)S^2 \,], \tag{7.14}$$

where a_* and b_* are solutions of $a_*^2 f(a_*) = b_*^2 f(b_*)$ and $\int_{a_*}^{b_*} f(x)dx = 1 - \alpha$. Numerical values of a_* and b_* can be obtained (Tate and Klett, 1959). Note that this interval is not equal-tail.

If $\theta = \sigma^2$ and μ is known, then a better pivotal quantity is T/σ^2, where $T = \sum_{i=1}^n (X_i - \mu)^2$. One can show (exercise) that if we replace $(n-1)S^2$ by T and f by the p.d.f. of the chi-square distribution χ^2_n, then the resulting interval has shorter expected length than that of the interval in (7.14).

Suppose that we need a confidence interval for $\theta = \sigma$ when μ is unknown. Consider the class of confidence intervals

$$\left[\, b^{-1/2}\sqrt{n-1}\, S, a^{-1/2}\sqrt{n-1}\, S \,\right]$$

with $\int_a^b f(x)dx = 1 - \alpha$. The shortest-length interval, however, is not the one with the endpoints equal to the square roots of the endpoints of the interval (7.14) (Exercise 26(c)). ∎

Note that Theorem 7.3(ii) cannot be applied to obtain the result in Example 7.13 unless $n = 1$, since the p.d.f. of $X_{(n)}/\theta$ is strictly increasing when $n > 1$. A result similar to Theorem 7.3, which can be applied to Example 7.13, is given in Exercise 28.

The result in Theorem 7.3 can also be applied to justify the idea of HPD credible sets in Bayesian analysis (Exercise 30).

If a confidence interval has the shortest length within a class of confidence intervals, then its expected length is also the shortest within the

same class, provided that its expected length is finite. In a problem where a shortest-length confidence interval does not exist, we may have to use the expected length as the criterion in comparing confidence intervals. For instance, the expected length of the interval in (7.13) is always shorter than that of the interval in (7.12), whereas the probability that the interval in (7.12) is shorter than the interval in (7.13) is positive for any fixed n. Another example is the interval $[X_{(n)}, \alpha^{-1/n}X_{(n)}]$ in Example 7.13. Although we are not able to say that this interval has the shortest length among all confidence intervals for θ with confidence coefficient $1-\alpha$, we can show that it has the shortest expected length, using the results in Theorems 7.4 and 7.6 (§7.2.2).

For one-sided confidence intervals (confidence bounds) of a real-valued θ, their lengths may be infinity. We can use the distance between the confidence bound and θ as a criterion in comparing confidence bounds, which is equivalent to comparing the tightness of confidence bounds. Let $\underline{\theta}_j$, $j = 1, 2$, be two lower confidence bounds for θ with the same confidence coefficient. If $\underline{\theta}_1 - \theta \geq \underline{\theta}_2 - \theta$ is always true, then $\underline{\theta}_1 \geq \underline{\theta}_2$ and $\underline{\theta}_1$ is tighter (more informative) than $\underline{\theta}_2$. Again, since $\underline{\theta}_j$ are random, we may have to consider $E(\underline{\theta}_j - \theta)$ and choose $\underline{\theta}_1$ if $E(\underline{\theta}_1) \geq E(\underline{\theta}_2)$. As a specific example, consider i.i.d. $X_1, ..., X_n$ from $N(\theta, 1)$. If we use the pivotal quantity $\bar{X} - \mu$, then $\underline{\theta}_1 = \bar{X} - \Phi^{-1}(1-\alpha)/\sqrt{n}$. If we use the pivotal quantity $X_1 - \mu$, then $\underline{\theta}_2 = X_1 - \Phi^{-1}(1-\alpha)$. Clearly $E(\underline{\theta}_1) \geq E(\underline{\theta}_2)$. Although $\underline{\theta}_1$ is intuitively preferred, $\underline{\theta}_1 < \underline{\theta}_2$ with a positive probability for any fixed $n > 1$.

Some ideas discussed previously can be extended to the comparison of confidence sets for multivariate θ. For bounded confidence sets in $\mathcal{R}^k$, for example, we may consider their volumes (Lebesgue measures). However, in multivariate cases it is difficult to compare the volumes of confidence sets with different shapes. Some results about expected volumes of confidence sets are given in Theorem 7.6.

7.2.2 UMA and UMAU confidence sets

For a confidence set obtained by inverting the acceptance regions of some UMP or UMPU tests, it is expected that the confidence set inherits some optimality property.

Definition 7.2. Let $\theta \in \Theta$ be an unknown parameter and Θ' be a subset of Θ that does not contain the true parameter value θ. A confidence set $C(X)$ for θ with confidence coefficient $1-\alpha$ is said to be Θ'-*uniformly most accurate* (UMA) if and only if for any other confidence set $C_1(X)$ with confidence coefficient $1-\alpha$,

$$P\big(\theta' \in C(X)\big) \leq P\big(\theta' \in C_1(X)\big) \qquad \text{for all } \theta' \in \Theta'. \tag{7.15}$$

$C(X)$ is UMA if $\Theta' = \{\theta\}^c$ (the set containing all false values). ∎

The probabilities in (7.15) are probabilities of covering false values. Intuitively, confidence sets with small probabilities of covering wrong parameter values are preferred. The reason why we sometimes need to consider a Θ' different from $\{\theta\}^c$ is that for some confidence sets such as one-sided confidence intervals, we do not need to worry about the probabilities of covering some false values. For example, if we consider a lower confidence bound for a real-valued θ, we are asserting that θ is larger than a certain value and we only need to worry about covering values of θ that are too small. Thus, $\Theta' = \{\theta' \in \Theta : \theta' < \theta\}$. A similar discussion leads to the consideration of $\Theta' = \{\theta' \in \Theta : \theta' > \theta\}$ for upper confidence bounds.

Theorem 7.4. Let $C(X)$ be a confidence set for θ obtained by inverting the acceptance regions of nonrandomized tests T_{θ_0} for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \in \Theta_{\theta_0}$. Suppose that for each θ_0, T_{θ_0} is UMP of size α. Then $C(X)$ is Θ'-UMA with confidence coefficient $1 - \alpha$, where $\Theta' = \{\theta' : \theta \in \Theta_{\theta'}\}$.
Proof. The fact that $C(X)$ has confidence coefficient $1 - \alpha$ follows from Theorem 7.2. Let $C_1(X)$ be another confidence set with confidence coefficient $1 - \alpha$. By Proposition 7.2, the test $T_{1\theta_0}(X) = 1 - I_{A_1(\theta_0)}(X)$ with $A_1(\theta_0) = \{x : \theta_0 \in C_1(x)\}$ has size α for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \in \Theta_{\theta_0}$. For any $\theta' \in \Theta'$,

$$\begin{aligned} P\big(\theta' \in C(X)\big) &= P\big(T_{\theta'}(X) = 0\big) \\ &= 1 - P\big(T_{\theta'}(X) = 1\big) \\ &\le 1 - P\big(T_{1\theta'}(X) = 1\big) \\ &= P\big(X \in A_1(\theta')\big) \\ &= P\big(\theta' \in C_1(X)\big), \end{aligned}$$

where the second equality follows from the fact that $T_{\theta'}$ is nonrandomized and the inequality follows from the fact that $T_{\theta'}$ is UMP. ∎

Theorem 7.4 can be applied to construct UMA confidence bounds in problems where the population is in a one-parameter parametric family with monotone likelihood ratio so that UMP tests exist (Theorem 6.2). It can also be applied to a few cases to construct two-sided UMA confidence intervals. For example, the confidence interval $[X_{(n)}, \alpha^{-1/n} X_{(n)}]$ in Example 7.13 is UMA (exercise).

As we discussed in §6.2, in many problems there are UMPU tests but not UMP tests. This leads to the following definition.

Definition 7.3. Let $C(X)$ be a confidence set for θ with confidence coefficient $1-\alpha$ and Θ' be a subset of Θ that does not contain the true parameter

value θ.
(i) If $P(\theta' \in C(X)) \leq 1-\alpha$ for all $\theta' \in \Theta'$, then $C(X)$ is said to be Θ'-*unbiased* (unbiased if $\Theta' = \{\theta\}^c$).
(ii) If $C(X)$ is Θ'-unbiased and (7.15) holds for any other Θ'-unbiased confidence set $C_1(X)$ with confidence coefficient $1-\alpha$, then $C(X)$ is said to be Θ'-*uniformly most accurate unbiased* (UMAU). $C(X)$ is UMAU if $\Theta' = \{\theta\}^c$. ∎

Theorem 7.5. Let $C(X)$ be a confidence set for θ obtained by inverting the acceptance regions of nonrandomized tests T_{θ_0} for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \in \Theta_{\theta_0}$. If T_{θ_0} is unbiased of size α for each θ_0, then $C(X)$ is Θ'-unbiased with confidence coefficient $1-\alpha$, where $\Theta' = \{\theta' : \theta \in \Theta_{\theta'}\}$; if T_{θ_0} is also UMPU for each θ_0, then $C(X)$ is Θ'-UMAU. ∎

The proof of Theorem 7.5 is very similar to that of Theorem 7.4.

It follows from Theorem 7.5 and the results in §6.2 that the confidence intervals in (7.12), (7.13) and (7.14) are UMAU, since they can be obtained by inverting acceptance regions of UMPU tests (Exercise 14).

Example 7.15. Consider the normal linear model in Example 7.9 and the parameter $\theta = \beta l^\tau$, where $l \in \mathcal{R}(Z)$. From §6.2.3, the nonrandomized test with acceptance region

$$A(\theta_0) = \left\{X : \hat{\beta} l^\tau - \theta_0 > t_{n-r,\alpha}\sqrt{l(Z^\tau Z)^- l^\tau SSR/(n-r)}\right\}$$

is UMPU with size α for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta < \theta_0$, where $\hat{\beta}$ is the LSE of β and $t_{n-r,\alpha}$ is the $(1-\alpha)$th quantile of the t-distribution t_{n-r}. Inverting $A(\theta)$ we obtain the following Θ'-UMAU upper confidence bound with confidence coefficient $1-\alpha$ and $\Theta' = (\theta, \infty)$:

$$\overline{\theta} = \hat{\beta} l^\tau - t_{n-r,\alpha}\sqrt{l(Z^\tau Z)^- l^\tau SSR/(n-r)}.$$

A UMAU confidence interval for θ can be similarly obtained.

If $\theta = \beta L^\tau$ with L described in Example 7.9 and $s > 1$, then θ is multivariate. It can be shown that the confidence set derived in Example 7.9 is unbiased (exercise), but it may not be UMAU. ∎

The volume of a confidence set $C(X)$ for $\theta \in \mathcal{R}^k$ when $X = x$ is defined to be $\text{vol}(C(x)) = \int_{C(x)} d\theta'$ (if it is well defined), which is the Lebesgue measure of the set $C(x)$. In particular, if θ is real-valued and $C(X) = [\underline{\theta}(X), \overline{\theta}(X)]$ is a confidence interval, then $\text{vol}(C(x))$ is simply the length of $C(x)$. The next result reveals a relationship between the expected volume (length) and the probability of covering a false value of a confidence set (interval).

Theorem 7.6 (Pratt's theorem). Let X be a sample from P and $C(X)$ be a confidence set for $\theta \in \mathcal{R}^k$. Suppose that for each x in the range of X, $\text{vol}(C(x)) = \int_{C(x)} d\theta'$ is well defined. Then the expected volume of $C(X)$ is

$$E[\text{vol}(C(X))] = \int_{\theta \neq \theta'} P\big(\theta' \in C(X)\big) d\theta'. \tag{7.16}$$

Proof. By Fubini's theorem,

$$\begin{aligned}
E[\text{vol}(C(X))] &= \int \text{vol}(C(X)) dP \\
&= \int \left[\int_{C(x)} d\theta' \right] dP(x) \\
&= \int \int_{\theta' \in C(x)} d\theta' dP(x) \\
&= \int \left[\int_{\theta' \in C(x)} dP(x) \right] d\theta' \\
&= \int P\big(\theta' \in C(X)\big) d\theta' \\
&= \int_{\theta \neq \theta'} P\big(\theta' \in C(X)\big) d\theta'.
\end{aligned}$$

This proves the result. ∎

It follows from Theorem 7.6 that if $C(X)$ is UMA (or UMAU) with confidence coefficient $1 - \alpha$, then it has the smallest expected volume among all confidence sets (or all unbiased confidence sets) with confidence coefficient $1 - \alpha$. For example, the confidence intervals (7.13) in Example 7.14 (when σ^2 is known) and $[X_{(n)}, \alpha^{-1/n} X_{(n)}]$ in Example 7.13 have the shortest expected length among all confidence intervals with confidence coefficient $1 - \alpha$; the confidence intervals (7.12) and (7.14) have the shortest expected lengths among all unbiased confidence intervals with confidence coefficient $1 - \alpha$.

7.2.3 Randomized confidence sets

Applications of Theorems 7.4 and 7.5 require that $C(X)$ be obtained by inverting acceptance regions of nonrandomized tests. Thus, these results cannot be directly applied to discrete problems. In fact, in discrete problems inverting acceptance regions of randomized tests may not lead to a confidence set with a given confidence coefficient. Note that randomization is used in hypothesis testing to obtain tests with a given size. Thus, the

same idea can be applied to confidence sets, i.e., we may consider *randomized* confidence sets.

Suppose that we invert acceptance regions of randomized tests T_{θ_0} that reject $H_0 : \theta = \theta_0$ with probability $T_{\theta_0}(x)$ when $X = x$. Let U be a random variable that is independent of X and has the uniform distribution $U(0,1)$. Then the test $\tilde{T}_{\theta_0}(X,U) = I_{(U,1]}(T_{\theta_0})$ has the same power function as T_{θ_0} and is "nonrandomized" if U is viewed as part of the sample. Let

$$A_U(\theta_0) = \{(x,U) : U \geq T_{\theta_0}(x)\}$$

be the acceptance region of $\tilde{T}_{\theta_0}(X,U)$. If T_{θ_0} has size α for all θ_0, then inverting $A_U(\theta)$ we obtain a confidence set

$$C(X,U) = \{\theta : (X,U) \in A_U(\theta)\}$$

having confidence coefficient $1-\alpha$, since

$$P\big(\theta \in C(X,U)\big) = E\big[P\big(U \geq T_\theta(X)|X\big)\big] = E\big[1 - T_\theta(X)\big].$$

If T_{θ_0} is UMP (or UMPU) for each θ_0, then $C(X,U)$ is UMA (or UMAU). However, $C(X,U)$ is a randomized confidence set since it is still random when we observe $X = x$.

When T_{θ_0} is integer-valued, we can use the method in the following example to derive $C(X,U)$.

Example 7.16. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $p = P(X_i = 1)$. The confidence coefficient of $(\underline{p}, 1]$ may not be $1-\alpha$, where $\underline{p}$ is given by (7.5).

From Example 6.2 and the previous discussion, a randomized UMP test for testing $H_0 : p = p_0$ versus $H_1 : p > p_0$ can be constructed based on $Y = \sum_{i=1}^n X_i$ and U, a random variable that is independent of Y and has the uniform distribution $U(0,1)$. Since Y is integer-valued and $U \in (0,1)$, $W = Y + U$ is equivalent to (Y,U). Hence, we can also construct a UMP test based on W. It can be shown (exercise) that W has the following Lebesgue p.d.f.:

$$f_p(w) = \binom{n}{[w]} p^{[w]}(1-p)^{n-[w]} I_{(0,n+1)}(w), \tag{7.17}$$

where $[w]$ is the integer part of w, and that the family $\{f_p : p \in (0,1)\}$ has monotone likelihood ratio in W. It follows from Theorem 6.2 that the test $\tilde{T}_{p_0}(Y,U) = I_{(c(p_0),n+1)}(W)$ is UMP of size α for testing $H_0 : p = p_0$ versus $H_1 : p > p_0$, where $\alpha = \int_{c(p_0)}^{n+1} f_{p_0}(w)dw$. Inverting the acceptance regions of $\tilde{T}_p(Y,U)$, a lower confidence bound $\underline{p}_1$ for p is the solution of

$$\int_{Y+U}^{n+1} \binom{n}{[w]} p^{[w]}(1-p)^{n-[w]} dw = \alpha$$

($\underline{p}_1 = 0$ if $Y = 0$ and $U < \alpha$). This lower confidence bound has confidence coefficient $1 - \alpha$ and is Θ'-UMA with $\Theta' = (0, p)$. ∎

Using a randomized confidence set, we can achieve the purpose of obtaining a confidence set with a given confidence coefficient as well as some optimality properties such as UMA, UMAU, or shortest expected length. On the other hand, randomization may not be desired in practical problems.

7.2.4 Invariant confidence sets

Let $C(X)$ be a confidence set for θ and g be a one-to-one transformation. The invariance principle requires that $C(x)$ change in a specified way when x is transformed to $g(x)$.

Definition 7.4. Let $\mathcal{G}$ be a group of one-to-one transformations of X such that $\mathcal{P}$ is invariant (Definition 6.5). Let $\theta = \theta(P)$ be a parameter with range Θ. Assume that $\tilde{g}(\theta) = \theta(P_{g(X)})$ is well defined for any g, i.e., $\tilde{g}$ is a transformation on Θ induced by g ($\tilde{g} = \bar{g}$ given in Definition 6.5 if $\mathcal{P}$ is indexed by θ).
(i) A confidence set $C(X)$ is *invariant* under $\mathcal{G}$ if and only if $\theta \in C(x)$ is equivalent to $\tilde{g}(\theta) \in C(g(x))$ for every x in the range of X, $\theta \in \Theta$, and $g \in \mathcal{G}$.
(ii) $C(X)$ is Θ'-*uniformly most accurate invariant* (UMAI) with confidence coefficient $1 - \alpha$ if and only if $C(X)$ is invariant with confidence coefficient $1 - \alpha$ and (7.15) holds for any other invariant confidence set $C_1(X)$ with confidence coefficient $1 - \alpha$. $C(X)$ is UMAI if $\Theta' = \{\theta\}^c$. ∎

Example 7.17. Consider the confidence intervals in Example 7.14. Let $\mathcal{G} = \{g_{r,c} : r > 0, c \in \mathcal{R}\}$ with $g_{r,c}(x) = (rx_1 + c, ..., rx_n + c)$. Let $\theta = \mu$. Then $\bar{g}_{r,c}(\mu, \sigma^2) = (r\mu + c, r^2\sigma^2)$ and $\tilde{g}(\mu) = r\mu + c$. Clearly, confidence interval (7.12) is invariant under $\mathcal{G}$.

When σ^2 is known, the family $\mathcal{P}$ is not invariant under $\mathcal{G}$ and we consider $\mathcal{G}_1 = \{g_{1,c} : c \in \mathcal{R}\}$. Then both confidence intervals (7.12) and (7.13) are invariant under $\mathcal{G}_1$.

Suppose now that $\theta = \sigma^2$. For $g_{r,c} \in \mathcal{G}$, $\tilde{g}(\sigma^2) = r^2\sigma^2$. Hence confidence interval (7.14) is invariant under $\mathcal{G}$. ∎

If a confidence set $C(X)$ is UMA and invariant, then it is UMAI. If $C(X)$ is UMAU and invariant, it is not so obvious whether it is UMAI, since a UMAI confidence set (if it exists) is not necessarily unbiased. The following result may be used to construct a UMAI confidence set.

Theorem 7.7. Suppose that for each $\theta_0 \in \Theta$, $A(\theta_0)$ is the acceptance

region of a nonrandomized UMPI test of size α for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \in \Theta_{\theta_0}$ under $\mathcal{G}_{\theta_0}$ and that for any θ_0 and $g \in \mathcal{G}_{\theta_0}$, $\tilde{g}$, the transformation on Θ induced by g, is well defined. If $C(X) = \{\theta : x \in A(\theta)\}$ is invariant under $\mathcal{G}$, the smallest group containing $\cup_{\theta\in\Theta}\mathcal{G}_\theta$, then it is Θ'-UMAI with confidence coefficient $1 - \alpha$, where $\Theta' = \{\theta' : \theta \in \Theta_{\theta'}\}$. ∎

The proofs of Theorem 7.7 and the following result are given as exercises.

Proposition 7.3. Let $\mathcal{P}$ be a parametric family indexed by θ and $\mathcal{G}$ be a group of transformations such that $\tilde{g}$ is well defined by $P_{\tilde{g}(\theta)} = P_{g(X)}$. Suppose that, for any $\theta, \theta' \in \Theta$, there is a $g \in \mathcal{G}$ such that $\tilde{g}(\theta) = \theta'$. Then, for any invariant confidence set $C(X)$, $P(\theta \in C(X))$ is a constant. ∎

Example 7.18. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with unknown μ and σ^2. Consider the problem of setting a lower confidence bound for $\theta = \mu/\sigma$ and $\mathcal{G} = \{g_r : r > 0\}$ with $g_r(x) = rx$. From Example 6.17, a nonrandomized UMPI test of size α for $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ has the acceptance region $A(\theta_0) = \{x : t(x) \le c(\theta_0)\}$, where $t(X) = \sqrt{n}\bar{X}/S$ and $c(\theta)$ is the $(1-\alpha)$th quantile of the noncentral t-distribution $t_{n-1}(\sqrt{n}\theta)$. Applying Theorem 7.7 with $\mathcal{G}_{\theta_0} = \mathcal{G}$ for all θ_0, one can show (exercise) that the solution of $\int_{t(x)}^{\infty} f_\theta(u)du = \alpha$ is a Θ'-UMAI lower confidence bound for θ with confidence coefficient $1 - \alpha$, where f_θ is the Lebesgue p.d.f. of the noncentral t-distribution $t_{n-1}(\sqrt{n}\theta)$ and $\Theta' = (-\infty, \theta)$. ∎

Example 7.19. Consider again the confidence intervals in Example 7.14. In Example 7.17, confidence interval (7.12) is shown to be invariant under $\mathcal{G} = \{g_{r,c} : r > 0, c \in \mathcal{R}\}$ with $g_{r,c}(x) = (rx_1 + c, ..., rx_n + c)$. Although confidence interval (7.12) is UMAU, it is not obvious whether it is UMAI. This interval can be obtained by inverting $A(\mu_0) = \{x : |\bar{X} - \mu_0| \le t_{1-\alpha/2}S/\sqrt{n}\}$, which is the acceptance region of a nonrandomized test UMP among unbiased and invariant tests of size α for $H_0 : \mu = \mu_0$ versus $H_1 : \mu \ne \mu_0$, under $\mathcal{G}_{\mu_0} = \{h_{r,\mu_0} : r > 0\}$ with $h_{r,\mu_0}(x) = (r(x_1 - \mu_0) + \mu_0, ..., r(x_n - \mu_0) + \mu_0)$ (exercise). Note that the testing problem $H_0 : \mu = \mu_0$ versus $H_1 : \mu \ne \mu_0$ is not invariant under $\mathcal{G}$. Since $\mathcal{G}$ is the smallest group containing $\cup_{\mu_0\in\mathcal{R}}\mathcal{G}_{\mu_0}$ (exercise), by Theorem 7.7, interval (7.12) is UMA among unbiased and invariant confidence intervals with confidence coefficient $1 - \alpha$, under $\mathcal{G}$.

Using similar arguments one can show (exercise) that confidence intervals (7.13) and (7.14) are UMA among unbiased and invariant confidence intervals with confidence coefficient $1 - \alpha$, under $\mathcal{G}_1$ (in Example 7.17) and $\mathcal{G}$, respectively. ∎

When UMPI tests are randomized, one can construct randomized UMAI confidence sets, using the techniques introduced in Theorem 7.7 and §7.2.3.

7.3 Asymptotic Confidence Sets

In some problems, especially in nonparametric problems, it is difficult to find a reasonable confidence set with a given confidence coefficient or significance level $1-\alpha$. A common approach is to find a confidence set whose confidence coefficient or significance level is nearly $1-\alpha$ when the sample size n is large. According to Definition 2.4, a confidence set $C(X)$ for θ has asymptotic significance level $1-\alpha$ if $\liminf_n P(\theta \in C(X)) \geq 1-\alpha$ for any $P \in \mathcal{P}$. If $\lim_{n\to\infty} P(\theta \in C(X)) = 1-\alpha$ for any $P \in \mathcal{P}$, then $C(X)$ is a $1-\alpha$ *asymptotically correct* confidence set. Note that asymptotic correctness is not the same as having limiting confidence coefficient $1-\alpha$ (Definition 2.4).

7.3.1 Asymptotically pivotal quantities

A known Borel function of (X,θ), $\Re_n(X,\theta)$, is said to be *asymptotically pivotal* if the limiting distribution of $\Re_n(X,\theta)$ does not depend on P. Like a pivotal quantity in constructing confidence sets (§7.1.1) with a given confidence coefficient or significance level, an asymptotically pivotal quantity can be used in constructing asymptotically correct confidence sets.

Most asymptotically pivotal quantities are of the form $(\hat{\theta}_n - \theta)\hat{V}_n^{-1/2}$, where $\hat{\theta}_n$ is an estimator of θ that is asymptotically normal, i.e.,

$$(\hat{\theta}_n - \theta)V_n^{-1/2} \to_d N_k(0, I_k), \tag{7.18}$$

and $\hat{V}_n$ is an estimator of the asymptotic covariance matrix V_n and is consistent according to Definition 5.4. The resulting $1-\alpha$ asymptotically correct confidence sets are of the form

$$C(X) = \{\theta : \|(\hat{\theta}_n - \theta)\hat{V}_n^{-1/2}\|^2 \leq \chi^2_{k,\alpha}\}, \tag{7.19}$$

where $\chi^2_{k,\alpha}$ is the $(1-\alpha)$th quantile of the chi-square distribution χ^2_k. If θ is real-valued ($k=1$), then $C(X)$ in (7.19) is a $1-\alpha$ asymptotically correct confidence interval. When $k>1$, $C(X)$ in (7.19) is an ellipsoid.

Example 7.20 (Functions of means). Suppose that $X_1, ..., X_n$ are i.i.d. random vectors having a c.d.f. F on $\mathcal{R}^d$ and that the unknown parameter of interest is $\theta = g(\mu)$, where $\mu = E(X_1)$ and g is a known differentiable function from $\mathcal{R}^d$ to $\mathcal{R}^k$, $k \leq d$. From the CLT (Theorem 1.12) and the result in §5.5.1, (7.18) holds with $\hat{\theta}_n = g(\bar{X})$ and $\hat{V}_n$ given by (5.102). Thus, $C(X)$ in (7.19) is a $1-\alpha$ asymptotically correct confidence set for θ. ∎

Example 7.21 (Statistical functionals). Suppose that $X_1, ..., X_n$ are i.i.d. random vectors having a c.d.f. F on $\mathcal{R}^d$ and that the unknown parameter of interest is $\theta = \mathtt{T}(F)$, where $\mathtt{T}$ is a k-vector-valued functional. Let F_n be

the empirical c.d.f. defined by (5.1) and $\hat{\theta}_n = \mathtt{T}(F_n)$. Suppose that each component of $\mathtt{T}$ is ϱ_∞-Hadamard differentiable with an influence function satisfying (5.33) and that the conditions in Theorem 5.15 hold. Then, by Theorems 5.5 and 5.15 and the discussions in §5.2.1, (7.18) holds with $\hat{V}_n$ given by (5.105) and $C(X)$ in (7.19) is a $1-\alpha$ asymptotically correct confidence set for θ. ∎

Example 7.22 (Linear models). Consider linear model (3.25): $X = \beta Z^\tau + \varepsilon$, where ε has i.i.d. components with mean 0 and variance σ^2. Assume that Z is of full rank and that the conditions in Theorem 3.12 hold. It follows from Theorem 1.9(iii) and Theorem 3.12 that (7.18) holds with $\hat{\theta}_n = \hat{\beta}$ and $\hat{V}_n = (n-p)^{-1}SSR(Z^\tau Z)^{-1}$ (see §5.5.1). Thus, a $1-\alpha$ asymptotically correct confidence set for β is

$$C(X) = \{\beta : (\hat{\beta}-\beta)(Z^\tau Z)(\hat{\beta}-\beta)^\tau \le \chi^2_{p,\alpha} SSR/(n-p)\}.$$

Note that this confidence set is different from the one in Example 7.9 derived under the normality assumption on ε. ∎

The problems in the previous three examples are nonparametric. The method of using asymptotically pivotal quantities can also be applied to parametric problems. Note that in a parametric problem where the unknown parameter θ is multivariate, a confidence set for θ with a given confidence coefficient may be difficult or impossible to obtain.

Typically, in a given problem there exist many different asymptotically pivotal quantities that lead to different $1-\alpha$ asymptotically correct confidence sets for θ. Intuitively, if two asymptotic confidence sets are constructed using (7.18) with two different estimators, $\hat{\theta}_{1n}$ and $\hat{\theta}_{2n}$, and if $\hat{\theta}_{1n}$ is asymptotically more efficient than $\hat{\theta}_{2n}$ (§4.5.1), then the confidence set based on $\hat{\theta}_{1n}$ should be better than the one based on $\hat{\theta}_{2n}$ in some sense. This is formally stated in the following result.

Proposition 7.4. Let $C_j(X)$, $j = 1, 2$, be the confidence sets given in (7.19) with $\hat{\theta}_n = \hat{\theta}_{jn}$ and $\hat{V}_n = \hat{V}_{jn}$, $j = 1, 2$, respectively. Suppose that for each j, (7.18) holds for $\hat{\theta}_{jn}$ and $\hat{V}_{jn}$ is consistent for V_{jn}, the asymptotic covariance matrix of $\hat{\theta}_{jn}$. If $\mathrm{Det}(V_{1n}) < \mathrm{Det}(V_{2n})$ for sufficiently large n, where $\mathrm{Det}(A)$ is the determinant of A, then

$$P\big(\mathrm{vol}(C_1(X)) < \mathrm{vol}(C_2(X))\big) \to 1.$$

Proof. The result follows from the consistency of $\hat{V}_{jn}$ and the fact that the volume of the ellipsoid $C(X)$ defined by (7.19) is equal to

$$\mathrm{vol}(C(X)) = \frac{\pi^{k/2}(\chi^2_{p,\alpha})^{k/2}[\mathrm{Det}(\hat{V}_n)]^{1/2}}{\Gamma(1+k/2)}. \quad ∎$$

If $\hat{\theta}_{1n}$ is asymptotically more efficient than $\hat{\theta}_{2n}$ (§4.5.1), then $\mathrm{Det}(V_{1n}) \leq \mathrm{Det}(V_{2n})$. Hence, Proposition 7.4 indicates that a more efficient estimator of θ results in a better confidence set of the form (7.19) in terms of volume. If $\hat{\theta}_n$ is asymptotically efficient (optimal in the sense of having the smallest asymptotic covariance matrix; see Definition 4.4), then the confidence set $C(X)$ in (7.19) is asymptotically optimal (in terms of volume) among the confidence sets of the form (7.19).

Asymptotically correct confidence sets for θ can also be constructed by inverting acceptance regions of asymptotic tests for testing $H_0 : \theta = \theta_0$ versus some H_1. If asymptotic tests are constructed using asymptotically pivotal quantities (see §6.4.2, §6.4.3, and §6.5.4), the resulting confidence sets are almost the same as those based on asymptotically pivotal quantities.

7.3.2 Confidence sets based on likelihoods

As we discussed in §7.3.1, a $1-\alpha$ asymptotically correct confidence set for θ is asymptotically optimal in some sense if it is based on a point estimator $\hat{\theta}_n$ which is asymptotically efficient. In parametric problems, it is shown in §4.5 that MLE's or RLE's of θ are asymptotically efficient. Thus, in this section we study more closely the asymptotic confidence sets based on MLE's and RLE's or, more generally, based on likelihoods.

Consider first the case where $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a parametric family dominated by a σ-finite measure, where $\Theta \subset \mathcal{R}^k$. Let $\ell(\theta)$ be the likelihood function based on the observation $X = x$. The acceptance region of the LR test defined in §6.4.1 with $\Theta_0 = \{\theta_0\}$ is

$$A(\theta_0) = \{x : \ell(\theta_0) \geq e^{-c_\alpha/2}\ell(\hat{\theta})\},$$

where $\ell(\hat{\theta}) = \sup_{\theta\in\Theta} \ell(\theta)$ and c_α is a constant related to the significance level α. Under the conditions of Theorem 6.5, if c_α is chosen to be $\chi^2_{k,\alpha}$, the $(1-\alpha)$th quantile of the chi-square distribution χ^2_k, then

$$C(X) = \{\theta \in \Theta : \ell(\theta) \geq e^{-c_\alpha/2}\ell(\hat{\theta})\} \tag{7.20}$$

is a $1-\alpha$ asymptotically correct confidence set. Note that this confidence set and the one given by (7.19) are generally different.

In many cases $\ell(\theta)$ is a convex function of θ and, therefore, the set defined by (7.20) is a bounded set in $\mathcal{R}^k$; in particular, $C(X)$ in (7.20) is a bounded interval when $k = 1$.

Let $\theta = (\vartheta, \varphi)$, where ϑ is $r \times 1$. Then a $1-\alpha$ asymptotically correct confidence set for ϑ is also given by (7.20) with $\ell(\theta)$ replaced by $\ell(\vartheta, \hat{\varphi}) = \sup_\varphi \ell(\vartheta, \varphi)$ and c_α replaced by $\chi^2_{r,\alpha}$, the $(1-\alpha)$th quantile of the chi-square distribution χ^2_r.

In §6.4.2 we discussed two asymptotic tests closely related to the LR test: Wald's test and Rao's score test. When $\Theta_0 = \{\theta_0\}$, Wald's test has acceptance region

$$A(\theta_0) = \{x : (\hat{\theta} - \theta_0) I_n(\hat{\theta})(\hat{\theta} - \theta_0)^\tau \leq \chi^2_{k,\alpha}\}, \tag{7.21}$$

where $\hat{\theta}$ is an MLE or RLE of θ and $I_n(\theta)$ is the Fisher information matrix based on X. The confidence set obtained by inverting $A(\theta)$ in (7.21) is the same as that in (7.19) with $\hat{V}_n = [I_n(\hat{\theta})]^{-1}$, which is consistent for $[I_n(\theta)]^{-1}$, the asymptotic covariance matrix of $\hat{\theta}$ (e.g., Theorem 4.17 or 4.18).

When $\Theta_0 = \{\theta_0\}$, Rao's score test has the acceptance region

$$A(\theta_0) = \{x : s_n(\theta_0)[I_n(\theta_0)]^{-1}[s_n(\theta_0)]^\tau \leq \chi^2_{k,\alpha}\}, \tag{7.22}$$

where $s_n(\theta) = \partial \log \ell(\theta)/\partial\theta$. The confidence set obtained by inverting $A(\theta)$ in (7.22) is also $1-\alpha$ asymptotically correct, but is generally different from the previous two confidence sets. To illustrate these likelihood-based confidence sets and their differences, we consider the following two examples.

Example 7.23. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $p = P(X_i = 1)$. Since confidence sets for p with a given confidence coefficient are usually randomized (§7.2.3), asymptotically correct confidence sets may be considered when n is large.

The likelihood ratio for testing $H_0 : p = p_0$ versus $H_1 : p \neq p_0$ is

$$\lambda(Y) = p_0^Y (1-p_0)^{n-Y} / \hat{p}^Y (1-\hat{p})^{n-Y},$$

where $Y = \sum_{i=1}^n X_i$ and $\hat{p} = Y/n$ is the MLE of p. The confidence set (7.20) is then equal to

$$C_1(X) = \{p : p^Y (1-p)^{n-Y} \geq e^{-c_\alpha/2} \hat{p}^Y (1-\hat{p})^{n-Y}\}.$$

When $0 < Y < n$, $p^Y(1-p)^{n-Y}$ is strictly convex and equals 0 if $p = 0$ or 1 and, hence, $C_1(X) = [\underline{p}, \overline{p}]$ with $0 < \underline{p} < \overline{p} < 1$. When $Y = 0$, $(1-p)^n$ is strictly decreasing and, therefore, $C_1(X) = (0, \overline{p}]$ with $0 < \overline{p} < 1$. Similarly, when $Y = n$, $C_1(X) = [\underline{p}, 1)$ with $0 < \underline{p} < 1$.

The confidence set obtained by inverting acceptance regions of Wald's tests is simply

$$C_2(X) = [\hat{p} - \Phi^{-1}(1-\alpha/2)\sqrt{\hat{p}(1-\hat{p})/n},\ \hat{p} + \Phi^{-1}(1-\alpha/2)\sqrt{\hat{p}(1-\hat{p})/n}\,],$$

since $I_n(p) = n/[p(1-p)]$ and $(\chi^2_{1,\alpha})^{1/2} = \Phi^{-1}(1-\alpha/2)$. Note that

$$s_n(p) = \frac{Y}{p} - \frac{n-Y}{1-p} = \frac{Y - pn}{p(1-p)}$$

and

$$s_n(p)[I_n(p)]^{-1}[s_n(p)]^\tau = \frac{(Y-pn)^2}{p^2(1-p)^2}\frac{p(1-p)}{n} = \frac{n(\hat{p}-p)^2}{p(1-p)}.$$

Hence, the confidence set obtained by inverting acceptance regions of Rao's score tests is

$$C_3(X) = \{p : n(\hat{p}-p)^2 \le p(1-p)\chi^2_{1,\alpha}\}.$$

It can be shown (exercise) that $C_3(X) = [p_-, p_+]$ with

$$p_\pm = \frac{2Y + \chi^2_{1,\alpha} \pm \sqrt{\chi^2_{1,\alpha}[4n\hat{p}(1-\hat{p}) + \chi^2_{1,\alpha}]}}{2(n+\chi^2_{1,\alpha})}. \quad \blacksquare$$

Example 7.24. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \varphi)$ with unknown $\theta = (\mu, \varphi)$. Consider the problem of constructing a $1-\alpha$ asymptotically correct confidence set for θ. The log-likelihood function is

$$\log \ell(\theta) = -\frac{1}{2\varphi}\sum_{i=1}^n (X_i-\mu)^2 - \frac{n}{2}\log\varphi - \frac{n}{2}\log(2\pi).$$

Since $(\bar{X}, \hat{\varphi})$ is the MLE of θ, where $\hat{\varphi} = (n-1)S^2/n$, the confidence set based on the LR tests is

$$C_1(X) = \left\{\theta : \frac{1}{\varphi}\sum_{i=1}^n (X_i-\mu)^2 + n\log\varphi \le \chi^2_{2,\alpha} + n + n\log\hat{\varphi}\right\}.$$

Note that

$$s_n(\theta) = \left(\frac{n(\bar{X}-\mu)}{\varphi}, \frac{1}{2\varphi^2}\sum_{i=1}^n (X_i-\mu)^2 - \frac{n}{2\varphi}\right)$$

and

$$I_n(\theta) = \begin{pmatrix} \frac{n}{\varphi} & 0 \\ 0 & \frac{n}{2\varphi^2} \end{pmatrix}.$$

Hence, the confidence set based on Wald's tests is

$$C_2(X) = \left\{\theta : \frac{(\bar{X}-\mu)^2}{\hat{\varphi}} + \frac{(\hat{\varphi}-\varphi)^2}{2\hat{\varphi}^2} \le \frac{\chi^2_{2,\alpha}}{n}\right\},$$

which is an ellipsoid in $\mathcal{R}^2$, and the confidence set based on Rao's score tests is

$$C_3(X) = \left\{\theta : \frac{(\bar{X}-\mu)^2}{\varphi} + \frac{1}{2}\left[\frac{1}{n\varphi}\sum_{i=1}^n (X_i-\mu)^2 - 1\right]^2 \le \frac{\chi^2_{2,\alpha}}{n}\right\}.$$

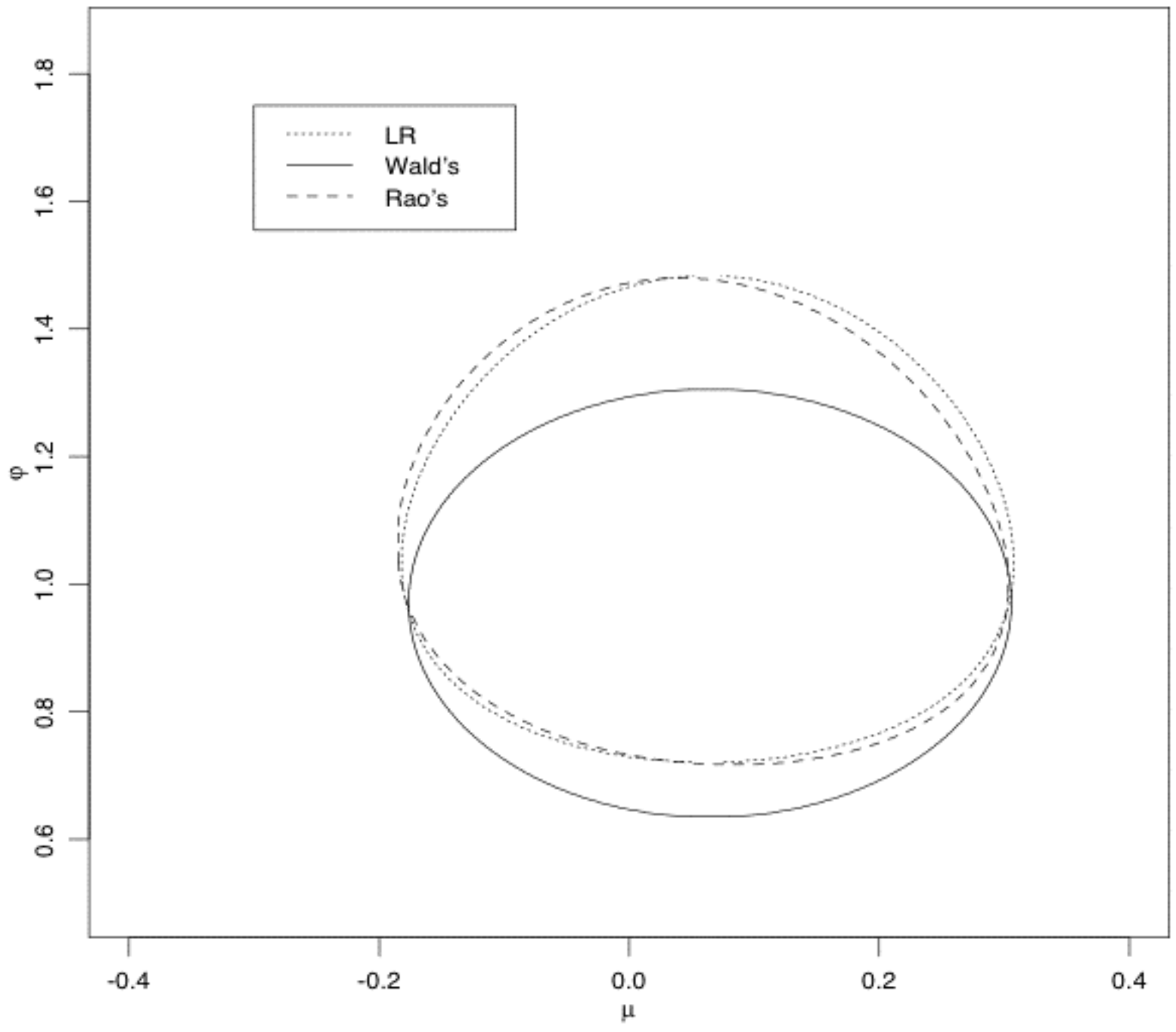

Figure 7.2: Confidence sets obtained by inverting LR, Wald's, and Rao's score tests in Example 7.24

In general, $C_j(X)$, $j = 1, 2, 3$, are different. An example of these three confidence sets is given in Figure 7.2, where $n = 100$, $\mu = 0$, and $\varphi = 1$. ∎

In nonparametric problems, asymptotic confidence sets can be obtained by inverting acceptance regions of empirical likelihood ratio tests or profile empirical likelihood ratio tests (§6.5.3). We consider the following problem as an example. Let $X_1, ..., X_n$ be i.i.d. from F and $\theta = (\vartheta, \varphi)$ be a k-vector of unknown parameters defined by $E[\psi(X_1, \theta)] = 0$, where ψ is a known function. Using the empirical likelihood

$$\ell(G) = \prod_{i=1}^{n} p_i \quad \text{subject to} \quad p_i \geq 0, \ \sum_{i=1}^{n} p_i = 1, \ \sum_{i=1}^{n} p_i \psi(x_i, \theta) = 0,$$

we can obtain a confidence set for ϑ by inverting acceptance regions of the profile empirical likelihood ratio tests based on the ratio $\lambda_n(X)$ in (6.86). This leads to the confidence set defined by

$$C(X) = \left\{ \vartheta : \prod_{i=1}^{n} \frac{1 + \psi(x_i, \hat{\theta})[\xi_n(\hat{\theta})]^\tau}{1 + \psi(x_i, \vartheta, \hat{\varphi})[\zeta_n(\vartheta, \hat{\varphi})]^\tau} \geq e^{-\chi^2_{r,\alpha}/2} \right\}, \tag{7.23}$$

where the notation is the same as that in (6.86) and $\chi^2_{r,\alpha}$ is the $(1-\alpha)$th quantile of the chi-square distribution χ^2_r with $r =$ the dimension of ϑ. By Theorem 6.11, this confidence set is $1-\alpha$ asymptotically correct. Inverting the function of ϑ in (7.23) may be complicated, but $C(X)$ can usually be obtained numerically. More discussions about confidence sets based on empirical likelihoods can be found in Owen (1988, 1990, 1991), Chen and Qin (1993), Qin (1993), and Qin and Lawless (1994).

7.3.3 Results for quantiles

Let $X_1, ..., X_n$ be i.i.d. from a continuous c.d.f. F on $\mathcal{R}$ and let $\theta = F^{-1}(p)$ be the pth quantile of F, $0 < p < 1$. The general methods we previously discussed can be applied to obtain a confidence set for θ, but we introduce here a method that works specially for quantile problems.

In fact, for any given α, it is possible to derive a confidence interval (or bound) for θ with confidence coefficient $1-\alpha$ (Exercise 68), but the numerical computation of such a confidence interval may be cumbersome. We focus on asymptotic confidence intervals for θ.

Our result is based on the following result due to Bahadur (1966). Its proof is omitted.

Theorem 7.8. Let $X_1, ..., X_n$ be i.i.d. from a continuous c.d.f. F on $\mathcal{R}$ which is twice differentiable at $\theta = F^{-1}(p)$, $0 < p < 1$, with $F'(\theta) > 0$. Let $\{k_n\}$ be a sequence of integers satisfying $1 \le k_n \le n$ and $k_n/n = p + o\big((\log n)^\delta/\sqrt{n}\big)$ for some $\delta > 0$. Let F_n be the empirical c.d.f. defined in (5.1). Then

$$X_{(k_n)} = \theta + \frac{(k_n/n) - F_n(\theta)}{F'(\theta)} + O\left(\frac{(\log n)^{(1+\delta)/2}}{n^{3/4}}\right) \quad \text{a.s.} \quad \blacksquare$$

The result in Theorem 7.8 is a refinement of the Bahadur representation in Theorem 5.11. The following corollary of Theorem 7.8 is useful in statistics. Let $\hat{\theta}_n = F_n^{-1}(p)$ be the sample pth quantile.

Corollary 7.1. Assume the conditions in Theorem 7.8 and $k_n/n = p + cn^{-1/2} + o(n^{-1/2})$ with a constant c. Then

$$\sqrt{n}(X_{(k_n)} - \hat{\theta}_n) \to_{a.s.} c/F'(\theta). \quad \blacksquare$$

The proof of Corollary 7.1 is left as an exercise. Using Corollary 7.1, we can obtain a confidence interval for θ with limiting confidence coefficient $1-\alpha$ (Definition 2.14) for any given $\alpha \in (0, \frac{1}{2})$.

Corollary 7.2. Assume the conditions in Theorem 7.8. Let $\{k_{1n}\}$ and $\{k_{2n}\}$ be two sequences of integers satisfying $1 \le k_{1n} < k_{2n} \le n$,

$$k_{1n}/n = p - \Phi^{-1}(1-\alpha/2)\sqrt{p(1-p)/n} + o(n^{-1/2}),$$

and

$$k_{2n}/n = p + \Phi^{-1}(1-\alpha/2)\sqrt{p(1-p)/n} + o(n^{-1/2}).$$

Then the confidence interval $C(X) = [X_{(k_{1n})}, X_{(k_{2n})}]$ has the property that $P(\theta \in C(X))$ does not depend on P and

$$\lim_{n\to\infty} \inf_{P\in\mathcal{P}} P\big(\theta \in C(X)\big) = \lim_{n\to\infty} P\big(\theta \in C(X)\big) = 1-\alpha. \tag{7.24}$$

Furthermore,

$$\text{the length of } C(X) = \frac{2\Phi^{-1}(1-\alpha/2)\sqrt{p(1-p)}}{F'(\theta)\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right) \quad \text{a.s.}$$

Proof. Note that $P(X_{(k_{1n})} \le \theta \le X_{(k_{2n})}) = P(U_{(k_{1n})} \le p \le U_{(k_{2n})})$, where $U_{(k)}$ is the kth order statistic based on a sample $U_1, ..., U_n$ i.i.d. from the uniform distribution $U(0,1)$ (Exercise 68). Hence, $P(\theta \in C(X))$ does not depend on P and the first equality in (7.24) holds.

By Corollary 7.1, Theorem 5.10, and Slutsky's theorem,

$$\begin{aligned}
P(X_{(k_{1n})} > \theta) &= P\left(\hat\theta_n - \Phi^{-1}(1-\alpha/2)\frac{\sqrt{p(1-p)}}{F'(\theta)\sqrt{n}} + o_p(n^{-1/2}) > \theta\right) \\
&= P\left(\frac{\sqrt{n}(\hat\theta_n - \theta)}{\sqrt{p(1-p)}/F'(\theta)} + o_p(1) > \Phi^{-1}(1-\alpha/2)\right) \\
&\to \Phi\big(\Phi^{-1}(1-\alpha/2)\big) \\
&= \alpha/2.
\end{aligned}$$

Similarly,

$$P(X_{(k_{2n})} < \theta) \to \alpha/2.$$

Hence the second equality in (7.24) holds. The result for the length of $C(X)$ follows directly from Corollary 7.1. ∎

The confidence interval $[X_{(k_{1n})}, X_{(k_{2n})}]$ given in Corollary 7.2 is called Woodruff's (1952) interval. It has limiting confidence coefficient $1-\alpha$, a property that is stronger than the $1-\alpha$ asymptotic correctness.

From Theorem 5.10, if $F'(\theta)$ exists and is positive, then

$$\sqrt{n}(\hat\theta_n - \theta) \to_d N\left(0, \tfrac{p(1-p)}{[F'(\theta)]^2}\right).$$

If the derivative $F'(\theta)$ has a consistent estimator $\hat{d}_n$ obtained using some method such as one of those introduced in §5.1.3, then (7.18) holds with $\hat{V}_n = p(1-p)/\hat{d}_n^2$ and the method introduced in §7.3.1 can be applied to derive the following $1-\alpha$ asymptotically correct confidence interval:

$$C(X) = \left[\hat{\theta}_n - \Phi^{-1}(1-\alpha/2)\frac{\sqrt{p(1-p)}}{\hat{d}_n\sqrt{n}},\ \hat{\theta}_n + \Phi^{-1}(1-\alpha/2)\frac{\sqrt{p(1-p)}}{\hat{d}_n\sqrt{n}}\right].$$

The length of $C(X)$ is asymptotically almost the same as Woodruff's interval. However, $C(X)$ depends on the estimated derivative $\hat{d}_n$ and it is usually difficult to obtain a precise estimator $\hat{d}_n$.

7.4 Bootstrap Confidence Sets

In this section we study how to use the bootstrap method introduced in §5.5.3 to construct asymptotically correct confidence sets. There are two main advantages of using the bootstrap method. First, as we can see from previous sections, constructing confidence sets having a given confidence coefficient or being asymptotically correct requires some theoretical derivations. The bootstrap method replaces these derivations by some routine computations. Second, confidence intervals (especially one-sided confidence intervals) constructed using the bootstrap method may be asymptotically more accurate than those asymptotic confidence sets discussed in §7.3.

We use the notation in §5.5.3. Let $X = (X_1, ..., X_n)$ be a sample from P. We focus on the case where X_i's are i.i.d. so that P is specified by a c.d.f. F on $\mathcal{R}^d$, although some results discussed here can be extended to non-i.i.d. cases. Also, we assume that F is estimated by the empirical c.d.f. F_n defined in (5.1) (which means that no assumption is imposed on F and the problem is nonparametric) so that $\hat{P}$ in §5.5.3 is the population corresponding to F_n. A bootstrap sample $X^* = (X_1^*, ..., X_n^*)$ is obtained from $\hat{P}$, i.e., X_i^*'s are i.i.d. from F_n. Some other bootstrap sampling procedures are described in Exercises 72 and 75-77. Let θ be a parameter of interest, $\hat{\theta}_n$ be an estimator of θ, and $\hat{\theta}_n^*$ be the bootstrap analogue of $\hat{\theta}_n$, i.e., $\hat{\theta}_n^*$ is the same as $\hat{\theta}_n$ except that X is replaced by the bootstrap sample X^*.

7.4.1 Construction of bootstrap confidence intervals

We now introduce several different ways of constructing bootstrap confidence intervals for a real-valued θ. Some ideas can be extended to the construction of bootstrap confidence sets for multivariate θ. We mainly consider lower confidence bounds. Upper confidence bounds and two-sided confidence intervals can be similarly obtained.

The bootstrap percentile

Define

$$K_B(x) = P_*(\hat{\theta}_n^* \leq x), \tag{7.25}$$

where P_* denotes the distribution of X^* conditional on X. For a given $\alpha \in (0, \frac{1}{2})$, the *bootstrap percentile* method (Efron, 1981) gives the following lower confidence bound for θ:

$$\underline{\theta}_{BP} = K_B^{-1}(\alpha). \tag{7.26}$$

The name percentile comes from the fact that $K_B^{-1}(\alpha)$ is a percentile of the bootstrap distribution K_B in (7.25).

For most cases, the computation of $\underline{\theta}_{BP}$ requires numerical approximations such as the Monte Carlo approximation described in §5.5.3.

We now provide a justification of the bootstrap percentile method that allows us to see what assumptions are required for a good performance of a bootstrap percentile confidence set. Suppose that there exists an increasing transformation $\phi_n(x)$ such that

$$P\big(\hat{\phi}_n - \phi_n(\theta) \leq x\big) = \Psi(x) \tag{7.27}$$

holds for all possible F (including $F = F_n$), where $\hat{\phi}_n = \phi_n(\hat{\theta}_n)$ and Ψ is a c.d.f. that is continuous, increasing, and symmetric about 0. When $\Psi = \Phi$, the standard normal distribution, the function ϕ_n is called the normalizing and variance stabilizing transformation. If ϕ_n and Ψ in (7.27) can be derived, then the following lower confidence bound for θ has confidence coefficient $1 - \alpha$:

$$\underline{\theta}_E = \phi_n^{-1}(\hat{\phi}_n + z_\alpha),$$

where $z_\alpha = \Psi^{-1}(\alpha)$.

We now show that $\underline{\theta}_{BP} = \underline{\theta}_E$ and, therefore, we can still use this lower confidence bound without deriving ϕ_n and Ψ. Let $w_n = \phi_n(\underline{\theta}_{BP}) - \hat{\phi}_n$. From the fact that assumption (7.27) holds when F is replaced by F_n,

$$\Psi(w_n) = P_*(\hat{\phi}_n^* - \hat{\phi}_n \leq w_n) = P_*(\hat{\theta}_n^* \leq \underline{\theta}_{BP}) = \alpha,$$

where $\hat{\phi}_n^* = \phi_n(\hat{\theta}_n^*)$ and the last equality follows from the definition of $\underline{\theta}_{BP}$ and the assumption on Ψ. Hence $w_n = z_\alpha = \Psi^{-1}(\alpha)$ and

$$\underline{\theta}_{BP} = \phi_n^{-1}(\hat{\phi}_n + z_\alpha) = \underline{\theta}_E.$$

Thus, the bootstrap percentile lower confidence bound $\underline{\theta}_{BP}$ has confidence coefficient $1 - \alpha$ for all n if assumption (7.27) holds exactly for all n. If assumption (7.27) holds approximately for large n, then $\underline{\theta}_{BP}$ is $1 - \alpha$ asymptotically correct (see Theorem 7.9 in §7.4.2) and its performance depends on how good the approximation is.

The bootstrap bias-corrected percentile

Efron (1981) considered the following assumption that is more general than assumption (7.27):

$$P\big(\hat{\phi}_n - \phi_n(\theta) + z_0 \le x\big) = \Psi(x), \tag{7.28}$$

where ϕ_n and Ψ are the same as those in (7.27) and z_0 is a constant that may depend on F and n. When $z_0 = 0$, (7.28) reduces to (7.27). Since $\Psi(0) = \frac{1}{2}$, z_0 is a kind of "bias" of $\hat{\phi}_n$. If ϕ_n, z_0, and Ψ in (7.28) can be derived, then a lower confidence bound for θ with confidence coefficient $1-\alpha$ is

$$\underline{\theta}_E = \phi_n^{-1}(\hat{\phi}_n + z_\alpha + z_0).$$

Applying assumption (7.28) to $F = F_n$, we obtain that

$$K_B(\hat{\theta}_n) = P_*(\hat{\phi}_n^* - \hat{\phi}_n + z_0 \le z_0) = \Psi(z_0),$$

where K_B is given in (7.25). This implies

$$z_0 = \Psi^{-1}\big(K_B(\hat{\theta}_n)\big). \tag{7.29}$$

Also from (7.28),

$$\begin{aligned} 1-\alpha &= \Psi(-z_\alpha) \\ &= \Psi\big(\hat{\phi}_n - \phi_n(\underline{\theta}_E) + z_0\big) \\ &= P_*\big(\hat{\phi}_n^* - \hat{\phi}_n \le \hat{\phi}_n - \phi_n(\underline{\theta}_E)\big) \\ &= P_*\big(\hat{\theta}_n^* \le \phi_n^{-1}(\hat{\phi}_n - z_\alpha - z_0)\big), \end{aligned}$$

which implies

$$\phi_n^{-1}(\hat{\phi}_n - z_\alpha - z_0) = K_B^{-1}(1-\alpha).$$

Since this equation holds for any α, it implies that for $0 < x < 1$,

$$K_B^{-1}(x) = \phi_n^{-1}\big(\hat{\phi}_n + \Psi^{-1}(x) - z_0\big). \tag{7.30}$$

By the definition of $\underline{\theta}_E$ and (7.30),

$$\underline{\theta}_E = K_B^{-1}\big(\Psi(z_\alpha + 2z_0)\big).$$

Assuming that Ψ is known (e.g., $\Psi = \Phi$) and using (7.29), Efron (1981) obtained the bootstrap bias-corrected (BC) percentile lower confidence bound for θ:

$$\underline{\theta}_{BC} = K_B^{-1}\big(\Psi\big(z_\alpha + 2\Psi^{-1}(K_B(\hat{\theta}_n))\big)\big), \tag{7.31}$$

which is a percentile of the bootstrap distribution K_B. Note that $\underline{\theta}_{BC}$ reduces to $\underline{\theta}_{BP}$ if $K_B(\hat{\theta}_n) = \frac{1}{2}$, i.e., $\hat{\theta}_n$ is the median of the bootstrap

distribution K_B. Hence, the bootstrap BC percentile method is a bias-corrected version of the bootstrap percentile method and the bias-correction is represented by $2\Psi^{-1}(K_B(\hat{\theta}_n))$. If (7.28) holds exactly, then $\underline{\theta}_{BC}$ has confidence coefficient $1-\alpha$ for all n. If (7.28) holds approximately, then $\underline{\theta}_{BC}$ is $1-\alpha$ asymptotically correct.

The bootstrap BC percentile method improves the bootstrap percentile method by taking a bias into account. This is supported by the theoretical result in §7.4.2. However, there are still many cases where assumption (7.28) cannot be fulfilled nicely and the bootstrap BC percentile method does not work well. Efron (1987) proposed a bootstrap accelerated bias-corrected (BC_a) percentile method (see Exercise 73) that improves the bootstrap BC percentile method. However, applications of the bootstrap BC_a percentile method involve some derivations that may be very complicated. See Efron (1987) and Efron and Tibshirani (1993) for details.

The hybrid bootstrap

Suppose that $\hat{\theta}_n$ is asymptotically normal, i.e., (7.18) holds with $V_n = \sigma_F^2/n$. Let H_n be the c.d.f. of $\sqrt{n}(\hat{\theta}_n - \theta)$ and

$$\hat{H}_B(x) = P_*\big(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \le x\big)$$

be its bootstrap estimator defined in (5.116). From the results in Theorem 5.20, for any $t \in (0,1)$, $\hat{H}_B^{-1}(t) - H_n^{-1}(t) \to_p 0$. Treating the quantile of $\hat{H}_B$ as the quantile of H_n, we obtain the following *hybrid bootstrap* lower confidence bound for θ:

$$\underline{\theta}_{HB} = \hat{\theta}_n - n^{-1/2}\hat{H}_B^{-1}(1-\alpha). \tag{7.32}$$

The bootstrap-t

Suppose that (7.18) holds with $V_n = \sigma_F^2/n$ and $\hat{\sigma}_F^2$ is a consistent estimator of σ_F^2. The bootstrap-t method is based on $t(X,\theta) = \sqrt{n}(\hat{\theta}_n - \theta)/\hat{\sigma}_F$, which is often called a studentized "statistic". If the distribution G_n of $t(X,\theta)$ is known (i.e., $t(X,\theta)$ is pivotal), then a confidence interval for θ with confidence coefficient $1-\alpha$ can be obtained (§7.1.1). If G_n is unknown, it can be estimated by the bootstrap estimator

$$\hat{G}_B(x) = P_*\big(t(X^*, \hat{\theta}_n) \le x\big),$$

where $t(X^*, \hat{\theta}_n) = \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}_F^*$ and $\hat{\sigma}_F^*$ is the bootstrap analogue of $\hat{\sigma}_F$. Treating the quantile of $\hat{G}_B$ as the quantile of G_n, we obtain the following *bootstrap-t* lower confidence bound for θ:

$$\underline{\theta}_{BT} = \hat{\theta}_n - n^{-1/2}\hat{\sigma}_F\hat{G}_B^{-1}(1-\alpha). \tag{7.33}$$

Although it is shown in §7.4.2 that $\underline{\theta}_{BT}$ in (7.33) is more accurate than $\underline{\theta}_{BP}$ in (7.26), $\underline{\theta}_{BC}$ in (7.31), and $\underline{\theta}_{HB}$ in (7.32), the use of the bootstrap-t method requires a consistent variance estimator $\hat{\sigma}_F^2$.

7.4.2 Asymptotic correctness and accuracy

From the construction of the hybrid bootstrap and bootstrap-t confidence bounds, $\underline{\theta}_{HB}$ is $1-\alpha$ asymptotically correct if $\varrho_\infty(\hat{H}_B, H_n) \to_p 0$, and $\underline{\theta}_{BT}$ is $1-\alpha$ asymptotically correct if $\varrho_\infty(\hat{G}_B, G_n) \to_p 0$. On the other hand, the asymptotic correctness of the bootstrap percentile (with or without bias-correction or acceleration) confidence bounds requires slightly more.

Theorem 7.9. Suppose that $\varrho_\infty(\hat{H}_B, H_n) \to_p 0$ and

$$\lim_{n\to\infty} \rho_\infty(H_n, H) = 0, \tag{7.34}$$

where H is a c.d.f. on $\mathcal{R}$ that is continuous, strictly increasing, and symmetric about 0. Then $\underline{\theta}_{BP}$ in (7.26) and $\underline{\theta}_{BC}$ in (7.31) are $1-\alpha$ asymptotically correct.

Proof. The result for $\underline{\theta}_{BP}$ follows from

$$\begin{aligned}
P\big(\underline{\theta}_{BP} \le \theta\big) &= P\big(\alpha \le K_B(\theta)\big) \\
&= P\big(\alpha \le H_B\big(\sqrt{n}(\theta - \hat{\theta}_n)\big)\big) \\
&= P\big(\sqrt{n}(\hat{\theta}_n - \theta) \le -\hat{H}_B^{-1}(\alpha)\big) \\
&= P\big(\sqrt{n}(\hat{\theta}_n - \theta) \le -H^{-1}(\alpha)\big) + o(1) \\
&= H\big(-H^{-1}(\alpha)\big) + o(1) \\
&= 1 - \alpha + o(1).
\end{aligned}$$

The result for $\underline{\theta}_{BC}$ follows from the previous result and

$$z_0 = \Psi^{-1}\big(K_B(\hat{\theta}_n)\big) = \Psi^{-1}\big(\hat{H}_B(0)\big) \to_p \Psi^{-1}\big(H(0)\big) = 0. \quad \blacksquare$$

Theorem 7.9 can be obviously extended to the case of upper confidence bounds or two-sided confidence intervals. The result also holds for the bootstrap BC_a percentile confidence intervals.

Note that H in (7.34) is not the same as Ψ in assumption (7.28). Usually $H(x) = \Phi(x/\sigma_F)$ for some $\sigma_F > 0$, whereas $\Psi = \Phi$. Also, condition (7.34) is much weaker than assumption (7.28), since the latter requires variance stabilizing.

It is not surprising that all bootstrap methods introduced in §7.3.1 produce asymptotically correct confidence sets. To compare various bootstrap

confidence intervals and other asymptotic confidence intervals, we now consider the convergence rate of their coverage probability.

Definition 7.5. A confidence set $C(X)$ for θ is said to be lth-order (asymptotically) accurate if and only if $P(\theta \in C(X)) = 1 - \alpha + O(n^{-l/2})$, where l is a positive integer. ∎

We now discuss the results in Hall (1988) and sketch the proofs. Consider the case of $\theta = g(\mu)$, $\mu = EX_1$, and $\hat{\theta}_n = g(\bar{X})$, where g is continuously differentiable from $\mathcal{R}^d$ to $\mathcal{R}$ with $\nabla g(\mu) \neq 0$. The asymptotic variance of $\sqrt{n}(\hat{\theta}_n - \theta)$ is $\sigma_F^2 = \nabla g(\mu)\mathrm{Var}(X_1)[\nabla g(\mu)]^\tau$ and a consistent estimator of σ_F^2 is $\hat{\sigma}_F^2 = \frac{n-1}{n}\nabla g(\bar{X})S^2[\nabla g(\bar{X})]^\tau$, where S^2 is the sample covariance matrix. Let G_n be the distribution of $\sqrt{n}(\hat{\theta}_n - \theta)/\hat{\sigma}_F$. If G_n is known, then a lower confidence bound for θ with confidence coefficient $1 - \alpha$ is:

$$\underline{\theta}_E = \hat{\theta}_n - n^{-1/2}\hat{\sigma}_F G_n^{-1}(1 - \alpha), \tag{7.35}$$

which is not useful if G_n is unknown.

Suppose that $G_n^{-1}(t)$ has the following Cornish-Fisher expansion on a compact interval:

$$G_n^{-1}(t) = z_t + \frac{q_1(z_t, F)}{\sqrt{n}} + \frac{q_2(z_t, F)}{n} + o\left(\frac{1}{n}\right), \tag{7.36}$$

which is a type of an inverse of the following Edgeworth expansion:

$$G_n(x) = \Phi(x) - \left[\frac{q_1(x, F)}{\sqrt{n}} + \frac{q_3(x, F)}{n}\right]\Phi'(x) + o\left(\frac{1}{n}\right), \tag{7.37}$$

where Φ and Φ' are, respectively, the c.d.f. and p.d.f. of the standard normal distribution, q_j's are some functions depending on F, and $z_t = \Phi^{-1}(t)$. More details about these expansions can be found, for example, in Hall (1992). Let $\hat{G}_B$ be the bootstrap estimator of G_n defined in §7.4.1. Under some conditions (Hall, 1992), $\hat{G}_B^{-1}$ admits expansion (7.36) with F replaced by F_n for almost all sequences $X_1, X_2, \ldots$. Hence the bootstrap-t lower confidence bound in (7.33) can be written as

$$\underline{\theta}_{BT} = \hat{\theta}_n - \frac{\hat{\sigma}_F}{\sqrt{n}}\left[z_{1-\alpha} + \sum_{j=1}^{2}\frac{q_j(z_{1-\alpha}, F_n)}{n^{j/2}} + o\left(\frac{1}{n}\right)\right] \text{ a.s.} \tag{7.38}$$

Under some moment conditions, $q_j(x, F_n) - q_j(x, F) = O_p(n^{-1/2})$ for each x, $j = 1, 2$. Then, comparing (7.35), (7.36), and (7.38), we obtain that

$$\underline{\theta}_{BT} - \underline{\theta}_E = O_p(n^{-3/2}). \tag{7.39}$$

Furthermore,

$$\begin{aligned}
P(\underline{\theta}_{BT} \le \theta) &= P\left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_F/\sqrt{n}} \le \hat{G}_B^{-1}(1-\alpha)\right) \\
&= P\left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_F/\sqrt{n}} \le z_{1-\alpha} + \sum_{j=1}^{2} \frac{q_j(z_{1-\alpha}, F_n)}{n^{j/2}}\right) + o\left(\frac{1}{n}\right) \\
&= P\left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_F/\sqrt{n}} \le z_{1-\alpha} + \sum_{j=1}^{2} \frac{q_j(z_{1-\alpha}, F)}{n^{j/2}}\right) \\
&\quad + \frac{\psi(z_{1-\alpha})}{n} + o\left(\frac{1}{n}\right) \\
&= 1 - \alpha + \frac{\psi(z_{1-\alpha})\Phi'(z_{1-\alpha})}{n} + o\left(\frac{1}{n}\right), \qquad (7.40)
\end{aligned}$$

where $\psi(x)$ is a polynomial whose coefficients are functions of moments of F and the last two equalities can be justified by a somewhat complicated argument (Hall, 1992).

Result (7.40) implies that $\underline{\theta}_{BT}$ is second-order accurate according to Definition 7.5. The same can be concluded for the bootstrap-t upper confidence bound and the two-sided bootstrap-t confidence interval for θ.

Next, we consider the hybrid bootstrap lower confidence bound $\underline{\theta}_{HB}$ given by (7.32). Let $\tilde{H}_B$ be the bootstrap estimator of $\tilde{H}_n$, the distribution of $\sqrt{n}(\hat{\theta}_n - \theta)/\sigma_F$. Then $\hat{H}_B^{-1}(1-\alpha) = \hat{\sigma}_F \tilde{H}_B^{-1}(1-\alpha)$ and

$$\underline{\theta}_{HB} = \hat{\theta}_n - n^{-1/2}\hat{\sigma}_F \tilde{H}_B^{-1}(1-\alpha),$$

which can be viewed as a bootstrap approximation to

$$\underline{\theta}_{H} = \hat{\theta}_n - n^{-1/2}\hat{\sigma}_F \tilde{H}_n^{-1}(1-\alpha).$$

Note that $\underline{\theta}_H$ does not have confidence coefficient $1-\alpha$, since it is obtained by muddling up $G_n^{-1}(1-\alpha)$ and $\tilde{H}_n^{-1}(1-\alpha)$. Similar to G_n^{-1}, under some conditions $\tilde{H}_n^{-1}$ admits a Cornish-Fisher expansion

$$\tilde{H}_n^{-1}(t) = z_t + \frac{\tilde{q}_1(z_t, F)}{\sqrt{n}} + \frac{\tilde{q}_2(z_t, F)}{n} + o\left(\frac{1}{n}\right) \qquad (7.41)$$

and $\tilde{H}_B^{-1}$ admits the same expansion (7.41) with F replaced by F_n for almost all $X_1, X_2, \ldots$. Then

$$\underline{\theta}_{HB} = \hat{\theta}_n - \frac{\hat{\sigma}_F}{\sqrt{n}}\left[z_{1-\alpha} + \sum_{j=1}^{2} \frac{\tilde{q}_j(z_{1-\alpha}, F_n)}{n^{j/2}} + o\left(\frac{1}{n}\right)\right] \text{ a.s.} \qquad (7.42)$$

and, by (7.35),

$$\underline{\theta}_{HB} - \underline{\theta}_E = O_p(n^{-1}), \tag{7.43}$$

since $q_1(x, F)$ and $\tilde{q}_1(x, F)$ are usually different. Results (7.39) and (7.43) imply that $\underline{\theta}_{HB}$ is not as close to $\underline{\theta}_E$ as $\underline{\theta}_{BT}$. Similar to (7.40), we can show that (Hall, 1992)

$$\begin{aligned}
P(\underline{\theta}_{HB} \le \theta) &= P\left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_F/\sqrt{n}} \le \tilde{H}_B^{-1}(1-\alpha)\right) \\
&= P\left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_F/\sqrt{n}} \le z_{1-\alpha} + \sum_{j=1}^{2} \frac{\tilde{q}_j(z_{1-\alpha}, F_n)}{n^{j/2}}\right) + o\left(\frac{1}{n}\right) \\
&= P\left(\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_F/\sqrt{n}} \le z_{1-\alpha} + \sum_{j=1}^{2} \frac{\tilde{q}_j(z_{1-\alpha}, F)}{n^{j/2}}\right) + O\left(\frac{1}{n}\right) \\
&= 1 - \alpha + \frac{\tilde{\psi}(z_{1-\alpha})\Phi'(z_{1-\alpha})}{\sqrt{n}} + O\left(\frac{1}{n}\right),
\end{aligned} \tag{7.44}$$

where $\tilde{\psi}(x)$ is an even polynomial. This implies that when $\tilde{\psi} \not\equiv 0$, $\underline{\theta}_{HB}$ is only first-order accurate according to Definition 7.5.

The same conclusion can be drawn for the hybrid bootstrap upper confidence bounds. However, the two-sided hybrid bootstrap confidence interval

$$[\,\underline{\theta}_{HB}, \overline{\theta}_{HB}\,] = [\,\hat{\theta}_n - n^{-1/2}\hat{H}_B^{-1}(1-\alpha),\, \hat{\theta}_n - n^{-1/2}\hat{H}_B^{-1}(\alpha)\,]$$

is second-order accurate, as is the two-sided bootstrap-t confidence interval, since

$$\begin{aligned}
P(\underline{\theta}_{HB} \le \theta \le \overline{\theta}_{HB}) &= P(\theta \le \overline{\theta}_{HB}) - P(\theta < \underline{\theta}_{HB}) \\
&= 1 - \alpha + n^{-1/2}\tilde{\psi}(z_{1-\alpha})\Phi'(z_{1-\alpha}) \\
&\quad - \alpha - n^{-1/2}\tilde{\psi}(z_\alpha)\Phi'(z_\alpha) + O(n^{-1}) \\
&= 1 - 2\alpha + O(n^{-1})
\end{aligned}$$

by the fact that $\tilde{\psi}$ and Φ' are even functions and $z_{1-\alpha} = -z_\alpha$.

For the bootstrap percentile lower confidence bound in (7.26),

$$\underline{\theta}_{BP} = K_B^{-1}(\alpha) = \hat{\theta}_n + n^{-1/2}\hat{H}_B^{-1}(\alpha).$$

Comparing $\underline{\theta}_{BP}$ with $\underline{\theta}_{BT}$ and $\underline{\theta}_{HB}$, we find that the bootstrap percentile method muddles up not only $\tilde{H}_B^{-1}(1-\alpha)$ and $\hat{G}_B^{-1}(1-\alpha)$, but also $\hat{H}_B^{-1}(\alpha)$ and $-\hat{H}_B^{-1}(1-\alpha)$. If $\hat{H}_B$ is asymptotically symmetric about 0, then the bootstrap percentile method is equivalent to the hybrid bootstrap method and, therefore, one-sided bootstrap percentile confidence intervals are only

first-order accurate and the two-sided bootstrap percentile confidence interval is second-order accurate.

Since $\hat{\theta}_n$ is asymptotically normal, we can use $\Psi = \Phi$ for the bootstrap BC percentile method. Let $\tilde{\alpha}_n = \Phi(z_\alpha + 2z_0)$. Then the bootstrap BC percentile lower confidence bound given by (7.31) is just the $\tilde{\alpha}_n$th quantile of K_B in (7.25). Using the Edgeworth expansion, we obtain that

$$K_B(\hat{\theta}_n) = \tilde{H}_B(0) = \Phi(0) + \frac{\tilde{q}(0, F_n)\Phi'(0)}{\sqrt{n}} + O_p\left(\frac{1}{n}\right)$$

with some function $\tilde{q}$ and, therefore,

$$\tilde{\alpha}_n = \alpha + \frac{2\tilde{q}(0, F_n)\Phi'(z_\alpha)}{\sqrt{n}} + O_p\left(\frac{1}{n}\right).$$

This result and the Cornish-Fisher expansion for $\tilde{H}_B^{-1}$ imply

$$\tilde{H}_B^{-1}(\tilde{\alpha}_n) = z_\alpha + \frac{2\tilde{q}(0, F_n) + \tilde{q}_1(z_\alpha, F_n)}{\sqrt{n}} + O_p\left(\frac{1}{n}\right).$$

Then from (7.31) and $K_B^{-1}(\tilde{\alpha}_n) = \hat{\theta}_n + n^{-1/2}\hat{\sigma}_F\tilde{H}_B^{-1}(\tilde{\alpha}_n)$,

$$\underline{\theta}_{BC} = \hat{\theta}_n + \frac{\hat{\sigma}_F}{\sqrt{n}}\left[z_\alpha + \frac{2\tilde{q}(0, F_n) + \tilde{q}_1(z_\alpha, F_n)}{\sqrt{n}} + O_p\left(\frac{1}{n}\right)\right]. \tag{7.45}$$

Comparing (7.35) with (7.45), we conclude that

$$\underline{\theta}_{BC} - \underline{\theta}_E = O_p(n^{-1}).$$

It also follows from (7.45) that

$$P(\underline{\theta}_{BC} \le \theta) = 1 - \alpha + \frac{\bar{\psi}(z_{1-\alpha})\Phi'(z_{1-\alpha})}{\sqrt{n}} + O\left(\frac{1}{n}\right) \tag{7.46}$$

with an even polynomial $\bar{\psi}(x)$. Hence $\underline{\theta}_{BC}$ is first-order accurate in general. In fact,

$$\underline{\theta}_{BC} - \underline{\theta}_{BP} = 2\tilde{q}(0, F_n)\hat{\sigma}_F n^{-1} + O_p(n^{-3/2})$$

and, therefore, the bootstrap BC percentile and the bootstrap percentile confidence intervals have the same order of accuracy. The bootstrap BC percentile method, however, is a partial improvement over the bootstrap percentile method in the sense that the absolute value of $\bar{\psi}(z_{1-\alpha})$ in (7.46) is smaller than that of $\tilde{\psi}(z_{1-\alpha})$ in (7.44) (see Example 7.25).

While the bootstrap BC percentile method does not improve the bootstrap percentile method in terms of accuracy order, Hall (1988) showed that the bootstrap BC$_a$ percentile method in Efron (1987) produces second-order

accurate one-sided and two-sided confidence intervals and that (7.39) holds with $\underline{\theta}_{BT}$ replaced by the bootstrap BC_a percentile lower confidence bound.

We have considered the order of asymptotic accuracy for all bootstrap confidence intervals introduced in §7.4.1. In summary, all two-sided confidence intervals are second-order accurate; the one-sided bootstrap-t and bootstrap BC_a percentile confidence intervals are second-order accurate whereas the one-sided bootstrap percentile, bootstrap BC percentile, and hybrid bootstrap confidence intervals are first-order accurate; however, the latter three are simpler than the former two.

What is the order of accuracy of the asymptotic confidence intervals derived in §7.3.1? The lower confidence bound derived in §7.3.1 is

$$\underline{\theta}_N = \hat{\theta}_n - n^{-1/2}\hat{\sigma}_F z_{1-\alpha}. \tag{7.47}$$

From (7.42), we obtain that

$$\underline{\theta}_N - \underline{\theta}_{HB} = O_p(n^{-1}).$$

Hence, confidence intervals obtained using the method in §7.3.1 have the same order of accuracy as the hybrid bootstrap confidence intervals.

Example 7.25. Suppose that $d = 1$ and $g(x) = x$. Under some conditions (see Hall (1992)) expansions (7.36) and (7.37) hold with $q_1(x, F) = -\gamma(2x^2 + 1)/6$, $q_2(x, F) = x[(x^2 + 3)/4 - \kappa(x^2 - 3)/12 + 5\gamma^2(4x^2 - 1)/72]$, and $q_3(x, F) = x[(x^2 + 3)/4 + \gamma^2(x^4 + 2x^2 - 3)/18 - \kappa(x^2 - 3)/12]$, where $\gamma = E(X_1 - \mu)^3/\sigma^3$ (skewness), $\kappa = E(X_1 - \mu)^4/\sigma^4 - 3$ (kurtosis), and $\sigma^2 = \mathrm{Var}(X_1)$. Also, expansion (7.41) holds with $\tilde{q}_1(x, F) = \gamma(x^2 - 1)/6$ and $\tilde{q}_2(x, F) = x[\kappa(x^2 - 3)/24 - \gamma^2(2x^2 - 5)/36]$.

The function ψ in (7.40) is equal to $x(1+2x^2)(\kappa-3\gamma^2/2)/6$; the function $\tilde{\psi}$ in (7.44) is equal to $\gamma x^2/2$; and the function $\bar{\psi}(x)$ in (7.46) is equal to $\gamma(x^2 + 2)/6$ (see Liu and Singh (1987)). If $\gamma \neq 0$, then $\underline{\theta}_{HB}$, $\underline{\theta}_{BC}$, and $\underline{\theta}_N$ are first-order accurate. In this example, we can still compare their relative performances in terms of the convergence speed of the coverage probability. Let

$$e(\underline{\theta}) = P(\underline{\theta} \leq \theta) - (1 - \alpha)$$

be the error in coverage probability for the lower confidence bound $\underline{\theta}$. It can be shown (exercise) that

$$|e(\underline{\theta}_{HB})| = |e(\underline{\theta}_N)| + C_n(z_\alpha, F) + o(n^{-1/2}) \tag{7.48}$$

and

$$|e(\underline{\theta}_N)| = |e(\underline{\theta}_{BC})| + C_n(z_\alpha, F) + o(n^{-1/2}), \tag{7.49}$$

where $C_n(x, F) = |\gamma|(x^2 - 1)\Phi'(x)/(6\sqrt{n})$. Assume $\gamma \neq 0$. When $z_\alpha^2 > 1$, which is usually the case in practice, $C_n(z_\alpha, F) > 0$ and, therefore, $\underline{\theta}_{BC}$ is

better than $\underline{\theta}_N$, which is better than $\underline{\theta}_{HB}$. The use of $\underline{\theta}_N$ requires a variance estimator $\hat{\sigma}_F^2$, which is not required by the bootstrap BC percentile and hybrid bootstrap methods. When a variance estimator is available, we can use the bootstrap-t lower confidence bound, which is second-order accurate even when $\gamma \neq 0$. ∎

7.4.3 High-order accurate bootstrap confidence sets

Can we obtain a confidence interval with a higher order accuracy than the second-order accuracy? The answer is affirmative under some conditions. Suppose that we construct a confidence set for θ (not necessarily real-valued) based on an asymptotically pivotal quantity $\Re_n$ that admits an Edgeworth expansion

$$P(\Re_n \leq x) = \Phi(x) + \sum_{j=1}^{l} \frac{\eta_j(x)\Phi'(x)}{n^{j/2}} + O\left(\frac{1}{n^{(l+1)/2}}\right),$$

where η_j are some polynomials whose coefficients depend on F. Then there exist polynomials ζ_j such that

$$P\left(\Re_n \leq x - \sum_{j=1}^{l} \frac{\zeta_j(x)}{n^{j/2}}\right) = \Phi(x) + O\left(\frac{1}{n^{(l+1)/2}}\right) \tag{7.50}$$

(Hall, 1992). Let $\hat{\zeta}_j$ be consistent estimators of ζ_j. Then (7.50) still holds with ζ_j replaced by $\hat{\zeta}_j$ and we can obtain an $(l+1)$th-order accurate confidence set

$$C_{EDG}^{(l)}(X) = \left\{\theta : \ \Re_n \leq z_{1-\alpha} - \sum_{j=1}^{l} \frac{\hat{\zeta}_j(z_{1-\alpha})}{n^{j/2}}\right\}. \tag{7.51}$$

To use $C_{EDG}^{(l)}(X)$, we need to derive the polynomials ζ_j's, which is complicated. The bootstrap method can be used to obtain confidence sets as accurate as $C_{EDG}^{(l)}(X)$ without requiring the theoretical derivation of ζ_j's but requiring some extra extensive computations.

The bootstrap prepivoting and bootstrap inverting

The hybrid bootstrap and the bootstrap-t are based on the bootstrap distribution estimators for $\sqrt{n}(\hat{\theta}_n - \theta)$ and $\sqrt{n}(\hat{\theta}_n - \theta)/\hat{\sigma}_F$, respectively. Beran (1987) argued that the reason why the bootstrap-t is better than the hybrid bootstrap is that $\sqrt{n}(\hat{\theta}_n - \theta)/\hat{\sigma}_F$ is more pivotal than $\sqrt{n}(\hat{\theta}_n - \theta)$ in the

sense that the distribution of $\sqrt{n}(\hat{\theta}_n - \theta)/\hat{\sigma}_F$ is less dependent on the unknown F. The bootstrap-t method, however, requires a variance estimator $\hat{\sigma}_F^2$. Beran (1987) suggested the following method called *bootstrap prepivoting*. Let $\Re_n^{(0)}$ be a random function (such as $\sqrt{n}(\hat{\theta}_n - \theta)$ or $\sqrt{n}(\hat{\theta}_n - \theta)/\hat{\sigma}_F$) used to construct a confidence set for $\theta \in \mathcal{R}^k$, $H_n^{(0)}$ be the distribution of $\Re_n^{(0)}$, and let $\hat{H}_B^{(0)}$ be the bootstrap estimator of $H_n^{(0)}$. Define

$$\Re_n^{(1)} = \hat{H}_B^{(0)}(\Re_n^{(0)}). \tag{7.52}$$

If $H_n^{(0)}$ is continuous and if we replace $\hat{H}_B^{(0)}$ in (7.52) by $H_n^{(0)}$, then $\Re_n^{(1)}$ has the uniform distribution $U(0,1)$. Hence, it is expected that $\Re_n^{(1)}$ is more pivotal than $\Re_n^{(0)}$. Let $\hat{H}_B^{(1)}$ be the bootstrap estimator of $H_n^{(1)}$, the distribution of $\Re_n^{(1)}$. Then $\Re_n^{(2)} = \hat{H}_B^{(1)}(\Re_n^{(1)})$ is more pivotal than $\Re_n^{(1)}$. In general, let $H_n^{(j)}$ be the distribution of $\Re_n^{(j)}$ and $\hat{H}_B^{(j)}$ be the bootstrap estimator of $H_n^{(j)}$, $j = 0, 1, 2, ...$ Then we can use the following confidence sets for θ:

$$C_{PREB}^{(j)}(X) = \{\theta : \Re_n^{(j)} \leq (\hat{H}_B^{(j)})^{-1}(1-\alpha)\}, \quad j = 0, 1, 2, ... \tag{7.53}$$

Note that for each j, $C_{PREB}^{(j)}(X)$ is a hybrid bootstrap confidence set based on $\Re_n^{(j)}$. Since $\Re_n^{(j+1)}$ is more pivotal than $\Re_n^{(j)}$, we obtain a sequence of confidence sets with increasing accuracies. 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 $C_{PREB}^{(1)}(X)$ based on $\Re_n^{(0)} = \sqrt{n}(\hat{\theta}_n - \theta)$ is second-order accurate, and the two-sided confidence interval $C_{PREB}^{(1)}(X)$ based on $\Re_n^{(0)} = \sqrt{n}|\hat{\theta}_n - \theta|$ is third-order accurate. Hence, bootstrap prepivoting with one iteration improves the hybrid bootstrap method. It is expected that the one-sided confidence interval $C_{PREB}^{(2)}(X)$ based on $\Re_n^{(0)} = \sqrt{n}(\hat{\theta}_n - \theta)$ is third-order accurate, i.e., it is better than the one-sided bootstrap-t or bootstrap BC_a percentile confidence interval. More detailed discussion can be found in Beran (1987).

It seems that, using this iterative method, we can start with a $\Re_n^{(0)}$ and obtain a bootstrap confidence set that is as accurate as we want it to be. However, more computations are required for higher stage bootstrapping and, therefore, the practical implementation of this method is very hard or even impossible, with current computational ability. We explain this with the computation of $C_{PREB}^{(1)}(X)$ based on $\Re_n^{(0)} = \Re(X, F)$. Suppose that we use the Monte Carlo approximation. Let $\{X_{1b}^*, ..., X_{nb}^*\}$ be i.i.d. samples from F_n, $b = 1, ..., B_1$. Then $\hat{H}_B^{(0)}$ is approximated by $\hat{H}_B^{(0,B_1)}$, the empirical distribution of $\{\Re_{nb}^{(0)*} : b = 1, ..., B_1\}$, where $\Re_{nb}^{(0)*} = \Re(X_{1b}^*, ..., X_{nb}^*, F_n)$. For each b, let F_{nb}^* be the empirical distribution of

$X_{1b}^*, ..., X_{nb}^*$, $\{X_{1bj}^{**}, ..., X_{nbj}^{**}\}$ be i.i.d. samples from F_{nb}^*, $j = 1, ..., B_2$, H_b^* be the empirical c.d.f. of $\{\Re_n(X_{1bj}^{**}, ..., X_{nbj}^{**}, F_{nb}^*), j = 1, ..., B_2\}$, and $z_b^* = H_b^*(\Re_{nb}^{(0)*})$. Then $\hat{H}_B^{(1)}$ can be approximated by $\hat{H}_B^{(1,B_1B_2)}$, the empirical distribution of $\{z_b^*, b = 1, ..., B_1\}$, and the confidence set $C_{PREB}^{(1)}(X)$ can be approximated by

$$\big\{\theta : \Re(X, F) \leq (\hat{H}_B^{(0,B_1)})^{-1}\big((\hat{H}_B^{(1,B_1B_2)})^{-1}(1-\alpha)\big)\big\}.$$

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_1B_2, which is why this method is also called the double bootstrap. If each stage requires 1,000 bootstrap replicates, then the total number of bootstrap replicates is 1,000,000! Similarly, to compute $C_{PREB}^{(j)}(X)$ we need $(1,000)^{j+1}$ bootstrap replicates, $j = 2, 3, ...$, which limits the application of the bootstrap prepivoting method.

A very similar method, *bootstrap inverting*, is given in Hall (1992). Instead of using (7.53), we define

$$C_{INVB}^{(j)}(X) = \{\theta : \Re_n^{(j)} \leq (\hat{H}_B^{(j)})^{-1}(1-\alpha)\}, \quad j = 0, 1, 2, \ldots,$$

where

$$\Re_n^{(j)} = \Re_n^{(j-1)} - (\hat{H}_B^{(j-1)})^{-1}(1-\alpha), \quad j = 1, 2, \ldots,$$

and $\hat{H}_B^{(j)}$ is the bootstrap estimator of the distribution of $\Re_n^{(j)}$. For each $j \geq 1$, $C_{INVB}^{(j)}(X)$, $C_{PREB}^{(j)}(X)$ in (7.53), and $C_{EDG}^{(j)}(X)$ in (7.51) have the same order of accuracy. Thus, the analytic derivation of ζ_j's in the Edgeworth expansion (7.50) can be accomplished by bootstrap prepivoting or bootstrap inverting. Bootstrap prepivoting and bootstrap inverting require the same amount of computation and both of them are special cases of a general iterative bootstrap introduced by Hall and Martin (1988).

Bootstrap calibrating

Suppose that we want a confidence set $C(X)$ with confidence coefficient $1-\alpha$, which is called the *nominal level*. The basic idea of *bootstrap calibrating* is to improve $C(X)$ by adjusting its nominal level. Let π_n be the actual coverage probability of $C(X)$. The value of π_n can be estimated by a bootstrap estimator $\hat{\pi}_n$. If we find that $\hat{\pi}_n$ is far from $1-\alpha$, then we construct a confidence set $C_1(X)$ with nominal level $1-\tilde{\alpha}$ so that the coverage probability of $C_1(X)$ is closer to $1-\alpha$ than π_n. Bootstrap calibrating can be used iteratively as follows: estimate the true coverage probability of $C_1(X)$; if the difference between $1-\alpha$ and the estimated coverage probability of $C_1(X)$ is still large, we can adjust the nominal level again and construct a new calibrated confidence set $C_2(X)$.

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 (1987, 1991) in the case where the initial confidence sets are obtained by using the method in §7.3.1. Consider first the lower confidence bound $\underline{\theta}_N$ defined in (7.47). The coverage probability $\pi_n = P(\underline{\theta}_N \le \theta)$ can be estimated by the bootstrap estimator (approximated by Monte Carlo if necessary)

$$\hat{\pi}_n = \hat{G}_B(z_{1-\alpha}) = P_*\big(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}_F^* \le z_{1-\alpha}\big).$$

When the bootstrap distribution has an Edgeworth expansion, we have

$$\hat{\pi}_n = 1 - \alpha + \left[\frac{q_1(z_{1-\alpha}, F_n)}{\sqrt{n}} + \frac{q_2(z_{1-\alpha}, F_n)}{n}\right]\Phi'(z_{1-\alpha}) + O_p\left(\frac{1}{n^{3/2}}\right).$$

Let h be any increasing, unbounded, and twice differentiable function on the interval $(0,1)$ and

$$\tilde{\alpha} = 1 - h^{-1}\big(h(1-\alpha) - \delta\big),$$

where

$$\begin{aligned}\delta &= h(\hat{\pi}_n) - h(1-\alpha)\\ &= \left[\frac{q_1(z_{1-\alpha}, F_n)}{\sqrt{n}} + \frac{q_2(z_{1-\alpha}, F_n)}{n}\right]\Phi'(z_{1-\alpha})h'(1-\alpha)\\ &\quad + \frac{[q_1(z_{1-\alpha}, F_n)\Phi'(z_{1-\alpha})]^2}{2n}h''(1-\alpha) + O_p\left(\frac{1}{n^{3/2}}\right). \qquad (7.54)\end{aligned}$$

The bootstrap calibration lower confidence bound is

$$\underline{\theta}_{CLB} = \hat{\theta}_n - n^{-1/2}\hat{\sigma}_F z_{1-\tilde{\alpha}}.$$

By (7.54),

$$1 - \tilde{\alpha} = 1 - \alpha + \frac{q_1(z_{1-\alpha}, F_n)\Phi'(z_{1-\alpha})}{\sqrt{n}} + O_p\left(\frac{1}{n}\right) \qquad (7.55)$$

and

$$z_{1-\tilde{\alpha}} = z_{1-\alpha} + \frac{q_1(z_{1-\alpha}, F_n)}{\sqrt{n}} + O_p\left(\frac{1}{n}\right) \qquad (7.56)$$

(exercise). Thus,

$$\underline{\theta}_{CLB} = \hat{\theta}_n - \frac{\hat{\sigma}_F}{\sqrt{n}}\left[z_{1-\alpha} + \frac{q_1(z_{1-\alpha}, F_n)}{\sqrt{n}} + O_p\left(\frac{1}{n}\right)\right]. \qquad (7.57)$$

Comparing (7.57) with (7.38), we find that

$$\underline{\theta}_{CLB} - \underline{\theta}_{BT} = O_p(n^{-3/2}).$$

Thus, $\underline{\theta}_{CLB}$ is second-order accurate.

We can take $[\,\underline{\theta}_{CLB}, \overline{\theta}_{CLB}\,]$ as a two-sided confidence interval; it is still second-order accurate. By calibrating directly the equal-tail two-sided confidence interval

$$[\,\underline{\theta}_N, \overline{\theta}_N\,] = [\,\hat{\theta}_n - n^{-1/2}\hat{\sigma}_F z_{1-\alpha}, \hat{\theta}_n + n^{-1/2}\hat{\sigma}_F z_{1-\alpha}\,], \tag{7.58}$$

we can obtain a higher order accurate confidence interval. Let $\hat{\pi}_n$ be the bootstrap estimator of the coverage probability $P\{\underline{\theta}_N \le \theta \le \overline{\theta}_N\}$, $\delta = h(\hat{\pi}_n) - h(1-2\alpha)$, and $\tilde{\alpha} = [1 - h^{-1}\big(h(1-2\alpha) - \delta\big)]/2$. Then the two-sided bootstrap calibration confidence interval is the interval given by (7.58) with α replaced by $\tilde{\alpha}$. Loh (1991) showed that this confidence interval is fourth-order accurate. The length of this interval exceeds the length of the interval in (7.58) by an amount of order $O_p(n^{-3/2})$.

7.5 Simultaneous Confidence Intervals

So far we have studied confidence sets for a real-valued θ or a vector-valued θ with a finite dimension k. In some applications we need a confidence set for real-valued θ_t with $t \in \mathcal{T}$, where $\mathcal{T}$ is an index set that may contain infinitely many elements, for example, $\mathcal{T} = [0,1]$ or $\mathcal{T} = \mathcal{R}$.

Definition 7.6. Let X be a sample from $P \in \mathcal{P}$, θ_t, $t \in \mathcal{T}$, be real-valued parameters related to P, and let $C_t(X)$, $t \in \mathcal{T}$, be a class of (one-sided or two-sided) confidence intervals.
(i) If

$$\inf_{P\in\mathcal{P}} P\big(\theta_t \in C_t(X) \text{ for all } t \in \mathcal{T}\big) \ge 1-\alpha, \tag{7.59}$$

then $C_t(X)$, $t \in \mathcal{T}$, are level $1-\alpha$ *simultaneous confidence intervals* for θ_t, $t \in \mathcal{T}$. The left-hand side of (7.59) is called the confidence coefficient of $C_t(X)$, $t \in \mathcal{T}$.
(ii) If

$$\lim_{n\to\infty} P\big(\theta_t \in C_t(X) \text{ for all } t \in \mathcal{T}\big) \ge 1-\alpha, \tag{7.60}$$

then $C_t(X)$, $t \in \mathcal{T}$, are simultaneous confidence intervals for θ_t, $t \in \mathcal{T}$, with asymptotic significance level $1-\alpha$, and they are $1-\alpha$ asymptotically correct if the equality in (7.60) holds. ∎

If the index set $\mathcal{T}$ contains $k < \infty$ elements, then $\theta = (\theta_t, t \in \mathcal{T})$ is a k-vector and the methods studied in the previous sections can be applied to construct a level $1-\alpha$ confidence set $C(X)$ for θ. If $C(X)$ can be expressed as $\prod_{t\in\mathcal{T}} C_t(X)$ for some intervals $C_t(X)$, then $C_t(X)$, $t \in \mathcal{T}$, are level $1-\alpha$ simultaneous confidence intervals. This simple method, however,

does not always work. In this section we introduce some other commonly used methods for constructing simultaneous confidence intervals.

7.5.1 Bonferroni's method

Bonferroni's method, which works when $\mathcal{T}$ contains $k < \infty$ elements, is based on the following simple inequality for k events $A_1, ..., A_k$:

$$P\left(\bigcup_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} P(A_i) \tag{7.61}$$

(see Proposition 1.1). For each $t \in \mathcal{T}$, let $C_t(X)$ be a level $1-\alpha_t$ confidence interval for θ_t. If α_t's are chosen so that $\sum_{t\in\mathcal{T}} \alpha_t = \alpha$ (e.g., $\alpha_t = \alpha/k$ for all t), then Bonferroni's simultaneous confidence intervals are $C_t(X)$, $t \in \mathcal{T}$. It can be shown (exercise) that Bonferroni's intervals are of level $1-\alpha$, but they are not of confidence coefficient $1-\alpha$ even if all $C_t(X)$'s have confidence coefficient $1-\alpha_t$. Note that Bonferroni's method does not require that $C_t(X)$ be independent.

Example 7.26 (Multiple comparison in one-way ANOVA models). Consider the one-way ANOVA model in Example 6.18. If the hypothesis H_0 in (6.53) is rejected, one typically would like to compare μ_i's. One way to compare μ_i's is to consider simultaneous confidence intervals for $\mu_i - \mu_j$, $1 \leq i < j \leq m$. Since X_{ij}'s are independently normal, the sample means $\bar{X}_{i\cdot}$ are independently normal $N(\mu_i, \sigma^2/n_i)$, $i = 1, ..., m$, respectively, and they are independent of $SSR = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2$. Consequently, $(\bar{X}_{i\cdot} - \bar{X}_{j\cdot})/\sqrt{v_{ij}}$ has the t-distribution t_{n-m}, $1 \leq i < j \leq m$, where $v_{ij} = (n_i^{-1} + n_j^{-1})SSR/(n-m)$. For each (i,j), a confidence interval for $\mu_i - \mu_j$ with confidence coefficient $1-\alpha$ is

$$C_{ij,\alpha}(X) = [\,\bar{X}_{i\cdot} - \bar{X}_{j\cdot} - t_{n-m,\alpha/2}\sqrt{v_{ij}},\ \bar{X}_{i\cdot} - \bar{X}_{j\cdot} + t_{n-m,\alpha/2}\sqrt{v_{ij}}\,], \tag{7.62}$$

where $t_{n-m,\alpha}$ is the $(1-\alpha)$th quantile of the t-distribution t_{n-m}. One can show that $C_{ij,\alpha}(X)$ is actually UMAU (exercise). Bonferroni's level $1-\alpha$ simultaneous confidence intervals for $\mu_i - \mu_j$, $1 \leq i < j \leq m$, are $C_{ij,\alpha_*}(X)$, $1 \leq i < j \leq m$, where $\alpha_* = 2\alpha/[m(m-1)]$. When m is large, these confidence intervals are very conservative in the sense that the confidence coefficient of these intervals may be much larger than the given level $1-\alpha$ and these intervals may be too wide to be useful.

If the normality assumption is removed, then $C_{ij,\alpha}(X)$ is $1-\alpha$ asymptotically correct as $\min(n_1, ..., n_m) \to \infty$ and $\max(n_1, ..., n_m)/\min(n_1, ..., n_m) \to c < \infty$. Therefore, $C_{ij,\alpha_*}(X)$, $1 \leq i < j \leq m$, are simultaneous confidence intervals with asymptotic significance level $1-\alpha$.

One can establish similar results for the two-way balanced ANOVA models in Example 6.19 (exercise). ∎

7.5.2 Scheffé's method in linear models

Since multiple comparison in ANOVA models (or more generally, linear models) is one of the most important applications of simultaneous confidence intervals, we now introduce Scheffé's method for problems in linear models. Consider the normal linear model

$$X = N_n(\beta Z^\tau, \sigma^2 I_n), \tag{7.63}$$

where β is a p-vector of unknown parameters, $\sigma^2 > 0$ is unknown, and Z is an $n \times p$ known matrix of rank $r \le p$. Let L be an $s \times p$ matrix of rank $s \le r$. Suppose that $\mathcal{R}(L) \subset \mathcal{R}(Z)$ and we would like to construct simultaneous confidence intervals for $\beta L^\tau t^\tau$, where $t \in \mathcal{T} = \mathcal{R}^s - \{0\}$.

Let $\hat\beta$ be the LSE of β. Using the argument in Example 7.15, for each $t \in \mathcal{T}$, we can obtain the following confidence interval for $\beta L^\tau t^\tau$ with confidence coefficient $1-\alpha$:

$$\big[\, \hat\beta L^\tau t^\tau - t_{n-r,\alpha/2}\hat\sigma\sqrt{tDt^\tau},\ \hat\beta L^\tau t^\tau + t_{n-r,\alpha/2}\hat\sigma\sqrt{tDt^\tau}\,\big],$$

where $\hat\sigma^2 = \|X - \hat\beta Z^\tau\|^2/(n-r)$, $D = L(Z^\tau Z)^- L^\tau$, and $t_{n-r,\alpha}$ is the $(1-\alpha)$th quantile of the t-distribution t_{n-r}. However, these intervals are not level $1-\alpha$ simultaneous confidence intervals for $\beta L^\tau t^\tau$, $t \in \mathcal{T}$.

Scheffé's (1959) method of constructing simultaneous confidence intervals for $\beta L^\tau t^\tau$ is based on the following equality (exercise)

$$xA^{-1}x^\tau = \max_{y\in\mathcal{R}^k, y\neq 0} \frac{(xy^\tau)^2}{yAy^\tau}, \tag{7.64}$$

where $x \in \mathcal{R}^k$ and A is a $k \times k$ positive definite matrix.

Theorem 7.10. Assume normal linear model (7.63). Let L be an $s \times p$ matrix of rank $s \le r$. Assume that $\mathcal{R}(L) \subset \mathcal{R}(Z)$ and $D = L(Z^\tau Z)^- L^\tau$ is of full rank. Then

$$C_t(X) = \big[\, \hat\beta L^\tau t^\tau - \hat\sigma\sqrt{sc_\alpha tDt^\tau},\ \hat\beta L^\tau t^\tau + \hat\sigma\sqrt{sc_\alpha tDt^\tau}\,\big], \quad t \in \mathcal{T},$$

are simultaneous confidence intervals for $\beta L^\tau t^\tau$, $t \in \mathcal{T}$, with confidence coefficient $1-\alpha$, where $\hat\sigma^2 = \|X - \hat\beta Z^\tau\|^2/(n-r)$, $\mathcal{T} = \mathcal{R}^s - \{0\}$, and c_α is the $(1-\alpha)$th quantile of the F-distribution $F_{s,n-r}$.

Proof. Note that $\beta L^\tau t^\tau \in C_t(X)$ for all $t \in \mathcal{T}$ is equivalent to

$$\frac{(\hat\beta L^\tau - \beta L^\tau)D^{-1}(\hat\beta L^\tau - \beta L^\tau)^\tau}{sc_\alpha\hat\sigma^2} = \max_{t\in\mathcal{T}} \frac{(\hat\beta L^\tau t^\tau - \beta L^\tau t^\tau)^2}{sc_\alpha\hat\sigma^2 tDt^\tau} \le 1. \tag{7.65}$$

Then the result follows from the fact that c_α times the quantity on the left-hand side of (7.65) has the F-distribution $F_{s,n-r}$. ∎

If the normality assumption is removed but conditions in Theorem 3.12 are assumed, then Scheffé's intervals in Theorem 7.10 are $1-\alpha$ asymptotically correct (exercise).

The choice of the matrix L depends on the purpose of the analysis. One particular choice is $L = Z$, in which case $\beta L^\tau t^\tau$ is the mean of Xt^τ. When Z is of full rank, we can choose $L = I_p$, in which case $\{\beta L^\tau t^\tau : t \in \mathcal{T}\}$ is the class of all linear functions of β. Another L commonly used when Z is of full rank is the following $(p-1)\times p$ matrix:

$$L = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 1 & 0 & \cdots & 0 & -1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}. \tag{7.66}$$

It can be shown (exercise) that when L is given by (7.66),

$$\left\{\beta L^\tau t^\tau : t \in \mathcal{R}^{p-1} - \{0\}\right\} = \left\{\beta c^\tau : c \in \mathcal{R}^p - \{0\}, cJ^\tau = 0\right\}, \tag{7.67}$$

where J is the p-vector of ones. Functions βc^τ satisfying $cJ^\tau = 0$ are called *contrasts*. Therefore, setting simultaneous confidence intervals for $\beta L^\tau t^\tau$, $t \in \mathcal{T}$, with L given by (7.66) is the same as setting simultaneous confidence intervals for all nonzero contrasts.

Although Scheffé's intervals have confidence coefficient $1-\alpha$, they are too conservative if we are only interested in $\beta L^\tau t^\tau$ for t in a subset of $\mathcal{T}$. In a one-way ANOVA model (Example 7.26), for instance, multiple comparison can be carried out using Scheffé's intervals with $\beta = (\mu_1, ..., \mu_m)$, L given by (7.66), and $t \in \mathcal{T}_0$ which contains exactly $m(m-1)/2$ vectors (Exercise 90). The resulting Scheffé's intervals are (Exercise 90)

$$[\,\bar{X}_{i\cdot} - \bar{X}_{j\cdot} - \sqrt{sc_\alpha v_{ij}},\ \bar{X}_{i\cdot} - \bar{X}_{j\cdot} + \sqrt{sc_\alpha v_{ij}}\,], \quad t \in \mathcal{T}_0, \tag{7.68}$$

where $\bar{X}_{i\cdot}$ and v_{ij} are given in (7.62). Since $\mathcal{T}_0$ contains a much smaller number of elements than $\mathcal{T}$, the simultaneous confidence intervals in (7.68) are very conservative. In fact, they are often more conservative than Bonferroni's intervals derived in Example 7.26 (see Example 7.28). In the following example, however, Scheffé's intervals have confidence coefficient $1-\alpha$, although we consider $t \in \mathcal{T}_0 \subset \mathcal{T}$.

Example 7.27 (Simple linear regression). Consider the special case of model (7.63) where

$$X_i = N(\beta_0 + \beta_1 t_i, \sigma^2), \qquad i = 1, ..., n,$$

and $t_i \in \mathcal{R}$ satisfying $S_{tt} = \sum_{i=1}^n (t_i - \bar{t})^2 > 0$, $\bar{t} = n^{-1}\sum_{i=1}^n t_i$. In this case, we are usually interested in simultaneous confidence intervals for the regression function $\beta_0 + \beta_1 x$, $x \in \mathcal{R}$. Note that the result in Theorem 7.10 (with $L = I_2$) can be applied to obtain simultaneous confidence intervals for $\beta_0 y + \beta_1 x$, $(y, x) \in \mathcal{T} = \mathcal{R}^2 - \{0\}$. If we let $y \equiv 1$, Scheffé's intervals in Theorem 7.10 are

$$\big[\hat{\beta}_0 + \hat{\beta}_1 x - \hat{\sigma}\sqrt{2c_\alpha D(x)},\, \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\sigma}\sqrt{2c_\alpha D(x)}\big], \quad x \in \mathcal{R} \tag{7.69}$$

(exercise), where $D(x) = n^{-1} + (x - \bar{t})^2/S_{tt}$. Unless

$$\max_{x\in\mathcal{R}} \frac{(\hat{\beta}_0 + \hat{\beta}_1 x - \beta_0 - \beta_1 x)^2}{D(x)} = \max_{(y,x)\in\mathcal{T}} \frac{(\hat{\beta}_0 y + \hat{\beta}_1 x - \beta_0 y - \beta_1 x)^2}{(y,x)(Z^\tau Z)^{-1}(y,x)^\tau} \tag{7.70}$$

holds with probability 1, the confidence coefficient of the intervals in (7.69) is larger than $1-\alpha$. We now show that (7.70) actually holds with probability 1 so that the intervals in (7.69) have confidence coefficient $1 - \alpha$. First,

$$P\big(n(\hat{\beta}_0 - \beta_0) + n(\hat{\beta}_1 - \beta_1)\bar{t} \neq 0\big) = 1.$$

Second, it can be shown (exercise) that the maximum on the right-hand side of (7.70) is achieved at

$$(y, x) = \frac{(\hat{\beta}_0 - \beta_0, \hat{\beta}_1 - \beta_1)(Z^\tau Z)}{n(\hat{\beta}_0 - \beta_0) + n(\hat{\beta}_1 - \beta_1)\bar{t}}. \tag{7.71}$$

Finally, (7.70) holds since y in (7.71) is equal to 1 (exercise). ∎

7.5.3 Tukey's method in one-way ANOVA models

Consider the one-way ANOVA model in Example 6.18 (and Example 7.26). Note that both Bonferroni's and Scheffé's simultaneous confidence intervals for $\mu_i - \mu_j$, $1 \le i < j \le m$, are not of confidence coefficient $1 - \alpha$ and often too conservative. Tukey's method introduced next produces simultaneous confidence intervals for all nonzero contrasts (including the differences $\mu_i - \mu_j$, $1 \le i < j \le m$) with confidence coefficient $1 - \alpha$.

Let $\hat{\sigma}^2 = SSR/(n - m)$, where SSR is given in Example 7.26. The *studentized range* is defined to be

$$R_{st} = \max_{1\le i<j\le m} \frac{|(\bar{X}_{i\cdot} - \mu_i) - (\bar{X}_{j\cdot} - \mu_j)|}{\hat{\sigma}}. \tag{7.72}$$

Note that the distribution of R_{st} does not depend on any unknown parameter (exercise).

Theorem 7.11. Assume the one-way ANOVA model in Example 6.18. Let q_α be the $(1-\alpha)$th quantile of R_{st} in (7.72). Then Tukey's intervals

$$[\,\hat{\beta}c^\tau - q_\alpha\hat{\sigma}c_+,\ \hat{\beta}c^\tau + q_\alpha\hat{\sigma}c_+\,], \qquad c \in \mathcal{R}^m - \{0\}, cJ^\tau = 0,$$

are simultaneous confidence intervals for βc^τ, $c \in \mathcal{R}^m - \{0\}$, $cJ^\tau = 0$, with confidence coefficient $1-\alpha$, where c_+ is the sum of all positive components of c, $\beta = (\mu_1, ..., \mu_m)$, $\hat{\beta} = (\bar{X}_{1\cdot}, ..., \bar{X}_{m\cdot})$, and J is the m-vector of ones.

Proof. Let $Y_i = (\bar{X}_{i\cdot} - \mu_i)/\hat{\sigma}$ and $Y = (Y_1, ..., Y_m)$. Then the result follows if we can show that

$$\max_{1\le i<j\le m} |Y_i - Y_j| \le q_\alpha \tag{7.73}$$

is equivalent to

$$|Yc^\tau| \le q_\alpha c_+ \qquad \text{for all } c \in \mathcal{R}^m \text{ satisfying } cJ^\tau = 0, c \neq 0. \tag{7.74}$$

Let $c(i,j) = (c_1, ..., c_m)$ with $c_i = 1$, $c_j = -1$, and $c_l = 0$ for $l \neq i$ or $l \neq j$. Then $c(i,j)_+ = 1$ and $|Y[c(i,j)]^\tau| = |Y_i - Y_j|$ and, therefore, (7.74) implies (7.73).

Let $c = (c_1, ..., c_m)$ be a vector satisfying the conditions in (7.74). Define $-c_-$ to be the sum of negative components of c. Then

$$\begin{aligned}
|Yc^\tau| &= \frac{1}{c_+}\Big|c_+ \sum_{j:c_j<0} c_jY_j + c_- \sum_{i:c_i>0} c_iY_i\Big| \\
&= \frac{1}{c_+}\Big|\sum_{i:c_i>0}\sum_{j:c_j<0} c_ic_jY_j - \sum_{j:c_j<0}\sum_{i:c_i>0} c_ic_jY_i\Big| \\
&= \frac{1}{c_+}\Big|\sum_{i:c_i>0}\sum_{j:c_j<0} c_ic_j(Y_j - Y_i)\Big| \\
&\le \frac{1}{c_+}\sum_{i:c_i>0}\sum_{j:c_j<0} |c_ic_j||Y_j - Y_i| \\
&\le \max_{1\le i<j\le m}|Y_j - Y_i|\left(\frac{1}{c_+}\sum_{i:c_i>0}\sum_{j:c_j<0}|c_i||c_j|\right) \\
&= \max_{1\le i<j\le m}|Y_j - Y_i|c_+,
\end{aligned}$$

where the first and the last equalities follow from the fact that $c_- = c_+ \neq 0$. Hence (7.73) implies (7.74). ∎

Tukey's method works well when n_i's are all equal to n_0, in which case values of $\sqrt{n_0}q_\alpha$ can be found using tables or statistical software. When n_i's are unequal, some modifications are suggested; see Tukey (1977) and Milliken and Johnson (1992).

Example 7.28. We compare the t-type confidence intervals in (7.62), Bonferroni's, Scheffé's, and Tukey's simultaneous confidence intervals for $\mu_i - \mu_j$, $1 \le i < j \le 3$, based on the following data X_{ij} given in Mendenhall and Sincich (1995):

	$j=1$	2	3	4	5	6	7	8	9	10
$i=1$	148	76	393	520	236	134	55	166	415	153
2	513	264	433	94	535	327	214	135	280	304
3	335	643	216	536	128	723	258	380	594	465

In this example, $m = 3$, $n_i \equiv n_0 = 10$, $\bar{X}_{1\cdot} = 229.6$, $\bar{X}_{2\cdot} = 309.8$, $\bar{X}_{3\cdot} = 427.8$, and $\hat{\sigma} = 168.95$. Let $\alpha = 0.05$. For the t-type intervals in (7.62), $t_{0.975} = 2.05$. For Bonferroni's method, $\alpha_* = \alpha/3 = 0.017$ and $t_{0.983} = 2.55$. For Scheffé's method, $c_{0.05} = 3.35$ and $\sqrt{2c_{0.05}} = 2.59$. From Table 13 in Mendenhall and Sincich (1995, Appendix II), $\sqrt{n_0}q_{0.05} = 3.49$. The resulting confidence intervals are given as follows.

Method	Parameter $\mu_1 - \mu_2$	$\mu_1 - \mu_3$	$\mu_2 - \mu_3$	Length
t-type	$[-235.2, 74.6]$	$[-353.1, -43.3]$	$[-272.8, 37.0]$	309.8
Bonferroni	$[-273.0, 112.4]$	$[-390.9, -5.5]$	$[-310.6, 74.8]$	385.4
Scheffé	$[-276.0, 115.4]$	$[-393.9, -2.5]$	$[-313.6, 77.8]$	391.4
Tukey	$[-267.3, 106.7]$	$[-385.2, -11.2]$	$[-304.9, 69.1]$	374.0

Apperently, t-type intervals have the shortest length, but they are not simultaneous confidence intervals. Tukey's intervals in this example have the shortest length among simultaneous confidence intervals. Scheffé's intervals have the longest length. ∎

7.5.4 Confidence bands for c.d.f.'s

Let $X_1, ..., X_n$ be i.i.d. from a continuous c.d.f. F on $\mathcal{R}$. Consider the problem of setting simultaneous confidence intervals for $F(t)$, $t \in \mathcal{R}$. A class of simultaneous confidence intervals indexed by $t \in \mathcal{R}$ is called a confidence band. For example, the class of intervals in (7.69) is a confidence band with confidence coefficient $1 - \alpha$.

First, consider the case where F is in a parametric family, i.e., $F = F_\theta$, $\theta \in \Theta \subset \mathcal{R}^k$. If θ is real-valued and $F_\theta(t)$ is nonincreasing in θ for every t (e.g., when the parametric family has monotone likelihood ratio; see Lemma 6.3) and if $[\underline{\theta}, \overline{\theta}]$ is a confidence interval for θ with confidence coefficient (or significance level) $1 - \alpha$, then

$$[F_{\overline{\theta}}(t), F_{\underline{\theta}}(t)], \qquad t \in \mathcal{R},$$

are simultaneous confidence intervals for $F(t)$, $t \in \mathcal{R}$, with confidence coefficient (or significance level) $1-\alpha$. One-sided simultaneous confidence intervals can be similarly obtained.

When $F = F_\theta$ with a multivariate θ, there is no simple and general way of constructing a confidence band for $F(t)$, $t \in \mathcal{R}$. We consider an example.

Example 7.29. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$. Note that $F(t) = \Phi\left(\frac{t-\mu}{\sigma}\right)$. If μ is unknown and σ^2 is known, then, from the results in Example 7.14, a confidence band for $F(t)$, $t \in \mathcal{R}$, with confidence coefficient $1-\alpha$ is

$$\left[\Phi\left(\frac{t-\bar{X}}{\sigma} - \frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{n}}\right), \Phi\left(\frac{t-\bar{X}}{\sigma} + \frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{n}}\right)\right], \quad t \in \mathcal{R}.$$

A confidence band can be similarly obtained if σ^2 is unknown and μ is known.

Suppose now that both μ and σ^2 are unknown. In Example 7.18, we discussed how to obtain a lower confidence bound $\underline{\theta}$ for $\theta = \mu/\sigma$. An upper confidence bound $\overline{\theta}$ for θ can be similarly obtained. Suppose that both $\underline{\theta}$ and $\overline{\theta}$ have confidence coefficient $1-\alpha/4$. Using inequality (7.61), we can obtain the following level $1-\alpha$ confidence band for $F(t)$, $t \in \mathcal{R}$:

$$\left[\Phi\left(\frac{a_{n,\alpha}t}{S} - \overline{\theta}\right), \Phi\left(\frac{b_{n,\alpha}t}{S} - \underline{\theta}\right)\right], \quad t \in \mathcal{R},$$

where $a_{n,\alpha} = [\chi^2_{n-1,1-\alpha/4}/(n-1)]^{1/2}$, $b_{n,\alpha} = [\chi^2_{n-1,\alpha/4}/(n-1)]^{1/2}$, and $\chi^2_{n-1,\alpha}$ is the $(1-\alpha)$th quantile of the chi-square distribution χ^2_{n-1}. ∎

Consider now the case where F is in a nonparametric family. Let $D_n(F) = \sup_{t\in\mathcal{R}} |F_n(t) - F(t)|$, which is related to the Kolmogorov-Smirnov test statistics introduced in §6.5.2, where F_n is the empirical c.d.f. given by (5.1). From Theorem 6.10(i), there exists a c_α such that

$$P\big(D_n(F) \le c_\alpha\big) = 1-\alpha. \tag{7.75}$$

Then a confidence band for $F(t)$, $t \in \mathcal{R}$, with confidence coefficient $1-\alpha$ is given by

$$[\,F_n(t) - c_\alpha,\, F_n(t) + c_\alpha\,] \qquad t \in \mathcal{R}. \tag{7.76}$$

When n is large, we may approximate c_α using the asymptotic result in Theorem 6.10(ii), i.e., we can replace (7.75) by

$$\sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2 c_\alpha^2} = \frac{\alpha}{2}. \tag{7.77}$$

The resulting intervals in (7.76) have limiting confidence coefficient $1-\alpha$.

Using $D_n^+(F) = \sup_{t\in\mathcal{R}}[F_n(t) - F(t)]$ and the results in Theorem 6.10, we can also obtain one-sided simultaneous confidence intervals for $F(t)$, $t \in \mathcal{R}$, with confidence coefficient $1-\alpha$ or limiting confidence coefficient $1-\alpha$.

When n is small, it is possible that some intervals in (7.76) are not within the interval $[0,1]$. This is undesirable since $F(t) \in [0,1]$ for all t. One way to solve this problem is replacing $F_n(t) - c_\alpha$ and $F_n(t) + c_\alpha$ by

$$\max[F_n(t) - c_\alpha, 0] \qquad \text{and} \qquad \min[F_n(t) + c_\alpha, 1],$$

respectively. But the resulting intervals have a confidence coefficient larger than $1-\alpha$. The limiting confidence coefficient of these intervals is still $1-\alpha$ (exercise).

7.6 Exercises

1. Let $X_{i1}, ..., X_{in_i}$, $i = 1,2$, be two independent samples i.i.d. from $N(\mu_i, \sigma_i^2)$, $i = 1,2$, respectively. Let $\bar{X}_i$ and S_i^2 be the sample mean and sample variance of the ith sample, $i = 1,2$.
(a) Let $\theta = \mu_2 - \mu_1$. Assume that $\sigma_1 = \sigma_2$. Show that

$$t(X,\theta) = \frac{(\bar{X}_2 - \bar{X}_1 - \theta)/\sqrt{n_1^{-1} + n_2^{-1}}}{\sqrt{[(n_1-1)S_1^2 + (n_2-1)S_2^2]/(n_1+n_2-2)}}$$

is a pivotal quantity and construct a confidence interval for θ with confidence coefficient $1-\alpha$, using $t(X,\theta)$.
(b) Let $\theta = \sigma_2^2/\sigma_1^2$. Show that $\Re(X,\theta) = S_2^2/(\theta S_1^2)$ is a pivotal quantity and construct a confidence interval for θ with confidence coefficient $1-\alpha$, using $\Re(X,\theta)$.

2. Let X_i, $i = 1,2$, be independent with the p.d.f.'s $\lambda_i e^{-\lambda_i x} I_{(0,\infty)}(x)$, $i = 1,2$, respectively.
(a) Let $\theta = \lambda_1/\lambda_2$. Show that $\theta X_1/X_2$ is a pivotal quantity and construct a confidence interval for θ with confidence coefficient $1-\alpha$, using this pivotal quantity.
(b) Let $\theta = (\lambda_1, \lambda_2)$. Show that $\lambda_1 X_1 + \lambda_2 X_2$ is a pivotal quantity and construct a confidence set for θ with confidence coefficient $1-\alpha$, using this pivotal quantity.

3. In Example 7.3, show that the equation $n[\bar{Y}(\theta)]^2 = t^2_{n-1,\alpha/2} S^2(\theta)$ defines a parabola in θ and discuss when $C(X)$ is a finite interval, the complement of a finite interval, or the whole real line.

4. Let X be a sample from P in a parametric family indexed by θ. Suppose that $T(X)$ is a real-valued statistic with p.d.f. $f_\theta(t)$ and that $\Re(t,\theta)$ is a monotone function of t for each θ. Show that if

$$f_\theta(t) = g(\Re(t,\theta))\left|\frac{\partial}{\partial t}\Re(t,\theta)\right|$$

for some function g, then $\Re(T(X),\theta)$ is a pivotal quantity.

5. Let $X_1, ..., X_n$ be i.i.d. from $N(\theta,\theta)$ with an unknown $\theta > 0$. Find a pivotal quantity and use it to construct a confidence interval for θ.

6. Prove (7.3).

7. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(0,\theta)$ with an unknown $\theta > 0$.
(a) Using the pivotal quantity $\bar{X}/\theta$, construct a confidence interval for θ with confidence coefficient $1-\alpha$.
(b) Apply Theorem 7.1 with $T = \bar{X}$ to construct a confidence interval for θ with confidence coefficient $1-\alpha$.

8. Let $X_1, ..., X_n$ be i.i.d. random variables with the Lebesgue p.d.f. $\frac{a}{\theta}\left(\frac{x}{\theta}\right)^{a-1} I_{(0,\theta)}(x)$, where $a \geq 1$ is known and $\theta > 0$ is unknown.
(a) Apply Theorem 7.1 with $T = X_{(n)}$ to construct a confidence interval for θ with confidence coefficient $1-\alpha$. Compare the result with that in Example 7.2 when $a = 1$.
(b) Show that the confidence interval in (a) can also be obtained using a pivotal quantity.

9. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(a,1)$ with an unknown a.
(a) Construct a confidence interval for a with confidence coefficient $1-\alpha$ by using Theorem 7.1 with $T = X_{(1)}$.
(b) Show that the confidence interval in (a) can also be obtained using a pivotal quantity.

10. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $P(X_i = 1) = p$.
(a) Using Theorem 7.1 with $T = \sum_{i=1}^n X_i$, show that a level $1-\alpha$ confidence interval for p is

$$\left[\frac{1}{1+\frac{n-T+1}{T}F_{2(n-T+1),2T,\alpha_2}}, \frac{\frac{T+1}{n-T}F_{2(T+1),2(n-T),\alpha_1}}{1+\frac{T+1}{n-T}F_{2(T+1),2(n-T),\alpha_1}}\right],$$

where $F_{a,b,\alpha}$ is the $(1-\alpha)$th quantile of the F-distribution $F_{a,b}$, and $F_{a,0,\alpha}$ is defined to be ∞. (Hint: show that $P(T \geq t) = P(Y \leq p)$, where Y has the beta distribution $B(T, n-T+1)$.)

(b) Show that $(\underline{p}, 1]$ with $\underline{p}$ given by (7.5) is a level $1-\alpha$ confidence interval. Compare it with the interval obtained using the result in (a) with $\alpha_1 = 0$.

11. Let X be a sample of size 1 from the negative binomial distribution $NB(p, r)$ with a known r and an unknown $p \in (0, 1)$. Using Theorem 7.1 with $T = X$, show that a level $1-\alpha$ confidence interval for p is

$$\left[\frac{1}{1+\frac{T+1}{r+1}F_{2(T+1),2(r+1),\alpha_2}}, \frac{\frac{r+1}{T}F_{2(r+1),2T,\alpha_1}}{1+\frac{r+1}{T}F_{2(r+1),2T,\alpha_1}}\right],$$

where $F_{a,b,\alpha}$ is the same as that in the previous exercise.

12. Prove Proposition 7.2.

13. In Example 7.7, show that $c(\theta)$ and $c_i(\theta)$'s are nondecreasing in θ.

14. Show that the confidence intervals in Example 7.14 and Exercise 1 can also be obtained by inverting the acceptance regions of the tests for one-sample and two-sample problems in §6.2.3.

15. Let X_i, $i = 1, 2$, be independently distributed as the binomial distributions $Bi(p_i, n_i)$, $i = 1, 2$, respectively, where n_i's are known and p_i's are unknown. Show how to invert the acceptance regions of the UMPU tests in Example 6.11 to obtain a level $1-\alpha$ confidence interval for the odds ratio $\frac{p_2(1-p_1)}{p_1(1-p_2)}$.

16. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$.
(a) Suppose that $\sigma^2 = \gamma\mu^2$ with unknown $\gamma > 0$ and $\mu \in \mathcal{R}$. Obtain a confidence set for γ with confidence coefficient $1-\alpha$ by inverting the acceptance regions of LR tests for $H_0 : \gamma = \gamma_0$ versus $H_1 : \gamma \neq \gamma_0$.
(b) Repeat (a) when $\sigma^2 = \gamma\mu$ with unknown $\gamma > 0$ and $\mu > 0$.

17. Consider the problem in Example 6.17. Discuss how to construct a confidence interval for θ with confidence coefficient $1-\alpha$ by
(a) inverting the acceptance regions of the tests derived in Example 6.17;
(b) applying Theorem 7.1.

18. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. Construct a confidence interval for θ with confidence coefficient $1-\alpha$.

19. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $P(X_i = 1) = p$. Using the p.d.f. of the beta distribution $B(a, b)$ as the prior p.d.f., construct a level $1-\alpha$ HPD credible set for p.

20. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with an unknown $\theta = (\mu, \sigma^2)$. Consider the prior p.d.f. $\pi(\theta) = \pi_1(\mu|\sigma^2)\pi_2(\sigma^2)$, where $\pi_1(\mu|\sigma^2)$ is the p.d.f. of $N(\mu_0, \sigma_0^2\sigma^2)$,

$$\pi_2(\sigma^2) = \frac{1}{\Gamma(a)b^a}\left(\frac{1}{\sigma^2}\right)^{a+1} e^{-1/(b\sigma^2)} I_{(0,\infty)}(\sigma^2),$$

and μ_0, a, and b are known.
(a) Find the posterior of μ and construct a level $1-\alpha$ HPD credible set for μ.
(b) Show that the credible set in (a) approaches the confidence interval obtained in Example 7.14 as σ_0^2, a, and b approach some limits.

21. Let $X_1, ..., X_n$ be i.i.d. with a Lebesgue p.d.f. $\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$, where f is a known p.d.f. and μ and $\sigma > 0$ are unknown. Let X_0 be a future observation that is independent of X_i's and has the same distribution as X_i. Find a pivotal quantity $\Re(X, X_0)$ and construct a level $1-\alpha$ prediction set for X_0.

22. Let $X_1, ..., X_n$ be i.i.d. from a continuous c.d.f. F on $\mathcal{R}$ and X_0 be a future observation that is independent of X_i's and has the c.d.f. F. Suppose that F is strictly increasing in a neighborhood of $F^{-1}(\alpha/2)$ and a neighborhood of $F^{-1}(1-\alpha/2)$. Let F_n be the empirical c.d.f. defined by (5.1). Show that the prediction interval $C(X) = [F_n^{-1}(\alpha/2), F_n^{-1}(1-\alpha/2)]$ for X_0 satisfies $P(X_0 \in C(X)) \to 1-\alpha$, where P is the joint distribution of $(X_0, X_1, ..., X_n)$.

23. Let $X_1, ..., X_n$ be i.i.d. with a Lebesgue p.d.f. $f(x-\mu)$, where f is known and μ is unknown.
(a) If f is the p.d.f. of the standard normal distribution, show that the confidence interval $[\bar{X}-c_1, \bar{X}+c_1]$ is better than $[X_1-c_2, X_1+c_2]$ in terms of their lengths, where c_i's are chosen so that these confidence intervals have confidence coefficient $1-\alpha$.
(b) If f is the p.d.f. of the Cauchy distribution $C(0,1)$, show that the two confidence intervals in (a) have the same length.

24. Prove Theorem 7.3(ii).

25. Show that the expected length of the interval in (7.13) is shorter than the expected length of the interval in (7.12).

26. Consider Example 7.14.
(a) Suppose that $\theta = \sigma^2$ and μ is known. Let a_* and b_* be constants satisfying $a_*^2 g(a_*) = b_*^2 g(b_*) > 0$ and $\int_{a_*}^{b_*} g(x)dx = 1-\alpha$, where g is the p.d.f. of the chi-square distribution χ_n^2. Show that the interval $[b_*^{-1}T, a_*^{-1}T]$ has the shortest length within the class of intervals of

the form $[b^{-1}T, a^{-1}T]$, $\int_a^b g(x)dx = 1-\alpha$, where $T = \sum_{i=1}^n (X_i - \mu)^2$.
(b) Show that the expected length of the interval in (a) is shorter than the expected length of the interval in (7.14).
(c) Find the shortest-length interval for $\theta = \sigma$ within the class of confidence intervals of the form $[b^{-1/2}\sqrt{n-1}S, a^{-1/2}\sqrt{n-1}S]$, where $0 < a < b < \infty$ and $\int_a^b f(x)dx = 1-\alpha$.

27. Assume the conditions in Theorem 7.3(i). Assume further that f is symmetric. Show that a_* and b_* in Theorem 7.3(i) must satisfy $a_* = -b_*$.

28. Let f be a Lebesgue p.d.f. that is nonzero in $[x_-, x_+]$ and is 0 outside $[x_-, x_+]$, $-\infty \le x_- < x_+ \le \infty$.
(a) Suppose that f is strictly dereasing. Show that, among all intervals $[a, b]$ satisfying $\int_a^b f(x)dx = 1-\alpha$, the shortest interval is obtained by choosing $a = x_-$ and b so that $\int_{x_-}^b f(x)dx = 1-\alpha$.
(b) Obtain a result similar to that in (a) when f is strictly increasing. Use the result to show that the interval $[X_{(n)}, \alpha^{-1/n}X_{(n)}]$ in Example 7.13 has the shortest length among all intervals $[b^{-1}X_{(n)}, a^{-1}X_{(n)}]$.

29. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(a, 1)$ with an unknown a. Find a confidence interval for a having the shortest length within the class of confidence intervals $[X_{(1)} + c, X_{(1)} + d]$ with confidence coefficient $1-\alpha$.

30. Consider the HPD credible set $C(x)$ in (7.7) for a real-valued θ. Suppose that $p_x(\theta)$ is a unimodal Lebesgue p.d.f. and is not monotone. Show that $C(x)$ is an interval having the shortest length within the class of intervals $[a, b]$ satisfying $\int_a^b p_x(\theta)d\theta = 1-\alpha$.

31. Let X be a single observation from the gamma distribution $\Gamma(\alpha, \gamma)$ with a known α and an unknown γ. Find the shortest-length confidence interval within the class of confidence intervals $[b^{-1}X, a^{-1}X]$ with a given confidence coefficient.

32. Let $X_1, ..., X_n$ be i.i.d. with the Lebesgue p.d.f. $\theta x^{\theta-1}I_{(0,1)}(x)$, where $\theta > 0$ is unknown.
(a) Construct a confidence interval for θ with confidence coefficient $1-\alpha$, using a sufficient statistic.
(b) Discuss whether the confidence interval obtained in (a) has the shortest length within a class of confidence intervals.
(c) Discuss whether the confidence interval obtained in (a) is UMAU.

33. Let X be a single observation from the logistic distribution $LG(\mu, 1)$ with an unknown $\mu \in \mathcal{R}$. Find a Θ'-UMA upper confidence bound for μ with confidence coefficient $1-\alpha$, where $\Theta' = (\mu, \infty)$.

34. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(0, \theta)$ with an unknown $\theta > 0$. Find a Θ'-UMA lower confidence bound for θ with confidence coefficient $1 - \alpha$, where $\Theta' = (0, \theta)$.

35. Let X be a single observation from $N(\theta - 1, 1)$ if $\theta < 0$, $N(0, 1)$ if $\theta = 0$, and $N(\theta + 1, 1)$ if $\theta > 0$.
(a) Show that the distribution of X is in a family with monotone likelihood ratio.
(b) Construct a Θ'-UMA lower confidence bound for θ with confidence coeffcient $1 - \alpha$, where $\Theta' = (-\infty, \theta)$.

36. Show that the confidence set in Example 7.9 is unbiased.

37. In Example 7.13, show that the confidence interval $[X_{(n)}, \alpha^{-1/n} X_{(n)}]$ is UMA and has the shortest expected length among all confidence intervals for θ with confidence coefficient $1 - \alpha$.

38. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(a, \theta)$ with unknown a and θ. Find a UMAU confidence interval for a with confidence coefficient $1 - \alpha$.

39. Let Y and U be independent having the binomial distribution $Bi(p, n)$ and the uniform distribution $U(0, 1)$, respectively.
(a) Show that $W = Y + U$ has the Lebesgue p.d.f. $f_p(w)$ given by (7.17).
(b) Show that the family $\{f_p : p \in (0, 1)\}$ has monotone likelihood ratio in W.

40. Let $X_1, ..., X_n$ be i.i.d. from the Poisson distribution $P(\theta)$ with an unknown $\theta > 0$. Find a randomized UMA upper confidence bound for θ with confidence coefficient $1 - \alpha$.

41. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with unknown μ and σ^2.
(a) Show that $\overline{\theta} = \bar{X} + t_{n-1,\alpha} S/\sqrt{n}$ is a UMAU upper confidence bound for μ with confidence coefficient $1 - \alpha$, where $t_{n-1,\alpha}$ is the $(1 - \alpha)$th quantile of the t-distribution t_{n-1}.
(b) Show that the confidence bound in (a) can be derived by inverting acceptance regions of LR tests.

42. Prove Theorem 7.7 and Proposition 7.3.

43. Let $X_1, ..., X_n$ be i.i.d. with p.d.f. $f(x - \theta)$, where f is a known Lebesgue p.d.f. Show that the confidence interval $[\bar{X} - c_1, \bar{X} + c_2]$ has constant coverage probability, where c_1 and c_2 are constants.

44. Prove the claim in Example 7.18.

45. In Example 7.19, show that
(a) the testing problem is invariant under $\mathcal{G}_{\mu_0}$, but not $\mathcal{G}$;
(b) the nonrandomized test with acceptance region $A(\mu_0)$ is UMP among unbiased and invariant tests of size α, under $\mathcal{G}_{\mu_0}$;
(c) $\mathcal{G}$ is the smallest group containing $\cup_{\mu_0 \in \mathcal{R}} \mathcal{G}_{\mu_0}$.

46. In Example 7.17, show that intervals (7.13) and (7.14) are UMA among unbiased and invariant confidence intervals with confidence coefficient $1-\alpha$, under $\mathcal{G}_1$ and $\mathcal{G}$, respectively.

47. Let X_i, $i = 1, 2$, be independent with the exponential distributions $E(0, \theta_i)$, $i = 1, 2$, respectively.
(a) Show that $[\alpha Y/(2-\alpha), (2-\alpha)Y/\alpha]$ is a UMAU confidence interval for θ_2/θ_1 with confidence coefficient $1-\alpha$, where $Y = X_2/X_1$.
(b) Show that the confidence interval in (a) is also UMAI.

48. Let $X_1, ..., X_n$ be i.i.d. from a bivariate normal distribution with unknown mean and covariance matrix and let $R(X)$ be the sample correlation coefficient. Define $\underline{\rho} = C^{-1}(R(X))$, where $C(\rho)$ is determined by
$$P\big(R(X) \le C(\rho)\big) = 1 - \alpha$$
and ρ is the unknown correlation coefficient. Show that $\underline{\rho}$ is a Θ'-UMAI lower confidence bound for ρ with confidence coefficient $1-\alpha$, where $\Theta' = (-1, \rho)$.

49. Let $X_{i1}, ..., X_{in_i}$, $i = 1, 2$, be two independent samples i.i.d. from $N(\mu_i, \sigma^2)$, $i = 1, 2$, respectively, where μ_i's are unknown. Find a UMAI confidence interval for $\mu_2 - \mu_1$ with confidence coefficient $1-\alpha$ when (a) σ^2 is known; (b) σ^2 is unknown.

50. In Example 7.23, show that $C_3(X) = [p_-, p_+]$ with the given $p_\pm$. Compare the lengths of the confidence intervals $C_2(X)$ and $C_3(X)$.

51. Show that the confidence intervals in Example 7.14 can be derived by inverting acceptance regions of LR tests.

52. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(0, \theta)$ with an unknown $\theta > 0$.
(a) Show that $\Re(X, \theta) = \sqrt{n}(\bar{X} - \theta)/\theta$ is asymptotically pivotal. Construct a $1-\alpha$ asymptotically correct confidence interval for θ, using $\Re(X, \theta)$.
(b) Show that $\Re_1(X, \theta) = \sqrt{n}(\bar{X} - \theta)/\bar{X}$ is asymptotically pivotal. Construct a $1-\alpha$ asymptotically correct confidence interval for θ, using $\Re_1(X, \theta)$.
(c) Obtain $1-\alpha$ asymptotically correct confidence intervals for θ by

inverting acceptance regions of LR tests, Wald's tests, and Rao's score tests.

53. Let $X_1, ..., X_n$ be i.i.d. from the Poisson distribution $P(\theta)$ with an unknown $\theta > 0$.
(a) Show that $\Re(X, \theta) = (\bar{X} - \theta)/\sqrt{\theta/n}$ is asymptotically pivotal. Construct a $1 - \alpha$ asymptotically correct confidence interval for θ, using $\Re(X, \theta)$.
(b) Show that $\Re_1(X, \theta) = (\bar{X} - \theta)/\sqrt{\bar{X}/n}$ is asymptotically pivotal. Construct a $1 - \alpha$ asymptotically correct confidence interval for θ, using $\Re_1(X, \theta)$.
(c) Obtain $1 - \alpha$ asymptotically correct confidence intervals for θ by inverting acceptance regions of LR tests, Wald's tests, and Rao's score tests.

54. Let $X_{i1}, ..., X_{in_i}$, $i = 1, 2$, be two independent samples i.i.d. from $N(\mu_i, \sigma_i^2)$, $i = 1, 2$, respectively, where all parameters are unknown. Show that $\Re(X, \mu_1 - \mu_2) = (\bar{X}_{1\cdot} - \bar{X}_{2\cdot} - \mu_1 + \mu_2)/\sqrt{n_1^{-1}S_1^2 + n_2^{-1}S_2^2}$ is asymptotically pivotal, assuming that $n_1/n_2 \to c \in (0, \infty)$. Construct a $1 - \alpha$ asymptotically correct confidence interval for $\mu_1 - \mu_2$ using $\Re(X, \mu_1 - \mu_2)$.

55. Consider the problem in Example 5.15. Construct an asymptotically pivotal quantity and a $1 - \alpha$ asymptotically correct confidence set for θ. Compare this confidence set with those in Example 7.24.

56. Consider the problem in Example 3.21. Construct an asymptotically pivotal quantity and a $1 - \alpha$ asymptotically correct confidence set for μ_y/μ_x.

57. Consider the problem in Example 3.23. Construct an asymptotically pivotal quantity and a $1 - \alpha$ asymptotically correct confidence set for $R(t)$ with a fixed t.

58. Let U_n be a U-statistic based on i.i.d. $X_1, ..., X_n$ and the kernel $h(x_1, ..., x_m)$, and let $\theta = E(U_n)$. Construct an asymptotically pivotal quantity based on U_n and a $1 - \alpha$ asymptotically correct confidence set for θ.

59. Let $X_1, ..., X_n$ be i.i.d. from a c.d.f. F on $\mathcal{R}$ that is continuous and symmetric about θ. Let $\hat{\theta} = W/n - \frac{1}{2}$ and $\mathtt{T}(F) = \theta + \frac{1}{2}$, where W and $\mathtt{T}$ are given by (6.78) and (5.46), respectively. Construct a confidence interval for θ that has limiting confidence coefficient $1 - \alpha$.

60. Consider the linear model $X = \beta Z^\tau + \varepsilon$, where ε has independent components with mean 0 and Z is of full rank. Assume the conditions

in Theorem 3.12.
(a) Suppose that $\mathrm{Var}(\varepsilon) = \sigma^2 D$, where D is a known diagonal matrix and σ^2 is unknown. Find an asymptotically pivotal quantity and construct a $1-\alpha$ asymptotically correct confidence set for β.
(b) Suppose that $\mathrm{Var}(\varepsilon)$ is an unknown diagonal matrix. Find an asymptotically pivotal quantity and construct a $1-\alpha$ asymptotically correct confidence set for β.

61. Consider a GEE estimator $\hat{\theta}$ of θ described in §5.4.1. Discuss how to construct an asymptotically pivotal quantity and a $1-\alpha$ asymptotically correct confidence set for θ. (Hint: see §5.5.2.)

62. Suppose that $X_1, ..., X_n$ are i.i.d. from the negative binomial distribution $NB(p, r)$ with a known r and an unknown p. Obtain $1-\alpha$ asymptotically correct confidence intervals for p by inverting acceptance regions of LR tests, Wald's tests, and Rao's score tests.

63. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(a, \theta)$ with unknown a and θ. Find a $1-\alpha$ asymptotically correct confidence set by inverting acceptance regions of LR tests.

64. Let $X_{i1}, ..., X_{in_i}$, $i = 1, 2$, be two independent samples i.i.d. from $N(\mu_i, \sigma^2)$, $i = 1, 2$, respectively, where all parameters are unknown.
(a) Find $1-\alpha$ asymptotically correct confidence sets for (μ_1, μ_2) by inverting acceptance regions of LR tests, Wald's tests, and Rao's score tests.
(b) Repeat (a) for the parameter (μ_1, μ_2, σ^2).

65. Let $X_{i1}, ..., X_{in_i}$, $i = 1, 2$, be two independent samples i.i.d. from $N(\mu_i, \sigma_i^2)$, $i = 1, 2$, respectively, where all parameters are unknown.
(a) Find $1-\alpha$ asymptotically correct confidence sets for (μ_1, μ_2) by inverting acceptance regions of LR tests, Wald's tests, and Rao's score tests.
(b) Repeat (a) for the parameter $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$.

66. Let $X_{i1}, ..., X_{in_i}$, $i = 1, 2$, be two independent samples i.i.d. from the exponential distributions $E(0, \theta_i)$, $i = 1, 2$, respectively, where θ_i's are unknown. Find $1-\alpha$ asymptotically correct confidence sets for (θ_1, θ_2) by inverting acceptance regions of LR tests, Wald's tests, and Rao's score tests.

67. Consider the problem in Example 7.9. Find $1-\alpha$ asymptotically correct confidence sets for θ by inverting acceptance regions of LR tests, Wald's tests, and Rao's score tests. Which one is the same as that derived in Example 7.9?

68. Let $X_1, ..., X_n$ be i.i.d. from a continuous c.d.f. F on $\mathcal{R}$ and let $\theta = F^{-1}(p)$, $p \in (0,1)$.
(a) Show that $P(X_{(k_1)} \leq \theta \leq X_{(k_2)}) = P(U_{(k_1)} \leq p \leq U_{(k_2)})$, where $X_{(k)}$ is the kth order statistic and $U_{(k)}$ is the kth order statistic based on a sample $U_1, ..., U_n$ i.i.d. from the uniform distribution $U(0,1)$.
(b) Show that

$$P(U_{(k_1)} \leq p \leq U_{(k_2)}) = B_p(k_1, n-k_1+1) - B_p(k_2, n-k_2+1),$$

where

$$B_p(i,j) = \frac{\Gamma(i+j)}{\Gamma(i)\Gamma(j)} \int_0^p t^{i-1}(1-t)^{j-1} dt.$$

(c) Discuss how to obtain a confidence interval for θ with confidence coefficient $1-\alpha$.

69. Prove Corollary 7.1.

70. Assume the conditions in Corollary 7.1.
(a) Show that $\sqrt{n}(X_{(k_n)} - \theta)F'(\theta) \to_d N(c, p(1-p))$.
(b) Prove result (7.24) using the result in part (a).
(c) Construct a consistent estimator of the asymptotic variance of the sample median (see Example 6.27), using Woodruff's interval.

71. Show that $\underline{\theta}_{HB}$ in (7.32) is equal to $2\hat{\theta}_n - K_B^{-1}(1-\alpha)$, where K_B is defined in (7.25).

72. (Parametric bootstrapping in location-scale families). Let $X_1, ..., X_n$ be i.i.d. random variables with p.d.f. $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$, where f is a known Lebesgue p.d.f. and μ and $\sigma > 0$ are unknown. Let $X_1^*, ..., X_n^*$ be i.i.d. bootstrap data from the p.d.f. $\frac{1}{s} f\left(\frac{x-\bar{x}}{s}\right)$, where $\bar{x}$ and s^2 are the observed sample mean and sample variance, respectively.
(a) Suppose that we construct the bootstrap-t lower confidence bound (7.33) for μ using the parametric bootstrap data. Show that $\underline{\theta}_{BT}$ has confidence coefficient $1-\alpha$.
(b) Suppose that we construct the hybrid bootstrap lower confidence bound (7.32) for μ using the parametric bootstrap data. Show that $\underline{\theta}_{HB}$ does not necessarily have confidence coefficient $1-\alpha$.
(c) Suppose that f has mean 0 and variance 1. Show that $\underline{\theta}_{HB}$ in (b) is $1-\alpha$ asymptotically correct.

73. (The bootstrap BC_a percentile method). Suppose that we change assumption (7.28) to

$$P\left(\frac{\hat{\phi}_n - \phi_n(\theta)}{1 + a\phi_n(\theta)} + z_0 \leq x\right) = \Psi(x),$$

where a is an extra parameter called the acceleration constant.
(a) If ϕ_n, z_0, a, and Φ are known, show that the following lower confidence bound for θ has confidence coefficient $1-\alpha$:

$$\underline{\theta}_E = \phi_n^{-1}\big(\hat{\phi}_n + (z_\alpha + z_0)(1 + a\hat{\phi}_n)/[1 - a(z_\alpha + z_0)]\big).$$

(b) Show that $K_B^{-1}(x) = \phi_n^{-1}\big(\hat{\phi}_n + [\Phi^{-1}(x) - z_0](1 + a\hat{\phi}_n)\big)$, where K_B is defined in (7.25).
(c) Let $\underline{\theta}_{BC}(a) = K_B^{-1}\big(\Phi(z_0 + (z_\alpha + z_0)/[1 - a(z_\alpha + z_0)])\big)$. Show that $\underline{\theta}_{BC}(a) = \underline{\theta}_E$ in part (a). (The bootstrap BC_a percentile lower confidence bound for θ is $\underline{\theta}_{BC}(\hat{a})$, where $\hat{a}$ is an estimator of a.)

74. (Automatic bootstrap percentile). Let $\mathcal{P} = \{P_\theta : \theta \in \mathcal{R}\}$ be a parametric family. Define $K_\theta(x) = P_\theta(\hat{\theta}_n \le x)$, where $\hat{\theta}_n$ is an estimator of θ. Let θ_0 be a given value of θ and $\theta_1 = K_{\theta_0}^{-1}(1-\alpha)$. The *automatic bootstrap percentile* lower confidence bound for θ is defined to be

$$\underline{\theta}_{ABP} = K_{\hat{\theta}_n}^{-1}(K_{\theta_1}(\theta_0)).$$

Assume the assumption in the previous exercise. Show that $\underline{\theta}_{ABP}$ has confidence coefficient $1-\alpha$.

75. (Bootstrapping residuals). Consider linear model (3.25): $X = \beta Z^\tau + \varepsilon$, where Z is of full rank and ε is a vector of i.i.d. random variables having mean 0 and variance σ^2. Let $r_i = X_i - \hat{\beta} Z_i^\tau$ be the ith residual, where $\hat{\beta}$ is the LSE of β. Assume that the average of r_i's is always 0. Let $\varepsilon_1^*, ..., \varepsilon_n^*$ be i.i.d. bootstrap data from the empirical c.d.f. putting mass n^{-1} on each r_i. Define $X_i^* = \hat{\beta} Z_i^\tau + \varepsilon_i^*$, $i = 1, ..., n$.
(a) Find an expression for $\hat{\beta}^*$, the bootstrap analogue of $\hat{\beta}$. Calculate $E(\hat{\beta}^*|X)$ and $\mathrm{Var}(\hat{\beta}^*|X)$.
(b) Using $(\hat{\beta} - \beta)l^\tau$ and the idea in §7.4.1, construct a hybrid bootstrap lower confidence bound for βl^τ, where $l \in \mathcal{R}^p$.
(c) Discuss when the lower confidence bound in (b) is $1-\alpha$ asymptotically correct.
(d) Describe how to construct a bootstrap-t lower confidence bound for βl^τ.
(e) Describe how to construct a hybrid bootstrap confidence set for β, using the idea in §7.4.1.

76. (Bootstrapping pairs). Consider linear model (3.25): $X = \beta Z^\tau + \varepsilon$, where Z is of full rank and ε is a vector of independent random variables having mean 0 and finite variances. Let $(X_1^*, Z_1^*), ..., (X_n^*, Z_n^*)$ be i.i.d. bootstrap data from the empirical c.d.f. putting mass n^{-1} on each (X_i, Z_i). Define $\hat{\beta}^* = \sum_{i=1}^n X_i^* Z_i^* (Z^\tau Z)^{-1}$. Repeat (a)-(e) of the previous exercise.

77. (External bootstrapping or wild bootstrapping). Assume the model in the previous exercise. Let $\varepsilon_1^*, ..., \varepsilon_n^*$ be i.i.d. random variables with mean 0 and variance 1. Define the bootstrap data as $X_i^* = \hat{\beta} Z_i^\tau + |t_i|\varepsilon_i^*$, $i = 1, ..., n$, where $\hat{\beta}$ is the LSE of β, $t_i = (X_i - \hat{\beta} Z_i^\tau)/\sqrt{1 - h_i}$, and $h_i = Z_i(Z^\tau Z)^{-1} Z_i^\tau$. Repeat (a)-(e) of Exercise 75.

78. Prove (7.48) and (7.49).

79. Using the Edgeworth expansion given in Example 7.25, construct a third-order accurate lower confidence bound for μ. (Hint: use (7.50) and (7.51).)

80. Describe how to approximate $C_{PREB}^{(3)}(X)$ in (7.53), using the Monte Carlo method.

81. Prove (7.55) and (7.56).

82. Show that Bonferroni's simultaneous confidence intervals are of level $1 - \alpha$.

83. Let $C_{t,\alpha}(X)$ be a confidence interval for θ_t with confidence coefficient $1-\alpha$, $t = 1, ..., k$. Suppose that $C_{1,\alpha}(X), ..., C_{k,\alpha}(X)$ are independent for any α. Show how to construct simultaneous confidence intervals for θ_t, $t = 1, ..., k$, with confidence coefficient $1 - \alpha$.

84. Show that $C_{ij,\alpha}(X)$ in (7.62) is UMAU for $\mu_i - \mu_j$.

85. Consider the two-way balanced ANOVA model in Example 6.19. Using Bonferroni's method, obtain level $1 - \alpha$ simultaneous confidence intervals for
(a) α_i, $i = 1, ..., a - 1$;
(b) μ_{ij}, $i = 1, ..., a$, $j = 1, ..., b$.

86. Prove (7.64). (Hint: use the Cauchy-Schwarz inequality.)

87. Let $x \in \mathcal{R}^k$, $y \in \mathcal{R}^k$, and A be a $k \times k$ positive definite matrix.
(a) Suppose that $yA^{-1}x^\tau = 0$. Show that

$$xA^{-1}x^\tau = \max_{c \in \mathcal{R}^k, c \neq 0, cy^\tau = 0} \frac{(cx^\tau)^2}{cAc^\tau}.$$

(b) Assume model (7.63) with a full rank Z. Using the result in (a), construct simultaneous confidence intervals (with confidence coefficient $1 - \alpha$) for βc^τ, $c \in \mathcal{R}^p$, $c \neq 0$, $cy^\tau = 0$, where $y \in \mathcal{R}^p$ satisfying $yZ^\tau Z = 0$.

88. Assume the conditions in Theorem 3.12. Show that Scheffé's intervals in Theorem 7.10 are $1 - \alpha$ asymptotically correct.

89. Prove (7.67).

90. Find explicitly the $m(m-1)/2$ vectors in the set $\mathcal{T}_0$ in (7.68) so that $\{\beta L^\tau t^\tau : t \in \mathcal{T}_0\}$ is exactly the same as $\mu_i - \mu_j$, $1 \le i < j \le m$. Show that the intervals in (7.68) are Scheffé's simultaneous confidence intervals.

91. In Example 7.27, show that
(a) Scheffé's intervals in Theorem 7.10 with $t = (1, x)$ and $L = I_2$ are of the form (7.69);
(b) the maximum on the right-hand side of (7.70) is achieved at (y, x) given by (7.71);
(c) y in (7.71) is equal to 1.

92. Consider the two-way balanced ANOVA model in Example 6.19. Using Scheffé's method, obtain level $1-\alpha$ simultaneous confidence intervals for α_i's, β_j's, and γ_{ij}'s.

93. Let $X_{ij} = N(\mu + \alpha_i + \beta_j, \sigma^2)$, $i = 1, ..., a$, $j = 1, ..., b$, be independent, where $\sum_{i=1}^a \alpha_i = 0$ and $\sum_{j=1}^b \beta_j = 0$. Construct level $1-\alpha$ simultaneous confidence intervals for all linear combinations of α_i's and β_j's, using
(a) Bonferroni's method;
(b) Scheffé's method.

94. Assume model (7.63) with $\beta = (\beta_0, \beta_1, \beta_2)$ and $Z_i = (1, t_i, t_i^2)$, where $t_i \in \mathcal{R}$, $\sum_{i=1}^n t_i = 0$, $\sum_{i=1}^n t_i^2 = 1$, and $\sum_{i=1}^n t_i^3 = 0$.
(a) Construct a confidence ellipsoid for (β_1, β_2) with confidence coefficient $1-\alpha$;
(b) Construct simultaneous confidence intervals for all linear combinations of β_1 and β_2, with confidence coefficient $1-\alpha$.

95. Show that the distribution of R_{st} in (7.72) does not depend on any unknown parameter.

96. For $\alpha = 0.05$, obtain numerically the t-type confidence intervals in (7.62), Bonferroni's, Scheffé's, and Tukey's simultaneous confidence intervals for $\mu_i - \mu_j$, $1 \le i < j \le 4$, based on the following data X_{ij} from a one-way ANOVA model ($q_{0.05} = 4.45$):

	$j=1$	2	3	4	5	6
$i=1$	0.08	0.10	0.09	0.07	0.09	0.06
2	0.15	0.09	0.11	0.10	0.08	0.13
3	0.13	0.10	0.15	0.09	0.09	0.17
4	0.05	0.11	0.07	0.09	0.11	0.08

97. (Dunnett's simultaneous confidence intervals). Let X_{0j} $(j = 1, ..., n_0)$ and X_{ij} $(i = 1, ..., m,\ j = 1, ..., n_0)$ represent independent measurements on a standard and m competing new treatments. Suppose that $X_{ij} = N(\mu_i, \sigma^2)$ with unknown μ_i and $\sigma^2 > 0$, $i = 0, 1, ..., m$. Let $\bar{X}_{i\cdot}$ be the sample mean based on X_{ij}, $j = 1, ..., n_0$, and $\hat{\sigma}^2 = [(m+1)(n_0-1)]^{-1} \sum_{i=0}^m \sum_{j=1}^{n_0} (X_{ij} - \bar{X}_{i\cdot})^2$.
(a) Show that the distribution of

$$R_{st} = \max_{i=1,...,m} |(\bar{X}_{i\cdot} - \mu_i) - (\bar{X}_{0\cdot} - \mu_0)|/\hat{\sigma}$$

does not depend on any unknown parameter.
(b) Show that

$$\left[\sum_{i=0}^m c_i \bar{X}_{i\cdot} - q_\alpha \hat{\sigma} \sum_{i=1}^m |c_i|,\ \sum_{i=0}^m c_i \bar{X}_{i\cdot} + q_\alpha \hat{\sigma} \sum_{i=1}^m |c_i|\right]$$

for all $c_0, c_1, ..., c_m$ satisfying $\sum_{i=0}^m c_i = 0$ are simultaneous confidence intervals for $\sum_{i=0}^m c_i \mu_i$ with confidence coefficient $1 - \alpha$, where q_α is the $(1-\alpha)$th quantile of R_{st}.

98. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution $U(0, \theta)$, where $\theta > 0$ is unknown. Construct a confidence band for the c.d.f. of X_1 with confidence coefficient $1 - \alpha$.

99. Let $X_1, ..., X_n$ be i.i.d. with the p.d.f. $\frac{1}{\sigma} f\left(\frac{t-\mu}{\sigma}\right)$, where f is a known Lebesgue p.d.f. (a location-scale family). Let F be the c.d.f. of X_1.
(a) Suppose that μ is unknown and $\sigma > 0$ is known. Construct simultaneous confidence intervals for $F(t)$, $t \in \mathcal{R}$, with confidence coefficient $1 - \alpha$.
(b) Suppose that μ is known and $\sigma > 0$ is unknown. Construct simultaneous confidence intervals for $F(t)$, $t \in \mathcal{R}$, with confidence coefficient $1 - \alpha$.
(c) Suppose that both μ and $\sigma > 0$ are unknown. Construct level $1 - \alpha$ simultaneous confidence intervals for $F(t)$, $t \in \mathcal{R}$.

100. Let $X_1, ..., X_n$ be i.i.d. from F on $\mathcal{R}$ and F_n be the empirical c.d.f. Show that the intervals

$$\big[\max[F_n(t) - c_\alpha, 0],\ \min[F_n(t) + c_\alpha, 1]\big], \qquad t \in \mathcal{R},$$

form a confidence band for $F(t)$, $t \in \mathcal{R}$, with limiting confidence coefficient $1 - \alpha$, where c_α is given by (7.77).

References

We provide some references for further readings of the topics covered in this book.

For a general probability theory, Billingsley (1986) and Chung (1974) are suggested, although there are many standard textbooks. An asymptotic theory for statistics can be found in Serfling (1980), Shorack and Wellner (1986), Sen and Singer (1993), and Barndorff-Nielsen and Cox (1994).

More discussions of fundamentals of statistical decision theory and inference can be found in many textbooks on mathematical statistics, such as Cramér (1946), Wald (1950), Savage (1954), Ferguson (1967), Rao (1973), Rohatgi (1976), Bickel and Doksum (1977), Lehmann (1986), Casella and Berger (1990), and Barndorff-Nielsen and Cox (1994). Discussions and proofs for results related to sufficiency and completeness can be found in Rao (1945), Blackwell (1947), Hodges and Lehmann (1950), Lehmann and Scheffé (1950), and Basu (1955). More results for exponential families are given in Barndorff-Nielsen (1979).

The theory of UMVUE in §3.1.1 and §3.1.2 is mainly based on Chapter 2 of Lehmann (1983). More results on information inequalities can be found in Cramér (1946), Rao (1973), Lehmann (1983), and Pitman (1979). The theory of U-statistics and the method of projection can be found in Hoeffding (1948), Randles and Wolfe (1979), and Serfling (1980). The related theory for V-statistics is given in von Mises (1947), Serfling (1980), and Sen (1981). Three excellent textbooks for the theory of LSE are Scheffé (1959), Searle (1971), and Rao (1973). Additional materials for sample surveys can be found in Basu (1958), Godambe (1958), Cochran (1977), Särndal, Swensson, and Wretman (1992), and Ghosh and Meeden (1997).

Excellent textbooks for the Bayesian theory include Lindley (1965), Box and Tiao (1973), Berger (1985), and Schervish (1995). For Bayesian computation and Markov chain Monte Carlo, more discussions can be found in references cited in §4.1.4. More general results on invariance in estimation and testing problems are provided by Ferguson (1967) and Lehmann (1983, 1986). The theory of shrinkage estimation was established by Stein (1956)

and James and Stein (1961); Lehmann (1983) and Berger (1985) provide excellent discussions on this topic. The method of likelihood has more than 200 years of history (Edwards, 1974). An excellent textbook on the MLE in generalized linear models is McCullagh and Nelder (1989). Asymptotic properties for MLE can be found in Cramér (1946), Serfling (1980), and Sen and Singer (1993). Asymptotic results for the MLE in generalized linear models are provided by Fahrmeir and Kaufmann (1985).

An excellent book containing results for empirical c.d.f.'s and their properties is Shorack and Wellner (1986). References for empirical likelihoods are provided in §5.1.2 and §6.5.3. More results in density estimation can be found, for example, in Rosenblatt (1971) and Silverman (1986). More discussions on statistical functionals can be found in von Mises (1947), Serfling (1980), Fernholz (1983), Sen and Singer (1993), and Shao and Tu (1995). Two textbooks for robust statistics are Huber (1981) and Hampel et al. (1986). A general discussion of L-estimators and sample quantiles can be found in Serfling (1980) and Sen (1981). L-estimators in linear models are covered by Bickel (1973), Puri and Sen (1985), Welsh (1987), and He and Shao (1996). Some references on generalized estimation equations and quasi-likelihoods are Godambe and Heyde (1987), Godambe and Thompson (1989), McCullagh and Nelder (1989), and Diggle, Liang, and Zeger (1994). Two textbooks containing materials on variance estimation are Efron and Tibshirani (1993) and Shao and Tu (1995).

The theory of UMP, UMPU, and UMPI tests in Chapter 6 is mainly based on Lehmann (1986) and Chapter 5 of Ferguson (1967). Berger (1985) contains discussion on Bayesian tests. Results on large sample tests and chi-square tests can be found in Serfling (1980) and Sen and Singer (1993). Two textbooks on nonparametric tests are Lehmann (1975) and Randles and Wolfe (1979).

Further materials on confidence sets can be found in Ferguson (1967), Bickel and Doksum (1977), Lehmann (1986), and Casella and Berger (1990). More results on asymptotic confidence sets based on likelihoods can be found in Serfling (1980). The theory of bootstrap confidence sets is covered by Hall (1992), Efron and Tibshirani (1993), and Shao and Tu (1995). Further discussions on simultaneous confidence intervals can be found in Scheffé (1959), Lehmann (1986), and Tukey (1977).

The following references are those cited in this book. Many additional references can be found in Lehmann (1983, 1986).

Arvesen, J. N. (1969). Jackknifing U-statistics. *Ann. Math. Statist.*, **40**, 2076-2100.

Bahadur, R. R. (1957). On unbiased estimates of uniformly minimum variance. *Sankhyā*, **18**, 211-224.

Bahadur, R. R. (1964). On Fisher's bound for asymptotic variances. *Ann. Math. Statist.*, **35**, 1545-1552.

Bahadur, R. R. (1966). A note on quantiles in large samples. *Ann. Math. Statist.*, **37**, 577-580.

Barndorff-Nielsen, O. E. (1979). *Information and Exponential Families in Statistical Theory.* Wiley, New York.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics.* Chapman & Hall, London.

Basag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statist. Science*, **10**, 3-66.

Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhyā*, **15**, 377-380.

Basu, D. (1958). On sampling with and without replacement. *Sankhyā*, **20**, 287-294.

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika*, **74**, 151-173.

Berger, J. O. (1976). Inadmissibility results for generalized Bayes estimators of coordinates of a location vector. *Ann. Statist.*, **4**, 302-333.

Berger, J. O. (1980). Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters. *Ann. Statist.*, **8**, 545-571.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, second edition. Springer-Verlag, New York.

Bickel, P. J. (1973). On some analogues to linear combinations of order statistics in the linear model. *Ann. Statist.*, **1**, 597-616.

Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics.* Holden Day, San Francisco.

Bickel, P. J. and Yahav, J. A. (1969). Some contributions to the asymptotic theory of Bayes solutions. *Z. Wahrsch. verw. Geb.*, **11**, 257-276.

Billingsley, P. (1986). *Probability and Measure*, second edition. Wiley, New York.

Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *Ann. Math. Statist.*, **18**, 105-110.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA.

Brown, L. D. (1966). On the admissibility of invariant estimators of one or more location parameters. *Ann. Math. Statist.*, **37**, 1087-1136.

Brown, L. D. and Fox, M. (1974). Admissibility in statistical problems involving a location or scale parameter. *Ann. Statist.*, **2**, 248-266.

Carroll, R. J. (1982). Adapting for heteroscedasticity in linear models. *Ann. Statist.*, **10**, 1224-1233.

Carroll, R. J. and Cline, D. B. H. (1988). An asymptotic theory for weighted least-squares with weights estimated by replication. *Biometrika*, **75**, 35-43.

Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Wadsworth, Belmont, CA.

Chan, K. S. (1993). Asymptotic behavior of the Gibbs sampler. *J. Amer. Statist. Assoc.*, **88**, 320-325.

Chen, J. and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, **80**, 107-116.

Chen, J. and Shao, J. (1993). Iterative weighted least squares estimators. *Ann. Statist.*, **21**, 1071-1092.

Chung, K. L. (1974). *A Course in Probability Theory*, second edition. Academic Press, New York.

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.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ.

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data*. Clarendon Press, Oxford.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*, second edition. Wiley, New York.

Durbin, J. (1973). *Distribution Theory for Tests Based on the Sample Distribution Function*. SIAM, Philadelphia, PA.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, **27**, 642-669.

Edwards, A. W. F. (1974). The history of likelihood. *Internat. Statist. Rev.*, **42**, 4-15.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, **7**, 1-26.

Efron, B. (1981). Nonparametric standard errors and confidence intervals (with discussions). *Canadian J. Statist.*, **9**, 139-172.

Efron, B. (1987). Better bootstrap confidence intervals (with discussions). *J. Amer. Statist. Assoc.*, **82**, 171-200.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors — An empirical Bayes approach. *J. Amer. Statist. Assoc.*, **68**, 117-130.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, New York.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.*, **13**, 342-368.

Farrell, R. H. (1964). Estimators of a location parameter in the absolutely continuous case. *Ann. Math. Statist.*, **35**, 949-998.

Farrell, R. H. (1968a). Towards a theory of generalized Bayes tests. *Ann. Math. Statist.*, **38**, 1-22.

Farrell, R. H. (1968b). On a necessary and sufficient condition for admissibility of estimators when strictly convex loss is used. *Ann. Math. Statist.*, **38**, 23-28.

Ferguson, T. S. (1967). *Mathematical Statistics.* Academic Press, New York.

Fernholz, L. T. (1983). *Von Mises Calculus for Statistical Functionals.* Lecture Notes in Statistics, **19**, Springer-Verlag, New York.

Fuller, W. A. (1996). *Introduction to Statistical Time Series*, second edition. Wiley, New York.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.*, **85**, 398-409.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, **57**, 1317-1339.

Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *J. R. Statist. Soc.*, B, **56**, 261-274.

Ghosh, M. and Meeden, G. (1997). *Bayesian Methods in Finite Population Sampling.* Chapman & Hall, London.

Godambe, V. P. (1958). A unified theory of sampling from finite populations. *J. R. Statist. Soc.*, B, **17**, 269-278.

Godambe, V. P. and Heyde, C. C. (1987). Quasi-likelihood and optimal estimation. *Internat. Statist. Rev.*, **55**, 231-244.

Godambe, V. P. and Thompson, M. E. (1989). An extension of quasi-likelihood estimation (with discussion). *J. Statist. Plann. Inference*, **22**, 137-172.

Hall, P. (1988). Theoretical comparisons of bootstrap confidence intervals (with discussions). *Ann. Statist.*, **16**, 927-953.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

Hall, P. and Martin, M. A. (1988). On bootstrap resampling and iteration. *Biometrika*, **75**, 661-671.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, **62**, 1179-1186.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley, New York.

He, X. and Shao, Q.-M. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Ann. Statist.*, **24**, 2608-2630.

Hodges, J. L., Jr. and Lehmann, E. L. (1950). Some problems in minimax point estimation. *Ann. Math. Statist.*, **21**, 182-197.

Hoeffding, W. (1948). A class of statistics with asymptotic normal distribution. *Ann. Math. Statist.*, **19**, 293-325.

Hogg, R. V. and Tanis, E. A. (1993). *Probability and Statistical Inference*, fourth edition. Macmillan, New York.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73-101.

Huber, P. J. (1981). *Robust Statistics.* Wiley, New York.

Ibragimov, I. A. and Has'minskii, R. Z. (1972). *Statistical Estimation: Asymptotic Theory.* Springer-Verlag, New York.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, **1**, 311-319. University of California Press, CA.

Jeffreys, H. (1939, 1948, 1961). *The Theory of Probability.* Oxford University Press, Oxford.

Jones, M. C. (1991). Kernel density estimation for length biased data. *Biometrika*, **78**, 511-519.

Kalbfleisch, J. D. and Prentice, R. T. (1980). *The Statistical Analysis of Failure Time Data.* Wiley, New York.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, **53**, 457-481.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Ann. Math. Statist.*, **27**, 887-906.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giorn. Inst. Ital. Attuari*, **4**, 83-91.

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. of Calif. Publ. in Statist.*, **1**, 277-330.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden Day, San Francisco.

Lehmann, E. L. (1983). *Theory of Point Estimation.* Springer-Verlag, New York.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, second edition. Springer-Verlag, New York.

Lehmann, E. L. and Scheffé, H. (1950). Completeness, similar regions and unbiased estimation. *Sankhyā*, **10**, 305-340.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

Lindley, D. V. (1965). *Introduction to Probability and Statistics.* Cambridge University Press, Cambridge.

Liu, R. Y. and Singh, K. (1987). On a partial correction by the bootstrap. *Ann. Statist.*, **15**, 1713-1718.

Loh, W.-Y. (1987). Calibrating confidence coefficients. *J. Amer. Statist. Assoc.*, **82**, 155-162.

Loh, W.-Y. (1991). Bootstrap calibration for confidence interval construction and selection. *Statist. Sinica*, **1**, 479-495.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, second edition. Chapman & Hall, London.

Mendenhall, W. and Sincich, T. (1995). *Statistics for Engineering and the Sciences*, fourth edition. Prentice-Hall, Englewood Cliffs, NJ.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chemical Physics*, **21**, 1087-1091.

Milliken, G. A. and Johnson, D. E. (1992). *Analysis of Messy Data, Vol. 1: Designed Experiments.* Chapman & Hall, London.

Moore, D. S. and Spruill, M. C. (1975). Unified large-sample theory of general chi-squared statistics for test of fit. *Ann. Statist.*, **3**, 599-616.

Müller, H.-G. and Stadrmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.*, **15**, 610-625.

Natanson, I. P. (1961). *Theory of Functions of a Real Variable*, Vol. 1, rev. edition. Ungar, New York.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators.* Cambridge Univ. Press, New York.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237-249.

Owen, A. B. (1990). Empirical likelihood confidence regions. *Ann. Statist.*, **18**, 90-120.

Owen, A. B. (1991). Empirical likelihood for linear models. *Ann. Statist.*, **19**, 1725-1747.

Pitman, E. J. G. (1979). *Some Basic Theory for Statistical Inference.* Chapman & Hall, London.

Puri, M. L. and Sen, P. K. (1985). *Nonparametric Methods in General Linear Models.* Wiley, New York.

Qin, J. (1993). Empirical likelihood in biased sample problems. *Ann. Statist.*, **21**, 1182-1196.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**, 300-325.

Quenouille, M. (1949). Approximation tests of correlation in time series. *J. R. Statist. Soc.*, B, **11**, 18-84.

Randles, R. H. and Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics.* Wiley, New York.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calc. Math. Soc.*, **37**, 81-91.

Rao, C. R. (1947). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proc. Comb. Phil. Soc.*, **44**, 50-57.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, second edition. Wiley, New York.

Rohatgi, V. K. (1976). *An Introduction to Probability Theory and Mathematical Statistics.* Wiley, New York.

Rosenblatt, M. (1971). Curve estimates. *Ann. Math. Statist.*, **42**, 1815-1842.

Royden, H. L. (1968). *Real Analysis*, second edition. Macmillan, New York.

Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer-Verlag, New York.

Savage, S. L. (1954). *The Foundations of Statistics.* Wiley, New York.

Scheffé, H. (1959). *Analysis of Variance.* Wiley, New York.

Schervish, M. J. (1995). *Theory of Statistics.* Springer-Verlag, New York.

Searle, S. R. (1971). *Linear Models.* Wiley, New York.

Sen, P. K. (1981). *Sequential Nonparametrics: Invariance Principles and Statistical Inference.* Wiley, New York.

Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics.* Chapman & Hall, London.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

Shao, J. (1989). Monte Carlo approximations in Bayesian decision theory. *J. Amer. Statist. Assoc.*, **84**, 727-732.

Shao, J. (1993). Differentiability of statistical functionals and consistency of the jackknife. *Ann. Statist.*, **21**, 61-75.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap.* Springer-Verlag, New York.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London.

Smirnov, N. V. (1944). An approximation to the distribution laws of random quantiles determined by empirical data. *Uspehi Mat. Nauk*, **10**, 179-206.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *J. R. Statist. Soc.*, B, **51**, 47-60.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Proc. Third Berkeley Symp. Math. Statist. Prob.*, **1**, 197-206. University of California Press, CA.

Stein, C. (1959). The admissibility of Pitman's estimator for a single location parameter. *Ann. Math. Statist.*, **30**, 970-979.

Stone, C. J. (1974). Asymptotic properties of estimators of a location parameter. *Ann. Statist.*, **2**, 1127-1137.

Stone, C. J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.*, **5**, 595-645.

Strawderman, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Statist.*, **42**, 385-388.

Tanner, M. A. (1996). *Tools for Statistical Inference*, third edition. Springer-Verlag, New York.

Tate, R. F. and Klett, G. W. (1959). Optimal confidence intervals for the variance of a normal distribution. *J. Amer. Statist. Assoc.*, **54**, 674-682.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussions). *Ann. Statist.*, **22**, 1701-1762.

Tsui, K.-W. (1981). Simultaneous estimation of several Poisson parameters under squared error loss. *Ann. Inst. Statist. Math.*, **10**, 299-326.

Tukey, J. (1958). Bias and confidence in not quite large samples. *Ann. Math. Statist.*, **29**, 614.

Tukey, J. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, MA.

Vardi, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.*, **13**, 178-203.

von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functionals. *Ann. Math. Statist.*, **18**, 309-348.

Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia, PA.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.*, **54**, 426-482.

Wald, A. (1950). *Statistical Decision Functions.* Wiley, New York.

Weerahandi, S. (1995). *Exact Statistical Methods for Data Analysis.* Springer-Verlag, New York.

Welsh, A. H. (1987). The trimmed mean in the linear model. *Ann. Statist.*, **15**, 20-36.

Woodruff, R. S. (1952). Confidence intervals for medians and other position measures. *J. Amer. Statist. Assoc.*, **47**, 635-646.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussions). *Ann. Statist.*, **14**, 1261-1350.

Appendix A

Abbreviations

a.e.: almost everywhere.

amse: asymptotic mean squared error.

ANOVA: analysis of variance.

a.s.: almost surely.

BC: bias-corrected.

BC_a: accelerated bias-corrected.

BLUE: best linear unbiased estimator.

c.d.f.: cumulative distribution function.

ch.f.: characteristic function.

CLT: central limit theorem.

GEE: generalized estimation equation.

GLM: generalized linear model.

HPD: highest posterior density.

i.i.d.: independent and identically distributed.

LR: likelihood ratio.

LSE: least squares estimator.

MCMC: Markov chain Monte Carlo.

MELE: maximum empirical likelihood estimator.

m.g.f.: moment generating function.

MLE: maximum likelihood estimator.

MQLE: maximum quasi-likelihood estimator.

MRIE: minimum risk invariant estimator.

mse: mean squared error.

p.d.f.: probability density function.

RLE: root of likelihood equation.

SLLN: strong law of large numbers.

UMA: uniformly most accurate.

UMAI: uniformly most accurate invariant.

UMAU: uniformly most accurate unbiased.

UMP: uniformly most powerful.

UMPI: uniformly most powerful invariant.

UMPU: uniformly most powerful unbiased.

UMVUE: uniformly minimum variance unbiased estimator.

WLLN: weak law of large numbers.

w.r.t.: with respect to.

Appendix B

Notation

$\mathcal{R}$: the real line.

$\mathcal{R}^k$: the k-dimensional Euclidean space.

$\mathcal{B}$: the Borel σ-field on $\mathcal{R}$.

$\mathcal{B}^k$: the Borel σ-field on $\mathcal{R}^k$.

(a, b) and $[a, b]$: the open and closed intervals from a to b.

$\{a, b\}$: the set consisting of the elements a and b.

A^τ: the transpose of a matrix or a vector A.

$\mathrm{Det}(A)$: the determinant of a matrix A.

$\mathrm{tr}(A)$: the trace of a matrix A.

A^{-1}: the inverse of a matrix A.

$A^{1/2}$: the square root of a nonnegative definite matrix A defined by $A^{1/2}A^{1/2} = A$.

$A^{-1/2}$: the inverse of $A^{1/2}$.

I_k: the $k \times k$ identity matrix.

$\|x\|$: the Euclidean norm of a vector $x \in \mathcal{R}^k$, $\|x\|^2 = xx^\tau$.

I_A: the indicator function of the set A.

A^c: the complement of the set A.

$P(A)$: the probability of the set A.

$\{a_n\}$: a sequence of vectors or random vectors $a_1, a_2, \ldots$.

$a_n \to a$: $\{a_n\}$ converges to a as n increases to ∞.

$\to_{a.s.}$: convergence almost surely.

$\to_p$: convergence in probability.

$\to_d$: convergence in distribution.

g', g'', and $g^{(k)}$: the first-, second-, and kth-order derivatives of a function g on $\mathcal{R}$.

$g(x+)$ or $g(x-)$: the right or the left limit of the function g at x.

$\partial g/\partial x$ or ∇g: the partial derivative row vector of the function g on $\mathcal{R}^k$.

$\partial^2 g/\partial x \partial x^\tau$ or $\nabla^2 g$: the second-order partial derivative matrix of the function g on $\mathcal{R}^k$.

$F^{-1}(p)$: the pth quantile of a c.d.f. F, $F^{-1}(t) = \inf\{x : F(x) \geq t\}$.

$E(X)$ or EX: the expectation of a random variable (vector or matrix) X.

$\mathrm{Var}(X)$: the variance (covariance matrix) of a random variable (vector) X.

$\mathrm{Cov}(X, Y)$: the covariance between random variables X and Y.

$\mathcal{P}$: a family containing the population P that generates data

$b_T(P)$: the bias of an estimator T under population P.

$\tilde{b}_T(P)$: an asymptotic bias of an estimator T under population P.

$\mathrm{mse}_T(P)$: the mse of an estimator T under population P.

$R_T(P)$: the risk of an estimator T under population P.

$\mathrm{amse}_T(P)$: an asymptotic mse of an estimator T under population P.

$e_{T'_n,T_n}(P)$: the asymptotic relative efficiency of T'_n w.r.t. T_n.

$\alpha_T(P)$: probability of type I error for a test T.

$\beta_T(P)$: power function for a test T.

$X_{(i)}$: the ith order statistic of X_1, ..., X_n.

$\bar{X}$: the sample mean of $X_1, ..., X_n$, $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$.

S^2 : the sample variance (covariance matrix) of $X_1, ..., X_n$, $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^\tau (X_i - \bar{X})}{n-1}$.

F_n: the empirical c.d.f. based on $X_1, ..., X_n$.

$N(\mu, \sigma^2)$: the one-dimensional normal distribution or random variable with mean μ and variance σ^2.

$N_k(\mu, \Sigma)$: the k-dimensional normal distribution or random vector with mean vector μ and covariance matrix Σ.

$\Phi(x)$: the standard normal c.d.f.

z_α: the αth quantile of the standard normal distribution.

χ^2_r: a random variable having the chi-square distribution χ^2_r.

$\chi^2_{r,\alpha}$: the $(1-\alpha)$th quantile of the chi-square distribution χ^2_r.

$t_{r,\alpha}$: the $(1-\alpha)$th quantile of the t-distribution t_r.

$F_{a,b,\alpha}$: the $(1-\alpha)$th quantile of the F-distribution $F_{a,b}$.

Author Index

Arvesen, J. N., 332, 494

Bahadur, R. R., 78, 80, 250, 451, 494, 495

Barndorff-Nielsen, O. E., 493, 495

Basag, J., 212, 495

Basu, D., 493, 495

Beran, R., 463, 464, 495

Berger, J. O., 92, 198-200, 207, 208, 235, 259, 393, 432, 493, 494, 495

Berger, R. L., vii, 493, 494, 496

Bickel, P. J., 259, 312, 493, 494, 495

Billingsley, P., 1, 46, 493, 495

Blackwell, D., 493, 495

Box, G. E. P., 198, 493, 496

Brown, L. D., 217, 235, 496

Carroll, R. J., 181, 496

Casella, G., vii, 493, 494, 496

Chan, K. S., 209, 210, 496

Chen, J., 181, 286, 403, 451, 496

Chung, K. L., 1, 16, 46, 493, 496

Clarke, B. R., 301, 496

Cline, D. B. H., 181, 496

Cochran, W. G., 161, 163, 167, 168, 493, 496

Cox, D. R., 493, 495

Cramér, H., 493, 494, 496

Diggle, P. J., 494, 496

Doksum, K. A., 493, 494, 495

Draper, N. R., 148, 496

Durbin, J., 401, 496

Dvoretzky, A., 279, 497

Edwards, A. W. F., 494, 497

Efron, B., 233, 336, 454-456, 461, 494, 497

Fahrmeir, L., 494, 497

Farrell, R. H., 202, 217, 497

Ferguson, T. S., vii, 493, 494, 497

Fernholz, L. T., 293, 294, 307, 494, 497

Fox, M., 217, 496

Fuller, W. A., 247, 497

Gelfand, A. E., 209, 497

Geweke, J., 208, 498

Geyer, C. J., 240, 498

Ghosh, M., 493, 498

Godambe, V. P., 493, 494, 498

Green, P., 495

Hall, P., 336, 458-463, 465, 494, 498

Hampel, F. R., 292, 299, 494, 498

Has'minskii, R. Z., 259, 499

He, X., 324, 494, 498

Heyde, C. C., 494, 498

Higdon, D., 495

Hodges, J. L., Jr., 493, 498

Hoeffding, W., 493, 498

Hogg, R. V., 61, 498

Huber, P. J., 294, 299, 313, 494, 499

Ibragimov, I. A., 259, 499

James, W., 231, 494, 499

Jeffreys, H., 198, 499

Johnson, D. E., 472, 500

Jones, M. C., 290, 499

Kalbfleisch, J. D., 287, 499

Kaplan, E. L., 288, 499

Kaufmann, H., 494, 497

Kiefer, J., 279, 282, 497, 499

Klett, G. W., 437, 502

Kolmogorov, A. N., 1, 400, 499

Lawless, J., 282, 317, 403, 451, 501

Le Cam, L., 249, 499

Lehmann, E. L., vii, 68, 80, 98, 107, 137, 196, 223, 233, 235, 259, 310, 349, 360, 372, 379, 394-397, 408, 415, 416, 493, 494, 498, 499

Liang, K.-Y., 315, 494, 496, 499

Lindley, D. V., 493, 500

Liu, R. Y., 462, 500

Loh, W.-Y., 466, 467, 500

Martin, M. A., 465, 498

McCullagh, P., 244, 494, 500

Meeden, G., 493, 498

Meier, P., 288, 499

Mendenhall, W., 473, 500

Mengersen, K., 495

Metropolis, N., 211, 500

Milliken, G. A., 472, 500

Moore, D. S., 391, 500

Morris, C., 233, 497

Müller, H.-G., 181, 500

Natanson, I. P., 295, 500

Nelder, J. A., 244, 494, 500

Nummelin, E., 209, 212, 500

Owen, A. B., 282, 402, 451, 500

Pitman, E. J. G., 138, 493, 500

Prentice, R. T., 287, 499

Puri, M. L., 494, 501

Qin, J., 282, 286, 287, 317, 403, 451, 496, 501

Quenouille, M., 330, 501

Randles, R. H., 493, 494, 501

Rao, C. R., 155, 386, 493, 501

Rohatgi, V. K., vii, 493, 501

Ronchetti, E. M., 498

Rosenblatt, M., 494, 501

Rosenbluth, A. W., 500

Rosenbluth, M. N., 500

Rousseeuw, P. J., 498

Royden, H. L., 3, 501

Särndal, C. E., 161, 493, 501

Savage, L. J., 493, 501

Scheffé, H., 29, 80, 469, 493, 494, 501

Schervish, M. J., 493, 501

Searle, S. R., 29, 148, 493, 501

Sen, P. K., 38, 386, 493, 494, 501

Serfling, R. J., 38, 49, 145, 146, 281, 304, 308, 324, 387, 406, 493, 494, 502

Shao, J., 141, 181, 208, 301, 332, 335, 336, 494, 496, 502

Shao, Q.-M., 324, 494, 498

Shorack, G. R., 49, 278, 288, 493, 494, 502

Silverman, B. W., 290, 494, 502

Sincich, T., 473, 500

Singer, J. M., 38, 386, 493, 494, 501

Singh, K., 462, 500

Smirnov, N. V., 400, 502

Smith, A. F. M., 209, 497

Smith, H., 148, 496

Smyth, G. K., 246, 502

Spruill, M. C., 391, 500

Stadrmüller, U., 181, 500

Stahel, W. A., 498

Stein, C., 217, 231, 233, 493, 499, 502

Stone, C. J., 257, 290, 502

Strawderman, W. E., 233, 502

Swensson, B., 161, 493, 501

Tanis, E. A., 61, 498

Tanner, M. A., 208, 212, 502

Tate, R. F., 437, 502

Teller, A. H., 500

Thompson, M. E., 494, 498

Tiao, G. C., 198, 493, 496

Tibshirani, R. J., 336, 456, 494

Tierney, L., 209, 210, 212, 503

Tsui, K.-W., 235, 503

Tu, D., 141, 332, 335, 336, 494, 502

Tukey, J., 330, 472, 494, 503

Vardi, Y., 287, 503

von Mises, R., 493, 494, 503

Wahba, G., 290, 503

Wald, A., 386, 387, 493, 503

Weerahandi, S., 98, 503

Wellner, J. A., 49, 278, 288, 493, 494, 502

Welsh, A. H., 494, 503

Wolfe, D. A., 493, 494, 501

Wolfowitz, J., 279, 282, 497, 499

Woodruff, R. S., 452, 503

Wretman, J., 161, 493, 501

Wu, C. F. J., 332, 503

Yahav, J. A., 259, 495

Zeger, S. L., 315, 494, 496, 499

Subject Index

A

Absolute continuity, 14-15

Absolute error loss, 216

Absolute moment, 26

Acceptance region, 427

Action, 83; *see also* decision

Action space, 83

Admissibility, 86-89, 104, 120, 202-203, 217, 223, 226-228, 231, 233, 235, 241, 263-264, 269-271, 273

Almost every where (a.e.), 11

Almost surely (a.s.), 11

Alternative hypothesis, 85

Analysis of variance (ANOVA), 29, 151-152, 161, 187, 376-377, 468-472, 486-487

Ancillary, 79, 82, 118, 130

Approximate unbiasedness, 105-106

Asymptotic accuracy, 458-464, 467, 486

Asymptotic bias, 105-107, 125, 138, 171, 178, 191, 325

Asymptotic confidence set *or* interval, 445-451, 453, 458, 462

Asymptotic correctness: of confidence sets, 445-449, 451-454, 456-457, 481-485; of simultaneous confidence intervals, 467-468, 470, 486; of tests, 110-111, 404-406

Asymptotic covariance matrix, 248-249, 259, 325-327, 329-330, 404, 445-448

Asymptotic criteria, 102

Asymptotic efficiency, 241, 251-255, 257-260, 275-277, 281, 285, 308, 312, 315, 446-447; of tests, 406

Asymptotic expectation, 105-107, 313

Asymptotic inference, 109, 327

Asymptotic mean squared error (amse), 107-109, 125, 138-139, 146-147, 170-173, 178, 186, 191-192, 248, 309, 325

Asymptotic normality, 45, 47-49, 71-72, 104-105, 109, 114, 125, 176, 246, 248, 250, 255, 259, 278, 284, 288-289, 291-292, 302, 304, 306, 308-310, 312, 315, 321-324, 385-386, 388,

397, 400, 404, 434, 445, 452, 456, 461, 484

Asymptotic optimality, 139, 248; of tests, 387

Asymptotic relative efficiency, 108-109, 125, 178, 180, 190, 192, 224, 273-274, 309-310, 323, 340-342

Asymptotic significance level: of confidence sets, 111-112, 126, 445; of simultaneous confidence intervals, 467-468; of tests, 110-112, 126, 380, 384, 386, 389-390, 394, 401-404, 420

Asymptotic test, 386, 404, 406, 420, 447-448

Asymptotic unbiasedness, 106, 127, 170-173, 204, 249, 259

Asymptotic variance, 107, 109, 126, 181, 245, 248, 325, 335-336, 343, 397, 405, 456, 458, 484; *see also* asymptotic covariance matrix

Asymptotically pivotal quantity, 445-447, 463, 481-483

Automatic bootstrap percentile, 485

Autoregressive time series, 247-248

B

Bahadur representation, 307, 344, 451

Basu's theorem, 82, 130, 216, 218, 221-222, 363-366, 368-369

Bayes action, 195-198, 201, 207, 229, 261-263, 270, 392

Bayes estimator, 201-207, 223-225, 228-231, 241, 259-260, 263-265, 269

Bayes factor, 392-393, 419

Bayes formula, 194-195

Bayes risk, 90, 201-203, 225

Bayes rule, 90-91, 122, 124, 193, 201-202

Bayes test, 392-393, 419

Bayesian approach, 92, 193, 195, 201, 236, 392, 430

Bayesian hypothesis testing, 262, 392-393

Behrens-Fisher problem, 366

Bernoulli variable, *see* binary random variable

Best linear unbiased estimator (BLUE), 155-159, 170, 172, 179-181, 189-191, 277

Beta distribution, 20, 112, 115-116, 185, 192, 224, 261, 264, 350, 366, 411-412, 418-419, 476-477

Bias, 89, 92-93, 105-107, 121-122, 124, 126, 176-177, 202-203, 214-215, 218, 230, 263-264, 267, 325

Biased sampling, 286

Binary random variable, 92-93, 107, 119, 126, 182, 186, 224, 237, 243, 261, 263-264, 270, 348, 352, 394, 409-410, 419, 429, 442, 448, 476-477

Binomial distribution, 18, 37-38, 48, 54, 57-60, 67-69, 73, 88, 93, 121, 174, 185, 242-243, 247, 261, 271, 274, 278, 288, 305-306, 348, 350, 360-362, 382, 411, 418, 429, 477, 480

Biostatistics, 287, 314

Bivariate normal distribution, 273, 275, 368, 412-413, 416-417, 424, 481; *see also* multivariate normal distribution *and* normal distribution

Bonferroni's method *or* intervals, 468, 470-471, 473, 486-487

Bootstrap, 326, 329, 334, 405, 453, 457, 463

Bootstrap accelerated bias-corrected (BC_a) percentile, 456-457, 461-462, 464, 484-485

Bootstrap bias-corrected (BC) percentile, 455-457, 461-463

Bootstrap calibrating, 465-467

Bootstrap confidence set *or* interval, 335, 453-467, 484-485

Bootstrap data, 335, 484-486

Bootstrap distribution estimator, 335-336, 344

Bootstrap inverting, 465

Bootstrap percentile, 454-457, 460-462

Bootstrap prepivoting, 464-465

Bootstrap sample, 334, 453

Bootstrap sampling procedure, 453

Bootstrap-t, 456-460, 463-464, 484-485

Bootstrap variance estimator, 335, 344

Bootstrapping pairs, 485

Bootstrapping residuals, 485

Borel-Cantelli's lemma, 50

Borel function, 6

Borel σ-field, 2

Bounded completeness, 80, 82, 116-118, 357

Bounded in probability, 42

C

χ^2-statistic, 388-392

χ^2-test, 387, 389-392, 401

Cartesian product, 5

Cauchy distribution, 20, 23-24, 42, 54, 113, 116, 119, 125, 185, 273, 275, 308-309, 341, 351, 407, 434, 478

Cauchy-Schwarz' inequality, 331, 486

Censored data, 287, 338

Censoring times, 287

Central limit theorem (CLT), 47-49, 71-72, 103-104, 109, 111, 147, 173, 253, 257, 278, 284, 289, 292, 306, 324, 326-327, 385, 388, 405, 445

Central moment, 26, 175

Change of variables, 13

Characteristic function (ch.f.), 20, 27; properties of, 28

Chebyshev's inequality, 51

Chi-square distribution, 19-20, 23-24, 42, 45, 54, 71-72, 82-83, 100, 115, 130, 146-147, 154, 178, 192, 221, 232-234, 364, 367, 373, 379, 384, 386, 391, 401-402, 404, 412-413, 426, 433, 437, 445, 447, 451, 474, 478; *see also* noncentral chi-square distribution

Cluster, 165, 243

Cluster sampling, 165

Cochran's theorem, 29, 54, 377, 416

Comparison of two treatments, 360, 365

Completeness, 79-82, 116-118, 122, 128-133, 140, 153, 162-164, 174, 184, 190, 229, 233, 304, 357, 359, 423, 426; of order statistics, 81

Completeness of a class of decision rules, 120

Composite hypothesis, 349

Conditional distribution, 36

Conditional expectation, 30; properties of, 33

Conditional likelihood, 246-247, 274

Conditional p.d.f., 32

Conditional probability, 30; properties of, 33

Confidence band, 473-474, 488

Confidence bound, 99-100, 428-429, 438-440, 442-444, 451, 453-464, 466, 479-481, 484-486

Confidence coefficient, 99-101, 112, 123, 421-425, 427-428, 430, 432-434, 437-446, 448, 451, 453-456, 458-459, 468-469, 473, 475-481, 484-486; of simultaneous confidence intervals, 467-475, 486-488

Confidence interval, 99-100, 112, 123, 126, 422-426, 428-429, 432-441, 443, 445, 451-453, 456-462, 464, 467-469, 475-479, 481-484, 486-487; *see also* confidence bound

Confidence set, 92, 99-101, 109, 111-112, 123, 421-424, 426-427, 430-433, 438-439, 441-451, 453, 464-465, 467, 475, 477, 480, 483, 485; properties of, 434; *see also* confidence bound *and* confidence interval

Conjugate prior, 197, 230, 261

Contrast, 470-471

Consistency: of point estimators, 102-106, 123-125, 138, 147, 159, 173, 179, 180-181, 191-192, 202, 204-205, 207, 252-254, 257-259, 263-264, 273, 276, 278-279, 305, 308-309, 312, 315, 317, 320-324, 329, 340, 453, 456, 458, 463; of tests, 110-111, 126, 387, 394, 404, 406, 419-420; of variance estimators, 326-330, 332-336, 343, 404-406, 420, 445-446, 448, 457

Contingency table, 361-362, 391, 419

Continuous c.d.f., 9

Convergence almost surely, 38

Convergence in distribution *or* in law, 39

Convergence in L_p, 39

Convergence in probability, 38

Convolution, 294

Cornish-Fisher expansion, 458-459, 461

Corollary 1.1, 45; 1.2, 48; 1.3, 48-49

Corollary 3.1, 134; 3.2, 143; 3.3, 157

Corollary 4.1, 219; 4.2, 221; 4.3, 228

Corollary 5.1, 308

Corollary 6.1, 352

Corollary 7.1, 451; 7.2, 452

Correlation coefficient, 26, 114, 121, 273, 315, 325, 343, 416-417, 481

Countable set, 4

Counting measure, 3

Covariance, 26-27, 72, 113-114, 137, 170, 275, 326, 329, 417, 424, 481

Covariate, 148, 242, 314, 367, 432

Coverage probability, 99, 462, 465-467, 480; convergence speed of, 462

Cramér-Rao low bound, 135-139, 152, 180, 186, 273

Cramér-von Mises test, 400-401, 420

Cramér-Wold device, 41

Credible interval, 431

Credible set, 430-432, 437, 477-479

Critical region, *see* rejection region

Cumulative distribution function (c.d.f.), 4

D

Data reuse method, *see* resampling method

Decision, 83, 195; *see also* action

Decision rule, 83, 120, 193, 201-203, 229

Decision theory, 61-65, 83, 92-93, 95, 101-102, 104, 128, 229

Definition 1.1, 2; 1.2, 3; 1.3, 6; 1.4, 9-10; 1.5, 27; 1.6, 30; 1.7, 34; 1.8, 38; 1.9, 42

Definition 2.1, 64; 2.2, 66; 2.3, 69; 2.4, 73; 2.5, 77-78; 2.6, 80; 2.7, 86; 2.8, 89; 2.9, 89-90; 2.10, 102-103; 2.11, 105; 2.12, 108; 2.13, 110; 2.14, 111

Definition 3.1, 127; 3.2, 140; 3.3, 144; 3.4, 148

Definition 4.1, 195; 4.2, 214; 4.3, 236; 4.4, 251

Definition 5.1, 278; 5.2, 291-292; 5.3, 295; 5.4, 326

Definition 6.1, 346; 6.2, 349-350; 6.3, 356; 6.4, 356; 6.5, 369-370; 6.6, 380

Definition 7.1, 421; 7.2, 438-439; 7.3, 439-440; 7.4, 443; 7.5, 458; 7.6, 467

Delta-method, 45

DeMorgan's law, 2

Density estimation, 288, 290

Differentiability *or* differential of functionals, 291-292, 294

Dirichlet distribution, 230

Discrete c.d.f., 8

Discrete p.d.f., 15

Discrete random variable, 8

Discrete uniform distribution, 17-18, 236, 408, 417

Dispersion measure, 93

Dispersion parameter, 242-243

Distance, 278; Mallows', 280

Distribution, 8

Distribution-free test, 394

Dominated convergence theorem, 12, 41, 47, 56, 250, 280, 289, 298

Dominated family, 64

Double bootstrap, 465

Double exponential distribution, 21, 24, 113, 116, 173-174, 185-186, 261, 266, 270, 309, 340, 342, 347-348, 350-351

Dunnett's interval, 488

Dvoretzky-Kiefer-Wolfowitz's inequality, 279, 305, 344

E

Edgeworth expansion, 458, 461, 463-464, 466, 486

Empirical Bayes, 198-201, 231, 263

Empirical c.d.f., 93, 109, 125, 140, 278-285, 288, 291, 307, 311, 334, 336-337, 344, 398, 401, 446, 451, 453, 464-465, 474, 478, 485, 488

Empirical likelihood, 282, 285-286, 316, 401-402, 450; method of, 282-283, 290, 315-316, 451

Empirical likelihood ratio, 401-402, 420

Empirical likelihood ratio test, 401-403, 450

Equal-tailed confidence interval, 423, 436-437, 467

Equal-tailed test, 364

Equicontinuity, 318-320, 342

Estimability, 127-128, 150, 152-154, 162, 164, 179, 190, 229

Estimation, 84

Estimator, *see* point estimator

Euclidean space, 2

Event, 2

Example 1.1, 3; 1.2, 3; 1.3, 8; 1.4, 9; 1.5, 11; 1.6, 11; 1.7, 12; 1.8, 12-13; 1.9, 14; 1.10, 15; 1.11, 15-16; 1.12, 17; 1.13, 19; 1.14, 22-23; 1.15, 23-24; 1.16, 27; 1.17, 29; 1.18, 30-32; 1.19, 33-34; 1.20, 36; 1.21, 37-38; 1.22, 40-41; 1.23, 42-43; 1.24, 45; 1.25, 47; 1.26, 48

Example 2.1, 62-63; 2.2, 63; 2.3, 63; 2.4, 64-65; 2.5, 67; 2.6, 67; 2.7, 68; 2.8, 71-72; 2.9, 72; 2.10, 73-74; 2.11, 77; 2.12, 77; 2.13, 78; 2.14, 79; 2.15, 80-81; 2.16, 81; 2.17, 81-82; 2.18, 82-83; 2.19, 84; 2.20, 84-85; 2.21, 85; 2.22, 88; 2.23, 88-89; 2.24, 90; 2.25, 90-91; 2.26, 93-94; 2.27, 94; 2.28, 96-97; 2.29, 97-98; 2.30, 98; 2.31, 99; 2.32, 100-101; 2.33, 103; 2.34, 104; 2.35, 107; 2.36, 109; 2.37, 110-111

Example 3.1, 128; 3.2, 128-129; 3.3, 129-130; 3.4, 130-131; 3.5, 131-132; 3.6, 132; 3.7, 133-134; 3.8, 134; 3.9, 136; 3.10, 138; 3.11, 143-144; 3.12, 151; 3.13, 151; 3.14, 152; 3.15, 154; 3.16, 157; 3.17, 158; 3.18, 158-159; 3.19, 163; 3.20, 164-165; 3.21, 170-171; 3.22, 171; 3.23, 171-172; 3.24, 173-174; 3.25, 174; 3.26, 174-175; 3.27, 175; 3.28, 178; 3.29, 181; 3.30, 181

Example 4.1, 196; 4.2, 197; 4.3, 198; 4.4, 199; 4.5, 200-201; 4.6, 203; 4.7, 204; 4.8, 205; 4.9, 206-207; 4.10, 210; 4.11, 216; 4.12, 216; 4.13, 216-217; 4.14, 219; 4.15, 219; 4.16, 221; 4.17, 221-222; 4.18, 224; 4.19, 226; 4.20, 228; 4.21, 229; 4.22, 229; 4.23, 230; 4.24, 230; 4.25, 230; 4.26, 230; 4.27, 234; 4.28, 235-236; 4.29, 237; 4.30, 238-239; 4.31, 239; 4.32, 239; 4.33, 239-240; 4.34, 240; 4.35, 242-243; 4.36, 244-245; 4.37, 245; 4.38, 249; 4.39, 254, 4.40, 258

Example 5.1, 285-286; 5.2, 286-287; 5.3, 287-288; 5.4, 290; 5.5, 294; 5.6, 297; 5.7, 299-300; 5.8, 304; 5.9, 308; 5.10, 309-310; 5.11, 320; 5.12, 321; 5.13, 323; 5.14, 325; 5.15, 327; 5.16, 328

Example 6.1, 347-348; 6.2, 348; 6.3, 350; 6.4, 350-351; 6.5, 351; 6.6, 352; 6.7, 352; 6.8, 352-353; 6.9, 353; 6.10, 355; 6.11, 360-361; 6.12, 361-362; 6.13, 371; 6.14, 371-372; 6.15, 372-373; 6.16, 373; 6.17, 374; 6.18, 376-377; 6.19, 377-378; 6.20, 382; 6.21, 382-383; 6.22, 387; 6.23, 389; 6.24, 391-392; 6.25, 393; 6.26, 403; 6.27, 405-406

Example 7.1, 422-423; 7.2, 423-424; 7.3, 424; 7.4, 424; 7.5, 426; 7.6, 426; 7.7, 428-429; 7.8, 429; 7.9, 430; 7.10, 430; 7.11, 431-432; 7.12, 432-433; 7.13, 434-435; 7.14, 436-437; 7.15, 440; 7.16, 442-443; 7.17, 443; 7.18, 444; 7.19, 444; 7.20, 445; 7.21, 445-446; 7.22, 446; 7.23, 448-449; 7.24, 449-450; 7.25, 462-463; 7.26, 468-469; 7.27, 470-471; 7.28, 473; 7.29, 474

Expectation *or* expected value, 11, 26; *see also* mean

Explanatory variable, *see* covariate

Exponential distribution, 9, 20, 40, 54, 65, 70-71, 94, 112-115, 118-119, 123, 182-183, 185, 192, 204, 216, 261, 267-268, 270, 272-274, 350-351, 408, 413-415, 417-418, 476, 479-481, 483

Exponential family, 66-68, 76, 79-82, 112-114, 116, 131, 137-138, 149, 227, 235, 241-242, 247, 254, 260, 350, 352, 357-359,

362, 382, 428-429; canonical form for, 66; full rank of, 67 natural parameter *and* natural parameter space, 66

External bootstrapping, 486

F

F-distribution, 21, 23-24, 365, 373, 375-376, 379, 416, 424, 430, 469-470, 476; *see also* noncentral F-distribution

F-test, 365-366

Factor, 377-378; interaction of, 378

Factorial experiment, 152

Factorization theorem, 74, 162

Fatou's lemma, 12, 56

Fieller's interval, 424

Figure 2.1, 95; 2.2, 97; 2.3, 101

Figure 4.1, 225

Figure 5.1, 290

Figure 7.1, 428; 7.2, 450

First-order ancillary, 79

First-stage sampling, 168

Fisher's exact test, 362, 392

Fisher information, 135-137, 185, 188, 227, 249, 251, 257, 384-387, 448

Fisher-scoring, 240, 245, 258, 272, 276

Fréchet differentiability, 292-298, 301, 336, 338-339

Frequentist approach, 193, 201

Fubini's theorem, 13, 32, 34, 52, 75, 194-195, 441

G

Gamma distribution, 20, 29, 65, 71, 112-115, 118, 130, 185, 192, 196-197, 204-205, 210, 235, 239, 261, 264, 268, 350, 410-411, 479

Gâteaux differentiability, 291-293, 327-328, 332, 338, 341

Gauss-Markov's theorem, 155

Generalized Bayes, 197-199, 201-202, 204, 235, 262, 264, 270

Generalized estimating equations (GEE), 312-315, 317, 320-321, 342, 403, 483

Generalized estimating equations (GEE) estimator, 313, 317, 320-321, 323-324, 329, 333-334, 343

Generalized inverse, 149

Generalized linear model (GLM), 241-243, 245-246, 254, 258, 274, 314, 321, 387

Generating function, 25

Geometric distribution, 18, 117, 262, 269, 410

Gibbs sampler, 209-210, 265

Gini's mean difference, 141, 144, 147, 297

Goodness of fit test, 389, 391, 401

Group of transformations, 369

H

Hadamard differentiability, 292-296, 300-302, 307, 327-328, 332, 336, 338-340, 397, 446

Hierarchical Bayes, 199-201, 263

Highest posterior density (HPD), 431, 437, 477-479

Histogram, 61, 289

Hodges-Lehmann's estimator, 302, 304, 340, 405-406

Hoeffding's theorem, 142

Hölder's inequality, 281, 319

Horvitz-Thompson estimator, 165, 167-169, 190-191, 277, 286

Hybrid bootstrap, 456-457, 459-460, 462-464, 484-485

Hypergeometric distribution, 18, 112, 239, 269, 351, 361-362

Hyperparameter, 198-200

Hypothesis testing, 84, 89, 92, 95, 99-100, 110, 122, 231, 302, 345, 441

I

Importance function, 208

Importance sampling, 207, 212

Improper prior, 197-198, 200-201, 259, 262, 264, 269, 393, 431-432

Independence, 22, 34

Independence chain, 212

Independent and identically distributed (i.i.d.), 45

Indicator function, 6

Induced likelihood function, 271

Induced measure, 8

Inference, 61-65, 92, 95, 101, 195, 235, 325-326

Information inequality, 136, 227

Influence function, 292-293, 300, 302, 304, 307, 309, 327-328, 338-339, 341, 446

Integrability, 11

Integral, 9

Integration, 9; by parts, 52

Interquartile range, 308

Interval estimator *or* set estimator, 92, 99; *see also* confidence set

Invariance, 89, 92, 122, 213; in linear models, 222-223, 268; in location problems, 213-217; in location-scale problems, 89-90, 219-223; in scale problems, 217-219; in testing hypothesis, 369, 372, 374, 396; of confidence sets, 443

Invariant confidence set, 443-444, 481

Invariant decision problem, 90, 121-122, 230, 268

Invariant decision rule, 90, 121, 218

Invariant family, 89, 121, 218, 222, 268, 369, 443

Invariant estimator, 213-223, 230, 257, 266-268, 273

Invariant test, 122, 356, 370, 373-374, 444, 481

Invariant testing problem, 369, 375, 379, 414-416, 481

Inverse gamma distribution, 261-262, 265

Inverse image, 6

Iterative bootstrap, 465

J

Jackknife, 141, 147, 326, 329-334, 344, 405

Jacobian, 19, 23

James-Stein's estimator, 231, 233-234

Jensen's inequality, 53, 88, 281

Joint c.d.f., 6

Joint p.d.f., 22

K

Kaplan-Meier's estimator, 288

Kernel density estimator, 288, 290

Kolmogorov-Smirnov statistic, 399

Kolmogorov-Smirnov test, 399-401, 420, 474

Kurtosis, 462

L

L-functional, 296-297, 328, 339

L-estimator, 294, 296-298, 304, 310-312, 328, 332, 339, 341

L_p distance, 280-281

Lagrange multiplier, 282-283, 286, 316

Law, *see* distribution

Least absolute deviation estimator, *see* minimum L_p distance estimator

Least squares estimator (LSE), 148-160, 170-172, 179-181, 187, 189-190, 223, 229, 234, 244-245, 268-269, 274, 276-277, 311-315, 328, 332, 368, 376-377, 382, 420, 424, 430, 433, 440, 469, 485-486

Lebesgue integral, 11

Lebesgue measure, 3

Lebesgue p.d.f., 16

Lemma 3.1, 144; 3.2, 146; 3.3, 161

Lemma 4.1, 205; 4.2, 212; 4.3, 226

Lemma 5.1, 279; 5.2, 294-295; 5.3, 318

Lemma 6.1, 349; 6.2, 349; 6.3, 350; 6.4, 354-355; 6.5, 357; 6.6, 357; 6.7, 363

Length of a confidence interval, 100, 123, 434-438, 452-453, 467, 473, 478-481

Level of significance: in hypothesis testing, 96-100, 110, 123, 345-346, 356, 370, 395, 417-418, 427; of a confidence set, 99-100, 123, 421-422, 425, 427-431, 445, 447, 468, 473, 476-477; of a prediction set, 432-433, 478; of simultaneous confidence intervals, 467-469, 474, 486, 488

Lévy-Cramér continuity theorem, 41

Liapunov's condition, 48, 60, 324, 338

Life-time testing, 63-65, 93, 287, 314

Likelihood, 92, 386; method of, 281, 447-448

Likelihood function, 236, 239-241, 243, 245-247, 259, 285, 342, 349, 380, 382, 393, 431, 447, 449

Likelihood equation, 237-241, 243, 246, 272-276, 314, 316, 320-321

Likelihood ratio, 380, 383, 385, 401, 448

Likelihood ratio (LR) test, 380-387, 390, 392, 417-418, 420, 447-450, 477, 480-483

Limiting confidence coefficient, 111-112, 126, 445, 451-452, 474-475, 482, 488

Limiting size, 110-112, 394, 397-398, 400-401, 404, 406, 420

Lindeberg's CLT, 47, 322

Lindeberg's condition, 48, 60, 322, 344

Linear function of order statistics, 304-305; *see also* L-estimator

Linear model, 148, 179, 206, 222, 229, 234, 241-242, 268-269, 274, 276, 311, 314, 328, 332, 367, 374, 382, 412, 415, 420, 424, 430, 433, 440, 446, 469, 482, 485; with random coefficients, 157

Link function, 242-245, 274

Location family, 69-70, 90, 114, 121, 213, 226, 235, 267, 273, 434

Location-scale family, 69-70, 89, 114, 136, 188, 213, 217, 219, 257, 369, 371, 414, 422, 484, 488

Log-distribution, 18, 184, 192, 350

Log-likelihood equation, *see* likelihood equation

Log-normal distribution, 21, 65, 113, 115, 192, 272

Logistic distribution, 21, 173-174, 185, 272, 275-276, 309, 340, 351, 479

Longitudinal data, 314

Loss function, 83-90, 92-93, 95, 120-122, 128, 195, 197, 213-215, 217-224, 226, 228-231, 233-235, 241, 261-263, 266-270; convexity of, 87-88, 94, 102, 128, 195, 215-216, 218, 222-223, 226, 229, 266; invariance of, 90, 217-218, 220, 267; *see also* absolute error loss, squared error loss, *and* 0-1 loss

M

M-functional, 298, 299-300, 339

M-estimator, 294, 298, 301, 313, 315, 321, 323, 328, 332, 342-343

Marcinkiewicz SLLN, 46

Marginal p.d.f. *or* distribution, 22, 195, 198-200

Markov chain, 208, 240; properties of, 208-209, 265

Markov chain Monte Carlo (MCMC), 207-212, 240

Maximal invariant, 370-375, 379, 396-397, 414, 417, 419

Maximum empirical likelihood estimator (MELE), 282-283, 285-287, 338, 401-402

Maximum likelihood estimator (MLE), 236-241, 243-247, 252, 254, 257, 260, 271-276, 299, 308, 313, 315-316, 342, 381, 384, 386-387, 390-391, 447-449

Maximum profile likelihood estimator, 316, 342

Maximum quasi-likelihood estimator (MQLE), 246, 274, 313, 315, 343

Maximum likelihood method, 235

Mean, 26-27, 61, 71, 84, 88, 92, 103, 112, 114, 121-122, 124-126, 132, 143-144, 148, 158, 164, 170-172, 184, 188, 191, 231, 288, 405, 416-417, 445-446, 481, 484; *see also* expectation *and* expected value

Mean absolute error, 93

Mean squared error (mse), 93-95, 107, 109, 121-122, 125, 127-128, 138-139, 143, 164, 170, 176, 189, 225, 241, 273, 288, 325; consistency in, 103-104, 124, 143, 160

Measurable function, 6

Measurable space, 2

Measure, 2-3; monotonicity of, 4; subadditivity of, 4; continuity of, 4

Measure space, 3

Measurement problem, 62, 64, 84

Median, 61, 122, 261, 266, 308, 356, 402, 405, 455

Metric, *see* distance

Metropolis algorithm, 211-212

Minimal completeness, 120

Minimal sufficiency, 77-80, 116, 118, 304

Minimax estimator, 223-231, 233, 235, 269-271

Minimaxity, 90-91

Minimum L_p distance estimator, 299

Minimum risk invariant estimator (MRIE), 214-223, 226, 228, 230-231, 257, 266-268, 270, 308

Mixture distribution, 240, 306

Moment, 25; method of, 173, 175, 193, 199, 263, 342

Moment estimator, 173-176, 192, 199

Moment generating function (m.g.f.), 18, 20-21, 27; properties of, 28

Monotone convergence theorem, 12, 56

Monotone likelihood ratio, 349-351, 353-354, 373-374, 379, 407-408, 415, 425-426, 439, 442, 473, 480

Monte Carlo, 207-208, 212, 335, 454, 464, 466, 486

Multinomial distribution, 68, 184, 230, 362, 387-388, 391, 419

Multiple comparison, 468-470

Multivariate normal distribution *or* p.d.f., 22, 27, 54, 64, 70, 113, 149, 172, 210, 231-232, 259, 261, 268, 274, 367, 374-375, 385, 424, 430, 433, 469; *see also* asymptotic normality *and* normal distribution

N

Nearest neighbor method, 290

Negative binomial distribution, 18, 54, 57, 112-113, 115, 183, 192, 264, 270, 350, 382, 411, 418, 477, 483

Negative part, 10

Newton-Raphson method, 240, 245, 257, 276, 334

Neyman structure, 357-358

Neyman-Pearson's lemma, 251, 346, 349

Nominal level, 465-466

Noncentral chi-square distribution, 24-25, 29, 54; *see also* chi-square distribution

Noncentral F-distribution, 24-25, 54, 375, 415; *see also* F-distribution

Noncentral t-distribution, 24-25, 54, 364, 374, 410, 415, 444; *see also* t-distribution

Noncentrality parameter, 24

Noninformative prior, 197-198, 200, 210, 261, 432

Nonlinear regression, 245, 314

Nonparametric family, 65

Nonparametric likelihood function, 281, 402

Nonparametric maximum likelihood estimator, 282

Nonparametric method, 65

Nonparametric model, 65

Nonparametric test, 394

Norm, 291-295

Normal distribution *or* p.d.f., 19-20, 45, 49, 58, 64, 67, 71-72, 80, 82, 91, 96, 100, 111, 114, 116-119, 121-122, 125-126, 130, 138, 151, 154-155, 157-158, 160, 173-174, 180, 182, 185-186, 191, 198-201, 203, 205-206, 216, 221, 225-226, 228, 238, 242, 244-245, 247, 249, 261-263, 265, 269, 272-275, 306, 308-309, 315, 340, 342, 350, 352, 355, 362-363, 365, 372-374, 376-378, 382, 390, 393, 395, 407-408, 410, 412, 414, 416-418, 423-424, 431, 433-434, 436, 438, 440, 444, 446, 449, 468, 470-471, 474-478, 480, 482-483, 487-488; *see also* asymptotic normality, bivariate normal distribution, multivariate normal distribution, *and* standard normal distribution

Normalizing and variance stabilizing transformation, 454

Nuisance parameter, 242, 246

Null hypothesis, 85

O

One-sample problem, 363, 383, 396, 477

One-sample t-test, 364, 367-368, 383, 405, 410, 414, 420

One-sample Wilcoxon statistic, 141, 143, 147

One-sided confidence interval, *see* confidence bound

One-sided hypothesis, 351, 368, 399, 406, 409-410

One-step MLE, 257, 276, 333

Optimality in risk, 84, 86, 88-90, 95, 120-121, 170, 193

Order statistics, 72, 77, 81, 84, 104, 114-116, 122, 124, 132-133, 140, 162, 213, 216, 218, 222, 239, 287, 304, 308, 311, 335, 350, 353, 371, 399, 408-409, 413, 423, 425, 435, 437, 439, 451-452, 476, 479, 484; p.d.f. of, 72; sufficiency of, 77

Outcome, 1

Over-dispersion, 243

P

p-value, 97-98, 123, 393, 407, 410

Pairwise independence, 35-36, 57

Parameter space, 64

Parametric bootstrapping, 484

Parametric family, 64, 193; identifiability of, 64, 149

Parametric method, 65

Parametric model, 64, 193

Pareto distribution, 21, 115, 175, 183, 185, 261, 267, 274, 409, 417

Partition, 7

Permutation test, 395-397

Pitman's estimator, 215, 217-218, 226, 267

Pivotal quantity, 421-426, 430, 432-438, 445, 456, 475-476, 478

Point estimator, 92, 99, 102-105, 127, 173, 193, 248, 325, 434-435, 445, 447

Point mass, 17

Poisson distribution, 18, 52, 57-60, 109, 112, 115, 121, 128-129, 182-183, 185-186, 196-197, 229, 235, 245, 247, 261, 270, 274, 350, 352-353, 360-361, 382, 410-411, 418-419, 426, 429, 480, 482

Pólya's theorem, 39, 306

Polynomial regression, 151, 171, 187

Population, 61

Positive part, 10

Posterior distribution *or* p.d.f., 193-197, 200, 204-207, 210, 230, 236, 259, 261-264, 270, 392-393, 431, 478

Power function, 345, 348, 351, 353-354, 356, 359-360, 364, 372, 380, 406, 409-410, 412, 419, 442

Power series distribution, 113, 131

Power set, 2

Pratt's theorem, 441

Prediction, 33, 91, 188, 432

Prediction interval *or* set, 432-433, 478

Predictor, 33, 91, 432

Prior distribution *or* p.d.f., 193-207, 210, 223-225, 228, 230-231, 236, 259-261, 263-265, 269, 392-393, 419, 431, 477-478

Probability density function (p.d.f.), 15

Probability measure, 3

Probability space, 3

Product-limit estimator, 288

Product measure, 5

Product σ-field, 5

Product space, 5

Profile likelihood, 316, 342

Profile empirical likelihood, 316, 402-403

Profile empirical likelihood ratio test, 402-403, 420, 450

Projection, 51, 144-145, 186; method of, 144-145

Projection matrix, 367, 385, 388-389

Proportional allocation, 165

Proposition 1.1, 4; 1.2, 4; 1.3, 5; 1.4, 7-8; 1.5, 11; 1.6, 11; 1.7, 16; 1.8, 19; 1.9, 27; 1.10, 28; 1.11, 32; 1.12, 33; 1.13, 34; 1.14, 35; 1.15, 36-37; 1.16, 39; 1.17, 41

Proposition 2.1, 80; 2.2, 87; 2.3, 105-106; 2.4, 108

Proposition 3.1, 136; 3.2, 137; 3.3, 138; 3.4, 158; 3.5, 177

Proposition 4.1, 195; 4.2, 203; 4.3, 213; 4.4, 214; 4.5, 218; 4.6, 220

Proposition 5.1, 296; 5.2, 317; 5.3, 317; 5.4, 318; 5.5, 320; 5.6, 321

Proposition 6.1, 349; 6.2, 370; 6.3, 372; 6.4, 373; 6.5, 381

Proposition 7.1, 424; 7.2, 427; 7.3, 444; 7.4, 446

Pseudo-likelihood equation, 315

Q

Quantile, 291, 294, 304, 341, 451

Quasi-likelihood, 246, 314, 315, 320-321, 325

R

R-estimator, 294, 302, 304, 340

Radon-Nikodym derivative *or* density, 15, 75; properties of, 16

Radon-Nikodym theorem, 14-15

Random censorship model, 287

Random effect, 158, 188, 190, 378-379, 417

Random element, 6

Random experiment, 1

Random variable, 6

Random vector, 6

Random walk chain, 212

Randomized confidence set, 442-444, 448, 480

Randomized decision rule, 86-87, 98, 195; risk of, 86

Randomized estimator, 119, 268

Randomized test, 98, 345, 347-348, 352, 381, 427, 441-442, 444

Rank statistics, 294, 301-302, 372, 396-397, 426

Rank test, 396-397

Rao's score test, *see* score test

Rao-Blackwell's theorem, 88, 94, 128, 133, 229

Ratio estimator, 170-171, 343

Regression M-estimator, 313

Rejection region, 85, 97

Repeated measurements, 314

Replication method, *see* resampling method

Resampling estimator, 335

Resampling method, 329

Residual, 154, 207, 311, 485

Riemann integral, 11, 13

Risk, 83-93, 95, 119, 193, 202-203, 214-215, 218, 220-221, 223-226, 230-231, 233-234, 266, 269, 271

Robustness, 155-156, 159, 277, 292-293, 298, 300-302, 304, 308-310, 312, 323

Root of the likelihood equation (RLE), 252-254, 257, 273, 275-276, 313, 386-387, 447-448

S

σ-field, 2; generated by a collection of sets, 2; generated by a measurable function, 6

σ-finite measure, 5

Sample, 62

Sample central moment, 176, 277, 344

Sample correlation coefficient, 114, 126, 171, 191, 369, 413, 416, 481

Sample covariance matrix, 326, 330, 334-335, 424, 458

Sample mean, 62, 71-72, 81-82, 84, 88, 91-92, 94, 96, 99-100, 103, 105, 107, 114, 118-120, 125, 140, 151, 170, 172, 198-199, 203, 207, 213, 221, 224-225, 297, 299, 308-310, 323, 340-342, 365, 369, 377, 405, 422-423, 445, 458, 475, 480, 484, 488

Sample median, 309-310, 334-335, 341, 405, 484

Sample moment, 140, 173, 277, 291, 294

Sample quantile, 291, 294, 304-307, 340, 344, 351; distribution of, 305

Sample size, 62

Sample space, 1

Sample standard deviation, 217-218

Sample variance, 62, 71-72, 81-82, 100, 103, 114, 118, 121, 125-126, 141, 164, 176, 187, 191, 207, 217, 221, 326, 344, 364-365, 405, 423-424, 475, 480, 484; *see also* sample covariance matrix

Scale family, 69-70, 114, 213, 217, 273, 423

Scheffé's method *or* intervals, 469-471, 473, 486-487

Scheffé's theorem, 41, 114, 182, 341

Score function, 254-255, 259, 419

Score test, 386-387, 418, 448-450, 482-483

Scoring, 254

Second-stage sampling, 168

Shortest-length confidence interval, 434-435, 437-438, 479

Shrinkage estimator, 231, 233-235

Sign test, 394-395, 419

Signed rank statistic, 301, 430; one-sample Wilcoxon's, 301-302, 406, 419-420

Signed rank test, 396-398; Wilcoxon's, 396

Significance level, *see* level of significance

Similar test, 356-358

Simple function, 7

Simple hypothesis, 346, 349

Simple linear regression, 151, 161, 187, 470

Simple random sampling, 63, 162-165, 167-169, 190-191, 285

Simultaneous confidence intervals, 467-475, 486-488

Simultaneous estimation, 229-231, 235

Single-stage sampling, 167

Size, 96, 98-99, 110-111, 123, 345-349, 351-354, 356-359, 370, 373-376, 379-380, 383, 385-386, 394-395, 397, 400-401, 406-419, 427-429, 439-442, 444, 481

Skewness, 462

Slutsky's theorem, 43, 72, 103-105, 147, 179, 253, 257, 260, 292, 452

Smooth splines, 290

Squared error loss, 84, 88, 90-91, 93-94, 118-121, 128, 196, 198-199, 203-207, 215-218, 224-229, 259, 261, 263-267, 269-271

Standard deviation, 26

Standard normal distribution *or* p.d.f., 22-24, 41-42, 49, 58, 96, 172, 191, 243, 261, 268, 306, 309, 322, 347-348, 366, 393,

407, 431, 437, 452, 454, 458, 478; *see also* asymptotic normality *and* normal distribution

Statistic, 70; distribution of, 71

Statistical computing, 207

Statistical decision theory, *see* decision theory

Statistical functional, 291, 294, 296, 307, 327, 445

Statistical inference, *see* inference

Statistical model, 64

Stepwise function, 8

Stochastic order, 42

Stratified sampling, 163-165, 169, 190-191

Strong law of large numbers (SLLN), 45-47, 72, 103, 147, 173, 176, 182, 207, 252, 254, 280, 284, 327-328, 330-331, 335, 405

Studentized range, 471

Substitution, 92, 173, 179; in variance estimation, 326-329

Sufficiency, 73-82, 86-88, 92, 94, 96, 99-100, 115-117, 122, 128-133, 140, 153, 162-164, 174, 184, 190, 229, 233, 246-247, 261, 267, 271, 304, 346, 356-357, 359, 372-375, 392, 394, 396-397, 409, 422-423, 426, 434, 479

Sup-norm distance, 278-281

Superefficiency, 251

Survey, 37, 63, 65, 94, 161, 285, 332

Survival analysis, 287

Survival distribution, 287

Survival times, 287

Symmetric c.d.f. *or* p.d.f., 24

Systematic sampling, 168-169, 191

T

t-distribution, 21, 23-24, 41, 58, 123, 185, 262, 264, 309-310, 340, 364, 367-369, 383, 412-413, 423-424, 433, 436, 440, 468-469, 480; *see also* noncentral t-distribution

t-type confidence interval, 473, 487

Table 1.1, 18; 1.2, 20-21

Test, 85, 95, 109, 112, 119, 345, 427, 434-435, 477

Testing independence, 362, 368, 391, 412, 419

Theorem 1.1, 12; 1.2, 13; 1.3, 13-14; 1.4, 14-15; 1.5, 29; 1.6, 30; 1.7, 36; 1.8, 39-40; 1.9, 41; 1.10, 42; 1.11, 43; 1.12, 43-44; 1.13, 45-46; 1.14, 46; 1.15, 47

Theorem 2.1, 68; 2.2, 74; 2.3, 78; 2.4, 82; 2.5, 87-88; 2.6, 106; 2.7, 109

Theorem 3.1, 128; 3.2, 132; 3.3, 135; 3.4, 142; 3.5, 146; 3.6, 150; 3.7, 152; 3.8, 153; 3.9, 155; 3.10, 156; 3.11, 159-160; 3.12, 160; 3.13, 162; 3.14, 164; 3.15, 165-166; 3.16, 178; 3.17, 179

Theorem 4.1, 194; 4.2, 202; 4.3, 202; 4.4, 208-209; 4.5, 215; 4.6, 215, 4.7, 218; 4.8, 218; 4.9, 221; 4.10, 223; 4.11, 223; 4.12, 225; 4.13, 226; 4.14, 227; 4.15, 231; 4.16, 249-250; 4.17, 252; 4.18, 254-255; 4.19, 257-258; 4.20, 259

Theorem 5.1, 279; 5.2, 280; 5.3, 282; 5.4, 284; 5.5, 293; 5.6, 297; 5.7, 300; 5.8, 302; 5.9, 305; 5.10, 306; 5.11, 307; 5.12, 312; 5.13, 321; 5.14, 323; 5.15, 328; 5.16, 329; 5.17, 330; 5.18, 332; 5.19, 334; 5.20, 336

Theorem 6.1, 346; 6.2, 351; 6.3, 353-354; 6.4, 358-359; 6.5, 384; 6.6, 386; 6.7, 387; 6.8, 388; 6.9, 390; 6.10, 399-400; 6.11, 402; 6.12, 404

Theorem 7.1, 425; 7.2, 427; 7.3, 435-436; 7.4, 439; 7.5, 440; 7.6, 441; 7.7, 443-444; 7.8, 451; 7.9, 457; 7.10, 469; 7.11, 472

Trimmed sample mean, 297, 299, 310-311, 341-342, 405

Truncation family, 77, 113, 274

Tukey's method *or* intervals, 471-473, 487

Tukey's model, 309

Two-sample linear rank statistic, 302

Two-sample problem, 222, 365, 383, 396, 401, 412, 477

Two-sample rank test, 397-398, 420; Wilcoxon's, 397

Two-sample t-test, 367-368, 383, 395, 397, 415, 420

Two-sided hypothesis, 353, 364, 410

Two-stage sampling, 165, 168, 191

Two-way additive model, 187, 416

Type I error, 95, 98, 110, 345, 351

Type II error, 95, 98, 110, 345, 406

U

U-statistic, 140-147, 170, 176-178, 186-187, 192, 277, 328, 332, 343, 482

Unbiased confidence set, 440-441, 443-444, 480-481

Unbiased estimator, 89, 92-94, 105-107, 109, 121-122, 124, 127-129, 132-136, 138, 140, 150, 153, 155-156, 159, 161-162, 164-165, 169-171, 173, 179, 181, 191-192, 215-216, 223, 225, 229-230, 267, 278, 388

Unbiased test, 356-357, 359, 369, 373-374, 395, 410, 415, 440, 444, 481

Unbiasedness: in estimation, 89, 92, 105, 127, 203-204; in hypothesis testing, 356, 369, 374, 396; of confidence set, 440

Uncorrelated random variables, 27

Uniform distribution, 9, 20, 60, 67, 70, 78, 81, 118, 122, 126, 128, 133-134, 144, 174, 182-183, 185, 192, 198, 216, 221, 239, 261, 268, 270, 273, 304, 341, 350-351, 353, 382, 399, 408,

410, 413-414, 420, 423, 425, 434, 442, 452, 464, 477, 480, 484, 488

Uniform integrability, 40, 58, 109, 125-126, 139, 146-147, 187

Uniformly minimum risk unbiased estimator, 128

Uniformly minimum variance unbiased estimator (UMVUE), 127-140, 152-154, 156-158, 162-164, 170, 174, 179, 182-188, 190, 192, 204, 215, 223-224, 229, 231, 241, 264, 270, 273-274, 277-278, 308

Uniformly most accurate (UMA) confidence set, 438-439, 441-444, 479-481

Uniformly most accurate invariant (UMAI) confidence set, 443-444, 481

Uniformly most accurate unbiased (UMAU) confidence set, 440-444, 468, 479-481, 486

Uniformly most powerful invariant (UMPI) test, 370-381, 383, 394, 414-417, 419, 430, 444

Uniformly most powerful (UMP) test, 346-357, 359-360, 369-372, 374, 380-382, 392, 394-396, 406-409, 415, 428-429, 438-439, 442, 444, 481

Uniformly most powerful unbiased (UMPU) test, 356-360, 362-369, 373-374, 376, 380-381, 392, 394-395, 410-413, 415, 417-418, 428-429, 438-440, 442, 477

Unimodality, 435-437, 479

V

V-statistic, 176-178, 192, 277, 295, 400

Variance, 18, 20-21, 26-27, 71-72, 84, 88, 93-94, 103, 113-114, 122, 124-127, 132, 135, 137-140, 142-145, 155, 158-159, 166-172, 177, 186, 190-191, 214, 231, 266, 278, 288, 298, 309-310, 325, 335, 339, 395, 405, 416, 424, 433, 446, 484

Variance estimation, 325-326, 329, 404

Variance estimator, 141, 167, 326-334, 343-344, 457, 463-464; *see also* bootstrap variance estimator *and* jackknife

Volume of a confidence set, 440-441

W

Wald's test, 386-387, 418-419, 448-450, 482-483

Watson-Royall's theorem, 162

Weak law of large numbers (WLLN), 45-47, 333

Weibull distribution, 21, 65, 113, 115, 258, 273, 418

Weighted jackknife variance estimator, 332

Weighted least squares estimator, 179-181

Wild bootstrapping, 486

Winsorized sample mean, 299

With replacement, 112, 122, 285

Without replacement, 63, 163, 165, 190, 285

Working correlation matrix, 315, 320

Woodruff's interval, 405, 452-453, 484

Z

0-1 loss, 85, 89, 95, 119

Springer Texts in Statistics *(continued from page ii)*

Noether: Introduction to Statistics: The Nonparametric Way
Peters: Counting for Something: Statistical Principles and Personalities
Pfeiffer: Probability for Applications
Pitman: Probability
Rawlings, Pantula and Dickey: Applied Regression Analysis
Robert: The Bayesian Choice: A Decision-Theoretic Motivation
Santner and Duffy: The Statistical Analysis of Discrete Data
Saville and Wood: Statistical Methods: The Geometric Approach
Sen and Srivastava: Regression Analysis: Theory, Methods, and Applications
Shao: Mathematical Statistics
Whittle: Probability via Expectation, Third Edition
Zacks: Introduction to Reliability Analysis: Probability Models and Statistical Methods