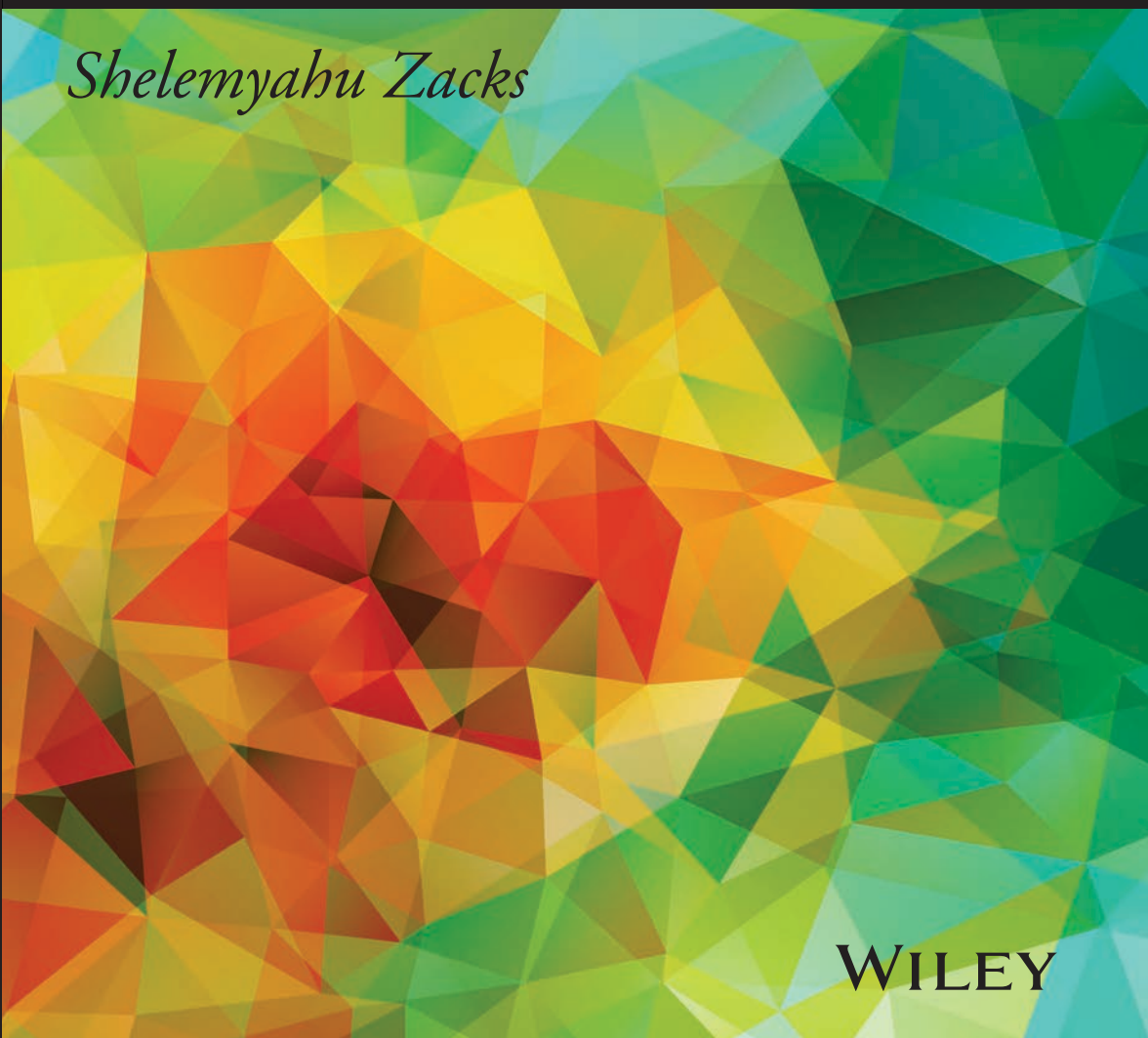
Wiley Series in Probability and Statistics

Examples and Problems in Mathematical Statistics



Examples and Problems in Mathematical Statistics

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

Examples and Problems in Mathematical Statistics

SHELEMYAHU ZACKS

Department of Mathematical Sciences
Binghamton University
Binghamton, NY

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Zacks, Shelemyahu, 1932- author.
Examples and problems in mathematical statistics / Shelemyahu Zacks.
pages cm
Summary: “This book presents examples that illustrate the theory of mathematical statistics and details how to apply the methods for solving problems” – Provided by publisher.
Includes bibliographical references and index.
ISBN 978-1-118-60550-9 (hardback)
1. Mathematical statistics–Problems, exercises, etc. I. Title.
QC32.Z265 2013
519.5–dc23

2013034492

Printed in the United States of America

ISBN: 9781118605509

10 9 8 7 6 5 4 3 2 1

To my wife Hanna,
our sons Yuval and David,
and their families, with love.

Contents

Preface **xv**

List of Random Variables **xvii**

List of Abbreviations **xix**

1 Basic Probability Theory **1**

PART I: THEORY, 1

1.1 Operations on Sets, 1
1.2 Algebra and σ-Fields, 2
1.3 Probability Spaces, 4
1.4 Conditional Probabilities and Independence, 6
1.5 Random Variables and Their Distributions, 8
1.6 The Lebesgue and Stieltjes Integrals, 12
1.6.1 General Definition of Expected Value: The Lebesgue Integral, 12
1.6.2 The Stieltjes–Riemann Integral, 17
1.6.3 Mixtures of Discrete and Absolutely Continuous Distributions, 19
1.6.4 Quantiles of Distributions, 19
1.6.5 Transformations, 20
1.7 Joint Distributions, Conditional Distributions and Independence, 21
1.7.1 Joint Distributions, 21
1.7.2 Conditional Expectations: General Definition, 23
1.7.3 Independence, 26
1.8 Moments and Related Functionals, 26
1.9 Modes of Convergence, 35
1.10 Weak Convergence, 39
1.11 Laws of Large Numbers, 41

1.11.1 The Weak Law of Large Numbers (WLLN), 41
1.11.2 The Strong Law of Large Numbers (SLLN), 42
1.12 Central Limit Theorem, 44
1.13 Miscellaneous Results, 47
1.13.1 Law of the Iterated Logarithm, 48
1.13.2 Uniform Integrability, 48
1.13.3 Inequalities, 52
1.13.4 The Delta Method, 53
1.13.5 The Symbols o_p and O_p, 55
1.13.6 The Empirical Distribution and Sample Quantiles, 55
PART II: EXAMPLES, 56
PART III: PROBLEMS, 73
PART IV: SOLUTIONS TO SELECTED PROBLEMS, 93

2 Statistical Distributions 106

PART I: THEORY, 106
2.1 Introductory Remarks, 106
2.2 Families of Discrete Distributions, 106
2.2.1 Binomial Distributions, 106
2.2.2 Hypergeometric Distributions, 107
2.2.3 Poisson Distributions, 108
2.2.4 Geometric, Pascal, and Negative Binomial Distributions, 108
2.3 Some Families of Continuous Distributions, 109
2.3.1 Rectangular Distributions, 109
2.3.2 Beta Distributions, 111
2.3.3 Gamma Distributions, 111
2.3.4 Weibull and Extreme Value Distributions, 112
2.3.5 Normal Distributions, 113
2.3.6 Normal Approximations, 114
2.4 Transformations, 118
2.4.1 One-to-One Transformations of Several Variables, 118
2.4.2 Distribution of Sums, 118
2.4.3 Distribution of Ratios, 118
2.5 Variances and Covariances of Sample Moments, 120
2.6 Discrete Multivariate Distributions, 122
2.6.1 The Multinomial Distribution, 122
2.6.2 Multivariate Negative Binomial, 123
2.6.3 Multivariate Hypergeometric Distributions, 124

2.7 Multinormal Distributions, 125
2.7.1 Basic Theory, 125
2.7.2 Distribution of Subvectors and Distributions of Linear Forms, 127
2.7.3 Independence of Linear Forms, 129
2.8 Distributions of Symmetric Quadratic Forms of Normal Variables, 130
2.9 Independence of Linear and Quadratic Forms of Normal Variables, 132
2.10 The Order Statistics, 133
2.11 t-Distributions, 135
2.12 F-Distributions, 138
2.13 The Distribution of the Sample Correlation, 142
2.14 Exponential Type Families, 144
2.15 Approximating the Distribution of the Sample Mean: Edgeworth and Saddlepoint Approximations, 146
2.15.1 Edgeworth Expansion, 147
2.15.2 Saddlepoint Approximation, 149
PART II: EXAMPLES, 150
PART III: PROBLEMS, 167
PART IV: SOLUTIONS TO SELECTED PROBLEMS, 181

3 Sufficient Statistics and the Information in Samples 191

PART I: THEORY, 191
3.1 Introduction, 191
3.2 Definition and Characterization of Sufficient Statistics, 192
3.2.1 Introductory Discussion, 192
3.2.2 Theoretical Formulation, 194
3.3 Likelihood Functions and Minimal Sufficient Statistics, 200
3.4 Sufficient Statistics and Exponential Type Families, 202
3.5 Sufficiency and Completeness, 203
3.6 Sufficiency and Ancillarity, 205
3.7 Information Functions and Sufficiency, 206
3.7.1 The Fisher Information, 206
3.7.2 The Kullback–Leibler Information, 210
3.8 The Fisher Information Matrix, 212
3.9 Sensitivity to Changes in Parameters, 214
3.9.1 The Hellinger Distance, 214
PART II: EXAMPLES, 216
PART III: PROBLEMS, 230
PART IV: SOLUTIONS TO SELECTED PROBLEMS, 236

4 Testing Statistical Hypotheses 246

PART I: THEORY, 246

4.1 The General Framework, 246

4.2 The Neyman–Pearson Fundamental Lemma, 248

4.3 Testing One-Sided Composite Hypotheses in MLR Models, 251

4.4 Testing Two-Sided Hypotheses in One-Parameter Exponential Families, 254

4.5 Testing Composite Hypotheses with Nuisance Parameters—Unbiased Tests, 256

4.6 Likelihood Ratio Tests, 260

4.6.1 Testing in Normal Regression Theory, 261

4.6.2 Comparison of Normal Means: The Analysis of Variance, 265

4.7 The Analysis of Contingency Tables, 271

4.7.1 The Structure of Multi-Way Contingency Tables and the Statistical Model, 271

4.7.2 Testing the Significance of Association, 271

4.7.3 The Analysis of 2×2 Tables, 273

4.7.4 Likelihood Ratio Tests for Categorical Data, 274

4.8 Sequential Testing of Hypotheses, 275

4.8.1 The Wald Sequential Probability Ratio Test, 276

PART II: EXAMPLES, 283

PART III: PROBLEMS, 298

PART IV: SOLUTIONS TO SELECTED PROBLEMS, 307

5 Statistical Estimation 321

PART I: THEORY, 321

5.1 General Discussion, 321

5.2 Unbiased Estimators, 322

5.2.1 General Definition and Example, 322

5.2.2 Minimum Variance Unbiased Estimators, 322

5.2.3 The Cramér–Rao Lower Bound for the One-Parameter Case, 323

5.2.4 Extension of the Cramér–Rao Inequality to Multiparameter Cases, 326

5.2.5 General Inequalities of the Cramér–Rao Type, 327

5.3 The Efficiency of Unbiased Estimators in Regular Cases, 328

5.4 Best Linear Unbiased and Least-Squares Estimators, 331

5.4.1 BLUEs of the Mean, 331

5.4.2 Least-Squares and BLUEs in Linear Models, 332

5.4.3 Best Linear Combinations of Order Statistics, 334

5.5 Stabilizing the LSE: Ridge Regressions, 335
5.6 Maximum Likelihood Estimators, 337
5.6.1 Definition and Examples, 337
5.6.2 MLEs in Exponential Type Families, 338
5.6.3 The Invariance Principle, 338
5.6.4 MLE of the Parameters of Tolerance Distributions, 339
5.7 Equivariant Estimators, 341
5.7.1 The Structure of Equivariant Estimators, 341
5.7.2 Minimum MSE Equivariant Estimators, 343
5.7.3 Minimum Risk Equivariant Estimators, 343
5.7.4 The Pitman Estimators, 344
5.8 Estimating Equations, 346
5.8.1 Moment-Equations Estimators, 346
5.8.2 General Theory of Estimating Functions, 347
5.9 Pretest Estimators, 349
5.10 Robust Estimation of the Location and Scale Parameters of Symmetric Distributions, 349
PART II: EXAMPLES, 353
PART III: PROBLEMS, 381
PART IV: SOLUTIONS OF SELECTED PROBLEMS, 393

6 Confidence and Tolerance Intervals 406

PART I: THEORY, 406
6.1 General Introduction, 406
6.2 The Construction of Confidence Intervals, 407
6.3 Optimal Confidence Intervals, 408
6.4 Tolerance Intervals, 410
6.5 Distribution Free Confidence and Tolerance Intervals, 412
6.6 Simultaneous Confidence Intervals, 414
6.7 Two-Stage and Sequential Sampling for Fixed Width Confidence Intervals, 417
PART II: EXAMPLES, 421
PART III: PROBLEMS, 429
PART IV: SOLUTION TO SELECTED PROBLEMS, 433

7 Large Sample Theory for Estimation and Testing 439

PART I: THEORY, 439
7.1 Consistency of Estimators and Tests, 439
7.2 Consistency of the MLE, 440

7.3 Asymptotic Normality and Efficiency of Consistent Estimators, 442
7.4 Second-Order Efficiency of BAN Estimators, 444
7.5 Large Sample Confidence Intervals, 445
7.6 Edgeworth and Saddlepoint Approximations to the Distribution of the MLE: One-Parameter Canonical Exponential Families, 446
7.7 Large Sample Tests, 448
7.8 Pitman's Asymptotic Efficiency of Tests, 449
7.9 Asymptotic Properties of Sample Quantiles, 451
PART II: EXAMPLES, 454
PART III: PROBLEMS, 475
PART IV: SOLUTION OF SELECTED PROBLEMS, 479

8 Bayesian Analysis in Testing and Estimation 485

PART I: THEORY, 485
8.1 The Bayesian Framework, 486
8.1.1 Prior, Posterior, and Predictive Distributions, 486
8.1.2 Noninformative and Improper Prior Distributions, 487
8.1.3 Risk Functions and Bayes Procedures, 489
8.2 Bayesian Testing of Hypothesis, 491
8.2.1 Testing Simple Hypothesis, 491
8.2.2 Testing Composite Hypotheses, 493
8.2.3 Bayes Sequential Testing of Hypotheses, 495
8.3 Bayesian Credibility and Prediction Intervals, 501
8.3.1 Credibility Intervals, 501
8.3.2 Prediction Intervals, 501
8.4 Bayesian Estimation, 502
8.4.1 General Discussion and Examples, 502
8.4.2 Hierarchical Models, 502
8.4.3 The Normal Dynamic Linear Model, 504
8.5 Approximation Methods, 506
8.5.1 Analytical Approximations, 506
8.5.2 Numerical Approximations, 508
8.6 Empirical Bayes Estimators, 513
PART II: EXAMPLES, 514
PART III: PROBLEMS, 549
PART IV: SOLUTIONS OF SELECTED PROBLEMS, 557

9 Advanced Topics in Estimation Theory 563

PART I: THEORY, 563
9.1 Minimax Estimators, 563

9.2 Minimum Risk Equivariant, Bayes Equivariant, and Structural Estimators, 565
9.2.1 Formal Bayes Estimators for Invariant Priors, 566
9.2.2 Equivariant Estimators Based on Structural Distributions, 568
9.3 The Admissibility of Estimators, 570
9.3.1 Some Basic Results, 570
9.3.2 The Inadmissibility of Some Commonly Used Estimators, 575
9.3.3 Minimax and Admissible Estimators of the Location Parameter, 582
9.3.4 The Relationship of Empirical Bayes and Stein-Type Estimators of the Location Parameter in the Normal Case, 584
PART II: EXAMPLES, 585
PART III: PROBLEMS, 592
PART IV: SOLUTIONS OF SELECTED PROBLEMS, 596

References **601**

Author Index **613**

Subject Index **617**

Preface

I have been teaching probability and mathematical statistics to graduate students for close to 50 years. In my career I realized that the most difficult task for students is solving problems. Bright students can generally grasp the theory easier than apply it. In order to overcome this hurdle, I used to write examples of solutions to problems and hand it to my students. I often wrote examples for the students based on my published research. Over the years I have accumulated a large number of such examples and problems. This book is aimed at sharing these examples and problems with the population of students, researchers, and teachers.

The book consists of nine chapters. Each chapter has four parts. The first part contains a short presentation of the theory. This is required especially for establishing notation and to provide a quick overview of the important results and references. The second part consists of examples. The examples follow the theoretical presentation. The third part consists of problems for solution, arranged by the corresponding sections of the theory part. The fourth part presents solutions to some selected problems. The solutions are generally not as detailed as the examples, but as such these are examples of solutions. I tried to demonstrate how to apply known results in order to solve problems elegantly. All together there are in the book 167 examples and 431 problems.

The emphasis in the book is on statistical inference. The first chapter on probability is especially important for students who have not had a course on advanced probability. Chapter Two is on the theory of distribution functions. This is basic to all developments in the book, and from my experience, it is important for all students to master this calculus of distributions. The chapter covers multivariate distributions, especially the multivariate normal; conditional distributions; techniques of determining variances and covariances of sample moments; the theory of exponential families; Edgeworth expansions and saddle-point approximations; and more. Chapter Three covers the theory of sufficient statistics, completeness of families of distributions, and the information in samples. In particular, it presents the Fisher information, the Kullback–Leibler information, and the Hellinger distance. Chapter Four provides a strong foundation in the theory of testing statistical hypotheses. The Wald SPRT is

discussed there too. Chapter Five is focused on optimal point estimation of different kinds. Pitman estimators and equivariant estimators are also discussed. Chapter Six covers problems of efficient confidence intervals, in particular the problem of determining fixed-width confidence intervals by two-stage or sequential sampling. Chapter Seven covers techniques of large sample approximations, useful in estimation and testing. Chapter Eight is devoted to Bayesian analysis, including empirical Bayes theory. It highlights computational approximations by numerical analysis and simulations. Finally, Chapter Nine presents a few more advanced topics, such as minimaxity, admissibility, structural distributions, and the Stein-type estimators.

I would like to acknowledge with gratitude the contributions of my many ex-students, who toiled through these examples and problems and gave me their important feedback. In particular, I am very grateful and indebted to my colleagues, Professors A. Schick, Q. Yu, S. De, and A. Polunchenko, who carefully read parts of this book and provided important comments. Mrs. Marge Pratt skillfully typed several drafts of this book with patience and grace. To her I extend my heartfelt thanks. Finally, I would like to thank my wife Hanna for giving me the conditions and encouragement to do research and engage in scholarly writing.

SHELEMYAHU ZACKS

List of Random Variables

$B(n, p)$	Binomial, with parameters n and p
$E(\mu)$	Exponential with parameter μ
$EV(\lambda, \alpha)$	Extreme value with parameters λ and α
$F(v_1, v_2)$	Central F with parameters v_1 and v_2
$F(n_1, n_2; \lambda)$	Noncentral F with parameters v_1, v_2, λ
$G(\lambda, p)$	Gamma with parameters λ and p
$H(M, N, n)$	Hyper-geometric with parameters M, N, n
$N(\mu, V)$	Multinormal with mean vector μ and covariance matrix V
$N(\mu, \sigma)$	Normal with mean μ and σ
$NB(\psi, v)$	Negative-binomial with parameters ψ, and v
$P(\lambda)$	Poisson with parameter λ
$R(a, b)$	Rectangular (uniform) with parameters a and b
$t[n; \lambda]$	Noncentral Student's t with parameters n and λ
$t[n; \xi, V]$	Multivariate t with parameters n, ξ and V
$t[n]$	Student's t with n degrees of freedom
$W(\lambda, \alpha)$	Weibul with parameters λ and α
$\beta(p, q)$	Beta with parameters p and q
$\chi^2[n, \lambda]$	Noncentral chi-squared with parameters n and λ
$\chi^2[n]$	Chi-squared with n degrees of freedom

List of Abbreviations

a.s.	Almost surely
ANOVA	Analysis of variance
c.d.f.	Cumulative distribution function
$\text{cov}(x, y)$	Covariance of X and Y
CI	Confidence interval
CLT	Central limit theorem
CP	Coverage probability
CR	Cramer Rao regularity conditions
$E\{X \mid Y\}$	Conditional expected value of X, given Y
$E\{X\}$	Expected value of X
FIM	Fisher information matrix
i.i.d.	Independent identically distributed
LBUE	Linear best unbiased estimate
LCL	Lower confidence limit
m.g.f.	Moment generating function
m.s.s.	Minimal sufficient statistics
MEE	Moments equations estimator
MLE	Maximum likelihood estimator
MLR	Monotone likelihood ratio
MP	Most powerful
MSE	Mean squared error
MVU	Minimum variance unbiased
OC	Operating characteristic
p.d.f.	Probability density function
p.g.f.	Probability generating function
$P\{E \mid A\}$	Conditional probability of E, given A
$P\{E\}$	Probability of E
PTE	Pre-test estimator
r.v.	Random variable
RHS	Right-hand side
s.v.	Stopping variable

SE	Standard error
SLLN	Strong law of large numbers
SPRT	Sequential probability ratio test
tr$\{A\}$	trace of the matrix A
UCL	Upper control limit
UMP	Uniformly most powerful
UMPI	Uniformly most powerful invariant
UMPU	Uniformly most powerful unbiased
UMVU	Uniformly minimum variance unbiased
$V\{X \mid Y\}$	Conditional variance of X, given Y
$V\{X\}$	Variance of X
w.r.t.	With respect to
WLLN	Weak law of large numbers

CHAPTER 1

Basic Probability Theory

PART I: THEORY

It is assumed that the reader has had a course in elementary probability. In this chapter we discuss more advanced material, which is required for further developments.

1.1 OPERATIONS ON SETS

Let $\mathcal{S}$ denote a **sample space**. Let E_1, E_2 be subsets of $\mathcal{S}$. We denote the **union** by $E_1 \cup E_2$ and the intersection by $E_1 \cap E_2$. $\bar{E} = \mathcal{S} - E$ denotes the **complement** of E. By DeMorgan's laws $\overline{E_1 \cup E_2} = \bar{E}_1 \cap \bar{E}_2$ and $\overline{E_1 \cap E_2} = \bar{E}_1 \cup \bar{E}_2$.

Given a sequence of sets $\{E_n, n \geq 1\}$ (finite or infinite), we define

$$\sup_{n\geq 1} E_n = \bigcup_{n\geq 1} E_n, \quad \inf_{n\geq 1} E_n = \bigcap_{n\geq 1} E_n. \tag{1.1.1}$$

Furthermore, $\liminf_{n\to\infty}$ and $\limsup_{n\to\infty}$ are defined as

$$\liminf_{n\to\infty} E_n = \bigcup_{n\geq 1}\bigcap_{k\geq n} E_k, \quad \limsup_{n\to\infty} E_n = \bigcap_{n\geq 1}\bigcup_{k\geq n} E_k. \tag{1.1.2}$$

If a point of $\mathcal{S}$ belongs to $\limsup_{n\to\infty} E_n$, it belongs to infinitely many sets E_n. The sets $\liminf_{n\to\infty} E_n$ and $\limsup_{n\to\infty} E_n$ always exist and

$$\liminf_{n\to\infty} E_n \subset \limsup_{n\to\infty} E_n. \tag{1.1.3}$$

Examples and Problems in Mathematical Statistics, First Edition. Shelemyahu Zacks.

If $\liminf_{n\to\infty} E_n = \limsup_{n\to\infty} E_n$, we say that a limit of $\{E_n, n \geq 1\}$ exists. In this case,

$$\lim_{n\to\infty} E_n = \liminf_{n\to\infty} E_n = \limsup_{n\to\infty} E_n. \tag{1.1.4}$$

A sequence $\{E_n, n \geq 1\}$ is called **monotone increasing** if $E_n \subset E_{n+1}$ for all $n \geq 1$. In this case $\lim_{n\to\infty} E_n = \bigcup_{n=1}^{\infty} E_n$. The sequence is **monotone decreasing** if $E_n \supset E_{n+1}$, for all $n \geq 1$. In this case $\lim_{n\to\infty} E_n = \bigcap_{n=1}^{\infty} E_n$. We conclude this section with the definition of a **partition** of the sample space. A collection of sets $\mathcal{D} = \{E_1, \ldots, E_k\}$ is called a finite **partition** of $\mathcal{S}$ if all elements of $\mathcal{D}$ are **pairwise disjoint** and their union is $\mathcal{S}$, i.e., $E_i \cap E_j = \emptyset$ for all $i \neq j$; $E_i, E_j \in \mathcal{D}$; and $\bigcup_{i=1}^{k} E_i = \mathcal{S}$. If $\mathcal{D}$ contains a countable number of sets that are mutually exclusive and $\bigcup_{i=1}^{\infty} E_i = \mathcal{S}$, we say that $\mathcal{D}$ is a countable partition.

1.2 ALGEBRA AND σ-FIELDS

Let $\mathcal{S}$ be a sample space. An algebra $\mathcal{A}$ is a collection of subsets of $\mathcal{S}$ satisfying

$$\begin{aligned} &\textbf{(i)} && \mathcal{S} \in \mathcal{A}; \\ &\textbf{(ii)} && \text{if } E \in \mathcal{A} \text{ then } \bar{E} \in \mathcal{A}; \\ &\textbf{(iii)} && \text{if } E_1, E_2 \in \mathcal{A} \text{ then } E_1 \cup E_2 \in \mathcal{A}. \end{aligned} \tag{1.2.1}$$

We consider $\emptyset = \bar{\mathcal{S}}$. Thus, (i) and (ii) imply that $\emptyset \in \mathcal{A}$. Also, if $E_1, E_2 \in \mathcal{A}$ then $E_1 \cap E_2 \in \mathcal{A}$.

The **trivial algebra** is $\mathcal{A}_0 = \{\emptyset, \mathcal{S}\}$. An algebra $\mathcal{A}_1$ is a subalgebra of $\mathcal{A}_2$ if all sets of $\mathcal{A}_1$ are contained in $\mathcal{A}_2$. We denote this inclusion by $\mathcal{A}_1 \subset \mathcal{A}_2$. Thus, the trivial algebra $\mathcal{A}_0$ is a subalgebra of every algebra $\mathcal{A}$. We will denote by $\mathcal{A}(\mathcal{S})$, the algebra generated by all subsets of $\mathcal{S}$ (see Example 1.1).

If a sample space $\mathcal{S}$ has a finite number of points n, say $1 \leq n < \infty$, then the collection of all subsets of $\mathcal{S}$ is called the **discrete algebra** generated by the elementary events of $\mathcal{S}$. It contains 2^n events.

Let $\mathcal{D}$ be a partition of $\mathcal{S}$ having k, $2 \leq k$, disjoint sets. Then, the algebra generated by $\mathcal{D}$, $\mathcal{A}(\mathcal{D})$, is the algebra containing all the $2^k - 1$ unions of the elements of $\mathcal{D}$ and the empty set.

An algebra on $\mathcal{S}$ is called a σ-**field** if, in addition to being an algebra, the following holds.

(iv) If $E_n \in \mathcal{A}$, $n \geq 1$, then $\bigcup_{n=1}^{\infty} E_n \in \mathcal{A}$.

We will denote a σ-field by $\mathcal{F}$. In a σ-field $\mathcal{F}$ the supremum, infinum, limsup, and liminf of any sequence of events belong to $\mathcal{F}$. If $\mathcal{S}$ is finite, the discrete algebra $\mathcal{A}(\mathcal{S})$ is a σ-field. In Example 1.3 we show an algebra that is not a σ-field.

The minimal σ-field containing the algebra generated by $\{(-\infty, x], -\infty < x < \infty\}$ is called the **Borel** σ-**field** on the real line $\mathbb{R}$.

A sample space $\mathcal{S}$, with a σ-field $\mathcal{F}$, $(\mathcal{S}, \mathcal{F})$ is called a **measurable space**.

The following lemmas establish the existence of smallest σ-field containing a given collection of sets.

Lemma 1.2.1. *Let $\mathcal{E}$ be a collection of subsets of a sample space $\mathcal{S}$. Then, there exists a smallest σ-field $\mathcal{F}(\mathcal{E})$, containing the elements of $\mathcal{E}$.*

Proof. The algebra of all subsets of $\mathcal{S}$, $\mathcal{A}(\mathcal{S})$ obviously contains all elements of $\mathcal{E}$. Similarly, the σ-field $\mathcal{F}$ containing all subsets of $\mathcal{S}$, contains all elements of $\mathcal{E}$. Define the σ-field $\mathcal{F}(\mathcal{E})$ to be the **intersection** of all σ-fields, which contain all elements of $\mathcal{E}$. Obviously, $\mathcal{F}(\mathcal{E})$ is an algebra. QED

A collection $\mathcal{M}$ of subsets of $\mathcal{S}$ is called a **monotonic class** if the limit of any monotone sequence in $\mathcal{M}$ belongs to $\mathcal{M}$.

If $\mathcal{E}$ is a collection of subsets of $\mathcal{S}$, let $\mathcal{M}^*(\mathcal{E})$ denote the smallest monotonic class containing $\mathcal{E}$.

Lemma 1.2.2. *A necessary and sufficient condition of an algebra $\mathcal{A}$ to be a σ-field is that it is a monotonic class.*

Proof. (**i**) Obviously, if $\mathcal{A}$ is a σ-field, it is a monotonic class.

(**ii**) Let $\mathcal{A}$ be a monotonic class.

Let $E_n \in \mathcal{A}, n \geq 1$. Define $B_n = \bigcup_{i=1}^{n} E_i$. Obviously $B_n \subset B_{n+1}$ for all $n \geq 1$. Hence $\lim_{n\to\infty} B_n = \bigcup_{n=1}^{\infty} B_n \in \mathcal{A}$. But $\bigcup_{n=1}^{\infty} B_n = \bigcup_{n=1}^{\infty} E_n$. Thus, $\sup_{n\geq 1} E_n \in \mathcal{A}$. Similarly, $\bigcap_{n=1}^{\infty} E_n \in \mathcal{A}$. Thus, $\mathcal{A}$ is a σ-field. QED

Theorem 1.2.1. *Let $\mathcal{A}$ be an algebra. Then $\mathcal{M}^*(\mathcal{A}) = \mathcal{F}(\mathcal{A})$, where $\mathcal{F}(\mathcal{A})$ is the smallest σ-field containing $\mathcal{A}$.*

Proof. See Shiryayev (1984, p. 139).

The measurable space $(\mathbb{R}, \mathcal{B})$, where $\mathbb{R}$ is the real line and $\mathcal{B} = \mathcal{F}(\mathbb{R})$, called the **Borel measurable space**, plays a most important role in the theory of statistics. Another important measurable space is $(\mathbb{R}^n, \mathcal{B}^n)$, $n \geq 2$, where $\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}$ is the Euclidean n-space, and $\mathcal{B}^n = \mathcal{B} \times \cdots \times \mathcal{B}$ is the smallest σ-field containing $\mathbb{R}^n$, $\emptyset$, and all n-dimensional rectangles $I = I_1 \times \cdots \times I_n$, where

$$I_i = (a_i, b_i], \quad i = 1, \ldots, n, \quad -\infty < a_i < b_i < \infty.$$

The measurable space $(\mathbb{R}^\infty, \mathcal{B}^\infty)$ is used as a basis for probability models of experiments with infinitely many trials. $\mathbb{R}^\infty$ is the space of ordered sequences $\mathbf{x} = (x_1, x_2, \ldots)$, $-\infty < x_n < \infty$, $n = 1, 2, \ldots$. Consider the cylinder sets

$$\mathcal{C}(I_1 \times \cdots \times I_n) = \{\mathbf{x} : x_i \in I_i,\ i = 1, \ldots, n\}$$

and

$$\mathcal{C}(B_1 \times \cdots \times B_n) = \{\mathbf{x} : x_i \in B_i,\ i = 1, \ldots, n\}$$

where B_i are Borel sets, i.e., $B_i \in \mathcal{B}$. The smallest σ-field containing all these cylinder sets, $n \geq 1$, is $\mathcal{B}(\mathbb{R}^\infty)$. Examples of Borel sets in $\mathcal{B}(\mathbb{R}^\infty)$ are

(a) $\{\mathbf{x} : \mathbf{x} \in \mathbb{R}^\infty, \sup_{n \geq 1} x_n > a\}$

or

(b) $\{\mathbf{x} : \mathbf{x} \in \mathbb{R}^\infty, \limsup_{n \to \infty} x_n \leq a\}$.

1.3 PROBABILITY SPACES

Given a measurable space $(\mathcal{S}, \mathcal{F})$, a **probability model** ascribes a countably additive function P on $\mathcal{F}$, which assigns a probability $P\{A\}$ to all sets $A \in \mathcal{F}$. This function should satisfy the following properties.

$$\begin{aligned} &\text{(A.1)} \quad \text{If } A \in \mathcal{F} \text{ then } 0 \leq P\{A\} \leq 1. \\ &\text{(A.2)} \quad P\{\mathcal{S}\} = 1. \qquad (1.3.1) \\ &\text{(A.3)} \quad \text{If } \{E_n, n \geq 1\} \in \mathcal{F} \text{ is a sequence of } \textbf{disjoint} \\ &\qquad \text{sets in } \mathcal{F}, \text{ then } P\left\{\bigcup_{n=1}^{\infty} E_n\right\} = \sum_{n=1}^{\infty} P\{E_n\}. \qquad (1.3.2) \end{aligned}$$

Recall that if $A \subset B$ then $P\{A\} \leq P\{B\}$, and $P\{\bar{A}\} = 1 - P\{A\}$. Other properties will be given in the examples and problems. In the sequel we often write AB for $A \cap B$.

Theorem 1.3.1. *Let $(\mathcal{S}, \mathcal{F}, P)$ be a probability space, where $\mathcal{F}$ is a σ-field of subsets of $\mathcal{S}$ and P a probability function. Then*

(i) *if $B_n \subset B_{n+1}$, $n \geq 1$, $B_n \in \mathcal{F}$, then*

$$P\left\{\lim_{n\to\infty} B_n\right\} = \lim_{n\to\infty} P\{B_n\}. \tag{1.3.3}$$

(ii) *if $B_n \supset B_{n+1}$, $n \geq 1$, $B_n \in \mathcal{F}$, then*

$$P\left\{\lim_{n\to\infty} B_n\right\} = \lim_{n\to\infty} P\{B_n\}. \tag{1.3.4}$$

Proof. (**i**) Since $B_n \subset B_{n+1}$, $\lim_{n\to\infty} B_n = \bigcup_{n=1}^{\infty} B_n$. Moreover,

$$P\left\{\bigcup_{n=1}^{\infty} B_n\right\} = P\{B_1\} + \sum_{n=2}^{\infty} P\{B_n - B_{n-1}\}. \tag{1.3.5}$$

Notice that for $n \geq 2$, since $\bar{B}_n B_{n-1} = \emptyset$,

$$\begin{aligned} P\{B_n - B_{n-1}\} &= P\{B_n \bar{B}_{n-1}\} \\ &= P\{B_n\} - P\{B_n B_{n-1}\} = P\{B_n\} - P\{B_{n-1}\}. \end{aligned} \tag{1.3.6}$$

Also, in (1.3.5)

$$\begin{aligned} P\{B_1\} + \sum_{n=2}^{\infty} P\{B_n - B_{n-1}\} &= \lim_{N\to\infty} \left(P\{B_1\} + \sum_{n=2}^{N} (P\{B_n\} - P\{B_{n-1}\}) \right) \\ &= \lim_{N\to\infty} P\{B_N\}. \end{aligned} \tag{1.3.7}$$

Thus, Equation (1.3.3) is proven.

(**ii**) Since $B_n \supset B_{n+1}$, $n \geq 1$, $\bar{B}_n \subset \bar{B}_{n+1}$, $n \geq 1$. $\lim_{n\to\infty} B_n = \bigcap_{n=1}^{\infty} B_n$. Hence,

$$\begin{aligned} P\left(\lim_{n\to\infty} B_n\right) &= 1 - P\left\{\overline{\bigcap_{n=1}^{\infty} B_n}\right\} \\ &= 1 - P\left\{\bigcup_{n=1}^{\infty} \bar{B}_n\right\} \\ &= 1 - \lim_{n\to\infty} P\{\bar{B}_n\} = \lim_{n\to\infty} P\{B_n\}. \end{aligned}$$

QED

Sets in a probability space are called events.

1.4 CONDITIONAL PROBABILITIES AND INDEPENDENCE

The conditional probability of an event $A \in \mathcal{F}$ given an event $B \in \mathcal{F}$ such that $P\{B\} > 0$, is defined as

$$P\{A \mid B\} = \frac{P\{A \cap B\}}{P\{B\}}. \tag{1.4.1}$$

We see first that $P\{\cdot \mid B\}$ is a probability function on $\mathcal{F}$. Indeed, for every $A \in \mathcal{F}$, $0 \leq P\{A \mid B\} \leq 1$. Moreover, $P\{\mathcal{S} \mid B\} = 1$ and if A_1 and A_2 are disjoint events in $\mathcal{F}$, then

$$\begin{aligned} P\{A_1 \cup A_2 \mid B\} &= \frac{P\{(A_1 \cup A_2)B\}}{P\{B\}} \\ &= \frac{P\{A_1 B\} + P\{A_2 B\}}{P\{B\}} = P\{A_1 \mid B\} + P\{A_2 \mid B\}. \end{aligned} \tag{1.4.2}$$

If $P\{B\} > 0$ and $P\{A\} \neq P\{A \mid B\}$, we say that the events A and B are **dependent**. On the other hand, if $P\{A\} = P\{A \mid B\}$ we say that A and B are **independent** events. Notice that two events are independent if and only if

$$P\{AB\} = P\{A\}P\{B\}. \tag{1.4.3}$$

Given n events in $\mathcal{A}$, namely $A_1, \ldots, A_n$, we say that they are **pairwise** independent if $P\{A_i A_j\} = P\{A_i\}P\{A_j\}$ for any $i \neq j$. The events are said to be independent in triplets if

$$P\{A_i A_j A_k\} = P\{A_i\}P\{A_j\}P\{A_k\}$$

for any $i \neq j \neq k$. Example 1.4 shows that pairwise independence does not imply independence in triplets.

Given n events $A_1, \ldots, A_n$ of $\mathcal{F}$, we say that they are **independent** if, for any $2 \leq k \leq n$ and any k-tuple $(1 \leq i_1 < i_2 < \cdots < i_k \leq n)$,

$$P\left\{\bigcap_{j=1}^{k} A_{i_j}\right\} = \prod_{j=1}^{k} P\{A_{i_j}\}. \tag{1.4.4}$$

Events in an infinite sequence $\{A_1, A_2, \ldots\}$ are said to be **independent** if $\{A_1, \ldots, A_n\}$ are independent, for each $n \geq 2$. Given a sequence of events $A_1, A_2, \ldots$ of a σ-field $\mathcal{F}$, we have seen that

$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k.$$

This event means that points w in $\limsup_{n\to\infty} A_n$ belong to infinitely many of the events $\{A_n\}$. Thus, the event $\limsup_{n\to\infty} A_n$ is denoted also as $\{A_n, \text{i.o.}\}$, where i.o. stands for "infinitely often."

The following important theorem, known as the **Borel–Cantelli Lemma**, gives conditions under which $P\{A_n, \text{i.o.}\}$ is either 0 or 1.

Theorem 1.4.1 (Borel–Cantelli). *Let $\{A_n\}$ be a sequence of sets in $\mathcal{F}$.*

(i) *If $\sum_{n=1}^{\infty} P\{A_n\} < \infty$, then $P\{A_n, \text{i.o.}\} = 0$.*

(ii) *If $\sum_{n=1}^{\infty} P\{A_n\} = \infty$ and $\{A_n\}$ are independent, then $P\{A_n, \text{i.o.}\} = 1$.*

Proof. (**i**) Notice that $B_n = \bigcup_{k=n}^{\infty} A_k$ is a decreasing sequence. Thus

$$P\{A_n, \text{i.o.}\} = P\left\{\bigcap_{n=1}^{\infty} B_n\right\} = \lim_{n\to\infty} P\{B_n\}.$$

But

$$P\{B_n\} = P\left\{\bigcup_{k=n}^{\infty} A_k\right\} \leq \sum_{k=n}^{\infty} P\{A_k\}.$$

The assumption that $\sum_{n=1}^{\infty} P\{A_n\} < \infty$ implies that $\lim_{n\to\infty} \sum_{k=n}^{\infty} P\{A_k\} = 0$.

(**ii**) Since $A_1, A_2, \ldots$ are independent, $\bar{A}_1, \bar{A}_2, \ldots$ are independent. This implies that

$$P\left\{\bigcap_{k=1}^{\infty} \bar{A}_k\right\} = \prod_{k=1}^{\infty} P\{\bar{A}_k\} = \prod_{k=1}^{\infty}(1 - P\{A_k\}).$$

If $0 < x \leq 1$ then $\log(1 - x) \leq -x$. Thus,

$$\begin{aligned}\log \prod_{k=1}^{\infty}(1 - P\{A_k\}) &= \sum_{k=1}^{\infty} \log(1 - P\{A_k\}) \\ &\leq -\sum_{k=1}^{\infty} P\{A_k\} = -\infty\end{aligned}$$

since $\sum_{n=1}^{\infty} P\{A_n\} = \infty$. Thus $P\left\{\bigcap_{k=1}^{\infty} \bar{A}_k\right\} = 0$ for all $n \geq 1$. This implies that $P\{A_n, \text{i.o.}\} = 1$. QED

We conclude this section with the celebrated **Bayes Theorem**.

Let $\mathcal{D} = \{B_i, i \in J\}$ be a partition of $\mathcal{S}$, where J is an index set having a finite or countable number of elements. Let $B_j \in \mathcal{F}$ and $P\{B_j\} > 0$ for all $j \in J$. Let $A \in \mathcal{F}$, $P\{A\} > 0$. We are interested in the conditional probabilities $P\{B_j \mid A\}$, $j \in J$.

Theorem 1.4.2 (Bayes).

$$P\{B_j \mid A\} = \frac{P\{B_j\}P\{A \mid B_j\}}{\sum_{j' \in J} P\{B_{j'}\}P\{A \mid B_{j'}\}}. \tag{1.4.5}$$

Proof. Left as an exercise. QED

Bayes Theorem is widely used in scientific inference. Examples of the application of Bayes Theorem are given in many elementary books. Advanced examples of Bayesian inference will be given in later chapters.

1.5 RANDOM VARIABLES AND THEIR DISTRIBUTIONS

Random variables are finite real value functions on the sample space $\mathcal{S}$, such that measurable subsets of $\mathcal{F}$ are mapped into Borel sets on the real line and thus can be

assigned probability measures. The situation is simple if $\mathcal{S}$ contains only a finite or countably infinite number of points.

In the general case, $\mathcal{S}$ might contain non-countable infinitely many points. Even if $\mathcal{S}$ is the space of all infinite binary sequences $w = (i_1, i_2, \ldots)$, the number of points in $\mathcal{S}$ is non-countable. To make our theory rich enough, we will require that the probability space will be $(\mathcal{S}, \mathcal{F}, P)$, where $\mathcal{F}$ is a σ-field. A random variable X is a finite real value function on $\mathcal{S}$. We wish to define the distribution function of X, on $\mathbb{R}$, as

$$F_X(x) = P\{w : X(w) \leq x\}. \tag{1.5.1}$$

For this purpose, we must require that every Borel set on $\mathbb{R}$ has a measurable inverse image with respect to $\mathcal{F}$. More specifically, given $(\mathcal{S}, \mathcal{F}, P)$, let $(\mathbb{R}, \mathcal{B})$ be Borel measurable space where $\mathbb{R}$ is the real line and $\mathcal{B}$ the Borel σ-field of subsets of $\mathbb{R}$. A subset of $(\mathbb{R}, B)$ is called a Borel set if B belongs to $\mathcal{B}$. Let $X : \mathcal{S} \to \mathbb{R}$. The inverse image of a Borel set B with respect to X is

$$X^{-1}(B) = \{w : X(w) \in B\}. \tag{1.5.2}$$

A function $X : \mathcal{S} \to \mathbb{R}$ is called $\mathcal{F}$-measurable if $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}$. Thus, a **random variable with respect to** $(\mathcal{S}, \mathcal{F}, P)$ **is an** $\mathcal{F}$**-measurable function on** $\mathcal{S}$. The class $\mathcal{F}_X = \{X^{-1}(B) : B \in \mathcal{B}\}$ is also a σ-field, generated by the random variable X. Notice that $\mathcal{F}_X \subset \mathcal{F}$.

By definition, every random variable X has a distribution function F_X. The **probability measure** $P_X\{\cdot\}$ induced by X on $(\mathbb{R}, B)$ is

$$P_X\{B\} = P\{X^{-1}(B)\}, \quad B \in \mathcal{B}. \tag{1.5.3}$$

A distribution function F_X is a real value function satisfying the properties

(i) $\lim_{x \to -\infty} F_X(x) = 0$;
(ii) $\lim_{x \to \infty} F_X(x) = 1$;
(iii) If $x_1 < x_2$ then $F_X(x_1) \leq F_X(x_2)$; and
(iv) $\lim_{\epsilon \downarrow 0} F_X(x + \epsilon) = F_X(x)$, and $\lim_{\epsilon \uparrow 0} F(x - \epsilon) = F_X(x-)$, all $-\infty < x < \infty$.

Thus, a distribution function F is right-continuous.

Given a distribution function F_X, we obtain from (1.5.1), for every $-\infty < a < b < \infty$,

$$P\{w : a < X(w) \leq b\} = F_X(b) - F_X(a) \tag{1.5.4}$$

and

$$P\{w : X(w) = x_0\} = F_X(x_0) - F_X(x_0-). \tag{1.5.5}$$

Thus, if F_X is continuous at a point x_0, then $P\{w : X(w) = x_0\} = 0$. If X is a random variable, then $Y = g(X)$ is a random variable only if g is $\mathcal{B}$-(Borel) measurable, i.e., for any $B \in \mathcal{B}$, $g^{-1}(B) \in \mathcal{B}$. Thus, if $Y = g(X)$, g is $\mathcal{B}$-measurable and X $\mathcal{F}$-measurable, then Y is also $\mathcal{F}$-measurable. The distribution function of Y is

$$F_Y(y) = P\{w : g(X(w)) \leq y\}. \tag{1.5.6}$$

Any two random variables X, Y having the same distribution are **equivalent**. We denote this by $Y \sim X$.

A distribution function F may have a countable number of distinct points of discontinuity. If x_0 is a point of discontinuity, $F(x_0) - F(x_0-) > 0$. In between points of discontinuity, F is continuous. If F assumes a constant value between points of discontinuity (step function), it is called **discrete**. Formally, let $-\infty < x_1 < x_2 < \cdots < \infty$ be points of discontinuity of F. Let $I_A(x)$ denote the indicator function of a set A, i.e.,

$$I_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{otherwise.} \end{cases}$$

Then a discrete F can be written as

$$\begin{aligned} F_{\mathrm{d}}(x) &= \sum_{i=1}^{\infty} I_{[x_i, x_{i+1})}(x) F(x_i) \\ &= \sum_{\{x_i \leq x\}} (F(x_i) - F(x_i-)). \end{aligned} \tag{1.5.7}$$

Let μ_1 and μ_2 be measures on $(\mathbb{R}, \mathcal{B})$. We say that μ_1 is **absolutely continuous** with respect to μ_2, and write $\mu_1 \ll \mu_2$, if $B \in \mathcal{B}$ and $\mu_2(B) = 0$ then $\mu_1(B) = 0$. Let λ denote the Lebesgue measure on $(\mathbb{R}, \mathcal{B})$. For every interval $(a, b]$, $-\infty < a < b < \infty$, $\lambda((a, b]) = b - a$. The celebrated **Radon–Nikodym** Theorem (see Shiryayev, 1984, p. 194) states that if $\mu_1 \ll \mu_2$ and μ_1, μ_2 are σ-finite measures on $(\mathbb{R}, \mathcal{B})$, there exists a $\mathcal{B}$-measurable nonnegative function $f(x)$ so that, for each $B \in \mathcal{B}$,

$$\mu_1(B) = \int_B f(x) d\mu_2(x) \tag{1.5.8}$$

where the **Lebesgue integral** in (1.5.8) will be discussed later. In particular, if P_c is absolutely continuous with respect to the Lebesgue measure λ, then there exists a function $f \geq 0$ so that

$$P_c\{B\} = \int_B f(x)\lambda(x), \quad B \in \mathcal{B}. \tag{1.5.9}$$

Moreover,

$$F_c(x) = \int_{-\infty}^{x} f(y)dy, \quad -\infty < x < \infty. \tag{1.5.10}$$

A distribution function F is called **absolutely continuous** if there exists a non-negative function f such that

$$F(\xi) = \int_{-\infty}^{\xi} f(x)dx, \quad -\infty < \xi < \infty. \tag{1.5.11}$$

The function f, which can be represented for "almost all x" by the derivative of F, is called the **probability density function** (p.d.f.) corresponding to F.

If F is absolutely continuous, then $f(x) = \dfrac{d}{dx}F(x)$ "almost everywhere." The term "almost everywhere" or "almost all" x means for all x values, excluding maybe on a set N of Lebesgue measure zero. Moreover, the probability assigned to any interval $(\alpha, \beta]$, $\alpha \leq \beta$, is

$$P\{\alpha < X \leq \beta\} = F(\beta) - F(\alpha) = \int_{\alpha}^{\beta} f(x)dx. \tag{1.5.12}$$

Due to the continuity of F we can also write

$$P\{\alpha < X \leq \beta\} = P\{\alpha \leq X \leq \beta\}.$$

Often the density functions f are Riemann integrable, and the above integrals are Riemann integrals. Otherwise, these are all Lebesgue integrals, which are defined in the next section.

There are continuous distribution functions that are not absolutely continuous. Such distributions are called **singular**. An example of a singular distribution is the **Cantor distribution** (see Shiryayev, 1984, p. 155).

Finally, every distribution function $F(x)$ is a **mixture** of the three types of distributions—discrete distribution $F_{\mathrm{d}}(\cdot)$, absolutely continuous distributions $F_{\mathrm{ac}}(\cdot)$, and singular distributions $F_{\mathrm{s}}(\cdot)$. That is, for some $0 \leq p_1, p_2, p_3 \leq 1$ such that $p_1 + p_2 + p_3 = 1$,

$$F(x) = p_1 F_{\mathrm{d}}(x) + p_2 F_{\mathrm{ac}}(x) + p_3 F_{\mathrm{s}}(x).$$

In this book we treat only mixtures of $F_{\mathrm{d}}(x)$ and $F_{\mathrm{ac}}(x)$.

1.6 THE LEBESGUE AND STIELTJES INTEGRALS

1.6.1 General Definition of Expected Value: The Lebesgue Integral

Let $(\mathcal{S}, \mathcal{F}, P)$ be a probability space. If X is a random variable, we wish to define the integral

$$E\{X\} = \int_{\mathcal{S}} X(w)P(dw). \tag{1.6.1}$$

We define first $E\{X\}$ for nonnegative random variables, i.e., $X(w) \geq 0$ for all $w \in \mathcal{S}$. Generally, $X = X^+ - X^-$, where $X^+(w) = \max(0, X(w))$ and $X^-(w) = -\min(0, X(w))$.

Given a nonnegative random variable X we construct for a given finite integer n the events

$$A_{k,n} = \left\{ w : \frac{k-1}{2^n} \leq X(w) < \frac{k}{2^n} \right\}, \quad k = 1, 2, \ldots, n2^n$$

and

$$A_{n2^n+1,n} = \{w : X(w) \geq n\}.$$

These events form a partition of $\mathcal{S}$. Let X_n, $n \geq 1$, be the discrete random variable defined as

$$X_n(w) = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} I_{A_{k,n}}(w) + n I_{A_{n2^n+1,n}}(w). \tag{1.6.2}$$

Notice that for each w, $X_n(w) \leq X_{n+1}(w) \leq \ldots \leq X(w)$ for all n. Also, if $w \in A_{k,n}$, $k = 1, \ldots, n2^n$, then $|X(w) - X_n(w)| \leq \frac{1}{2^n}$. Moreover, $A_{n2^n+1,n} \supset A_{(n+1)2^{n+1},n+1}$, all $n \geq 1$. Thus

$$\begin{aligned} \lim_{n\to\infty} A_{n2^n+1,n} &= \bigcap_{n=1}^{\infty} \{w : X(w) \geq n\} \\ &= \emptyset. \end{aligned}$$

Thus for all $w \in \mathcal{S}$

$$\lim_{n\to\infty} X_n(w) = X(w). \tag{1.6.3}$$

Now, for each discrete random variable $X_n(w)$

$$E\{X_n\} = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} P\{A_{k,n}\} + nP\{w : X(w) > n\}. \tag{1.6.4}$$

Obviously $E\{X_n\} \le n$, and $E\{X_{n+1}\} \ge E\{X_n\}$. Thus, $\lim_{n\to\infty} E\{X_n\}$ exists (it might be $+\infty$). Accordingly, **the Lebesgue integral** is defined as

$$\begin{aligned} E\{X\} &= \int X(w) P\{dw\} \\ &= \lim_{n\to\infty} E\{X_n\}. \end{aligned} \tag{1.6.5}$$

The Lebesgue integral may exist when the Riemann integral does not. For example, consider the probability space $(\mathcal{I}, \mathcal{B}, P)$ where $\mathcal{I} = \{x : 0 \le x \le 1\}$, $\mathcal{B}$ the Borel σ-field on $\mathcal{I}$, and P the Lebesgue measure on $[\mathcal{B}]$. Define

$$f(x) = \begin{cases} 0, & \text{if } x \text{ is irrational on } [0, 1] \\ 1, & \text{if } x \text{ is rational on } [0, 1]. \end{cases}$$

Let $B_0 = \{x : 0 \le x \le 1, \ f(x) = 0\}$, $B_1 = [0, 1] - B_0$. The Lebesgue integral of f is

$$\int_0^1 f(x)dx = 0 \cdot P\{B_0\} + 1 \cdot P\{B_1\} = 0,$$

since the Lebesgue measure of B_1 is zero. On the other hand, the Riemann integral of $f(x)$ does not exist. Notice that, contrary to the construction of the Riemann integral, the Lebesgue integral $\int f(x)P\{dx\}$ of a nonnegative function f is obtained by partitioning the **range** of the function f to 2^n subintervals $\mathcal{D}_n = \{B_j^{(n)}\}$ and constructing a discrete random variable $\hat{f}_n = \sum_{j=1}^{2^n} f_{n,j}^* I\{x \in B_j^{(n)}\}$, where $f_{n,j} = \inf\{f(x) : x \in B_j^{(n)}\}$. The expected value of $\hat{f}_n$ is $E\{\hat{f}_n\} = \sum_{j=1}^{2^n} f_{n,j}^* P(X \in B_j^{(n)})$. The sequence $\{E\{\hat{f}_n\}, n \ge 1\}$ is nondecreasing, and its limit exists (might be $+\infty$). Generally, we define

$$E\{X\} = E\{X^+\} - E\{X^-\} \tag{1.6.6}$$

if either $E\{X^+\} < \infty$ or $E\{X^-\} < \infty$.

If $E\{X^+\} = \infty$ and $E\{X^-\} = \infty$, we say that $E\{X\}$ does not exist. As a special case, if F is absolutely continuous with density f, then

$$E\{X\} = \int_{-\infty}^{\infty} x f(x) dx$$

provided $\int_{-\infty}^{\infty} |x| f(x) dx < \infty$. If F is discrete then

$$E\{X\} = \sum_{n=1}^{\infty} x_n P\{X = x_n\}$$

provided it is absolutely convergent.

From the definition (1.6.4), it is obvious that if $P\{X(w) \geq 0\} = 1$ then $E\{X\} \geq 0$. This immediately implies that if X and Y are two random variables such that $P\{w : X(w) \geq Y(w)\} = 1$, then $E\{X - Y\} \geq 0$. Also, if $E\{X\}$ exists then, for all $A \in \mathcal{F}$,

$$E\{|X| I_A(X)\} \leq E\{|X|\},$$

and $E\{X I_A(X)\}$ exists. If $E\{X\}$ is finite, $E\{X I_A(X)\}$ is also finite. From the definition of expectation we immediately obtain that for any finite constant c,

$$\begin{aligned} E\{cX\} &= cE\{X\}, \\ E\{X + Y\} &= E\{X\} + E\{Y\}. \end{aligned} \tag{1.6.7}$$

Equation (1.6.7) implies that the expected value is a **linear functional**, i.e., if $X_1, \ldots, X_n$ are random variables on $(\mathcal{S}, \mathcal{F}, P)$ and $\beta_0, \beta_1, \ldots, \beta_n$ are finite constants, then, if all expectations exist,

$$E\left\{\beta_0 + \sum_{i=1}^{n} \beta_i X_i\right\} = \beta_0 + \sum_{i=1}^{n} \beta_i E\{X_i\}. \tag{1.6.8}$$

We present now a few basic theorems on the convergence of the expectations of sequences of random variables.

Theorem 1.6.1 (Monotone Convergence). *Let $\{X_n\}$ be a monotone sequence of random variables and Y a random variable.*

(i) *Suppose that $X_n(w) \underset{n\to\infty}{\nearrow} X(w)$, $X_n(w) \geq Y(w)$ for all n and all $w \in \mathcal{S}$, and $E\{Y\} > -\infty$. Then*

$$\lim_{n\to\infty} E\{X_n\} = E\{X\}.$$

(ii) *If* $X_n(w) \underset{n\to\infty}{\searrow} X(w)$, $X_n(w) \le Y(w)$, *for all* n *and all* $w \in \mathcal{S}$, *and* $E\{Y\} < \infty$, *then*

$$\lim_{n\to\infty} E\{X_n\} = E\{X\}.$$

Proof. See Shiryayev (1984, p. 184). QED

Corollary 1.6.1. *If* $X_1, X_2, \ldots$ *are nonnegative random variables, then*

$$E\left\{\sum_{n=1}^{\infty} X_n\right\} = \sum_{n=1}^{\infty} E\{X_n\}. \tag{1.6.9}$$

Theorem 1.6.2 (Fatou). *Let* X_n, $n \ge 1$ *and* Y *be random variables.*

(i) *If* $X_n(w) \ge Y(w)$, $n \ge 1$, *for each* w *and* $E\{Y\} > -\infty$, *then*

$$E\left\{\underline{\lim}_{n\to\infty} X_n\right\} \le \underline{\lim}_{n\to\infty} E\{X_n\};$$

(ii) *if* $X_n(w) \le Y(w)$, $n \ge 1$, *for each* w *and* $E\{Y\} < \infty$, *then*

$$\overline{\lim}_{n\to\infty} E\{X_n\} \le E\left\{\overline{\lim}_{n\to\infty} X_n\right\};$$

(iii) *if* $|X_n(w)| \le Y(w)$ *for each* w, *and* $E\{Y\} < \infty$, *then*

$$E\left\{\underline{\lim}_{n\to\infty} X_n\right\} \le \underline{\lim}_{n\to\infty} E\{X_n\} \le \overline{\lim}_{n\to\infty} E\{X_n\} \le E\left\{\overline{\lim}_{n\to\infty} X_n\right\}. \tag{1.6.10}$$

Proof. (i)

$$\underline{\lim}_{n\to\infty} X_n(w) = \lim_{n\to\infty} \inf_{m\ge n} X_m(w).$$

The sequence $Z_n(w) = \inf_{m\ge n} X_m(w)$, $n \ge 1$ is monotonically increasing for each w, and $Z_n(w) \ge Y(w)$, $n \ge 1$. Hence, by Theorem 1.6.1,

$$\lim_{n\to\infty} E\{Z_n\} = E\left\{\lim_{n\to\infty} Z_n\right\}.$$

Or

$$E\left\{\underline{\lim}_{n\to\infty} X_n\right\} = \underline{\lim}_{n\to\infty} E\{Z_n\} = \underline{\lim}_{n\to\infty} E\{Z_n\} \leq \underline{\lim}_{n\to\infty} E\{X_n\}.$$

The proof of (ii) is obtained by defining $Z_n(w) = \sup_{m\geq n} X_m(w)$, and applying the previous theorem. Part (iii) is a result of (i) and (ii). QED

Theorem 1.6.3 (Lebesgue Dominated Convergence). *Let Y, X, X_n, $n \geq 1$, be random variables such that $|X_n(w)| \leq Y(w)$, $n \geq 1$ for almost all w, and $E\{Y\} < \infty$. Assume also that $P\left\{w : \lim_{n\to\infty} X_n(w) = X(w)\right\} = 1$. Then $E\{|X|\} < \infty$ and*

$$\lim_{n\to\infty} E\{X_n\} = E\left\{\lim_{n\to\infty} X_n\right\}, \tag{1.6.11}$$

and

$$\lim_{n\to\infty} E\{|X_n - X|\} = 0. \tag{1.6.12}$$

Proof. By Fatou's Theorem (Theorem 1.6.2)

$$E\left\{\underline{\lim}_{n\to\infty} X_n\right\} \leq \underline{\lim}_{n\to\infty} E\{X_n\} \leq \overline{\lim}_{n\to\infty} E\{X_n\} \leq E\left\{\overline{\lim}_{n\to\infty} X_n\right\}.$$

But since $\lim_{n\to\infty} X_n(w) = X(w)$, with probability 1,

$$E\{X\} = E\left\{\lim_{n\to\infty} X_n\right\} = \lim_{n\to\infty} E\{X_n\}.$$

Moreover, $|X(w)| < Y(w)$ for almost all w (with probability 1). Hence, $E\{|X|\} < \infty$. Finally, since $|X_n(w) - X(w)| \leq 2Y(w)$, with probability 1

$$\lim_{n\to\infty} E\{|X_n - X|\} = E\left\{\lim_{n\to\infty} |X_n - X|\right\} = 0.$$

QED

We conclude this section with a theorem on change of variables under Lebesgue integrals.

Theorem 1.6.4. *Let X be a random variable with respect to $(\mathcal{S}, \mathcal{F}, P)$. Let $g : \mathbb{R} \to \mathbb{R}$ be a Borel measurable function. Then for each $B \in \mathcal{B}$,*

$$\int_B g(x) P_X\{dx\} = \int_{X^{-1}(B)} g(X(w)) P\{dw\}. \tag{1.6.13}$$

The proof of the theorem is based on the following steps.

1. If $A \in \mathcal{B}$ and $g(x) = I_A(x)$ then

$$\begin{aligned}\int_B g(x)P_X\{dx\} &= \int_B I_A(x)P_X\{dx\} = P_X\{A \cap B\} \\ &= P\{w : X^{-1}(A) \cap X^{-1}(B)\} \\ &= \int_{X^{-1}(B)} g(X(w))P\{dw\}.\end{aligned}$$

2. Show that Equation (1.6.13) holds for simple random variables.

3. Follow the steps of the definition of the Lebesgue integral.

1.6.2 The Stieltjes–Riemann Integral

Let g be a function of a real variable and F a distribution function. Let $(\alpha, \beta]$ be a half-closed interval. Let

$$\alpha = x_0 < x_1 < \cdots < x_{n-1} < x_n = \beta$$

be a partition of $(\alpha, \beta]$ to n subintervals $(x_{i-1}, x_i]$, $i = 1, \ldots, n$. In each subinterval choose x_i', $x_{i-1} < x_i' \leq x_i$ and consider the sum

$$S_n = \sum_{i=1}^{n} g(x_i')[F(x_i) - F(x_{i-1})]. \tag{1.6.14}$$

If, as $n \to \infty$, $\max_{1 \leq i \leq n} |x_i - x_{i-1}| \to 0$ and if $\lim_{n \to \infty} S_n$ exists (finite) independently of the partitions, then the limit is called the **Stieltjes–Riemann integral** of g with respect to F. We denote this integral as

$$\int_\alpha^\beta g(x)dF(x).$$

This integral has the usual linear properties, i.e.,

(i) $\int_\alpha^\beta cg(x)dF(x) = c\int_\alpha^\beta g(x)dF(x)$;

(ii) $\int_\alpha^\beta (g_1(x) + g_2(x))dF(x) = \int_\alpha^\beta g_1(x)dF(x) + \int_\alpha^\beta g_2(x)dF(x)$; (1.6.15)

and

(iii) $$\int_\alpha^\beta g(x)d(\gamma F_1(x) + \delta F_2(x)) = \gamma \int_\alpha^\beta g(x)dF_1(x) + \delta \int_\alpha^\beta g(x)dF_2(x).$$

One can integrate by parts, if all expressions exist, according to the formula

$$\int_\alpha^\beta g(x)dF(x) = [g(\beta)F(\beta) - g(\alpha)F(\alpha)] - \int_\alpha^\beta g'(x)F(x)dx, \quad (1.6.16)$$

where $g'(x)$ is the derivative of $g(x)$. If F is strictly discrete, with jump points $-\infty < \xi_1 < \xi_2 < \cdots < \infty$,

$$\int_\alpha^\beta g(x)dF(x) = \sum_{j=1}^{\infty} I\{\alpha < \xi_j \le \beta\} g(\xi_j) p_j, \quad (1.6.17)$$

where $p_j = F(\xi_j) - F(\xi_j -)$, $j = 1, 2, \ldots$. If F is absolutely continuous, then at almost all points,

$$F(x + dx) - F(x) = f(x)dx + o(dx),$$

as $dx \to 0$. Thus, in the absolutely continuous case

$$\int_\alpha^\beta g(x)dF(x) = \int_\alpha^\beta g(x)f(x)dx. \quad (1.6.18)$$

Finally, the **improper Stieltjes–Riemann integral**, if it exists, is

$$\int_{-\infty}^{\infty} g(x)dF(x) = \lim_{\substack{\beta\to\infty \\ \alpha\to-\infty}} \int_\alpha^\beta g(x)dF(x). \quad (1.6.19)$$

If B is a set obtained by union and complementation of a sequence of intervals, we can write, by setting $g(x) = I\{x \in B\}$,

$$\begin{aligned} P\{B\} &= \int_{-\infty}^{\infty} I\{x \in B\}dF(x) \\ &= \int_B dF(x), \end{aligned} \quad (1.6.20)$$

where F is either discrete or absolutely continuous.

1.6.3 Mixtures of Discrete and Absolutely Continuous Distributions

Let F_{d} be a discrete distribution and let F_{ac} be an absolutely continuous distribution function. Then for all α $0 \leq \alpha \leq 1$,

$$F(x) = \alpha F_{\mathrm{d}}(x) + (1 - \alpha) F_{\mathrm{ac}}(x) \tag{1.6.21}$$

is also a distribution function, which is a mixture of the two types. Thus, for such mixtures, if $-\infty < \xi_1 < \xi_2 < \cdots < \infty$ are the jump points of F_{d}, then for every $-\infty < \gamma \leq \delta < \infty$ and $B = (\gamma, \delta]$,

$$\begin{aligned} P\{B\} &= \int_{\gamma}^{\delta} dF(x) \\ &= \alpha \sum_{j=1}^{\infty} I\{\gamma < \xi_j \leq \delta\} dF_{\mathrm{d}}(\xi_j) + (1 - \alpha) \int_{\gamma}^{\delta} dF_{\mathrm{ac}}(x). \end{aligned} \tag{1.6.22}$$

Moreover, if $B^+ = [\gamma, \delta]$ then

$$P\{B^+\} = P\{B\} + dF_{\mathrm{d}}(\gamma).$$

The expected value of X, when $F(x) = pF_{\mathrm{d}}(x) + (1 - p)F_{\mathrm{ac}}(x)$ is,

$$E\{X\} = p \sum_{\{j\}} \xi_j f_{\mathrm{d}}(\xi_j) + (1 - p) \int_{-\infty}^{\infty} x f_{\mathrm{ac}}(x) dx, \tag{1.6.23}$$

where $\{\xi_j\}$ is the set of jump points of F_{d}; f_{d} and f_{ac} are the corresponding p.d.f.s. We assume here that the sum and the integral are absolutely convergent.

1.6.4 Quantiles of Distributions

The p-**quantiles** or **fractiles** of distribution functions are inverse points of the distributions. More specifically, the p-quantile of a distribution function F, designated by x_p or $F^{-1}(p)$, is the smallest value of x at which $F(x)$ is greater or equal to p, i.e.,

$$x_p = F^{-1}(p) = \inf\{x : F(x) \geq p\}. \tag{1.6.24}$$

The inverse function defined in this fashion is unique. The **median** of a distribution, $x_{.5}$, is an important parameter characterizing the **location** of the distribution. The **lower** and **upper quartiles** are the .25- and .75-quantiles. The difference between these quantiles, $R_Q = x_{.75} - x_{.25}$, is called the **interquartile range**. It serves as one of the measures of **dispersion** of distribution functions.

1.6.5 Transformations

From the distribution function $F(x) = \alpha F_{\mathrm{d}}(x) + (1-\alpha)F_{\mathrm{ac}}(x)$, $0 \le \alpha \le 1$, we can derive the distribution function of a transformed random variable $Y = g(X)$, which is

$$\begin{aligned} F_Y(y) &= P\{g(X) \le y\} \\ &= P\{X \in B_y\} \\ &= \alpha \sum_{j=1}^{\infty} I\{\xi_j \in B_y\} dF_{\mathrm{d}}(\xi_j) + (1-\alpha) \int_{B_y} dF_{\mathrm{ac}}(x) \end{aligned} \tag{1.6.25}$$

where

$$B_y = \{x : g(x) \le y\}.$$

In particular, if F is absolutely continuous and if g is a strictly increasing differentiable function, then the p.d.f. of Y, $h(y)$, is

$$f_Y(y) = f_X(g^{-1}(y)) \left(\frac{d}{dy} g^{-1}(y) \right), \tag{1.6.26}$$

where $g^{-1}(y)$ is the inverse function. If $g'(x) < 0$ for all x, then

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right|. \tag{1.6.27}$$

Suppose that X is a continuous random variable with p.d.f. $f(x)$. Let $g(x)$ be a differentiable function that is not necessarily one-to-one, like $g(x) = x^2$. Excluding cases where $g(x)$ is a constant over an interval, like the indicator function, let $m(y)$ denote the number of roots of the equation $g(x) = y$. Let $\xi_j(y)$, $j = 1, \ldots, m(y)$ denote the roots of this equation. Then the p.d.f. of $Y = g(x)$ is

$$f_Y(y) = \sum_{j=1}^{m(y)} f_X(\xi_j(y)) \cdot \frac{1}{|g'(\xi_j(y))|} \tag{1.6.28}$$

if $m(y) > 0$ and zero otherwise.

1.7 JOINT DISTRIBUTIONS, CONDITIONAL DISTRIBUTIONS AND INDEPENDENCE

1.7.1 Joint Distributions

Let $(X_1, \ldots, X_k)$ be a vector of k random variables defined on the same probability space. These random variables represent variables observed in the same experiment. The joint distribution function of these random variables is a real value function F of k real arguments $(\xi_1, \ldots, \xi_k)$ such that

$$F(\xi_1, \ldots, \xi_k) = P\{X_1 \leq \xi_1, \ldots, X_k \leq \xi_k\}. \tag{1.7.1}$$

The joint distribution of two random variables is called a **bivariate distribution function**.

Every bivariate distribution function F has the following properties.

$$\begin{aligned}
&\textbf{(i)} \quad \lim_{\xi_1 \to -\infty} F(\xi_1, \xi_2) = \lim_{\xi_2 \to -\infty} F(\xi_1, \xi_2) = 0, \quad \text{for all } \xi_1, \xi_2; \\
&\textbf{(ii)} \quad \lim_{\xi_1 \to \infty} \lim_{\xi_2 \to \infty} F(\xi_1, \xi_2) = 1; \\
&\textbf{(iii)} \quad \lim_{\epsilon \downarrow 0} F(\xi_1 + \epsilon, \xi_2 + \epsilon) = F(\xi_1, \xi_2) \text{ for all } (\xi_1, \xi_2); \\
&\textbf{(iv)} \quad \text{for any } a < b,\ c < d,\ F(b, d) - F(a, d) - F(b, c) + F(a, c) \geq 0.
\end{aligned} \tag{1.7.2}$$

Property (iii) is the right continuity of $F(\xi_1, \xi_2)$. Property (iv) means that the probability of every rectangle is nonnegative. Moreover, the total increase of $F(\xi_1, \xi_2)$ is from 0 to 1. The similar properties are required in cases of a larger number of variables.

Given a bivariate distribution function F. The univariate distributions of X_1 and X_2 are F_1 and F_2 where

$$F_1(x) = \lim_{y \to \infty} F(x, y), \quad F_2(y) = \lim_{x \to \infty} F(x, y). \tag{1.7.3}$$

F_1 and F_2 are called the **marginal distributions** of X_1 and X_2, respectively. In cases of joint distributions of three variables, we can distinguish between three marginal bivariate distributions and three marginal univariate distributions. As in the univariate case, multivariate distributions are either discrete, absolutely continuous, singular, or mixtures of the three main types. In the discrete case there are at most a

countable number of points $\{(\xi_1^{(j)}, \ldots, \xi_k^{(j)}), j = 1, 2, \ldots\}$ on which the distribution concentrates. In this case the joint probability function is

$$p(x_1, \ldots, x_k) = \begin{cases} P\{X_1 = \xi_1^{(j)}, \ldots, X_k = \xi_k^{(j)}\}, & \text{if } (x_1, \ldots, x_k) = (\xi_1^{(j)}, \ldots, \xi_k^{(j)}) \\ & j = 1, 2, \ldots \\ 0, & \text{otherwise.} \end{cases} \tag{1.7.4}$$

Such a discrete p.d.f. can be written as

$$p(x_1, \ldots, x_k) = \sum_{j=1}^{\infty} I\{(x_1, \ldots, x_k) = (\xi_1^{(j)}, \ldots, \xi_k^{(j)})\} p_j,$$

where $p_j = P\{X_1 = \xi_1^{(j)}, \ldots, X_k = \xi_k^{(j)}\}$.

In the absolutely continuous case there exists a nonnegative function $f(x_1, \ldots, x_k)$ such that

$$F(\xi_1, \ldots, \xi_k) = \int_{-\infty}^{\xi_1} \cdots \int_{-\infty}^{\xi_k} f(x_1, \ldots, x_k) dx_1 \ldots dx_k. \tag{1.7.5}$$

The function $f(x_1, \ldots, x_k)$ is called the **joint density function**.

The marginal probability or density functions of single variables or of a subvector of variables can be obtained by summing (in the discrete case) or integrating, in the absolutely continuous case, the joint distribution functions (densities) with respect to the variables that are not under consideration, over their range of variation.

Although the presentation here is in terms of k discrete or k absolutely continuous random variables, the joint distributions can involve some discrete and some continuous variables, or mixtures.

If X_1 has an absolutely continuous marginal distribution and X_2 is discrete, we can introduce the function $N(B)$ on $\mathcal{B}$, which counts the number of jump points of X_2 that belong to B. $N(B)$ is a σ-finite measure. Let $\lambda(B)$ be the Lebesgue measure on $\mathcal{B}$. Consider the σ-finite measure on $\mathcal{B}^{(2)}$, $\mu(B_1 \times B_2) = \lambda(B_1)N(B_2)$. If X_1 is absolutely continuous and X_2 discrete, their joint probability measure $P_{\mathbf{X}}$ is absolutely continuous with respect to μ. There exists then a nonnegative function $f_{\mathbf{X}}$ such that

$$F_{\mathbf{X}}(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_{\mathbf{X}}(y_1, y_2) dy_1 dN(y_2).$$

The function $f_{\mathbf{X}}$ is a joint p.d.f. of X_1, X_2 with respect to μ. The joint p.d.f. $f_{\mathbf{X}}$ is positive only at jump point of X_2.

If $X_1, \ldots, X_k$ have a joint distribution with p.d.f. $f(x_1, \ldots, x_k)$, the **expected** value of a function $g(X_1, \ldots, X_k)$ is defined as

$$E\{g(X_1, \ldots, X_k)\} = \int g(x_1, \ldots, x_k) dF(x_1, \ldots, x_k). \tag{1.7.6}$$

We have used here the conventional notation for Stieltjes integrals.

Notice that if (X, Y) have a joint distribution function $F(x, y)$ and if X is discrete with jump points of $F_1(x)$ at $\xi_1, \xi_2, \ldots$, and Y is absolutely continuous, then, as in the previous example,

$$\int g(x, y) dF(x, y) = \sum_{j=1}^{\infty} \int g(\xi_j, y) f(\xi_j, y) dy$$

where $f(x, y)$ is the joint p.d.f. A similar formula holds for the case of X, absolutely continuous and Y, discrete.

1.7.2 Conditional Expectations: General Definition

Let $X(w) \geq 0$, for all $w \in \mathcal{S}$, be a random variable with respect to $(\mathcal{S}, \mathcal{F}, P)$. Consider a σ-field $\mathcal{G}$, $\mathcal{G} \subset \mathcal{F}$. The conditional expectation of X given $\mathcal{G}$ is defined as a $\mathcal{G}$-measurable random variable $E\{X \mid \mathcal{G}\}$ satisfying

$$\int_A X(w) P\{dw\} = \int_A E\{X \mid \mathcal{G}\}(w) P\{dw\} \tag{1.7.7}$$

for all $A \in \mathcal{G}$. Generally, $E\{X \mid \mathcal{G}\}$ is defined if $\min\{E\{X^+ \mid \mathcal{G}\}, E\{X^- \mid \mathcal{G}\}\} < \infty$ and $E\{X \mid \mathcal{G}\} = E\{X^+ \mid \mathcal{G}\} - E\{X^- \mid \mathcal{G}\}$. To see that such conditional expectations exist, where $X(w) \geq 0$ for all w, consider the σ-finite measure on $\mathcal{G}$,

$$Q(A) = \int_A X(w) P\{dw\}, \quad A \in \mathcal{G}. \tag{1.7.8}$$

Obviously $Q \ll P$ and by Radon–Nikodym Theorem, there exists a nonnegative, $\mathcal{G}$-measurable random variable $E\{X \mid \mathcal{G}\}$ such that

$$Q(A) = \int_A E\{X \mid \mathcal{G}\}(w) P\{dw\}. \tag{1.7.9}$$

According to the Radon–Nikodym Theorem, $E\{X \mid \mathcal{G}\}$ is determined only up to a set of P-measure zero.

If $B \in \mathcal{F}$ and $X(w) = I_B(w)$, then $E\{X \mid \mathcal{G}\} = P\{B \mid \mathcal{G}\}$ and according to (1.6.13),

$$\begin{aligned} P\{A \cap B\} &= \int_A I_B(w) P\{dw\} \\ &= \int_A P\{B \mid \mathcal{G}\} P\{dw\}. \end{aligned} \tag{1.7.10}$$

Notice also that if X is $\mathcal{G}$-measurable then $X = E\{X \mid \mathcal{G}\}$ with probability 1.

On the other hand, if $\mathcal{G} = \{\emptyset, \mathcal{S}\}$ is the trivial algebra, then $E\{X \mid \mathcal{G}\} = E\{X\}$ with probability 1.

From the definition (1.7.7), since $\mathcal{S} \in \mathcal{G}$,

$$\begin{aligned} E\{X\} &= \int_{\mathcal{S}} X(w) P\{dw\} \\ &= \int_{\mathcal{S}} E\{X \mid \mathcal{G}\} P\{dw\}. \end{aligned}$$

This is the law of iterated expectation; namely, for all $\mathcal{G} \subset \mathcal{F}$,

$$E\{X\} = E\{E\{X \mid \mathcal{G}\}\}. \tag{1.7.11}$$

Furthermore, if X and Y are two random variables on $(\mathcal{S}, \mathcal{F}, P)$, the collection of all sets $\{Y^{-1}(B), B \in \mathcal{B}\}$, is a σ-field generated by Y. Let $\mathcal{F}_Y$ denote this σ-field. Since Y is a random variable, $\mathcal{F}_Y \subset \mathcal{F}$. We define

$$E\{X \mid Y\} = E\{X \mid \mathcal{F}_Y\}. \tag{1.7.12}$$

Let y_0 be such that $f_Y(y_0) > 0$.

Consider the $\mathcal{F}_Y$-measurable set $A_\delta = \{w : y_0 < Y(w) \leq y_0 + \delta\}$. According to (1.7.7)

$$\begin{aligned} \int_{A\delta} X(w) P(dw) &= \int_{-\infty}^{\infty} \int_{y_0}^{y_0+\delta} x f_{XY}(x, y) dx dy \\ &= \int_{y_0}^{y_0+\delta} E\{X \mid Y = y\} f_Y(y) dy. \end{aligned} \tag{1.7.13}$$

The left-hand side of (1.7.13) is, if $E\{|X|\} < \infty$,

$$\begin{aligned} \int_{-\infty}^{\infty} x \int_{y_0}^{y_0+\delta} f_{XY}(x, y) dy dx &= \int_{y_0}^{y_0+\delta} f_Y(y_0) \int_{-\infty}^{\infty} x \frac{f_{XY}(x, y)}{f_Y(y_0)} dx dy \\ &= f_Y(y_0)\delta \int_{-\infty}^{\infty} x \frac{f_{XY}(x, y_0)}{f_Y(y_0)} dx + o(\delta), \quad \text{as } \delta \to 0 \end{aligned}$$

where $\lim_{\delta \to 0} \frac{o(\delta)}{\delta} = 0$. The right-hand side of (1.7.13) is

$$\int_{y_0}^{y_0+\delta} E\{X \mid Y = y\} f_Y(y) dy = E\{X \mid Y = y_0\} f_Y(y_0)\delta + o(\delta), \quad \text{as } \delta \to 0.$$

Dividing both sides of (1.7.13) by $f_Y(y_0)\delta$, we obtain that

$$\begin{aligned} E\{X \mid Y = y_0\} &= \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y_0) dx \\ &= \int_{-\infty}^{\infty} x \frac{f_{XY}(x, y_0)}{f_X(y_0)} dx. \end{aligned}$$

We therefore define for $f_Y(y_0) > 0$

$$f_{X|Y}(x \mid y_0) = \frac{f_{XY}(x, y_0)}{f_Y(y_0)}. \tag{1.7.14}$$

More generally, for $k > 2$ let $f(x_1, \ldots, x_k)$ denote the joint p.d.f. of $(X_1, \ldots, X_k)$. Let $1 \leq r < k$ and $g(x_1, \ldots, x_r)$ denote the marginal joint p.d.f. of $(X_1, \ldots, X_r)$. Suppose that $(\xi_1, \ldots, \xi_r)$ is a point at which $g(\xi_1, \ldots, \xi_r) > 0$. The **conditional** p.d.f. of $X_{r+1}, \ldots, X_k$ given $\{X_1 = \xi_1, \ldots, X_r = \xi_r\}$ is defined as

$$h(x_{r+1}, \ldots, x_k \mid \xi_1, \ldots, \xi_r) = \frac{f(\xi_1, \ldots, \xi_r, x_{r+1}, \ldots, x_k)}{g(\xi_1, \ldots, \xi_r)}. \tag{1.7.15}$$

We remark that conditional distribution functions are not defined on points $(\xi_1, \ldots, \xi_r)$ such that $g(\xi_1, \ldots, \xi_r) = 0$. However, it is easy to verify that the probability associated with this set of points is zero. Thus, the definition presented here is sufficiently general for statistical purposes. Notice that $f(x_{r+1}, \ldots, x_k \mid \xi_1, \ldots, \xi_r)$ is, for a fixed point $(\xi_1, \ldots, \xi_r)$ at which it is well defined, a nonnegative function of $(x_{r+1}, \ldots, x_k)$ and that

$$\int dF(x_{r+1}, \ldots, x_k \mid \xi_1, \ldots, \xi_r) = 1.$$

Thus, $f(x_{r+1}, \ldots, x_k \mid \xi_1, \ldots, \xi_r)$ is indeed a joint p.d.f. of $(X_{r+1}, \ldots, X_k)$. The point $(\xi_1, \ldots, \xi_r)$ can be considered a parameter of the conditional distribution.

If $\psi(X_{r+1}, \ldots, X_k)$ is an (integrable) function of $(X_{r+1}, \ldots, X_k)$, the **conditional expectation** of $\psi(X_{r+1}, \ldots, X_k)$ given $\{X_1 = \xi_1, \ldots, X_r = \xi_r\}$ is

$$E\{\psi(X_{r+1}, \ldots, X_k) \mid \xi_1, \ldots, \xi_r\} = \int \psi(x_{r+1}, \ldots, x_k) dF(x_{r+1}, \ldots, x_k \mid \xi_1, \ldots, \xi_r). \tag{1.7.16}$$

This conditional expectation exists if the integral is absolutely convergent.

1.7.3 Independence

Random variables $X_1, \ldots, X_n$, on the same probability space, are called **mutually independent** if, for any Borel sets $B_1, \ldots, B_n$,

$$P\{w : X_1(w) \in B_1, \ldots, X_n(w) \in B_n\} = \prod_{j=1}^{n} P\{w : X_j \in B_j\}. \quad (1.7.17)$$

Accordingly, the joint distribution function of any k-tuple $(X_{i_1}, \ldots, X_{i_k})$ is a product of their marginal distributions. In particular,

$$F_{X_1 \cdots X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i). \quad (1.7.18)$$

Equation (1.7.18) implies that if $X_1, \ldots, X_n$ have a joint p.d.f. $f_{\mathbf{X}}(x_1, \ldots, x_n)$ and if they are independent, then

$$f_{\mathbf{X}}(x_1, \ldots, x_n) = \prod_{j=1}^{n} f_{X_j}(x_j). \quad (1.7.19)$$

Moreover, if $g(X_1, \ldots, X_n) = \prod_{j=1}^{n} g_j(X_j)$, where $g(x_1, \ldots, x_n)$ is $\mathcal{B}^{(n)}$-measurable and $g_j(x)$ are $\mathcal{B}$-measurable, then under independence

$$E\{g(X_1, \ldots, X_n)\} = \prod_{j=1}^{n} E\{g_j(X_j)\}. \quad (1.7.20)$$

Probability models with independence structure play an important role in statistical theory. From (1.7.12) and (1.7.21), we imply that if $\mathbf{X}^{(r)} = (X_1, \ldots, X_r)$ and $\mathbf{Y}^{(r)} = (X_{r+1}, \ldots, X_n)$ are independent subvectors, then the conditional distribution of $\mathbf{X}^{(r)}$ given $\mathbf{Y}^{(r)}$ is independent of $\mathbf{Y}^{(r)}$, i.e.,

$$f(x_1, \ldots, x_r \mid x_{r+1}, \ldots, x_n) = f(x_1, \ldots, x_r) \quad (1.7.21)$$

with probability one.

1.8 MOMENTS AND RELATED FUNCTIONALS

A **moment** of **order** r, $r = 1, 2, \ldots$, of a distribution $F(x)$ is

$$\mu_r = E\{X^r\}. \quad (1.8.1)$$

The moments of $Y = X - \mu_1$ are called **central moments** and those of $|X|$ are called **absolute** moments. It is simple to prove that the existence of an absolute moment of order r, $r > 0$, implies the existence of all moments of order s, $0 < s \leq r$, (see Section 1.13.3).

Let $\mu_r^* = E\{(X - \mu_1)^r\}$, $r = 1, 2, \ldots$ denote the rth central moment of a distribution. From the binomial expansion and the linear properties of the expectation operator we obtain the relationship between moments (about the origin) μ_r and center moments m_r

$$\mu_r^* = \sum_{j=0}^{r} (-1)^j \binom{r}{j} \mu_{r-j} \mu_i^j, \quad r = 1, 2, \ldots \tag{1.8.2}$$

where $\mu_0 \equiv 1$.

A distribution function F is called **symmetric about a point** ξ_0 **if its** p.d.f. **is symmetric about** ξ_0, i.e.,

$$f(\xi_0 + h) = f(\xi_0 - h), \quad \text{all } 0 \leq h < \infty.$$

From this definition we immediately obtain the following results.

(i) If F is symmetric about ξ_0 and $E\{|X|\} < \infty$, then $\xi_0 = E\{X\}$.

(ii) If F is symmetric, then all **central** moments of **odd order** are zero, i.e., $E\{(X - E\{X\})^{2m+1}\} = 0, m = 0, 1, \ldots$, provided $E|X|^{2m+1} < \infty$.

The central moment of the second order occupies a central role in the theory of statistics and is called the **variance** of X. The variance is denoted by $V\{X\}$. The square-root of the variance, called the **standard deviation**, is a measure of dispersion around the expected value. We denote the standard deviation by σ. The variance of X is equal to

$$V\{X\} = E\{X^2\} - (E\{X\})^2. \tag{1.8.3}$$

The variance is always nonnegative, and hence for every distribution having a finite second moment $E\{X^2\} \geq (E\{X\})^2$. One can easily verify from the definition that if X is a random variable and a and b are constants, then $V\{a + bX\} = b^2 V\{X\}$.

The variance is equal to zero if and only if the distribution function is concentrated at one point (a degenerate distribution).

A famous inequality, called the **Chebychev inequality**, relates the probability of X concentrating around its mean, and the standard deviation σ.

Theorem 1.8.1 (Chebychev). *If F_X has a finite standard deviation σ, then, for every $a > 0$,*

$$P\{w : |X(w) - \mu| \leq a\} \geq 1 - \frac{\sigma^2}{a^2}, \tag{1.8.4}$$

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

Proof.

$$\begin{aligned}\sigma^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 dF_X(x) \\ &= \int_{\{|x-\mu|\le a\}} (x-\mu)^2 dF_X(x) + \int_{\{|x-\mu|>a\}} (x-\mu)^2 dF_X(x) \\ &\ge a^2 P\{w : |X(w)-\mu| > a\}.\end{aligned} \tag{1.8.5}$$

Hence,

$$P\{w : |X(w)-\mu| \le a\} = 1 - P\{w : |X(w)-\mu| > a\} \ge 1 - \frac{\sigma^2}{a^2}.$$

QED

Notice that in the proof of the theorem, we used the Riemann–Stieltjes integral. The theorem is true for **any** type of distribution for which $0 \le \sigma < \infty$. The Chebychev inequality is a crude inequality. Various types of better inequalities are available, under additional assumptions (see Zelen and Severv, 1968; Rohatgi, 1976, p. 102).

The **moment generating function** (m.g.f.) of a random variable X, denoted by M, is defined as

$$M(t) = E\{\exp(tX)\}, \tag{1.8.6}$$

where t is such that $M(t) < \infty$. Obviously, at $t = 0$, $M(0) = 1$. However, $M(t)$ may not exist when $t \neq 0$. Assume that $M(t)$ exists for all t in some interval (a, b), $a < 0 < b$. There is a one-to-one correspondence between the distribution function F and the moment generating function M. M is analytic on (a, b), and can be differentiated under the expectation integral. Thus

$$\frac{d^r}{dt^r} M(t) = E\{X^r \exp\{tX\}\}, \quad r = 1, 2, \ldots. \tag{1.8.7}$$

Under this assumption the rth derivative of $M(t)$ evaluated at $t = 0$ yields the moment of order r.

To overcome the problem of M being undefined in certain cases, it is useful to use the **characteristic function**

$$\phi(t) = E\{e^{itX}\}, \tag{1.8.8}$$

where $i = \sqrt{-1}$. The characteristic function exists for all t since

$$|\phi(t)| = \left| \int_{-\infty}^{\infty} e^{itx} dF(x) \right| \le \int_{-\infty}^{\infty} |e^{itx}| dF(x) = 1. \tag{1.8.9}$$

Indeed, $|e^{itx}| = 1$ for all x and all t.

If X assumes nonnegative integer values, it is often useful to use the **probability generating function** (p.g.f.)

$$G(t) = \sum_{j=0}^{\infty} t^j p_j, \tag{1.8.10}$$

which is convergent if $|t| < 1$. Moreover, given a p.g.f. of a nonnegative integer value random variable X, its p.d.f. can be obtained by the formula

$$P\{w : X(w) = k\} = \frac{1}{k!} \frac{d^k}{dt^k} G(t)\Big|_{t=0}. \tag{1.8.11}$$

The logarithm of the moment generating function is called **cumulants generating function**. We denote this generating function by K. K exists for all t for which M is finite. Both M and K are analytic functions in the interior of their domains of convergence. Thus we can write for t close to zero

$$K(t) = \log M(t) = \sum_{j=0}^{\infty} \frac{\kappa_j}{j!} t^j \tag{1.8.12}$$

The coefficients $\{\kappa_j\}$ are called **cumulants**. Notice that $\kappa_0 = 0$, and κ_j, $j \geq 1$, can be obtained by differentiating $K(t)$ j times, and setting $t = 0$. Generally, the relationships between the cumulants and the moments of a distribution are, for $j = 1, \ldots, 4$

$$\begin{aligned} \kappa_1 &= \mu_1 \\ \kappa_2 &= \mu_2 - \mu_1^2 = \mu_2^* \\ \kappa_3 &= \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3 = \mu_3^* \\ \kappa_4 &= \mu_4^* - 3(\mu_2^*)^2. \end{aligned} \tag{1.8.13}$$

The following two indices

$$\beta_1 = \frac{\mu_3^*}{\sigma^3} \tag{1.8.14}$$

and

$$\beta_2 = \frac{\mu_4^*}{\sigma^4}, \tag{1.8.15}$$

where $\sigma^2 = \mu_2^*$ is the variance, are called coefficients of **skewness** (asymmetry) and **kurtosis** (steepness), respectively. If the distribution is symmetric, then $\beta_1 = 0$. If $\beta_1 > 0$ we say that the distribution is positively skewed; if $\beta_1 < 0$, it is negatively

skewed. If $\beta_2 > 3$ we say that the distribution is **steep**, and if $\beta_2 < 3$ we say that the distribution is **flat**.

The following equation is called the law of total variance.

If $E\{X^2\} < \infty$ then

$$V\{X\} = E\{V\{X \mid Y\}\} + V\{E\{X \mid Y\}\}, \tag{1.8.16}$$

where $V\{X \mid Y\}$ denotes the conditional variance of X given Y.

It is often the case that it is easier to find the conditional mean and variance, $E\{X \mid Y\}$ and $V\{X \mid Y\}$, than to find $E\{X\}$ and $V\{X\}$ directly. In such cases, formula (1.8.16) becomes very handy.

The product central moment of two variables (X, Y) is called the **covariance** and denoted by $\text{cov}(X, Y)$. More specifically

$$\begin{aligned} \text{cov}(X, Y) &= E\{[X - E\{X\}][Y - E\{Y\}]\} \\ &= E\{XY\} - E\{X\}E\{Y\}. \end{aligned} \tag{1.8.17}$$

Notice that $\text{cov}(X, Y) = \text{cov}(Y, X)$, and $\text{cov}(X, X) = V\{X\}$. Notice that if X is a random variable having a finite first moment and a is any finite **constant**, then $\text{cov}(a, X) = 0$. Furthermore, whenever the second moments of X and Y exist the covariance exists. This follows from **the Schwarz inequality** (see Section 1.13.3), i.e., if F is the joint distribution of (X, Y) and F_X, F_Y are the marginal distributions of X and Y, respectively, then

$$\left(\int g(x)h(y)dF(x, y)\right)^2 \leq \left(\int g^2(x)dF_X(x)\right)\left(\int h^2(y)dF_Y(y)\right) \tag{1.8.18}$$

whenever $E\{g^2(X)\}$ and $E\{h^2(Y)\}$ are finite. In particular, for any two random variables having second moments

$$\text{cov}^2(X, Y) \leq V\{X\}V\{Y\}.$$

The ratio

$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{V\{X\}V\{Y\}}} \tag{1.8.19}$$

is called the **coefficient of correlation** (Pearson's product moment correlation). From (1.8.18) we deduce that $-1 \leq \rho \leq 1$. The sign of ρ is that of $\text{cov}(X, Y)$.

The **m.g.f.** of a multivariate distribution is a function of k variables

$$M(t_1, \ldots, t_k) = E\left\{\exp\left\{\sum_{i=1}^{k} t_i X_i\right\}\right\}. \tag{1.8.20}$$

Let $X_1, \ldots, X_k$ be random variables having a joint distribution. Consider the linear transformation $Y = \sum_{j=1}^{k} \beta_j X_j$, where $\beta_1, \ldots, \beta_k$ are constants. Some formulae for the moments and covariances of such linear functions are developed here. Assume that all the moments under consideration exist. Starting with the expected value of Y we prove:

$$E\left\{\sum_{i=1}^{k} \beta_i X_i\right\} = \sum_{i=1}^{k} \beta_i E\{X_i\}. \tag{1.8.21}$$

This result is a direct implication of the definition of the integral as a linear operator.

Let $\mathbf{X}$ denote a random vector in a column form and $\mathbf{X}'$ its transpose. The expected value of a random vector $\mathbf{X}' = (X_1, \ldots, X_k)$ is defined as the corresponding vector of expected values, i.e.,

$$E\{\mathbf{X}'\} = (E\{X_1\}, \ldots, E\{X_k\}). \tag{1.8.22}$$

Furthermore, let $\mathbf{\Sigma}$ denote a $k \times k$ matrix with elements that are the variances and covariances of the components of $\mathbf{X}$. In symbols

$$\mathbf{\Sigma} = (\sigma_{ij}; i, j = 1, \ldots, k), \tag{1.8.23}$$

where $\sigma_{ij} = \text{cov}(X_i, X_j)$, $\sigma_{ii} = V\{X_i\}$. If $Y = \boldsymbol{\beta}'\mathbf{X}$ where $\boldsymbol{\beta}$ is a vector of constants, then

$$\begin{aligned} V\{Y\} &= \boldsymbol{\beta}' \mathbf{\Sigma} \boldsymbol{\beta} \\ &= \sum_i \sum_j \beta_i \beta_j \sigma_{ij} \\ &= \sum_{i=1}^{k} \beta_i^2 \sigma_{ii} + \sum \sum_{i \neq j} \beta_i \beta_j \sigma_{ij}. \end{aligned} \tag{1.8.24}$$

The result given by (1.8.24) can be generalized in the following manner. Let $Y_1 = \boldsymbol{\beta}'\mathbf{X}$ and $Y_2 = \boldsymbol{\alpha}'\mathbf{X}$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are arbitrary constant vectors. Then

$$\text{cov}(Y_1, Y_2) = \boldsymbol{\alpha}' \mathbf{\Sigma} \boldsymbol{\beta}. \tag{1.8.25}$$

Finally, if $\mathbf{X}$ is a k-dimensional random vector with covariance matrix $\mathbf{\Sigma}$ and $\mathbf{Y}$ is an m-dimensional vector $\mathbf{Y} = A\mathbf{X}$, where A is an $m \times k$ matrix of constants, then the covariance matrix of $\mathbf{Y}$ is

$$V[\mathbf{Y}] = A \mathbf{\Sigma} A'. \tag{1.8.26}$$

In addition, if the covariance matrix of $\mathbf{X}$ is Σ, then the covariance matrix of $\mathbf{Y} = \boldsymbol{\xi} + A\mathbf{X}$ is $\mathbf{V}$, where $\boldsymbol{\xi}$ is a vector of constants, and A is a matrix of constants. Finally, if $\mathbf{Y} = A\mathbf{X}$ and $\mathbf{Z} = B\mathbf{X}$, where A and B are matrices of constants with compatible dimensions, then the covariance matrix of $\mathbf{Y}$ and $\mathbf{Z}$ is

$$C[\mathbf{Y}, \mathbf{Z}] = A\Sigma B'. \tag{1.8.27}$$

We conclude this section with an important theorem concerning a characteristic function. Recall that ϕ is generally a complex valued function on $\mathbb{R}$, i.e.,

$$\phi(t) = \int_{-\infty}^{\infty} \cos(tx) dF(x) + i \int_{-\infty}^{\infty} \sin(tx) dF(x).$$

Theorem 1.8.2. *A characteristic function ϕ, of a distribution function F, has the following properties.*

- ***(i)*** $|\phi(t)| \leq \phi(0) = 1$;
- ***(ii)*** *$\phi(t)$ is a uniformly continuous function of t, on $\mathbb{R}$;*
- ***(iii)*** $\phi(t) = \overline{\phi(-t)}$, *where $\bar{z}$ denotes the complex conjugate of z;*
- ***(iv)*** *$\phi(t)$ is real valued if and only if F is symmetric around $x_0 = 0$;*
- ***(v)*** *if $E\{|X|^n\} < \infty$ for some $n \geq 1$, then the rth order derivative $\phi^{(r)}(t)$ exists for every $1 \leq r \leq n$, and*

$$\phi^{(r)}(t) = \int_{-\infty}^{\infty} (ix)^r e^{itx} dF(x), \tag{1.8.28}$$

$$\mu_r = \frac{1}{i^r} \phi^{(r)}(0), \tag{1.8.29}$$

and

$$\phi(t) = \sum_{j=1}^{n} \frac{(it)^j}{j!} \mu_j + \frac{(it)^n}{n!} R_n(t), \tag{1.8.30}$$

where $|R_n(t)| \leq 3E\{|X|^n\}$, $R_n(t) \to 0$ as $t \to 0$;

- ***(vi)*** *if $\phi^{(2n)}(0)$ exists and is finite, then $\mu_{2n} < \infty$;*
- ***(vii)*** *if $E\{|X|^n\} < \infty$ for all $n \geq 1$ and*

$$\overline{\lim_{n\to\infty}} \left(\frac{E\{|X|^n\}}{n!} \right)^{1/n} = \frac{1}{R} < \infty, \tag{1.8.31}$$

then

$$\phi(t) = \sum_{n=0}^{\infty} \frac{(it)^n}{n!}\mu_n, \quad |t| < R. \tag{1.8.32}$$

Proof. The proof of (**i**) and (**ii**) is based on the fact that $|e^{itx}| = 1$ for all t and all x. Now, $\int e^{-itx}dF(x) = \phi(-t) = \overline{\phi(t)}$. Hence (**iii**) is proven.

(**iv**) Suppose $F(x)$ is symmetric around $x_0 = 0$. Then $dF(x) = dF(-x)$ for all x. Therefore, since $\sin(-tx) = -\sin(tx)$ for all x, $\int_{-\infty}^{\infty} \sin(tx)dF(x) = 0$, and $\phi(t)$ is real. If $\phi(t)$ is real, $\phi(t) = \overline{\phi(t)}$. Hence $\phi_X(t) = \phi_{-X}(t)$. Thus, by the one-to-one correspondence between ϕ and F, for any Borel set B, $P\{X \in B\} = P\{-X \in B\} = P\{X \in -B\}$. This implies that F is symmetric about the origin.

(**v**) If $E\{|X|^n\} < \infty$, then $E\{|X|^r\} < \infty$ for all $1 \le r \le n$. Consider

$$\frac{\phi(t+h) - \phi(t)}{h} = E\left\{e^{itX}\left(\frac{e^{ihX} - 1}{h}\right)\right\}.$$

Since $\left|\frac{e^{ihx} - 1}{h}\right| \le |x|$, and $E\{|X|\} < \infty$, we obtain from the Dominated Convergence Theorem that

$$\begin{aligned}\phi^{(1)}(t) &= \lim_{h\to 0}\left(\frac{\phi(t+h) - \phi(t)}{h}\right) \\ &= E\left\{e^{itX}\lim_{h\to 0}\frac{e^{ihX} - 1}{h}\right\} \\ &= iE\{Xe^{itX}\}.\end{aligned}$$

Hence $\mu_1 = \frac{1}{i}\phi^{(1)}(0)$.

Equations (1.8.28)–(1.8.29) follow by induction. Taylor expansion of e^{iy} yields

$$e^{iy} = \sum_{k=0}^{n-1}\frac{(iy)^k}{k!} + \frac{(iy)^n}{n!}(\cos(\theta_1 y) + i\sin(\theta_2 y)),$$

where $|\theta_1| \le 1$ and $|\theta_2| \le 1$. Hence

$$\begin{aligned}\phi(t) &= E\{e^{itX}\} \\ &= \sum_{k=0}^{n-1}\frac{(it)^k}{k!}\mu_k + \frac{(it)^n}{n!}(\mu_n + R_n(t)),\end{aligned}$$

where

$$R_n(t) = E\{X^n(\cos(\theta_1 tX) + i\sin(\theta_2 tX) - 1)\}.$$

Since $|\cos(ty)| \le 1$, $|\sin(ty)| \le 1$, evidently $R_n(t) \le 3E\{|X|^n\}$. Also, by dominated convergence, $\lim_{t\to 0} R_n(t) = 0$.

(**vi**) By induction on n. Suppose $\phi^{(2)}(0)$ exists. By L'Hospital's rule,

$$\begin{aligned}
\phi^{(2)}(0) &= \lim_{h\to 0} \frac{1}{2}\left[\frac{\phi'(2h) - \phi'(0)}{2h} + \frac{\phi'(0) - \phi'(-2h)}{2h}\right] \\
&= \lim_{h\to 0} \frac{1}{4h^2}[\phi(2h) - 2\phi(0) + \phi(-2h)] \\
&= \lim_{h\to 0} \int \left(\frac{e^{ihx} - e^{-ihx}}{2h}\right)^2 dF(x) \\
&= -\lim_{h\to 0} \int \left(\frac{\sin(hx)}{hx}\right)^2 x^2 dF(x).
\end{aligned}$$

By Fatou's Lemma,

$$\begin{aligned}
\phi^{(2)}(0) &\le -\int \lim_{h\to 0} \left(\frac{\sin(hx)}{hx}\right)^2 x^2 dF(x) \\
&= -\int x^2 dF(x) = -\mu_2.
\end{aligned}$$

Thus, $\mu_2 \le -\phi^{(2)}(0) < \infty$. Assume that $0 < \mu_{2k} < \infty$. Then, by (v),

$$\begin{aligned}
\phi^{(2k)}(t) &= \int (ix)^{2k} e^{itx} dF(x) \\
&= (-1)^k \int e^{itx} dG(x),
\end{aligned}$$

where $dG(x) = x^{2k} dF(x)$, or

$$G(x) = \int_{-\infty}^{x} u^{2k} dF(u).$$

Notice that $G(\infty) = \mu_{2k} < \infty$. Thus, $\dfrac{(-1)^k \phi^{(2k)}(t)}{G(\infty)}$ is the characteristic function of the distribution $G(x)/G(\infty)$. Since $\dfrac{1}{G(\infty)} > 0$, $\int x^{2h+2} dF(x) = \int x^2 dG(x) < \infty$. This proves that $\mu_{2k} < \infty$ for all $k = 1, \ldots, n$.

(**vii**) Assuming (1.8.31), if $0 < t_0 < R$, $\overline{\lim_{n\to\infty}} \frac{(E\{|X|^n\})^{1/n}}{n!} < \frac{1}{t_0}$. Therefore,

$$\overline{\lim_{n\to\infty}} \frac{(E\{|X|^n\}t_0^n)^{1/n}}{n!} < 1.$$

By Stirling's approximation, $\lim_{n\to\infty} (n!)^{1/n} = 1$. Thus, for $0 < t_0 < R$,

$$\overline{\lim_{n\to\infty}} \left(\frac{E\{|X|^n\}t_0^n)}{n!} \right)^{1/n} < 1.$$

Accordingly, by Cauchy's test, $\sum_{n=1}^{\infty} \frac{E\{|X|^n\}t_0^n}{n!} < \infty$. By (iv), for any $n \geq 1$, for any t, $|t| \leq t_0$

$$\phi(t) = \sum_{k=0}^{n} \frac{(it)^n}{n!}\mu_n + R_n^*(t),$$

where $|R_n^*(t)| \leq 3\frac{|t|^n}{n!}E\{|X|^n\}$. Thus, for every t, $|t| < R$, $\overline{\lim_{n\to\infty}} |R_n^*(t)| = 0$, which implies that

$$\phi(t) = \sum_{k=1}^{\infty} \frac{(it)^k}{k!}\mu_k, \quad \text{for all } |t| < R.$$

QED

1.9 MODES OF CONVERGENCE

In this section we formulate many definitions and results in terms of random vectors $\mathbf{X} = (X_1, X_2, \cdots, X_k)'$, $1 \leq k < \infty$. The notation $||\mathbf{X}||$ is used for the Euclidean norm, i.e., $||\mathbf{x}||^2 = \sum_{i=1}^{k} x_i^2$.

We discuss here four modes of convergence of sequences of random vectors to a random vector.

- **(i)** Convergence in distribution, $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$;
- **(ii)** Convergence in probability, $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$;
- **(iii)** Convergence in rth mean, $\mathbf{X}_n \xrightarrow{r} \mathbf{X}$; and
- **(iv)** Convergence almost surely, $\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$.

A sequence $\mathbf{X}_n$ *is said to converge in distribution to* $\mathbf{X}$, $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ *if the corresponding distribution functions* F_n *and* F *satisfy*

$$\lim_{n\to\infty} \int g(\mathbf{x}) dF_n(\mathbf{x}) = \int g(\mathbf{x}) dF(\mathbf{x}) \tag{1.9.1}$$

for every continuous bounded function g *on* $\mathbb{R}^k$.

One can show that this definition is equivalent to the following statement.

A sequence $\{\mathbf{X}_n\}$ *converges in distribution to* $\mathbf{X}$, $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ *if* $\lim_{n\to\infty} F_n(\mathbf{x}) = F(\mathbf{x})$ *at all continuity points* $\mathbf{x}$ *of* F.

If $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ we say that F_n converges to F weakly. The notation is $F_n \xrightarrow{w} F$ or $F_n \Rightarrow F$.

We define now convergence in probability.

A sequence $\{\mathbf{X}_n\}$ *converges in probability to* $\mathbf{X}$, $\mathbf{X}_n \xrightarrow[n\to\infty]{p} X$ *if, for each* $\epsilon > 0$,

$$\lim_{n\to\infty} P\{||\mathbf{X}_n - \mathbf{X}|| > \epsilon\} = 0. \tag{1.9.2}$$

We define now convergence in rth mean.

A sequence of random vectors $\{\mathbf{X}_n\}$ *converges in* r*th mean,* $r > 0$, *to* $\mathbf{X}$, $\mathbf{X}_n \xrightarrow{r} \mathbf{X}$ *if* $E\{||\mathbf{X}_n - \mathbf{X}||^r\} \to 0$ *as* $n \to \infty$.

A fourth mode of convergence is

A sequence of random vectors $\{\mathbf{X}_n\}$ *converges almost-surely to* $\mathbf{X}$, $\mathbf{X}_n \xrightarrow{\text{a.s.}} X$, *as* $n \to \infty$ *if*

$$P\{\lim_{n\to\infty} \mathbf{X}_n = \mathbf{X}\} = 1. \tag{1.9.3}$$

The following is an equivalent definition.

$\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$ *as* $n \to \infty$ *if and only if, for any* $\epsilon > 0$,

$$\lim_{n\to\infty} P\{||\mathbf{X}_m - \mathbf{X}|| < \epsilon, \ \forall m \geq n\} = 1. \tag{1.9.4}$$

Equation (1.9.4) is equivalent to

$$P\{\overline{\lim_{n\to\infty}} ||\mathbf{X}_n - \mathbf{X}|| < \epsilon\} = 1.$$

But,

$$\begin{aligned} P\{\overline{\lim_{n\to\infty}} ||\mathbf{X}_n - \mathbf{X}|| < \epsilon\} &= 1 - P\{\overline{\lim_{n\to\infty}} ||\mathbf{X}_n - \mathbf{X}|| \geq \epsilon\} \\ &= 1 - P\{||\mathbf{X}_n - \mathbf{X}|| \geq \epsilon, \ \text{i.o.}\}. \end{aligned}$$

By the Borel–Cantelli Lemma (Theorem 1.4.1), a sufficient condition for $\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$ is

$$\sum_{n=1}^{\infty} P\{||\mathbf{X}_n - \mathbf{X}|| \geq \epsilon\} < \infty \tag{1.9.5}$$

for all $\epsilon > 0$.

Theorem 1.9.1. *Let $\{\mathbf{X}_n\}$ be a sequence of random vectors. Then*

(a) $\mathbf{X}_n \xrightarrow{a.s.} \mathbf{X}$ *implies* $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$.
(b) $\mathbf{X}_n \xrightarrow{r} \mathbf{X}$, $r > 0$, *implies* $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$.
(c) $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$ *implies* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$.

Proof. (a) Since $\mathbf{X}_n \xrightarrow{\text{a.s.}} \mathbf{X}$, for any $\epsilon > 0$,

$$\begin{aligned} 0 &= P\{\overline{\lim_{n\to\infty}} ||\mathbf{X}_n - \mathbf{X}|| \geq \epsilon\} \\ &= \lim_{n\to\infty} P\left\{\bigcup_{m\geq n} ||\mathbf{X}_m - \mathbf{X}|| \geq \epsilon\right\} \qquad (1.9.6) \\ &\geq \lim_{n\to\infty} P\{||\mathbf{X}_n - \mathbf{X}|| \geq \epsilon\}. \end{aligned}$$

The inequality (1.9.6) implies that $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$.

(b) It can be immediately shown that, for any $\epsilon > 0$,

$$E\{||\mathbf{X}_n - \mathbf{X}||^r\} \geq \epsilon^r P\{||\mathbf{X}_n - \mathbf{X}|| \geq \epsilon\}.$$

Thus, $\mathbf{X}_n \xrightarrow{r} \mathbf{X}$ implies $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$.

(c) Let $\epsilon > 0$. If $\mathbf{X}_n \leq \mathbf{x}_0$ then either $\mathbf{X} \leq \mathbf{x}_0 + \epsilon\mathbf{1}$, where $\mathbf{1} = (1, \ldots, 1)'$, or $||\mathbf{X}_n - \mathbf{X}|| > \epsilon$. Thus, for all n,

$$F_n(\mathbf{x}_0) \leq F(\mathbf{x}_0 + \epsilon\mathbf{1}) + P\{||\mathbf{X}_n - \mathbf{X}|| > \epsilon\}.$$

Similarly,

$$F(\mathbf{x}_0 - \epsilon\mathbf{1}) \leq F_n(\mathbf{x}_0) + P\{||\mathbf{X}_n - \mathbf{X}|| > \epsilon\}.$$

Finally, since $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$,

$$F(\mathbf{x}_0 - \epsilon\mathbf{1}) \leq \underline{\lim_{n\to\infty}} F_n(\mathbf{x}_0) \leq \overline{\lim_{n\to\infty}} F_n(\mathbf{x}_0) \leq F(\mathbf{x}_0 + \epsilon\mathbf{1}).$$

Thus, if $\mathbf{x}_0$ is a continuity point of F, by letting $\epsilon \to 0$, we obtain

$$\lim_{n\to\infty} F_n(\mathbf{x}_0) = F(\mathbf{x}_0).$$

QED

Theorem 1.9.2. *Let $\{\mathbf{X}_n\}$ be a sequence of random vectors. Then*

- ***(a)*** *if $\mathbf{c} \in \mathbb{R}^k$, then $\mathbf{X}_n \xrightarrow{d} \mathbf{c}$ implies $\mathbf{X}_n \xrightarrow{p} \mathbf{c}$;*
- ***(b)*** *if $\mathbf{X}_n \xrightarrow{a.s.} \mathbf{X}$ and $||\mathbf{X}_n||^r \le Z$, for some $r > 0$ and some (positive) random variable Z, with $E\{Z\} < \infty$, then $\mathbf{X}_n \xrightarrow{r} \mathbf{X}$.*

For proof, see Ferguson (1996, p. 9). Part (b) is implied also from Theorem 1.13.3.

Theorem 1.9.3. *Let $\{X_n\}$ be a sequence of nonnegative random variables such that $X_n \xrightarrow{a.s.} X$ and $E\{X_n\} \to E\{X\}$, $E\{X\} < \infty$. Then*

$$E\{|X_n - X|\} \to 0, \quad as \ n \to \infty. \tag{1.9.7}$$

Proof. Since $E\{X_n\} \to E\{X\} < \infty$, for sufficiently large n, $E\{X_n\} < \infty$. For such n,

$$\begin{aligned} E\{|X_n - X|\} &= E\{(X - X_n)I\{X \ge X_n\}\} + E\{(X_n - X)I\{X_n > X\}\} \\ &= 2E\{(X - X_n)I\{X \ge X_n\}\} + E\{X - X_n\}. \end{aligned}$$

But,

$$0 \le (X - X_n)I\{X \ge X_n\} < X.$$

Therefore, by the Lebesgue Dominated Convergence Theorem,

$$\lim_{n\to\infty} E\{(X - X_n)I\{X \ge X_n\}\} = 0.$$

This implies (1.9.7). QED

1.10 WEAK CONVERGENCE

The following theorem plays a major role in weak convergence.

Theorem 1.10.1. *The following conditions are equivalent.*

- ***(a)*** $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$;
- ***(b)*** $E\{g(\mathbf{X}_n)\} \to E\{g(\mathbf{X})\}$, *for all continuous functions,* g, *that vanish outside a compact set;*
- ***(c)*** $E\{g(\mathbf{X}_n)\} \to E\{g(\mathbf{X})\}$, *for all continuous bounded functions* g;
- ***(d)*** $E\{g(\mathbf{X}_n)\} \to E\{g(\mathbf{X})\}$, *for all measurable functions* g *such that* $P\{\mathbf{X} \in C(g)\} = 1$, *where* $C(g)$ *is the set of all points at which* g *is continuous.*

For proof, see Ferguson (1996, pp. 14–16).

Theorem 1.10.2. *Let* $\{\mathbf{X}_n\}$ *be a sequence of random vectors in* $\mathbb{R}^k$, *and* $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$. *Then*

- ***(i)*** $\mathbf{f}(\mathbf{X}_n) \xrightarrow{d} \mathbf{f}(\mathbf{X})$;
- ***(ii)*** *if* $\{\mathbf{Y}_n\}$ *is a sequence such that* $\mathbf{X}_n - \mathbf{Y}_n \xrightarrow{p} \mathbf{0}$, *then* $\mathbf{Y}_n \xrightarrow{d} \mathbf{X}$;
- ***(iii)*** *if* $\mathbf{X}_n \in \mathbb{R}^k$ *and* $\mathbf{Y}_n \in \mathbb{R}^l$ *and* $\mathbf{Y}_n \xrightarrow{d} \mathbf{c}$, *then*

$$\begin{pmatrix} \mathbf{X}_n \\ \mathbf{Y}_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \mathbf{X} \\ \mathbf{c} \end{pmatrix}.$$

Proof. (**i**) Let $g : \mathbb{R}^l \to \mathbb{R}$ be bounded and continuous. Let $h(\mathbf{x}) = g(\mathbf{f}(\mathbf{x}))$. If $\mathbf{x}$ is a continuity point of $\mathbf{f}$, then $\mathbf{x}$ is a continuity point of h, i.e., $C(\mathbf{f}) \subset C(h)$. Hence $P\{X \in C(h)\} = 1$. By Theorem 1.10.1 (c), it is sufficient to show that $E\{g(\mathbf{f}(\mathbf{X}_n))\} \to E\{g(\mathbf{f}(\mathbf{X}))\}$. Theorem 1.10.1 (d) implies, since $P\{X \in C(h)\} = 1$ and $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$, that $E\{h(\mathbf{X}_n)\} \to E\{h(\mathbf{X})\}$.

(**ii**) According to Theorem 1.10.1 (b), let g be a continuous function on $\mathbb{R}^k$ vanishing outside a compact set. Thus g is uniformly continuous and bounded. Let $\epsilon > 0$, find $\delta > 0$ such that, if $||\mathbf{x} - \mathbf{y}|| < \delta$ then $|g(\mathbf{x}) - g(\mathbf{y})| < \epsilon$. Also, g is bounded, say $|g(x)| \le B < \infty$. Thus,

$$\begin{aligned} |E\{g(\mathbf{Y}_n)\} - E\{g(\mathbf{X})\}| &\le |E\{g(\mathbf{Y}_n)\} - E\{g(\mathbf{X}_n)\}| + |E\{g(\mathbf{X}_n)\} - E\{g(\mathbf{X})\}| \\ &\le E\{|g(\mathbf{Y}_n) - g(\mathbf{X}_n)| I\{||\mathbf{X}_n - \mathbf{Y}_n|| \le \delta\}\} \\ &+ E\{|g(\mathbf{Y}_n) - g(\mathbf{X}_n)| I\{||\mathbf{X}_n - \mathbf{Y}_n|| > \delta\} \\ &+ |E\{g(\mathbf{X}_n)\} - E\{g(\mathbf{X})\}| \\ &\le \epsilon + 2B P\{||\mathbf{X}_n - \mathbf{Y}_n|| > \delta\} \\ &+ |E\{g(\mathbf{X}_n)\} - E\{g(\mathbf{X})\}| \underset{n\to\infty}{\longrightarrow} \epsilon. \end{aligned}$$

Hence $\mathbf{Y}_n \xrightarrow{d} \mathbf{X}$.

(**iii**)

$$P\left\{\left|\left|\begin{pmatrix}\mathbf{X}_n\\ \mathbf{Y}_n\end{pmatrix}-\begin{pmatrix}\mathbf{X}_n\\ \mathbf{c}\end{pmatrix}\right|\right|>\epsilon\right\}=P\{||\mathbf{Y}_n-\mathbf{c}||>\epsilon\}\to 0,\quad \text{as } n\to\infty.$$

Hence, from part (ii), $\begin{pmatrix}\mathbf{X}_n\\ \mathbf{Y}_n\end{pmatrix} \xrightarrow{d} \begin{pmatrix}\mathbf{X}\\ \mathbf{c}\end{pmatrix}$. QED

As a special case of the above theorem we get

Theorem 1.10.3 (Slutsky's Theorem). *Let $\{X_n\}$ and $\{Y_n\}$ be sequences of random variables, $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$. Then*

$$\begin{aligned}
&\textit{(i)} \quad X_n + Y_n \xrightarrow{d} X + c;\\
&\textit{(ii)} \quad X_n Y_n \xrightarrow{d} cX;\\
&\textit{(iii)} \quad \textit{if } c \neq 0 \textit{ then } \frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}.
\end{aligned} \tag{1.10.1}$$

A sequence of distribution functions may not converge to a distribution function. For example, let X_n be random variables with

$$F_n(x) = \begin{cases} 0, & x < -n \\ \dfrac{1}{2}, & -n \le x < n \\ 1, & n \le x. \end{cases}$$

Then, $\lim_{n\to\infty} F_n(x) = \frac{1}{2}$ for all x. $F(x) = \frac{1}{2}$ for all x is not a distribution function. In this example, half of the probability mass escapes to $-\infty$ and half the mass escapes to $+\infty$. In order to avoid such situations, we require from collections (families) of probability distributions to be **tight**.

Let $\mathcal{F} = \{F_u, u \in \mathcal{U}\}$ be a family of distribution functions on $\mathbb{R}^k$. $\mathcal{F}$ is **tight** if, for any $\epsilon > 0$, there exists a **compact** set $C \subset \mathbb{R}^k$ such that

$$\sup_{u\in\mathcal{U}} \int I\{\mathbf{x} \in \mathbb{R}^k - C\} dF_u(\mathbf{x}) < \epsilon.$$

In the above, the sequence $F_n(x)$ is not tight.

If $\mathcal{F}$ is **tight**, then every sequence of distributions of $\mathcal{F}$ contains a subsequence converging weakly to a distribution function. (see Shiryayev, 1984, p. 315).

Theorem 1.10.4. *Let $\{F_n\}$ be a tight family of distribution functions on $\mathbb{R}$. A necessary and sufficient condition for $F_n \Rightarrow F$ is that, for each $t \in \mathbb{R}$, $\lim_{n\to\infty} \phi_n(t)$ exists, where $\phi_n(t) = \int e^{itx} dF_n(x)$ is the characteristic function corresponding to F_n.*

For proof, see Shiryayev (1984, p. 321).

Theorem 1.10.5 (Continuity Theorem). *Let $\{F_n\}$ be a sequence of distribution functions and $\{\phi_n\}$ the corresponding sequence of characteristic functions. Let F be a distribution function, with characteristic function ϕ. Then $F_n \Rightarrow F$ if and only if $\phi_n(\mathbf{t}) \to \phi(\mathbf{t})$ for all $\mathbf{t} \in \mathbb{R}^k$. (Shiryayev, 1984, p. 322).*

1.11 LAWS OF LARGE NUMBERS

1.11.1 The Weak Law of Large Numbers (WLLN)

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots$ be a sequence of identically distributed uncorrelated random vectors. Let $\boldsymbol{\mu} = E\{\mathbf{X}_1\}$ and let $\mathfrak{\Sigma} = E\{(\mathbf{X}_1 - \mu)(\mathbf{X}_1 - \mu)'\}$ be finite. Then the means $\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$ converge in probability to $\boldsymbol{\mu}$, i.e.,

$$\bar{\mathbf{X}}_n \xrightarrow{p} \mu \quad \text{as } n \to \infty. \bar{\mathbf{X}}_n \xrightarrow{p} \boldsymbol{\mu} \quad \text{as } n \to \infty. \tag{1.11.1}$$

The proof is simple. Since $\text{cov}(\mathbf{X}_n, \mathbf{X}_{n'}) = \mathbf{0}$ for all $n \neq n'$, the covariance matrix of $\bar{\mathbf{X}}_n$ is $\frac{1}{n}\mathfrak{\Sigma}$. Moreover, since $E\{\bar{\mathbf{X}}_n\} = \mu$,

$$E\{||\bar{\mathbf{X}}_n - \boldsymbol{\mu}||^2\} = \frac{1}{n} tr.\{\mathfrak{\Sigma}\} \to 0 \quad \text{as } n \to \infty.$$

Hence $\bar{\mathbf{X}}_n \xrightarrow{2} \mu$, which implies that $\bar{\mathbf{X}}_n \xrightarrow{p} \mu$. Here $tr.\{\mathfrak{\Sigma}\}$ denotes the trace of $\mathfrak{\Sigma}$.

If $\mathbf{X}_1, \mathbf{X}_2, \ldots$ are independent, and identically distributed, with $E\{\mathbf{X}_1\} = \boldsymbol{\mu}$, then the characteristic function of $\bar{\mathbf{X}}_n$ is

$$\phi_{\bar{\mathbf{X}}_n}(\mathbf{t}) = \left(\phi\left(\frac{\mathbf{t}}{n}\right)\right)^n, \tag{1.11.2}$$

where $\phi(\mathbf{t})$ is the characteristic function of $\mathbf{X}_1$. Fix $\mathbf{t}$. Then for large values of n,

$$\phi\left(\frac{\mathbf{t}}{n}\right) = 1 + \frac{i}{n}\mathbf{t}'\boldsymbol{\mu} + o\left(\frac{1}{n}\right), \quad \text{as } n \to \infty.$$

Therefore,

$$\phi_{\bar{\mathbf{X}}_n}(\mathbf{t}) = \left(1 + \frac{i}{n}\mathbf{t}'\boldsymbol{\mu} + o\left(\frac{1}{n}\right)\right)^n \to e^{i\mathbf{t}'\boldsymbol{\mu}}. \tag{1.11.3}$$

$\phi(\mathbf{t}) = e^{i\mathbf{t}'\mu}$ is the characteristic function of $\mathbf{X}$, where $P\{\mathbf{X} = \boldsymbol{\mu}\} = 1$. Thus, since $e^{i\mathbf{t}'\boldsymbol{\mu}}$ is continuous at $\mathbf{t} = \mathbf{0}$, $\bar{\mathbf{X}}_n \xrightarrow{d} \boldsymbol{\mu}$. This implies that $\bar{\mathbf{X}}_n \xrightarrow{p} \mu$ (left as an exercise).

1.11.2 The Strong Law of Large Numbers (SLLN)

Strong laws of large numbers, for independent random variables having finite expected values are of the form

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i) \xrightarrow{\text{a.s.}} 0, \quad \text{as } n \to \infty,$$

where $\mu_i = E\{X_i\}$.

Theorem 1.11.1 (Cantelli). *Let $\{X_n\}$ be a sequence of independent random variables having uniformly bounded fourth-central moments, i.e.,*

$$0 \le E(X_n - \mu_n)^4 \le C < \infty \tag{1.11.4}$$

for all $n \ge 1$. Then

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i) \xrightarrow[n\to\infty]{a.s.} 0. \tag{1.11.5}$$

Proof. Without loss of generality, we can assume that $\mu_n = E\{X_n\} = 0$ for all $n \ge 1$.

$$\begin{aligned} E\left\{\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^4\right\} &= \frac{1}{n^4}\left\{\sum_{i=1}^{n}E\{X_i^4\}\right. \\ &\quad + 4\sum\sum_{i\neq j}E\{X_i^3X_j\} + 3\sum\sum_{i\neq j}E\{X_i^2X_j^2\} \\ &\quad \left. + 6\sum\sum\sum_{i\neq j\neq k}E\{X_i^2X_jX_k\} + \sum\sum\sum\sum_{i\neq j\neq k\neq l}E\{X_iX_jX_kX_l\}\right\} \\ &= \frac{1}{n^4}\sum_{i=1}^{n}\mu_{4,i} + \frac{3}{n^4}\sum\sum_{i\neq j}\sigma_i^2\sigma_j^2, \end{aligned}$$

where $\mu_{4,i} = E\{X_i^4\}$ and $\sigma_i^2 = E\{X_i^2\}$. By the Schwarz inequality, $\sigma_i^2\sigma_j^2 \leq (\mu_{4,i} \cdot \mu_{4j})^{1/2}$ for all $i \neq j$. Hence,

$$E\{\bar{X}_n^4\} \leq \frac{C}{n^3} + \frac{3n(n-1)C}{n^4} = O\left(\frac{1}{n^2}\right).$$

By Chebychev's inequality,

$$\begin{aligned} P\{|\bar{X}_n| \geq \epsilon\} &= P\{\bar{X}_n^4 \geq \epsilon^4\} \\ &\leq \frac{E\{\bar{X}_n^4\}}{\epsilon^4}. \end{aligned}$$

Hence, for any $\epsilon > 0$,

$$\sum_{n=1}^{\infty} P\{|\bar{X}_n| \geq \epsilon\} \leq C^* \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty,$$

where C^* is some positive finite constant. Finally, by the Borel–Cantelli Lemma (Theorem 1.4.1),

$$P\{|\bar{X}_n| \geq \epsilon, \text{i.o.}\} = 0.$$

Thus, $P\{|\bar{X}_n| < \epsilon, \text{i.o.}\} = 1$. QED

Cantelli's Theorem is quite stringent, in the sense, that it requires the existence of the fourth moments of the independent random variables. Kolmogorov had relaxed this condition and proved that, if the random variables have finite variances, $0 < \sigma_n^2 < \infty$ and

$$\sum_{n=1}^{\infty} \frac{\sigma_n^2}{n^2} < \infty, \tag{1.11.6}$$

then $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i) \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$.

If the random variables are **independent** and **identically distributed** (i.i.d.), then Kolmogorov showed that $E\{|X_1|\} < \infty$ is sufficient for the strong law of large numbers. To prove Kolmogorov's strong law of large numbers one has to develop more theoretical results. We refer the reader to more advanced probability books (see Shiryayev, 1984).

1.12 CENTRAL LIMIT THEOREM

The Central Limit Theorem (CLT) states that, under general valid conditions, the distributions of properly normalized sample means converge weakly to the standard normal distribution.

A continuous random variable Z is said to have a standard normal distribution, and we denote it $Z \sim N(0, 1)$ if its distribution function is absolutely continuous, having a p.d.f.

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \quad -\infty < x < \infty. \tag{1.12.1}$$

The c.d.f. of $N(0, 1)$, called the **standard normal integral** is

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}y^2} dy. \tag{1.12.2}$$

The general family of normal distributions is studied in Chapter 2. Here we just mention that if $Z \sim N(0, 1)$, the moments of Z are

$$\mu_r = \begin{cases} \dfrac{(2k)!}{2^k k!}, & \text{if } r = 2k \\ 0, & \text{if } r = 2k + 1. \end{cases} \tag{1.12.3}$$

The characteristic function of $N(0, 1)$ is

$$\begin{aligned} \phi(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2 + itx} dx \\ &= e^{-\frac{1}{2}t^2}, \quad -\infty < t < \infty. \end{aligned} \tag{1.12.4}$$

A random vector $\bar{Z} = (Z_1, \ldots, Z_k)'$ is said to have a multivariate normal distribution with mean $\boldsymbol{\mu} = E\{\mathbf{Z}\} = \mathbf{0}$ and covariance matrix V (see Chapter 2), $\mathbf{Z} \sim N(\mathbf{0}, V)$ if the p.d.f. of $\mathbf{Z}$ is

$$f(\mathbf{Z}; V) = \frac{1}{(2\pi)^{k/2} |V|^{1/2}} \exp\left\{ -\frac{1}{2} \mathbf{Z}' V^{-1} \mathbf{Z} \right\}.$$

The corresponding characteristic function is

$$\phi_{\mathbf{Z}}(\mathbf{t}) = \exp\left\{ -\frac{1}{2} \mathbf{t}' V \mathbf{t} \right\}, \tag{1.12.5}$$

$\mathbf{t} \in \mathbb{R}^k$.

Using the method of characteristic functions, with the continuity theorem we prove the following simple two versions of the CLT. A proof of the Central Limit Theorem, which is not based on the continuity theorem of characteristic functions, can be obtained by the method of Stein (1986) for approximating expected values or probabilities.

Theorem 1.12.1 (CLT). *Let $\{X_n\}$ be a sequence of i.i.d. random variables having a finite positive variance, i.e., $\mu = E\{X_1\}$, $V\{X_1\} = \sigma^2$, $0 < \sigma^2 < \infty$. Then*

$$\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow{d} N(0,1), \quad as \ \ n \to \infty. \tag{1.12.6}$$

Proof. Notice that $\sqrt{n}\,\dfrac{\bar{X}_n - \mu}{\sigma} = \dfrac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i$, where $Z_i = \dfrac{X_i - \mu}{\sigma}$, $i \geq 1$. Moreover, $E\{Z_i\} = 0$ and $V\{Z_i\} = 1, i \geq 1$. Let $\phi_Z(t)$ be the characteristic function of Z_1. Then, since $E\{Z\} = 0$, $V\{Z\} = 1$, (1.8.33) implies that

$$\phi_Z(t) = 1 - \frac{t^2}{2} + o(t), \quad \text{as} \ \ t \to 0.$$

Accordingly, since $\{Z_n\}$ are i.i.d.,

$$\begin{aligned}\phi_{\sqrt{n}\,\bar{Z}_n}(t) &= \phi_Z^n\left(\frac{t}{\sqrt{n}}\right) \\ &= \left(1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \\ &\to e^{-t^2/2} \ \text{ as } \ n \to \infty.\end{aligned}$$

Hence, $\sqrt{n}\,\bar{Z}_n \xrightarrow{d} N(0,1)$. QED

Theorem 1.12.1 can be generalized to random vector. Let $\bar{\mathbf{X}}_n = \dfrac{1}{n}\sum_{j=1}^{n} \mathbf{X}_j$, $n \geq 1$. The generalized CLT is the following theorem.

Theorem 1.12.2. *Let $\{\mathbf{X}_n\}$ be a sequence of i.i.d. random vectors with $E\{\mathbf{X}_n\} = 0$, and covariance matrix $E\{\mathbf{X}_n \mathbf{X}'_n\} = \mathbf{V}$, $n \geq 1$, where $\mathbf{V}$ is positive definite with finite eigenvalues. Then*

$$\sqrt{n}\,\bar{\mathbf{X}}_n \xrightarrow{d} N(\mathbf{0}, \mathbf{V}). \tag{1.12.7}$$

Proof. Let $\phi_{\mathbf{X}}(\mathbf{t})$ be the characteristic function of $\mathbf{X}_1$. Then, since $E\{\mathbf{X}_1\} = \mathbf{0}$,

$$\begin{aligned}\phi_{\sqrt{n}\,\bar{X}_n}(\mathbf{t}) &= \phi_{\mathbf{X}}^n\left(\frac{\mathbf{t}}{\sqrt{n}}\right) \\ &= \left(1 - \frac{1}{2n}\mathbf{t}'V\mathbf{t} + o\left(\frac{\mathbf{t}}{\sqrt{n}}\right)\right)^n\end{aligned}$$

as $n \to \infty$. Hence

$$\lim_{n\to\infty} \phi_{\sqrt{n}\,\bar{X}_n}(\mathbf{t}) = \exp\left\{-\frac{1}{2}\mathbf{t}'\mathbf{V}\mathbf{t}\right\}, \quad \mathbf{t} \in \mathbb{R}^k.$$

QED

When the random variables are independent but not identically distributed, we need a stronger version of the CLT. The following celebrated CLT is sufficient for most purposes.

Theorem 1.12.3 (Lindeberg–Feller). *Consider a triangular array of random variables* $\{X_{n,k}\}$, $k = 1, \ldots, n$, $n \geq 1$ *such that, for each* $n \geq 1$, $\{X_{n,k}, k = 1, \ldots, n\}$ *are independent, with* $E\{X_{n,k}\} = 0$ *and* $V\{X_{n,k}\} = \sigma_{n,k}^2$. *Let* $S_n = \sum_{k=1}^{n} X_{n,k}$ *and* $B_n^2 = \sum_{k=1}^{n} \sigma_{n,k}^2$. *Assume that* $B_n > 0$ *for each* $n \geq 1$, *and* $B_n \nearrow \infty$, *as* $n \to \infty$. *If, for every* $\epsilon > 0$,

$$\frac{1}{B_n^2} \sum_{k=1}^{n} E\{X_{n,k}^2 I\{|X_{n,k}| > \epsilon B_n\}\} \to 0 \tag{1.12.8}$$

as $n \to \infty$, *then* $S_n/B_n \xrightarrow{d} N(0,1)$ *as* $n \to \infty$. *Conversely, if* $\max_{1\leq k\leq n} \frac{\sigma_{n,k}^2}{B_n^2} \to 0$ *as* $n \to \infty$ *and* $S_n/B_n \xrightarrow{d} N(0,1)$, *then (1.12.8) holds.*

For a proof, see Shiryayev (1984, p. 326). The following theorem, known as Lyapunov's Theorem, is weaker than the Lindeberg–Feller Theorem, but is often sufficient to establish the CLT.

Theorem 1.12.4 (Lyapunov). *Let $\{X_n\}$ be a sequence of independent random variables. Assume that $E\{X_n\} = 0$, $V\{X_n\} > 0$ and $E\{|X_n|^3\} < \infty$, for all $n \geq 1$. Moreover, assume that $B_n^2 = \sum_{j=1}^{n} V\{X_j\} \nearrow \infty$. Under the condition*

$$\frac{1}{B_n^3} \sum_{j=1}^{n} E\{|X_j|^3\} \to 0 \ \ as \ \ n \to \infty, \tag{1.12.9}$$

the CLT holds, i.e., $S_n/B_n \xrightarrow{d} N(0,1)$ as $n \to \infty$.

Proof. It is sufficient to prove that (1.12.9) implies the Lindberg–Feller condition (1.12.8). Indeed,

$$\begin{aligned} E\{|X_j|^3\} &= \int_{-\infty}^{\infty} |x|^3 dF_j(x) \\ &\geq \int_{\{x:|x|>\epsilon B_n\}} |x|^3 dF_j(x) \\ &\geq \epsilon B_n \int_{\{x:|x|>B_n\epsilon\}} x^2 dF_j(x). \end{aligned}$$

Thus,

$$\frac{1}{B_n^2} \sum_{j=1}^{n} \int_{\{x:|x|>\epsilon B_n\}} x^2 dF_j(x) \leq \frac{1}{\epsilon} \cdot \frac{1}{B_n^3} \sum_{j=1}^{n} E\{|X_j|^3\} \to 0.$$

QED

Stein (1986, p. 97) proved, using a novel approximation to expectation, that if $X_1, X_2, \ldots$ are independent and identically distributed, with $EX_1 = 0$, $EX_1^2 = 1$ and $\gamma = E\{|X_1|^3\} < \infty$, then, for all $-\infty < x < \infty$ and all $n = 1, 2, \ldots$,

$$\left| P\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \leq x \right\} - \Phi(x) \right| \leq \frac{6\gamma}{\sqrt{n}},$$

where $\Phi(x)$ is the c.d.f. of $N(0,1)$. This immediately implies the CLT and shows that the convergence is uniform in x.

1.13 MISCELLANEOUS RESULTS

In this section we review additional results.

1.13.1 Law of the Iterated Logarithm

We denote by $\log_2(x)$ the function $\log(\log(x))$, $x > e$.

Theorem 1.13.1. *Let $\{X_n\}$ be a sequence of i.i.d. random variables, such that $E\{X_1\} = 0$ and $V\{X_1\} = \sigma^2$, $0 < \sigma < \infty$. Let $S_n = \sum_{i=1}^{n} X_i$. Then*

$$P\left\{\overline{\lim_{n\to\infty}} \frac{|S_n|}{\psi(n)} = 1\right\} = 1, \tag{1.13.1}$$

where $\psi(n) = (2\sigma^2 n \log_2(n))^{1/2}$, $n \geq 3$.

For proof, in the normal case, see Shiryayev (1984, p. 372).

The theorem means the sequence $|S_n|$ will cross the boundary $\psi(n)$, $n \geq 3$, only a finite number of times, with probability 1, as $n \to \infty$. Notice that although $E\{S_n\} = 0$, $n \geq 1$, the variance of S_n is $V\{S_n\} = n\sigma^2$ and $P\{|S_n| \nearrow \infty\} = 1$. However, if we consider $\frac{S_n}{n}$ then by the SLLN, $\frac{S_n}{n} \xrightarrow{\text{a.s.}} 0$. If we divide only by $\sqrt{n}$ then, by the CLT, $\frac{S_n}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1)$. The law of the iterated logarithm says that, for every $\epsilon > 0$, $P\left\{\frac{|S_n|}{\sigma\sqrt{n}} > (1+\epsilon)\sqrt{2\log_2(n)}, i.o.\right\} = 0$. This means, that the fluctuations of S_n are not too wild. In Example 1.19 we see that if $\{X_n\}$ are i.i.d. with $P\{X_1 = 1\} = P\{X_1 = -1\} = \frac{1}{2}$, then $\frac{S_n}{n} \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$. But n goes to infinity faster than $\sqrt{n\log_2(n)}$. Thus, by (1.13.1), if we consider the sequence $W_n = \frac{S_n}{\sqrt{2n\log_2(n)}}$ then $P\{|W_n| < 1 + \epsilon, i.o.\} = 1$. $\{W_n\}$ fluctuates between -1 and 1 almost always.

1.13.2 Uniform Integrability

A sequence of random variables $\{X_n\}$ is **uniformly integrable** if

$$\lim_{c\to\infty} \sup_{n\geq 1} E\{|X_n| I\{|X_n| > c\}\} = 0. \tag{1.13.2}$$

Clearly, if $|X_n| \leq Y$ for all $n \geq 1$ and $E\{Y\} < \infty$, then $\{X_n\}$ is a uniformly integrable sequence. Indeed, $|X_n| I\{|X_n| > c\} \leq |Y| I\{|Y| > c\}$ for all $n \geq 1$. Hence,

$$\sup_{n\geq 1} E\{|X_n| I\{|X_n| > c\}\} \leq E\{|Y| I\{|Y| > c\}\} \to 0$$

as $c \to \infty$ since $E\{Y\} < \infty$.

Theorem 1.13.2. *Let $\{X_n\}$ be uniformly integrable. Then,*

(i) $$E\{\underline{\lim}_{n\to\infty} X_n\} \le \underline{\lim}_{n\to\infty} E\{X_n\} \le \overline{\lim}_{n\to\infty} E\{X_n\} \le E\{\overline{\lim}_{n\to\infty} X_n\}; \quad (1.13.3)$$

(ii) *if in addition $X_n \xrightarrow{a.s.} X$, as $n \to \infty$, then X is integrable and*

$$\lim_{n\to\infty} E\{X_n\} = E\{X\}, \quad (1.13.4)$$

$$\lim_{n\to\infty} E\{|X_n - X|\} = 0. \quad (1.13.5)$$

Proof. (**i**) For every $c > 0$

$$E\{X_n\} = E\{X_n I\{X_n < -c\}\} + E\{X_n I\{X_n \ge -c\}\}. \quad (1.13.6)$$

By uniform integrability, for every $\epsilon > 0$, take c sufficiently large so that

$$\sup_{n\ge 1} |E\{X_n I\{X_n < -c\}\}| < \epsilon.$$

By Fatou's Lemma (Theorem 1.6.2),

$$\underline{\lim}_{n\to\infty} E\{X_n I\{X_n \ge -c\}\} \ge E\left\{\underline{\lim}_{n\to\infty} X_n I\{X_n \ge -c\}\right\}. \quad (1.13.7)$$

But $X_n I\{X_n \ge -c\}\} \ge X_n$. Therefore,

$$\underline{\lim}_{n\to\infty} E\{X_n I\{X_n \ge -c\}\} \ge E\{\underline{\lim}_{n\to\infty} X_n\}. \quad (1.13.8)$$

From (1.13.6)–(1.13.8), we obtain

$$\underline{\lim}_{n\to\infty} E\{X_n\} \ge E\left\{\underline{\lim}_{n\to\infty} X_n\right\} - \epsilon. \quad (1.13.9)$$

In a similar way, we show that

$$\overline{\lim}_{n\to\infty} E\{X_n\} \le E\{\overline{\lim} X_n\} + \epsilon. \quad (1.13.10)$$

Since ϵ is arbitrary we obtain (1.13.3). Part (ii) is obtained from (i) as in the Dominated Convergence Theorem (Theorem 1.6.3). QED

Theorem 1.13.3. *Let $X_n \ge 0$, $n \ge 1$, and $X_n \xrightarrow{a.s.} X$, $E\{X_n\} < \infty$. Then $E\{X_n\} \to E\{X\}$ if and only if $\{X_n\}$ is uniformly integrable.*

Proof. The sufficiency follows from part (ii) of the previous theorem.

To prove necessity, let

$$A = \{a : F_X(a) - F_X(a-) > 0\}.$$

Then, for each $c \notin A$

$$X_n I\{X_n < c\} \xrightarrow{\text{a.s.}} X I\{X < c\}.$$

The family $\{X_n I\{X_n < c\}\}$ is uniformly integrable. Hence, by sufficiency,

$$\lim_{n\to\infty} E\{X_n I\{X_n < c\}\} = E\{X I\{X < c\}\}$$

for $c \notin A$, $n \to \infty$. A has a countable number of jump points. Since $E\{X\} < \infty$, we can choose $c_0 \notin A$ sufficiently large so that, for a given $\epsilon > 0$, $E\{X I\{X \geq c_0\}\} < \dfrac{\epsilon}{2}$. Choose $N_0(\epsilon)$ sufficiently large so that, for $n \geq N_0(\epsilon)$,

$$E\{X_n I\{X_n \geq c_0\}\} \leq E\{X I\{X \geq c_0\}\} + \frac{\epsilon}{2}.$$

Choose $c_1 > c_0$ sufficiently large so that $E\{X_n I\{X_n \geq c_1\}\} \leq \epsilon$, $n \leq N_0$. Then $\sup_n E\{X_n I\{X_n \geq c_1\}\} \leq \epsilon$. QED

Lemma 1.13.1. *If $\{X_n\}$ is a sequence of uniformly integrable random variables, then*

$$\sup_{n\geq 1} E\{|X_n|\} < \infty. \tag{1.13.11}$$

Proof.

$$\begin{aligned}
\sup_{n\geq 1} E\{|X|_n\} &= \sup_{n\geq 1}(E\{|X_n| I\{|X_n| > c\}\} + E\{|X_n| I\{|X_n| \leq c\}\}) \\
&\leq \sup_{n\geq 1} E\{|X_n| I\{|X_n| > c\}\} + \sup_{n\geq 1} E\{|X_n| I\{|X_n| \leq c\}\} \\
&\leq \epsilon + c,
\end{aligned}$$

for $0 < c < \infty$ sufficiently large. QED

Theorem 1.13.4. *A necessary and sufficient condition for a sequence* $\{X_n\}$ *to be uniformly integrable is that*

$$\sup_{n\geq 1} E\{|X_n|\} \leq B < \infty \tag{1.13.12}$$

and

$$\sup_{n\geq 1} E\{|X_n| I_A\} \to 0 \ \textit{when} \ \ P\{A\} \to 0. \tag{1.13.13}$$

Proof. (**i**) **Necessity**: Condition (1.13.12) was proven in the previous lemma. Furthermore, for any $0 < c < \infty$,

$$\begin{aligned} E\{|X_n| I_A\} &= E\{|X_n| I\{A \cap \{|X_n| \geq c\}\}\} \\ &\quad + E\{|X_n| I\{A \cap \{|X_n| < c\}\}\} \\ &\leq E\{|X_n| I\{|X_n| \geq c\}\} + cP(A). \end{aligned} \tag{1.13.14}$$

Choose c sufficiently large, so that $E\{|X_n| I\{|X_n| \geq c\}\} < \frac{\epsilon}{2}$ and A so that $P\{A\} < \frac{\epsilon}{2c}$, then $E\{|X_n| I_A\} < \epsilon$. This proves the necessity of (1.13.13).

(**ii**) **Sufficiency**: Let $\epsilon > 0$ be given. Choose $\delta(\epsilon)$ so that $P\{A\} < \delta(\epsilon)$, and $\sup_{n\geq 1} E\{|X_n| I_A\} \leq \epsilon$.

By Chebychev's inequality, for every $c > 0$,

$$P\{|X_n| \geq c\} \leq \frac{E\{|X_n|\}}{c}, \quad n \geq 1.$$

Hence,

$$\sup_{n\geq 1} P\{|X_n| \geq c\} \leq \frac{1}{c} \sup_{n\geq 1} E\{|X_n|\} \leq \frac{B}{c}. \tag{1.13.15}$$

The right-hand side of (1.13.15) goes to zero, when $c \to \infty$. Choose c sufficiently large so that $P\{|X_n| \geq c\} < \epsilon$. Such a value of c exists, independently of n, due to (1.13.15). Let $A = \left\{ \bigcup_{n=1}^{\infty} |X_n| \geq c \right\}$. For sufficiently large c, $P\{A\} < \epsilon$ and, therefore,

$$\sup_{n\geq 1} E\{|X_n| I\{|X_n| \geq c\}\} \leq E\{|X_n| I_A\} \to 0$$

as $c \to \infty$. This establishes the uniform integrability of $\{X_n\}$. QED

Notice that according to Theorem 1.13.3, if $E|X_n|^r < \infty$, $r \geq 1$ and $X_n \xrightarrow{\text{a.s.}} X$, $\lim_{n\to\infty} E\{X_n^r\} = E\{X^r\}$ if and only if $\{X_n\}$ is a uniformly integrable sequence.

1.13.3 Inequalities

In previous sections we established several inequalities. The Chebychev inequality, the Kolmogorov inequality. In this section we establish some useful additional inequalities.

1. **The Schwarz Inequality**

Let (X, Y) be random variables with joint distribution function F_{XY} and marginal distribution functions F_X and F_Y, respectively. Then, for every Borel measurable and integrable functions g and h, such that $E\{g^2(X)\} < \infty$ and $E\{h^2(Y)\} < \infty$,

$$\left| \int g(x)h(y)dF_{XY}(x, y) \right| \leq \left(\int g^2(x)dF_X(x) \right)^{1/2} \left(\int h^2(y)dF_Y(y) \right)^{1/2}. \quad (1.13.16)$$

To prove (1.13.16), consider the random variable $Q(t) = (g(X) + th(Y))^2$, $-\infty < t < \infty$. Obviously, $Q(t) \geq 0$, for all t, $-\infty < t < \infty$. Moreover,

$$E\{Q(t)\} = E\{g^2(X)\} + 2tE\{g(X)h(Y)\} + t^2E\{h^2(Y)\} \geq 0$$

for all t. But, $E\{Q(t)\} \geq 0$ for all t if and only if

$$(E\{g(X)h(Y)\})^2 \leq E\{g^2(X)\}E\{h^2(Y)\}.$$

This establishes (1.13.16).

2. **Jensen's Inequality**

A function $g : \mathbb{R} \to \mathbb{R}$ is called **convex** if, for any $-\infty < x < y < \infty$ and $0 \leq \alpha \leq 1$,

$$g(\alpha x + (1 - \alpha)y) \leq \alpha g(x) + (1 - \alpha)g(y).$$

Suppose X is a random variable and $E\{|X|\} < \infty$. Then, if g is convex,

$$g(E\{X\}) \leq E\{g(X)\}. \quad (1.13.17)$$

To prove (1.13.17), notice that since g is convex, for every x_0, $-\infty < x_0 < \infty$, $g(x) \geq g(x_0) + (x - x_0)g^*(x_0)$ for all x, $-\infty < x < \infty$, where $g^*(x_0)$ is finite. Substitute $x_0 = E\{X\}$. Then

$$g(X) \geq g(E\{X\}) + (X - E\{X\})g^*(E\{X\})$$

with probability one. Since $E\{X - E\{X\}\} = 0$, we obtain (1.13.17).

3. **Lyapunov's Inequality**
If $0 < s < r$ and $E\{|X|^r\} < \infty$, then

$$(E\{|X|^s\})^{1/s} \leq (E\{|X|^r\})^{1/r}. \tag{1.13.18}$$

To establish this inequality, let $t = r/s$. Notice that $g(x) = |x|^t$ is convex, since $t > 1$. Let $\xi = E\{|X|^s\}$, and $(|X|^s)^t = |X|^r$. Thus, by Jensen's inequality,

$$\begin{aligned} g(\xi) = (E|X|^s)^{r/s} &\leq E\{g(|X|^s)\} \\ &= E\{|X|^r\}. \end{aligned}$$

Hence, $E\{|X|^s\}^{1/s} \leq (E\{|X|^r\})^{1/r}$. As a result of Lyapunov's inequality we have the following chain of inequalities among absolute moments.

$$E\{|X|\} \leq (E\{X^2\})^{1/2} \leq (E\{|X|^3\})^{1/3} \leq \cdots. \tag{1.13.19}$$

4. **Hölder's Inequality**
Let $1 < p < \infty$ and $1 < q < \infty$, such that $\dfrac{1}{p} + \dfrac{1}{q} = 1$. $E\{|X|^p\} < \infty$ and $E\{|Y|^q\} < \infty$. Then

$$E\{|XY|\} \leq (E\{|X|^p\})^{1/p}(E\{|Y|^q\})^{1/q}. \tag{1.13.20}$$

Notice that the Schwarz inequality is a special case of Holder's inequality for $p = q = 2$.

For proof, see Shiryayev (1984, p. 191).

5. **Minkowsky's Inequality**
If $E\{|X|^p\} < \infty$ and $E\{|Y|^p\} < \infty$ for some $1 \leq p < \infty$, then $E\{|X + Y|^p\} < \infty$ and

$$(E\{|X + Y|^p\})^{1/p} \leq (E|X|^p)^{1/p} + (E\{|Y|^p\})^{1/p}. \tag{1.13.21}$$

For proof, see Shiryayev (1984, p. 192).

1.13.4 The Delta Method

The delta method is designed to yield large sample approximations to nonlinear functions g of the sample mean $\bar{X}_n$ and its variance. More specifically, let $\{X_n\}$ be

a sequence of i.i.d. random variables. Assume that $0 < V\{X\} < \infty$. By the SLLN, $\bar{X}_n \xrightarrow{\text{a.s.}} \mu$, as $n \to \infty$, where $\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$, and by the CLT, $\sqrt{n}\,\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{d} N(0,1)$. Let $g : \mathbb{R} \to \mathbb{R}$ having third order continuous derivative. By the Taylor expansion of $g(\bar{X}_n)$ around μ,

$$g(\bar{X}_n) = g(\mu) + (\bar{X}_n - \mu)g^{(1)}(\mu) + \frac{1}{2}(\bar{X}_n - \mu)^2 g^{(2)}(\mu) + R_n, \quad (1.13.22)$$

where $R_n = \frac{1}{6}(\bar{X}_n - \mu)^3 g^{(3)}(\mu_n^*)$, where μ_n^* is a point between $\bar{X}_n$ and μ, i.e., $|\bar{X}_n - \mu_n^*| < |\bar{X}_n - \mu|$. Since we assumed that $g^{(3)}(x)$ is continuous, it is bounded on the closed interval $[\mu - \Delta, \mu + \Delta]$. Moreover, $g^{(3)}(\mu_n^*) \xrightarrow{\text{a.s.}} g^{(3)}(\mu)$, as $n \to \infty$. Thus $R_n \xrightarrow{p} 0$, as $n \to \infty$. The distribution of $g(\mu) + g^{(1)}(\mu)(\bar{X}_n - \mu)$ is asymptotically $N(g(\mu), (g^{(1)}(\mu))^2\sigma^2/n)$. $(\bar{X}_n - \mu)^2 \xrightarrow{2} 0$, as $n \to \infty$. Thus, $\sqrt{n}(g(\bar{X}_n) - g(\mu)) \xrightarrow{d} N(0, \sigma^2(g^{(1)}(\mu))^2)$. Thus, if $\bar{X}_n$ satisfies the CLT, an approximation to the expected value of $g(\bar{X}_n)$ is

$$E\{g(\bar{X}_n)\} \cong g(\mu) + \frac{\sigma^2}{2n} g^{(2)}(\mu). \quad (1.13.23)$$

An approximation to the variance of $g(\bar{X}_n)$ is

$$V\{g(\bar{X}_n)\} \cong \frac{\sigma^2}{n}(g^{(1)}(\mu))^2. \quad (1.13.24)$$

Furthermore, from (1.13.22)

$$\sqrt{n}(g(\bar{X}_n) - g(\mu)) = \sqrt{n}(\bar{X}_n - \mu)g^{(1)}(\mu) + D_n, \quad (1.13.25)$$

where

$$D_n = \frac{(\bar{X}_n - \mu)^2}{2} g^{(2)}(\mu_n^{**}), \quad (1.13.26)$$

and $|\mu_n^{**} - \bar{X}_n| \le |\mu - \bar{X}_n|$ with probability one. Thus, since $\bar{X}_n - \mu \to 0$ a.s., as $n \to \infty$, and since $|g^{(2)}(\mu_n^{**})|$ is bounded, $D_n \xrightarrow{p} 0$, as $n \to \infty$, then

$$\sqrt{n}\,\frac{g(\bar{X}_n) - g(\mu)}{\sigma |g^{(1)}(\mu)|} \xrightarrow{d} N(0,1). \quad (1.13.27)$$

1.13.5 The Symbols o_p and O_p

Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables, $Y_n > 0$ a.s. for all $n \geq 1$. We say that $X_n = o_p(Y_n)$, i.e., X_n is of a smaller order of magnitude than Y_n in probability if

$$\frac{X_n}{Y_n} \xrightarrow{p} 0 \quad \text{as } n \to \infty. \tag{1.13.28}$$

We say that $X_n = O_p(Y_n)$, i.e., X_n has the same order of magnitude in probability as Y_n if, for all $\epsilon > 0$, there exists K_ϵ such that $\sup_n P\left\{\left|\frac{X_n}{Y_n}\right| > K_\epsilon\right\} < \epsilon$.

One can verify the following relations.

$$\begin{aligned}
&\textbf{(i)} \quad o_p(1) + O_p(1) = O_p(1),\\
&\textbf{(ii)} \quad O_p(1) + O_p(1) = O_p(1),\\
&\textbf{(iii)} \quad o_p(1) + o_p(1) = o_p(1),\\
&\textbf{(iv)} \quad O_p(1) \cdot O_p(1) = O_p(1),\\
&\textbf{(v)} \quad o_p(1) \cdot O_p(1) = o_p(1).
\end{aligned} \tag{1.13.29}$$

1.13.6 The Empirical Distribution and Sample Quantiles

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having a distribution F. The function

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \leq x\} \tag{1.13.30}$$

is called the **empirical distribution function** (EDF).

Notice that $E\{I\{X_i \leq x\}\} = F(x)$. Thus, the SLLN implies that at each x, $F_n(x) \xrightarrow{\text{a.s.}} F(x)$ as $n \to \infty$. The question is whether this convergence is uniform in x. The answer is given by

Theorem 1.13.5 (Glivenko–Cantelli). *Let $X_1, X_2, X_3, \ldots$ be i.i.d. random variables. Then*

$$\sup_{-\infty<x<\infty} |F_n(x) - F(x)| \xrightarrow{\text{a.s.}} 0, \quad \text{as } n \to \infty. \tag{1.13.31}$$

For proof, see Sen and Singer (1993, p. 185).

The pth sample quantile $x_{n,p}$ is defined as

$$\begin{aligned} x_{n,p} &= F_n^{-1}(p) \\ &= \inf\{x : F_n(x) \geq p\} \end{aligned} \tag{1.13.32}$$

for $0 < p < 1$, where $F_n(x)$ is the EDF. When $F(x)$ is continuous then, the points of increase of $F_n(x)$ are the order statistics $X_{(1:n)} < \cdots < X_{(n:n)}$ with probability one. Also, $F_n(X_{(i:n)}) = \dfrac{i}{n}$, $i = 1, \ldots, n$. Thus,

$$\begin{aligned} &x_{n,p} = X_{(i(p):n)}, \quad \text{where} \\ &i(p) = \text{smallest integer } i \text{ such that } i \geq pn. \end{aligned} \tag{1.13.33}$$

Theorem 1.13.6. *Let F be a continuous distribution function, and $\xi_p = F^{-1}(p)$, and suppose that $F(\xi_p) = p$ and for any $\epsilon > 0$, $F(\xi_p - \epsilon) < p < F(\xi_p + \epsilon)$. Let $X_1, \ldots, X_n$ be i.i.d. random variables from this distribution. Then*

$$x_{n,p} \xrightarrow{a.s.} \xi_p \text{ as } n \to \infty.$$

For proof, see Sen and Singer (1993, p. 167).

The following theorem establishes the asymptotic normality of $x_{n,p}$.

Theorem 1.13.7. *Let $F(x)$ be an absolutely continuous distribution, with continuous p.d.f. $f(x)$. Let p, $0 < p < 1$, $\xi_p = F^{-1}(p)$ and $f(\xi_p) > 0$. Then*

$$\sqrt{n}(x_{n,p} - \xi_p) \xrightarrow{d} N\left(0, \frac{p(1-p)}{f^2(\xi_p)}\right). \tag{1.13.34}$$

For proof, see Sen and Singer (1993, p. 168).

The results of Theorems 1.13.6–1.13.7 will be used in Chapter 7 to establish the asymptotic relative efficiency of the sample median, relative to the sample mean.

PART II: EXAMPLES

Example 1.1. We illustrate here two algebras.

The sample space is finite

$$\mathcal{S} = \{1, 2, \ldots, 10\}.$$

Let $E_1 = \{1, 2\}$, $E_2 = \{9, 10\}$. The algebra generated by E_1 and E_2, $\mathcal{A}_1$, contains the events

$$\mathcal{A}_1 = \{\mathcal{S}, \emptyset, E_1, \bar{E}_1, E_2, \bar{E}_2, E_1 \cup E_2, \overline{E_1 \cup E_2}\}.$$

The algebra generated by the partition $\mathcal{D} = \{E_1, E_2, E_3, E_4\}$, where $E_1 = \{1, 2\}$, $E_2 = \{9, 10\}$, $E_3 = \{3, 4, 5\}$, $E_4 = \{6, 7, 8\}$ contains the $2^4 = 16$ events

$$\mathcal{A}_2 = \{\mathcal{S}, \emptyset, E_1, E_2, E_3, E_4, E_1 \cup E_2, E_1 \cup E_3, E_1 \cup E_4, E_2 \cup E_3, E_2 \cup E_4,$$
$$E_3 \cup E_4, E_1 \cup E_2 \cup E_3, E_1 \cup E_2 \cup E_4, E_1 \cup E_3 \cup E_4, E_2 \cup E_3 \cup E_4\}.$$

Notice that the complement of each set in $\mathcal{A}_2$ is in $\mathcal{A}_2$. $\mathcal{A}_1 \subset \mathcal{A}_2$. Also, $\mathcal{A}_2 \subset \mathcal{A}(\mathcal{S})$. ■

Example 1.2. In this example we consider a **random walk** on the integers. Consider an experiment in which a particle is initially at the origin, 0. In the first trial the particle moves to $+1$ or to -1. In the second trial it moves either one integer to the right or one integer to the left. The experiment consists of $2n$ such trials ($1 \leq n < \infty$). The sample space $\mathcal{S}$ is finite and there are 2^{2n} points in $\mathcal{S}$, i.e., $\mathcal{S} = \{(i_1, \ldots, i_{2n}) : i_j = \pm 1, j = 1, \ldots, 2n\}$. Let $E_j = \left\{(i_1, \ldots, i_{2n}) : \sum_{k=1}^{2n} i_k = j\right\}$, $j = 0, \pm 2, \pm, \cdots, \pm 2n$. E_j is the event that, at the end of the experiment, the particle is at the integer j. Obviously, $-2n \leq j \leq 2n$. It is simple to show that j must be an even integer $j = \pm 2k$, $k = 0, 1, \ldots, n$. Thus, $\mathcal{D} = \{E_{2k}, k = 0, \pm 1, \ldots, \pm n\}$ is a partition of $\mathcal{S}$. The event E_{2k} consists of all elementary events in which there are $(n + k)$ $+1$s and $(n - k)$ -1s. Thus, E_{2k} is the union of $\binom{2n}{n+k}$ points of $\mathcal{S}$, $k = 0, \pm 1, \ldots, \pm n$.

The algebra generated by $\mathcal{D}$, $\mathcal{A}(\mathcal{D})$, consists of $\emptyset$ and $2^{2n+1} - 1$ unions of the elements of $\mathcal{D}$. ■

Example 1.3. Let $\mathcal{S}$ be the real line, i.e., $\mathcal{S} = \{x : -\infty < x < \infty\}$. We construct an algebra $\mathcal{A}$ generated by half-closed intervals: $E_x = (-\infty, x]$, $-\infty < x < \infty$. Notice that, for $x < y$, $E_x \cup E_y = (-\infty, y]$. The complement of E_x is $\bar{E}_x = (x, \infty)$. We will adopt the convention that $(x, \infty) \equiv (x, \infty]$.

Consider the sequence of intervals $E_n = \left(-\infty, 1 - \frac{1}{n}\right], n \geq 1$. All $E_n \in \mathcal{A}$. However, $\bigcup_{n=1}^{\infty} E_n = (-\infty, 1)$. Thus $\lim_{n \to \infty} E_n$ **does not** belong to $\mathcal{A}$. $\mathcal{A}$ is **not** a σ-field. In order to make $\mathcal{A}$ into a σ-field we have to add to it all limit sets of sequences of events in $\mathcal{A}$. ■

Example 1.4. We illustrate here three events that are only pairwise independent.

Let $\mathcal{S} = \{1, 2, 3, 4\}$, with $P(w) = \frac{1}{4}$, for all $w \in \mathcal{S}$. Define the three events

$$A_1 = \{1, 2\}, \quad A_2 = \{1, 3\}, \quad A_3 = \{1, 4\}.$$

$P\{A_i\} = \frac{1}{2}, i = 1, 2, 3.$

$$A_1 \cap A_2 = \{1\}.$$
$$A_1 \cap A_3 = \{1\}.$$
$$A_2 \cap A_3 = \{1\}.$$

Thus

$$P\{A_1 \cap A_2\} = \frac{1}{4} = P\{A_1\}P\{A_2\}.$$
$$P\{A_1 \cap A_3\} = \frac{1}{4} = P\{A_1\}P\{A_3\}.$$
$$P\{A_2 \cap A_3\} = \frac{1}{4} = P\{A_2\}P\{A_3\}.$$

Thus, A_1, A_2, A_3 are pairwise independent. On the other hand,

$$A_1 \cap A_2 \cap A_3 = \{1\}$$

and

$$P\{A_1 \cap A_2 \cap A_3\} = \frac{1}{4} \neq P\{A_1\}P\{A_2\}P\{A_3\} = \frac{1}{8}.$$

Thus, the triplet (A_1, A_2, A_3) is not independent. ■

Example 1.5. An infinite sequence of trials, in which each trial results in either "success" S or "failure" F is called **Bernoulli trials** if all trials are independent and the probability of success in each trial is the same. More specifically, consider the sample space of countable sequences of Ss and Fs, i.e.,

$$\mathcal{S} = \{(i_1, i_2, \ldots) : i_j = S, F,\ \ j = 1, 2, \ldots\}.$$

Let

$$E_j = \{(i_1, i_2, \ldots) : i_j = S\},\ \ j = 1, 2, \ldots.$$

We assume that $\{E_1, E_2, \ldots, E_n\}$ are mutually independent for all $n \geq 2$ and $P\{E_j\} = p$ for all $j = 1, 2, \ldots, 0 < p < 1$.

The points of $\mathcal{S}$ represent an infinite sequence of **Bernoulli trials**. Consider the events

$$\begin{aligned} A_j &= \{(i_1, i_2, \ldots) : i_j = S, i_{j+1} = F, i_{j+2} = S\} \\ &= E_j \cap \bar{E}_{j+1} \cap E_{j+2} \end{aligned}$$

$j = 1, 2, \ldots$. $\{A_j\}$ are not independent.

Let $B_j = \{A_{3j+1}\}$, $j \geq 0$. The sequence $\{B_j,\ j \geq 1\}$ consists of mutually independent events. Moreover, $P(B_j) = p^2(1-p)$ for all $j = 1, 2, \ldots$. Thus, $\sum_{j=1}^{\infty} P(B_j) = \infty$ and the Borel–Cantelli Lemma implies that $P\{B_n, \text{i.o.}\} = 1$. That is, the pattern SFS will occur infinitely many times in a sequence of Bernoulli trials, with probability one. ■

Example 1.6. Let $\mathcal{S}$ be the sample space of $N = 2^n$ binary sequences of size n, $n < \infty$, i.e.,

$$\mathcal{S} = \{(i_1, \ldots, i_n) : i_j = 0, 1,\ \ j = 1, \ldots, n\}.$$

We assign the points $w = (i_1, \ldots, i_n)$ of $\mathcal{S}$, equal probabilities, i.e., $P\{(i_1, \ldots, i_n)\} = 2^{-n}$. Consider the partition $\mathcal{D} = \{B_0, B_1, \ldots, B_n\}$ to $k = n + 1$ disjoint events, such that

$$B_j = \{(i_1, \ldots, i_n) : \sum_{l=1}^{n} i_l = j\}, \quad j = 0, \ldots, n.$$

B_j is the set of all points having exactly j ones and $(n - j)$ zeros. We define the discrete random variable corresponding to $\mathcal{D}$ as

$$X(w) = \sum_{j=0}^{n} j I_{B_j}(w).$$

The jump points of $X(w)$ are $\{0, 1, \ldots, n\}$. The probability distribution function of $X(w)$ is

$$f_X(x) = \sum_{j=0}^{n} I_{\{j\}}(x) P\{B_j\}.$$

It is easy to verify that

$$P\{B_j\} = \binom{n}{j} 2^{-n}, \quad j = 0, 1, \ldots, n$$

where

$$\binom{n}{j} = \frac{n!}{j!(n-j)!}, \quad j = 0, 1, \ldots, n.$$

Thus,

$$f_X(x) = \sum_{j=0}^{n} I_{\{j\}}(x)\binom{n}{x}2^{-n}.$$

The distribution function (c.d.f.) is given by

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0 \\ \sum_{j=0}^{[x]}\binom{n}{j}2^{-n}, & \end{cases}$$

where $[x]$ is the maximal integer value smaller or equal to x. The distribution function illustrated here is called a **binomial distribution** (see Section 2.2.1). ■

Example 1.7. Consider the random variable of Example 1.6. In that example $X(w) \in \{0, 1, \ldots, n\}$ and $f_X(j) = \binom{n}{j}2^{-n}$, $j = 0, \ldots, n$. Accordingly,

$$E\{X\} = \sum_{j=0}^{n} j\binom{n}{j}2^{-n} = \frac{n}{2}\sum_{j=0}^{n-1}\binom{n-1}{j}2^{-(n-1)} = \frac{n}{2}.$$

■

Example 1.8. Let $(\mathcal{S}, \mathcal{F}, P)$ be a probability space where $\mathcal{S} = \{0, 1, 2, \ldots\}$. $\mathcal{F}$ is the σ-field of all subsets of $\mathcal{S}$. Consider $X(w) = w$, with probability function

$$\begin{aligned} p_j &= P\{w : X(w) = j\} \\ &= e^{-\lambda}\frac{\lambda^j}{j!}, \quad j = 0, 1, 2, \ldots \end{aligned}$$

for some λ, $0 < \lambda < \infty$. $0 < p_j < \infty$ for all j, and since $\sum_{j=0}^{\infty}\frac{\lambda^j}{j} = e^{\lambda}$, $\sum_{j=0}^{\infty} p_j = 1$.

Consider the partition $\mathcal{D} = \{A_1, A_2, A_3\}$ where $A_1 = \{w : 0 \le w \le 10\}$, $A_2 = \{w : 10 < w \le 20\}$ and $A_3 = \{w : w \ge 21\}$. The probabilities of these sets are

$$q_1 = P\{A_1\} = e^{-\lambda} \sum_{j=0}^{10} \frac{\lambda^j}{j!},$$

$$q_2 = P\{A_2\} = e^{-\lambda} \sum_{j=11}^{20} \frac{\lambda^j}{j!}, \quad \text{and}$$

$$q_3 = P\{A_3\} = e^{-\lambda} \sum_{j=21}^{\infty} \frac{\lambda^j}{j!}.$$

The conditional distributions of X given A_i $i = 1, 2, 3$ are

$$f_{X|A_i}(x) = \frac{\frac{\lambda^x}{x!} I_{A_i}(x)}{\sum_{j=b_{i-1}}^{b_i - 1} \frac{\lambda^j}{j!}}, \quad i = 1, 2, 3$$

where $b_0 = 0, b_1 = 11, b_2 = 21, b_3 = \infty$.

The conditional expectations are

$$E\{X \mid A_i\} = \lambda \frac{\sum_{j=(b_{i-1}-1)^+}^{b_i - 2} \frac{\lambda^j}{j!}}{\sum_{j=b_{i-1}}^{b_i - 1} \frac{\lambda^j}{j!}}, \quad i = 1, 2, 3$$

where $a+ = \max(a, 0)$. $E\{X \mid \mathcal{D}\}$ is a random variable, which obtains the values $E\{X \mid A_1\}$ with probability q_1, $E\{X \mid A_2\}$ with probability q_2, and $E\{X \mid A_3\}$ with probability q_3. ■

Example 1.9. Consider two discrete random variables X, Y on $(\mathcal{S}, \mathcal{F}, P)$ such that the jump points of X and Y are the nonnegative integers $\{0, 1, 2, \ldots\}$. The joint probability function of (X, Y) is

$$f_{XY}(x, y) = \begin{cases} e^{-\lambda} \dfrac{\lambda^y}{(y+1)!}, & x = 0, 1, \ldots, y;\ y = 0, 1, 2, \ldots \\ 0, & \text{otherwise,} \end{cases}$$

where λ, $0 < \lambda < \infty$, is a specified parameter.

First, we have to check that

$$\sum_{x=0}^{\infty}\sum_{y=0}^{\infty} f_{XY}(x, y) = 1.$$

Indeed,

$$\begin{aligned} f_Y(y) &= \sum_{x=0}^{y} f_{XY}(x, y) \\ &= e^{-\lambda}\frac{\lambda^y}{y!}, \quad y = 0, 1, \ldots \end{aligned}$$

and

$$\sum_{y=0}^{\infty} e^{-\lambda}\frac{\lambda^y}{y!} = e^{-\lambda}\cdot e^{\lambda} = 1.$$

The conditional p.d.f. of X given $\{Y = y\}$, $y = 0, 1, \ldots$ is

$$f_{X|Y}(x \mid y) = \begin{cases} \dfrac{1}{1+y}, & x = 0, 1, \ldots, y \\ 0, & \text{otherwise.} \end{cases}$$

Hence,

$$\begin{aligned} E\{X \mid Y = y\} &= \frac{1}{1+y}\sum_{x=0}^{y} x \\ &= \frac{y}{2}, \quad y = 0, 1, \ldots \end{aligned}$$

and, as a random variable,

$$E\{X \mid Y\} = \frac{Y}{2}.$$

Finally,

$$E\{E\{X \mid Y\}\} = \sum_{y=0}^{\infty}\frac{y}{2}e^{-\lambda}\frac{\lambda^y}{y!} = \frac{\lambda}{2}.$$

■

Example 1.10. In this example we show an absolutely continuous distribution for which $E\{X\}$ **does not** exist.

Let $F(x) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(x)$. This is called the **Cauchy distribution**. The density function (p.d.f.) is

$$f(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}, \quad -\infty < x < \infty.$$

It is a symmetric density around $x = 0$, in the sense that $f(x) = f(-x)$ for all x. The expected value of X having this distribution does not exist. Indeed,

$$\begin{aligned}\int_{-\infty}^{\infty} |x| f(x)dx &= \frac{2}{\pi}\int_0^{\infty} \frac{x}{1+x^2}dx \\ &= \frac{1}{\pi}\lim_{T\to\infty} \log(1+T^2) = \infty.\end{aligned}$$

■

Example 1.11. We show here a mixture of discrete and absolutely continuous distributions.

Let

$$F_{\text{ac}}(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1 - \exp\{-\lambda x\}, & \text{if } x \geq 0 \end{cases}$$

$$F_{\text{d}}(x) = \begin{cases} 0, & \text{if } x < 0 \\ e^{-\mu}\sum_{j=0}^{[x]} \frac{\mu^j}{j!}, & \text{if } x \geq 0 \end{cases}$$

where $[x]$ designates the maximal integer not exceeding x; λ and μ are real positive numbers. The mixed distribution is, for $0 \leq \alpha \leq 1$,

$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ \alpha e^{-\mu}\sum_{j=0}^{[x]} \frac{\mu^j}{j!} + (1-\alpha)[1 - \exp(-\lambda x)], & \text{if } x \geq 0. \end{cases}$$

This distribution function can be applied with appropriate values of α, λ, and μ for modeling the length of telephone conversations. It has discontinuities at the nonnegative integers and is continuous elsewhere. ■

Example 1.12. Densities derived after transformations.

Let X be a random variable having an absolutely continuous distribution with p.d.f. f_X.

A. If $Y = X^2$, the number of roots are

$$m(y) = \begin{cases} 0, & \text{if } y < 0 \\ 1, & \text{if } y = 0 \\ 2, & \text{if } y > 0. \end{cases}$$

Thus, the density of Y is

$$f_Y(y) = \begin{cases} 0, & y \le 0 \\ \dfrac{1}{2\sqrt{y}}[f_X(\sqrt{y}) + f_X(-\sqrt{y})], & y > 0. \end{cases}$$

B. If $Y = \cos X$

$$m(y) = \begin{cases} 0, & \text{if } |y| > 1 \\ \infty, & \text{if } |y| \le 1. \end{cases}$$

For every y, such that $|y| < 1$, let $\xi(y)$ be the value of $\cos^{-1}(y)$ in the interval $(0, \pi)$. Then, if $f(x)$ is the p.d.f. of X, the p.d.f. of $Y = \cos X$ is, for $|y| < 1$,

$$f_Y(y) = \frac{1}{\sqrt{1-y^2}} \sum_{j=0}^{\infty} \{f_X(\xi(y) + 2\pi j) + \\ + f_X(\xi(y) - 2\pi j) + f_X(-\xi(y) + 2\pi j) + f_X(-\xi(y) - 2\pi j)\}.$$

The density does not exist for $|y| \ge 1$. ■

Example 1.13. Three cases of joint p.d.f.

A. Both X_1, X_2 are discrete, with jump points on $\{0, 1, 2, \ldots\}$. Their joint p.d.f. for $0 < \lambda < \infty$ is,

$$f_{X_1X_2}(x_1, x_2) = \binom{x_2}{x_1} 2^{-x_2} e^{-\lambda} \frac{\lambda^{x_2}}{x_2!},$$

for $x_1 = 0, \ldots, x_2$, $x_2 = 0, 1, \ldots$. The marginal p.d.f. are

$$f_{X_1}(x_1) = e^{-\lambda/2} \frac{(\lambda/2)^{x_1}}{x_1!}, \quad x_1 = 0, 1, \ldots \text{and}$$

$$f_{X_2}(x_2) = e^{-\lambda} \frac{\lambda^{x_2}}{x_2!}, \quad x_2 = 0, 1, \ldots.$$

B. Both X_1 and X_2 are absolutely continuous, with joint p.d.f.

$$f_{X_1X_2}(x, y) = 2I_{(0,1)}(x) I_{(0,x)}(y).$$

The marginal distributions of X_1 and X_2 are

$$f_{X_1}(x) = 2x I_{(0,1)}(x) \text{ and}$$
$$f_{X_2}(y) = 2(1-y) I_{(0,1)}(y).$$

C. X_1 is discrete with jump points $\{0, 1, 2, \ldots\}$ and X_2 absolutely continuous. The joint p.d.f., with respect to the σ-finite measure $dN(x_1)dy$ is, for $0 < \lambda < \infty$,

$$f_{X_1X_2}(x, y) = e^{-\lambda}\frac{\lambda^x}{x!} \cdot \frac{1}{1+x} I\{x = 0, 1, \ldots\} I_{(0,1+x)}(y).$$

The marginal p.d.f. of X_1, is

$$f_X(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots.$$

The marginal p.d.f. of X_2 is

$$f_{X_2}(y) = \frac{1}{\lambda}\sum_{n=0}^{\infty}\left(1 - e^{-\lambda}\sum_{j=0}^{n}\frac{\lambda^j}{j!}\right) I_{(n,n+1)}(y).$$

■

Example 1.14. Suppose that X, Y are positive random variables, having a joint p.d.f.

$$f_{XY}(x, y) = \frac{1}{y}\lambda e^{-\lambda y} I_{(0,y)}(x), \quad 0 < y < \infty, \ \ 0 < x < y, \ \ 0 < \lambda < \infty.$$

The marginal p.d.f. of X is

$$f_X(x) = \lambda \int_x^{\infty} \frac{1}{y} e^{-\lambda y} dy$$
$$= \lambda E_1(\lambda x),$$

where $E_1(\xi) = \displaystyle\int_{\xi}^{\infty} \frac{1}{u} e^{-u} du$ is called the exponential integral, which is finite for all $\xi > 0$. Thus, according to (1.6.62), for $x_0 > 0$,

$$f_{Y|X}(y \mid x_0) = \frac{\dfrac{1}{y} e^{-\lambda y} I_{(x_0,\infty)}(y)}{E_1(\lambda x_0)}.$$

Finally, for $x_0 > 0$,

$$E\{Y \mid X = x_0\} = \frac{\int_{x_0}^{\infty} e^{-\lambda y} dy}{E_1(\lambda x_0)}$$
$$= \frac{e^{-\lambda x_0}}{\lambda E_1(\lambda x_0)}.$$

■

Example 1.15. In this example we show a distribution function whose m.g.f., M, exists only on an interval $(-\infty, t_0)$. Let

$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1 - e^{-\lambda x}, & \text{if } x \geq 0, \end{cases}$$

where $0 < \lambda < \infty$. The m.g.f. is

$$M(t) = \lambda \int_0^{\infty} e^{tx - \lambda x} dx$$
$$= \frac{\lambda}{\lambda - t} = \left(1 - \frac{t}{\lambda}\right)^{-1}, \quad -\infty < t < \lambda.$$

The integral in $M(t)$ is ∞ if $t \geq \lambda$. Thus, the domain of convergence of M is $(-\infty, \lambda)$. ■

Example 1.16. Let

$$X_i = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } (1 - p) \end{cases}$$

$i = 1, \ldots, n$. We assume also that $X_1, \ldots, X_n$ are independent. We wish to derive the p.d.f. of $S_n = \sum_{i=1}^{n} X_i$. The p.g.f. of S_n is, due to independence, when $q = 1 - p$,

$$E\{t^{S_n}\} = E\left\{t^{\sum_{i=1}^{n} X_i}\right\}$$
$$= \prod_{i=1}^{n} E\{t^{X_i}\}$$
$$= (pt + q)^n, \quad -\infty < t < \infty.$$

Since all X_i have the same distribution. Binomial expansion yields

$$E\{t^{S_n}\} = \sum_{j=0}^{n} \binom{n}{j} p^j (1-p)^{n-j} t^j.$$

Since two polynomials of degree n are equal for **all** t only if their coefficients are equal, we obtain

$$P\{S_n = j\} = \binom{n}{j} p^j (1-p)^{n-j}, \quad j = 0, \ldots, n.$$

The distribution of S_n is called the **binomial distribution**. ■

Example 1.17. In Example 1.13 Part C, the conditional p.d.f. of X_2 given $\{X_1 = x\}$ is

$$f_{X_2|X_1}(y \mid x) = \frac{1}{1+x} I_{(0,1+x)}(y).$$

This is called the **uniform distribution** on $(0, 1+x)$. It is easy to find that

$$E\{Y \mid X = x\} = \frac{1+x}{2}$$

and

$$V\{Y \mid X = x\} = \frac{(1+x)^2}{12}.$$

Since the p.d.f. of X is

$$P\{X = x\} = e^{-\lambda} \frac{\lambda x}{x!}, \quad x = 0, 1, 2, \ldots$$

the law of iterated expectation yields

$$\begin{aligned} E\{Y\} &= E\{E\{Y \mid X\}\} \\ &= E\left\{\frac{1}{2} + \frac{1}{2}X\right\} \\ &= \frac{1}{2} + \frac{\lambda}{2}, \end{aligned}$$

since $E\{X\} = \lambda$.

The law of total variance yields

$$\begin{aligned} V\{Y\} &= V\{E\{Y \mid X\}\} + E\{V\{Y \mid X\}\} \\ &= V\left\{\frac{1}{2} + \frac{1}{2}X\right\} + E\left\{\frac{(1+X)^2}{12}\right\} \\ &= \frac{1}{4}V\{X\} + \frac{1}{12}E\{1 + 2X + X^2\} \\ &= \frac{1}{4}\lambda + \frac{1}{12}(1 + 2\lambda + \lambda(1+\lambda)) \\ &= \frac{1}{12}(1+\lambda)^2 + \frac{\lambda}{3}. \end{aligned}$$

To verify these results, prove that $E\{X\} = \lambda$, $V\{X\} = \lambda$ and $E\{X^2\} = \lambda(1+\lambda)$. We also used the result that $V\{a + bX\} = b^2 V\{X\}$. ■

Example 1.18. Let X_1, X_2, X_3 be uncorrelated random variables, having the same variance σ^2, i.e.,

$$\mathfrak{V} = \sigma^2 I.$$

Consider the linear transformations

$$\begin{aligned} Y_1 &= X_1 + X_2, \\ Y_2 &= X_1 + X_3, \end{aligned}$$

and

$$Y_3 = X_2 + X_3.$$

In matrix notation

$$\mathbf{Y} = A\mathbf{X},$$

where

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

The variance–covariance matrix of $\mathbf{Y}$, according to (1.8.30) is

$$\begin{aligned} V[\mathbf{Y}] &= A\Sigma A' \\ &= \sigma^2 AA' \\ &= \sigma^2 \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}. \end{aligned}$$

From this we obtain that correlations of Y_i, Y_j for $i \neq j$ and $\rho_{ij} = \frac{1}{2}$. ■

Example 1.19. We illustrate here convergence in distribution.

A. Let $X_1, X_2, \ldots$ be random variables with distribution functions

$$F_n(x) = \begin{cases} 0, & \text{if } x < 0 \\ \frac{1}{n} + \left(1 - \frac{1}{n}\right)\left(1 - e^{-x}\right), & \text{if } x \geq 0. \end{cases}$$

$X_n \xrightarrow{d} X$, where the distribution of X is

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-x}, & x \geq 0. \end{cases}$$

B. X_n are random variables with

$$F_n(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-nx}, & x \geq 0 \end{cases}$$

and $F(x) = I\{x \geq 0\}$. $X_n \xrightarrow{d} X$. Notice that $F(x)$ is discontinuous at $x = 0$. But, for all $x \neq 0$ $\lim_{n\to\infty} F_n(x) = F(x)$.

C. $\mathbf{X}_n$ are random vectors, i.e.,

$$\mathbf{X}_n = (X_{1n}, X_{2n}), \quad n \geq 1.$$

The function $I_x(a, b)$, for $0 < a, b < \infty, 0 \leq x \leq 1$, is called the incomplete beta function ratio and is given by

$$I_x(a, b) = \frac{\int_0^x u^{a-1}(1-u)^{b-1}du}{B(a, b)},$$

where $B(a,b) = \int_0^1 u^{a-1}(1-u)^{b-1}du$. In terms of these functions, the marginal distribution of X_{1n} and X_{2n} are

$$F_{1n}(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{n} + \left(1 - \frac{1}{n}\right) I_x(a,b), & 0 \le x \le 1 \\ 1, & 1 < x \end{cases}$$

and

$$F_{2n}(y) = \begin{cases} 0, & y < 0 \\ \left(1 - \frac{1}{n}\right) I_y(a,b), & 0 \le y < 1 \\ 1, & 1 \le y \end{cases}$$

where $0 < a, b < \infty$. The joint distribution of (X_{1n}, X_{2n}) is $F_n(x, y) = F_{1n}(x)F_{2n}(y)$, $n \ge 1$. The random vectors $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$, where $F(\mathbf{x})$ is

$$F(x,y) = \begin{cases} 0, & x < 0 \text{ or } y < 0 \\ I_x(a,b)I_y(a,b), & 0 \le x, y \le 1 \\ I_x(a,b), & 0 \le x \le 1, y > 1 \\ I_y(a,b), & 1 < x, 0 \le y \le 1 \\ 1, & 1 < x, 1 < y. \end{cases}$$

■

Example 1.20. Convergence in probability.

Let $\mathbf{X}_n = (X_{1n}, X_{2n})$, where $X_{i,n}$ $(i = 1, 2)$ are independent and have a distribution

$$F_n(x) = \begin{cases} 0, & x < 0 \\ nx, & 0 < x < \frac{1}{n} \\ 1, & \frac{1}{n} \le x. \end{cases}$$

Fix an $\epsilon > 0$ and let $N(\epsilon) = \left[\frac{2}{\epsilon}\right]$, then for every $n > N(\epsilon)$,

$$P[(X_{1,n}^2 + X_{2,n}^2)^{1/2} < \epsilon] = 1.$$

Thus, $\mathbf{X}_n \xrightarrow{p} \mathbf{0}$. ■

Example 1.21. Convergence in mean square.

Let $\{X_n\}$ be a sequence of random variables such that

$$E\{X_n\} = 1 + \frac{a}{n}, \quad 0 < a < \infty \text{ and}$$
$$V\{X_n\} = \frac{b}{n}, \qquad 0 < b < \infty.$$

Then, $X_n \xrightarrow{2} 1$, as $n \to \infty$. Indeed, $E\{(X_n - 1)^2\} = \frac{a^2}{n^2} + \frac{b}{n} \to 0$, as $n \to \infty$. ■

Example 1.22. Central Limit Theorem.

A. Let $\{X_n\}$, $n \geq 1$ be a sequence of i.i.d. random variables, $P\{X_n = 1\} = P\{X_n = -1\} = \frac{1}{2}$. Thus, $E\{X_n\} = 0$ and $V\{X_n\} = 1$, $n \geq 1$. Thus $\sqrt{n}\,\bar{X}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{d} N(0, 1)$. It is interesting to note that for these random variables, when $S_n = \sum_{i=1}^{n} X_i$, $\frac{1}{\sqrt{n}} S_n \xrightarrow{d} N(0, 1)$, while $\frac{S_n}{n} \xrightarrow{\text{a.s.}} 0$.

B. Let $\{X_n\}$ be i.i.d, having a rectangular p.d.f.

$$f(x) = 1_{(0,1)}(x).$$

In this case, $E\{X_1\} = \frac{1}{2}$ and $V\{X_1\} = \frac{1}{12}$. Thus,

$$\sqrt{n}\,\frac{\bar{X}_n - \frac{1}{2}}{\sqrt{\frac{1}{12}}} \xrightarrow[n\to\infty]{d} N(0, 1).$$

Notice that if $n = 12$, then if $S_{12} = \sum_{i=1}^{12} X_i$, then $S_{12} - 6$ might have a distribution close to that of $N(0, 1)$. Early simulation programs were based on this. ■

Example 1.23. Application of Lyapunov's Theorem.

Let $\{X_n\}$ be a sequence of independent random variables, with distribution functions

$$F_n(x) = \begin{cases} 0, & x < 0 \\ 1 - \exp\{-x/n\}, & x \geq 0 \end{cases}$$

$n \geq 1$. Thus, $E\{X_n\} = n$, $V\{X_n\} = n^2$, and $E\{X_n^3\} = 6n^3$. Thus, $B_n^2 = \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$, $n \geq 1$. In addition,

$$\sum_{k=1}^{n} E\{X_k^3\} = 6 \sum_{k=1}^{n} k^3 = O(n^4).$$

Thus,

$$\lim_{n \to \infty} \frac{\sum_{k=1}^{n} E\{X_k^3\}}{B_n^3} = 0.$$

It follows from Lyapunov's Theorem that

$$\sqrt{6}\, \frac{\sum_{k=1}^{n} (X_k - k)}{\sqrt{n(n+1)(2n+1)}} \xrightarrow{d} N(0, 1).$$

■

Example 1.24. Variance stabilizing transformation.

Let $\{X_n\}$ be i.i.d. binary random variables, such that $P\{X_n = 1\} = p$, and $P\{X_n = 0\} = 1 - p$. It is easy to verify that $\mu = E\{X_1\} = p$ and $V\{X_1\} = p(1-p)$. Hence, by the CLT, $\sqrt{n}\, \frac{\bar{X}_n - p}{\sqrt{p(1-p)}} \xrightarrow{d} N(0, 1)$, as $n \to \infty$. Consider the transformation

$$g(\bar{X}_n) = 2 \sin^{-1} \sqrt{\bar{X}_n}.$$

The derivative of $g(x)$ is

$$g^{(1)}(x) = \frac{2}{\sqrt{1-x}} \cdot \frac{1}{2\sqrt{x}} = \frac{1}{\sqrt{x(1-x)}}.$$

Hence $V\{X_1\}(g^{(1)}(p))^2 = 1$.

It follows that

$$\sqrt{n}(2 \sin^{-1}(\sqrt{\bar{X}_n}) - 2 \sin^{-1}(\sqrt{p})) \xrightarrow{d} N(0, 1).$$

$g^{(2)}(x) = -\frac{1}{2}\frac{1-2x}{(x(1-x))^{3/2}}$. Hence, by the delta method,

$$E\{g(\bar{X}_n)\} \cong 2\sin^{-1}(\sqrt{p}) - \frac{1-2p}{4n(p(1-p))^{1/2}}.$$

This approximation is very ineffective if p is close to zero or close to 1. If p is close to $\frac{1}{2}$, the second term on the right-hand side is close to zero. ■

Example 1.25. A. Let $X_1, X_2, \ldots$ be i.i.d. random variables having a finite variance $0 < \sigma^2 < \infty$. Since $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$, we say that $\bar{X}_n - \mu = O_p\left(\frac{1}{\sqrt{n}}\right)$ as $n \to \infty$. Thus, if $c_n \nearrow \infty$ but $c_n = o(\sqrt{n})$, then $c_n(\bar{X}_n - \mu) \xrightarrow{p} 0$. Hence $\bar{X}_n - \mu = o_p(c_n)$, as $n \to \infty$.

B. Let $X_1, X_2, \ldots, X_n$ be i.i.d. having a common exponential distribution with p.d.f.

$$f(x;\mu) = \begin{cases} 0, & \text{if } x < 0 \\ \mu e^{-\mu x}, & \text{if } x \geq 0 \end{cases}$$

$0 < \mu < \infty$. Let $Y_n = \min[X_i, i = 1, \ldots, n]$ be the first order statistic in a random sample of size n (see Section 2.10). The p.d.f. of Y_n is

$$f_n(y;\mu) = \begin{cases} 0, & \text{if } y < 0 \\ n\mu e^{-n\mu y}, & \text{if } y \geq 0. \end{cases}$$

Thus $nY_n \sim X_1$ for all n. Accordingly, $Y_n = O_p\left(\frac{1}{n}\right)$ as $n \to \infty$. It is easy to see that $\sqrt{n}\, Y_n \xrightarrow{p} 0$. Indeed, for any given $\epsilon > 0$,

$$P\{\sqrt{n}\, Y_n > \epsilon\} = e^{-\sqrt{n}\,\mu\epsilon} \to 0 \quad \text{as } n \to \infty.$$

Thus, $Y_n = o_p\left(\frac{1}{\sqrt{n}}\right)$ as $n \to \infty$. ■

PART III: PROBLEMS

Section 1.1

1.1.1 Show that $A \cup B = B \cup A$ and $AB = BA$.

1.1.2 Prove that $A \cup B = A \cup B\bar{A}$, $(A \cup B) - AB = A\bar{B} \cup \bar{A}B$.

1.1.3 Show that if $A \subset B$ then $A \cup B = B$ and $A \cap B = A$.

1.1.4 Prove DeMorgan's laws, i.e., $\overline{A \cup B} = \bar{A} \cap \bar{B}$ or $\overline{A \cap B} = \bar{A} \cup \bar{B}$.

1.1.5 Show that for every $n \geq 2$, $\overline{\left(\bigcup_{i=1}^{n} A_i\right)} = \bigcap_{i=1}^{n} \bar{A}_i$.

1.1.6 Show that if $A_1 \subset \cdots \subset A_N$ then $\sup_{1 \leq n \leq N} A_n = A_N$ and $\inf_{1 \leq n \leq N} A_n = A_1$.

1.1.7 Find $\lim_{n \to \infty} \left[0, 1 - \frac{1}{n}\right)$.

1.1.8 Find $\lim_{n \to \infty} \left(0, \frac{1}{n}\right)$.

1.1.9 Show that if $\mathcal{D} = \{A_1, \ldots, A_k\}$ is a partition of $\mathcal{S}$ then, for every B, $B = \bigcup_{i=1}^{n} A_i B$.

1.1.10 Prove that $\underline{\lim}_{n \to \infty} A_n \subset \overline{\lim}_{n \to \infty} A_n$.

1.1.11 Prove that $\bigcup_{n=1}^{\infty} A_n = \lim_{n \to \infty} \bigcup_{j=1}^{n} A_j$ and $\bigcap_{n=1}^{\infty} A_n = \lim_{n \to \infty} \bigcap_{j=1}^{n} A_j$.

1.1.12 Show that if $\{A_n\}$ is a sequence of pairwise disjoint sets, then $\lim_{n \to \infty} \bigcup_{j=n}^{\infty} A_j = \phi$.

1.1.13 Prove that $\overline{\lim}_{n \to \infty} (A_n \cup B_n) = \overline{\lim}_{n \to \infty} A_n \cup \overline{\lim}_{n \to \infty} B_n$.

1.1.14 Show that if $\{a_n\}$ is a sequence of nonnegative real numbers, then $\sup_{n \geq 1} [0, a_n) = [0, \sup_{n \geq 1} a_n)$.

1.1.15 Let $A \triangle B = A\bar{B} \cup B\bar{A}$ (symmetric difference). Let $\{A_n\}$ be a sequence of disjoint events; define $B_1 = A_1$, $B_{n+1} = B_n \triangle A_{n+1}$, $n \geq 1$. Prove that $\lim B_n = \bigcup_{n=1}^{\infty} A_n$.

1.1.16 Verify

(i) $A \triangle B = \bar{A} \triangle \bar{B}$.

(ii) $C = A \triangle B$ if and only if $A = B \triangle C$.

(iii) $\left(\bigcup_{n=1}^{\infty} A_n\right) \triangle \left(\bigcup_{n=1}^{\infty} B_n\right) \subset \bigcup_{n=1}^{\infty} (A_n \triangle B_n)$.

1.1.17 Prove that $\overline{\overline{\lim}}_{n\to\infty} A_n = \underline{\lim}_{n\to\infty} \bar{A}_n$.

Section 1.2

1.2.1 Let $\mathcal{A}$ be an algebra over $\mathcal{S}$. Show that if $A_1, A_2 \in \mathcal{A}$ then $A_1 A_2 \in \mathcal{A}$.

1.2.2 Let $\mathcal{S} = \{-, \ldots, -2, -1, 0, 1, 2, \ldots\}$ be the set of all integers. A set $A \subset \mathcal{S}$ is called symmetric if $A = -A$. Prove that the collection $\mathcal{A}$ of all symmetric subsets of $\mathcal{S}$ is an algebra.

1.2.3 Let $\mathcal{S} = \{-, \ldots, -2, -1, 0, 1, 2 \ldots\}$. Let $\mathcal{A}_1$ be the algebra of symmetric subsets of $\mathcal{S}$, and let $\mathcal{A}_2$ be the algebra generated by sets $A_n = \{-2, -1, i_1, \ldots, i_n\}$, $n \geq 1$, where $i_j \geq 0$, $j = 1, \ldots, n$.

(i) Show that $\mathcal{A}_3 = \mathcal{A}_1 \cap \mathcal{A}_2$ is an algebra.

(ii) Show that $\mathcal{A}_4 = \mathcal{A}_1 \cup \mathcal{A}_2$ is **not** an algebra.

1.2.4 Show that if $\mathcal{A}$ is a σ-field, $A_n \subset A_{n+1}$, for all $n \geq 1$, then $\lim_{n\to\infty} \bar{A}_n \in \mathcal{A}$.

Section 1.3

1.3.1 Let $F(x) = P\{(-\infty, x]\}$. Verify

(a) $P\{(a, b]\} = F(b) - F(a)$.

(b) $P\{(a, b)\} = F(b-) - F(a)$.

(c) $P\{[a, b)\} = F(b-) - F(a-)$.

1.3.2 Prove that $P\{A \cup B\} = P\{A\} + P\{B\bar{A}\}$.

1.3.3 A point (X, Y) is chosen in the unit square. Thus, $\mathcal{S} = \{(x, y) : 0 \leq x, y \leq 1\}$. Let $\mathcal{B}$ be the Borel σ-field on $\mathcal{S}$. For a Borel set B, we define

$$P\{B\} = \int\int_B dx dy.$$

Compute the probabilities of

$$B = \left\{(x, y) : x > \frac{1}{2}\right\}$$

$$C = \left\{(x, y) : x^2 + y^2 \leq 1\right\}$$

$$D = \{(x, y) : x + y \leq 1\}$$

$P\{D \cap B\}$, $P\{D \cap C\}$, $P\{C \cap B\}$.

1.3.4 Let $\mathcal{S} = \{x : 0 \le x < \infty\}$ and $\mathcal{B}$ the Borel σ-field on $\mathcal{S}$, generated by the sets $[0, x)$, $0 < x < \infty$. The probability function on $\mathcal{B}$ is $P\{B\} = \lambda \int_B e^{-\lambda x} dx$, for some $0 < \lambda < \infty$. Compute the probabilities

(i) $P\{X \le 1/\lambda\}$.

(ii) $P\left\{\frac{1}{\lambda} \le X \le \frac{2}{\lambda}\right\}$.

(iii) Let $B_n = \left[0, \left(1 + \frac{1}{n}\right)/\lambda\right)$. Compute $\lim_{n\to\infty} P\{B_n\}$ and show that it is equal to $P\left\{\lim_{n\to\infty} B_n\right\}$.

1.3.5 Consider an experiment in which independent trials are conducted sequentially. Let R_i be the result of the ith trial. $P\{R_i = 1\} = p$, $P\{R_i = 0\} = 1 - p$. The trials stop when $(R_1, R_2, \ldots, R_N)$ contains exactly two 1s. Notice that in this case, the number of trials N is random. Describe the sample space. Let w_n be a point of $\mathcal{S}$, which contains exactly n trials. $w_n = \{(i_1, \ldots, i_{n-1}, 1)\}$, $n \ge 2$, where $\sum_{j=1}^{n-1} i_j = 1$. Let $E_n = \{(i_1, \ldots, i_{n-1}, 1) : \sum_{j=1}^{n-1} i_j = 1\}$.

(i) Show that $\mathcal{D} = \{E_2, E_3, \ldots\}$ is a countable partition of $\mathcal{S}$.

(ii) Show that $P\{E_n\} = (n-1)p^2q^{n-2}$, where $0 < p < 1, q = 1 - p$, and prove that $\sum_{n=2}^{\infty} P\{E_n\} = 1$.

(iii) What is the probability that the experiment will require at least 5 trials?

1.3.6 In a parking lot there are 12 parking spaces. What is the probability that when you arrive, assuming cars fill the spaces at random, there will be four adjacent spaces vacant, while all other spaces filled?

Section 1.4

1.4.1 Show that if A and B are independent, then $\bar{A}$ and $\bar{B}$, A and $\bar{B}$, $\bar{A}$ and B are independent.

1.4.2 Show that if three events are mutually independent, then if we replace any event with its complement, the new collection is still mutually independent.

1.4.3 Two digits are chosen from the set $\mathcal{P} = \{0, 1, \ldots, 9\}$, without replacement. The order of choice is immaterial. The probability function assigns every possible set of two the same probability. Let $A_i (i = 0, \ldots, 9)$ be the event that the chosen set contains the digit i. Show that for any $i \ne j$, A_i and A_j are **not** independent.

1.4.4 Let $A_1, \ldots, A_n$ be mutually independent events. Show that

$$P\left\{\bigcup_{i=1}^{n} A_i\right\} = 1 - \prod_{i=1}^{n} P\{\bar{A}_i\}.$$

1.4.5 If an event A is independent of itself, then $P\{A\} = 0$ or $P(A) = 1$.

1.4.6 Consider the random walk model of Example 1.2.

(i) What is the probability that after n steps the particle will be on a positive integer?

(ii) Compute the probability that after $n = 7$ steps the particle will be at $x = 1$.

(iii) Let p be the probability that in each trial the particle goes one step to the right. Let A_n be the event that the particle returns to the origin after n steps. Compute $P\{A_n\}$ and show, by using the Borel–Cantelli Lemma, that if $p \neq \frac{1}{2}$ then $P\{A_n, i.o.\} = 0$.

1.4.7 Prove that

(i) $\sum_{k=0}^{n} \binom{n}{k} = 2^k.$

(ii) $\sum_{k=0}^{n} \binom{M}{k}\binom{N-M}{n-k} = \binom{N}{n}.$

(iii) $\sum_{k=0}^{n} k\binom{M}{k}\binom{N-M}{n-k} = n\frac{M}{N}\binom{N}{n} = M\binom{N-1}{n-1}.$

1.4.8 What is the probability that the birthdays of $n = 12$ randomly chosen people will fall in 12 different calendar months?

1.4.9 A stick is broken at random into three pieces. What is the probability that these pieces can form a triangle?

1.4.10 There are $n = 10$ particles and $m = 5$ cells. Particles are assigned to the cells at random.

(i) What is the probability that each cell contains at least one particle?

(ii) What is the probability that all 10 particles are assigned to the first 3 cells?

Section 1.5

1.5.1 Let F be a discrete distribution concentrated on the jump points $-\infty < \xi_1 < \xi_2 < \cdots < \infty$. Let $p_i = dF(\xi_i)$, $i = 1, 2, \ldots$. Define the function

$$U(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0. \end{cases}$$

(i) Show that, for all $-\infty < x < \infty$

$$F(x) = \sum_{i=1}^{\infty} p_i U(x - \xi_i)$$
$$= \sum_{i=1}^{\infty} p_i I(\xi_i \leq x).$$

(ii) For $h > 0$, define

$$D_h U(x) = \frac{1}{h}[U(x + h) - U(x)] = \frac{1}{h}[I(x \geq -h) - I(x \geq 0)].$$

Show that

$$\int_{-\infty}^{\infty} \sum_{i=1}^{\infty} p_i D_h U(x - \xi_i) dx = 1 \quad \text{for all } h > 0.$$

(iii) Show that for any continuous function $g(x)$, such that $\sum_{i=1}^{\infty} p_i |g(\xi_i)| < \infty$,

$$\lim_{h \to 0} \int_{-\infty}^{\infty} \sum_{i=1}^{\infty} p_i g(x) D_h U(x - \xi_i) dx = \sum_{i=1}^{\infty} p_i g(\xi_i).$$

1.5.2 Let X be a random variable having a discrete distribution, with jump points $\xi_i = i$, and $p_i = dF(\xi_i) = e^{-2}\frac{2^i}{i!}$, $i = 0, 1, 2, \ldots$. Let $Y = X^3$. Determine the p.d.f. of Y.

1.5.3 Let X be a discrete random variable assuming the values $\{1, 2, \ldots, n\}$ with probabilities

$$p_i = \frac{2i}{n(n+1)}, \quad i = 1, \ldots, n.$$

(i) Find $E\{X\}$.

(ii) Let $g(X) = X^2$; find the p.d.f. of $g(X)$.

1.5.4 Consider a discrete random variable X, with jump points on $\{1, 2, \ldots\}$ and p.d.f.

$$f_X(n) = \frac{c}{n^2}, \quad n = 1, 2, \ldots$$

where c is a normalizing constant.

(i) Does $E\{X\}$ exist?

(ii) Does $E\{X/\log X\}$ exist?

1.5.5 Let X be a discrete random variable whose distribution has jump points at $\{x_1, x_2, \ldots, x_k\}$, $1 \leq k \leq \infty$. Assume also that $E\{|X|\} < \infty$. Show that for any linear transformation $Y = \alpha + \beta x$, $\beta \neq 0$, $-\infty < \alpha < \infty$, $E\{Y\} = \alpha + \beta E\{X\}$. (The result is trivially true for $\beta = 0$).

1.5.6 Consider two discrete random variables (X, Y) having a joint p.d.f.

$$f_{XY}(j, n) = \frac{e^{-\lambda}}{j!(n-j)!}\left(\frac{p}{1-p}\right)^j (\lambda(1-p))^n, \quad j = 0, 1, \ldots, n,$$
$$n = 0, 1, 2, \ldots.$$

(i) Find the marginal p.d.f. of X.

(ii) Find the marginal p.d.f. of Y.

(iii) Find the conditional p.d.f. $f_{X|Y}(j \mid n)$, $n = 0, 1, \ldots$.

(iv) Find the conditional p.d.f. $f_{Y|X}(n \mid j)$, $j = 0, 1, \ldots$.

(v) Find $E\{Y \mid X = j\}$, $j = 0, 1, \ldots$.

(vi) Show that $E\{Y\} = E\{E\{Y \mid X\}\}$.

1.5.7 Let X be a discrete random variable, $X \in \{0, 1, 2, \ldots\}$ with p.d.f.

$$f_X(n) = e^{-n} - e^{-(n+1)}, \quad n = 0, 1, \ldots.$$

Consider the partition $\mathcal{D} = \{A_1, A_2, A_3\}$, where

$$A_1 = \{w : X(w) < 2\},$$
$$A_2 = \{w : 2 \leq X(w) < 4\},$$
$$A_3 = \{w : 4 \leq X(w)\}.$$

(i) Find the conditional p.d.f.

$$f_{X|\mathcal{D}}(x \mid A_i), \quad i = 1, 2, 3.$$

(ii) Find the conditional expectations $E\{X \mid A_i\}$, $i = 1, 2, 3$.

(iii) Specify the random variable $E\{X \mid \mathcal{D}\}$.

1.5.8 For a given λ, $0 < \lambda < \infty$, define the function $P(j;\lambda) = e^{-\lambda}\sum_{l=0}^{j}\frac{\lambda^l}{l!}$.

(i) Show that, for a fixed nonnegative integer j, $F_j(x)$ is a distribution function, where

$$F_j(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1 - P(j-1;x), & \text{if } x \geq 0 \end{cases}$$

and where $P(j;0) = I\{j \geq 0\}$.

(ii) Show that $F_j(x)$ is absolutely continuous and find its p.d.f.

(iii) Find $E\{X\}$ according to $F_j(x)$.

1.5.9 Let X have an absolutely continuous distribution function with p.d.f.

$$f(x) = \begin{cases} 3x^2, & \text{if } 0 \leq x \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

Find $E\{e^{-X}\}$.

Section 1.6

1.6.1 Consider the absolutely continuous distribution

$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ x, & \text{if } 0 \leq x < 1 \\ 1, & \text{if } 1 \leq x \end{cases}$$

of a random variable X. By considering the sequences of simple functions

$$X_n(w) = \sum_{i=1}^{n} \frac{i-1}{n} I\left\{\frac{i-1}{n} \leq X(w) < \frac{i}{n}\right\}, \quad n \geq 1$$

and

$$X_n^2(w) = \sum_{i=1}^{n} \left(\frac{i-1}{n}\right)^2 I\left\{\frac{i-1}{n} \leq X(w) < \frac{i}{n}\right\}, \quad n \geq 1,$$

show that

$$\lim_{n\to\infty} E\{X_n\} = \int_0^1 x dx = \frac{1}{2}$$

and

$$\lim_{n\to\infty} E\{X_n^2\} = \int_0^1 x^2 dx = \frac{1}{3}.$$

1.6.2 Let X be a random variable having an absolutely continuous distribution F, such that $F(0) = 0$ and $F(1) = 1$. Let f be the corresponding p.d.f.

(i) Show that the Lebesgue integral

$$\int_0^1 x\,P\{dx\} = \lim_{n\to\infty} \sum_{i=1}^{2^n} \frac{i-1}{2^n}\left[F\left(\frac{i}{2^n}\right) - F\left(\frac{i-1}{2^n}\right)\right].$$

(ii) If the p.d.f. f is continuous on $(0, 1)$, then

$$\int_0^1 x\,P\{dx\} = \int_0^1 x f(x) dx,$$

which is the Riemann integral.

1.6.3 Let X, Y be independent identically distributed random variables and let $E\{X\}$ exist. Show that

$$E\{X \mid X + Y\} = E\{Y \mid X + Y\} = \frac{X+Y}{2} \text{ a.s.}$$

1.6.4 Let $X_1, \ldots, X_n$ be i.i.d. random variables and let $E\{X_1\}$ exist. Let $S_n = \sum_{j=1}^{n} X_j$. Then, $E\{X_1 \mid S_n\} = \frac{S_n}{n}$, a.s.

1.6.5 Let

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0 \\ \frac{1}{4}, & \text{if } x = 0 \\ \frac{1}{4} + \frac{1}{2}x^3, & \text{if } 0 < x < 1 \\ 1, & \text{if } 1 \le x. \end{cases}$$

Find $E\{X\}$ and $E\{X^2\}$.

1.6.6 Let $X_1, \ldots, X_n$ be Bernoulli random variables with $P\{X_i = 1\} = p$. If $n = 100$, how large should p be so that $P\{S_n < 100\} < 0.1$, when $S_n = \sum_{i=1}^{n} X_i$?

1.6.7 Prove that if $E\{|X|\} < \infty$, then, for every $A \in \mathcal{F}$,

$$E\{|X| I_A(X)\} \leq E\{|X|\}.$$

1.6.8 Prove that if $E\{|X|\} < \infty$ and $E\{|Y|\} < \infty$, then $E\{X + Y\} = E\{X\} + E\{Y\}$.

1.6.9 Let $\{X_n\}$ be a sequence of i.i.d. random variables with common c.d.f.

$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1 - e^{-x}, & \text{if } x \geq 0. \end{cases}$$

Let $S_n = \sum_{i=1}^{n} X_i$.

(i) Use the Borel–Cantelli Lemma to show that $\lim_{n \to \infty} S_n = \infty$ a.s.

(ii) What is $\lim_{n \to \infty} E\left\{\frac{S_n}{1 + S_n}\right\}$?

1.6.10 Consider the distribution function F of Example 1.11, with $\alpha = .9$, $\lambda = .1$, and $\mu = 1$.

(i) Determine the lower quartile, the median, and the upper quartile of $F_{ac}(x)$.

(ii) Tabulate the values of $F_{\mathrm{d}}(x)$ for $x = 0, 1, 2, \ldots$ and determine the lower quartile, median, and upper quartile of $F_{\mathrm{d}}(x)$.

(iii) Determine the values of the median and the interquartile range IQR of $F(x)$.

(iv) Determine $P\{0 < X < 3\}$.

1.6.11 Consider the Cauchy distribution with p.d.f.

$$f(x; \mu, \sigma) = \frac{1}{\pi \sigma} \cdot \frac{1}{1 + (x - \mu)^2 / \sigma^2}, \quad -\infty < x < \infty,$$

with $\mu = 10$ and $\sigma = 2$.

(i) Write the formula of the c.d.f. $F(x)$.

(ii) Determine the values of the median and the interquartile range of $F(x)$.

1.6.12 Let X be a random variable having the p.d.f. $f(x) = e^{-x}$, $x \geq 0$. Determine the p.d.f. and the median of

(i) $Y = \log X$,

(ii) $Y = \exp\{-X\}$.

1.6.13 Let X be a random variable having a p.d.f. $f(x) = \frac{1}{\pi}, -\frac{\pi}{2} \leq x \leq \frac{\pi}{2}$. Determine the p.d.f. and the median of

(i) $Y = \sin X$,

(ii) $Y = \cos X$,

(iii) $Y = \tan X$.

1.6.14 Prove that if $E\{|X|\} < \infty$ then

$$E\{X\} = -\int_{-\infty}^{0} F(x)dx + \int_{0}^{\infty} (1 - F(x))dx.$$

1.6.15 Apply the result of the previous problem to derive the expected value of a random variable X having an exponential distribution, i.e.,

$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1 - e^{-\lambda x}, & \text{if } x \geq 0. \end{cases}$$

1.6.16 Prove that if $F(x)$ is symmetric around η, i.e.,

$$F(\eta - x) = 1 - F(\eta + x-), \quad \text{for all } 0 \leq x < \infty,$$

then $E\{X\} = \eta$, provided $E\{|X|\} < \infty$.

Section 1.7

1.7.1 Let (X, Y) be random variables having a joint p.d.f.

$$f_{XY}(x, y) = \begin{cases} 1, & \text{if } -1 < x < 1, 0 < y < 1 - |x| \\ 0, & \text{otherwise.} \end{cases}$$

(i) Find the marginal p.d.f. of Y.

(ii) Find the conditional p.d.f. of X given $\{Y = y\}, 0 < y < 1$.

1.7.2 Consider random variables $\{X, Y\}$. X is a discrete random variable with jump points $\{0, 1, 2, \ldots\}$. The marginal p.d.f. of X is $f_X(x) = e^{-\lambda}\frac{\lambda^x}{x!}, x = 0, 1, \ldots, 0 < \lambda < \infty$. The conditional distribution of Y given $\{X = x\}$, $x \geq 1$, is

$$F_{Y|X}(y \mid x) = \begin{cases} 0, & y < 0 \\ y/x, & 0 \leq y \leq x \\ 1, & x < y. \end{cases}$$

When $\{X = 0\}$

$$F_{Y|X}(y \mid 0) = \begin{cases} 0, & y < 0 \\ 1, & y \geq 0. \end{cases}$$

(i) Find $E\{Y\}$.

(ii) Show that the c.d.f. of Y has discontinuity at $y = 0$, and $F_Y(0) - F_Y(0-) = e^{-\lambda}$.

(iii) For each $0 < y < \infty$, $F'_Y(y) = f_Y(y)$, where $\int_0^\infty f_Y(y)dy = 1 - e^{-\lambda}$. Show that, for $y > 0$,

$$f_Y(y) = \sum_{n=1}^{\infty} I\{n-1 < y < n\} e^{-\lambda} \sum_{x=n}^{\infty} \frac{1}{x} \cdot \frac{\lambda^x}{x!},$$

and prove that $\int_0^\infty f_Y(y)dy = 1 - e^{-\lambda}$.

(iv) Derive the conditional p.d.f. of X given $\{Y = y\}$, $0 < y < \infty$, and find $E\{X \mid Y = y\}$.

1.7.3 Show that if X, Y are independent random variables, $E\{|X|\} < \infty$ and $E\{|Y| < \infty\}$, then $E\{XY\} = E\{X\}E\{Y\}$. More generally, if g, h are integrable, then if X, Y are independent, then

$$E\{g(X)h(Y)\} = E\{g(X)\}E\{h(Y)\}.$$

1.7.4 Show that if X, Y are independent, absolutely continuous, with p.d.f. f_X and f_Y, respectively, then the p.d.f. of $T = X + Y$ is

$$f_T(t) = \int_{-\infty}^{\infty} f_X(x) f_Y(t - x)dx.$$

[f_T is the **convolution** of f_X and f_Y.]

Section 1.8

1.8.1 Prove that if $E\{|X|^r\}$ exists, $r \geq 1$, then $\lim_{a \to \infty} (a)^r P\{|X| \geq a\} = 0$.

1.8.2 Let X_1, X_2 be i.i.d. random variables with $E\{X_1^2\} < \infty$. Find the correlation between X_1 and $T = X_1 + X_n$.

1.8.3 Let $X_1, \ldots, X_n$ be i.i.d. random variables; find the correlation between X_1 and the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

1.8.4 Let X have an absolutely continuous distribution with p.d.f.

$$f_X(x) = \begin{cases} 0, & \text{if } x < 0 \\ \dfrac{\lambda^m}{(m-1)!} x^{m-1} e^{-\lambda x}, & \text{if } x \geq 0 \end{cases}$$

where $0 < \lambda < \infty$ and m is an integer, $m \geq 2$.

(i) Derive the m.g.f. of X. What is its domain of convergence?

(ii) Show, by differentiating the m.g.f. $M(t)$, that $E\{X^r\} = \dfrac{m(m+1)\cdots(m+r-1)}{\lambda^r}$, $r \geq 1$.

(iii) Obtain the first four central moments of X.

(iv) Find the coefficients of skewness β_1 and kurtosis β_2.

1.8.5 Let X have an absolutely continuous distribution with p.d.f.

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise.} \end{cases}$$

(i) What is the m.g.f. of X?

(ii) Obtain $E\{X\}$ and $V\{X\}$ by differentiating the m.g.f.

1.8.6 Random variables X_1, X_2, X_3 have the covariance matrix

$$\mathfrak{V} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}.$$

Find the variance of $Y = 5x_1 - 2x_2 + 3x_3$.

1.8.7 Random variables $X_1, \ldots, X_n$ have the covariance matrix

$$\mathfrak{V} = I + J,$$

where J is an $n \times n$ matrix of 1s. Find the variance of $\bar{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

1.8.8 Let X have a p.d.f.

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \quad -\infty < x < \infty.$$

Find the characteristic function ϕ of X.

1.8.9 Let $X_1, \ldots, X_n$ be i.i.d., having a common characteristic function ϕ. Find the characteristic function of $\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$.

1.8.10 If ϕ is a characteristic function of an absolutely continuous distribution, its p.d.f. is

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi(t)dt.$$

Show that the p.d.f. corresponding to

$$\phi(t) = \begin{cases} 1 - |t|, & |t| \leq 1 \\ 0, & |t| > 1 \end{cases}$$

is

$$f(x) = \frac{1 - \cos x}{\pi x^2}, \quad |x| \leq \frac{\pi}{2}.$$

1.8.11 Find the m.g.f. of a random variable whose p.d.f. is

$$f_X(x) = \begin{cases} \dfrac{a - |x|}{a^2}, & \text{if } |x| \leq a \\ 0, & \text{if } |x| > a, \end{cases}$$

$0 < a < \infty$.

1.8.12 Prove that if ϕ is a characteristic function, then $|\phi(t)|^2$ is a characteristic function.

1.8.13 Prove that if ϕ is a characteristic function, then

(i) $\lim_{|t|\to\infty} \phi(t) = 0$ if X has an absolutely continuous distribution.

(ii) $\limsup_{|t|\to\infty} |\phi(t)| = 1$ if X is discrete.

1.8.14 Let X be a discrete random variable with p.d.f.

$$f(x) = \begin{cases} e^{-\lambda}\dfrac{\lambda^x}{x!}, & x = 0, 1, \ldots \\ 0, & \text{otherwise.} \end{cases}$$

Find the p.g.f. of X.

Section 1.9

1.9.1 Let F_n, $n \geq 1$, be the c.d.f. of a discrete uniform distribution on $\left\{\frac{1}{n}, \frac{2}{n}, \ldots, 1\right\}$. Show that $F_n(x) \xrightarrow{d} F(x)$, as $n \to \infty$, where

$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ x, & \text{if } 0 \leq x \leq 1 \\ 1, & \text{if } 1 < x. \end{cases}$$

1.9.2 Let $B(j; n, p)$ denote the c.d.f. of the binomial distribution with p.d.f.

$$b(j; n, p) = \binom{n}{j} p^j (1-p)^{n-j}, \quad j = 0, 1, \ldots, n,$$

where $0 < p < 1$. Consider the sequence of binomial distributions

$$F_n(x) = B\left([x]; n, \frac{1}{2n}\right) I\{0 \leq x \leq n\} + I\{x > n\}, \quad n \geq 1.$$

What is the weak limit of $F_n(x)$?

1.9.3 Let $X_1, X_2, \ldots, X_n, \ldots$ be i.i.d. random variables such that $V\{X_1\} = \sigma^2 < \infty$, and $\mu = E\{X_1\}$. Use Chebychev's inequality to prove that $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu$ as $n \to \infty$.

1.9.4 Let $X_1, X_2, \ldots$ be a sequence of binary random variables, such that $P\{X_n = 1\} = \frac{1}{n}$, and $P\{X_n = 0\} = 1 - \frac{1}{n}$, $n \geq 1$.

(i) Show that $X_n \xrightarrow{r} 0$ as $n \to \infty$, for any $r \geq 1$.

(ii) Show from the definition that $X_n \xrightarrow{p} 0$ as $n \to \infty$.

(iii) Show that if $\{X_n\}$ are independent, then $P\{X_n = 1, i.o.\} = 1$. Thus, $X_n \not\to 0$ a.s.

1.9.5 Let $\epsilon_1, \epsilon_2, \ldots$ be independent r.v., such that $E\{\epsilon_n\} = \mu$ and $V\{\epsilon_n\} = \sigma^2$ for all $n \geq 1$. Let $X_1 = \epsilon_1$ and for $n \geq 2$, let $X_n = \beta X_{n-1} + \epsilon_n$, where $-1 < \beta < 1$. Show that $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{2} \frac{\mu}{1-\beta}$, as $n \to \infty$.

1.9.6 Prove that convergence in the rth mean, for some $r > 0$ implies convergence in the sth mean, for all $0 < s < r$.

1.9.7 Let $X_1, X_2, \ldots, X_n, \ldots$ be i.i.d. random variables having a common rectangular distribution $R(0, \theta), 0 < \theta < \infty$. Let $X_{(n)} = \max\{X_1, \ldots, X_n\}$. Let $\epsilon > 0$. Show that $\sum_{n=1}^{\infty} P_\theta\{X_{(n)} < \theta - \epsilon\} < \infty$. Hence, by the Borel–Cantelli Lemma, $X_{(n)} \xrightarrow{\text{a.s.}} \theta$, as $n \to \infty$. The $R(0, \theta)$ distribution is

$$F_\theta(x) = \begin{cases} 0, & \text{if } x < 0 \\ \dfrac{x}{\theta}, & \text{if } 0 \le x \le \theta \\ 1, & \text{if } \theta < x \end{cases}$$

where $0 < \theta < \infty$.

1.9.8 Show that if $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{p} Y$, then $P\{w : X(w) \neq Y(w)\} = 0$.

1.9.9 Let $X_n \xrightarrow{p} X$, $Y_n \xrightarrow{p} Y$, $P\{w : X(w) \neq Y(w)\} = 0$. Then, for every $\epsilon > 0$,

$$P\{|X_n - Y_n| \ge \epsilon\} \to 0, \quad \text{as } n \to \infty.$$

1.9.10 Show that if $X_n \xrightarrow{d} C$ as $n \to \infty$, where C is a constant, then $X_n \xrightarrow{p} C$.

1.9.11 Let $\{X_n\}$ be such that, for any $p > 0$, $\sum_{n=1}^{\infty} E\{|X_n|^p\} < \infty$. Show that $X_n \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$.

1.9.12 Let $\{X_n\}$ be a sequence of i.i.d. random variables. Show that $E\{|X_1|\} < \infty$ if and only if $\sum_{n=1}^{\infty} P\{|X_1| > \epsilon \cdot n\} < \infty$. Show that $E|X_1| < \infty$ if and only if $\dfrac{X_n}{n} \xrightarrow{\text{a.s.}} 0$.

Section 1.10

1.10.1 Show that if X_n has a p.d.f. f_n and X has a p.d.f. $g(x)$ and if $\int |f_n(x) - g(x)|dx \to 0$ as $n \to \infty$, then $\sup_B |P_n\{B\} - P\{B\}| \to 0$ as $n \to \infty$, for all Borel sets B. (Ferguson, 1996, p. 12).

1.10.2 Show that if $\mathbf{a}'\mathbf{X}_n \xrightarrow{d} \mathbf{a}'\mathbf{X}$ as $n \to \infty$, for all vectors $\mathbf{a}$, then $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ (Ferguson, 1996, p. 18).

1.10.3 Let $\{X_n\}$ be a sequence of i.i.d. random variables. Let $Z_n = \sqrt{n}(\bar{X}_n - \mu)$, $n \geq 1$, where $\mu = E\{X_1\}$ and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Let $V\{X_1\} < \infty$. Show that $\{Z_n\}$ is tight.

1.10.4 Let $B(n, p)$ designate a discrete random variable, having a binomial distribution with parameter (n, p). Show that $\left\{B\left(n, \frac{1}{2n}\right)\right\}$ is tight.

1.10.5 Let $P(\lambda)$ designate a discrete random variable, which assumes on $\{0, 1, 2, \ldots\}$ the p.d.f. $f(x) = e^{-\lambda}\frac{\lambda^x}{x!}$, $x = 0, 1, \ldots, 0 < \lambda < \infty$. Using the continuity theorem prove that $B(n, p_n) \xrightarrow{d} P(\lambda)$ if $\lim_{n\to\infty} np_n = \lambda$.

1.10.6 Let $X_n \sim B\left(n, \frac{1}{2n}\right)$, $n \geq 1$. Compute $\lim_{n\to\infty} E\{e^{-X_n}\}$.

Section 1.11

1.11.1 (Khinchin WLLN). Use the continuity theorem to prove that if X_1, $X_2, \ldots, X_n, \ldots$ are i.i.d. random variables, then $\bar{X}_n \xrightarrow{p} \mu$, where $\mu = E\{X_1\}$.

1.11.2 (Markov WLLN). Prove that if $X_1, X_2, \ldots, X_n, \ldots$ are independent random variables and if $\mu_k = E\{X_k\}$ exists, for all $k \geq 1$, and $E|X_k - \mu_k|^{1+\delta} < \infty$ for some $\delta > 0$, all $k \geq 1$, then $\frac{1}{n^{1+\delta}}\sum_{k=1}^{n} E|X_k - \mu_k|^{1+\delta} \to 0$ as $n \to \infty$ implies that $\frac{1}{n}\sum_{k=1}^{n}(X_k - \mu_k) \xrightarrow{p} 0$ as $n \to \infty$.

1.11.3 Let $\{\mathbf{X}_n\}$ be a sequence of random vectors. Prove that if $\bar{\mathbf{X}}_n \xrightarrow{d} \boldsymbol{\mu}$ then $\bar{\mathbf{X}}_n \xrightarrow{p} \boldsymbol{\mu}$, where $\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{j=1}^{n}\mathbf{X}_j$ and $\boldsymbol{\mu} = E\{\mathbf{X}_1\}$.

1.11.4 Let $\{X_n\}$ be a sequence of i.i.d. random variables having a common p.d.f.

$$f(x) = \begin{cases} 0, & \text{if } x < 0 \\ \frac{\lambda^m}{(m-1)!}x^{m-1}e^{-\lambda x}, & \text{if } x \geq 0, \end{cases}$$

where $0 < \lambda < \infty, m = 1, 2, \ldots$. Use Cantelli's Theorem (Theorem 1.11.1) to prove that $\bar{X}_n \xrightarrow{\text{a.s.}} \frac{m}{\lambda}$, as $n \to \infty$.

1.11.5 Let $\{X_n\}$ be a sequence of independent random variables where

$$X_n \sim R(-n, n)/n$$

and $R(-n, n)$ is a random variable having a uniform distribution on $(-n, n)$, i.e.,

$$f_n(x) = \frac{1}{2n} 1_{(-n,n)}(x).$$

Show that $\bar{X}_n \xrightarrow{\text{a.s.}} 0$, as $n \to \infty$. [Prove that condition (1.11.6) holds].

1.11.6 Let $\{X_n\}$ be a sequence of i.i.d. random variables, such that $|X_n| \leq C$ a.s., for all $n \geq 1$. Show that $\bar{X}_n \xrightarrow{\text{a.s.}} \mu$ as $n \to \infty$, where $\mu = E\{X_1\}$.

1.11.7 Let $\{X_n\}$ be a sequence of independent random variables, such that

$$P\{X_n = \pm 1\} = \frac{1}{2}\left(1 - \frac{1}{2^n}\right)$$

and

$$P\{X_n = \pm n\} = \frac{1}{2} \cdot \frac{1}{2^n}, \quad n \geq 1.$$

Prove that $\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} 0$, as $n \to \infty$.

Section 1.12

1.12.1 Let $X \sim P(\lambda)$, i.e.,

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, \ldots.$$

Apply the continuity theorem to show that

$$\frac{X - \lambda}{\sqrt{\lambda}} \xrightarrow{d} N(0, 1), \quad \text{as } \lambda \to \infty.$$

1.12.2 Let $\{X_n\}$ be a sequence of i.i.d. discrete random variables, and $X_1 \sim P(\lambda)$. Show that

$$\frac{S_n - n\lambda}{\sqrt{n\lambda}} \xrightarrow{d} N(0, 1), \quad \text{as } n \to \infty.$$

What is the relation between problems 1 and 2?

1.12.3 Let $\{X_n\}$ be i.i.d., binary random variables, $P\{X_n = 1\} = P\{X_n = 0\} = \frac{1}{2}$, $n \geq 1$. Show that

$$\frac{\sum_{i=1}^{n} i X_i - \frac{n(n+1)}{4}}{B_n} \xrightarrow{d} N(0,1), \quad \text{as } n \to \infty,$$

where $B_n^2 = \frac{n(n+1)(2n+1)}{24}$, $n \geq 1$.

1.12.4 Consider a sequence $\{X_n\}$ of independent discrete random variables, $P\{X_n = n\} = P\{X_n = -n\} = \frac{1}{2}$, $n \geq 1$. Show that this sequence satisfies the CLT, in the sense that

$$\frac{\sqrt{6}\, S_n}{\sqrt{n(n+1)(2n+1)}} \xrightarrow{d} N(0,1), \quad \text{as } n \to \infty.$$

1.12.5 Let $\{X_n\}$ be a sequence of i.i.d. random variables, having a common absolutely continuous distribution with p.d.f.

$$f(x) = \begin{cases} \dfrac{1}{2|x| \log^2 |x|}, & \text{if } |x| < \dfrac{1}{e} \\ 0, & \text{if } |x| \geq \dfrac{1}{e}. \end{cases}$$

Show that this sequence satisfies the CLT, i.e.,

$$\sqrt{n}\, \frac{\bar{X}_n}{\sigma} \xrightarrow{d} N(0,1), \quad \text{as } n \to \infty,$$

where $\sigma^2 = V\{X\}$.

1.12.6 (**i**) Show that

$$\frac{(G(1,n) - n)}{\sqrt{n}} \xrightarrow{d} N(0,1), \quad \text{as } n \to \infty$$

where $G(1,n)$ is an absolutely continuous random variable with a p.d.f.

$$g_n(x) = \begin{cases} 0, & \text{if } x < 0 \\ \dfrac{1}{(n-1)!} x^{n-1} e^{-x}, & x \geq 0. \end{cases}$$

(**ii**) Show that, for large n,

$$g_n(n) = \frac{1}{(n-1)!} n^{n-1} e^{-n} \approx \frac{1}{\sqrt{2\pi}\ \sqrt{n}}.$$

Or

$$n! \approx \sqrt{2\pi}\ n^{n+\frac{1}{2}} e^{-n} \ \text{ as } \ n \to \infty.$$

This is the famous **Stirling approximation**.

Section 1.13

1.13.1 Let $X_n \sim R(-n, n), n \geq 1$. Is the sequence $\{X_n\}$ uniformly integrable?

1.13.2 Let $Z_n = \dfrac{X_n - n}{\sqrt{n}} \sim N(0, 1), n \geq 1$. Show that $\{Z_n\}$ is uniformly integrable.

1.13.3 Let $\{X_1, X_2, \ldots, X_n, \ldots\}$ and $\{Y_1, Y_2, \ldots, Y_n, \ldots\}$ be two independent sequences of i.i.d. random variables. Assume that $0 < V\{X_1\} = \sigma_x^2 < \infty$, $0 < V\{Y_1\} = \sigma_y^2 < \infty$. Let $f(x, y)$ be a continuous function on R^2, having continuous partial derivatives. Find the limiting distribution of $\sqrt{n}(f(\bar{X}_n, \bar{Y}_n) - f(\xi, \eta))$, where $\xi = E\{X_1\}, \eta = E\{Y_1\}$. In particular, find the limiting distribution of $R_n = \bar{X}_n / \bar{Y}_n$, when $\eta > 0$.

1.13.4 We say that $X \sim E(\mu), 0 < \mu < \infty$, if its p.d.f. is

$$f(x) = \begin{cases} 0, & \text{if } x < 0 \\ \mu e^{-\mu x}, & \text{if } x \geq 0. \end{cases}$$

Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of i.i.d. random variables, $X_1 \sim E(\mu), 0 < \mu < \infty$. Let $\bar{X}_n = \dfrac{1}{n} \sum_{i=1}^{n} X_i$.

(a) Compute $V\{e^{\bar{X}_n}\}$ exactly.
(b) Approximate $V\{e^{\bar{X}_n}\}$ by the delta method.

1.13.5 Let $\{X_n\}$ be i.i.d. Bernoulli random variables, i.e., $X_1 \sim B(1, p), 0 < p < 1$. Let $\hat{p}_n = \dfrac{1}{n} \sum_{i=1}^{n} X_i$ and

$$W_n = \log\left(\frac{\hat{p}_n}{1 - \hat{p}_n}\right).$$

Use the delta method to find an approximation, for large values of n, of

(i) $E\{W_n\}$

(ii) $V\{W_n\}$.

Find the asymptotic distribution of $\sqrt{n}\left(W_n - \log\left(\frac{p}{1-p}\right)\right)$.

1.13.6 Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having a common continuous distribution function $F(x)$. Let $F_n(x)$ be the empirical distribution function. Fix a value x_0 such that $0 < F_n(x_0) < 1$.

(i) Show that $nF_n(x_0) \sim B(n, F(x_0))$.

(ii) What is the asymptotic distribution of $F_n(x_0)$ as $n \to \infty$?

1.13.7 Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having a standard Cauchy distribution. What is the asymptotic distribution of the sample median $F_n^{-1}\left(\frac{1}{2}\right)$?

PART IV: SOLUTIONS TO SELECTED PROBLEMS

1.1.5 For $n = 2$, $\overline{A_1 \cup A_2} = \bar{A}_1 \cap \bar{A}_2$. By induction on n, assume that $\overline{\bigcup_{i=1}^{k} A_i} = \bigcap_{i=1}^{k} \bar{A}_i$ for all $k = 2, \ldots, n$. For $k = n+1$,

$$\overline{\bigcup_{k=1}^{n+1} A_i} = \overline{\left(\bigcup_{i=1}^{n} A_i\right) \cup A_{n+1}} = \overline{\left(\bigcup_{i=1}^{n} A_i\right)} \cap \bar{A}_{n+1}$$
$$= \bigcap_{i=1}^{n} \bar{A}_i \cap \bar{A}_{n+1} = \bigcap_{i=1}^{n+1} \bar{A}_i.$$

1.1.10 We have to prove that $\left(\underline{\lim}_{n\to\infty} A_n\right) \subset \left(\overline{\lim}_{n\to\infty} A_n\right)$. For an elementary event $w \in \mathcal{S}$, let

$$I_{A_n}(w) = \begin{cases} 1, & \text{if } w \in A_n \\ 0, & \text{if } w \notin A_n. \end{cases}$$

Thus, if $w \in \underline{\lim}_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_n$, there exists an integer $K(w)$ such that

$$\prod_{n \geq K(w)}^{\infty} I_{A_n}(w) = 1.$$

Accordingly, for all $n \geq 1$, $w \in \bigcup_{k=n}^{\infty} A_k$. Here $w \in \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k = \overline{\lim}_{n\to\infty} A_n$.

1.1.15 Let $\{A_n\}$ be a sequence of disjoint events. For all $n \geq 1$, we define

$$\begin{aligned} B_n &= B_{n-1} \Delta A_n \\ &= B_{n-1} \bar{A}_n \cup \bar{B}_{n-1} A_n \end{aligned}$$

and

$$\begin{aligned} B_1 &= A_1 \\ B_2 &= \bar{A}_1 A_2 \cup A_1 \bar{A}_2 \\ B_3 &= \overline{(\bar{A}_1 A_2 \cup A_1 \bar{A}_2)} A_3 \cup (\bar{A}_1 A_2 \cup A_1 \bar{A}_2) \bar{A}_3 \\ &= (\overline{\bar{A}_1 A_2} \cap \overline{A_1 \bar{A}_2}) A_3 \cup \bar{A}_1 A_2 \bar{A}_3 \cup A_1 \bar{A}_2 \bar{A}_3 \\ &= (A_1 \cup \bar{A}_2)(\bar{A}_1 \cup A_2) A_3 \cup \bar{A}_1 A_2 \bar{A}_3 \cup A_1 \bar{A}_2 \bar{A}_3 \\ &= A_1 A_2 A_3 \cup \bar{A}_1 \bar{A}_2 A_3 \cup \bar{A}_1 A_2 \bar{A}_3 \cup A_1 \bar{A}_2 \bar{A}_3. \end{aligned}$$

By induction on n we prove that, for all $n \geq 2$,

$$B_n = \left(\bigcap_{j=1}^{n} A_j \right) \cup \left(\bigcup_{i=1}^{n} A_i \left(\bigcap_{j \neq i} \bar{A}_j \right) \right) = \bigcup_{i=1}^{n} A_i.$$

Hence $B_n \subset B_{n+1}$ for all $n \geq 1$ and $\lim_{n \to \infty} B_n = \bigcup_{n=1}^{\infty} A_n$.

1.2.2 The sample space $\mathcal{S} = \mathbb{Z}$, the set of all integers. A is a symmetric set in $\mathcal{S}$, if $A = -A$. Let $\mathcal{A} = \{\text{collection of all symmetric sets}\}$. $\phi \in \mathcal{A}$. If $A \in \mathcal{A}$ then $\bar{A} \in \mathcal{A}$. Indeed $-\bar{A} = -\mathcal{S} - (-A) = \mathcal{S} - A = \bar{A}$. Thus, $\bar{A} \in \mathcal{A}$. Moreover, if $A, B \in \mathcal{A}$ then $A \cup B \in \mathcal{A}$. Thus, $\mathcal{A}$ is an algebra.

1.2.3 $\mathcal{S} = \mathbb{Z}$. Let $\mathcal{A}_1 = \{\text{generated by symmetric sets}\}$. $\mathcal{A}_2 = \{\text{generated by } (-2, -1, i_1, \ldots, i_n),\ n \geq 1,\ i_j \in \mathbb{N}\ \forall j = 1, \ldots, n\}$. Notice that if $A = (-2, -1, i_1, \ldots, i_n)$ then $\bar{A} = \{(\cdots, -4, -3, \mathbb{N} - (i_1, \ldots, i_n))\} \in \mathcal{A}_2$, and $\mathcal{S} = A \cup \bar{A}_1 \in A_2$. $\mathcal{A}_2$ is an algebra. $\mathcal{A}_3 = \mathcal{A}_1 \cap \mathcal{A}_2$. If $B \in \mathcal{A}_3$ it must be symmetric and also $B \in \mathcal{A}_2$. Thus, $B = (-2, -1, 1, 2)$ or $B = (\cdots, -4, -3, 3, 4, \ldots)$. Thus, B and $\bar{B}$ are in $\mathcal{A}_3$, so $\mathcal{S} = (B \cup \bar{B}) \in \mathcal{A}_3$ and so is ϕ. Thus, $\mathcal{A}$ is an algebra.

Let $\mathcal{A}_4 = \mathcal{A}_1 \cup \mathcal{A}_2$. Let $A = \{-2, -1, 3, 7\}$ and $B = \{-3, 3\}$. Then $A \cup B = \{-3, -2, -1, 3; 7\}$. But $A \cup B$ does not belong to $\mathcal{A}_1$ neither to $\mathcal{A}_2$. Thus $A \cup B \notin \mathcal{A}_4$. $\mathcal{A}_4$ is not an algebra.

1.3.5 The sample space is

$$\mathcal{S} = \{(i_1, \ldots, i_{n-1}, 1) : \sum_{j=1}^{n-1} i_j = 1, \quad n \geq 2\}.$$

(i) Let $E_n = \left\{(i_1, \ldots, i_{n-1}, 1) : \sum_{j=1}^{n-1} i_j = 1\right\}$, $n = 2, 3, \ldots$. For $j \neq k$, $E_j \cap E_k = \emptyset$. Also $\bigcup_{n=l}^{\infty} E_n = \mathcal{S}$. Thus, $\mathcal{D} = \{E_2, E_3, \ldots\}$ is a countable partition of $\mathcal{S}$.

(ii) All elementary events $w_n = (i_1, \ldots, i_{n-1}, 1) \in E_n$ are equally probable and $P\{w_n\} = p^2 q^{n-2}$. There are $\binom{n-1}{1} = n - 1$ such elementary events in E_n. Thus, $P\{E_n\} = (n-1)p^2 q^{n-2}$. Moreover,

$$\begin{aligned}\sum_{n=2}^{\infty} P\{E_n\} &= p^2 \sum_{n=2}^{\infty} (n-1) q^{n-2} \\ &= p^2 \sum_{l=1}^{\infty} l q^{k-1} = 1.\end{aligned}$$

Indeed,

$$\begin{aligned}\sum_{l=1}^{\infty} l q^{l-1} &= \sum_{l=1}^{\infty} \frac{d}{dq} q^l \\ &= \frac{d}{dq}\left(\frac{q}{1-q}\right) \\ &= \frac{1}{(1-q)^2} = \frac{1}{p^2}.\end{aligned}$$

(iii) The probability that the experiment requires at least 5 trials is the probability that in the first 4 trials there is at most 1 success, which is $1 - p^2(1 + 2q + 3q^2)$.

1.4.6 Let X_n denote the position of the particle after n steps.

(i) If $n = 2k$, the particle after n steps could be, on the positive side only on even integers $2, 4, 6, \ldots, 2k$. If $n = 2k + 1$, the particle could be after n steps on the positive side only on an odd integer $1, 3, 5, \ldots, 2k + 1$.

Let p be the probability of step to the right ($0 < p < 1$) and $q = 1 - p$ of step to the left. If $n = 2k + 1$,

$$P\{X_n = 2j + 1\} = \binom{2j}{j} p^{2k+1-j} q^j, \quad j = 0, \ldots, k.$$

Thus, if $n = 2k + 1$,

$$P\{X_n > 0\} = \sum_{j=0}^{k} \binom{2j}{j} p^{2k+1-j} q^j.$$

In this solution, we assumed that all steps are independent (see Section 1.7). If $n = 2k$ the formula can be obtained in a similar manner.

(ii) $P\{X_7 = 1\} = \binom{6}{3} p^4 q^3$. If $p = \dfrac{1}{2}$, then $P\{X_7 = 1\} = \dfrac{\binom{6}{3}}{2^7} = 0.15625$.

(iii) The probability of returning to the origin after n steps is

$$P\{X_n = 0\} = \begin{cases} 0, & \text{if } n = 2k + 1 \\ \binom{2k}{k} p^k q^k, & \text{if } n = 2k. \end{cases}$$

Let $A_n = \{X_n = 0\}$. Then, $\sum_{k=0}^{\infty} P\{A_{2k+1}\} = 0$ and when $p = \dfrac{1}{2}$,

$$\sum_{k=0}^{\infty} P\{A_{2k}\} = \sum_{k=0}^{\infty} \binom{2k}{k} \frac{1}{2^{2k}} = \sum_{k=0}^{\infty} \frac{(2k)!}{(k!)^2} \cdot \frac{1}{4^k} = \infty.$$

Thus, by the Borel–Cantelli Lemma, if $p = \dfrac{1}{2}$, $P\{A_n i.o.\} = 1$. On the other hand, if $p \neq \dfrac{1}{2}$,

$$\sum_{k=0}^{\infty} \binom{2k}{k} (pq)^k = \frac{4pq}{\sqrt{1 - 4pq}\,(1 + \sqrt{1 - 4pq})} < \infty.$$

Thus, if $p \neq \dfrac{1}{2}$, $P\{A_n i.o.\} = 0$.

1.5.1 $F(x)$ is a discrete distribution with jump points at $-\infty < \xi_1 < \xi_2 < \cdots < \infty$. $p_i = d(F\xi_i)$, $i = 1, 2, \ldots$. $U(x) = I(x \geq 0)$.

(i)
$$U(x-\xi_i) = I(x \geq \xi_i)$$
$$F(x) = \sum_{\xi_i \leq x} p_i = \sum_{i=1}^{\infty} p_i U(x-\xi_i).$$

(ii) For $h > 0$,
$$D_h U(x) = \frac{1}{h}[U(x+h) - U(x)].$$
$U(x+h) = 1$ if $x \geq -h$. Thus,
$$D_h U(x) = \frac{1}{h} I(-h \leq x < 0)$$
$$\int_{-\infty}^{\infty} \sum_{i=1}^{\infty} p_i D_h U(x-\xi_i)dx = \sum_{i=1}^{\infty} p_i \frac{1}{h} \int_{-h+x-\xi_i}^{x-\xi_i} du = \sum_{i=1}^{\infty} p_i = 1.$$

(iii)
$$\lim_{h\to 0} \int_{-\infty}^{\infty} \sum_{i=1}^{\infty} p_i g(x) D_h U(x-\xi_i)dx = \sum_{i=1}^{\infty} p_i \lim_{h\downarrow 0} \int_{\xi_i - h}^{\xi_i} \frac{g(x)}{h} dx$$
$$= \sum_{i=1}^{\infty} p_i \lim_{h\downarrow 0} \frac{G(\xi_i) - G(\xi_i - h)}{h} = \sum_{i=1}^{\infty} p_i g(\xi_i)$$

Here, $G(\xi_i) = \int_{-\infty}^{\xi_i} g(x)dx$; $\frac{d}{d\xi_i} G(\xi_i) = g(\xi_i)$.

1.5.6 The joint p.d.f. of two discrete random variables is
$$f_{X,Y}(j,n) = \frac{e^{-\lambda}}{j!(n-j)!} \left(\frac{p}{1-p}\right)^j (\lambda(1-p))^n, \; j = 0, \ldots, n \; n = 0, 1, \ldots.$$

(i) The marginal distribution of X is
$$f_X(j) = \sum_{n=j}^{\infty} f_{X,Y}(j,n)$$
$$= \frac{p^j(\lambda(1-p))^j}{(1-p)^j j!} e^{-\lambda} \sum_{n=j}^{\infty} \frac{(\lambda(1-p))^{n-j}}{(n-j)!}$$
$$= e^{-\lambda p} \frac{(\lambda p)^j}{j!} e^{-\lambda(1-p)} \sum_{n=0}^{\infty} \frac{(\lambda(1-p))^n}{n!}$$
$$= e^{-\lambda p} \frac{(\lambda p)^j}{j!}, \quad j = 0, 1, 2, \ldots.$$

(ii) The marginal p.d.f. of Y is

$$\begin{aligned} f_Y(n) &= \sum_{j=0}^{n} p_{X,Y}(j,n) \\ &= e^{-\lambda}\frac{\lambda^n}{n!}\sum_{j=0}^{n}\binom{n}{j}p^j(1-p)^{n-j} \\ &= e^{-\lambda}\frac{\lambda^n}{n!}, \quad n = 0, 1, \ldots. \end{aligned}$$

(iii) $$p_{X|Y}(j \mid n) = \frac{f_{X,Y}(j,n)}{f_Y(n)} = \binom{n}{j}p^j(1-p)^{n-j}, \quad j = 0, \ldots, n.$$

(iv) $$\begin{aligned} p_{Y|X}(n \mid j) &= \frac{f_{X,Y}(j,n)}{f_X(j)} \\ &= e^{-\lambda(1-p)}\frac{(\lambda(1-p))^{n-j}}{(n-j)!}, \quad n \geq j. \end{aligned}$$

(v) $$E(Y \mid X = j) = j + \lambda(1-p).$$

(vi) $$\begin{aligned} E\{Y\} = E\{E\{Y \mid X\}\} &= \lambda(1-p) + E\{X\} \\ &= \lambda(1-p) + \lambda p = \lambda. \end{aligned}$$

1.5.8

$$F_j(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1 - P(j-1;x), & \text{if } x \geq 0, \end{cases}$$

where $j \geq 1$, and $P(j-1;x) = e^{-x}\sum_{i=0}^{j-1}\frac{x^i}{x!}$.

(i) We have to show that, for each $j \geq 1$, $F_j(x)$ is a c.d.f.

(i) $0 \leq F_j(x) \leq 1$ for all $0 \leq x < \infty$.

(ii) $F_j(0) = 0$ and $\lim_{x\to\infty} F_j(x) = 1$.

(iii) We show now that $F_j(x)$ is strictly increasing in x. Indeed, for all $x > 0$

$$\begin{aligned} \frac{d}{dx}F_j(x) &= -\sum_{i=0}^{j-1}\frac{d}{dx}\left(e^{-x}\frac{x^i}{i!}\right) \\ &= e^{-x} + \sum_{i=1}^{j-1}\left(e^{-x}\frac{x^i}{i!} - e^{-x}\frac{x^{i-1}}{(i-1)!}\right) \\ &= e^{-x}\frac{x^{j-1}}{(j-1)!} > 0, \quad \text{for all } 0 < x < \infty. \end{aligned}$$

(ii) The density of $F_j(x)$ is

$$f_j(x) = \frac{x^{j-1}}{(j-1)!}e^{-x}, \quad j \geq 1, \quad x \geq 0.$$

$F_j(x)$ is absolutely continuous.

(iii)
$$E_j\{X\} = \int_0^\infty \frac{x^j}{(j-1)!}e^{-x}dx$$
$$= j\int_0^\infty \frac{x^j}{j!}e^{-x}dx = j.$$

1.6.3 X, Y are independent and identically distributed, $E|X| < \infty$.

$$E\{X \mid X+Y\} + E\{Y \mid X+Y\} = X+Y$$
$$E\{X \mid X+Y\} = E\{Y \mid X+Y\} = \frac{X+Y}{2}.$$

1.6.9
$$F(x) = \begin{cases} 0, & x < 0 \\ 1-e^{-x}, & x \geq 0. \end{cases}$$

(i) Let $A_n = \{X_n > 1\}$. The events A_n, $n \geq 1$, are independent. Also $P\{A_n\} = e^{-1}$. Hence, $\sum_{n=1}^{\infty} P\{A_n\} = \infty$. Thus, by the Borel–Cantelli Lemma, $P\{A_n i.o.\} = 1$. That is, $P\left(\lim_{n\to\infty} S_n = \infty\right) = 1$.

(ii) $\frac{S_n}{1+S_n} \geq 0$. This random variable is bounded by 1. Thus, by the Dominated Convergence Theorem, $\lim_{n\to\infty} E\left\{\frac{S_n}{1+S_n}\right\} = E\left\{\lim_{n\to\infty}\frac{S_n}{1+S_n}\right\} = 1$.

1.8.4
$$f_X(x) = \begin{cases} 0, & \text{if } x < 0 \\ \frac{\lambda^m}{(m-1)!}x^{m-1}e^{-\lambda x}, & x \geq 0; m \geq 2. \end{cases}$$

(i) The m.g.f. of X is

$$M(t) = \frac{\lambda^m}{(m-1)!}\int_0^\infty e^{-x(\lambda-t)}x^{m-1}dx$$
$$= \frac{\lambda^m}{(\lambda-t)^m} = \left(1-\frac{t}{\lambda}\right)^{-m}, \quad \text{for } t < \lambda.$$

The domain of convergence is $(-\infty, \lambda)$.

(ii)
$$M'(t) = \frac{m}{\lambda}\left(1 - \frac{t}{\lambda}\right)^{-m-1}$$
$$M''(t) = \frac{m(m+1)}{\lambda^2}\left(1 - \frac{t}{\lambda}\right)^{-(m+2)}$$
$$\vdots$$
$$M^{(r)}(t) = \frac{m(m+1)\cdots(m+r-1)}{\lambda^r}\left(1 - \frac{t}{\lambda}\right)^{-(m+r)}.$$

Thus, $\mu_r = M^{(r)}(t)\big|_{t=0} = \dfrac{m(m+1)\cdots(m+r-1)}{\lambda^r}$ $r \geq 1$.

(iii)
$$\mu_1 = \frac{m}{\lambda}$$
$$\mu_2 = \frac{m(m+1)}{\lambda^2}$$
$$\mu_3 = \frac{m(m+1)(m+2)}{\lambda^3}$$
$$\mu_4 = \frac{m(m+1)(m+2)(m+3)}{\lambda^4}.$$

The central moments are

$$\mu_1^* = 0$$
$$\mu_2^* = \mu_2 - \mu_1^2 = \frac{m}{\lambda^2}$$
$$\mu_3^* = \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3$$
$$= \frac{1}{\lambda^3}(m(m+1)(m+2) - 3m^2(m+1) + 2m^3)$$
$$= \frac{2m}{\lambda^3}$$
$$\mu_4^* = \mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^4$$
$$= \frac{1}{\lambda^4}(m(m+1)(m+2)(m+3) - 4m^2(m+1)(m+2)$$
$$+ 6m^3(m+1) - 3m^4)$$
$$= \frac{3m(m+2)}{\lambda^4}.$$

(iv)
$$\beta_1 = \frac{2m}{m^{3/2}} = \frac{2}{\sqrt{m}}; \quad \beta_2 = \frac{3m(m+2)}{m^2} = 3 + \frac{6}{m}.$$

1.8.11 The m.g.f. is

$$\begin{aligned} M_X(t) &= \frac{1}{a^2}\int_{-a}^{a} e^{tx}(a-|x|)dx \\ &= \frac{2(\cosh(at)-1)}{a^2t^2} \\ &= 1+\frac{1}{12}(at)^2+o(t), \quad \text{as } t\to 0. \end{aligned}$$

1.9.1

$$F_n(x)=\begin{cases} 0, & x<0 \\ \frac{j}{n}, & \frac{j}{n}\le x<\frac{j+1}{n},\ j=0,\ldots,n-1 \\ 1, & 1\le x. \end{cases}$$

$$F(x)=\begin{cases} 0, & x<0 \\ x, & 0<x<1 \\ 1, & 1\le x. \end{cases}$$

All points $-\infty < x < \infty$ are continuity points of $F(x)$. $\lim_{n\to\infty} F_n(x) = F(x)$, for all $x<0$ or $x>1$. $|F_n(x)-F(x)| \le \frac{1}{n}$ for all $0\le x\le 1$. Thus $F_n(x) \xrightarrow{w} F(x)$, as $n\to\infty$.

1.9.4

$$X_n=\begin{cases} 0, & \text{w.p. } \left(1-\frac{1}{n}\right) \\ 1, & \text{w.p. } \frac{1}{n} \end{cases}, n\ge 1.$$

(i) $E\{|X_n|^r\} = \frac{1}{n}1 = \frac{1}{n}$ for all $r>0$. Thus, $X_n \xrightarrow[n\to\infty]{r} 0$, for all $r>0$.

(ii) $P\{|X_n|>\epsilon\} = \frac{1}{n}$ for all $n\ge 1$, any $\epsilon>0$. Thus, $X_n \xrightarrow[n\to\infty]{p} 0$.

(iii) Let $A_n=\{w: X_n(w)=1\}$; $P\{A_n\}=\frac{1}{n}$, $n\ge 0$. $\sum_{n=1}^{\infty}\frac{1}{n}=\infty$. Since $X_1, X_2, \ldots$ are independent, by Borel–Cantelli's Lemma, $P\{X_n = 1, i.o.\} = 1$. Thus $X_n \not\to 0$ a.s.

1.9.5 $\epsilon_1, \epsilon_2, \cdots$ independent r.v.s, such that $E(\epsilon_n)=\mu$, and $V\{\epsilon_n\}=\sigma^2$. $\forall n\ge 1$.

$$\begin{aligned} X_1 &= \epsilon_1, \\ X_n &= \beta X_{n-1}+\epsilon_n = \beta(\beta X_{n-2}+\epsilon_{n-1})+\epsilon_n \\ &= \cdots = \sum_{j=1}^{n}\beta^{n-j}\epsilon_j, \quad \forall n\ge 1, \quad |\beta|<1. \end{aligned}$$

Thus, $E\{X_n\} = \mu \sum_{j=0}^{n-1} \beta^j \xrightarrow[n\to\infty]{} \frac{\mu}{1-\beta}$.

$$\begin{aligned}\bar{X}_n &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{i}\beta^{i-j}\epsilon_j \\ &= \frac{1}{n}\sum_{j=1}^{n}\epsilon_j \sum_{i=j}^{n}\beta^{i-j} \\ &= \frac{1}{n}\sum_{j=1}^{n}\epsilon_j \frac{1-\beta^{n-j+1}}{1-\beta}.\end{aligned}$$

Since $\{\epsilon_n\}$ are independent,

$$\begin{aligned}V\{\bar{X}_n\} &= \frac{\sigma^2}{n^2(1-\beta)^2}\sum_{j=1}^{n}(1-\beta^{n-j+1})^2 \\ &= \frac{\sigma^2}{n(1-\beta)^2}\left(1 - 2\frac{\beta(1-\beta^{n+1})}{n(1-\beta)} + \frac{\beta^2(1-\beta^{2n+1})}{n(1-\beta^2)}\right) \to 0 \text{ as } n \to \infty.\end{aligned}$$

Furthermore,

$$\begin{aligned}E\left\{\left(\bar{X}_n - \frac{\mu}{1-\beta}\right)^2\right\} &= V\{\bar{X}_n\} + \left(E\{\bar{X}_n\} - \frac{\mu}{1-\beta}\right)^2 \\ \left(E\{\bar{X}_n\} - \frac{\mu}{1-\beta}\right)^2 &= \frac{\mu^2}{n^2(1-\beta)^2}(1-\beta^{n+1})^2 \to 0 \text{ as } n \to \infty.\end{aligned}$$

Hence, $\bar{X}_n \xrightarrow{2} \frac{\mu}{1-\beta}$.

1.9.7 $X_1, X_2, \ldots$ i.i.d. distributed like $R(0,\theta)$. $X_n = \max_{1\le i\le n}(X_i)$. Due to independence,

$$F_n(x) = P\{X_{(n)} \le x\} = \begin{cases} 0, & x < 0 \\ \left(\frac{x}{\theta}\right)^n, & 0 \le x < \theta \\ 1, & \theta \le x. \end{cases}$$

Accordingly, $P\{X_{(n)} < \theta - \epsilon\} = \left(1 - \frac{\epsilon}{\theta}\right)^n, \quad 0 < \epsilon < \theta.$ Thus, $\sum_{n=1}^{\infty} P\{X_{(n)} \le \theta - \epsilon\} < \infty,$ and $P\{X_{(n)} \le \theta - \epsilon, i.o.\} = 0.$ Hence, $X_{(n)} \to \theta$ a.s.

1.10.2 We are given that $\mathbf{a}'\mathbf{X}_n \xrightarrow{d} \mathbf{a}'X$ for all $\mathbf{a}$. Consider the m.g.f.s, by continuity theorem $M_{\mathbf{a}'\mathbf{X}_n}(t) = E\{e^{t\mathbf{a}'\mathbf{X}_n}\} \to E\{e^{ta=\mathbf{X}}\}$, for all t in the domain of convergence. Thus $E\{e^{(t\mathbf{a})'\mathbf{X}_n}\} \to E\{e^{(t\mathbf{a})'\mathbf{X}}\}$ for all $\boldsymbol{\beta} = t\mathbf{a}$. Thus, $\bar{X}_n \xrightarrow{d} \mathbf{X}$.

1.10.6 $X_n \sim B\left(n, \frac{1}{n}\right)$

$$E\{e^{-X_n}\} = \left(\frac{1}{n}e^{-1} + 1 - \frac{1}{n}\right)^n = \left(1 - \frac{1}{n}(1 - e^{-1})\right)^n \xrightarrow[n\to\infty]{} e^{-(1-e^{-1})}.$$

Thus, $\lim_{n\to\infty} M_{X_n}(-1) = M_x(-1)$, where $X \sim P(1)$.

1.11.1 **(i)**

$$M_{\bar{X}_n}(t) = \left(M_X\left(\frac{t}{n}\right)\right)^n = \left(E\left\{e^{\frac{t}{n}X}\right\}\right)^n = \left(1 + \frac{t}{n}E\{X\} + o\left(\frac{1}{n}\right)\right)^n = \left(1 + \frac{t}{n}\mu + o\left(\frac{1}{n}\right)\right)^n \xrightarrow[n\to\infty]{} e^{t\mu}, \quad \forall t.$$

$e^{t\mu}$ is the m.g.f. of the distribution

$$F(x) = \begin{cases} 0, & x < \mu \\ 1, & x \ge \mu. \end{cases}$$

Thus, by the continuity theorem, $\bar{X}_n \xrightarrow{d} \mu$ and, therefore, $\bar{X}_n \xrightarrow{p} \mu$, as $n \to \infty$.

1.11.5 $\{X_n\}$ are independent. For $\delta > 0$,

$$X_n \sim R(-n, n)/n.$$

The expected values are $E\{X_n\} = 0\ \forall n \geq 1$.

$$\sigma_n^2 = \frac{4n^2}{12n^2} = \frac{1}{3}$$
$$\sum_{n=1}^{\infty} \frac{\sigma_n^2}{n^2} < \infty.$$

Hence, by (1.11.6), $\bar{X}_n \xrightarrow{\text{a.s.}} 0$.

1.12.1
$$M_{\frac{X-\lambda}{\sqrt{\lambda}}} = E\left\{e^{t(\frac{X-\lambda}{\sqrt{\lambda}})}\right\} = e^{-\sqrt{\lambda}t-\lambda(1-e^{t/\sqrt{\lambda}})}.$$

$$1 - \exp\{t/\sqrt{\lambda}\} = 1 - \left(1 + \frac{t}{\sqrt{\lambda}} + \frac{t^2}{2\lambda} + \cdots\right)$$
$$= -\frac{t}{\sqrt{\lambda}} - \frac{t^2}{2\lambda} + \cdots.$$

Thus,

$$\sqrt{\lambda}t - \lambda(1 - e^{t\sqrt{\lambda}}) = \frac{t^2}{2} + O\left(\frac{1}{\sqrt{\lambda}}\right).$$

Hence,

$$M_{\frac{X-\lambda}{\sqrt{\lambda}}}(t) = \exp\left\{\frac{t^2}{2} + O\left(\frac{1}{\sqrt{\lambda}}\right)\right\} \to e^{t^2/2} \quad \text{as } \lambda \to \infty.$$

$M_Z(t) = e^{t^2/2}$ is the m.g.f. of $N(0, 1)$.

1.12.3
$$P(X_n = 1) = \frac{1}{2},$$
$$P(X_n = 0) = \frac{1}{2},$$
$$E\{X_n\} = \frac{1}{2}.$$

Let $Y_n = nX_n$; $E\{Y_n\} = \frac{n}{2}$, $E|Y_n|^3 = \frac{n^3}{2}$. Notice that $\sum_{i=1}^{n} i X_i - \frac{n(n+1)}{4} = \sum_{i=1}^{n}(Y_i - \mu_i)$, where $\mu_i = \frac{i}{2} = E\{Y_i\}$. $E|Y_i - \mu_i|^3 = \frac{i^3}{8}$. Accordingly,

$$\frac{\sum_{i=1}^{n} E\{|Y_i - \mu_i|^3\}}{\sum_{i=1}^{n} E\{(Y_i - \mu_i)^2\}^{3/2}} = \frac{\frac{1}{4}n^2(n+1)^2}{\left(\frac{1}{24}n(n+1)(2n+1)\right)^{3/2}} \to 0 \quad \text{as } n \to \infty.$$

Thus, by Lyapunov's Theorem,

$$\frac{\sum_{i=1}^{n} i X_i - \frac{n(n+1)}{4}}{\left(\frac{1}{24}n(n+1)(2n+1)\right)^{1/2}} \xrightarrow{d} N(0,1) \quad \text{as } n \to \infty.$$

1.13.5 $\{X_n\}$ i.i.d. $B(1, p)$, $0 < p < 1$.

$$\hat{p}_n = \frac{1}{n}\sum X_i$$

$$W_n = \log \frac{\hat{p}_n}{(1 - \hat{p}_n)}.$$

(i)
$$E\{W_n\} \cong \log \frac{p}{1-p} + \frac{1}{2n} p(1-p) W''(p)$$

$$W'(p) = \frac{(1-p)(1-p+p)}{p(1-p)^2} = \frac{1}{p(1-p)}$$

$$W''(p) = -\frac{(1-2p)}{p^2(1-p)^2}.$$

Thus,

$$E\{W_n\} \cong log\left(\frac{p}{1-p}\right) - \frac{1-2p}{2np(1-p)}.$$

(ii)
$$V\{W_n\} \cong \frac{p(1-p)}{n} \cdot \frac{1}{(p(1-p))^2}$$

$$= \frac{1}{np(1-p)}.$$

CHAPTER 2

Statistical Distributions

PART I: THEORY

2.1 INTRODUCTORY REMARKS

This chapter presents a systematic discussion of families of distribution functions, which are widely used in statistical modeling. We discuss univariate and multivariate distributions. A good part of the chapter is devoted to the distributions of sample statistics.

2.2 FAMILIES OF DISCRETE DISTRIBUTIONS

2.2.1 Binomial Distributions

Binomial distributions correspond to random variables that count the number of successes among N independent trials having the same probability of success. Such trials are called **Bernoulli trials**. The probabilistic model of Bernoulli trials is applicable in many situations, where it is reasonable to assume independence and constant success probability.

Binomial distributions have two parameters N (number of trials) and θ (success probability), where N is a positive integer and $0 < \theta < 1$. The probability distribution function is denoted by $b(i; N, \theta)$ and is

$$b(i; N, \theta) = \binom{N}{i} \theta^i (1-\theta)^{N-i}, \quad i = 0, 1, \ldots, N. \tag{2.2.1}$$

Examples and Problems in Mathematical Statistics, First Edition. Shelemyahu Zacks.

The c.d.f. is designated by $B(i; N, \theta)$, and is equal to $B(i; N, \theta) = \sum_{j=0}^{i} b(j; N, \theta)$. The Binomial distribution formula can also be expressed in terms of the **incomplete beta function** by

$$\sum_{j=a}^{N} b(j; N, \theta) = I_\theta(a, N - a + 1), \quad a = 1, \ldots, N \tag{2.2.2}$$

where

$$I_\xi(p, q) = \frac{1}{B(p, q)} \int_0^\xi u^{p-1}(1-u)^{q-1} du, \quad 0 \le \xi \le 1. \tag{2.2.3}$$

The parameters p and q are positive, i.e., $0 < p, q < \infty$; $B(p, q) = \int_0^1 u^{p-1}(1-u)^{q-1} du$ is the (complete) **beta function**. Or

$$B(i; N, \theta) = 1 - I_\theta(i + 1, N - i) = I_{1-\theta}(N - i, i + 1), \quad i = 0, \ldots, N - 1. \tag{2.2.4}$$

The quantiles $B^{-1}(p; N, \theta)$, $0 < p < 1$, can be easily determined by finding the smallest value of i at which $B(i; N, \theta) \ge p$.

2.2.2 Hypergeometric Distributions

The hypergeometric distributions are applicable when we sample at random **without replacement** from a finite population (collection) of N units, so that every possible sample of size n has equal selection probability, $1/\binom{N}{n}$. If X denotes the number of units in the sample having a certain attribute, and if M is the number of units in the population (before sampling) having the same attribute, then the distribution of X is **hypergeometric** with the probability density function (p.d.f.)

$$h(i; N, M, n) = \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}, \quad i = 0, \ldots, n. \tag{2.2.5}$$

The c.d.f. of the hypergeometric distribution will be denoted by $H(i; N, M, n)$. When n/N is sufficiently small (smaller than 0.1 for most practical applications), we can approximate $H(i; N, M, n)$ by $B(i; n, M/N)$. Better approximations (Johnson and Kotz, 1969, p. 148) are available, as well, as bounds on the error terms.

2.2.3 Poisson Distributions

Poisson distributions are applied when the random variables under consideration count the number of events occurring in a specified time period, or on a spatial area, and the observed processes satisfy the basic conditions of time (or space) homogeneity, independent increments, and no memory of the past (Feller, 1966, p. 566). The Poisson distribution is prevalent in numerous applications of statistics to engineering reliability, traffic flow, queuing and inventory theories, computer design, ecology, etc.

A random variable X is said to have a Poisson distribution with intensity λ, $0 < \lambda < \infty$, if it assumes only the nonnegative integers according to a probability distribution function

$$p(i;\lambda) = e^{-\lambda}\frac{\lambda^i}{i!}, \quad i = 0, 1 \ldots . \tag{2.2.6}$$

The c.d.f. of such a distribution is denoted by $P(i;\lambda)$.

The Poisson distribution can be obtained from the Binomial distribution by letting $N \to \infty$, $\theta \to 0$ so that $N\theta \to \lambda$, where $0 < \lambda < \infty$ (Feller, 1966, p. 153, or Problem 5 of Section 1.10). For this reason, the Poisson distribution can provide a good model in cases of counting events that occur very rarely (the number of cases of a rare disease per 100,000 in the population; the number of misprints per page in a book, etc.).

The Poisson c.d.f. can be determined from the incomplete gamma function according to the following formula

$$P(k;\lambda) = \frac{1}{\Gamma(k+1)}\int_{\lambda}^{\infty} x^k e^{-x}dx, \tag{2.2.7}$$

for all $k = 0, 1, \ldots$, where

$$\Gamma(p) = \int_0^{\infty} x^{p-1}e^{-x}dx, \quad p > 0 \tag{2.2.8}$$

is the gamma function.

2.2.4 Geometric, Pascal, and Negative Binomial Distributions

The geometric distribution is the distribution of the number of Bernoulli trials until the first success. This distribution has therefore many applications (the number of shots at a target until the first hit). The probability distribution function of a geometric random variable is

$$g(i;\theta) = \theta(1-\theta)^{i-1}, \quad i = 1, 2, \ldots \tag{2.2.9}$$

where θ, $0 < \theta < 1$, is the probability of success.

If the random variable counts the number of Bernoulli trials until the ν-th success, $\nu = 1, 2, \ldots$, we obtain the **Pascal distribution** with p.d.f.

$$g(i; \theta, \nu) = \binom{i-1}{\nu-1} \theta^{\nu} (1-\theta)^{i-\nu}, \quad i = \nu, \nu+1, \ldots. \tag{2.2.10}$$

The geometric distributions constitute a subfamily with $\nu = 1$. Another family of distributions of this type is that of the **Negative-Binomial** distributions. We designate by $NB(\psi, \nu), 0 < \psi < 1, 0 < \nu < \infty$, a random variable having a Negative-Binomial distribution if its p.d.f. is

$$nb(i; \psi, \nu) = \frac{\Gamma(\nu + i)}{\Gamma(i+1)\Gamma(\nu)} (1-\psi)^{\nu} \psi^{i}, \quad i = 0, 1, \ldots. \tag{2.2.11}$$

Notice that if X has the Pascal distribution with parameters ν and θ, then $X - \nu$ is distributed like $NB(1-\theta, \nu)$. The probability distribution of Negative-Binomial random variables assigns positive probabilities to all the nonnegative integers. It can therefore be applied as a model in cases of counting random variables where the Poisson assumptions are invalid. Moreover, as we show later, Negative-Binomial distributions may be obtained as averages of Poisson distributions. The family of Negative-Binomial distributions depend on two parameters and can therefore be fitted to a variety of empirical distributions better than the Poisson distributions. Examples of this nature can be found in logistics research in studies of population growth with immigration, etc.

The c.d.f. of the $NB(\psi, \nu)$, to be designated as $NB(i; \psi, \nu)$, can be determined by the incomplete beta function according to the formula

$$NB(k; \psi, \nu) = I_{1-\psi}(\nu, k+1), \quad k = 0, 1 \ldots. \tag{2.2.12}$$

A proof of this useful relationship is given in Example 2.3.

2.3 SOME FAMILIES OF CONTINUOUS DISTRIBUTIONS

2.3.1 Rectangular Distributions

A random variable X has a rectangular distribution over the interval (θ_1, θ_2), $-\infty < \theta_1 < \theta_2 < \infty$, if its p.d.f. is

$$f_R(x; \theta_1, \theta_2) = \begin{cases} \dfrac{1}{\theta_2 - \theta_1}, & \text{if } \theta_1 < x < \theta_2 \\ 0, & \text{otherwise.} \end{cases} \tag{2.3.1}$$

The family of all rectangular distributions is a two-parameter family. We denote r.v.s having these distributions by $R(\theta_1, \theta_2)$; $-\infty < \theta_1 < \theta_2 < \infty$. We note that if X is

distributed as $R(\theta_1, \theta_2)$, then X is equivalent to $\theta_1 + (\theta_2 - \theta_1)U$, where $U \sim R(0, 1)$. This can be easily verified by considering the distribution functions of $R(\theta_1, \theta_2)$ and of $R(0, 1)$, respectively. Accordingly, the parameter $\alpha = \theta_1$ can be considered a **location** parameter and $\beta = \theta_2 - \theta_1$ is a **scale** parameter. Let $f_U(x) = I\{0 \le x \le 1\}$ be the p.d.f. of the standard rectangular r.v. U. Thus, we can express the p.d.f. of $R(\theta_1, \theta_2)$ by the general presentation of p.d.f.s in the **location and scale parameter models**; namely

$$f_R(x; \theta_1, \theta_2) = \frac{1}{\theta_2 - \theta_1} f_U\left(\frac{x - \theta_1}{\theta_2 - \theta_1}\right), \quad -\infty \le x \le \infty. \tag{2.3.2}$$

The standard rectangular distribution function occupies an important place in the theory of statistics. One of the reasons is that if a random variable has an arbitrary **continuous** distribution function $F(x)$, then the transformed random variable $Y = F(X)$ is distributed as U. For each ξ, $0 < \xi < 1$, let

$$F_U^{-1}(\xi) = \inf\{x : F(x) = \xi\} = \xi. \tag{2.3.3}$$

Accordingly, since $F(x)$ is nondecreasing and continuous,

$$P\{F(X) \le \xi\} = P\{X \le F^{-1}(\xi)\} = F(F^{-1}(\xi)) = \xi. \tag{2.3.4}$$

The transformation $X \to F(X)$ is called the **Cumulative Probability Integral Transformation**.

Notice that the pth quantile of $R(\theta_1, \theta_2)$ is

$$R_p(\theta_1, \theta_2) = \theta_1 + p(\theta_2 - \theta_1). \tag{2.3.5}$$

The following has application in the theory of testing hypotheses.

If X has a discrete distribution $F(x)$ and if we define the function

$$H(x, \gamma) = F(x - 0) + \gamma[F(x) - F(x - 0)], \tag{2.3.6}$$

where $-\infty < x < \infty$ and $0 \le \gamma \le 1$, then $H(X, U)$ has a rectangular distribution as $R(0, 1)$, where U is also distributed like $R(0, 1)$, independently of X. We notice that if x is a jump point of $F(x)$, then $H(x, \gamma)$ assumes a value in the interval $[F(x - 0), F(x)]$. On the other hand, if x is not a jump point, then $H(x, \gamma) = F(x)$ for all γ. Thus, for every p, $0 \le p \le 1$,

$$H(x, \gamma) \le p \text{ if and only if}$$
$$x < F^{-1}(p) \text{ or } x = F^{-1}(p) \text{ and } \gamma \le \gamma(p),$$

where

$$\gamma(p) = \frac{p - F(F^{-1}(p) - 0)}{F(F^{-1}(p)) - F(F^{-1}(p) - 0)}. \tag{2.3.7}$$

Accordingly, for every p, $0 \leq p \leq 1$,

$$\begin{aligned} P\{H(X, U) \leq p\} &= P\{X < F^{-1}(p)\} + P\{U \leq \gamma(p)\} P\{X = F^{-1}(p)\} \\ &= F(F^{-1}(p) - 0) + \gamma(p)[F(F^{-1}(p)) - F(F^{-1}(p) - 0)] = p. \end{aligned} \tag{2.3.8}$$

2.3.2 Beta Distributions

The family of Beta distributions is a two-parameter family of continuous distributions concentrated over the interval [0, 1]. We denote these distributions by $\beta(p, q)$; $0 < p$, $q < \infty$. The p.d.f. of a $\beta(p, q)$ distribution is

$$f(x; p, q) = \frac{1}{B(p, q)} x^{p-1}(1 - x)^{q-1}, \quad 0 \leq x \leq 1. \tag{2.3.9}$$

The $R(0, 1)$ distribution is a special case. The distribution function (c.d.f.) of $\beta(p, q)$ coincides over the interval $(0, 1)$ with the incomplete Beta function (2.3.2). Notice that

$$I_\xi(p, q) = 1 - I_{1-\xi}(q, p), \quad \text{for all } 0 \leq \xi \leq 1. \tag{2.3.10}$$

Hence, the Beta distribution is symmetric about $x = .5$ if and only if $p = q$.

2.3.3 Gamma Distributions

The Gamma function $\Gamma(p)$ was defined in (2.2.8). On the basis of this function we define a two-parameter family of distribution functions. We say that a random variable X has a Gamma distribution with positive parameters λ and p, to be denoted by $G(\lambda, p)$, if its p.d.f. is

$$f(x; \lambda, p) = \frac{\lambda^p}{\Gamma(p)} x^{p-1} e^{-\lambda x}, \quad 0 \leq x \leq \infty. \tag{2.3.11}$$

λ^{-1} is a **scale** parameter, and p is called a **shape** parameter. A special important case is that of $p = 1$. In this case, the density reduces to

$$f(x; \lambda) = \lambda e^{-\lambda x}, \quad 0 \leq x \leq \infty. \tag{2.3.12}$$

This distribution is called the (negative) **exponential distribution**. Exponentially distributed r.v.s with parameter λ are denoted also as $E(\lambda)$.

The following relationship between Gamma distributions explains the role of the scale parameter λ^{-1}

$$G(\lambda, p) \sim \frac{1}{\lambda} G(1, p), \quad \text{for all } \lambda. \tag{2.3.13}$$

Indeed, from the definition of the gamma p.d.f. the following relationship holds for all ξ, $0 \le \xi \le \infty$,

$$\begin{aligned} P\{G(\lambda, p) \le \xi\} &= \frac{\lambda^p}{\Gamma(p)} \int_0^{\xi} x^{p-1} e^{-\lambda x} dx \\ &= \frac{1}{\Gamma(p)} \int_0^{\lambda\xi} x^{p-1} e^{-x} dx = P\left\{\frac{1}{\lambda} G(1, p) \le \xi\right\}. \end{aligned} \tag{2.3.14}$$

In the case of $\lambda = \frac{1}{2}$ and $p = \nu/2$, $\nu = 1, 2, \ldots$ the Gamma distribution is also called chi-squared distribution with ν degrees of freedom. The chi-squared random variables are denoted by $\chi^2[\nu]$, i.e.,

$$\chi^2[\nu] \sim G\left(\frac{1}{2}, \frac{\nu}{2}\right); \quad \nu = 1, 2, \ldots. \tag{2.3.15}$$

The reason for designating a special name for this subfamily of Gamma distributions will be explained later.

2.3.4 Weibull and Extreme Value Distributions

The family of Weibull distributions has been extensively applied to the theory of systems reliability as a model for lifetime distributions (Zacks, 1992). It is also used in the theory of survival distributions with biological applications (Gross and Clark, 1975). We say that a random variable X has a Weibull distribution with parameters (λ, α, ξ); $0 < \lambda$, $0 < \alpha < \infty$; $-\infty < \xi < \infty$, if $(X - \xi)^{\alpha} \sim G(\lambda, 1)$. Accordingly, $(X - \xi)^{\alpha}$ has an exponential distribution with a scale parameter λ^{-1}, ξ is a location parameter, i.e., the p.d.f. assumes positive values only for $x \ge \xi$. We will assume here, without loss of generality, that $\xi = 0$. The parameter α is called the **shape parameter**. The p.d.f. of X, for $\xi = 0$ is

$$f_W(x; \lambda, \alpha) = \lambda\alpha x^{\alpha-1} \exp\{-\lambda x^{\alpha}\}, \quad 0 \le x \le \infty, \tag{2.3.16}$$

and its c.d.f. is

$$F_W(x; \lambda, \alpha) = \begin{cases} 1 - \exp\{-\lambda x^{\alpha}\}, & x \ge 0 \\ 0, & x < 0. \end{cases} \tag{2.3.17}$$

The **extreme value** distribution (of Type I) is obtained from the Weibull distribution if we consider the distribution of $Y = -\log X$, where $X^{\alpha} \sim G(\lambda, 1)$. Accordingly, the c.d.f. of Y is

$$P\{Y \leq \eta\} = \exp\{-\lambda e^{-\alpha\eta}\}, \tag{2.3.18}$$

$-\infty < \eta < \infty$, and its p.d.f. is

$$f_{EV}(x; \lambda, \alpha) = \lambda\alpha \exp\{-\alpha x - \lambda e^{-\alpha x}\}, \tag{2.3.19}$$

$-\infty < x < \infty$.

Extreme value distributions have been applied in problems of testing strength of materials, maximal water flow in rivers, biomedical problems, etc. (Gumbel, 1958).

2.3.5 Normal Distributions

The normal distribution occupies a central role in statistical theory. Many of the statistical tests and estimation procedures are based on statistics that have distributions approximately normal in a large sample.

The family of normal distributions, to be designated by $N(\xi, \sigma^2)$, depends on two parameters. A **location** parameter ξ, $-\infty < \xi < \infty$ and a **scale** parameter σ, $0 < \sigma < \infty$. The p.d.f. of a normal distribution is

$$f(x; \xi, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\xi}{\sigma}\right)^2\right\}, \tag{2.3.20}$$

$-\infty < x < \infty$.

The normal distribution with $\xi = 0$ and $\sigma = 1$ is called the **standard normal distribution**. The standard normal p.d.f. is denoted by $\phi(x)$. Notice that $N(\xi, \sigma^2) \sim \xi + \sigma N(0, 1)$. Indeed, since $\sigma > 0$,

$$\begin{aligned} P\{N(\xi, \sigma^2) \leq x\} &= \frac{1}{\sqrt{2\pi}\,\sigma} \int_{-\infty}^{x} \exp\left\{-\frac{1}{2}\left(\frac{y-\xi}{\sigma}\right)^2\right\} dy \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\xi}{\sigma}} \exp\left\{-\frac{1}{2}z^2\right\} dz = P\{\xi + \sigma N(0, 1) \leq x\}. \end{aligned} \tag{2.3.21}$$

According to (2.3.21), the c.d.f. of $N(\xi, \sigma^2)$ can be computed on the basis of the standard c.d.f. The standard c.d.f. is denoted by $\Phi(x)$. It is also called the standard normal integral. Efficient numerical techniques are available for the computation of $\Phi(x)$. The function and its derivatives are tabulated. Efficient numerical approximations and asymptotic expansions are given in Abramowitz and Stegun (1968, p. 925).

The normal p.d.f. is **symmetric** about the location parameter ξ. From this symmetry, we deduce that

$$\begin{aligned} \phi(x) &= \phi(-x), \quad \text{all } -\infty < x < \infty \\ \Phi(-x) &= 1 - \Phi(x), \quad \text{all } -\infty < x < \infty. \end{aligned} \tag{2.3.22}$$

By a series expansion of $e^{-t^2/2}$ and direct integration, one can immediately derive the formula

$$\Phi(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \sum_{j=0}^{\infty} \frac{(-1)^j x^{2j+1}}{j!2^j(2j+1)}, \quad -\infty < x < \infty. \tag{2.3.23}$$

The computation according to this formula is often inefficient. An excellent computing formula was given by Zelen and Severo (1968), namely

$$\Phi(x) = 1 - \phi(x)[b_1 t + b_2 t^2 + \cdots + b_5 t^5] + \epsilon(x), \quad x \geq 0, \tag{2.3.24}$$

where $t = (1 + px)^{-1}$, $p = .2316419$; $b_1 = .3193815$; $b_2 = -.3565638$; $b_3 = 1.7814779$; $b_4 = -1.8212550$; $b_5 = 1.3302744$. The magnitude of the error term is $|\epsilon(x)| < 7.5 \cdot 10^{-8}$.

2.3.6 Normal Approximations

The normal distribution can be used in certain cases to approximate well, the cumulative probabilities of other distribution functions. Such approximations are very useful when it becomes too difficult to compute the exact cumulative probabilities of the distributions under consideration. For example, suppose $X \sim B(100, .35)$ and we have to compute the probability of the event $\{X \leq 88\}$. This requires the computation of the sum of 89 terms in

$$B(88; 100, .35) = \sum_{j=0}^{88} \binom{100}{j} (.35)^j (.65)^{100-j}.$$

Usually, such a numerical problem requires the use of some numerical approximation and/or the use of a computer. However, the cumulative probability $B(88 \mid 100, .35)$ can be easily approximated by the normal c.d.f. This approximation is based on the celebrated Central Limit Theorem, which was discussed in Section 1.12. Accordingly, if $X \sim B(n, \theta)$ and n is sufficiently large (relative to θ) then, for $0 \leq k_1 \leq k_2 \leq n$,

$$P\{k_1 \leq X \leq k_2\} \approx \Phi\left(\frac{k_2 + \frac{1}{2} - n\theta}{\sqrt{n\theta(1-\theta)}}\right) - \Phi\left(\frac{k_1 - \frac{1}{2} - n\theta}{\sqrt{n\theta(1-\theta)}}\right). \tag{2.3.25}$$

The symbol $\approx$ designates a large sample approximation.

The maximal possible error in using this approximation is less than $.14[n\theta(1-\theta)]^{-1/2}$ (Johnson and Kotz, 1969, p. 64). The approximation turns out to be quite good, even if n is not very large, if θ is close to $\theta_0 = .5$. In Table 2.1, we compare the

Table 2.1 Normal Approximation to the Binomial c.d.f. $n = 25$

	$\theta = .5$		$\theta = .4$		$\theta = .25$	
k	Exact	Approx.	Exact	Approx.	Exact	Approx.
0	0.000000	0.000001	0.000003	0.000053	0.000753	0.003956
1	0.000001	0.000005	0.000047	0.000260	0.006271	0.014120
2	0.000010	0.000032	0.000426	0.001100	0.031356	0.041632
3	0.000078	0.000159	0.002364	0.003982	0.095462	0.102012
4	0.000455	0.000687	0.009468	0.012372	0.212988	0.209462
5	0.002039	0.002555	0.029359	0.033096	0.377526	0.364517
6	0.007317	0.008198	0.073562	0.076521	0.560346	0.545964
7	0.021643	0.022750	0.153549	0.153717	0.725754	0.718149
8	0.053876	0.054799	0.273529	0.270146	0.849810	0.850651
9	0.114761	0.115070	0.424614	0.419128	0.927919	0.933337
10	0.212178	0.211855	0.585772	0.580872	0.969578	0.975176
11	0.345019	0.344578	0.732279	0.729854	0.988513	0.992343
12	0.500000	0.500000	0.846229	0.846283	0.995877	0.998054
13	0.654981	0.655422	0.922196	0.923479	0.998332	0.999594
14	0.787822	0.788145	0.965606	0.966904	0.999033	0.999931
15	0.885238	0.884930	0.986828	0.987628	0.999204	0.999990
16	0.946124	0.945201	0.995671	0.996018	0.999240	0.999999
17	0.978357	0.977250	0.998792	0.998900	0.999246	1.000000
18	0.992683	0.991802	0.999716	0.999740	0.999247	1.000000
19	0.997961	0.997445	0.999944	0.999947	0.999247	1.000000
20	0.999545	0.999313	0.999939	0.999991	0.999247	1.000000
21	0.999922	0.999841	0.999996	0.999999	0.999247	1.000000
22	0.999990	0.999968	0.999997	1.000000	0.999247	1.000000
23	0.999999	0.999995	0.999997	1.000000	0.999247	1.000000
24	1.000000	0.999999	0.999997	1.000000	0.999247	1.000000
25	1.000000	1.000000	0.999997	1.000000	0.999247	1.000000

numerically exact c.d.f. values of the Binomial distribution $B(k; n, \theta)$ with $n = 25$ (relatively small) and $\theta = .25, .40, .50$ to the approximation obtained from (2.3.25) with $k = k_2$ and $k_1 = 0$.

Considerable research has been done to improve the Normal approximation to the Binomial c.d.f. Some of the main results and references are provided in Johnson and Kotz (1969, p. 64).

In a similar manner, the normal approximation can be applied to approximate the Hypergeometric c.d.f. (Johnson and Kotz, 1969, p. 148); the Poisson c.d.f. (Johnson and Kotz, 1969, p. 99) and the Negative-Binomial c.d.f. (Johnson and Kotz, 1969, p. 127).

The normal distribution can provide also good approximations to the $G(\lambda, \nu)$ distributions, when ν is sufficiently large, and to other continuous distributions. For a summary of approximating formulae and references see Johnson and Kotz (1969) and Zelen and Severo (1968). In Table 2.2 we summarize important characteristics of the above distribution functions.

Table 2.2 Expectations, Variances and Moment Generating Functions of Selected Distributions

Distribution	$E\{X\}$	$V\{X\}$	$M(t)$
Binomial $B(n, \theta)$ $0 < \theta < 1$	$n\theta$	$n\theta(1-\theta)$	$[e^t\theta + (1-\theta)]^n$, all $-\infty < t < \infty$
Hypergeometric $H(N, M, n)$ $1 \le n \le N$ $0 \le M \le N$	$n\frac{M}{N}$	$n\frac{M}{N}\left(1-\frac{M}{N}\right)\cdot$ $\cdot\left(1-\frac{n-1}{N-1}\right)$	$\sum_{j=0}^{n} e^{tj}\frac{\binom{M}{j}\binom{N-M}{n-j}}{\binom{N}{n}}$, all $-\infty < t < \infty$
Poisson $P(\lambda)$ $0 < \lambda < \infty$	λ	λ	$e^{-\lambda(1-e^t)}$, all $-\infty < t < \infty$
Negative-Binomial $NB(\psi, \nu)$ $0 < \psi < 1; 0 < \nu < \infty$	$\nu\psi/(1-\psi)$	$\nu\psi/(1-\psi)^2$	$[(1-\psi)/(1-e^t\psi)]^\nu, t < -\log\psi$
Rectangular $R(\theta_1, \theta_2)$ $-\infty < \theta_1 < \theta_2 < \infty$	$\frac{\theta_1+\theta_2}{2}$	$\frac{(\theta_2-\theta_1)^2}{12}$	$\frac{e^{t\theta_1}}{t(\theta_2-\theta_1)}[e^{t(\theta_2-\theta_1)}-1], t \ne 0$ $1, \quad t = 0$
Beta $\beta(p, q)$ $0 < p, q < \infty$	$\frac{p}{p+q}$	$\frac{pq}{(p+q)^2(p+q+1)}$	$\sum_{j=0}^{\infty}\frac{t^j}{j!}\prod_{i=0}^{j-1}\frac{p+i}{p+q+i}, -\infty < t < \infty$

Gamma			
$G(\lambda, \nu)$	$\frac{\nu}{\lambda}$	$\frac{\nu}{\lambda^2}$	$\left(1 - \frac{t}{\lambda}\right)^{-\nu}, t < \lambda$
$0 < \lambda, \nu < \infty$			
Weibull		$\frac{1}{\lambda^{2/\alpha}}\left[\Gamma\left(1 + \frac{2}{\alpha}\right)\right.$	
$W(\lambda, \alpha) \sim$			
$G^{1/\alpha}(\lambda, 1)$	$\frac{\Gamma\left(1 + \frac{1}{\alpha}\right)}{\lambda^{1/\alpha}}$	$\left.-\Gamma^2\left(1 + \frac{1}{\alpha}\right)\right]$	$\sum_{j=0}^{\infty}\left(\left(\frac{t}{\lambda^{1/\alpha}}\right)^j \frac{\Gamma\left(1 + \frac{j}{\alpha}\right)}{\Gamma(1+j)}\right), t < \lambda^{1/\alpha}$
$0 < \alpha, \lambda < \infty$			
Extreme-Values			
$EV(\lambda, \alpha) \sim$			
$-\log G^{1/\alpha}(\lambda, 1)$	$\frac{1}{\alpha}(\log\lambda + \gamma)$	$\pi^2/6\alpha^2$	$\lambda^{t/\alpha}\Gamma\left(1 - \frac{t}{\alpha}\right), t < \alpha$
$0 < \lambda, \alpha < \infty$	$\gamma = .577216\ldots$ is the Euler constant		
Normal			
$N(\mu, \sigma^2)$			
$-\infty < \mu < \infty$	μ	σ^2	$e^{\mu t + \frac{1}{2}t^2\sigma^2}, -\infty < t < \infty$
$0 < \sigma < \infty$			

2.4 TRANSFORMATIONS

2.4.1 One-to-One Transformations of Several Variables

Let $X_1, \ldots, X_k$ be random variables of the continuous type with a joint p.d.f. $f(x_1, \ldots, x_k)$. Let $y_i = g_i(x_i, \ldots, x_k)$, $i = 1, \ldots, k$, be one-to-one transformations, and let $x_i = \psi_i(y_1, \ldots, y_k)$ $i = 1, \ldots, k$, be the inverse transformations. Assume that $\frac{\partial \psi_i}{\partial y_j}$ are continuous for all $i, j = 1, \ldots, k$ at all points $(y_1, \ldots, y_k)$. The Jacobian of the transformation is

$$J(y_1, \ldots, y_k) = \text{det.} \left(\frac{\partial \psi_i}{\partial y_j}; i, j = 1, \ldots, k \right); \tag{2.4.1}$$

where det.$(\cdot)$ denotes the determinant of the matrix of partial derivatives. Then the joint p.d.f. of $(Y_1, \ldots, Y_k)$ is

$$h(y_1, \ldots, y_k) = f(\psi_1(\mathbf{y}), \ldots, \psi_k(\mathbf{y}))|J(\mathbf{y})|, \quad \mathbf{y} = (y_1, \ldots, y_k). \tag{2.4.2}$$

2.4.2 Distribution of Sums

Let X_1, X_2 be absolutely continuous random variables with a joint p.d.f. $f(x_1, x_2)$. Consider the one-to-one transformation $Y_1 = X_1$, $Y_2 = X_1 + X_2$. It is easy to verify that $J(y_1, y_2) = 1$. Hence,

$$f_{Y_1, Y_2}(y_1, y_2) = f_{X_1, X_2}(y_1, y_2 - y_1).$$

Integrating over the range of Y_1 we obtain the marginal p.d.f. of Y_2, which is the required p.d.f. of the sum. Thus, if $g(y)$ denotes the p.d.f. of Y_2

$$g(y) = \int_{-\infty}^{\infty} f(x, y - x) dx. \tag{2.4.3}$$

If X_1 and X_2 are independent, having marginal p.d.f.s $f_1(x)$ and $f_2(x)$, the p.d.f. of the sum $g(y)$ is the **convolution** of $f_1(x)$ and $f_2(x)$, i.e.,

$$g(y) = \int_{-\infty}^{\infty} f_1(x) f_2(y - x) dx. \tag{2.4.4}$$

If X_1 is discrete, the integral in (2.4.4) is replaced by a sum over the jump points of $F_1(x)$. If there are more than two variables, the distribution of the sum can be found by a similar method.

2.4.3 Distribution of Ratios

Let X_1, X_2 be absolutely continuous with a joint p.d.f., $f(x_1, x_2)$. We wish to derive the p.d.f. of $R = X_1/X_2$. In the general case, X_2 can be positive or negative and

therefore we separate between the two cases. Over the set $-\infty < x_1 < \infty, 0 < x_2 < \infty$ the transformation $R = X_1/X_2$ and $Y = X_2$ is one-to-one. It is also the case over the set $-\infty < x_1 < \infty$, $-\infty < x_2 < 0$. The Jacobian of the inverse transformation is $J(y, r) = -y$. Hence, the p.d.f. of R is

$$h(r) = -\int_{-\infty}^{0} yf(yr, y)dy + \int_{0}^{\infty} yf(yr, y)dy. \tag{2.4.5}$$

The result of Example 2.2 has important applications.

Let $X_1, X_2, \ldots, X_k$ be independent random variables having gamma distributions with equal λ, i.e., $X_i \sim G(\lambda, \nu_i)$, $i = 1, \ldots, k$. Let $T = \sum_{i=1}^{k} X_i$ and for $i = 1, \ldots, k-1$

$$Y_i = X_i/T.$$

The marginal distribution of Y_i is $\beta\left(\nu_i, \sum_{j=1}^{k} \nu_j - \nu_i\right)$. The joint distribution of $\mathbf{Y} = (Y_1, \ldots, Y_{k-1})$ is called the **Dirichlet distribution**, $\mathcal{D}(\nu_1, \nu_2, \ldots, \nu_k)$, whose joint p.d.f. is

$$g(y_1, \ldots, y_{k-1}; \boldsymbol{\nu}) = \frac{\Gamma(\nu_1 + \cdots + \nu_k)}{\prod_{i=1}^{k} \Gamma(\nu_i)} \prod_{i=1}^{k-1} y_i^{\nu_i - 1} \left(1 - \sum_{j=1}^{k-1} y_j\right)^{\nu_k - 1}, \tag{2.4.6}$$

for $y_i \geq 0$, $\sum_{j=1}^{k-1} y_j \leq 1$.

The p.d.f. of $\mathcal{D}(\nu_1, \ldots, \nu_k)$ is a multivariate generalization of the beta distribution.

Let $\nu^* = \sum_{j=1}^{k} \nu_j$. One can immediately prove that for all $i, i' = 1, \ldots, k-1$

$$E\{Y_i Y_{i'}\} = \frac{\nu_i \nu_{i'}}{\nu^*(\nu^* + 1)}, \tag{2.4.7}$$

and thus

$$\text{cov}(Y_i, Y_{i'}) = -\frac{\nu_i \nu_{i'}}{\nu^{*2}(\nu^* + 1)} \tag{2.4.8}$$

Additional properties of the Dirichlet distributions are specified in the exercises.

2.5 VARIANCES AND COVARIANCES OF SAMPLE MOMENTS

A **random sample** is a set of n ($n \geq 1$) **independent and identically distributed** (i.i.d.) random variables, having a common distribution $F(x)$. We assume that F has all moments required in the following development. The rth moment of F, $r \geq 1$, is μ_r.

The rth **sample moment** is

$$\hat{\mu}_r = \frac{1}{n} \sum_{i=1}^{n} X_i^r. \tag{2.5.1}$$

We immediately obtain that

$$\begin{aligned} E\{\hat{\mu}_r\} &= \frac{1}{n} \sum_{i=1}^{n} E\{X_i^r\} \\ &= \mu_r, \quad r \geq 1, \end{aligned} \tag{2.5.2}$$

since all X_i are identically distributed. Notice that due to independence, $\text{cov}(X_i, X_j) = 0$ for all $i \neq j$. We present here a method for computing $V\{\hat{\mu}_r\}$ and $\text{cov}\{\hat{\mu}_r, \hat{\mu}_{r'}\}$ for $r \neq r'$. We consider expansions of the form $\left(\sum_{i=1}^{n} X_i^l\right)^k$, $l, k \geq 1$, in terms of **augmented symmetric functions** and introduce the following notation

$$(l) = \sum_{i=1}^{n} X_i^l, \tag{2.5.3}$$

$$[l_1 l_2] = \sum_{i \neq j} \sum X_i^{l_1} X_j^{l_2}, \tag{2.5.4}$$

$$[l_1 l_2 l_3] = \sum_{i \neq j \neq k} \sum \sum X_i^{l_1} X_j^{l_2} X_k^{l_3}, \tag{2.5.5}$$

etc. The sum of powers in such an expression is called the **weight** of [ ]. Thus, the weight of $[l_1 l_2 l_3]$ is $w = l_1 + l_2 + l_3$. In Table 2.3, we find expansions of $(l_1)^{\alpha_1}(l_2)^{\alpha_2} \cdots$ in terms of multi-sums $[l_1^{r_1} l_2^{r_2} \cdots]$. For additional values of coefficients for such expansions, see David and Kendall (1955). For example, to expand $\left(\sum_{i=1}^{n} X_i^3\right)\left(\sum_{i=1}^{n} X_i\right)^2$ the weight is $w = 5$, and according to Table 2.3, $(3)(1)^2 = [5] + 2[41] + [32] + [31^2]$.

Table 2.3 Augmented Symmetric Functions in Terms of Power-Series

Weight	( )	[ ]
2	(2)	[2]
	$(1)^2$	$[2]+[1^2]$
3	(3)	[3]
	(2)(1)	$[3]+[21]$
	$(1)^3$	$[3]+3[21]+[1^3]$
4	(4)	[4]
	(3)(1)	$[4]+[31]$
	$(2)^2$	$[4]+[2^2]$
	$(2)(1)^2$	$[4]+2[31]+[2^2]+[21^2]$
	$(1)^4$	$[4]+4[31]+3[2^2]+6[21^2]+[1^4]$
5	(5)	[5]
	(4)(1)	$[5]+[41]$
	(3)(2)	$[5]+[32]$
	$(3)(1)^2$	$[5]+2[41]+[32]+[31^2]$
	$(2)^2(1)$	$[5]+[41]+2[32]+[2^21]$
	$(2)(1)^3$	$[5]+3[41]+4[32]+3[31^2]+3[2^21]+[21^3]$
	$(1)^5$	$[5]+5[41]+10[32]+10[31^2]+15[2^21]+10[21^3]+[1^5]$
6	(6)	[6]
	(5)(1)	$[6]+[51]$
	(4)(2)	$[6]+[42]$
	$(4)(1^2)$	$[6]+2[51]+[42]+[41^2]$
	$(3)^2$	$[6]+[3^2]$
	(3)(2)(1)	$[6]+[51]+[42]+[3^2]+[321]$
	$(3)(1)^3$	$[6]+3[51]+3[42]+3[41^2]+[3^2]+3[321]+[31^3]$
	$(2)^3$	$[6]+3[42]+[2^3]$
	$(2)^2(1)^2$	$[6]+2[51]+3[42]+[41^2]+2[3^2]+4[321]+[2^3]+[2^21^2]$
	$(2)(1)^4$	$[6]+4[51]+7[42]+6[41^2]+4[3^2]+16[32]$ $+4[31^3]+3[2^3]+6[2^21^2]+[21^4]$
	$(1)^6$	$[6]+6[51]+15[42]+15[41^2]+10[3^2]+60[321]$ $+20[31^3]+15[2^3]+45[2^21^2]+15[21^4]+[1^6]$

(*) $[3^2]=[33]$, etc.
Source: Compiled from David and Kendall (1955).

Thus,

$$\left(\sum_i X_i^3\right)\left(\sum_i X_i\right)^2 = \sum_i X_i^5 + 2\sum_{i\neq j}\sum X_i^4 X_j + \sum_{i\neq j}\sum X_i^3 X_j^2 + \sum_{i\neq j\neq k}\sum\sum X_i^3 X_j X_k.$$

The expected values of such expansions are given in terms of product of the moments (independence) times the number of terms in the sum, e.g.,

$$E\left\{\sum\sum_{i\neq j} X_i^3 X_j^2\right\} = n(n-1)\mu_3\mu_2.$$

2.6 DISCRETE MULTIVARIATE DISTRIBUTIONS

2.6.1 The Multinomial Distribution

Consider an experiment in which the result of each trial belongs to one of k alternative categories. Let $\boldsymbol{\theta}' = (\theta_1, \ldots, \theta_k)$ be a probability vector, i.e., $0 < \theta_i < 1$ for all $i = 1, \ldots, k$ and $\sum_{i=1}^{k} \theta_i = 1$. θ_i designates the probability that the outcome of an individual trial belongs to the ith category. Consider n such independent trials, $n \geq 1$, and let $\mathbf{X} = (X_1, \ldots, X_k)$ be a random vector. X_i is the number of trials in which the ith category is realized, $\sum_{i=1}^{k} X_i = n$. The distribution of $\mathbf{X}$ is given by the multinomial probability distribution

$$p(j_1, \ldots, j_k; n, \theta) = \frac{n!}{j_1! \ldots j_k!} \prod_{i=1}^{k} \theta_i^{j_i}, \tag{2.6.1}$$

where $j_i = 0, 1, \ldots, n$ and $\sum_{i=1}^{k} j_i = n$. These terms are obtained by the multinomial expansion of $(\theta_1 + \cdots + \theta_k)^n$. Hence, their sum equals 1. We will designate the multinomial distribution based on n trials and probability vector $\boldsymbol{\theta}$ by $M(n, \boldsymbol{\theta})$. The binomial distribution is a special case, when $k = 2$. Moreover, the marginal distribution of X_i is the binomial $B(n, \theta_i)$. The joint marginal distribution of any pair $(X_i, X_{i'})$ where $1 \leq i < i' \leq k$ is the corresponding trinomial, with probability distribution function

$$p(j_i, j_{i'}) = \frac{n!}{j_i!\, j_{i'}!(n - j_i - j_{i'})!} \theta_i^{j_i} \theta_{i'}^{j_{i'}} (1 - \theta_i - \theta_{i'})^{n-\nu} \tag{2.6.2}$$
$$\text{where } \nu = j_i + j_{i'}.$$

We consider now the moments of the multinomial distribution. From the marginal Binomial distribution of the Xs we have

$$\begin{aligned} E\{X_i\} &= n\theta_i, \quad i = 1, \ldots, k \\ V\{X_i\} &= n\theta_i(1 - \theta_i), \quad i = 1, \ldots, k. \end{aligned} \tag{2.6.3}$$

To obtain the covariance of X_i, X_j, $i \neq j$ we proceed in the following manner. If $n = 1$ then $E\{X_i X_j\} = 0$ for all $i \neq j$, since only one of the components of $\mathbf{X}$ is one and all the others are zero. Hence, $E\{X_i X_j\} - E\{X_i\}E\{X_j\} = -\theta_i\theta_j$ if $i \neq j$. If $n > 1$, we obtain the result by considering the sum of n independent vectors. Thus,

$$\text{cov}(X_i, X_j) = -n\theta_i\theta_j, \quad \text{all } i \neq j. \tag{2.6.4}$$

We conclude the section with a remark about the joint moment generating function (m.g.f.) of the multinomial random vector $\mathbf{X}$. This function is defined in the following manner. Since $X_k = n - \sum_{i=1}^{k-1} X_i$, we define for every $k \geq 2$

$$M(t_1, \ldots, t_{k-1}) = E\left\{\exp \sum_{i=1}^{k-1} t_i X_i\right\}. \tag{2.6.5}$$

One can prove by induction on k that

$$M(t_1, \ldots, t_{k-1}) = \left[\sum_{i=1}^{k-1} \theta_i e^{t_i} + \left(1 - \sum_{i=1}^{k-1} \theta_i\right)\right]^n. \tag{2.6.6}$$

2.6.2 Multivariate Negative Binomial

Let $\mathbf{X} = (X_1, \ldots, X_k)$ be a k-dimensional random vector. Each random variable, X_i, $i = 1, \ldots, k$, can assume only nonnegative integers. Their joint probability distribution function is given by

$$g(j_1, \ldots, j_k; \boldsymbol{\theta}, \nu) = \frac{\Gamma\left(\nu + \sum_{i=1}^{k} j_i\right)}{\Gamma(\nu)\prod_{i=1}^{k}\Gamma(j_i + 1)} \left(1 - \sum_{i=1}^{k} \theta_i\right)^{\nu} \prod_{i=1}^{k} \theta_i^{j_i}, \tag{2.6.7}$$

where $j_1, \ldots, j_k = 0, 1, \ldots$; $0 < \nu < \infty$, $0 < \theta_i < 1$ for each $i = 1, \ldots, k$ and $\sum_{i=1}^{k} \theta_i < 1$. We develop here the basic theory for the case of $k = 2$. (For $k = 1$ the distribution reduces to the univariate NB(θ, ν). Summing first with respect to j_2 we obtain

$$\sum_{j_2=0}^{\infty} g(j_1, j_2; \theta_1, \theta_2, \nu) = \frac{\Gamma(\nu + j_1)(1 - \theta_1 - \theta_2)^{\nu}\theta_1^{j_1}}{(1 - \theta_2)^{\nu + j_1}\Gamma(\nu)\Gamma(j_1 + 1)}. \tag{2.6.8}$$

Hence, the marginal of X_i is

$$P\{X_1 = j_1\} = nb\left(j_1; \frac{\theta_1}{1-\theta_2}, \nu\right), \quad j_1 = 0, 1, \ldots \tag{2.6.9}$$

where $nb(j; \psi, \nu)$ is the p.d.f. of the negative binomial NB(ψ, ν). By dividing the joint probability distribution function $g(j_1, j_2; \theta_1, \theta_2, \nu)$ by $nb\left(j_1; \frac{\theta_1}{1-\theta_2}, \nu\right)$, we obtain that the conditional distribution of X_2 given X_1 is the negative binomial NB(θ_2, $\nu + X_1$). Accordingly, if NB(θ_1, θ_2, ν) designates a bivariate negative binomial with parameters (θ_1, θ_2, ν), then the expected value of X_i is given by

$$E\{X_i\} = \nu\theta_i/(1 - \theta_1 - \theta_2), \quad i = 1, 2. \tag{2.6.10}$$

The variance of the marginal distribution is

$$V\{X_1\} = \nu\theta_1(1 - \theta_2)/(1 - \theta_1 - \theta_2)^2. \tag{2.6.11}$$

Finally, to obtain the covariance between X_1 and X_2 we determine first

$$\begin{aligned} E\{X_1X_2\} &= E\{X_1E\{X_2 \mid X_1\}\} \\ &= \frac{\theta_2}{1-\theta_2}E\{X_1(\nu + X_2)\} = \nu(\nu+1)\frac{\theta_1\theta_2}{(1-\theta_1-\theta_2)^2}. \end{aligned} \tag{2.6.12}$$

Therefore,

$$\text{cov}(X_1, X_2) = \frac{\nu\theta_1\theta_2}{(1-\theta_1-\theta_2)^2}. \tag{2.6.13}$$

We notice that, contrary to the multinomial case, the covariances of any two components of the multivariate negative binomial vector are all positive.

2.6.3 Multivariate Hypergeometric Distributions

This family of k-variate distributions is derived by a straightforward generalization of the univariate model. Accordingly, suppose that a finite population of elements contain M_1 of type 1, M_2 of type 2, ..., M_k of type k and $N - \sum_{i=1}^{k} M_i$ of other types. A sample of n elements is drawn at random and without replacement from this

population. Let X_i, $i = 1, \ldots, k$ denote the number of elements of type i observed in the sample. The p.d.f. of $\mathbf{X} = (X_1, \ldots, X_k)$ is

$$f(x_1, \ldots, x_k; N, M_1, \ldots, M_k, n) = \frac{\prod_{i=1}^{k} \binom{M_i}{x_i} \binom{N - \Sigma M_i}{n - \Sigma x_i}}{\binom{N}{n}}, \tag{2.6.14}$$

$$x_i = 0, 1, \ldots (i = 1, \ldots, k), \sum_{i=1}^{k} x_i \leq n.$$

One immediately obtains that the marginal distributions of the components of X are hypergeometric distributions, with parameters (N, M_i, n), $i = 1, \ldots, k$. If we designate by $H(N, M_1, \ldots, M_k, n)$ the multivariate hypergeometric distribution, then the conditional distribution of $(X_{r+1}, \ldots, X_k)$ given $(X_1 = j_1, \ldots, X_r = j_r)$ is the hypergeometric $H\left(N - \sum_{i=1}^{r} M_i, M_{r+1}, \ldots, M_k, n - \sum_{i=1}^{r} j_i\right)$. Using this result and the law of the iterated expectation we obtain the following result, for all $i \neq j$,

$$\operatorname{cov}(X_i, X_j) = -n \left(1 - \frac{n-1}{N-1}\right) \frac{M_i}{N} \cdot \frac{M_j}{N}. \tag{2.6.15}$$

This result is similar to that of the multinomial (2.6.4), which corresponds to sampling with replacement.

2.7 MULTINORMAL DISTRIBUTIONS

2.7.1 Basic Theory

A random vector $(X_1, \ldots, X_k)$ of the continuous type has a k-variate multinormal distribution if its joint p.d.f. can be expressed in vector and matrix notation as

$$f(x_1, \ldots, x_k) = \frac{1}{(2\pi)^{k/2} |V|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\xi})' V^{-1} (\mathbf{x} - \boldsymbol{\xi})\right\} \tag{2.7.1}$$

for $-\infty < \xi_i < \infty$, $i = 1, \ldots, k$. Here, $\mathbf{x} = (x_1, \ldots, x_k)'$, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k)'$. V is a $k \times k$ symmetric positive definite matrix and $|V|$ is the determinant of V. We introduce the notation $\mathbf{X} \sim N(\boldsymbol{\xi}, V)$. We notice that the k-variate multinormal p.d.f. (2.7.1) is symmetric about the point $\boldsymbol{\xi}$. Hence, $\boldsymbol{\xi}$ is the expected value (mean) vector of $\mathbf{X}$. Moreover, all the moments of $\mathbf{X}$ exist.

The m.g.f. of $\mathbf{X}$ is

$$M(t_1, \ldots, t_k) = E\{\exp(\mathbf{t}'\mathbf{X})\} = \exp\left(\frac{1}{2}\mathbf{t}'V\mathbf{t} + \mathbf{t}'\boldsymbol{\xi}\right). \tag{2.7.2}$$

To establish formula (2.7.2) we can assume, without loss of generality, that $\boldsymbol{\xi} = \mathbf{0}$, since if $M_{\mathbf{X}}(t)$ is the m.g.f. of $\mathbf{X}$ and $\mathbf{Y} = \mathbf{X} + \mathbf{b}$, then the m.g.f. of Y is $M_{\mathbf{Y}}(\mathbf{t}) = \exp(\mathbf{t}'\mathbf{b})M_{\mathbf{X}}(t)$. Thus, we have to determine

$$M(\mathbf{t}) = \frac{1}{(2\pi)^{k/2}|V|^{1/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(\mathbf{t}'\mathbf{x} - \frac{1}{2}\mathbf{x}'V^{-1}\mathbf{x}\right) \prod_{i=1}^{k} dx_i. \tag{2.7.3}$$

Since V is positive definite, there exists a nonsingular matrix D such that $V = DD'$. Consider the transformation $\mathbf{Y} = D^{-1}\mathbf{X}$; then $\mathbf{x}'V^{-1}\mathbf{x} = \mathbf{y}'\mathbf{y}$ and $\mathbf{t}'\mathbf{x} = \mathbf{t}'D\mathbf{y}$. Therefore,

$$-\frac{1}{2}\mathbf{x}'V^{-1}\mathbf{x} + t'\mathbf{x} = -\frac{1}{2}(\mathbf{y} - D'\mathbf{t})'(y - D'\mathbf{t}) + \frac{1}{2}t'V\mathbf{t}. \tag{2.7.4}$$

Finally, the Jacobian of the transformation is $|D|$ and

$$M(\mathbf{t}) = \exp\left(\frac{1}{2}\mathbf{t}'V\mathbf{t}\right) \cdot \frac{|D|}{(2\pi)^{k/2}|V|^{1/2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(\mathbf{y} - D'\mathbf{t})'(\mathbf{y} - D'\mathbf{t})\right) \prod_{i=1}^{k} dy_i. \tag{2.7.5}$$

Since $|D| = |V|^{1/2}$ and $(2\pi)^{-k/2}$ times the multiple integral on the right-hand side is equal to one, we establish (2.7.2). In order to determine the variance–covariance matrix of $\mathbf{X}$ we can assume, without loss of generality, that its expected value is zero. Accordingly, for all i, j,

$$\operatorname{cov}(X_i, X_j) = \frac{\partial^2}{\partial t_i \partial t_j} M(\mathbf{t})\Big|_{\mathbf{t}=\mathbf{0}}. \tag{2.7.6}$$

From (2.7.2) and (2.7.6), we obtain that $\operatorname{cov}(X_i, X_j) = \sigma_{ij}$. $(i, j = 1, \ldots, k)$, where σ_{ij} is the (i, j)th element of V. Thus, V is the variance–covariance matrix of X.

A k-variate multinormal distribution is called **standard** if $\xi_i = 0$ and $\sigma_{ii} = 1$ for all $i = 1, \ldots, k$. In this case, the variance matrix will be denoted by R since its elements are the correlations between the components of $\mathbf{X}$. A standard normal vector is often denoted by $\mathbf{Z}$, its joint p.d.f. and c.d.f. by $\phi_k(\mathbf{z} \mid R)$ and $\Phi_k(\mathbf{z} \mid R)$, respectively.

2.7.2 Distribution of Subvectors and Distributions of Linear Forms

In this section we present several basic results without proofs. The proofs are straightforward and the reader is referred to Anderson (1958) and Graybill (1961).

Suppose that a k-dimensional vector $\mathbf{X}$ has a multinormal distribution $N(\boldsymbol{\mu}, V)$. We consider the two subvectors $\mathbf{Y}$ and $\mathbf{Z}$, i.e., $\mathbf{X}' = (\mathbf{Y}', \mathbf{Z}')$, where $\mathbf{Y}$ is r-dimensional, $1 \leq r < k$.

Partition correspondingly the expectation vector $\boldsymbol{\xi}$ to $\boldsymbol{\xi}' = (\boldsymbol{\eta}', \boldsymbol{\zeta}')$ and the covariance matrix to

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}$$

The following results are fundamental to the multinormal theory.

(i) $\mathbf{Y} \sim N(\boldsymbol{\eta}, V_{11})$
(ii) $\mathbf{Z} \sim N(\boldsymbol{\zeta}, V_{22})$
(iii) $\mathbf{Y} \mid \mathbf{Z} \sim N(\boldsymbol{\eta} + V_{12}V_{22}^{-1}(\mathbf{Z} - \boldsymbol{\zeta}), V_{11} - V_{12}V_{22}^{-1}V_{21})$,

and an analogous formula can be obtained for the conditional distribution of $\mathbf{Z}$ given $\mathbf{Y}$.

The conditional expectation

$$E\{\mathbf{Y} \mid \mathbf{Z}\} = \boldsymbol{\eta} + V_{12}V_{22}^{-1}(\mathbf{Z} - \boldsymbol{\zeta}) \tag{2.7.7}$$

is called the linear **regression** of $\mathbf{Y}$ on $\mathbf{Z}$. The conditional covariance matrix

$$\mathfrak{D}(\mathbf{Y} \mid \mathbf{Z}) = V_{11} - V_{12}V_{22}^{-1}V_{21} \tag{2.7.8}$$

represents the variances and covariances of the components of $\mathbf{Y}$ around the linear regression hyperplane. The above results have the following converse counterpart. Suppose that $\mathbf{Y}$ and $\mathbf{Z}$ are two vectors such that

(i) $\mathbf{Y} \mid \mathbf{Z} \sim N(A\mathbf{Z}, V)$

and

(ii) $\mathbf{Z} \sim N(\zeta, D)$;

then the marginal distribution of $\mathbf{Y}$ is the multinormal

$$\mathbf{Y} \sim N(A\boldsymbol{\zeta}, V + ADA')$$

and the joint distribution of $\mathbf{Y}$ and $\mathbf{Z}$ is the multinormal, with expectation vector $(\zeta' A', \zeta')'$ and a covariance matrix

$$\begin{pmatrix} V + ADA' & AD \\ DA' & D \end{pmatrix}.$$

Finally, if $\mathbf{X} \sim N(\boldsymbol{\xi}, V)$ and $\mathbf{Y} = \mathbf{b} + A\mathbf{X}$, then $\mathbf{Y} \sim N(\mathbf{b} + A\boldsymbol{\xi}, AVA')$. That is, **every linear combination of normally distributed random variables is normally distributed**.

In the case of $k = 2$, the multinormal distribution is called a **bivariate normal distribution**. The joint p.d.f. of a bivariate normal distribution is

$$f(x, y; \xi, \mu, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\xi}{\sigma_1}\right)^2 - 2\rho\frac{x-\xi}{\sigma_1}\cdot\frac{y-\eta}{\sigma_2} + \left(\frac{y-\eta}{\sigma_2}\right)^2\right]\right\}, \tag{2.7.9}$$

$-\infty < x, y < \infty$.

The parameters ξ and η are the expectations, and σ_1^2 and σ_2^2 are the variances of X and Y, respectively. ρ is the coefficient of correlation.

The conditional distribution of Y given $\{X = x\}$ is normal with conditional expectation

$$E\{Y \mid x\} = \eta + \beta(x - \xi), \tag{2.7.10}$$

where $\beta = \rho\sigma_2/\sigma_1$. The conditional variance is

$$\sigma^2_{y|x} = \sigma_2^2(1 - \rho^2). \tag{2.7.11}$$

These formulae are special cases of (2.7.7) and (2.7.8). Since the joint p.d.f. of (X, Y) can be written as the product of the conditional p.d.f. of Y given X, with the marginal p.d.f. of X, we obtain the expression,

$$f(x, y; \xi, \eta, \sigma_1, \sigma_2, \rho) = \frac{1}{\sigma_1}\phi\left(\frac{x-\xi}{\sigma_1}\right) \cdot \frac{1}{\sigma_2\sqrt{1-\rho^2}}\phi\left(\frac{y-\eta-\beta(x-\xi)}{\sigma_2\sqrt{1-\rho^2}}\right). \tag{2.7.12}$$

This expression can serve also as a basis for an algorithm to compute the Bivariate-Normal c.d.f., i.e.,

$$P\{X \le x_0, Y \le y_0\} = \int_{-\infty}^{(x_0-\xi)/\sigma_1} \phi(x)\Phi\left(\frac{y_0 - \eta - \rho\sigma_2 x}{\sigma_2\sqrt{1-\rho^2}}\right) dx. \quad (2.7.13)$$

Let Z_1, Z_2 and Z_3 have a joint standard Trivariate-Normal distribution, with a correlation matrix

$$R = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}.$$

The **conditional** Bivariate-Normal distribution of (Z_1, Z_2) given Z_3 has a covariance matrix

$$V = \begin{pmatrix} 1-\rho_{13}^2 & \rho_{12}-\rho_{13}\rho_{23} \\ \rho_{12}-\rho_{13}\rho_{23} & 1-\rho_{23}^2 \end{pmatrix}. \quad (2.7.14)$$

The conditional correlation between Z_1 and Z_2, given Z_3 can be determined from (2.7.14). It is called the **partial correlation** of Z_1, Z_2 under Z_3 and is given by

$$\rho_{12.3} = \frac{\rho_{12}-\rho_{13}\rho_{23}}{\sqrt{(1-\rho_{13}^2)(1-\rho_{23}^2)}}. \quad (2.7.15)$$

2.7.3 Independence of Linear Forms

Let $\mathbf{X} = (X_1, \ldots, X_k)'$ be a multinormal random vector. Without loss of generality, assume that $E\{\mathbf{X}\} = \mathbf{0}$. Let V be the covariance matrix of $\mathbf{X}$. We investigate first the conditions under which two linear functions $Y_1 = \boldsymbol{\alpha}'\mathbf{X}$ and $Y_2 = \boldsymbol{\beta}'\mathbf{X}$ are independent.

Let $\mathbf{Y} = (Y_1, Y_2)'$, $A = \begin{pmatrix} \boldsymbol{\alpha}' \\ \boldsymbol{\beta}' \end{pmatrix}$. That is, A is a $2 \times k$ matrix and $\mathbf{Y} = A\mathbf{X}$. $\mathbf{Y}$ has a bivariate normal distribution with a covariance matrix AVA'. Y_1 and Y_2 are independent if and only if $\mathrm{cov}(Y_1, Y_2) = 0$. Moreover, $\mathrm{cov}(Y_1, Y_2) = \boldsymbol{\alpha}'V\boldsymbol{\beta}$. Since V is positive definite there exists a nonsingular matrix C such that $V = CC'$. Accordingly, $\mathrm{cov}(Y_1, Y_2) = 0$ if and only if $(C'\boldsymbol{\alpha})'(C'\boldsymbol{\beta}) = 0$. This means that the vectors $C'\boldsymbol{\alpha}$ and $C'\boldsymbol{\beta}$ should be orthogonal. This condition is generalized in a similar fashion to cases where Y_1 and Y_2 are vectors. Accordingly, if $\mathbf{Y}_1 = A\mathbf{X}$ and $\mathbf{Y}_2 = B\mathbf{X}$, then $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent if and only if $AVB' = 0$. In other words, the column vectors of CA' should be mutually orthogonal to the column vectors of $C'B$.

2.8 DISTRIBUTIONS OF SYMMETRIC QUADRATIC FORMS OF NORMAL VARIABLES

In this section, we study the distributions of **symmetric** quadratic forms in normal random variables. We start from the simplest case.

Case A:

$$X \sim N(0, \sigma^2), \;\; Q = X^2.$$

Assume first that $\sigma^2 = 1$. The density of X is then $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)$. Therefore, the p.d.f. of Q is

$$\begin{aligned} f_Q(y) &= \frac{1}{2\sqrt{2\pi}} y^{-1/2} \left[\exp\left(-\frac{1}{2}(\sqrt{y})^2 \right) + \exp\left(-\frac{1}{2}(-\sqrt{y})^2 \right) \right] \\ &= \frac{1}{\sqrt{2}\,\Gamma(\frac{1}{2})} y^{-1/2} e^{-\frac{1}{2}y}, \quad 0 \le y \le \infty, \end{aligned} \tag{2.8.1}$$

since $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Comparing $f_Q(y)$ with the p.d.f. of the gamma distributions, we conclude that if $\sigma^2 = 1$ then $Q \sim G(\frac{1}{2}, \frac{1}{2}) \sim \chi^2[1]$. In the more general case of arbitrary σ^2, $Q \sim \sigma^2 \chi^2[1]$.

Case B:

$$X \sim N(\xi, \sigma^2), \;\; Q = X^2.$$

This is a more complicated situation. We shall prove that the p.d.f. of Q (and so its c.d.f. and m.g.f.) is, at each point, the expected value of the p.d.f. (or c.d.f. or m.g.f.) of $\sigma^2 \chi^2[1 + 2J]$, where J is a Poisson random variable with mean

$$\lambda = \frac{1}{2\sigma^2} \xi^2. \tag{2.8.2}$$

Such an expectation of distributions is called a **mixture**. The distribution of Q when $\sigma^2 = 1$ is called a **noncentral chi-squared** with 1 degree of freedom and parameter of noncentrality λ. In symbols $Q \sim \chi^2[1; \lambda]$. When $\lambda = 0$, the noncentral chi-squared coincides with the chi-squared, which is also called **central chi-squared**. The proof is obtained by determining first the m.g.f. of Q. As before, assume that $\sigma^2 = 1$. Then,

$$M_Q(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(tx^2 - \frac{1}{2}(x - \xi)^2 \right) dx. \tag{2.8.3}$$

Write, for all $t < \frac{1}{2}$,

$$tx^2 - \frac{1}{2}(x-\xi)^2 = \frac{1}{2}(1-2t)\left(x - \frac{\xi}{1-2t}\right)^2 + \xi^2\frac{t}{1-2t}. \tag{2.8.4}$$

Thus,

$$M_Q(t) = \exp\left(\xi^2\frac{t}{1-2t}\right)(1-2t)^{-1/2}, \quad t < 1/2. \tag{2.8.5}$$

Furthermore, $t/(1-2t) = -\frac{1}{2} + \frac{1}{2}(1-2t)^{-1}$. Hence,

$$M_Q(t) = e^{-\frac{1}{2}\xi^2}\sum_{j=0}^{\infty}\frac{\xi^{2j}}{j!2^j}(1-2t)^{-(\frac{1}{2}+j)}. \tag{2.8.6}$$

According to Table 2.2, $(1-2t)^{-(\frac{1}{2}+j)}$ is the m.g.f. of $\chi^2[1+2j]$. Thus, according to (2.8.6) the m.g.f. of $\chi^2[1;\lambda]$ is the mixture of the m.g.f.s of $\chi^2[1+2J]$, where J has a Poisson distribution, with mean λ as in (2.8.2). This implies that the distribution of $\chi^2[1;\lambda]$ is the marginal distribution of X in a model where (X, J) have a joint distribution, such that the conditional distribution of X given $\{J = j\}$ is like that of $\chi^2[1+2j]$ and the marginal distribution of J is Poisson with expectation λ. From Table 2.2, we obtain that $E\{\chi^2[\nu]\} = \nu$ and $V\{\chi^2[\nu]\} = 2\nu$. Hence, by the laws of the iterated expectation and total variance

$$E\{\chi^2[1;\lambda]\} = 1 + 2\lambda \tag{2.8.7}$$

and

$$V\{\chi^2[1;\lambda]\} = 2(1+4\lambda). \tag{2.8.8}$$

Case C:

$X_1, \ldots, X_n$ are independent; $X_i \sim N(\xi_i, \sigma^2)$, $i = 1, \ldots, n$,

$$Q = \sum_{i=1}^{n} X_i^2.$$

It is required that all the variances σ^2 are the same. As proven in Case B,

$$X_i^2 \sim \sigma^2\chi^2[1;\lambda_1] \sim \sigma^2\chi^2[1+2J_i], \quad i = 1, \ldots, n \tag{2.8.9}$$

where $J_i \sim P(\lambda_i)$.

Consider first the conditional distribution of Q given $(J_1, \ldots, J_n)$. From the result on the sum of independent chi-squared random variables, we infer

$$Q \mid (J_1, \ldots, J_n) \sim \sigma^2 \chi^2 \left[n + 2 \sum_{i=1}^{n} J_i \right], \tag{2.8.10}$$

where $Q \mid (J_1, \ldots, J_n)$ denotes the conditional equivalence of the random variables. Furthermore, since the original X_is are independent, so are the J_is and therefore

$$J_1 + \cdots + J_n \sim P(\lambda_1 + \cdots + \lambda_n). \tag{2.8.11}$$

Hence, the marginal distribution of Q is the mixture of $\sigma^2 \chi^2[n + 2M]$ where $M \sim P(\lambda_1 + \cdots + \lambda_n)$. We have thus proven that

$$Q \sim \sigma^2 \chi^2[n; \lambda_1 + \cdots + \lambda_n]. \tag{2.8.12}$$

Case D:

$$\mathbf{X} \sim N(\boldsymbol{\xi}, V) \text{ and } Q = \mathbf{X}' A \mathbf{X},$$

where A is a real symmetric matrix. The following is an important result.

$$Q \sim \chi^2[r; \lambda], \text{ with } \lambda = \frac{1}{2} \boldsymbol{\xi}' A \boldsymbol{\xi} \tag{2.8.13}$$

if and only if VA is an **idempotent** matrix of rank r (Graybill, 1961). The proof is based on the fact that every positive definite matrix V can be expressed as $V = CC'$, where C is nonsingular. If $\mathbf{Y} = C^{-1}\mathbf{X}$ then $\mathbf{Y} \sim N(C^{-1}\xi, I)$ and $\mathbf{X}' A \mathbf{X} = \mathbf{Y}' C' A C \mathbf{Y}$. $C'AC$ is idempotent if and only if VA is idempotent.

The following are important facts about real symmetric idempotent matrices.

(i) A is idempotent if $A^2 = A$.

(ii) All eigenvalues of A are either 1 or 0.

(iii) Rank $(A) = \text{tr.}\{A\}$, where $\text{tr.}\{A\} = \sum_{i=1}^{n} A_{ii}$, is the sum of the diagonal elements of A.

(iv) The only nonsingular idemptotent matrix is the identity matrix I.

2.9 INDEPENDENCE OF LINEAR AND QUADRATIC FORMS OF NORMAL VARIABLES

Without loss of generality, we assume that $\mathbf{X} \sim N(0, I)$. Indeed, if $\mathbf{X} \sim N(0, V)$ and $V = CC'$ make the transformation $\mathbf{X}^* = C^{-1}\mathbf{X}$, then $\mathbf{X}^* \sim N(0, I)$. Let $\mathbf{Y} = B\mathbf{X}$

and $Q = \mathbf{X}'A\mathbf{X}$, where A is **idempotent** of rank r, $1 \leq r \leq k$. B is an $n \times k$ matrix of full rank, $1 \leq n \leq k$.

Theorem 2.9.1. $\mathbf{Y}$ *and* $\mathbf{Q}$ *are independent if and only if*

$$BA = 0. \tag{2.9.1}$$

For proof, see Graybill (1961, Ch. 4).

Suppose now that we have m quadratic forms $\mathbf{X}'B_i\mathbf{X}$ in a multinormal vector $\mathbf{X} \sim N(\boldsymbol{\xi}, I)$.

Theorem 2.9.2. *If* $\mathbf{X} \sim N(\boldsymbol{\xi}, I)$ *the set of positive semidefinite quadratic forms* $\mathbf{X}'B_i\mathbf{X}$ $(i = 1, \ldots, m)$ *are jointly independent and* $\mathbf{X}'B_i\mathbf{X} \sim \chi^2[r_i; \lambda_i]$, *where* r_i *is the rank of* B_i *and* $\lambda_i = \frac{1}{2}\boldsymbol{\xi}'B_i\boldsymbol{\xi}$, *if any two of the following three conditions are satisfied.*

1. *Each* B_i *is idempotent* $(i = 1, \ldots, m)$;
2. $\sum_{j=1}^{m} B_j$ *is idempotent;*
3. $B_iB_j = 0$ *for all* $i \neq j$.

This theorem has many applications in the theory of regression analysis, as will be shown later.

2.10 THE ORDER STATISTICS

Let $X_1, \ldots, X_n$ be a set of random variables (having a joint distribution). The **order statistic** is

$$S(X_1, \ldots, X_n) = (X_{(1)}, X_{(2)}, \ldots, X_{(n)}), \tag{2.10.1}$$

where $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$.

If $X_1, \ldots, X_n$ are independent random variables having an identical absolutely continuous distribution function $F(x)$ with p.d.f. $f(x)$, then the p.d.f. of the order statistic is

$$f(x_{(1)}, \ldots, x_{(n)}) = n! \prod_{i=1}^{n} f(x_{(i)}). \tag{2.10.2}$$

To obtain the p.d.f. of the ith order statistic $X_{(i)}$, $i = 1, \ldots, n$, we can integrate (2.10.2) over the set

$$S_i(\xi) = \{-\infty \le X_{(1)} \le X_{(2)} \le \cdots \le X_{(i-1)} \le \xi \le X_{(i+1)} \le \cdots \le X_{(n)} \le \infty\}. \tag{2.10.3}$$

This integration yields the p.d.f.

$$f_{(i)}(\xi) = \frac{n!}{(i-1)!(n-i)!} f(\xi)(F(\xi))^{i-1}(1 - F(\xi))^{n-i}, \tag{2.10.4}$$

$-\infty < \xi < \infty$. We can obtain this result also by a nice probabilistic argument. Indeed, for all dx sufficiently small, the trinomial model yields

$$P\{\xi - dx < X_{(i)} \le \xi + dx\} = \frac{n!}{(i-1)!(n-i)!} 2f(\xi)[F(\xi - dx)]^{i-1}[1 - F(\xi + dx)]^{n-i} dx + o(dx), \tag{2.10.5}$$

where $o(dx)$ is a function of dx that approaches zero at a faster rate than dx, i.e., $o(dx)/dx \to 0$ as $dx \to 0$.

Dividing (2.10.5) by $2dx$ and taking the limit as $dx \to 0$, we obtain (2.10.4). The joint p.d.f. of $(X_{(i)}, X_{(j)})$ with $1 \le i < j \le n$ is obtained similarly as

$$\begin{aligned} f_{(i),(j)}(x, y) = {} & \frac{n!}{(i-1)!(j-1-i)!(n-j)!} f(x) f(y)[F(x)]^{i-1} \cdot \\ & [F(y) - F(x)]^{j-1-i} \cdot [1 - F(y)]^{n-j}, \\ & -\infty < x < y < \infty. \end{aligned} \tag{2.10.6}$$

In a similar fashion we can write the joint p.d.f. of any set of order statistics. From the joint p.d.f.s of order statistics we can derive the distribution of various functions of the order statistics. In particular, consider the **sample median** and the **sample range**.

The sample median is defined as

$$M_e = \begin{cases} (X_{(m)} + X_{(m+1)})/2, & \text{if } n = 2m \\ X_{(m+1)}, & \text{if } n = 2m + 1. \end{cases} \tag{2.10.7}$$

That is, half of the sample values are smaller than the median and half of them are greater. The **sample range** R_n is defined as

$$R_n = X_{(n)} - X_{(1)}. \tag{2.10.8}$$

In the case of absolutely continuous independent r.v.s, having a common density $f(x)$, the density $g(x)$ of the sample median is

$$g(x) = \begin{cases} \frac{(2m+1)!}{(m!)^2} f(x)F^m(x)[1-F(x)]^m, & \text{if } n = 2m+1 \\ \frac{(2m)!}{[(m-1)!]^2} \int_{-\infty}^{x} f(u)f(2x-u)F^{m-1}(u)[1-F(2x-u)]^{m-1}du, & \text{if } n = 2m. \end{cases} \tag{2.10.9}$$

We derive now the distribution of the sample range R_n. Starting with the joint p.d.f. of $(X_{(1)}, X_{(n)})$

$$f(x, y) = n(n-1)f(x)f(y)[F(y) - F(x)]^{n-2}, \quad x \le y, \tag{2.10.10}$$

we make the transformation $u = x, r = y - x$.

The Jacobian of this transformation is $J = 1$ and the joint density of (u, r) is

$$g(u, r) = n(n-1)f(u)f(u+r)[F(u+r) - F(u)]^{n-2}. \tag{2.10.11}$$

Accordingly, the density of R_n is

$$h(r) = n(n-1) \int_{-\infty}^{\infty} f(u)f(u+r)[F(u+r) - F(u)]^{n-2}du. \tag{2.10.12}$$

For a comprehensive development of the theory of order statistics and interesting applications, see the books of David (1970) and Gumbel (1958).

2.11 *t*-DISTRIBUTIONS

In many problems of statistical inference, one considers the distribution of the ratio of a statistic, which is normally distributed to its **standard-error** (the square root of its variance). Such ratios have distributions called the t-distributions. More specifically, let $U \sim N(0, 1)$ and $W \sim (\chi^2[\nu]/\nu)^{1/2}$, where U and W are independent. The distribution of U/W is called the "student's t-distribution." We denote this statistic by $t[\nu]$ and say that U/W is distributed as a (central) $t[\nu]$ with ν degrees of freedom.

An example for the application of this distribution is the following. Let $X_1, \ldots, X_n$ be i.i.d. from a $N(\xi, \sigma^2)$ distribution. We have proven that the sample mean $\bar{X}$ is distributed as $N\left(\xi, \frac{\sigma^2}{n}\right)$ and is independent of the sample variance S^2, where $S^2 \sim \sigma^2\chi^2[n-1]/(n-1)$. Hence,

$$\frac{\bar{X} - \xi}{S}\sqrt{n} \sim \frac{N(0, 1)}{(\chi^2[n-1]/(n-1))^{1/2}} \sim t[n-1]. \tag{2.11.1}$$

To find the moments of $t[\nu]$ we observe that, since the numerator and denominator are independent,

$$E\{(t[\nu])^r\} = E\{U^r\} \cdot E\{(\chi^2[\nu]/\nu)^{-r/2}\}. \tag{2.11.2}$$

Thus, all the existing odd moments of $t[\nu]$ are equal to zero, since $E\{U^r\} = 0$ for all $r = 2m + 1$. The existence of $E\{(t[\nu])^r\}$ depends on the existence of $E\{(\chi^2[\nu]/\nu)^{-r/2}\}$. We have

$$E\{(\chi^2[\nu]/\nu)^{-r/2}\} = \left(\frac{\nu}{2}\right)^{r/2} \Gamma\left(\frac{\nu}{2} - \frac{r}{2}\right) / \Gamma\left(\frac{\nu}{2}\right). \tag{2.11.3}$$

Accordingly, a necessary and sufficient condition for the existence of $E\{(t[\nu])^r\}$ is $\nu > r$. Thus, if $\nu > 2$ we obtain that

$$E\{t^2[\nu]\} = \nu/(\nu - 2). \tag{2.11.4}$$

This is also the variance of $t[\nu]$. We notice that $V\{t[\nu]\} \to 1$ as $\nu \to \infty$. It is not difficult to derive the p.d.f. of $t[\nu]$, which is

$$f(t; \nu) = \frac{1}{\sqrt{\nu}\, B(\frac{1}{2}, \frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}, \quad -\infty < t < \infty. \tag{2.11.5}$$

The c.d.f. of $t[\nu]$ can be expressed in terms of the incomplete beta function. Due to the symmetry of the distribution around the origin

$$P\{t[\nu] \leq t\} = 1 - P\{t[\nu] \leq -t\}, \quad t < 0. \tag{2.11.6}$$

We consider now the distribution of $(U + \xi)/W$, where ξ is any real number. This ratio is called the **noncentral** t with ν degrees of freedom, and parameter of noncentrality ξ. This variable is the ratio of two independent random variables namely $N(\xi, 1)$ to $(\chi^2[\nu]/\nu)^{1/2}$. If we denote the noncentral t by $t[\nu; \xi]$, then

$$t[\nu; \xi] \sim (N(0, 1) + \xi)/(\chi^2[\nu]/\nu)^{1/2}. \tag{2.11.7}$$

Since the random variables in the numerator and denominator of (2.11.7) are independent, one obtains

$$E\{t[\nu; \xi]\} = \xi \left(\frac{\nu}{2}\right)^{1/2} \frac{\Gamma(\frac{\nu}{2} - \frac{1}{2})}{\Gamma(\nu/2)}, \tag{2.11.8}$$

and that the central moments of orders 2 and 3 are

$$\mu_2^* = V\{t[\nu;\xi]\} = \frac{\nu}{\nu-2}(1+\xi^2) - \xi^2\frac{\nu}{2}\left(\frac{\Gamma\left(\frac{\nu}{2}-\frac{1}{2}\right)}{\Gamma(\nu/2)}\right)^2 \tag{2.11.9}$$

and

$$\mu_3^* = \xi\left(\frac{\nu}{2}\right)^{1/2}\frac{\Gamma\left(\frac{\nu}{2}-\frac{1}{2}\right)}{\Gamma(\nu/2)}\left(\frac{\nu(2\nu-3+\xi^2)}{(\nu-2)(\nu-3)} - 2\mu_2^*\right). \tag{2.11.10}$$

This shows that the $t[\nu;\xi]$ is not symmetric. Furthermore, since $U+\xi \sim -U+\xi$ we obtain that, for all $-\infty < \xi < \infty$,

$$P\{t[\nu;\xi] \geq t\} = P\{t[\nu;-\xi] \leq -t\}. \tag{2.11.11}$$

In particular, we have seen this in the central case ($\xi = 0$). The formulae of the p.d.f. and the c.d.f. of the noncentral $t[\nu;\xi]$ are quite complicated. There exists a variety of formulae for numerical computations. We shall not present these formulae here; the interested reader is referred to Johnson and Kotz (1969, Ch. 31). In the following section, we provide a representation of these distributions in terms of mixtures of beta distributions.

The univariate t-distribution can be generalized to a multivariate-t in a variety of ways. Consider an m-dimensional random vector $\mathbf{X}$ having a multinomial distribution $N(\boldsymbol{\xi}, \sigma^2 R)$, where R is a correlation matrix. This is the case when all components of $\mathbf{X}$ have the same variance σ^2. Recall that the marginal distribution of

$$Y_i = (X_i - \xi_i) \sim N(0, \sigma^2), \quad i = 1, \ldots, m.$$

Thus, if $S^2 \sim \sigma^2\chi^2[\nu]/\nu$ independently of $Y_1, \ldots, Y_m$, then

$$t_i = \frac{X_i - \xi_i}{S}, \quad i = 1, \ldots, m$$

have the marginal t-distributions $t[\nu]$. The p.d.f. of the multivariate distribution of $\mathbf{t} = \frac{1}{S}\mathbf{Y}$ is given by

$$f(t_1, \ldots, t_m) = \frac{\Gamma\left(\frac{1}{2}(\nu+m)\right)}{(\pi\nu)^{m/2}\Gamma\left(\frac{\nu}{2}\right)|R|^{1/2}}\left(1 + \frac{1}{\nu}\mathbf{t}'R^{-1}\mathbf{t}\right)^{-\frac{\nu+m}{2}}. \tag{2.11.12}$$

Generally, we say that $\mathbf{X}$ has a $t[\nu; \boldsymbol{\xi}, \boldsymbol{\Sigma}]$ distribution if its multivariate p.d.f. is

$$f(x_1, \ldots, x_m; \boldsymbol{\xi}, \boldsymbol{\Sigma}) \propto \left(1 + \frac{1}{\nu}(\mathbf{x} - \boldsymbol{\xi})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\xi})\right)^{-\frac{\nu+m}{2}}. \quad (2.11.13)$$

This distribution has applications in Bayesian analysis, as shown in Chapter 8.

2.12 *F*-DISTRIBUTIONS

The F-distributions are obtained by considering the distributions of ratios of two independent variance estimators based on normally distributed random variables. As such, these distributions have various important applications, especially in the analysis of variance and regression (Section 4.6). We introduce now the F-distributions formally. Let $\chi^2[\nu_1]$ and $\chi^2[\nu_2]$ be two **independent** chi-squared random variables with ν_1 and ν_2 degrees of freedom, respectively. The ratio

$$F[\nu_1, \nu_2] \sim \frac{\chi^2[\nu_1]/\nu_1}{\chi^2[\nu_2]/\nu_2} \quad (2.12.1)$$

is called an F-random variable with ν_1 and ν_2 degrees of freedom. It is a straightforward matter to derive the p.d.f. of $F[\nu_1, \nu_2]$, which is given by

$$f(x; \nu_1, \nu_2) = \frac{\nu_1^{\nu_1/2} \nu_2^{\nu_2/2}}{B\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \cdot \frac{x^{\nu_1/2-1}}{(\nu_2 + \nu_1 x)^{\nu_1/2+\nu_2/2}}. \quad (2.12.2)$$

The cumulative distribution function can be computed by means of the incomplete beta function ratio according to the following formula

$$P\{F[\nu_1, \nu_2] \leq \xi\} = I_{R(\xi)}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right), \quad (2.12.3)$$

where

$$R(\xi) = \xi \frac{\nu_1}{\nu_2} \Big/ \left(1 + \frac{\nu_1}{\nu_2}\xi\right). \quad (2.12.4)$$

In order to derive this formula, we recall that if $G\left(1, \frac{\nu_1}{2}\right)$ and $G\left(1, \frac{\nu_2}{2}\right)$ are two **independent** gamma random variables, then (see Example 2.2)

$$G\left(1, \frac{\nu_1}{2}\right) \Big/ \left[G\left(1, \frac{\nu_1}{2}\right) + G\left(1, \frac{\nu_2}{2}\right)\right] \sim \beta\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right). \quad (2.12.5)$$

Hence,

$$G\left(1, \frac{\nu_1}{2}\right) \Big/ G\left(1, \frac{\nu_2}{2}\right) \sim \frac{1}{\beta\left(\frac{\nu_2}{2}, \frac{\nu_1}{2}\right)} - 1. \tag{2.12.6}$$

We thus obtain

$$\begin{aligned} P\{F[\nu_1, \nu_2] \leq \xi\} &= P\left\{\frac{\nu_2}{\nu_1}\left[\frac{1}{\beta\left(\frac{\nu_2}{2}, \frac{\nu_1}{2}\right)} - 1\right] \leq \xi\right\} \\ &= P\left\{\frac{1}{\beta\left(\frac{\nu_2}{2}, \frac{\nu_1}{2}\right)} \leq 1 + \frac{\nu_1}{\nu_2}\xi\right\} = I_{R(\xi)}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right). \end{aligned} \tag{2.12.7}$$

For testing statistical hypotheses, especially for the analysis of variance and regression, one needs quantiles of the $F[\nu_1, \nu_2]$ distribution. These quantiles are denoted by $F_p[\nu_1, \nu_2]$ and are tabulated in various statistical tables. It is easy to establish the following relationship between the quantiles of $F[\nu_1, \nu_2]$ and those of $F[\nu_2, \nu_1]$, namely,

$$F_\gamma[\nu_1, \nu_2] = 1/F_{1-\gamma}[\nu_2, \nu_1], \quad 0 < \gamma < 1. \tag{2.12.8}$$

The quantiles of the $F[\nu_1, \nu_2]$ distribution can also be determined by those of the beta distribution by employing formula (2.12.5). If we denote by $\beta_\gamma(p, q)$ the values of x for which $I_x(p, q) = \gamma$, we obtain from (2.12.4) that

$$F_\gamma[\nu_1, \nu_2] = \frac{\nu_2}{\nu_1}\beta_\gamma\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right) \Big/ \left[1 - \beta_\gamma\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)\right]. \tag{2.12.9}$$

The moments of $F[\nu_1, \nu_2]$ are obtained in the following manner. For a positive integer r

$$E\{(F[\nu_1, \nu_2])^r\} = \left(\frac{\nu_2}{\nu_1}\right)^r \frac{\Gamma\left(\frac{\nu_2}{2} - r\right)\Gamma\left(\frac{\nu_1}{2} + r\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}. \tag{2.12.10}$$

We realize that the rth moment of $F[\nu_1, \nu_2]$ exists if and only if $\nu_2 > 2r$. In particular,

$$E\{F[\nu_1, \nu_2]\} = \nu_2/(\nu_2 - 2). \tag{2.12.11}$$

Similarly, if $\nu_2 > 4$ then

$$V\{F[\nu_1, \nu_2]\} = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}. \tag{2.12.12}$$

In various occasions one may be interested in an F-like statistic, in which the ratio consists of a noncentral chi-squared in the numerator. In this case the statistic is called a **noncentral** F. More specifically, let $\chi^2[\nu_1; \lambda]$ be a noncentral chi-squared with ν_1 degrees of freedom and a parameter of noncentrality λ. Let $\chi^2[\nu_2]$ be a central chi-squared with ν_2 degrees of freedom, independent of the noncentral chi-squared. Then

$$F[\nu_1, \nu_2; \lambda] \sim \frac{\chi^2[\nu_1; \lambda]/\nu_1}{\chi^2[\nu_2]/\nu_2} \tag{2.12.13}$$

is called a noncentral $F[\nu_1, \nu_2; \lambda]$ statistic. We have proven earlier that $\chi^2[\nu_1; \lambda] \sim \chi^2[\nu_1 + 2J]$, where J has a Poisson distribution with expected value λ. For this reason, we can represent the noncentral $F[\nu_1, \nu_2; \lambda]$ as a mixture of central F statistics.

$$\begin{aligned} F[\nu_1, \nu_2; \lambda] &\sim \frac{\nu_1 + 2J}{\nu_1} \cdot \frac{\chi^2[\nu_1 + 2J]/(\nu_1 + 2J)}{\chi^2[\nu_2]/\nu_2} \\ &\sim \frac{\nu_1 + 2J}{\nu_1} F[\nu_1 + 2J, \nu_2], \end{aligned} \tag{2.12.14}$$

where $J \sim P(\lambda)$. Various results concerning the c.d.f. of $F[\nu_1, \nu_2; \lambda]$, its moments, etc., can be obtained from relationship (2.12.14). The c.d.f. of the noncentral F statistic is

$$P\{F[\nu_1, \nu_2; \lambda] \le \xi\} = e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} P\{F[\nu_1 + 2j, \nu_2] \le \nu_1 \xi/(\nu_1 + 2j)\}. \tag{2.12.15}$$

Furthermore, following (2.12.3) we obtain

$$P\{F[\nu_1, \nu_2; \lambda] \le \xi\} = e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} I_{R(\xi)}\left(\frac{\nu_1}{2} + j, \frac{\nu_2}{2}\right),$$

where

$$R(\xi) = \frac{\nu_1}{\nu_2}\xi \Big/ \left(1 + \frac{\nu_1}{\nu_2}\xi\right).$$

As in the central case, the moments of the noncentral F are obtained by employing the law of the iterated expectation and (2.12.14). Thus,

$$E\{F[\nu_1, \nu_2; \lambda]\} = E\left\{\frac{\nu_1 + 2J}{\nu_1} F[\nu_1 + 2J, \nu_2]\right\}. \tag{2.12.16}$$

However, for all $j = 0, 1, \ldots, E\{F[\nu_1 + 2j, \nu_2]\} = \nu_2/(\nu_2 - 2)$. Hence,

$$\begin{aligned} &E\{F[\nu_1, \nu_2; \lambda]\} = \nu_2(\nu_1 + 2\lambda)/((\nu_1(\nu_2 - 2)), \\ &V\left\{\frac{\nu_1 + 2J}{\nu_1} F[\nu_1 + 2J, \nu_2] \mid J = j\right\} \\ &\quad = \frac{(\nu_1 + 2j)^2}{\nu_1^2} \cdot \frac{2\nu_2^2}{(\nu_2 - 2)^2(\nu_2 - 4)} \cdot \frac{\nu_1 + \nu_2 + 2j - 2}{\nu_1 + 2j}. \end{aligned} \tag{2.12.17}$$

Hence, applying the law of the total variance

$$V\{F[\nu_1, \nu_2; \lambda]\} = \frac{2\nu_2^2(\nu_1 + 2\lambda)(\nu_1 + 6\lambda + \nu_2 - 2)}{\nu_1^2(\nu_2 - 2)^2(\nu_2 - 4)} + \frac{4\lambda\nu_2^2}{\nu_1^2(\nu_2 - 2)^2}. \tag{2.12.18}$$

We conclude the section with the following observation on the relationship between t- and the F-distributions. According to the definition of $t[\nu]$ we immediately obtain that

$$t^2[\nu] \sim N^2(0, 1)/(\chi^2[\nu]/\nu) \sim F[1, \nu]. \tag{2.12.19}$$

Hence,

$$P\{-t \le t[\nu] \le t\} = P\{F[1, \nu] \le t^2\} = I_{t^2/(\nu+t^2)}\left(\frac{1}{2}, \frac{\nu}{2}\right). \tag{2.12.20}$$

Moreover, due to the symmetry of the $t[\nu]$ distribution, for $t > 0$ we have $2P\{t[\nu] \le t\} = 1 + P\{F[1, \nu] \le t^2\}$, or

$$P\{t[\nu] \le t\} = \frac{1}{2}\left(1 + I_{\frac{t^2}{\nu+t^2}}\left(\frac{1}{2}, \frac{\nu}{2}\right)\right). \tag{2.12.21}$$

In a similar manner we obtain a representation for $P\{|t[\nu, \xi]| \le t\}$. Indeed, $(N(0, 1) + \xi)^2 \sim \chi^2[1; \lambda]$ where $\lambda = \frac{1}{2}\xi^2$. Thus, according to (2.12.16)

$$P\{-t \le t[\nu; \xi] \le t\} = e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} I_{t^2/(\nu+t^2)}\left(\frac{1}{2} + j, \frac{\nu}{2}\right). \tag{2.12.22}$$

2.13 THE DISTRIBUTION OF THE SAMPLE CORRELATION

Consider a sample of n i.i.d. vectors $(X_1, Y_1), \ldots, (X_n, Y_n)$ that have a common bivariate normal distribution

$$N\left(\begin{pmatrix}\xi\\ \eta\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_{12} & \sigma_2^2\end{pmatrix}\right).$$

In this section we develop the distributions of the following sample statistics.

(i) The **sample correlation coefficient**

$$r = SPD_{XY}/(SSD_X \cdot SSD_Y)^{1/2}; \tag{2.13.1}$$

(ii) The **sample coefficient of regression**

$$b = SPD_{XY}/SSD_X \tag{2.13.2}$$

where

$$\begin{aligned} SSD_X &= \sum_{i=1}^{n}(X_i - \bar{X})^2 = \mathbf{X}'\left(I - \frac{1}{n}J\right)\mathbf{X},\\ SPD_{XY} &= \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \mathbf{Y}'\left(I - \frac{1}{n}J\right)\mathbf{X},\\ SSD_Y &= \sum_{i=1}^{n}(Y_i - Y)^2 = \mathbf{Y}'\left(I - \frac{1}{n}J\right)\mathbf{Y}. \end{aligned} \tag{2.13.3}$$

As mentioned earlier, the joint density of (X, Y) can be written as

$$f(x, y) = \frac{1}{\sigma_1\sigma_2\sqrt{1-\rho^2}}\phi\left(\frac{x-\xi}{\sigma_1}\right)\phi\left(\frac{y-\eta-\beta(x-\xi)}{\sigma_2\sqrt{1-\rho^2}}\right), \tag{2.13.4}$$

where $\beta = \rho\sigma_2/\sigma_1$. Hence, if we make the transformation

$$\begin{aligned} U_i &= X_i - \xi,\\ V_i &= Y_i - \eta - \beta(X_i - \xi), \quad i = 1, \ldots, n \end{aligned} \tag{2.13.5}$$

then U_i and V_i are **independent** random variables, $U_i \sim N(0, \sigma_1^2)$ and $V_i \sim N(0, \sigma_2^2 \cdot (1-\rho^2))$. We consider now the distributions of the variables

$$\begin{aligned} W_1 &= SPD_{UV}/\sigma_2[(1-\rho^2)SSD_U]^{1/2}, \\ W_2 &= (SSD_V - SPD_{UV}^2/SSD_U)\sigma_2^2(1-\rho^2), \\ W_3 &= SSD_U/\sigma_1^2, \end{aligned} \tag{2.13.6}$$

where SSD_U, SPD_{UV} and SSD_V are defined as in (2.13.3) in terms of (U_i, V_i), $i = 1, \ldots, n$. Let $\mathbf{U} = (U_1, \ldots, U_n)'$ and $\mathbf{V} = (V_1, \ldots, V_n)'$. We notice that the conditional distribution of $SPD_{UV} = \mathbf{V}'\left(I - \frac{1}{n}J\right)\mathbf{U}$ given $\mathbf{U}$ is the normal $N(0, \sigma_2^2(1-\rho^2) \cdot SSD_U)$. Hence, the conditional distribution of W_1 given $\mathbf{U}$ is $N(0, 1)$. This implies that W_1 is $N(0, 1)$, independently of $\mathbf{U}$. Furthermore, W_1 and W_3 are independent, and $W_3 \sim \chi^2[n-1]$. We consider now the variable W_2. It is easy to check

$$SSD_V - SPD_{UV}^2/SSD_U = \mathbf{V}'\left(A - \frac{1}{SSD_U}A\mathbf{U}\mathbf{U}'A'\right)\mathbf{V}, \tag{2.13.7}$$

where $A = I - \frac{1}{n}J$. A is idempotent and so is $B = A - \frac{1}{SSD_U}A\mathbf{U}\mathbf{U}'A$. Furthermore, the rank of B is $n-2$. Hence, the conditional distribution of $SSD_V - SPD_{VU}^2/SSD_U$ given $\mathbf{U}$ is like that of $\sigma_2^2(1-\rho^2)\chi^2[n-2]$. This implies that the distribution of W_2 is like that of $\chi^2[n-2]$. Obviously W_2 and W_3 are independent. We show now that W_1 and W_2 are independent. Since $SPD_{UV} = \mathbf{V}'A\mathbf{U}$ and since $BA\mathbf{U} = (A - \frac{1}{SSD_U}A\mathbf{U}\mathbf{U}'A)A\mathbf{U} = A\mathbf{U} - \frac{1}{SSD_U}A\mathbf{U} \cdot SSD_U = 0$ we obtain that, for any given $\mathbf{U}$, SPD_{UV} and $SSD_V - SPD_{UV}^2/SSD_U$ are conditionally independent. Moreover, since the conditional distributions of $SPD_{UV}/(SSD_U)^{1/2}$ and of $SSD_V - SPD_{UV}^2/SSD_U$ are independent of $\mathbf{U}$, W_1 and W_2 are independent. The variables W_1, W_2, and W_3 can be written in terms of SSD_X, SPD_{XY}, and SSD_Y in the following manner.

$$\begin{aligned} W_1 &= (SPD_{XY} - \beta SPD_X)/[\sigma_2^2(1-\rho^2)SSD_X]^{1/2}, \\ W_2 &= (SSD_Y - SPD_{XY}^2/SSD_X)/\sigma_2^2(1-\rho^2), \\ W_3 &= SSD_X/\sigma_1^2. \end{aligned} \tag{2.13.8}$$

Or, equivalently,

$$\begin{aligned} W_1 &= \frac{r\sqrt{SSD_Y}}{\sigma_2\sqrt{1-\rho^2}} - \frac{\rho\sqrt{SSD_X}}{\sigma_1\sqrt{1-\rho^2}}, \\ W_2 &= SSD_Y(1-r^2)/\sigma_2^2(1-\rho^2), \\ W_3 &= SSD_X/\sigma_1^2. \end{aligned} \tag{2.13.9}$$

From (2.13.9) one obtains that

$$W_1 = \frac{r}{\sqrt{1-r^2}}\sqrt{W_2} - \frac{\rho}{\sqrt{1-\rho^2}}\sqrt{W_3}. \tag{2.13.10}$$

An immediate conclusion is that, when $\rho = 0$,

$$\frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \sim \frac{N(0,1)}{(\chi^2[n-2]/(n-2))^{1/2}} \sim t[n-2]. \tag{2.13.11}$$

This result has important applications in testing the significance of the correlation coefficient. Generally, one can prove that the p.d.f. of r is

$$f(r;\rho) = \frac{2^{n-3}}{\pi(n-3)!}(1-\rho^2)^{\frac{n-1}{2}}(1-r^2)^{\frac{n-4}{2}} \sum_{j=0}^{\infty} \Gamma^2\left(\frac{n+j-1}{2}\right)\frac{(2\rho r)^j}{j!}. \tag{2.13.12}$$

2.14 EXPONENTIAL TYPE FAMILIES

A family of distribution $\mathcal{F}$, having density functions $f(x;\boldsymbol{\theta})$ with respect to some σ-finite measure μ, is called a k-parameter exponential type family if

$$f(x;\boldsymbol{\theta}) = h(x)A(\boldsymbol{\theta})\exp\{\psi_1(\boldsymbol{\theta})U_1(x) + \cdots + \psi_k(\boldsymbol{\theta})U_k(x)\}, \tag{2.14.1}$$

$-\infty < x < \infty, \boldsymbol{\theta} \in \Theta$. Here $\psi_i(\boldsymbol{\theta}), i = 1, \ldots, k$ are functions of the parameters and $U_i(x), i = 1, \ldots, k$ are functions of the observations.

In terms of the parameters $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_k)'$ and the statistics $\mathbf{U} = (U_1(x), \ldots, U_k(x))'$, the p.d.f of a k-parameter exponential type distribution can be written as

$$f(x;\boldsymbol{\psi}) = h^*(\mathbf{U}(x))\exp\{-K(\boldsymbol{\psi})\}\exp\{\boldsymbol{\psi}'\mathbf{U}(x)\}, \tag{2.14.2}$$

where $K(\boldsymbol{\psi}) = -\log A^*(\boldsymbol{\psi})$. Notice that $h^*(\mathbf{U}(x)) > 0$ for all x on the **support set** of $\mathcal{F}$, namely the closure of the smallest Borel set S, such that $P_{\boldsymbol{\psi}}\{S\} = 1$ for all $\boldsymbol{\psi}$. If $h^*(\mathbf{U}(x))$ does not depend on $\boldsymbol{\psi}$, we say that the exponential type family $\mathcal{F}$ is **regular**. Define the domain of convergence to be

$$\Omega^* = \left\{\boldsymbol{\psi} : \int h^*(\mathbf{U}(x))\exp\{-\boldsymbol{\psi}'\mathbf{U}(x)\}d\mu(x) < \infty\right\}. \tag{2.14.3}$$

The family $\mathcal{F}$ is called **full** if the parameter space Ω coincides with Ω^*. Formula (2.14.2) is called the **canonical** form of the p.d.f.; $\boldsymbol{\psi}$ are called the **canonical** (or **natural**) parameters. The statistics $U_i(x)(i = 1, \ldots, k)$ are called **canonical statistics**.

The family $\mathcal{F}$ is said to be of **order** k if $(1, \psi_1, \ldots, \psi_k)$ are **linearly independent** functions of $\boldsymbol{\theta}$. Indeed if, for example, $\psi_k = \alpha_0 + \sum_{j=1}^{k-1} \alpha_j \psi_j$, for some $\alpha_0, \ldots, \alpha_{k-1}$, which are not all zero, then by the reparametrization to

$$\psi_j' = (1 + \alpha_j)\psi_j, \quad j = 1, \ldots, k-1,$$

we reduce the number of canonical parameters to $k-1$. If $(1, \psi_1, \ldots, \psi_k)$ are linearly independent, the exponential type family is called **minimal**.

The following is an important theorem.

Theorem 2.14.1. *If Equation (2.14.2) is a minimal representation then*

(i) *Ω^* is a convex set, and $K(\boldsymbol{\psi})$ is strictly convex function on Ω^*.*

(ii) *$K(\boldsymbol{\psi})$ is a lower semicontinuous function on $\mathbb{R}^k$, and continuous in the interior of Ω^*.*

For proof, see Brown (1986, p. 19).

Let

$$\lambda(\boldsymbol{\psi}) = \int h^*(\mathbf{U}(x)) \exp\{\boldsymbol{\psi}' U(\mathbf{x})\} d\mu(x). \tag{2.14.4}$$

Accordingly, $\lambda(\boldsymbol{\psi}) = \exp\{K(\boldsymbol{\psi})\}$ or $K(\boldsymbol{\psi}) = \log \lambda(\boldsymbol{\psi})$. $\lambda(\boldsymbol{\psi})$ is an **analytic** function on the interior of Ω^* (see Brown, 1986, p. 32). Thus, $\lambda(\boldsymbol{\psi})$ can be differentiated repeatedly under the integral sign and we have for nonnegative integers l_i, such that $\sum_{i=1}^{k} l_i = l$,

$$\frac{\partial^l}{\prod_{i=1}^{k} \partial \psi_i^{l_i}} \lambda(\boldsymbol{\psi}) = \int \prod_{i=1}^{k} (U_i(x))^{l_i} h^*(\mathbf{U}(x)) \cdot \exp\{\boldsymbol{\psi}' \mathbf{U}(\mathbf{x})\} d\mu(x). \tag{2.14.5}$$

The m.g.f. of the canonical p.d.f. (2.14.2) is, for $\boldsymbol{\psi}$ in Ω^*,

$$\begin{aligned} M(\mathbf{t}; \boldsymbol{\psi}) &= \int h^*(\mathbf{U}) e^{-K(\psi) + (\boldsymbol{\psi} + \mathbf{t})' \mathbf{U}} d\mu^*(\mathbf{U}) \\ &= \exp\{-K(\boldsymbol{\psi}) + K(\boldsymbol{\psi} + \mathbf{t})\} \end{aligned} \tag{2.14.6}$$

for $\mathbf{t}$ sufficiently close to 0. The logarithm of $M(\mathbf{t};\psi)$, the **cumulants generating function**, is given here by

$$K^*(\mathbf{t};\boldsymbol{\psi}) = -K(\boldsymbol{\psi}) + K(\boldsymbol{\psi}+\mathbf{t}). \tag{2.14.7}$$

Accordingly,

$$\begin{aligned} E_\psi\{\mathbf{U}\} &= \nabla K^*(\mathbf{t};\boldsymbol{\psi})\big|_{\mathbf{t}=0} \\ &= \nabla K(\boldsymbol{\psi}), \end{aligned} \tag{2.14.8}$$

where ∇ denotes the gradient vector, i.e.,

$$\nabla K(\boldsymbol{\psi}) = \begin{pmatrix} \dfrac{\partial}{\partial\psi_1} K(\boldsymbol{\psi}) \\ \vdots \\ \dfrac{\partial}{\partial\psi_k} K(\boldsymbol{\psi}) \end{pmatrix}.$$

Similarly, the covariance matrix of $\mathbf{U}$ is

$$V_\psi(\mathbf{U}) = \left(\frac{\partial^2}{\partial\psi_i \partial\psi_j} K(\boldsymbol{\psi}); i, j = 1, \ldots, k \right). \tag{2.14.9}$$

Higher order cumulants can be obtained by additional differentiation of $K(\boldsymbol{\psi})$. We conclude this section with several comments.

1. The marginal distributions of canonical statistics are canonical exponential type distributions.
2. The conditional distribution of a subvector of canonical exponential type statistics, given the other canonical statistics, is also a canonical exponential type distribution.
3. The dimension of Ω^* in a minimal canonical exponential family of order k might be smaller than k. In this case we call $\mathcal{F}$ a **curved** exponential family (Efron, 1975, 1978).

2.15 APPROXIMATING THE DISTRIBUTION OF THE SAMPLE MEAN: EDGEWORTH AND SADDLEPOINT APPROXIMATIONS

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having a distribution, with all required moments existing.

2.15.1 Edgeworth Expansion

The **Edgeworth Expansion** of the distribution of $W_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$, which is developed below, may yield more satisfactory approximation than that of the normal. This expansion is based on the following development.

The p.d.f. of the standard normal distribution, $\phi(x)$, has continuous derivatives of all orders everywhere. By repeated differentiation we obtain

$$\begin{aligned} \phi^{(1)}(x) &= -x\phi(x) \\ \phi^{(2)}(x) &= (x^2 - 1)\phi(x), \end{aligned} \tag{2.15.1}$$

and generally, for $j \geq 1$,

$$\phi^{(j)}(x) = (-1)^j H_j(x)\phi(x), \tag{2.15.2}$$

where $H_j(x)$ is a polynomial of order j, called the **Chebychev–Hermite** polynomial. These polynomials can be obtained recursively by the formula, $j \geq 2$,

$$H_j(x) = xH_{j-1}(x) - (j-1)H_{j-2}(x), \tag{2.15.3}$$

where $H_0(x) \equiv 1$ and $H_1(x) = x$.

From this recursive relation one can prove by induction, that an even order polynomial $H_{2m}(x)$, $m \geq 1$, contains only terms with even powers of x, and an odd order polynomial, $H_{2m+1}(x)$, $n \geq 0$, contains only terms with odd powers of x. One can also show that

$$\int_{-\infty}^{\infty} H_j(x)\phi(x)dx = 0, \quad \text{for all } j \geq 1. \tag{2.15.4}$$

Furthermore, one can prove the orthogonality property

$$\int_{-\infty}^{\infty} H_j(x)H_k(x)\phi(x)dx = \begin{cases} 0, & \text{if } j \neq k \\ j!, & \text{if } j = k. \end{cases} \tag{2.15.5}$$

Thus, the system $\{H_j(x),\ j = 0, 1, \ldots\}$ of Chebychev–Hermite polynomials constitutes an orthogonal base for representing every continuous, integrable function $f(x)$ as

$$f(x) = \sum_{j=0}^{\infty} c_j H_j(x)\phi(x), \tag{2.15.6}$$

where, according to (2.15.5),

$$c_j = \frac{1}{j!}\int_{-\infty}^{\infty} H_j(x)f(x)dx, \quad j \geq 0. \tag{2.15.7}$$

In particular, if $f(x)$ is a p.d.f. of an absolutely continuous distribution, having all moments, then, for all $-\infty < x < \infty$,

$$f(x) = \phi(x) + \sum_{j=1}^{\infty} c_j H_j(x)\phi(x). \tag{2.15.8}$$

Moreover,

$$\begin{aligned} c_1 &= \int x f(x)dx = \mu_1, \\ c_2 &= \frac{1}{2}\int (x^2 - 1) f(x)dx = \frac{1}{2}(\mu_2 - 1), \\ c_3 &= \frac{1}{6}(\mu_3 - 3\mu_1), \end{aligned}$$

etc. If X is a standardized random variable, i.e., $\mu_1 = 0$ and $\mu_2 = \mu_2^* = 1$, then its p.d.f. $f(x)$ can be approximated by the formula

$$f(x) \cong \phi(x) + \frac{1}{6}\mu_3^*(x^3 - 3x)\phi(x) + \frac{1}{24}(\mu_4^* - 3)(x^4 - 6x^2 + 3)\phi(x), \tag{2.15.9}$$

which involves the first four terms of the expansion (2.15.8). For the standardized sample mean $W_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$,

$$\mu_{3,n}^* = E\{W_n^3\} = \frac{\beta_1}{\sqrt{n}}, \tag{2.15.10}$$

and

$$\mu_{4,n}^* = E\{W_n^4\} = 3 + \frac{\beta_2 - 3}{n}, \tag{2.15.11}$$

where β_1 and β_2 are the coefficients of skewness and kurtosis.

The same type of approximation with additional terms is known as the **Edgeworth expansion**. The Edgeworth approximation to the c.d.f. of W_n is

$$\begin{aligned} P\{W_n \le x\} \cong \Phi(x) &- \frac{\beta_1}{6\sqrt{n}}(x^2 - 1)\phi(x) \\ &- \frac{x}{n}\left[\frac{\beta_2 - 3}{24}(x^2 - 3) + \frac{\beta_1^2}{72}(x^4 - 10x^2 + 15)\right]\phi(x). \end{aligned} \tag{2.15.12}$$

The remainder term in this approximation is of a smaller order of magnitude than $\frac{1}{n}$, i.e., $o\left(\frac{1}{n}\right)$. One can obviously expand the distribution with additional terms to obtain a higher order of accuracy. Notice that the standard CLT can be proven by taking limits, as $n \to \infty$, of the two sides of (2.15.12).

We conclude this section with the remark that Equation (2.15.9) could serve to approximate the p.d.f. of any standardized random variable, having a continuous, integrable p.d.f., provided the moments exist.

2.15.2 Saddlepoint Approximation

As before, let $X_1, \ldots, X_n$ be i.i.d. random variables having a common density $f(x)$. We wish to approximate the p.d.f of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Let $M(t)$ be the m.g.f. of t, assumed to exist for all t in $(-\infty, t_0)$, for some $0 < t_0 < \infty$. Let $K(t) = \log M(t)$ be the corresponding cumulants generating function.

We construct a family of distributions $\mathcal{F} = \{f(x, \psi) : -\infty < \psi < t_0\}$ such that

$$f(x; \psi) = f(x) \exp\{\psi x - K(\psi)\}. \tag{2.15.13}$$

The family $\mathcal{F}$ is called an **exponential conjugate** to $f(x)$. Notice that $f(x; 0) = f(x)$, and that $\int_{-\infty}^{\infty} f(x; \psi) d\mu(x) = 1$ for all $\psi < t_0$.

Using the inversion formula for Laplace transforms, one gets the relationship

$$f_{\bar{X}}(x; \psi) = f_{\bar{X}}(x) \cdot \exp\{n(\psi x - K(\psi))\}, \tag{2.15.14}$$

where $f_{\bar{X}}(x; \psi)$ denotes the p.d.f. of the sample mean of n i.i.d. random variables from $f(x; \psi)$. The p.d.f. $f_{\bar{X}}(x; \psi)$ is now approximated by the expansion (2.15.9) with additional terms, and its modification for the standardized mean W_n. Accordingly,

$$f_{\bar{X}_n}(x; \psi) \cong \frac{\sqrt{n}}{\sigma(\psi)} \phi(z) \left[1 + \frac{\rho_3(\psi)}{6\sqrt{n}} H_3(z) + \frac{\rho_4(\psi)}{24n} H_4(z) + \frac{\rho_3^2(\psi)}{72n} H_6(z)\right] \tag{2.15.15}$$

where $\phi(z)$ is the p.d.f. of $N(0, 1)$, $z = \frac{x - \mu(\psi)}{\sigma(\psi)}\sqrt{n}$, $\rho_3(\psi) = \frac{K^{(3)}(\psi)}{(K^{(2)}(\psi))^{3/2}}$, and $\rho_4(\psi) = K^{(4)}(\psi)/(K^{(2)}(\psi))^2$. Furthermore, $\mu(\psi) = K'(\psi)$ and $\sigma^2(\psi) = K^{(2)}(\psi)$.

The objective is to approximate $f_{\bar{X}}(x)$. According to (2.15.14) and (2.15.15), we approximate $f_{\bar{X}}(x)$ by

$$\begin{aligned} f_{\bar{X}}(x) &= f_{\bar{X}}(x; \psi) \exp\{n[K(\psi) - \psi(x)]\} \\ &\cong \frac{\sqrt{n}}{\sigma(\psi)} \phi(z) \Bigg[1 + \frac{\rho_3(\psi)}{6\sqrt{n}} H_3(z) + \frac{\rho_4(\psi)}{24n} H_4(z) \\ &\quad + \frac{\rho_3^2(\psi)}{72n} H_6(z)\Bigg] \exp\{n[K(\psi) - \psi x]\}. \end{aligned} \tag{2.15.16}$$

The approximation is called a **saddlepoint approximation** if we substitute in (2.15.16) $\psi = \hat{\psi}$, where ψ is a point in $(-\infty, t_0)$ that maximizes $f(x;\psi)$. Thus, $\hat{\psi}$ is the root of the equation

$$K'(\psi) = x.$$

As we have seen in Section 2.14, $K(\psi)$ is strictly convex in the interior of $(-\infty, t_0)$. Thus, $K'(\psi)$ is strictly increasing in $(-\infty, t_0)$. Thus, if $\hat{\psi}$ exists then it is unique. Moreover, the value of z at $\psi = \hat{\psi}$ is $z = 0$. It follows that the saddlepoint approximation is

$$f_{\bar{X}}(x) = \frac{\sqrt{n}\, c}{(2\pi K^{(2)}(\hat{\psi}))^{1/2}} \exp\{n[K(\hat{\psi}) - \hat{\psi}x]\} \cdot \left\{1 + \frac{1}{n}\left[\frac{\rho_4(\hat{\psi})}{8} - \frac{5}{24}\rho_3^2(\hat{\psi})\right] + O\left(\frac{1}{n^2}\right)\right\}. \tag{2.15.17}$$

The coefficient c is introduced on the right-hand side of (2.15.17) for normalization. A lower order approximation is given by the formula

$$f_{\bar{X}}(x) \cong \frac{\sqrt{n}\, c}{(2\pi K^{(2)}(\hat{\psi}))^{1/2}} \exp\{n[K(\hat{\psi}) - \hat{\psi}x]\}. \tag{2.15.18}$$

The saddelpoint approximation to the tail of the c.d.f., i.e., $P\{\bar{X}_n \geq \xi\}$ is known to yield very accurate results. There is a famous Lugannani–Rice (1980) approximation to this tail probability. For additional reading, see Barndorff-Nielson and Cox (1979), Jensen (1995), Field and Ronchetti (1990), Reid (1988), and Skovgaard (1990).

PART II: EXAMPLES

Example 2.1. In this example we provide a few important results on the distributions of sums of **independent** random variables.

A. **Binomial**

If X_1 and X_2 are independent, $X_1 \sim B(N_1, \theta)$, $X_2 \sim B(N_2, \theta)$, then $X_1 + X_2 \sim B(N_1 + N_2, \theta)$. It is essential that the binomial distributions of X_1 and X_2 will have the same value of θ. The proof is obtained by multiplying the corresponding m.g.f.s.

B. **Poisson**

If $X_1 \sim P(\lambda_1)$ and $X_2 \sim P(\lambda_2)$ then, under independence, $X_1 + X_2 \sim P(\lambda_1 + \lambda_2)$.

C. **Negative-Binomial**

If $X_1 \sim NB(\psi, \nu_1)$ and $X_2 \sim NB(\psi, \nu_2)$ then, under independence, $X_1 + X_2 \sim NB(\psi, \nu_1 + \nu_2)$. It is essential that the two distributions will depend on the same ψ.

D. **Gamma**

If $X_1 \sim G(\lambda, \nu_1)$ and $X_2 \sim G(\lambda, \nu_2)$ then, under independence, $X_1 + X_2 \sim G(\lambda, \nu_1 + \nu_2)$. It is essential that the two values of the parameter λ will be the same. In particular,

$$\chi_1^2[\nu_1] + \chi_2^2[\nu_2] \sim \chi^2[\nu_1 + \nu_2]$$

for all $\nu_1, \nu_2 = 1, 2, \ldots$; where $\chi_i^2[\nu_i]$, $i = 1, 2$, denote two independent χ^2-random variables with ν_1 and ν_2 degrees of freedom, respectively. This result has important applications in the theory of normal regression analysis.

E. **Normal**

If $X_1 \sim N(\mu, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ and if X_1 and X_2 are independent, then $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. A generalization of this result to the case of possible dependence is given later. ■

Example 2.2. Using the theory of transformations, the following important result is derived. Let X_1 and X_2 be independent,

$$X_1 \sim G(\lambda, \nu_1) \text{ and } X_2 \sim G(\lambda, \nu_2),$$

then the ratio $R = X_1/(X_1 + X_2)$ has a beta distribution, $\beta(\nu_1, \nu_2)$, independent of λ. Furthermore, R and $T = X_1 + X_2$ are independent. Indeed, the joint p.d.f. of X_1 and X_2 is

$$f(X_1, X_2) = \frac{\lambda^{\nu_1+\nu_2}}{\Gamma(\nu_1)\Gamma(\nu_2)} x_1^{\nu_1-1} x_2^{\nu_2-1} \exp\{-\lambda(x_1 + x_2)\}, \quad 0 \le x_1,\ x_2 \le \infty.$$

Consider the transformation

$$\begin{aligned} X_1 &= X_1 \\ T &= X_1 + X_2. \end{aligned}$$

The Jacobian of this transformation is $J(x_1, t) = 1$. The joint p.d.f. of X_1 and T is then

$$g(x_1, t) = \frac{\lambda^{\nu_1+\nu_2}}{\Gamma(\nu_1)\Gamma(\nu_2)} x_1^{\nu_1-1} (t - x_1)^{\nu_2-1} \exp\{-\lambda t\}, \quad 0 \le x_1 \le t \le \infty.$$

We have seen in the previous example that $T = X_1 + X_2 \sim G(\lambda, \nu_1 + \nu_2)$. Thus, the marginal p.d.f. of T is

$$h(t) = \frac{\lambda^{\nu_1+\nu_2}}{\Gamma(\nu_1+\nu_2)} t^{\nu_1+\nu_2-1} e^{-\lambda t}, \quad 0 \le t < \infty.$$

Making now the transformation

$$t = t$$
$$r = x_1/t,$$

we see that the Jacobian is $J(r, t) = t$. Hence, from (2.4.8) and (2.4.9) the joint p.d.f. of r and t is, for $0 \le r \le 1$ and $0 \le t < \infty$,

$$g^*(r, t) = \frac{1}{B(\nu_1, \nu_2)} r^{\nu_1-1}(1-r)^{\nu_2-1} \cdot h(t).$$

This proves that $R \sim \beta(\nu_1, \nu_2)$ and that R and T are independent. ■

Example 2.3. Let (X, λ) be random variables, such that the conditional distribution of X given λ is Poisson with p.d.f.

$$p(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, \ldots$$

and $\lambda \sim G(\nu, \Lambda)$. Hence, the marginal p.d.f. of X is

$$\begin{aligned} p(x) &= \frac{\Lambda^\nu}{\Gamma(\nu)x!} \int_0^\infty \lambda^{x+\nu-1} e^{-\lambda(1+\Lambda)} d\lambda \\ &= \frac{\Gamma(x+\nu)}{\Gamma(\nu)x!} \frac{\Lambda^\nu}{(1+\Lambda)^{\nu+x}}, \quad x = 0, 1, \ldots. \end{aligned}$$

Let $\psi = \frac{1}{1+\Lambda}$, $1-\psi = \frac{\Lambda}{1+\Lambda}$. Then $p(x) = \frac{\Gamma(\nu+x)}{\Gamma(x+1)\Gamma(\nu)}(1-\psi)^\nu \psi^x$. Thus, $X \sim NB(\psi, \nu)$, and we get

$$NB(k; \psi, \nu) = \frac{\Lambda^\nu}{\Gamma(\nu)} \int_0^\infty \lambda^{\nu-1} \left(\sum_{l=0}^k e^{-\lambda} \frac{\lambda^l}{l!} \right) e^{-\lambda\Lambda} d_\lambda.$$

But,

$$\sum_{l=0}^k e^{-\lambda} \frac{\lambda^l}{l!} = 1 - P\{G(1, k+1) \le \lambda\}.$$

Hence,

$$NB(k;\psi,\nu) = 1 - P\{G(1,k+1) \le \frac{1}{\Lambda} G(1,\nu)\}$$

where $G(1,k+1)$ and $G(1,\nu)$ are independent.

Let $R = \dfrac{G(1,\nu)}{G(1,k+1)}$. According to Example 2.2,

$$U = \frac{G(1,\nu)}{G(1,k+1)+G(1,\nu)} \sim \beta(\nu, k+1).$$

But $U \sim \dfrac{R}{1+R}$; hence,

$$\begin{aligned} NB(k;\psi,\nu) &= 1 - P\{R \ge \Lambda\} \\ &= 1 - P\left\{U \ge \frac{\Lambda}{1+\Lambda}\right\} \\ &= P\{U \le 1-\psi\} = I_{1-\psi}(\nu, k+1). \end{aligned}$$

■

Example 2.4. Let $X_1, \ldots, X_n$ be i.i.d. random variables. Consider the linear and the quadratic functions

$$\text{sample mean:} \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\text{sample variance:} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

We compute first the variance of S^2. Notice first that S^2 does not change its value if we substitute $X_i' = X_i - \mu_1$ for X_i $(i = 1, \ldots, n)$. Thus, we can assume that $\mu_1 = 0$ and all moments are central moments.

Write $S^2 = \dfrac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$. Accordingly,

$$V\{S^2\} = \frac{1}{(n-1)^2}\left[V\left\{\sum_{i=1}^{n} X_i^2\right\} + n^2 V\{\bar{X}^2\} - 2n\mathrm{cov}\left(\sum_{i=1}^{n} X_i^2, \bar{X}^2\right)\right].$$

Now, since $X_1, \ldots, X_n$ are i.i.d.,

$$V\left\{\sum_{i=1}^{n} X_i^2\right\} = nV\{X_1^2\} = n(\mu_4 - \mu_2^2).$$

Also,

$$V\{\bar{X}^2\} = E\left\{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^4\right\} - \left(E\left\{\left(\frac{1}{n}\sum X_i\right)^2\right\}\right)^2$$

$$= \frac{1}{n^4} E\left\{\left(\sum_{i=1}^{n} X_i\right)^4\right\} - \frac{1}{n^4}\left(E\left\{\left(\sum X_i^2\right)^2\right\}\right)^2.$$

According to Table 2.3,

$$(1)^4 = [4] + 4[31] + 3[2^2] + 6[21^2] + [1^4].$$

Thus,

$$\left(\sum_{i=1}^{n} X_i\right)^4 = \sum_{i=1}^{n} X_i^4 + 4\sum\sum_{i\neq j} X_i^3 X_j + 3\sum\sum_{i\neq j} X_i^2 X_j^2$$

$$+ 6\sum\sum\sum_{i\neq j\neq k} X_i^2 X_j X_k + \sum\sum\sum\sum_{i\neq j\neq k\neq l} X_i X_j X_k X_l.$$

Therefore, since $\mu_1 = 0$, the independence implies that

$$E\left\{\left(\sum_{i=1}^{n} X_i\right)^4\right\} = n\mu_4 + 3n(n-1)\mu_2^2.$$

Also,

$$E\left\{\left(\sum_{i=1}^{n} X_i\right)^2\right\} = n\mu_2.$$

Thus,

$$n^2 V\{\bar{X}^2\} = \frac{1}{n^2}[n\mu_4 + 3n(n-1)\mu_2^2 - n^2\mu_2^2]$$

$$= \frac{1}{n}[\mu_4 + (2n-1)\mu_2^2].$$

At this stage we have to compute

$$\operatorname{cov}\left(\sum_{i=1}^{n} X_i^2, \bar{X}^2\right) = \frac{1}{n^2} E\left\{\left(\sum_{i=1}^{n} X_i^2\right)\left(\sum_{i=1}^{n} X_i\right)^2\right\} - E\left\{\sum_{i=1}^{n} X_i^2\right\} E\{\bar{X}^2\}.$$

From Table (2.3), $(2)(1)^2 = [4] + 2[31] + [2^2] + [21^2]$. Hence,

$$E\left\{\left(\sum_{i=1}^{n} X_i^2\right)\left(\sum_{i=1}^{n} X_i\right)^2\right\} = n\mu_4 + n(n-1)\mu_2^2$$

and

$$E\left\{\sum_{i=1}^{n} X_i^2\right\} E\{\bar{X}^2\} = \mu_2^2.$$

Therefore,

$$-2n \operatorname{cov}\left(\sum_{i=1}^{n} X_i^2, \bar{X}^2\right) = -2(\mu_4 - \mu_2^2).$$

Finally, substituting these terms we obtain

$$V\{S^2\} = \frac{\mu_4 - \mu_2^2}{n-1} - \frac{\mu_4 + \mu_2^2}{n(n-1)}.$$

■

Example 2.5. We develop now the formula for the covariance of $\bar{X}$ and S^2.

$$\begin{aligned}\operatorname{cov}(\bar{X}, S^2) &= \frac{1}{n(n-1)} \operatorname{cov}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2, \sum_{j=1}^{n} X_j\right) \\ &= \frac{1}{n(n-1)}\left[\operatorname{cov}\left(\sum_{i=1}^{n} X_i^2, \sum_{j=1}^{n} X_j\right) - n^2 \operatorname{cov}(\bar{X}^2, \bar{X})\right].\end{aligned}$$

First,

$$\operatorname{cov}\left(\sum_{i=1}^{n} X_i^2, \sum_{j=1}^{n} X_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{cov}(X_i^2, X_j)$$
$$= \sum_{i=1}^{n} \operatorname{cov}(X_i^2, X_i) = n\mu_3,$$

since the independence of X_i and X_j for all $i \neq j$ implies that $\operatorname{cov}(X_i^2, X_j) = 0$. Similarly,

$$\operatorname{cov}(\bar{X}^2, \bar{X}) = E\{\bar{X}^3\}$$
$$= \frac{1}{n^3} E\left\{\left(\sum_{i=1}^{n} X_i\right)^3\right\} = \frac{1}{n^2}\mu_3.$$

Thus, we obtain

$$\operatorname{cov}(\bar{X}, S^2) = \frac{1}{n}\mu_3.$$

Finally, if the distribution function $F(x)$ is symmetric about zero, $\mu_3 = 0$, and $\operatorname{cov}(\bar{X}, S^2) = 0$. ■

Example 2.6. The number of items, N, demanded in a given store during one week is a random variable having a Negative-Binomial distribution NB(ψ, ν); $0 < \psi < 1$ and $0 < \nu < \infty$. These items belong to k different classes. Let $\mathbf{X} = (X_1, \ldots, X_k)'$ denote a vector consisting of the number of items of each class demanded during the week. These are random variables such that $\sum_{i=1}^{k} X_i = N$ and the conditional distribution of $(X_1, \ldots, X_k)$ given N is the multinomial $M(N, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is the vector of probabilities; $0 < \theta_i < 1$, $\sum_{j=1}^{k} \theta_j = 1$. If we observe the $\mathbf{X}$ vectors over many weeks and construct the proportional frequencies of the $\mathbf{X}$ values in the various classes, we obtain an empirical distribution of these vectors. Under the assumption that the model and its parameters remain the same over the weeks we can fit to that empirical distribution the theoretical marginal distribution of $\mathbf{X}$. This marginal distribution is obtained in the following manner.

The m.g.f. of the conditional multinomial distribution of $\mathbf{X}^* = (X_1, \ldots, X_{k-1})'$ given N is

$$M_{\mathbf{X}^*|N}(t_1, \ldots, t_{k-1}) = \left[\sum_{i=1}^{k-1} \theta_i e^{t_i} + \left(1 - \sum_{i=1}^{k-1} \theta_i\right)\right]^N .$$

Hence, the m.g.f. of the marginal distribution of $\mathbf{X}^*$ is

$$M_{\mathbf{X}^*}(t_1, \ldots, t_{k-1}) = (1-\psi)^\nu \sum_{n=0}^{\infty} \frac{\Gamma(\nu+n)}{\Gamma(\nu)\Gamma(n+1)} \cdot$$

$$\left[\psi \sum_{i=1}^{k-1} \theta_i e^{t_i} + \psi\left(1 - \sum_{i=1}^{k-1} \theta_i\right)\right]^n = \left(\frac{1-\psi}{1-\psi\left(\sum_{i=1}^{k-1} \theta_i e^{t_i} + \left(1 - \sum_{i=1}^{k-1} \theta_i\right)\right)}\right)^\nu .$$

Or

$$M_{X^*}(t_1, \ldots, t_{k-1}) = \left(\frac{1 - \sum_{i=1}^{k-1} w_i}{1 - \sum_{i=1}^{k-1} w_i e^{t_i}}\right)^\nu ,$$

where

$$w_i = \frac{\psi\theta_i}{1 - \psi\left(1 - \sum_{i=1}^{k-1} \theta_i\right)}, \quad i = 1, \ldots, k-1.$$

This proves that $\mathbf{X}^*$ has the multivariate Negative-Binomial distribution. ■

Example 2.7. Consider a random variable X having a normal distribution, $N(\xi, \sigma^2)$. Let $\Phi(u)$ be the standard normal c.d.f. The transformed variable $Y = \Phi(X)$ is of interest in various problems of statistical inference in the fields of reliability, quality control, biostatistics, and others. In this example we study the first two moments of Y.

In the special case of $\xi = 0$ and $\sigma^2 = 1$, since $\Phi(u)$ is the c.d.f. of X, the above transformation yields a rectangular random variable, i.e., $Y \sim R(0, 1)$. In this case,

obviously $E\{Y\} = 1/2$ and $V\{Y\} = 1/12$. In the general case, we have according to the law of the iterated expectation

$$\begin{aligned} E\{Y\} &= E\{\Phi(X)\} \\ &= E\{P\{U \leq X \mid X\}\} \\ &= P\{U \leq X\}, \end{aligned}$$

where $U \sim N(0, 1)$, U and X are independent. Moreover, according to (2.7.7), $U - X \sim N(-\xi, 1 + \sigma^2)$. Therefore,

$$E\{Y\} = \Phi(\xi/\sqrt{(1 + \sigma^2)}).$$

In order to determine the variance of Y we observe first that, if U_1, U_2 are independent random variables identically distributed like $N(0, 1)$, then $P\{U_1 \leq x, U_2 \leq x\} = \Phi^2(x)$ for all $-\infty < x < \infty$. Thus,

$$\begin{aligned} E\{Y^2\} &= E\{\Phi^2(X)\} \\ &= P\{U_1 - X \leq 0, U_2 - X \leq 0\}, \end{aligned}$$

where U_1, U_2 and X are independent and $U_i \sim N(0, 1), i = 1, 2, U_1 - X$ and $U_2 - X$ have a joint **bivariate** normal distribution with mean vector $(-\xi, -\xi)$ and covariance matrix

$$V = \begin{pmatrix} 1 + \sigma^2 & \sigma^2 \\ \sigma^2 & 1 + \sigma^2 \end{pmatrix}.$$

Hence,

$$E\{Y^2\} = \Phi_2\left(\frac{\xi}{(1 + \sigma^2)^{1/2}}, \frac{\xi}{(1 + \sigma^2)^{1/2}}; \frac{\sigma^2}{1 + \sigma^2}\right).$$

Finally,

$$V\{Y\} = E\{Y^2\} - \Phi^2\left(\frac{\xi}{(1 + \sigma^2)^{1/2}}\right).$$

Generally, the nth moment of Y can be determined by the n-variate multinormal c.d.f. $\Phi_n\left(\frac{\xi}{(1 + \sigma^2)^{1/2}}, \ldots, \frac{\xi}{(1 + \sigma^2)^{1/2}}; R\right)$, where the correlation matrix R has off-diagonal elements $R_{ij} = \sigma^2/(1 + \sigma^2)$, for all $k \neq j$. We do not treat here the problem of computing the standard k-variate multinormal c.d.f. Computer routines are available for small values of k. The problem of the numerical evaluation is

generally difficult. Tables are available for the bivariate and the trivariate cases. For further comments on this issue see Johnson and Kotz (1972, pp. 83–132). ■

Example 2.8. Let $X_1, X_2, \ldots, X_n$ be i.i.d. $N(0, 1)$ r.v.s. The sample variance is defined as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad \text{where } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Let $Q = \Sigma(X_i - \bar{X})^2$. Define the matrix $J = \mathbf{1}\mathbf{1}'$, where $\mathbf{1}' = (1, \ldots, 1)$ is a vector of ones. Let $A = I - \frac{1}{n}J$, and $Q = \mathbf{X}'A\mathbf{X}$. It is easy to verify that A is an idempotent matrix. Indeed,

$$\left(I - \frac{1}{n}J\right)^2 = I - \frac{2}{n}J + \frac{1}{n^2}J^2 = I - \frac{1}{n}J.$$

The rank of A is $r = n - 1$. Thus, we obtained that $S^2 \sim \frac{1}{n-1}\chi^2[n-1]$. ■

Example 2.9. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a $N(\xi, \sigma^2)$ distribution. The sample mean is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and the sample variance is $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. In Section 2.5 we showed that if the distribution of the Xs is symmetric, then $\bar{X}$ and S^2 are uncorrelated. We prove here the stronger result that, in the normal case, $\bar{X}$ and S^2 are independent. Indeed,

$$\bar{X} \sim \xi + \frac{\sigma}{\sqrt{n}}\mathbf{1}'\mathbf{U}, \quad \text{where } \mathbf{U} = (U_1, \ldots, U_n)'$$

is distributed like $N(0, I)$. Moreover, $S^2 \sim \frac{\sigma^2}{n-1}\mathbf{U}'(I - \frac{1}{n}J)\mathbf{U}$. But,

$$\mathbf{1}'\left(I - \frac{1}{n}J\right) = \mathbf{0}'.$$

This implies the independence of $\bar{X}$ and S^2. ■

Example 2.10. Let $\mathbf{X}$ be a k-dimensional random vector having a multinormal distribution $N(A\beta, \sigma^2 I)$, where A is a $k \times r$ matrix of constants, $\boldsymbol{\beta}$ is an $r \times 1$ vector; $1 \le r \le k$, $0 < \sigma^2 < \infty$. We further assume that rank $(A) = r$, and the parameter

vector $\boldsymbol{\beta}$ is unknown. Consider the vector $\hat{\boldsymbol{\beta}}$ that minimizes the squared-norm $||\mathbf{X} - A\boldsymbol{\beta}||^2$, where $||\mathbf{X}||^2 = \Sigma X_i^2$. Such a vector $\hat{\boldsymbol{\beta}}$ is called the **least-squares estimate** of $\boldsymbol{\beta}$. The vector $\hat{\boldsymbol{\beta}}$ is determined so that

$$||\mathbf{X}||^2 = ||\mathbf{X} - A\hat{\boldsymbol{\beta}}||^2 + ||A\hat{\boldsymbol{\beta}}||^2 \equiv Q_1 + Q_2.$$

That is, $A\hat{\boldsymbol{\beta}}$ is the orthogonal projection of $\mathbf{X}$ on the subspace generated by the column vectors of A. Thus, the inner product of $(\mathbf{X} - A\hat{\boldsymbol{\beta}})$ and $A\hat{\boldsymbol{\beta}}$ should be zero. This implies that

$$\hat{\boldsymbol{\beta}} = (A'A)^{-1}A'\mathbf{X}.$$

The matrix $A'A$ is nonsingular, since A is of full rank. Substituting $\hat{\boldsymbol{\beta}}$ in the expressions for Q_1 and Q_2, we obtain

$$Q_1 = ||\mathbf{X} - A\hat{\boldsymbol{\beta}}||^2 = \mathbf{X}'(I - A(A'A)^{-1}A')\mathbf{X},$$

and

$$Q_2 = ||A\hat{\boldsymbol{\beta}}||^2 = \mathbf{X}'A(A'A)^{-1}A'\mathbf{X}.$$

We prove now that these quadratic forms are independent. Both

$$B_1 = I - A(A'A)^{-1}A' \text{ and } B_2 = A(A'A)^{-1}A'$$

are idempotent. The rank of B_1 is $k - r$ and that of B_2 is r. Moreover,

$$\begin{aligned} B_1B_2 &= (I - A(A'A)^{-1}A')A(A'A)^{-1}A' \\ &= A(A'A)^{-1}A' - A(A'A)^{-1}A' = 0. \end{aligned}$$

Thus, the conditions of Theorem 2.9.2 are satisfied and Q_1 is independent of Q_2. Moreover, $Q_1 \sim \sigma^2\chi^2[x - r; \lambda_1]$ and $Q_2 \sim \sigma^2\chi^2[r; \lambda_2]$ where

$$\lambda_1 = \frac{1}{2}\boldsymbol{\beta}'A'(I - A(A'A)^{-1}A')A\boldsymbol{\beta} = 0$$

and

$$\begin{aligned} \lambda_2 &= \frac{1}{2}\boldsymbol{\beta}'A'A(A'A)^{-1}A'A\boldsymbol{\beta} \\ &= \frac{1}{2}\boldsymbol{\beta}'(A'A)\boldsymbol{\beta}. \end{aligned}$$

■

Example 2.11. Let $X_1, \ldots, X_n$ be i.i.d. random variables from a rectangular $R(0, 1)$ distribution. The density of the ith order statistic is then

$$f_i(x) = \frac{1}{B(i, n-i+1)} x^{i-1}(1-x)^{n-i},$$

$0 \le x \le 1$. The p.d.f. of the sample median, for $n = 2m + 1$, is in this case

$$g_m(x) = \frac{1}{B(m+1, m+1)} x^m (1-x)^m, \quad 0 \le x \le 1.$$

The p.d.f. of the sample range is the $\beta(n-1, 2)$ density

$$h_n(r) = \frac{1}{B(n-1, 2)} r^{n-2}(1-r), \quad 0 \le r \le 1.$$

These results can be applied to test whether a sample of n observation is a realization of i.i.d. random variables having a specified continuous distribution, $F(x)$, since $Y = F(Y) \sim R(0, 1)$. ■

Example 2.12. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having a common exponential distribution $E(\lambda)$, $0 < \lambda < \infty$. Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ be the corresponding order statistic. The density of $X_{(1)}$ is

$$f_{X_{(1)}}(x; \lambda) = \lambda n e^{-\lambda n x}, \quad x \ge 0.$$

The joint density of $X_{(1)}$ and $X_{(2)}$ is

$$f_{X_{(1)}, X_{(2)}}(x, y) = n(n-1)\lambda^2 e^{-\lambda(x+y)} e^{-\lambda(n-2)y}, \quad 0 < x < y.$$

Let $U = X_{(2)} - X_{(1)}$. The joint density of $X_{(1)}$ and U is

$$f_{X_{(1)}, U}(x, u) = \lambda n e^{-\lambda n x} (n-1)\lambda e^{-\lambda(n-1)u}, \quad 0 < x < \infty,$$

and $0 < u < \infty$. Notice that $f_{X_{(1)}, U}(x, u) = f_{X_{(1)}}(x) \cdot f_U(u)$. Thus $X_{(1)}$ and U are independent, and U is distributed like the minimum of $(n-1)$ i.i.d. $E(\lambda)$ random variables. Similarly, by induction on $k = 2, 3, \ldots, n$, if $U_k = X_{(k)} - X_{(k-1)}$ then $X_{(k-1)}$ and U_k are independent and $U_k \sim E(\lambda(n-k+1))$. Thus, since $X_{(k)} = X_{(1)} + U_2 + \cdots + U_k$, $E\{X_{(k)}\} = \frac{1}{\lambda} \sum_{j=n-k+1}^{n} \frac{1}{j}$ and $V\{X_{(k)}\} = \frac{1}{\lambda^2} \sum_{j=n-k+1}^{n} \frac{1}{j^2}$, for all $k \ge 1$. ■

Example 2.13. Let $X \sim N(\mu, \sigma^2)$, $-\infty < \mu < \infty$, $0 < \sigma^2 < \infty$. The p.d.f. of X is

$$f(x; \mu, \sigma^2) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$
$$= \frac{1}{(2\pi)^{1/2}\sigma} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \cdot \exp\left\{\frac{\mu}{\sigma^2}x + \left(-\frac{1}{2\sigma^2}\right)x^2\right\}.$$

Let $h(x) = \frac{1}{\sqrt{2\pi}}$, $A(\mu, \sigma^2) = \exp\left\{-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log\sigma^2\right\}$, $\psi_1(\mu, \sigma^2) = \frac{\mu}{\sigma^2}$, $\psi_2(\mu, \sigma^2) = -\frac{1}{2\sigma^2}$, $U_1(x) = x$, and $U_2(x) = x^2$. We can write $f(x; \mu, \sigma^2)$ as a two-parameter exponential type family. By making the reparametrization $(\mu, \sigma^2) \to (\psi_1, \psi_2)$, the parameter space $\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, 0 < \sigma^2 < \infty\}$ is transformed to the parameter space

$$\Omega = \{(\psi_1, \psi_2) : -\infty < \psi_1 < \infty, \ -\infty < \psi_2 < 0\}.$$

In terms of (ψ_1, ψ_2) the density of X can be written as

$$f(x; \psi_1, \psi_2) = h(x)A(\psi_1, \psi_2)\exp\{\psi_1 x + \psi_2 x^2\}, \quad -\infty < x < \infty,$$

where $h(x) = 1/\sqrt{\pi}$ and

$$A(\psi_1, \psi_2) = \exp\left\{\frac{1}{4}\frac{\psi_1^2}{\psi_2} + \frac{1}{2}\log(-\psi_2)\right\}.$$

The p.d.f. of the standard normal distribution is obtained by substituting $\psi_1 = 0$, $\psi_2 = -\frac{1}{2}$. ■

Example 2.14. A simple example of a curved exponential family is

$$\mathcal{F} = \{N(\xi, c\xi^2), -\infty < \xi < \infty, c > 0 \text{ known}\}.$$

In this case,

$$f(x; \psi_1, \psi_2) = \frac{1}{\sqrt{\pi}} A^*(\psi_1, \psi_2)\exp\{\psi_1 x + \psi_2 x^2\},$$

with $\psi_2 = -\frac{c}{2}\psi_1^2$. ψ_1 and ψ_2 are linearly independent. The rank is $k = 2$ but

$$\Omega^* = \left\{(\psi_1, \psi_2) : \psi_2 = -\frac{c}{2}\psi_1^2, -\infty < \psi_1 < \infty\right\}.$$

The dimension of Ω^* is 1.

The following example shows a more interesting case of a regular exponential family of order $k = 3$. ■

Example 2.15. We consider here a model that is well known as the **Model II of Analysis of Variance**. This model will be discussed later in relation to the problem of estimation and testing **variance components**.

We are given $n \cdot k$ observations on random variables X_{ij} $(i = 1, \ldots, k; j = 1, \ldots, n)$. These random variables represent the results of an experiment performed in k blocks, each block containing n trials. In addition to the random component representing the experimental error, which affects the observations independently, there is also a random effect of the blocks. This block effect is the same on all the observations within a block, but is independent from one block to another. Accordingly, our model is

$$X_{ij} \sim \mu + a_i + e_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n$$

where e_{ij} are i.i.d. like $N(0, \sigma^2)$ and a_i are i.i.d. like $N(0, \tau^2)$.

We determine now the joint p.d.f. of the vector $\mathbf{X} = (X_{11}, \ldots, X_{1n}, X_{21}, \ldots, X_{2n}, \ldots, X_{k1}, \ldots, X_{kn})'$. The conditional distribution of $\mathbf{X}$ given $\mathbf{a} = (a_1, \ldots, a_k)'$ is the multinormal $N(\mu \mathbf{1}_{nk} + \boldsymbol{\xi}(a), \sigma^2 I_{nk})$, where $\boldsymbol{\xi}'(a) = (a_1 \mathbf{1}_n', a_2 \mathbf{1}_n', \ldots, a_k \mathbf{1}_n')$. Hence, the marginal distribution of $\mathbf{X}$ is the multinormal $N(\xi \mathbf{1}_{nk}, V)$, where the covariance matrix V is given by a matrix composed of k equal submatrices along the main diagonal and zeros elsewhere. That is, if $J_n = \mathbf{1}_n \mathbf{1}_n'$ is an $n \times n$ matrix of 1s,

$$V = \text{diag}\{\sigma^2 I_n + \tau^2 J_n, \ldots, \sigma^2 I_n + \tau^2 J_n\}.$$

The determinant of V is $(\sigma^2)^{kn} |I_n + \rho J_n|^k$, where $\rho = \tau^2/\sigma^2$. Moreover, let H be an orthogonal matrix whose first row vector is $\frac{1}{\sqrt{n}} \mathbf{1}_n'$. Then,

$$|I_n + \rho J_n| = |H(I_n + \rho J_n)H'| = (1 + n\rho).$$

Hence, $|V| = \sigma^{2nk}(1 + n\rho)^k$. The inverse of V is

$$V^{-1} = \text{diag}\left\{ \frac{1}{\sigma^2}(I_n + \rho J_n)^{-1}, \ldots, \frac{1}{\sigma^2}(I_n + \rho J_n)^{-1} \right\},$$

where $(I_n + \rho J_n)^{-1} = I_n - (\rho/(1 + n\rho))J_n$.

Accordingly, the joint p.d.f. of $\mathbf{X}$ is

$$f(\mathbf{x};\mu,\sigma^2,\tau^2) = \frac{1}{(2\pi)^{kn/2}|V|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu\mathbf{1}_{nk})'V^{-1}(\mathbf{x}-\mu\mathbf{1}_{nk})\right\}$$
$$= \frac{1}{(2\pi)^{kn/2}\sigma^{kn}(1+n\rho)^{k/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{x}-\mu\mathbf{1}_{nk})'(\mathbf{x}-\mu\mathbf{1}_{nk})\right.$$
$$\left. - \frac{\rho}{2\sigma^2(1+n\rho)}(\mathbf{x}-\mu\mathbf{1}_{nk})' \operatorname{diag}\{J_n,\ldots,J_n\}(\mathbf{x}-\mu\mathbf{1}_{nk})\right\}.$$

Furthermore,

$$(\mathbf{x}-\mu\mathbf{1}_{nk})'(\mathbf{x}-\mu\mathbf{1}_{nk}) = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\mu)^2$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2 + n\sum_{i=1}^{k}(\bar{x}_i-\mu)^2,$$

where $\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n}x_{ij}$, $i=1,\ldots,k$. Similarly,

$$(\mathbf{x}-\mu\mathbf{1}_{nk})' \operatorname{diag}\{J_n,\ldots,J_n\}(\mathbf{x}-\mu\mathbf{1}_{nk}) = n^2\sum_{i=1}^{k}(\bar{x}_i-\mu)^2.$$

Substituting these terms we obtain,

$$f(\mathbf{x};\mu,\sigma^2,\tau^2) = \frac{1}{(2\pi)^{kn/2}\sigma^{nk}(1+n\rho)^{k/2}} \cdot$$
$$\cdot \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2\right.$$
$$\left. - \frac{n}{2\sigma^2(1+n\rho)}\sum_{i=1}^{k}(\bar{x}_i-\bar{\bar{x}})^2 - \frac{nk}{2\sigma^2(1+n\rho)}(\bar{\bar{x}}-\mu)^2\right\},$$

where

$$\bar{\bar{x}} = \frac{1}{nk}\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ij}.$$

Define,

$$U_1(\mathbf{x}) = \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij}^2, \quad U_2(\mathbf{x}) = \sum_{i=1}^{k} \bar{x}_i^2, \quad U_3(\mathbf{x}) = \bar{\bar{x}},$$

and make the reparametrization

$$\psi_1 = -\frac{1}{2\sigma^2}, \quad \psi_2 = \frac{-n^2\rho}{2\sigma^2(1+n\rho)}, \quad \psi_3 = \frac{nk\mu(1+2n\rho)}{2\sigma^2(1+n\rho)}.$$

The joint p.d.f. of $\mathbf{X}$ can be expressed then as

$$f(\mathbf{x}; \psi_1, \psi_2, \psi_3) = \exp\{-nK(\boldsymbol{\psi})\} \cdot \exp\{\psi_1 U_1(\mathbf{x}) + \psi_2 U_2(\mathbf{x}) + \psi_3 U_3(\mathbf{x})\}.$$

The functions $U_1(\mathbf{x})$, $U_2(\mathbf{x})$, and $U_3(\mathbf{x})$ as well as $\psi_1(\theta)$, $\psi_2(\theta)$, and $\psi_3(\theta)$ are linearly independent. Hence, the order is $k = 3$, and the dimension of Ω^* is $d = 3$. ■

Example 2.16. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common gamma distribution $G(\lambda, \nu)$, $0 < \lambda, \nu < \infty$. For this distribution $\beta_1 = 2\nu$ and $\beta_2 = 6\nu$.

The sample mean $\bar{X}_n$ is distributed like $\frac{1}{n} G(\lambda, n\nu)$. The standardized mean is $W_n = \sqrt{n}\, \frac{\lambda \bar{X}_n - \nu}{\sqrt{\nu}}$. The exact c.d.f. of W_n is

$$P\{W_n \le x\} = P\{G(1, n\nu) \le n\nu + x\sqrt{n\nu}\}.$$

On the other hand, the Edgeworth approximation is

$$P\{W_n \le x\} \cong \Phi(x) - \frac{\nu}{3\sqrt{n}}(x^2 - 1)\phi(x) - \frac{x}{n}\left[\frac{2\nu - 1}{8}(x^2 - 3) + \frac{\nu^2}{18}(x^4 - 10x^2 + 15)\right]\phi(x).$$

In the following table, we compare the exact distribution of W_n with its Edgeworth expansion for the case of $\nu = 1$, $n = 10$, and $n = 20$. We see that for $n = 20$ the Edgeworth expansion yields a very good approximation, with a maximal relative error of -4.5% at $x = -2$. At $x = -1.5$ the relative error is 0.9%. At all other values of x the relative error is much smaller.

Exact Distribution and Edgeworth Approximation

	$n = 10$		$n = 20$	
x	Exact	Edgeworth	Exact	Edgeworth
−2.0	0.004635	0.001627	0.009767	0.009328
−1.5	0.042115	0.045289	0.051139	0.051603
−1.0	0.153437	0.160672	0.156254	0.156638
−0.5	0.336526	0.342605	0.328299	0.328311
0	0.542070	0.542052	0.529743	0.529735
.5	0.719103	0.713061	0.711048	0.711052
1.0	0.844642	0.839328	0.843086	0.843361
1.5	0.921395	0.920579	0.923890	0.924262
2.0	0.963145	0.964226	0.966650	0.966527

■

Example 2.17. Let $X_1, \ldots, X_n$ be i.i.d. distributed as $G(\lambda, \nu)$. $\bar{X}_n \sim \frac{1}{n} G(\lambda, n\nu)$. Accordingly,

$$f_{\bar{X}_n}(x) = \frac{(n\lambda)^{n\nu}}{\Gamma(n\nu)} x^{n\nu-1} e^{-n\lambda x}, \quad x > 0.$$

The cumulant generating function of $G(\lambda, \nu)$ is

$$K(\psi) = -\nu \log\left(1 - \frac{\psi}{\lambda}\right), \quad \psi < \lambda.$$

Thus,

$$K'(\psi) = \frac{\nu}{\lambda - \psi}, \quad \psi < \lambda$$

and

$$K''(\psi) = \frac{\nu}{(\lambda - \psi)^2}, \quad \psi < \lambda.$$

Accordingly, $\hat{\psi} = \lambda - \nu/x$ and

$$K''(\hat{\psi}) = x^2/\nu$$

$\exp\{n[K(\hat{\psi}) - \hat{\psi}x]\} = \exp\{n\nu - n\lambda x\} \cdot \left(\frac{\lambda x}{\nu}\right)^{n\nu}$. It follows from Equation (2.15.18) that the saddlepoint approximation is

$$f_{\bar{X}}(x) \cong \frac{\sqrt{n\nu}}{\sqrt{2\pi}} \left(\frac{\lambda}{\nu}\right)^{n\nu} e^{n\nu} \cdot x^{n\nu-1} e^{-n\lambda x}.$$

If we substitute in the exact formula the **Stirling approximation**, $\Gamma(n\nu) \cong \sqrt{2\pi}(n\nu)^{n\nu-\frac{1}{2}}e^{-n\nu}$, we obtain the saddlepoint approximation. ■

PART III: PROBLEMS

Section 2.2

2.2.1 Consider the binomial distribution with parameters $n, \theta, 0 < \theta < 1$.

Write an algorithm for the computation of $b(j \mid n, \theta)$ employing the recursive relationship

$$b(j; n, \theta) = \begin{cases} (1-\theta)^n & j = 0 \\ R_j(n, \theta)b(j-1; n, \theta), & j = 1, \ldots, N \end{cases}$$

where $R_j(n, \theta) = b(j; n, \theta)/b(j-1; n, \theta)$. Write the ratio $R_j(n, \theta)$ explicitly and find an expression for the **mode** of the distribution, i.e., $x_i^0 =$ smallest nonnegative integer for which $b(x^0; n, \theta) \geq b(j; n, \theta)$ for all $j = 0, \ldots, n$.

2.2.2 Prove formula (2.2.2).

2.2.3 Determine the median of the binomial distribution with $n = 15$ and $\theta = .75$.

2.2.4 Prove that when $n \to \infty, \theta \to 0$, but $n\theta \to \lambda, 0 < \lambda < \infty$, then

$$\lim_{\substack{n \to \infty \\ n\theta \to \lambda}} b(i; n, \theta) = p(i; \lambda), \quad i = 0, 1, \ldots$$

where $p(i; \lambda)$ is the p.d.f. of the Poisson distribution.

2.2.5 Establish formula (2.2.7).

2.2.6 Let X have the Pascal distribution with parameters ν (fixed positive integer) and $\theta, 0 < \theta < 1$. Employ the relationship between the Pascal distribution and the negative-binomial distribution to show that the median of X is $\nu + n_{.5}$, where $n_{.5} =$ least nonnegative integer n such that $I_\theta(\nu, n+1) \geq .5$. [This formula of the median is useful for writing a computer program and utilizing the computer's library subroutine function that computes $I_\theta(a, b)$.]

2.2.7 Apply formula (2.2.4) to prove the binomial c.d.f. $B(j; n, \theta)$ is a decreasing function of θ, for each $j = 0, 1, \ldots, n$.

2.2.8 Apply formula (2.2.12) to prove that the c.d.f. of the negative-binomial distribution, $NB(\psi, \nu)$, is strictly decreasing in ψ, for a fixed ν, for each $j = 0, 1, \ldots$.

2.2.9 Let $X \sim B(10^5, .0003)$. Apply the Poisson approximation to compute $P\{20 < X < 40\}$.

Section 2.3

2.3.1 Let U be a random variable having a rectangular distribution $R(0, 1)$. Let $\beta^{-1}(p \mid a, b)$, $0 < p < 1$, $0 < a, b < \infty$ denote the pth quantile of the $\beta(a, b)$ distribution. What is the distribution of $Y = \beta^{-1}(U; a, b)$?

2.3.2 Let X have a gamma distribution $G\left(\frac{1}{\beta}, k\right)$, $0 < \beta < \infty$, and k be a positive integer. Let $\chi_p^2[\nu]$ denote the pth quantile of the chi-squared distribution with ν degrees of freedom. Express the pth quantile of $G\left(\frac{1}{\beta}, k\right)$ in terms of the corresponding quantiles of the chi-squared distributions.

2.3.3 Let Y have the extreme value distribution (2.3.19). Derive formulae for the pth quantile of Y and for its interquartile range.

2.3.4 Let $n(x; \xi, \sigma^2)$ denote the p.d.f. of the normal distribution $N(\xi, \sigma^2)$. Prove that

$$\int_{-\infty}^{\infty} n(x; \xi, \sigma^2) dx = 1,$$

for all (ξ, σ^2); $-\infty < \xi < \infty, 0 < \sigma^2 < \infty$.

2.3.5 Let X have the binomial distribution with $n = 10^5$ and $\theta = 10^{-3}$. For large values of λ ($\lambda > 30$), the $N(\lambda, \lambda)$ distribution provides a good approximation to the c.d.f. of the Poisson distribution $P(\lambda)$. Apply this property to approximate the probability $P\{90 < X < 110\}$.

2.3.6 Let X have an exponential distribution $E(\lambda)$, $0 < \lambda < \infty$. Prove that for all $t > 0$, $E\{\exp\{-tX\}\} \geq \exp\{-t/\lambda\}$.

2.3.7 Let $X \sim R(0, 1)$ and $Y = -\log X$.

(i) Show that $E\{Y\} \geq \log 2$. [The logarithm is on the e base.]

(ii) Derive the distribution of Y and find $E\{Y\}$ exactly.

2.3.8 Determine the first four cumulants of the gamma distribution $G(\lambda, \nu)$, $0 < \lambda, \nu < \infty$. What are the coefficients of skewness and kurtosis?

2.3.9 Derive the coefficients of skewness and kurtosis of the log-normal distribution $LN(\mu, \sigma)$.

2.3.10 Derive the coefficients of skewness and kurtosis of the beta distribution $\beta(p, q)$.

Section 2.4

2.4.1 Let X and Y be independent random variables and $P\{Y \geq 0\} = 1$. Assume also that $E\{|X|\} < \infty$ and $E\left\{\frac{1}{Y}\right\} < \infty$. Apply the Jensen inequality and the law of the iterated expectation to prove that

$$E\left\{\frac{X}{Y}\right\} \geq E\{X\}/E\{Y\}, \quad \text{if } E\{X\} \geq 0,$$

$$\leq E\{X\}/E\{Y\}, \quad \text{if } E\{X\} \leq 0.$$

2.4.2 Prove that if X and Y are positive random variables and $E\{Y \mid X\} = bX$, $0 < b < \infty$, then

(i) $E\{Y/X\} = b$,

(ii) $E\{X/Y\} \geq 1/b$.

2.4.3 Let X and Y be independent random variables. Show that $\text{cov}(X + Y, X - Y) = 0$ if and only if $V\{X\} = V\{Y\}$.

2.4.4 Let X and Y be independent random variables having a common normal distribution $N(0, 1)$. Find the distribution of $R = X/Y$. Does $E\{R\}$ exist?

2.4.5 Let X and Y be independent random variables having a common **log-normal** distribution $LN(\mu, \sigma^2)$, i.e., $\log X \sim \log Y \sim N(\mu, \sigma^2)$.

(i) Prove that $XY \sim LN(2\mu, 2\sigma^2)$.

(ii) Show that $E\{XY\} = \exp\{2\mu + \sigma^2\}$.

(iii) What is $E\{X/Y\}$?

2.4.6 Let X have a binomial distribution $B(n, \theta)$ and let $U \sim R(0, 1)$ independently of X and $Y = X + U$.

(i) Show that Y has an absolutely continuous distribution with c.d.f.

$$F_Y(\eta) = \begin{cases} 0, & \text{if } \eta < 0 \\ (\eta - j)B(j; n, \theta) + (1 - \eta + j)B(j - 1; n, \theta), & \text{if } j \leq \eta < j + 1, \\ & j = 0, 1, \ldots, n \\ 1, & \text{if } \eta \geq n + 1 \end{cases}.$$

(ii) What are $E\{Y\}$ and $V\{Y\}$?

2.4.7 Suppose that the conditional distribution of X given θ is the binomial $B(n, \theta)$. Furthermore, assume that θ is a random variable having a beta distribution $\beta(p, q)$.

(i) What is the marginal p.d.f. of X?

(ii) What is the conditional p.d.f. of θ given $X = x$?

2.4.8 Prove that if the conditional distribution of X given λ is the Poisson $P(\lambda)$, and if λ has the gamma distribution as $G\left(\frac{1}{\tau}, \nu\right)$, then the marginal distribution of X is the negative-binomial $NB\left(\frac{\tau}{1+\tau}, \nu\right)$.

2.4.9 Let X and Y be independent random variables having a common exponential distribution, $E(\lambda)$. Let $U = X + Y$ and $W = X - Y$.

(i) Prove that the conditional distribution of W given U is the rectangular $R(-U, U)$.

(ii) Prove that the marginal distribution of W is the Laplace distribution, with p.d.f.

$$g(\omega; \lambda) = \frac{\lambda}{2} \exp\{-\lambda|\omega|\}, \quad -\infty < \omega < \infty.$$

2.4.10 Let X have a standard normal distribution as $N(0, 1)$. Let $Y = \Phi(X)$. Show that the correlation between X and Y is $\rho = (3/\pi)^{1/2}$. [Although Y is completely determined by X, i.e., $V\{Y \mid X\} = 0$ for all X, the correlation ρ is less than 1. This is due to the fact that Y is a nonlinear function of X.]

2.4.11 Let X and Y be independent standard normal random variables. What is the distribution of the distance of (X, Y) from the origin?

2.4.12 Let X have a $\chi^2[\nu]$ distribution. Let $Y = \delta X$, where $1 < \delta < \infty$. Express the m.g.f. of Y as a weighted average of the m.g.f.s of $\chi^2[\nu + 2j]$, $j = 0, 1, \ldots$, with weights given by $\omega_j = P[J = j]$, where $J \sim NB\left(1 - \frac{1}{\delta}, \frac{\nu}{2}\right)$. [The distribution of $Y = \delta X$ can be considered as an infinite mixture of $\chi^2[\nu + 2J]$ distributions, where J is a random variable having a negative-binomial distribution.]

2.4.13 Let X and Y be independent random variables; $X \sim \chi^2[\nu_1]$ and $Y \sim \delta\chi^2[\nu_2]$, $1 < \delta < \infty$. Use the result of the previous exercise to prove that $X + Y \sim \chi^2[\nu_1 + \nu_2 + 2J]$, where $J \sim NB\left(1 - \frac{1}{\delta}, \frac{\nu_2}{2}\right)$. [Hint: Multiply the m.g.f.s of X and Y or consider the conditional distribution of $X + Y$ given J, where J is independent of X and $Y \mid J \sim \chi^2[\nu_2 + 2J]$, $J \sim NB\left(1 - \frac{1}{\delta}, \frac{\nu_2}{2}\right)$.]

2.4.14 Let $X_1, \ldots, X_n$ be identically distributed independent random variables having a continuous distribution F symmetric around μ. Let $M(X_1, \ldots, X_n)$ and $Q(X_1, \ldots, X_n)$ be two functions of $\mathbf{X} = (X_1, \ldots, X_n)'$ satisfying:

(i) $E\{M(X_1, \ldots, X_n)\} = \mu$;

(ii) $E\{M^2(X_1, \ldots, X_n)\} < \infty$, $E\{Q^2(X_1, \ldots, X_n)\} < \infty$;

(iii) $M(-X_1, \ldots, -X_n) = -M(X_1, \ldots, X_n)$;

(iv) $M(X_1 + c, \ldots, X_n + c) = c + M(X_1, \ldots, X_n)$ for all constants c, $-\infty < c < \infty$;

(v) $Q(X_1 + c, \ldots, X_n + c) = Q(X_1, \ldots, X_n)$, all constants $c, -\infty < c < \infty$;

(vi) $Q(-X_1, \ldots, -X_n) = Q(X_1, \ldots, X_n)$;

then $\operatorname{cov}(M(X_1, \ldots, X_n), Q(X_1, \ldots, X_n)) = 0$.

2.4.15 Let $Y_1, \ldots, Y_{k-1}$ $(k \geq 3)$, $Y_i \geq 0$, $\sum_{i=1}^{k} Y_i \leq 1$, have a joint Dirichlet distribution $\mathcal{D}(\nu_1, \nu_2, \ldots, \nu_k)$, $0 < \nu_i < \infty$, $i = 1, \ldots, k$.

(i) Find the correlation between Y_i and $Y_{i'}$, $i \neq i'$.

(ii) Let $1 \leq m < k$. Show that

$$(Y_1, \ldots, Y_{m-1}) \sim \mathcal{D}\left(\nu_1, \ldots, \nu_{m-1}, \sum_{j=m}^{k} \nu_j\right).$$

(iii) For any $1 \leq m < k$, show the conditional law

$$(Y_1, \ldots, Y_{m-1} \mid Y_m, \ldots, Y_{k-1}) \sim \left(1 - \sum_{j=m}^{k-1} Y_j\right) \mathcal{D}(\nu_1, \ldots, \nu_{m-1}, \nu_k).$$

Section 2.5

2.5.1 Let $X_1, \ldots, X_n$ be i.i.d. random variables. $\hat{\mu}_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$ is the sample moment of order r. Assuming that all the required moments of the common distribution of X_i $(i = 1, \ldots, n)$ exist, find

(i) $E\{\hat{\mu}_r\}$;

(ii) $V\{\hat{\mu}_r\}$;

(iii) $\operatorname{cov}(\hat{\mu}_{r_1}, \hat{\mu}_{r_2})$, for $1 \leq r_1 < r_2$.

2.5.2 Let U_j, $j = 0, \pm 1, \pm 2, \ldots$, be a sequence of independent random variables, such that $E\{U_j\} = 0$ and $V\{U_j\} = \sigma^2$ for all j. Define the random

variables $X_t = \beta_0 + \beta_1 U_{t-1} + \beta_2 U_{t-2} + \cdots + \beta_p U_{t-p} + U_t$, $t = 0, 1, 2, \ldots$, where $\beta_0, \ldots, \beta_p$ are fixed constants. Derive

(i) $E\{X_t\}$, $t = 0, 1, \ldots$;

(ii) $V\{X_t\}$, $t = 0, 1, \ldots$;

(iii) $\text{cov}(X_t, X_{t+h})$, $t = 0, 1, \ldots$; h is a fixed positive integer.

[The sequence $\{X_i\}$ is called an **autoregressive time series** of order p. Notice that $E\{X_t\}$, $V\{X_t\}$, and $\text{cov}(X_t, X_{t+h})$ do not depend on t. Such a series is therefore called **covariance stationary**.]

2.5.3 Let $X_1, \ldots, X_n$ be random variables represented by the model $X_i = \mu_i + e_i$ $(i = 1, \ldots, n)$, where $e_1, \ldots, e_n$ are independent random variables, $E\{e_i\} = 0$ and $V\{e_i\} = \sigma^2$ for all $i = 1, \ldots, n$. Furthermore, let μ_1 be a constant and $\mu_i = \mu_{i-1} + J_i$ $(i = 1, 2, \ldots, n)$, where $J_2, \ldots, J_n$ are independent random variables, $J_i \sim B(1, p)$, $i = 2, \ldots, n$. Let $\mathbf{X} = (X_1, \ldots, X_n)'$. Determine

(i) $E\{\mathbf{X}\}$,

(ii) $\mathbb{\Sigma}(\mathbf{X})$. [The covariance matrix]

2.5.4 Let $X_1, \ldots, X_n$ be i.i.d. random variables. Assume that all required moments exist. Find

(i) $E\{\bar{X}^4\}$.

(ii) $E\{\bar{X}^5\}$.

(iii) $E\{(\bar{X} - \mu)^6\}$.

Section 2.6

2.6.1 Let (X_1, X_2, X_3) have the trinomial distribution with parameters $n = 20$, $\theta_1 = .3, \theta_2 = .6, \theta_3 = .1$.

(i) Determine the joint p.d.f. of (X_2, X_3).

(ii) Determine the conditional p.d.f. of X_1 given $X_2 = 5, X_3 = 7$.

2.6.2 Let $(X_1, \ldots, X_n)$ have a conditional multinomial distribution given N with parameters $N, \theta_1, \ldots, \theta_n$. Assume that N has a Poisson distribution $P(\lambda)$.

(i) Find the (non-conditional) joint distribution of $(X_1, \ldots, X_n)$.

(ii) What is the correlation of X_1 and X_2?

2.6.3 Let (X_1, X_2) have the bivariate negative-binomial distribution $NB(\theta_1, \theta_2, \nu)$.

(i) Determine the correlation coefficient ρ.

(ii) Determine the conditional expectation $E\{X_1 \mid X_2\}$ and the conditional variance $V\{X_1 \mid X_2\}$.

(iii) Compute the coefficient of determination $D^2 = 1 - E\{V\{X_1 \mid X_2\}\}/V\{X_1\}$ and compare D^2 to ρ^2.

2.6.4 Suppose that $(X_1, \ldots, X_k)$ has a k-variate hypergeometric distribution $H(N, M_1, \ldots, M_k, n)$. Determine the expected value and variance of $Y = \sum_{j=1}^{k} \beta_j X_j$, where β_j $(j = 1, \ldots, k)$ are arbitrary constants.

2.6.5 Suppose that $\mathbf{X}$ has a k-variate hypergeometric distribution $H(N, M_1, \ldots, M_k, n)$. Furthermore, assume that $(M_1, \ldots, M_k)$ has a multinomial distribution with parameters N and $(\theta_1, \ldots, \theta_k)$. Derive the (marginal, or expected) distribution of $\mathbf{X}$.

Section 2.7

2.7.1 Let (X, Y) have a bivariate normal distribution with mean vector (ξ, η) and covariance matrix

$$\mathfrak{V} = \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix}.$$

Make a linear transformation $(X, Y) \to (U, W)$ such that (U, W) are independent $N(0, 1)$.

2.7.2 Let (X_1, X_2) have a bivariate normal distribution $N(\boldsymbol{\xi}, \mathfrak{V})$. Define $Y = \alpha X_1 + \beta X_2$.

(i) Derive the formula of $E\{\Phi(Y)\}$.

(ii) What is $V\{\Phi(Y)\}$?

2.7.3 The following is a normal regression model discussed, more generally, in Chapter 5. $(x_1, Y_1), \ldots, (x_n, Y_n)$ are n pairs in which $x_1, \ldots, x_n$ are preassigned constants and $Y_1, \ldots, Y_n$ independent random variables. According to the model, $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$, $i = 1, \ldots, n$. Let $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and

$$\hat{\beta}_n = \sum_{i=1}^{n} (x_i - \bar{x}) Y_i / \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad \hat{\alpha}_n = \bar{Y}_n - \hat{\beta}_n \bar{x}_n.$$

$\hat{\alpha}_n$ and $\hat{\beta}_n$ are called the **least-squares** estimates of α and β, respectively. Derive the joint distribution of $(\hat{\alpha}_n, \hat{\beta}_n)$. [We assume that $\Sigma(x_i - \bar{x})^2 > 0$.]

2.7.4 Suppose that $\mathbf{X}$ is an m-dimensional random vector and $\mathbf{Y}$ is an r-dimensional one, $1 \le r \le m$. Furthermore, assume that the conditional distribution of $\mathbf{X}$

given $\mathbf{Y}$ is $N(A\mathbf{Y}, \Sigma)$ where A is an $m \times r$ matrix of constants. In addition, let $\mathbf{Y} \sim N(\eta, D)$.

(i) Determine the (marginal) joint distribution of $\mathbf{X}$;

(ii) Determine the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$.

2.7.5 Let (Z_1, Z_2) have a standard bivariate normal distribution with coefficient of correlation ρ. Suppose that Z_1 and Z_2 are unobservable and that the observable random variables are

$$X_i = \begin{cases} 0, & \text{if } Z_i \leq 0 \\ 1, & \text{if } Z_i > 0 \end{cases} \qquad i = 1, 2.$$

Let τ be the coefficient of correlation between X_1 and X_2. Prove that $\rho = \sin(\pi\tau/2)$. [τ is called the **tetrachoric** (four-entry) correlation.]

2.7.6 Let (Z_1, Z_2) have a standard bivariate normal distribution with coefficient of correlation $\rho = 1/2$. Prove that $P\{Z_1 \geq 0, Z_2 \leq 0\} = 1/6$.

2.7.7 Let (Z_1, Z_2, Z_3) have a standard trivariate normal distribution with a correlation matrix

$$R = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}.$$

Consider the linear regression of Z_1 on Z_3, namely $\hat{Z}_1 = \rho_{13}Z_3$ and the linear regression of Z_2 on Z_3, namely $\hat{Z}_2 = \rho_{23}Z_3$. Show that the correlation between $Z_1 - \hat{Z}_1$ and $Z_2 - \hat{Z}_2$ is the partial correlation $\rho_{12\cdot 3}$ given by (2.7.15).

2.7.8 Let (Z_1, Z_2) have a standard bivariate normal distribution with coefficient of correlation ρ. Let $\mu_{rs} = E\{Z_1^r Z_2^s\}$ denote the mixed moment of order (r, s). Show that

(i) $\mu_{12} = \mu_{21} = 0$;

(ii) $\mu_{13} = \mu_{31} = 3\rho$;

(iii) $\mu_{22} = 1 + 2\rho^2$;

(iv) $\mu_{14} = \mu_{41} = 0$;

(v) $\mu_{15} = \mu_{51} = 15\rho$;

(vi) $\mu_{24} = \mu_{42} = 3(1 + 4\rho^2)$;

(vii) $\mu_{33} = 3(3 + 2\rho^2)$.

2.7.9 A commonly tabulated function for the standard bivariate normal distribution is the upper orthant probability

$$L(h,k \mid \rho) = P\{h \le Z_1, k \le Z_2\}$$
$$= \frac{1}{2\pi\sqrt{(1-\rho^2)}} \int_h^\infty \int_k^\infty \exp\left\{-\frac{1}{2(1-\rho^2)}(z_1^2 - 2\rho z_1 z_2 + z_2^2)\right\} dz_1 dz_2.$$

Show that

(i) $\Phi_2(h,k;\rho) = 1 - L(h,-\infty \mid \rho) - L(-\infty,k \mid \rho) + L(h,k \mid \rho)$;

(ii) $L(h,k \mid \rho) = L(k,h \mid \rho)$;

(iii) $L(-h,k \mid \rho) + L(h,k \mid -\rho) = 1 - \Phi(k)$;

(iv) $L(-h,-k \mid \rho) - L(h,k \mid \rho) = 1 - \Phi(h) - \Phi(k)$;

(v) $L(0,0 \mid \rho) = \frac{1}{4} + \frac{1}{2\pi}\sin^{-1}(\rho)$.

2.7.10 Suppose that $\mathbf{X}$ has a multinormal distribution $N(\boldsymbol{\xi}, \Sigma)$. Let $X_{(n)} = \max_{1 \le i \le n}\{X_i\}$. Express the c.d.f. of $X_{(n)}$ in terms of the standard multinormal c.d.f., $\Phi_n(\mathbf{Z}; R)$.

2.7.11 Let $\mathbf{Z} = (Z_1, \ldots, Z_m)'$ have a standard multinormal distribution whose correlation matrix R has elements

$$\rho_{ij} = \begin{cases} 1, & \text{if } i = j \\ \lambda_i \lambda_j, & \text{if } i \ne j \end{cases}$$

where $|\lambda_i| \le 1$ $(i = 1, \ldots, m)$. Prove that

$$\Phi_m(h; R) = \int_{-\infty}^{\infty} \phi(u) \prod_{j=1}^{m} \Phi\left(\frac{h_j - \lambda_j u}{\sqrt{1-\lambda_j^2}}\right) du,$$

where $\phi(u)$ is the standard normal p.d.f. [Hint: Let $U_0, U_1, \ldots, U_m$ be independent $N(0,1)$ and let $Z_j = \lambda_j U_0 + \sqrt{(l-\lambda_j^2)}U_j$, $j = 1, \ldots, m$.]

2.7.12 Let $\mathbf{Z}$ have a standard m-dimensional multinormal distribution with a correlation matrix R whose off-diagonal elements are ρ, $0 < \rho < 1$. Show that

$$\Phi_m(\mathbf{0}; R) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2}[1 - \Phi(at)]^m dt,$$

where $0 < a < \infty$, $a^2 = 2\rho/(1-\rho)$. In particular, for $\rho = 1/2$, $\Phi_m(\mathbf{0}; R) = (1+m)^{-1}$.

Section 2.8

2.8.1 Let $X \sim N(0, \sigma^2)$ and $Q = X^2$. Prove that $Q \sim \sigma^2 \chi^2[1]$ by deriving the m.g.f. of Q.

2.8.2 Consider the normal regression model (Problem 3, Section 2.7). The sum of squares of deviation around the fitted regression line is

$$Q_{Y|X} = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2 = (1 - r^2) \sum_{i=1}^{n} (Y_i - \bar{Y})^2,$$

where r is the sample coefficient of correlation, i.e.,

$$r = \sum_{i=1}^{n} Y_i (x_i - \bar{x}) \Big/ \left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \right]^{1/2}.$$

Prove that $Q_{Y|X} \sim \sigma^2 \chi^2[n-2]$.

2.8.3 Let $\{Y_{ij}; i = 1, \ldots, I, j = 1, \ldots, J\}$ be a set of random variables. Consider the following two models (of ANOVA, discussed in Section 4.6.2).

Model I: Y_{ij} are mutually independent, and for each i $(i = 1, \ldots, I) Y_{ij} \sim N(\xi_i, \sigma^2)$ for all $j = 1, \ldots, J$. $\xi_1, \ldots, \xi_I$ are constants.

Model II: For each i $(i = 1, \ldots, I)$ the conditional distribution of Y_{ij} given ξ_i is $N(\xi_i, \sigma^2)$ for all $j = 1, \ldots, J$. Furthermore, given $\xi_1, \ldots, \xi_I$, Y_{ij} are conditionally independent. $\xi_1, \ldots, \xi_I$ are independent random variables having the common distribution $N(0, \tau^2)$. Define the quadratic forms

$$Q_1 = \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \bar{Y}_i)^2, \quad \bar{Y}_i = \frac{1}{J} \sum_{j=1}^{J} Y_{ij}, \quad i = 1, \ldots, I;$$

$$Q_2 = J \sum_{i=1}^{I} (\bar{Y}_i - \bar{\bar{Y}})^2, \quad \bar{\bar{Y}} = \frac{1}{I} \sum_{i=1}^{I} \bar{Y}_i.$$

Determine the distributions of Q_1 and Q_2 under the two different models.

2.8.4 Prove that if X_1 and X_2 are independent and $X_i \sim \chi^2[\nu_i; \lambda_i]$ $i = 1, 2$, then $X_1 + X_2 \sim \chi^2[\nu_1 + \nu_2; \lambda_1 + \lambda_2]$.

Section 2.9

2.9.1 Consider the statistics Q_1 and Q_2 of Problem 3, Section 2.8. Check whether they are independent.

2.9.2 Consider Example 2.5. Prove that the least-squares estimator $\hat{\boldsymbol{\beta}}$ is independent of Q_1.

2.9.3 Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent random vectors having a common bivariate normal distribution with $V\{X\} = V\{Y\} = 1$. Let

$$Q_1 = \frac{1}{1-\rho^2}\left\{\sum_{i=1}^{n}(X_i - \bar{X})^2 - 2\rho\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) + \rho^2\Sigma(Y_i - \bar{Y})^2\right\}$$

and

$$Q_2 = \frac{1}{1-\rho^2}(\bar{X} - \rho\bar{Y})^2,$$

where $\rho, -1 < \rho < 1$, is the correlation between X and Y. Prove that Q_1 and Q_2 are independent. [Hint: Consider the random variables $U_i = X_i - \rho Y_i$, $i = 1, \ldots, n$.]

2.9.4 Let $\mathbf{X}$ be an $n \times 1$ random vector having a multinormal distribution $N(\mu\mathbf{1}, \mathfrak{V})$ where

$$\mathfrak{V} = \sigma^2\begin{pmatrix} 1 & \rho & \ldots\rho \\ \rho & & \\ \vdots & \ddots & \vdots \\ \rho & \ldots & 1 \end{pmatrix} = \sigma^2(1-\rho)I + \sigma^2\rho J,$$

$J = \mathbf{1}\mathbf{1}'$. Prove that $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$ and $Q = \sum_{i=1}^{n}(X_i - \bar{X})^2$ are independent and find their distribution. [Hint: Apply the Helmert orthogonal transformation $\mathbf{Y} = H\mathbf{X}$, where H is an $n \times n$ orthogonal matrix with first row vector equal to $\frac{1}{\sqrt{n}}\mathbf{1}'$.]

Section 2.10

2.10.1 Let $X_1, \ldots, X_n$ be independent random variables having a common exponential distribution $E(\lambda)$, $0 < \lambda < \infty$. Let $X_{(1)} \leq \cdots \leq X_{(n)}$ be the order statistics.

- **(i)** Derive the p.d.f. of $X_{(1)}$.
- **(ii)** Derive the p.d.f. of $X_{(n)}$.
- **(iii)** Derive the joint p.d.f. of $(X_{(1)}, X_{(n)})$.
- **(iv)** Derive the formula for the coefficient of correlation between $X_{(1)}$ and $X_{(n)}$.

2.10.2 Let $X_1, \ldots, X_n$ be independent random variables having an identical continuous distribution $F(x)$. Let $X_{(1)} \leq \cdots \leq X_{(n)}$ be the order statistics. Find the distribution of $U = (F(X_{(n)}) - F(X_{(2)}))/(F(X_{(n)}) - F(X_{(1)}))$.

2.10.3 Derive the p.d.f. of the range $R = X_{(n)} - X_{(1)}$ of a sample of $n = 3$ independent random variables from a common $N(\mu, \sigma^2)$ distribution.

2.10.4 Let $X_1, \ldots, X_n$, where $n = 2m + 1$, be independent random variables having a common rectangular distribution $R(0, \theta)$, $0 < \theta < \infty$. Define the statistics $U = X_{(m)} - X_{(1)}$ and $W = X_{(n)} - X_{(m+1)}$. Find the joint p.d.f. of (U, W) and their coefficient of correlation.

2.10.5 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common continuous distribution symmetric about $x_0 = \mu$. Let $f_{(i)}(x)$ denote the p.d.f. of the ith order statistic, $i = 1, \ldots, n$. Show that $f_{(r)}(\mu + x) = f_{(n-r+1)}(\mu - x)$, all $x, r = 1, \ldots, n$.

2.10.6 Let $X_{(n)}$ be the maximum of a sample of size n of independent identically distributed random variables having a standard exponential distribution $E(1)$. Show that the c.d.f. of $Y_n = X_{(n)} - \log n$ converges, as $n \to \infty$, to $\exp\{-e^{-x}\}$, which is the extreme-value distribution of Type I (Section 2.3.4). [This result can be generalized to other distributions too. Under some general conditions on the distribution of X, the c.d.f. of $X_{(n)} + \log n$ converges to the extreme-value distribution of Type I (Galambos, 1978.)

2.10.7 Suppose that $X_{n,1}, \ldots, X_{n,k}$ are k independent identically distributed random variables having the distribution of the maximum of a random sample of size n from $R(0, 1)$. Let $V = \prod_{i=1}^{k} X_{n,i}$. Show that the p.d.f. of V is (David, 1970, p. 22)

$$g(v) = \frac{n^k}{\Gamma(k)} v^{n-1} (-\log v)^{k-1}, \quad 0 \leq v \leq 1.$$

Section 2.11

2.11.1 Let $X \sim t[10]$. Determine the value of the **coefficient of kurtosis** $\gamma = \mu_4^*/\mu_2^{*2}$.

2.11.2 Consider the normal regression model (Problem 3, Section 2.7 and Problem 2, Section 2.8). The **standard errors** of the least-squares estimates are defined as

$$S.E.\{\hat{\alpha}_n\} = S_{y|x} \left[\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i - \bar{x})^2} \right]^{1/2},$$

$$S.E.\{\hat{\beta}_n\} = S_{y|x}/(\Sigma(x_i - \bar{x})^2)^{1/2},$$

where $S^2_{y|x} = Q_{y|x}/(n-2)$. What are the distributions of $(\hat{\alpha}_n - \alpha)/S.E.\{\hat{\alpha}_n\}$ and of $(\hat{\beta}_n - \beta)/S.E.\{\hat{\beta}_n\}$?

2.11.3 Let $\Phi(u)$ be the standard normal integral and let $X \sim (\chi^2[1])^{1/2}$. Prove that $E\{\Phi(X)\} = 3/4$.

2.11.4 Derive the formulae (2.11.8)–(2.11.10).

2.11.5 Let $\mathbf{X} \sim N(\mu\mathbf{1}, \Sigma)$, with $\Sigma = \sigma^2(1-\rho)\left(I + \frac{\rho}{1-\rho}J\right)$, where $-\frac{1}{n-1} < \rho < 1$ (see Problem 4, Section 2.9). Let $\bar{X}$ and S be the (sample) mean and variance of the components of $\mathbf{X}$.

(i) Determine the distribution of $\bar{X}$.

(ii) Determine the distribution of S^2.

(iii) Prove that $\bar{X}$ and S^2 are independent.

(iv) Derive the distribution of $\sqrt{n}\,(\bar{X} - \mu)/S$.

2.11.6 Let $\mathbf{t}$ have the multivariate t-distribution $t[\nu; \boldsymbol{\xi}, \sigma^2 R]$. Show that the covariance matrix of $\mathbf{t}$ is $\Sigma(\mathbf{t}) = \frac{\nu}{\nu-2}R$, $\nu > 2$.

Section 2.12

2.12.1 Derive the p.d.f. (2.12.2) of $F[\nu_1, \nu_2]$.

2.12.2 Apply formulae (2.2.2) and (2.12.3) to derive the relationship between the binomial c.d.f. and that of the F-distribution, namely

$$B(a; n, \theta) = P\left\{F[2n-2a, 2a+2] \leq \frac{a+1}{N-a} \cdot \frac{1-\theta}{\theta}\right\}, \quad a = 0, \ldots, n-1. \tag{2.15.1}$$

Notice that this relationship can be used to compute the c.d.f. of a central-F distribution with both ν_1 and ν_2 even by means of the binomial distribution. For example, $P\{F[6, 8] \leq 8/3\} = B(3 \mid 6, \frac{1}{3}) = .89986$.

2.12.3 Derive formula (2.12.10).

2.12.4 Apply formula (2.12.15) to express the c.d.f. of $F[2m, 2k; \lambda]$ as a Poisson mixture of binomial distributions.

Section 2.13

2.13.1 Find the expected value and the variance of the sample correlation r when the parameter is ρ.

2.13.2 Show that when $\rho = 0$ then the distribution of the sample correlation r is symmetric around zero.

2.13.3 Express the quantiles of the sample correlation r, when $\rho = 0$, in terms of those of $t[n-2]$.

Section 2.14

2.14.1 Show that the families of binomial, Poisson, negative-binomial, and gamma distributions are exponential type families. In each case, identify the canonical parameters and the natural parameter space.

2.14.2 Show that the family of bivariate normal distributions is a five-parameter exponential type. What are the canonical parameters and the canonical variables?

2.14.3 Let $X \sim N(\mu, \sigma_1^2)$ and $Y \sim N(\mu, \sigma_2^2)$ where X and Y are independent. Show that the joint distribution of (X, Y) is a curved exponential family.

2.14.4 Consider n independent random variables, where $X_i \sim P(\mu(\alpha^i - \alpha^{i-1}))$, (Poisson), $(i = 1, \ldots, n)$. Show that their joint p.d.f. belongs to a two-parameter exponential family. What are the canonical parameters and what are the canonical statistics? The parameter space $\Theta = \{(\mu, \alpha) : 0 < \mu < \infty, 1 < \alpha < \infty\}$.

2.14.5 Let $T_1, T_2, \ldots, T_n$ be i.i.d. random variables having an exponential distribution $E(\lambda)$. Let $0 < t_0 < \infty$. We observed the censored variables $X_1, \ldots, X_n$ where

$$X_i = T_i I\{T_i < t_0\} + t_0 I\{T_i \geq t_0\},$$

$i = 1, \ldots, n$. Show that the joint distribution of $X_1, \ldots, X_n$ is a curved exponential family. What are the canonical statistics?

Section 2.15

2.15.1 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a binomial distribution $B(1, \theta)$. Compare the c.d.f. of $\bar{X}_n$, for $n = 10$ and $\theta = .3$ with the corresponding Edgeworth approximation.

2.15.2 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a Weibull distribution $G^{1/\alpha}(\lambda, 1)$. Approximate the distribution of $\bar{X}_n$ by the Edgeworth approximation for $n = 20, \alpha = 2.5, \lambda = 1$.

2.15.3 Approximate the c.d.f. of the Weibull, $G^{1/\alpha}(\lambda, 1)$, when $\alpha = 2.5$, $\lambda = \frac{1}{20}$ by an Edgeworth approximation with $n = 1$. Compare the exact c.d.f. to the approximation.

2.15.4 Let $\bar{X}_n$ be the sample mean of n i.i.d. random variables having a log-normal distribution, $LN(\mu, \sigma^2)$. Determine the saddlepoint approximation to the p.d.f. of $\bar{X}_n$.

PART IV: SOLUTIONS TO SELECTED PROBLEMS

2.2.2 Prove that $\sum_{j=a}^{n} b(j; n, \theta) = I_\theta(a, n - a + 1)$.

$$
\begin{aligned}
I_\theta(a, n-a+1) &= \frac{1}{B(a, n-a+1)} \int_0^\theta u^{a-1}(1-u)^{n-a} du \\
&= \frac{n!}{(a-1)!(n-a)!} \int_0^\theta u^{a-1}(1-u)^{n-a} du \\
&= \binom{n}{a}\theta^a(1-\theta)^{n-a} + \frac{n!}{a!(n-a-1)!} \int_0^\theta u^a(1-u)^{n-a-1} du \\
&= b(a; n, \theta) + I_\theta(a+1, n-a) \\
&= b(a; n, \theta) + b(a+1; n, \theta) + I_\theta(a+2, n-a-1) \\
&= \cdots = \sum_{j=a}^{N} b(j; n, \theta).
\end{aligned}
$$

2.2.6 $X \sim \text{Pascal}(\theta, \nu)$, i.e.,

$$P\{X = j\} = \binom{j-1}{\nu-1}\theta^\nu(1-\theta)^{j-\nu}, \quad j \geq \nu.$$

Let $k = j - \nu$, $k \geq 0$. Then

$$
\begin{aligned}
P\{X = j\} &= P\{X - \nu = k\} \\
&= \frac{\Gamma(k+\nu)}{\Gamma(\nu)k!}(1-\theta)^k\theta^\nu, \quad k \geq 0.
\end{aligned}
$$

Let $\psi = 1 - \theta$. Then

$$P\{X - \nu = k\} = \frac{\Gamma(k+\nu)}{\Gamma(\nu)k!}\psi^k(1-\psi)^\nu, \quad k \geq 0.$$

Thus, $X - \nu \sim \text{NB}(\psi, \nu)$ or $X \sim \nu + \text{NB}(\psi, \nu)$. The median of X is equal to $\nu+$ the median of $\text{NB}(\psi, \nu)$. Using the formula

$$NB(n \mid \psi, \nu) = I_{1-\psi}(\nu, n+1),$$

median of $\text{NB}(\psi, \nu) =$ least $n \geq 0$ such that $I_\theta(\nu, n+1) \geq 0.5$. Denote this median by $n_{.5}$. Then $X_{.5} = \nu + n_{.5}$.

2.3.1 $U \sim R(0,1)$. $Y \sim \beta^{-1}(U \mid a, b)$. For $0 < y < 1$

$$\begin{aligned} P[Y \leq y] &= P[\beta^{-1}(U \mid a, b) \leq y] \\ &= P\{U \leq \beta(y \mid a, b)\} = \beta(y \mid a, b) \\ &= I_y(a, b). \end{aligned}$$

That is, $\beta^{-1}(U; a, b) \sim \beta(a, b)$.

2.3.2 $\chi^2[\nu] \sim 2G\left(1, \frac{\nu}{2}\right)$.

$$G\left(\frac{1}{\beta}, k\right) \sim \beta G\left(1, \frac{2k}{2}\right) \sim \frac{\beta}{2}\chi[2k].$$

Hence, $G^{-1}\left(p; \frac{1}{\beta}, k\right) = \frac{\beta}{2}\chi_p^2[2k]$.

2.3.9 $X \sim \text{LN}(\mu, \sigma^2)$.

$$\begin{aligned} \mu_1 &= E\{X\} = E\{e^{N(\mu,\sigma^2)}\} = e^{\mu+\sigma^2/2} \\ \mu_2 &= E\{X^2\} = E\{e^{2N(\mu,\sigma^2)}\} = e^{2\mu+2\sigma^2} \\ \mu_3 &= E\{X^3\} = E\{e^{3N(\mu,\sigma^2)}\} = e^{3\mu+\frac{9}{2}\sigma^2} \\ \mu_4 &= E\{X^4\} = E\{e^{4N(\mu,\sigma^2)} = e^{4\mu+8\sigma^2} \\ \mu_2^* &= V\{X\} = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1) \\ \mu_3^* &= \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3 \\ &= e^{3\mu+\frac{9}{2}\sigma^2} - 3e^{3\mu+\frac{5}{2}\sigma^2} + 2e^{3\mu+\frac{3}{2}\sigma^2} \\ &= e^{3\mu+\frac{3}{2}\sigma^2}(e^{3\sigma^2} - 3e^{\sigma^2} + 2) \\ \mu_4^* &= \mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^4 \\ &= e^{4\mu+2\sigma^2}(e^{6\sigma^2} - 4e^{3\sigma^2} + 6e^{\sigma^2} - 3). \end{aligned}$$

Coefficient of skewness $\beta_1 = \dfrac{\mu_3^*}{(\mu_2^*)^{3/2}}$.

$$\beta_1 = \frac{e^{3\sigma^2} - 3e^{\sigma^2} + 2}{(e^{\sigma^2} - 1)^{3/2}}.$$

Coefficient of kurtosis

$$\begin{aligned}\beta_2 &= \frac{\mu_4^*}{(\mu_2^*)^2} \\ &= \frac{e^{6\sigma^2} - 4e^{3\sigma^2} + 6e^{\sigma^2} - 3}{(e^{\sigma^2} - 1)^2}.\end{aligned}$$

2.4.1 X, Y are independent r.v. $P\{Y > 0\} = 1$, $E\{|X|\} < \infty$.

$$E\left\{\frac{X}{Y}\right\} = E\{X\} \cdot E\left\{\frac{1}{Y}\right\}$$

By Jensen's inequality,

$$\begin{aligned}&\geq E\{X\}/E\{Y\}, \quad \text{if} \quad E\{X\} \geq 0 \\ &\leq E\{X\}/E\{Y\}, \quad \text{if} \quad E\{X\} \leq 0.\end{aligned}$$

2.4.4 X, Y are i.i.d. like $N(0, 1)$. Find the distribution of $R = X/Y$.

$$\begin{aligned}F_R(r) &= P\{X/Y \leq r\} \\ &= P\{X \leq rY, Y > 0\} \\ &+ P\{X \geq rY, Y < 0\} \\ &= \int_0^\infty \phi(y)\Phi(ry)dy \\ &+ \int_{-\infty}^0 \phi(y)(1 - \Phi(ry))dy \\ &= \int_0^\infty \phi(y)\Phi(ry)dy \\ &+ \int_0^\infty \phi(y)(1 - \Phi(-ry))dy \\ &= 2\int_0^\infty \phi(y)\Phi(ry)dy \\ &= \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(r).\end{aligned}$$

This is the standard Cauchy distribution with p.d.f.

$$f(r) = \frac{1}{\pi} \cdot \frac{1}{1+r^2}, \quad -\infty < r < \infty.$$

Moments of R do *not* exist.

2.4.9 X, Y i.i.d. $E(\lambda), U = X + Y, W = X - Y, U \sim G(\lambda, 2), f_U(u) = \lambda^2 u e^{-\lambda u}$.

(i) The joint p.d.f. of (X, Y) is $f(x, y) = \lambda^2 e^{-\lambda(x+y)}$, $0 < x, y < \infty$. Make the transformations

$$u = x + y \quad x = \frac{1}{2}(u + w)$$
$$w = x - y \quad y = \frac{1}{2}(u - w).$$

The Jacobian is

$$J = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}.$$

The joint p.d.f. of (U, W) is $g(u, w) = \frac{\lambda^2}{2} e^{-\lambda u}$, $0 < u < \infty$, $|w| < u$. The conditional density of W given U is

$$g(w \mid u) = \frac{\lambda^2 e^{-\lambda u}}{2\lambda^2 u e^{-\lambda u}} = \frac{1}{2u} I\{-u < w < u\}.$$

That is, $W \mid U \sim R(-U, U)$.

(ii) The marginal density of W is

$$f_W(w) = \frac{\lambda^2}{2} \int_{|w|}^{\infty} e^{-\lambda u} du = \frac{\lambda}{2} e^{-\lambda |w|}, \quad -\infty < w < \infty.$$

2.5.4 All moments exist. $\mu_i = E(X^i)$, $i = 1, \ldots, 4$.

(i)

$$
\begin{aligned}
E\{\bar{X}^4\} &= \frac{1}{n^4} E\left\{\left(\sum_i X_i\right)^4\right\} \\
&= \frac{1}{n^4} E\left\{\sum_{i=1}^{n} X_i^4 + 4\sum\sum_{i\neq j} X_i^3 X_j + 3\sum\sum_{i\neq j} X_i^2 X_j^2 \right. \\
&\quad \left. + 6\sum\sum\sum_{i\neq j\neq k} X_i^2 X_j X_k + \sum\sum\sum\sum_{i\neq j\neq k\neq l} X_i X_j X_k X_l\right\} \\
&= \frac{1}{n^4}[n\mu_4 + 4n(n-1)\mu_3\mu_1 + 3n(n-1)\mu_2^2 \\
&\quad + 6n(n-1)(n-2)\mu_2\mu_1^2 + n(n-1)(n-2)(n-3)\mu_1^4] \\
&= \mu_1^4 + \frac{1}{n}(6\mu_2\mu_1^2 - 6\mu_1^4) + \frac{1}{n^2}(4\mu_3\mu_1 + 3\mu_2^2 \\
&\quad - 18\mu_2\mu_1^2 + 11\mu_1^4) + \frac{1}{n^3}(\mu_4 - 4\mu_3\mu_1 - 3\mu_2^2 \\
&\quad + 12\mu_2\mu_1^2 - 6\mu_1^4).
\end{aligned}
$$

(ii)

$$
\begin{aligned}
E\{\bar{X}_5\} &= \frac{1}{n^5} E\left\{\sum_{i=1}^{n} X_i^5 + 5\sum\sum_{i\neq j} X_i^4 X_j \right. \\
&\quad + 10\sum\sum_{i\neq j} X_i^3 X_j^2 + 10\sum\sum\sum_{i\neq j\neq k} X_i^3 X_j X_k + 15\sum\sum\sum_{i\neq j\neq k} X_i^2 X_j^2 X_k \\
&\quad \left. + 10\sum\sum\sum\sum_{i\neq j\neq k\neq l} X_i^2 X_j X_k X_l + \sum\sum\sum\sum\sum_{i\neq j\neq k\neq l\neq m} X_i X_j X_k X_l X_m\right\}.
\end{aligned}
$$

$$
\begin{aligned}
E\{\bar{X}^5\} &= \frac{1}{n^5}\{n\mu_5 + 5n(n-1)\mu_4\mu_1 + 10n(n-1)\mu_3\mu_2 \\
&\quad + 10n(n-1)(n-2)\mu_3\mu_1^2 + 15n(n-1)(n-2)\mu_2^2\mu_1 \\
&\quad + 10n(n-1)(n-2)(n-3)\mu_2\mu_1^3 + n(n-1)(n-2)(n-3)(n-4)\mu_1^5\} \\
&= \mu_1^5 + \frac{1}{n}(10\mu_2\mu_1^3 - 10\mu_1^5) \\
&\quad + \frac{1}{n^2}(10\mu_3\mu_1^2 + 15\mu_2^2\mu_1 - 60\mu_2\mu_1^3 + 35\mu_1^5) \\
&\quad + \frac{1}{n^3}(5\mu_4\mu_1 + 10\mu_3\mu_2 - 30\mu_3\mu_1^2 - 45\mu_2^2\mu_1 \\
&\quad + 110\mu_2\mu_1^3 - 50\mu_1^5) \\
&\quad + \frac{1}{n^4}(\mu_5 - 5\mu_4\mu_1 - 10\mu_3\mu_2 + 20\mu_3\mu_1^2 \\
&\quad + 30\mu_2^2\mu_1 - 60\mu_2\mu_1^3 + 24\mu_1^5).
\end{aligned}
$$

(iii) Assume $\mu_1 = 0$,

$$
\begin{aligned}
E\{(\bar{X}-\mu_1)^6\} &= \frac{1}{n^6}E\left\{\sum_i X_i^6 + 6\sum\sum_{i\neq j} X_i^5 X_j \right.\\
&\quad + 15\sum\sum_{i\neq j} X_i^4 X_j^2 + 15\sum\sum\sum_{i\neq j\neq k} X_i^4 X_j X_k + 10\sum\sum_{i\neq j} X_i^3 X_j^3\\
&\quad + 60\sum\sum\sum_{i\neq j\neq k} X_i^3 X_j^2 X_k + 20\sum\sum\sum\sum_{i\neq j\neq k\neq l} X_i^3 X_j X_k X_l\\
&\quad + 15\sum\sum\sum_{i\neq j\neq k} X_i^2 X_j^2 X_k^2 + 45\sum\sum\sum\sum_{i\neq j\neq k\neq l} X_i^2 X_j^2 X_k X_l\\
&\quad + 15\sum\sum\sum\sum\sum_{i\neq j\neq k\neq l\neq m} X_i^2 X_j X_k X_l X_m\\
&\quad \left. + \sum\sum\sum\sum\sum\sum_{i\neq j\neq k\neq l\neq m\neq n} X_i X_j X_l X_k X_m X_n \right\}\\
&= \frac{1}{n^6}\{n\mu_6^* + 15n(n-1)\mu_4^*\mu_2^* + 10n(n-1)\mu_3^{*2}\\
&\quad + 15n(n-1)(n-2)\mu_2^{*3}\}.
\end{aligned}
$$

$$
\begin{aligned}
E\{(\bar{X}-\mu_1)^6\} &= \frac{1}{n^3}15\mu_2^{*3} + \frac{1}{n^4}(15\mu_4^*\mu_2^* + 10\mu_3^{*2} - 45\mu_2^{*3})\\
&\quad + \frac{1}{n^5}(\mu_6^* - 15\mu_4^*\mu_2^* - 10\mu_3^{*2} + 30\mu_2^{*3}).
\end{aligned}
$$

2.6.3 (X_1, X_2) has a bivariate NB$(\theta_1, \theta_2, \nu)$. The marginals $X_1 \sim \text{NB}\left(\frac{\theta_1}{1-\theta_2}, \nu\right)$, $X_2 \sim \text{NB}\left(\frac{\theta_2}{1-\theta_1}, \nu\right)$.

(i)

$$
\begin{aligned}
\text{cov}(X_1, X_2) &= \frac{\nu\theta_1\theta_2}{(1-\theta_1-\theta_2)^2}\\
V\{X_1\} &= \frac{\nu\theta_1(1-\theta_2)}{(1-\theta_1-\theta_2)^2}\\
V\{X_2\} &= \frac{\nu\theta_2(1-\theta_1}{(1-\theta_1-\theta_2)^2}\\
\rho_{X_1,X_2} &= \sqrt{\frac{\theta_1\theta_2}{(1-\theta_1)(1-\theta_2)}}.
\end{aligned}
$$

(ii) $X_1 \mid X_2 \sim \text{NB}(\theta_1, \nu + X_2)$.

$$E(X_1 \mid X_2) = (\nu + X_2)\frac{\theta_1}{1-\theta_1}$$

$$V(X_1 \mid X_2) = (\nu + X_2)\frac{\theta_1}{(1-\theta_1)^2}.$$

(iii)

$$D^2 = 1 - \frac{E\{V\{X_1 \mid X_2\}\}}{V\{X_1\}}$$

$$E\{V\{X_1 \mid X_2\}\} = \frac{\theta_1}{(1-\theta_1)^2}E\{\nu + X_2\}$$

$$X_2 \sim \text{NB}\left(\frac{\theta_2}{1-\theta_1}, \nu\right)$$

$$E\{X_2\} = \nu\frac{\theta_2}{1-\theta_1-\theta_2}$$

$$\nu + E\{X_2\} = \nu\left(1 + \frac{\theta_2}{1-\theta_1-\theta_2}\right) = \nu\frac{1-\theta_1}{1-\theta_1-\theta_2}$$

$$E\{V\{X_1 \mid X_2\}\} = \nu\frac{\theta_1}{(1-\theta_1)(1-\theta_1-\theta_2)}$$

$$V\{X_1\} = \frac{\nu\theta_1(1-\theta_2)}{(1-\theta_1-\theta_2)^2}$$

$$\frac{E\{V\{X_1 \mid X_2\}\}}{V\{X_1\}} = \frac{1-\theta_1-\theta_2}{(1-\theta_1)(1-\theta_2)}$$

$$D^2 = 1 - \frac{1-\theta_1-\theta_2}{(1-\theta_1)(1-\theta_2)} = \frac{\theta_1\theta_2}{(1-\theta_1)(1-\theta_2)}.$$

Notice that $\rho^2_{X_1,X_2} = D^2$.

2.7.4 **(i)** The conditional m.g.f. of $\mathbf{X}$ given $\mathbf{Y}$ is

$$M_{\mathbf{X}|\mathbf{Y}}(\mathbf{t}) = \exp\left(\mathbf{t}'A\mathbf{Y} + \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}\right).$$

Hence, the m.g.f. of $\mathbf{X}$ is

$$M_{\mathbf{X}}(t) = \exp\left(\frac{1}{2}\mathbf{t}'\Sigma\mathbf{t} + \mathbf{t}'A\eta + \mathbf{t}'ADA'\mathbf{t}\right).$$

Thus, $\mathbf{X} \sim N(A\eta, \Sigma + ADA')$.

(ii) The joint distribution of $\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix}$ is

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix} \sim N\left(\begin{bmatrix} A\eta \\ \eta \end{bmatrix}, \begin{bmatrix} \Sigma + ADA' & AD \\ DA' & D \end{bmatrix}\right).$$

It follows that the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ is

$$\mathbf{Y} \mid \mathbf{X} \sim N(\boldsymbol{\eta} + DA'(\Sigma + ADA')^{-1}(\mathbf{X} - A\boldsymbol{\eta}),\ D - DA'(\Sigma + ADA')^{-1}AD).$$

2.7.5

$$\begin{pmatrix}Z_1\\ Z_2\end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right).$$

$$X_i = \begin{cases}0, & \text{if } Z_i \le 0\\ 1, & \text{if } Z_i > 0.\end{cases} \qquad i = 1, 2.$$

$$E\{X_i\} = P\{X_i = 1\} = P\{Z_i > 0\} = \frac{1}{2}.$$

$$V\{X_i\} = \frac{1}{4}, \quad i = 1, 2.$$

$$cov(X_1, X_2) = E(X_1X_2) - \frac{1}{4}.$$

$$\begin{aligned} E(X_1X_2) &= P(Z_1 > 0, Z_2 > 0)\\ &= \frac{1}{\sqrt{2\pi}}\int_0^\infty e^{-\frac{1}{2}z^2}\Phi\left(\frac{\rho z}{\sqrt{1-\rho^2}}\right)dz\\ &= \frac{1}{2\pi}\int_0^\infty e^{-\frac{1}{2}z^2}\int_{-\infty}^{\frac{\rho z}{\sqrt{1-\rho^2}}} e^{-\frac{1}{2}y^2}dy\,dz\\ &= \frac{1}{4} + \frac{1}{2\pi}\int_0^\infty re^{-\frac{1}{2}r^2}dr\int_0^{\tan^{-1}\left(\frac{\rho}{\sqrt{1-\rho^2}}\right)}d\theta\\ &= \frac{1}{4} + \frac{1}{2\pi}\tan^{-1}\left(\frac{\rho}{\sqrt{1-\rho^2}}\right)\\ &= \frac{1}{4} + \frac{1}{2\pi}\sin^{-1}(\rho). \end{aligned}$$

Hence,

$$cov(X_1, X_2) = \frac{1}{2\pi}\sin^{-1}(\rho).$$

The tetrachoric correlation is

$$\tau = \frac{\frac{1}{2\pi}\sin^{-1}(\rho)}{1/4} = \frac{2}{\pi}\sin^{-1}(\rho)$$

or

$$\rho = \sin\left(\frac{\tau\pi}{2}\right).$$

2.7.11
$$R = \begin{pmatrix} 1 & & & \\ & 1 & & \lambda_i\lambda_j \\ & & \ddots & \\ & \lambda_i\lambda_j & & \\ & & & 1 \end{pmatrix}, \quad |\lambda_j| \le 1, \quad j = 1, \ldots, m.$$

Let $U_0, U_1, \ldots, U_m$ be i.i.d. $N(0,1)$ random variables. Define $Z_j = \lambda_j U_0 + \sqrt{1-\lambda_j^2}\, U_j$, $j = 1, \ldots, m$. Obviously, $E\{Z_j\} = 0$, $V\{Z_j\} = 1$, $j = 1, \ldots, m$, and $\text{cov}(Z_i, Z_j) = \lambda_i\lambda_j$ if $i \neq j$. Finally, since $U_0, \ldots, U_m$ are independent

$$\begin{aligned} P[Z_1 \le h_1, \ldots, Z_m \le h_m] &= P[\lambda_1 U_0 + \sqrt{1-\lambda_1^2}\, U_1 \le h_1, \ldots, \lambda_m U_0 \\ &\quad + \sqrt{1-\lambda_m^2}\, U_m \le h_m] \\ &= \int_{-\infty}^{\infty} \phi(u) \prod_{j=1}^{m} \Phi\left(\frac{h_j - \lambda_j u}{\sqrt{1-\lambda_j^2}}\right) du. \end{aligned}$$

2.10.1 $X_1, \ldots, X_n$ are i.i.d. $E(\lambda)$.

(i) $f_{X_{(1)}}(x) = n\lambda e^{-n\lambda x}$.

(ii) $f_{X_{(2)}}(x) = n\lambda e^{-\lambda x}(1 - e^{-\lambda x})^{n-1}$.

(iii) $f_{X_{(1)}, X_{(n)}}(x, y) = n(n-1)\lambda^2 e^{-\lambda(x+y)}(e^{-\lambda x} - e^{-\lambda y})^{n-2}$, $0 < x < y$.

As shown in Example 2.12,

(iv)
$$E(X_{(1)}) = \frac{1}{n\lambda}$$
$$V(X_{(1)}) = \frac{1}{(n\lambda)^2}$$
$$E\{X_{(n)}\} = \frac{1}{\lambda}\left(1 + \frac{1}{2} + \cdots + \frac{1}{n}\right)$$
$$V\{X_{(n)}\} = \frac{1}{\lambda^2}\sum_{i=1}^{n}\frac{1}{i^2}$$
$$cov(X_{(1)}, X_{(n)}) = E\{X_{(1)}X_{(n)}\} - \frac{1}{n\lambda^2}\sum_{i=1}^{n}\frac{1}{i}$$
$$E\{X_{(1)}X_{(n)}\} = \frac{1}{n\lambda^2}\left(\frac{2}{n} + \sum_{i=1}^{n-1}\frac{1}{i}\right).$$

Thus,

$$cov(X_{(1)}, X_{(n)}) = \frac{1}{n\lambda^2}\left(\frac{2}{n} - \frac{1}{n}\right) = \frac{1}{n^2\lambda^2}.$$

$$\rho_{X_{(1)},X_{(n)}} = \frac{1}{n\left(\sum_{i=1}^{n}\frac{1}{i^2}\right)^{1/2}}.$$

Also, as shown in Example 2.12, $X_{(1)}$ is independent of $X_{(n)} - X_{(1)}$. Thus,

$$cov(X_{(1)}, X_{(n)}) = cov(X_{(1)}, X_{(1)} + (X_{(n)} - X_{(1)})) = V\{X_{(1)}\} = \frac{1}{n^2\lambda^2}.$$

2.10.3 We have to derive the p.d.f. of $R = X_{(3)} - X_{(1)}$, where X_1, X_2, X_3 are i.i.d. $N(\mu, \sigma^2)$. Write $R = \sigma(Z_{(3)} - Z_{(1)})$, where $Z_1, \ldots, Z_3$ are i.i.d. $N(0, 1)$. The joint density of $(Z_{(1)}, Z_{(3)})$ is

$$f_{(1),(3)}(z_1, z_2) = \frac{6}{2\pi}e^{-\frac{1}{2}(z_1^2+z_3^2)}(\Phi(z_3) - \Phi(z_1)),$$

where $-\infty < z_1 < z_3 < \infty$.

Let $U = Z_{(3)} - Z_{(1)}$. The joint density of $(Z_{(1)}, U)$ is

$$g(z, u) = \frac{6}{2\pi}e^{-\frac{1}{4}u^2} \cdot \exp\left(-\left(z + \frac{1}{2}u\right)^2\right)(\Phi(z + u) - \Phi(z)),$$

$-\infty < z < \infty$, $0 < u < \infty$. Thus, the marginal density of U is

$$\begin{aligned} f_U(u) &= \frac{6}{\sqrt{2}\cdot\sqrt{2\pi}}e^{-\frac{1}{4}u^2} \cdot \frac{\sqrt{2}}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(z+\frac{1}{2})^2} \cdot (\Phi(z + u) - \Phi(z))dz \\ &= \frac{6}{2\sqrt{\pi}}e^{-\frac{1}{4}u^2}\left(\Phi\left(\frac{2u-1}{\sqrt{6}}\right) + \Phi\left(\frac{1}{\sqrt{6}}\right) - 1\right). \end{aligned}$$

2.11.3

$$E\{\Phi(X)\} = P\{U \le X\},$$

where U and X are independent. Hence,

$$E\{\Phi(X)\} = P\{t[1] \le 1\} = \frac{3}{4}.$$

CHAPTER 3

Sufficient Statistics and the Information in Samples

PART I: THEORY

3.1 INTRODUCTION

The problem of statistical inference is to draw conclusions from the observed sample on some characteristics of interest of the parent distribution of the random variables under consideration. For this purpose we formulate a **model** that presents our assumptions about the **family** of distributions to which the parent distribution belongs. For example, in an inventory management problem one of the important variables is the number of units of a certain item demanded every period by the customer. This is a random variable with an unknown distribution. We may be ready to assume that the distribution of the demand variable is Negative Binomial $NB(\psi, \nu)$. The statistical model specifies the possible range of the parameters, called the **parameter space**, and the corresponding family of distributions $\mathcal{F}$. In this example of an inventory system, the model may be

$$\mathcal{F} = \{NB(\psi, \nu); 0 < \psi < 1, 0 < \nu < \infty\}.$$

Such a model represents the case where the two parameters, ψ and ν, are unknown. The parameter space here is $\Theta = \{(\psi, \nu); 0 < \psi < 1, 0 < \nu < \infty\}$. Given a sample of n independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$, representing the weekly demand, the question is what can be said on the specific values of ψ and ν from the observed sample?

Every sample contains a certain amount of information on the parent distribution. Intuitively we understand that the larger the number of observations in the sample (on i.i.d. random variables) the more information it contains on the distribution

Examples and Problems in Mathematical Statistics, First Edition. Shelemyahu Zacks.

under consideration. Later in this chapter we will discuss two specific information functions, which are used in statistical design of experiments and data analysis. We start with the investigation of the question whether the sample data can be condensed by computing first the values of certain **statistics** without losing information. If such statistics exist they are called **sufficient statistics**. The term **statistic** will be used to indicate a function of the (observable) random variables that does not involve any function of the unknown parameters. The sample mean, sample variance, the sample order statistics, etc., are examples of statistics. As will be shown, the notion of sufficiency of statistics is strongly dependent on the model under consideration. For example, in the previously mentioned inventory example, as will be established later, if the value of the parameter ν is known, a sufficient statistic is the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. On the other hand, if ν is unknown, the sufficient statistic is the order statistic $(X_{(1)}, \ldots, X_{(n)})$. When ν is unknown, the sample mean $\bar{X}$ by itself does not contain all the information on ψ and ν. In the following section we provide a definition of sufficiency relative to a specified model and give a few examples.

3.2 DEFINITION AND CHARACTERIZATION OF SUFFICIENT STATISTICS

3.2.1 Introductory Discussion

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector having a joint c.d.f. $F_\theta(\mathbf{x})$ belonging to a family $\mathcal{F} = \{F_\theta(x); \theta \in \Theta\}$. Such a random vector may consist of n i.i.d. variables or of dependent random variables. Let $T(\mathbf{X}) = (T_1(\mathbf{X}), \ldots, T_r(\mathbf{X}))'$, $1 \le r \le n$ be a statistic based on $\mathbf{X}$. T could be real ($r = 1$) or vector valued ($r > 1$). The transformations $T_j(\mathbf{X})$, $j = 1, \ldots, r$ are not necessarily one-to-one. Let $f(\mathbf{x}; \theta)$ denote the (joint) probability density function (p.d.f.) of $\mathbf{X}$. In our notation here $T_i(\mathbf{X})$ is a concise expression for $T_i(X_1, \ldots, X_n)$. Similarly, $F_\theta(\mathbf{x})$ and $f(\mathbf{x}; \theta)$ represent the multivariate functions $F_\theta(x_1, \ldots, x_n)$ and $f(x_1, \ldots, x_n; \theta)$. As in the previous chapter, we assume throughout the present chapter that all the distribution functions belonging to the same family are either absolutely continuous, discrete, or mixtures of the two types.

Definition of Sufficiency. *Let $\mathcal{F}$ be a family of distribution functions and let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector having a distribution in $\mathcal{F}$. A statistic $T(\mathbf{X})$ is called sufficient with respect to $\mathcal{F}$ if the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X})$ is the same for all the elements of $\mathcal{F}$.*

Accordingly, if the joint p.d.f. of $\mathbf{X}$, $f(\mathbf{x}; \theta)$, depends on a parameter θ and $T(\mathbf{X})$ is a sufficient statistic with respect to $\mathcal{F}$, the conditional p.d.f. $h(\mathbf{x} \mid t)$ of $\mathbf{X}$ given $\{T(\mathbf{X}) = t\}$ is independent of θ. Since $f(\mathbf{x}; \theta) = h(\mathbf{x} \mid t)g(t; \theta)$, where $g(t; \theta)$ is the p.d.f. of $T(\mathbf{x})$, all the information on θ in $\mathbf{x}$ is summarized in $T(\mathbf{x})$.

The process of checking whether a given statistic is sufficient for some family following the above definition may be often very tedious. Generally the identification of sufficient statistics is done by the application of the following theorem. This celebrated theorem was given first by Fisher (1922) and Neyman (1935). We state the theorem here in terms appropriate for families of absolutely continuous or discrete distributions. For more general formulations see Section 3.2.2. For the purposes of our presentation we require that the family of distributions $\mathcal{F}$ consists of

(i) absolutely continuous distributions; or

(ii) discrete distributions, having jumps on a set of points $\{\xi_1, \xi_2, \ldots\}$ independent of θ, i.e., $\sum_{i=1}^{\infty} p(\xi_i; \theta) = 1$ for all $\theta \in \Theta$; or

(iii) mixtures of distributions satisfying (i) or (ii). Such families of distributions will be called **regular** (Bickel and Doksum, 1977, p. 61).

The families of discrete or absolutely continuous distributions discussed in Chapter 2 are all regular.

Theorem 3.2.1 (The Neyman–Fisher Factorization Theorem). *Let* $\mathbf{X}$ *be a random vector having a distribution belonging to a regular family* $\mathcal{F}$ *and having a joint p.d.f.* $f(x; \theta)$, $\theta \in \Theta$. *A statistic* $T(\mathbf{X})$ *is sufficient for* $\mathcal{F}$ *if and only if*

$$f(\mathbf{x}; \theta) = K(\mathbf{x}) g(T(\mathbf{x}); \theta), \tag{3.2.1}$$

where $K(\mathbf{x}) \geq 0$ *is independent of* θ *and* $g(T(\mathbf{x}); \theta) \geq 0$ *depends on* $\mathbf{x}$ *only through* $T(\mathbf{x})$.

Proof. We provide here a proof for the case of discrete distributions.

(i) Sufficiency:

We show that (3.2.1) implies that the conditional distribution of $\mathbf{X}$ given $\{T(\mathbf{X}) = t\}$ is independent of θ. The (marginal) p.d.f. of $T(\mathbf{X})$ is, according to (3.2.1),

$$\begin{aligned} g^*(t; \theta) &= \sum_{\{\mathbf{x}\}} I\{\mathbf{x}; T(\mathbf{x}) = t\} f(\mathbf{x}; \theta) \\ &= g(t; \theta) \sum_{\{\mathbf{x}\}} I\{\mathbf{x}; T(\mathbf{x}) = t\} K(\mathbf{x}). \end{aligned} \tag{3.2.2}$$

The joint p.d.f. of $\mathbf{X}$ and $T(\mathbf{X})$ is

$$p(\mathbf{x}, t; \theta) = I\{\mathbf{x}; T(\mathbf{x}) = t\} K(\mathbf{x}) g(t; \theta). \tag{3.2.3}$$

Hence, the conditional p.d.f. of $\mathbf{X}$, given $\{T(\mathbf{X}) = t\}$ at every point t such that $g^*(t;\theta) > 0$, is

$$\frac{p(\mathbf{x}, t;\theta)}{g^*(t;\theta)} = \frac{I\{\mathbf{x}; T(\mathbf{x}) = t\}K(\mathbf{x})}{\sum_{\{\mathbf{y}\}} I\{\mathbf{y}; T(\mathbf{y}) = t\}K(\mathbf{y})}. \tag{3.2.4}$$

This proves that $T(\mathbf{X})$ is sufficient for $\mathcal{F}$.

(ii) Necessity:

Suppose that $T(\mathbf{X})$ is sufficient for $\mathcal{F}$. Then, for every t at which the (marginal) p.d.f. of $T(\mathbf{X})$, $g^*(t;\theta)$, is positive we have,

$$\frac{p(x, t;\theta)}{g^*(t;\theta)} = I\{\mathbf{x}; T(\mathbf{x}) = t\}B(\mathbf{x}), \tag{3.2.5}$$

where $B(\mathbf{x}) \geq 0$ is independent of θ. Moreover, $\sum_{\{\mathbf{y}\}} I\{\mathbf{y}; T(\mathbf{y}) = t\}\ B(\mathbf{y}) = 1$ since (3.2.5) is a conditional p.d.f. Thus, for every $\mathbf{x}$,

$$p(\mathbf{x}, t;\theta) = I\{\mathbf{x}; T(\mathbf{x}) = t\}B(\mathbf{x})g^*(t;\theta). \tag{3.2.6}$$

Finally, since for every $\mathbf{x}$,

$$p(\mathbf{x}, t;\theta) = I\{\mathbf{x}; T(\mathbf{x}) = t\}f(\mathbf{x};\theta), \tag{3.2.7}$$

we obtain that

$$f(\mathbf{x};\theta) = B(\mathbf{x})g^*(T(\mathbf{x});\theta), \quad \text{for all } \mathbf{x}. \tag{3.2.8}$$

QED

3.2.2 Theoretical Formulation

3.2.2.1 Distributions and Measures

We generalize the definitions and proofs of this section by providing measure-theoretic formulation. Some of these concepts were discussed in Chapter 1. This material can be skipped by students who have not had real analysis.

Let $(\Omega, \mathcal{A}, P)$ be a probability space. A random variable X is a finite real value measurable function on this probability space, i.e., $X : \Omega \to \mathbb{R}$. Let $\mathcal{X}$ be the sample space (range of X), i.e., $\mathcal{X} = X(\Omega)$. Let $\mathcal{B}$ be the Borel σ-field on $\mathcal{X}$, and consider the probability space $(\mathcal{X}, \mathcal{B}, P^X)$ where, for each $B \in \mathcal{B}$, $P^X\{B\} = P\{X^{-1}(B)\}$. Since X is a random variable, $\mathcal{B}^X = \{A : A = X^{-1}(B), B \in \mathcal{B}\} \subset \mathcal{A}$.

The distribution function of X is

$$F_X(x) = P^X\{(-\infty, x]\}, \quad -\infty < x < \infty. \tag{3.2.9}$$

Let $X_1, X_2, \ldots, X_n$ be n random variables defined on the same probability space $(\Omega, \mathcal{A}, P)$. The joint distribution of $\mathbf{X} = (X_1, \ldots, X_n)'$ is a real value function of $\mathbb{R}^n$ defined as

$$F(x_1, \ldots, x_n) = P\left\{\bigcap_{i=1}^{n}[X_i \leq x_i]\right\}. \tag{3.2.10}$$

Consider the probability space $(\mathcal{X}^{(n)}, \mathcal{B}^{(n)}, P^{(n)})$ where $\mathcal{X}^{(n)} = \mathcal{X} \times \cdots \times \mathcal{X}$, $\mathcal{B}^{(n)} = \mathcal{B} \times \cdots \times \mathcal{B}$ (or the Borel σ-field generated by the intervals $(-\infty, x_1] \times \cdots (-\infty, x_n]$, $(x_1, \ldots, x_n) \in \mathbb{R}^n$) and for $B \in \mathcal{B}^{(n)}$

$$P^{(n)}\{B\} = \int_B dF(x_1, \ldots, x_n). \tag{3.2.11}$$

A function $h : \mathcal{X}^{(n)} \to \mathcal{R}$ is said to be $\mathcal{B}^{(n)}$-measurable if the sets $h^{-1}((-\infty, \zeta])$ are in $\mathcal{B}^{(n)}$ for all $-\infty < \zeta < \infty$. By the notation $h \in \mathcal{B}^{(n)}$ we mean that h is $\mathcal{B}^{(n)}$-measurable.

A **random sample** of size n is the realization of n i.i.d. random variables (see Chapter 2 for definition of independence).

To economize in notation, we will denote by bold $\mathbf{x}$ the vector $(x_1, \ldots, x_n)$, and by $F(\mathbf{x})$ the joint distribution of $(X_1, \ldots, X_n)$. Thus, for all $B \in \mathcal{B}^{(n)}$,

$$P^{(n)}\{B\} = P\{(X_1, \ldots, X_n) \in B\} = \int_B dF(\mathbf{x}). \tag{3.2.12}$$

This is a probability measure on $(\mathcal{X}^{(n)}, B^{(n)})$ induced by $F(\mathbf{x})$. Generally, a σ-**finite measure** μ on $\mathcal{B}^{(n)}$ is a **nonnegative** real value set function, i.e., $\mu : \mathcal{B}^{(n)} \to [0, \infty]$, such that

(i) $\mu(\phi) = 0$;

(ii) if $\{B_n\}_{n=1}^{\infty}$ is a sequence of mutually disjoint sets in $\mathcal{B}^{(n)}$, i.e., $B_i \cap B_j = \phi$ for any $i \neq j$, then

$$\mu\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mu(B_n);$$

(iii) there exists a partition of $\mathcal{X}^{(n)}$, $\{B_1, B_2, \ldots\}$ for which $\mu(B_i) < \infty$ for all $i = 1, 2, \ldots$.

The Lebesque measure $\int_B d\mathbf{x}$ is a σ-finite measure on $B^{(n)}$.

If there is a countable set of marked points in $\mathbb{R}^n$, $S = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots\}$, the **counting measure** is

$$\begin{aligned} N(B; S) &= \{\#\ \text{of marked points in}\ \ B\} \\ &= |B \cap S|. \end{aligned}$$

$N(B; S)$ is a σ-finite measure, and for any finite real value function $g(x)$

$$\int_B g(\mathbf{x}) dN(\mathbf{x}; S) = \sum_x I\{\mathbf{x} \in S \cap B\} g(\mathbf{x}).$$

Notice that if B is such that $N(B; S) = 0$ then $\int_B g(\mathbf{x}) dN(\mathbf{x}; S) = 0$. Similarly, if B is such that $\int_B d\mathbf{x} = 0$ then, for any positive integrable function $g(\mathbf{x})$, $\int_B g(\mathbf{x}) d\mathbf{x} = 0$. Moreover, $\nu(B) = \int_B g(\mathbf{x}) d\mathbf{x}$ and $\lambda(B) = \int_B g(\mathbf{x}) dN(\mathbf{x}; S)$ are σ-finite measures on $(\mathcal{X}^{(n)}, \mathcal{B}^{(n)})$.

Let ν and μ be two σ-finite measures defined on $(\mathcal{X}^{(n)}, \mathcal{B}^{(n)})$. We say that ν is **absolutely continuous** with respect to μ if $\mu(B) = 0$ implies that $\nu(B) = 0$. We denote this relationship by $\nu \ll \mu$. If $\nu \ll \mu$ and $\mu \ll \nu$, we say that ν and μ are **equivalent**, $\nu \equiv \mu$. We will use the notation $F \ll \mu$ if the probability measure P_F is absolutely continuous with respect to μ. If $F \ll \mu$ there exists a nonnegative function $f(\mathbf{x})$, which is $\mathcal{B}$ measurable, satisfying

$$P_F\{B\} = \int_B dF(\mathbf{x}) = \int_B f(\mathbf{x}) d\mu(\mathbf{x}). \tag{3.2.13}$$

$f(\mathbf{x})$ is called the (Radon–Nikodym) derivative of F with respect to μ or the generalized density p.d.f. of $F(\mathbf{x})$. We write

$$dF(\mathbf{x}) = f(\mathbf{x}) d\mu(\mathbf{x}) \tag{3.2.14}$$

or

$$f(\mathbf{x}) = \frac{dF(\mathbf{x})}{d\mu(\mathbf{x})}. \tag{3.2.15}$$

As discussed earlier, a **statistical model** is represented by a family $\mathcal{F}$ of distribution functions F_θ on $\mathcal{X}^{(n)}$, $\theta \in \Theta$. The family $\mathcal{F}$ is **dominated** by a σ-finite measure μ if $F_\theta \ll \mu$, for each $\theta \in \Theta$.

We consider only models of dominated families. A theorem in measure theory states that if $\mathcal{F} \ll \mu$ then there exists a countable sequence $\{F_{\theta_n}\}_{n=1}^{\infty} \subset \mathcal{F}$ such that

$$F^*(\mathbf{x}) = \sum_{n=1}^{\infty} \frac{1}{2^n} F_{\theta_n}(\mathbf{x}) \tag{3.2.16}$$

induces a probability measure P^*, which dominates $\mathcal{F}$.

A statistic $T(\mathbf{X})$ is a measurable function of the data $\mathbf{X}$. More precisely, let $T : \mathcal{X}^{(n)} \to \mathcal{T}^{(k)}$, $k \geq 1$ and let $\mathcal{C}^{(k)}$ be the Borel σ-field of subsets of $\mathcal{T}^{(k)}$. The function $T(\mathbf{X})$ is a **statistic** if, for every $C \in \mathcal{C}^{(k)}$, $T^{-1}(C) \in \mathcal{B}^{(n)}$. Let $\mathcal{B}^T = \{B : B = T^{-1}(C)$ for $C \in \mathcal{C}^{(k)}\}$. The probability measure P^T on $\mathcal{C}^{(k)}$, induced by P^X, is given by

$$P^T\{C\} = P^X\{T^{-1}(C)\}, \quad C \in \mathcal{C}^{(k)}. \tag{3.2.17}$$

Thus, the induced distribution function of T is $F^T(\mathbf{t})$, where $\mathbf{t} \in \mathbb{R}^k$ and

$$F^T(\mathbf{t}) = \int_{T^{-1}((-\infty,t_1]\times\cdots(-\infty,t_k])} dF(\mathbf{x}). \tag{3.2.18}$$

If $F \ll \mu$ then $F^T \ll \mu^T$, where $\mu^T(C) = \mu(T^{-1}(C))$ for all $C \in \mathcal{C}^{(k)}$. The generalized density (p.d.f.) of T with respect to μ^T is $g^T(\mathbf{t})$ where

$$dF^T(\mathbf{t}) = g^T(\mathbf{t}) d\mu^T(\mathbf{t}). \tag{3.2.19}$$

If $h(\mathbf{x})$ is $\mathcal{B}^{(n)}$ measurable and $\int |h(\mathbf{x})| dF(\mathbf{x}) < \infty$, then the conditional expectation of $h(\mathbf{X})$ given $\{T(\mathbf{X}) = \mathbf{t}\}$ is a $\mathcal{B}^T$ measurable function, $E_F\{h(\mathbf{X}) \mid T(\mathbf{X}) = \mathbf{t}\}$, for which

$$\begin{aligned} \int_{T^{-1}(C)} h(\mathbf{x}) dF(\mathbf{x}) &= \int_{T^{-1}(C)} E_F\{h(\mathbf{X}) \mid T(\mathbf{x})\} dF(\mathbf{x}) \\ &= \int_C E_F\{h(\mathbf{X}) \mid T(\mathbf{X}) = \mathbf{t}\} dF^T(\mathbf{t}) \end{aligned} \tag{3.2.20}$$

for all $C \in \mathcal{C}^{(k)}$. In particular, if $C = \mathcal{T}^{(k)}$ we obtain the law of the iterated expectation; namely

$$E_F\{h(\mathbf{X})\} = E_F\{E_F\{h(\mathbf{X}) \mid T(\mathbf{X})\}\}. \tag{3.2.21}$$

Notice that $E_F\{h(\mathbf{X}) \mid T(\mathbf{X})\}$ assumes a constant value on the coset $A(\mathbf{t}) = \{\mathbf{x} : T(\mathbf{x}) = \mathbf{t}\} = T^{-1}(\{\mathbf{t}\})$, $\mathbf{t} \in \mathcal{T}^{(k)}$.

3.2.2.2 Sufficient Statistics

Consider a statistical model $(\mathcal{X}^{(n)}, \mathcal{B}^{(n)}, \mathcal{F})$ where $\mathcal{F}$ is a family of joint distributions of the random sample. A statistic $T : (\mathcal{X}^{(n)}, \mathcal{B}^{(n)}, \mathcal{F}) \to (\mathcal{T}^{(k)}, \mathcal{C}^{(k)}, \mathcal{F}^T)$ is called **sufficient for** $\mathcal{F}$ if, for all $B \in \mathcal{B}^{(n)}$, $P_F\{B \mid T(\mathbf{X}) = \mathbf{t}\} = p(B; \mathbf{t})$ for all $F \in \mathcal{F}$. That is, the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X})$ is the same for all F in $\mathcal{F}$. Moreover, for a fixed $\mathbf{t}$, $p(B; \mathbf{t})$ is $\mathcal{B}^{(n)}$ measurable and for a fixed B, $p(B; t)$ is $\mathcal{C}^{(k)}$ measurable.

Theorem 3.2.2. *Let $(\mathcal{X}^{(n)}, \mathcal{B}^{(n)}, \mathcal{F})$ be a statistical model and $\mathcal{F} \ll \mu$. Let $\{F_{\theta_n}\}_{n=1}^{\infty} \subset \mathcal{F}$ such that $F^*(\mathbf{x}) = \sum_{n=1}^{\infty} \frac{1}{2^n} F_{\theta_n}(\mathbf{x})$ and $\mathcal{F} \ll P^*$. Then $T(\mathbf{X})$ is sufficient for $\mathcal{F}$ if and only if for each $\theta \in \Theta$ there exists a $\mathcal{B}^T$ measurable function $g_\theta(T(\mathbf{x}))$ such that, for each $B \in \mathcal{B}^{(n)}$*

$$P_\theta\{B\} = \int_B g_\theta(T(\mathbf{x}))dF^*(\mathbf{x}), \tag{3.2.22}$$

i.e.,

$$dF_\theta(\mathbf{x}) = g_\theta(T(\mathbf{x}))dF^*(\mathbf{x}). \tag{3.2.23}$$

Proof. (**i**) Assume that $T(\mathbf{X})$ is sufficient for $\mathcal{F}$. Accordingly, for each $B \in \mathcal{B}^{(n)}$,

$$P_\theta\{B \mid T(\mathbf{X})\} = p(B, T(\mathbf{X}))$$

for all $\theta \in \Theta$. Fix B in $\mathcal{B}^{(n)}$ and let $C \in \mathcal{C}^{(k)}$.

$$P_\theta\{B \cap T^{-1}(C)\} = \int_{T^{-1}(C)} p(B, T(\mathbf{x}))dF_\theta(\mathbf{x}),$$

for each $\theta \in \Theta$. In particular,

$$p(B, T(\mathbf{X})) = E^*\{I\{\mathbf{X} \in B\} \mid T(\mathbf{X})\}.$$

By the Radon–Nikodym Theorem, since $F_\theta \ll F^*$ for each θ, there exists a $\mathcal{B}^T$ measurable function $g_\theta(T(\mathbf{X}))$ so that, for every $C \in \mathcal{C}^{(k)}$,

$$P_\theta\{T^{-1}(C)\} = \int_{T^{-1}(C)} g_\theta(T(\mathbf{x}))dF^*(\mathbf{x}).$$

Now, for $B \in \mathcal{B}^{(n)}$ and $\theta \in \Theta$,

$$\begin{aligned} P_\theta\{B\} &= \int_{\mathcal{X}^{(n)}} p(B \mid T(\mathbf{x}))dF_\theta(\mathbf{x}) \\ &= \int_{T^{-1}(\mathcal{T}^{(k)})} E^*\{I\{\mathbf{X} \in B\} \mid T(\mathbf{x})\}g_\theta(T(\mathbf{x}))dF^*(\mathbf{x}) \\ &= \int_B g_\theta(T(\mathbf{x}))dF^*(\mathbf{x}). \end{aligned}$$

Hence, $dF_\theta(\mathbf{x})/dF^*(\mathbf{x}) = g_\theta(T(\mathbf{x}))$, which is $\mathcal{B}^T$ measurable.

(**ii**) Assume that there exists a $\mathcal{B}^T$ measurable function $g_\theta(T(\mathbf{x}))$ so that, for each $\theta \in \Theta$,

$$dF_\theta(\mathbf{x}) = g_\theta(T(\mathbf{x}))dF^*(\mathbf{x}).$$

Let $A \in \mathcal{B}^{(n)}$ and define the σ-finite measure $d\nu_A^{(\theta)}(\mathbf{x}) = I\{\mathbf{x} \in A\}dF_\theta(\mathbf{x})$. $\nu_A^{(\theta)} \ll F^*$. Thus,

$$\begin{aligned} d\nu_A^{(\theta)}(\mathbf{x})/dF^*(\mathbf{x}) &= P_\theta\{A \mid T(\mathbf{x})\}g_\theta(T(\mathbf{x})) \\ &= P^*\{A \mid T(\mathbf{x})\}g_\theta(T(\mathbf{x})). \end{aligned}$$

Thus, $P_\theta\{A \mid T(\mathbf{X})\} = P^*\{A \mid T(\mathbf{X})\}$ for all $\theta \in \Theta$. Therefore T is a sufficient statistic. QED

Theorem 3.2.3 (Abstract Formulation of the Neyman–Fisher Factorization Theorem). *Let $(\mathcal{X}^{(n)}, \mathcal{B}^{(n)}, \mathcal{F})$ be a statistical model with $\mathcal{F} \ll \mu$. Then $T(\mathbf{X})$ is sufficient for $\mathcal{F}$ if and only if*

$$dF_\theta(\mathbf{x}) = g_\theta(T(\mathbf{x}))h(\mathbf{x})d\mu(\mathbf{x}), \quad \theta \in \Theta \tag{3.2.24}$$

where $h \geq 0$ and $h \in \mathcal{B}^{(n)}$, $g_\theta \in \mathcal{B}^T$.

Proof. Since $\mathcal{F} \ll \mu$, $\exists\{F_{\theta_n}\}_{n=1}^\infty \subset \mathcal{F}$, such that $F^*(\mathbf{x}) = \sum_{n=1}^\infty \frac{1}{2^n} F_{\theta_n}(\mathbf{x})$ dominates $\mathcal{F}$. Hence, by the previous theorem, $T(\mathbf{X})$ is sufficient for $\mathcal{F}$ if and only if there exists a $\mathcal{B}^T$ measurable function $g_\theta(T(\mathbf{x}))$ so that

$$dF_\theta(x) = g_\theta(T(\mathbf{x}))dF^*(\mathbf{x}).$$

Let $f_{\theta_n}(\mathbf{x}) = dF_{\theta_n}(\mathbf{x})/d\mu(\mathbf{x})$ and set $h(\mathbf{x}) = \sum_{n=1}^{\infty} \frac{1}{2^n} f_{\theta_n}(\mathbf{x})$. The function $h(\mathbf{x}) \in \mathcal{B}^{(n)}$ and

$$dF_\theta(x) = g_\theta(T(\mathbf{x}))h(\mathbf{x})d\mu(\mathbf{x}).$$

QED

3.3 LIKELIHOOD FUNCTIONS AND MINIMAL SUFFICIENT STATISTICS

Consider a vector $\mathbf{X} = (X_1, \ldots, X_n)'$ of random variables having a joint c.d.f. $F_\theta(\mathbf{x})$ belonging to a family $\mathcal{F} = \{F_\theta(\mathbf{x}); \theta \in \Theta\}$. It is assumed that $\mathcal{F}$ is a regular family of distributions, i.e., $\mathcal{F} \ll \mu$, and, for each $\theta \in \Theta$, there exists $f(\mathbf{x}; \theta)$ such that

$$dF_\theta(\mathbf{x}) = f(\mathbf{x}; \theta)d\mu(\mathbf{x}).$$

$f(\mathbf{x}; \theta)$ is the joint p.d.f. of $\mathbf{X}$ with respect to $\mu(\mathbf{x})$. We define over the parameter space Θ a class of functions $L(\theta; \mathbf{X})$ called **likelihood functions**. The likelihood function of θ associated with a vector of random variables $\mathbf{X}$ is defined up to a positive factor of proportionality as

$$L(\theta; \mathbf{X}) \propto f(\mathbf{X}; \theta). \tag{3.3.1}$$

The factor of proportionality in (3.3.1) may depend on $\mathbf{X}$ but not on θ. Accordingly, we say that two likelihood functions $L_1(\theta; \mathbf{X})$ and $L_2(\theta; \mathbf{X})$ are equivalent, i.e., $L_1(\theta; \mathbf{X}) \sim L_2(\theta; \mathbf{X})$, if $L_1(\theta; \mathbf{X}) = A(\mathbf{X})L_2(\theta; \mathbf{X})$ where $A(\mathbf{X})$ is a positive function independent of θ. For example, suppose that $\mathbf{X} = (X_1, \ldots, X_n)'$ is a vector of i.i.d. random variables having a $N(\theta, 1)$ distribution, $-\infty < \theta < \infty$. The likelihood function of θ can be defined as

$$\begin{aligned} L_1(\theta; \mathbf{X}) &= \exp\left\{-\frac{1}{2}(\mathbf{X} - \theta\mathbf{1})'(\mathbf{X} - \theta\mathbf{1})\right\} \\ &= \exp\left\{-\frac{1}{2}Q\right\} \exp\left\{-\frac{n}{2}(\bar{X} - \theta)^2\right\}, \end{aligned} \tag{3.3.2}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $Q = \sum_{i=1}^{n}(X_i - \bar{X})^2$ and $\mathbf{1}' = (1, \ldots, 1)$ or as

$$L_2(\theta; \mathbf{X}) = \exp\left\{-\frac{n}{2}(T(\mathbf{X}) - \theta)^2\right\}, \tag{3.3.3}$$

where $T(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$. We see that for a given value of $\mathbf{X}$, $L_1(\theta;\mathbf{X}) \sim L_2(\theta;\mathbf{X})$. All the equivalent versions of a likelihood function $L(\theta;\mathbf{X})$ belong to the same equivalence class. They all represent similar functions of θ.

If $S(\mathbf{X})$ is a statistic having a p.d.f. $g^S(s;\theta)$, $(\theta \in \Theta)$, then the likelihood function of θ given $S(\mathbf{X}) = s$ is $L^S(\theta;s) \propto g^S(s;\theta)$. $L^S(\theta;s)$ may or may not have a shape similar to $L(\theta;\mathbf{X})$. From the Factorization Theorem we obtain that if $L(\theta;\mathbf{X}) \sim L^S(\theta;S(\mathbf{X}))$, for all $\mathbf{X}$, then $S(\mathbf{X})$ is a sufficient statistic for $\mathcal{F}$. The information on θ given by $\mathbf{X}$ can be reduced to $S(\mathbf{X})$ without changing the factor of the likelihood function that depends on θ. This factor is called the **kernel of the likelihood function**. In terms of the above example, if $T(\mathbf{X}) = \bar{X}$, since $\bar{X} \sim N\left(\theta, \frac{1}{n}\right)$, $L^{T(\mathbf{X})}(\theta;t) = \exp\left\{-\frac{n}{2}(t-\theta)^2\right\}$. Thus, for all $\mathbf{x}$ such that $T(\mathbf{x}) = t$, $L^X(\theta;t) \sim L_1(\theta;\mathbf{x}) \sim L_2(\theta;\mathbf{x})$. $\bar{X}$ is indeed a sufficient statistic. The likelihood function $L^T(\theta;T(\mathbf{x}))$ associated with any sufficient statistic for $\mathcal{F}$ is equivalent to the likelihood function $L(\theta;\mathbf{x})$ associated with $\mathbf{X}$. Thus, if $T(\mathbf{X})$ is a sufficient statistic, then the likelihood ratio

$$L^T(\theta;T(\mathbf{X}))/L(\theta;\mathbf{X})$$

is independent of θ. A sufficient statistic $T(\mathbf{X})$ is called **minimal** if it is a function of any other sufficient statistic $S(\mathbf{X})$. The question is how to determine whether a sufficient statistic $T(\mathbf{X})$ is minimal sufficient.

Every statistic $S(\mathbf{X})$ induces a partition of the sample space $\chi^{(n)}$ of the observable random vector $\mathbf{X}$. Such a partition is a collection of disjoint sets whose union is $\chi^{(n)}$. Each set in this partition is determined so that all its elements yield the same value of $S(\mathbf{X})$. Conversely, every partition of $\chi^{(n)}$ corresponds to some function of $\mathbf{X}$. Consider now the partition whose sets contain only $\mathbf{x}$ points having equivalent likelihood functions. More specifically, let $\mathbf{x}^0$ be a point in $\chi^{(n)}$. A coset of $\mathbf{x}^0$ in this partition is

$$C(\mathbf{x}^0) = \{\mathbf{x}; L(\theta;\mathbf{x}) \sim L(\theta;\mathbf{x}^0)\}. \tag{3.3.4}$$

The partition of $\chi^{(n)}$ is obtained by varying $\mathbf{x}^0$ over all the points of $\chi^{(n)}$. We call this partition the **equivalent-likelihood** partition. For example, in the $N(\theta, 1)$ case $-\infty < \theta < \infty$, each coset consists of vectors $\mathbf{x}$ having the same mean $\bar{x} = \frac{1}{n}\mathbf{1}'\mathbf{x}$. These means index the cosets of the equivalent-likelihood partitions. The statistic $T(\mathbf{X})$ corresponding to the equivalent-likelihood partition is called the **likelihood statistic**. This statistic is an index of the likelihood function $L(\theta;\mathbf{x})$. We show now that **the likelihood statistic $T(\mathbf{X})$ is a minimal sufficient statistic (m.s.s.)**.

Let $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ be two different points and let $T(\mathbf{x})$ be the likelihood statistic. Then, $T(\mathbf{x}^{(1)}) = T(\mathbf{x}^{(2)})$ if and only if $L(\theta;\mathbf{x}^{(1)}) \sim L(\theta;\mathbf{x}^{(2)})$. Accordingly, $L(\theta;\mathbf{X})$

is a function of $T(\mathbf{X})$, i.e., $f(\mathbf{X};\theta) = A(\mathbf{X})g^*(T(\mathbf{X});\theta)$. Hence, by the Factorization Theorem, $T(\mathbf{X})$ is a sufficient statistic. If $S(\mathbf{X})$ is any other sufficient statistic, then each coset of $S(\mathbf{X})$ is contained in a coset of $T(\mathbf{X})$. Indeed, if $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ are such that $S(\mathbf{x}^{(1)}) = S(\mathbf{x}^{(2)})$ and $f(\mathbf{x}^{(i)}, \theta) > 0$ $(i = 1, 2)$, we obtain from the Factorization Theorem that $f(\mathbf{x}^{(1)};\theta) = k(\mathbf{x}^{(1)})g(S(\mathbf{x}^{(1)});\theta) = k(\mathbf{x}^{(1)})g(S(\mathbf{x}^{(2)});\theta) = k(\mathbf{x}^{(1)})f(x^{(2)};\theta)/k(x^{(2)})$, where $k(\mathbf{x}^{(2)}) > 0$. That is, $L(\theta;\mathbf{X}^{(1)}) \sim L(\theta;X^{(2)})$ and hence $T(\mathbf{X}^{(1)}) = T(\mathbf{X}^{(2)})$. This proves that $T(\mathbf{X})$ is a function of $S(\mathbf{X})$ and therefore minimal sufficient.

The minimal sufficient statistic can be determined by determining the likelihood statistic or, equivalently, by determining the partition of $\chi^{(n)}$ having the property that $f(\mathbf{x}^{(1)};\theta)/f(\mathbf{x}^{(2)};\theta)$ is independent of θ for every two points at the same coset.

3.4 SUFFICIENT STATISTICS AND EXPONENTIAL TYPE FAMILIES

In Section 2.16 we discussed the k-parameter exponential type family of distributions. If $X_1, \ldots, X_n$ are i.i.d. random variables having a k-parameter exponential type distribution, then the joint p.d.f. of $\mathbf{X} = (X_1, \ldots, X_n)$, in its canonical form, is

$$f(\mathbf{x};\psi_1, \ldots, \psi_k) = \prod_{i=1}^{n} h(x_i) \cdot \exp\left\{ \psi_1 \sum_{i=1}^{n} U_1(x_i) + \cdots + \psi_k \sum_{i=1}^{n} U_k(x_i) - nK(\psi_1, \ldots, \psi_k) \right\}. \tag{3.4.1}$$

It follows that $T(\mathbf{X}) = \left(\sum_{i=1}^{n} U_1(X_i), \ldots, \sum_{i=1}^{n} U_k(X_i) \right)$ is a sufficient statistic. The statistic $T(\mathbf{X})$ is minimal sufficient if the parameters $\{\psi_1, \ldots, \psi_k\}$ are linearly independent. Otherwise, by reparametrization we can reduce the number of natural parameters and obtain an m.s.s. that is a function of $T(\mathbf{X})$.

Dynkin (1951) investigated the conditions under which the existence of an m.s.s., which is a nontrivial reduction of the sample data, implies that the family of distributions, $\mathcal{F}$, is of the exponential type. The following regularity conditions are called Dynkin's Regularity Conditions. In Dynkin's original paper, condition (iii) required only piecewise continuous differentiability. Brown (1964) showed that it is insufficient. We phrase (iii) as required by Brown.

Dynkin's Regularity Conditions

- **(i)** The family $\mathcal{F} = \{F_\theta(x); \theta \in \Theta\}$ is a regular parametric family. Θ is an open subset of the Euclidean space R^k.
- **(ii)** If $f(x;\theta)$ is the p.d.f. of $F_\theta(x)$, then $\mathcal{S} = \{x; f(x;\theta) > 0\}$ is independent of θ.

(iii) The p.d.f.s $f(x;\theta)$ are such that, for each $\theta \in \Theta$, $f(x;\theta)$ is a continuously differentiable function of x over χ.

(iv) The p.d.f.s $f(x;\theta)$ are differentiable with respect to θ for each $x \in \mathcal{S}$.

Theorem 3.4.1 (Dynkin's). *If the family $\mathcal{F}$ is regular in the sense of Dynkin, and if for a sample of $n \geq k$ i.i.d. random variables $U_1(\mathbf{X}), \ldots, U_k(\mathbf{X})$ are linearly independent sufficient statistics, then the p.d.f. of X is*

$$f(x;\theta) = h(x) \exp\left\{\sum_{i=1}^{k} \psi_i(\theta) U_i(x) + C(\theta)\right\},$$

where the functions $\psi_1(\theta), \ldots, \psi_k(\theta)$ are linearly independent.

For a proof of this theorem and further reading on the subject, see Dynkin (1951), Brown (1964), Denny (1967, 1969), Tan (1969), Schmetterer (1974, p. 215), and Zacks (1971, p. 60). The connection between sufficient statistics and the exponential family was further investigated by Borges and Pfanzagl (1965), and Pfanzagl (1972). A one dimensional version of the theorem is proven in Schervish (1995, p. 109).

3.5 SUFFICIENCY AND COMPLETENESS

A family of distribution functions $\mathcal{F} = \{F_\theta(x); \theta \in \Theta\}$ **is called complete if, for any integrable function** $h(X)$,

$$\int h(x) dF_\theta(x) = 0 \quad \text{for all} \quad \theta \in \Theta \tag{3.5.1}$$

implies that $P_\theta[h(X) = 0] = 1$ **for all** $\theta \in \Theta$.

A statistic $T(\mathbf{X})$ is called **complete sufficient** statistic if it is **sufficient** for a family $\mathcal{F}$, and if the family $\mathcal{F}^T$ of all the distributions of $T(\mathbf{X})$ corresponding to the distributions in $\mathcal{F}$ is **complete**.

Minimal sufficient statistics are not necessarily complete. To show it, consider the family of distributions of Example 3.6 with $\xi_1 = \xi_2 = \xi$. It is a four-parameter, exponential-type distribution and the m.s.s. is

$$T(\mathbf{X}, \mathbf{Y}) = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{m} Y_i, \sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} Y_i^2\right).$$

The family $\mathcal{F}^T$ is incomplete since $E_\theta\left\{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i\right\} = 0$ for all $\theta = (\xi, \sigma_1, \sigma_2)$. But $P_\theta\left\{\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i\right\} = 0$, all θ. The reason for this incompleteness is that when

$\xi_1 = \xi_2$ the four natural parameters are not independent. Notice that in this case the parameter space $\Omega = \{\psi = (\psi_1, \psi_2, \psi_3, \psi_4); \psi_1 = \psi_2\psi_3/\psi_4\}$ is three-dimensional.

Theorem 3.5.1. *If the parameter space Ω corresponding to a k-parameter exponential type family is k-dimensional, then the family of the minimal sufficient statistic is complete.*

The proof of this theorem is based on the analyticity of integrals of the type (2.16.4). For details, see Schervish (1995, p. 108).

From this theorem we immediately deduce that the following families are complete.

1. $B(N, \theta)$, $0 < \theta < 1$; N fixed.
2. $P(\lambda)$, $0 < \lambda < \infty$.
3. $NB(\psi, \nu)$, $0 < \psi < 1$; ν fixed.
4. $G(\lambda, \nu)$, $0 < \lambda < \infty$, $0 < \nu < \infty$.
5. $\beta(p, q)$, $0 < p, q < \infty$.
6. $N(\mu, \sigma^2)$, $-\infty < \mu < \infty$, $0 < \sigma < \infty$.
7. $M(N, \theta)$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, $0 < \sum_{i=1}^{n} \theta_i < 1$; N fixed.
8. $N(\boldsymbol{\mu}, V)$, $\boldsymbol{\mu} \in R^{(k)}$; V positive definite.

We define now a weaker notion of **boundedly complete** families. These are families for which if $h(x)$ is a bounded function and $E_\theta\{h(X)\} = 0$, for all $\theta \in \Theta$, then $P_\theta\{h(x) = 0\} = 1$, for all $\theta \in \Theta$. For an example of a boundedly complete family that is incomplete, see Fraser (1957, p. 25).

Theorem 3.5.2 (Bahadur). *If $T(\mathbf{X})$ is a boundedly complete sufficient statistic, then $T(\mathbf{X})$ is minimal.*

Proof. Suppose that $S(\mathbf{X})$ is a sufficient statistic. If $S(\mathbf{X}) = \psi(T(\mathbf{X}))$ then, for any Borel set $B \in \mathcal{B}$,

$$E\{P\{B \mid S(\mathbf{X})\} \mid T(\mathbf{X})\} = P\{B \mid S(\mathbf{X})\} \text{ a.s.}$$

Define

$$h(T) = E\{P\{B \mid S(\mathbf{X})\} \mid T(\mathbf{X})\} - P\{B \mid T(\mathbf{X})\}.$$

By the law of iterated expectation, $E_\theta\{h(T)\} = 0$, for all $\theta \in \Theta$. But since $T(\mathbf{X})$ is boundedly complete,

$$E\{P\{B \mid S(\mathbf{X})\} \mid T(\mathbf{X})\} = P(B \mid S(\mathbf{X})\} = P\{B \mid T(\mathbf{X})\} \quad \text{a.s.}$$

Hence, $T \in \mathcal{B}^S$, which means that T is a function of S. Hence $T(\mathbf{X})$ is an m.s.s. QED

3.6 SUFFICIENCY AND ANCILLARITY

A statistic $A(\mathbf{X})$ is called **ancillary** if its distribution does not depend on the particular parameter(s) specifying the distribution of $\mathbf{X}$. For example, suppose that $\mathbf{X} \sim N(\theta 1_n, I_n)$, $-\infty < \theta < \infty$. The statistic $\mathbf{U} = (X_2 - X_1, \ldots, X_n - X_1)$ is distributed like $N(\mathbf{O}_{n-1}, I_{n-1} + J_{n-1})$. Since the distribution of $\mathbf{U}$ does not depend on θ, $\mathbf{U}$ is ancillary for the family $\mathcal{F} = \{N(\theta \mathbf{1}, I), -\infty < \theta < \infty\}$. If $S(\mathbf{X})$ is a sufficient statistic for a family $\mathcal{F}$, the inference on θ can be based on the likelihood based on S. If $f_S(s;\theta)$ is the p.d.f. of S, and if $A(\mathbf{X})$ is ancillary for $\mathcal{F}$, with p.d.f. $h(a)$, one could write

$$p_S(s;\theta) = p^*_\theta(s \mid a)h(a), \tag{3.6.1}$$

where $p^*_\theta(s \mid a)$ is the conditional p.d.f. of S given $\{A = a\}$. One could claim that, for inferential objectives, one should consider the family of conditional p.d.f.s $\mathcal{F}^{S|A} = \{p^*_\theta(s \mid a), \theta \in \Theta\}$. However, the following theorem shows that if S is a complete sufficient statistic, conditioning on $A(\mathbf{X})$ does not yield anything different, since $p_S(s;\theta) = p^*_\theta(s \mid a)$, with probability one for each $\theta \in \Theta$.

Theorem 3.6.1 (Basu's Theorem). *Let $\mathbf{X} = (X_1, \ldots, X_n)'$ be a vector of i.i.d. random variables with a common distribution belonging to $\mathcal{F} = \{F_\theta(\mathbf{x}), \theta \in \Theta\}$. Let $T(\mathbf{X})$ be a boundedly complete sufficient statistic for $\mathcal{F}$. Furthermore, suppose that $A(\mathbf{X})$ is an ancillary statistic. Then $T(\mathbf{X})$ and $A(\mathbf{X})$ are independent.*

Proof. Let $C \in \mathcal{B}^A$, where $\mathcal{B}^A$ is the Borel σ-subfield induced by $A(\mathbf{X})$. Since the distribution of $A(\mathbf{X})$ is independent of θ, we can determine $P\{A(\mathbf{X}) \in C\}$ without any information on θ. Moreover, the conditional probability $P\{A(\mathbf{X}) \in C \mid T(X)\}$ is independent of θ since $T(\mathbf{X})$ is a sufficient statistic. Hence, $P\{A(\mathbf{X}) \in C \mid T(\mathbf{X})\} - P\{A(\mathbf{X}) \in C\}$ is a statistic depending on $T(\mathbf{X})$. According to the law of the iterated expectation,

$$E_\theta\{P\{A(\mathbf{X}) \in C \mid T(\mathbf{X})\} - P\{A(\mathbf{X}) \in C\}\} = 0, \quad \text{all } \theta \in \Theta. \tag{3.6.2}$$

Finally, since $T(x)$ is boundedly complete,

$$P\{A(\mathbf{X}) \in C \mid T(\mathbf{X})\} = P\{A(\mathbf{X}) \in C\} \tag{3.6.3}$$

with probability one for each θ. Thus, $A(\mathbf{X})$ and $T(\mathbf{X})$ are independent. QED

From Basu's Theorem, we can deduce that only if the sufficient statistic $S(\mathbf{X})$ is incomplete for $\mathcal{F}$, then an inference on $\boldsymbol{\theta}$, conditional on an ancillary statistic, can be meaningful. An example of such inference is given in Example 3.10.

3.7 INFORMATION FUNCTIONS AND SUFFICIENCY

In this section, we discuss two types of information functions used in statistical analysis: the Fisher information function and the Kullback–Leibler information function. These two information functions are somewhat related but designed to fulfill different roles. The Fisher information function is applied in various estimation problems, while the Kullback–Leibler information function has direct applications in the theory of testing hypotheses. Other types of information functions, based on the log likelihood function, are discussed by Basu (1975), Barndorff-Nielsen (1978).

3.7.1 The Fisher Information

We start with the Fisher information and consider parametric families of distribution functions with p.d.f.s $f(x;\theta)$, $\theta \in \Theta$, which depend only on one real parameter θ. A generalization to vector valued parameters is provided later.

Definition 3.7.1. *The Fisher information function for a family* $\mathcal{F} = \{F(x;\theta); \theta \in \Theta\}$, *where* $dF(x;\theta) = f(x;\theta)d\mu(x)$, *is*

$$I(\theta) = E_\theta \left\{ \left[\frac{\partial}{\partial\theta} \log f(X;\theta) \right]^2 \right\}. \tag{3.7.1}$$

Notice that according to this definition, $\dfrac{\partial}{\partial\theta} \log f(x;\theta)$ should exist with probability one, under F_θ, and its second moment should exist. The random variable $\dfrac{\partial}{\partial\theta} \log f(x;\theta)$ is called the **score function**. In Example 3.11 we show a few cases.

We develop now some properties of the Fisher information when the density functions in $\mathcal{F}$ satisfy the following set of **regularity conditions**.

(i) Θ is an open interval on the real line (could be the whole line);

(ii) $\dfrac{\partial}{\partial\theta} f(x;\theta)$ exists (finite) for every x and every $\theta \in \Theta$.

(iii) For each θ in Θ there exists a $\delta < 0$ and a positive integrable function $G(x;\theta)$ such that, for all ϕ in $(\theta - \delta, \theta + \delta)$,

$$\left| \frac{f(x;\phi) - f(x;\theta)}{\phi - \theta} \right| \leq G(x;\theta). \tag{3.7.2}$$

(iv) $0 < E_\theta \left\{ \left[\frac{\partial}{\partial \theta} \log f(X;\theta) \right]^2 \right\} < \infty$ for each $\theta \in \Theta$.

One can show that under condition (iii) (using the Lebesgue Dominated Convergence Theorem)

$$\begin{aligned} \frac{\partial}{\partial \theta} \int f(x;\theta) d\mu(x) &= \int \frac{\partial}{\partial \theta} f(x;\theta) d\mu(x) \\ &= E_\theta \left\{ \frac{\partial}{\partial \theta} \log f(X;\theta) \right\} = 0, \end{aligned} \tag{3.7.3}$$

for all $\theta \in \Theta$. Thus, under these regularity conditions,

$$I(\theta) = V_\theta \left\{ \frac{\partial}{\partial \theta} \log f(X;\theta) \right\}.$$

This may not be true if conditions (3.7.2) do not hold. Example 3.11 illustrates such a case where $X \sim R(0,\theta)$. Indeed, if $X \sim R(0,\theta)$ then

$$\begin{aligned} \frac{\partial}{\partial \theta} \int_0^\theta \frac{dx}{\theta} &= 0 \\ &\neq \int_0^\theta \frac{\partial}{\partial \theta} \frac{1}{\theta} dx = -\frac{1}{\theta}. \end{aligned}$$

Moreover, in that example $V_\theta \left\{ \frac{\partial}{\partial \theta} \log f(X;\theta) \right\} = 0$ for all θ. Returning back to cases where regularity conditions (3.7.2) are satisfied, we find that if $X_1, \ldots, X_n$ are i.i.d. and $I_n(\theta)$ is the Fisher information function based on their joint distribution,

$$I_n(\theta) = E_\theta \left\{ \left[\frac{\partial}{\partial \theta} \log f(\mathbf{X};\theta) \right]^2 \right\}. \tag{3.7.4}$$

Since $X_1, \ldots, X_n$ are i.i.d. random variables, then

$$\frac{\partial}{\partial \theta} \log f(\mathbf{X};\theta) = \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f(\mathbf{X}_i;\theta) \tag{3.7.5}$$

and due to (3.7.3),

$$I_n(\theta) = nI(\theta). \tag{3.7.6}$$

Thus, under the regularity conditions (3.7.2), $I(\theta)$ is an additive function.

We consider now the information available in a statistic $S = (S_1(\mathbf{X}), \ldots, S_r(\mathbf{X}))$, where $1 \leq r \leq n$. Let $g^S(y_1, \ldots, y_r; \theta)$ be the joint p.d.f. of S. The Fisher information function corresponding to S is analogously

$$I^S(\theta) = E_\theta \left\{ \left[\frac{\partial}{\partial \theta} \log g^S(Y_1, \ldots, Y_r; \theta) \right]^2 \right\}. \tag{3.7.7}$$

We obviously assume that the family of induced distributions of S satisfies the regularity conditions (i)–(iv). We show now that

$$I_n(\theta) \geq I^S(\theta), \quad \text{all } \theta \in \Theta. \tag{3.7.8}$$

We first show that

$$\frac{\partial}{\partial \theta} \log g^S(y; \theta) = E \left\{ \frac{\partial}{\partial \theta} \log f(\mathbf{X}; \theta) \mid S = y \right\}, \text{ a.s.} \tag{3.7.9}$$

We prove (3.7.9) first for the discrete case. The general proof follows. Let $A(y) = \{\mathbf{x}; S_1(\mathbf{x}) = y_1, \ldots, S_r(\mathbf{x}) = y_r\}$. The joint p.d.f. of S at y is given by

$$g^S(y; \theta) = \sum_{\mathbf{x}} I\{\mathbf{x}; \mathbf{x} \in A(y)\} f(\mathbf{x}; \theta), \tag{3.7.10}$$

where $f(\mathbf{x}; \theta)$ is the joint p.d.f. of $\mathbf{X}$. Accordingly,

$$E \left\{ \frac{\partial}{\partial \theta} \log f(\mathbf{X}; \theta) \mid Y = y \right\} = \sum_{\mathbf{x}} I\{\mathbf{x}; \mathbf{x} \in A(y)\} \left(\frac{\partial}{\partial \theta} \log f(\mathbf{x}; \theta) \right) \cdot f(\mathbf{x}; \theta) / g^S(y; \theta). \tag{3.7.11}$$

Furthermore, for each $\mathbf{x}$ such that $f(\mathbf{x}; \theta) > 0$ and according to regularity condition (iii),

$$\begin{aligned} E_\theta \left\{ \frac{\partial}{\partial \theta} \log f(\mathbf{X}; \theta) \mid Y = y \right\} &= \frac{1}{g^S(y; \theta)} \sum_{\mathbf{x}} I\{\mathbf{x}; \mathbf{x} \in A(y)\} \cdot \frac{\partial}{\partial \theta} f(\mathbf{x}; \theta) \\ &= \frac{\partial}{\partial \theta} g^S(y; \theta) / g^S(y; \theta) \\ &= \frac{\partial}{\partial \theta} \log g^S(y; \theta). \end{aligned} \tag{3.7.12}$$

To prove (3.7.9) generally, let $S : (\mathcal{X}, \mathcal{B}, \mathcal{F}) \to (\mathcal{S}, \Gamma, \mathcal{G})$ be a statistic and $\mathcal{F} \ll \mu$ and $\mathcal{F}$ be regular. Then, for any $C \in \Gamma$,

$$\begin{aligned}
&\int_C E_\theta \left\{ \frac{\partial}{\partial \theta} \log f(x;\theta) \mid S = s \right\} g(s;\theta) d\lambda(s) \\
&\quad = \int_{S^{-1}(C)} \left(\frac{\partial}{\partial \theta} \log f(\mathbf{x};\theta) \right) f(\mathbf{x};\theta) d\mu(\mathbf{x}) \\
&\quad = \int_{S^{-1}(C)} \frac{\partial}{\partial \theta} f(\mathbf{x};\theta) d\mu(\mathbf{x}) = \frac{\partial}{\partial \theta} \int_{S^{-1}(C)} f(\mathbf{x};\theta) d\mu(\mathbf{x}) \qquad (3.7.13) \\
&\quad = \frac{\partial}{\partial \theta} \int_C g(s;\theta) d\lambda(s) = \int_C \frac{\partial}{\partial \theta} g(s;\theta) d\lambda(s) \\
&\quad = \int_C \left(\frac{\partial}{\partial \theta} \log g(s;\theta) \right) g(s;\theta) d\lambda(s).
\end{aligned}$$

Since C is arbitrary, (3.7.9) is proven. Finally, to prove (3.7.8), write

$$\begin{aligned}
0 &\le E_\theta \left\{ \left[\frac{\partial}{\partial \theta} \log f(\mathbf{X};\theta) - \frac{\partial}{\partial \theta} \log g^S(Y;\theta) \right]^2 \right\} \\
&= I_n(\theta) + I^S(\theta) - 2E_\theta \left\{ \frac{\partial}{\partial \theta} \log f(\mathbf{X};\theta) \cdot \frac{\partial}{\partial \theta} \log g^S(Y;\theta) \right\} \qquad (3.7.14) \\
&= I_n(\theta) + I^S(\theta) - 2E_\theta \left\{ \frac{\partial}{\partial \theta} \log g^S(Y;\theta) \cdot \right. \\
&\quad \left. \cdot \; E_\theta \left\{ \frac{\partial}{\partial \theta} \log f(X;\theta) \mid Y \right\} \right\} = I_n(\theta) - I^S(\theta).
\end{aligned}$$

We prove now that **if $T(\mathbf{X})$is a sufficient statistic for $\mathcal{F}$, then**

$$I^T(\theta) = I_n(\theta), \quad \text{all } \theta \in \Theta.$$

Indeed, from the Factorization Theorem, if $T(\mathbf{X})$ is sufficient for $\mathcal{F}$ then $f(\mathbf{x};\theta) = K(\mathbf{x})g(T(\mathbf{x});\theta)$, for all $\theta \in \Theta$. Accordingly, $I_n(\theta) = E_\theta \left\{ \left[\frac{\partial}{\partial \theta} \log g(T(\mathbf{X});\theta) \right]^2 \right\}$.

On the other hand, the p.d.f. of $T(\mathbf{X})$ is $g^T(t;\theta) = A(t)g(t;\theta)$, all $\theta \in \Theta$. Hence, $\frac{\partial}{\partial\theta}\log g^T(t;\theta) = \frac{\partial}{\partial\theta}\log g(t;\theta)$ for all θ and all t. This implies that

$$\begin{aligned} I^T(\theta) &= E_\theta\left\{\left[\frac{\partial}{\partial\theta}\log g^T(T(\mathbf{X});\theta)\right]^2\right\} \\ &= E_\theta\left\{\left[\frac{\partial}{\partial\theta}\log f(T(\mathbf{X});\theta)\right]^2\right\} = I_n(\theta), \end{aligned} \tag{3.7.15}$$

for all $\theta \in \Theta$. Thus, we have proven that if a family of distributions, $\mathcal{F}$, admits a sufficient statistic, we can determine the amount of information in the sample from the distribution of the m.s.s.

Under regularity conditions (3.7.2), for any statistic $U(\mathbf{X})$,

$$\begin{aligned} I_n(\theta) &= V\left\{\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right\} \\ &= V\left\{E\left\{\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta) \mid U(\mathbf{X})\right\}\right\} \\ &+ E\left\{V\left\{\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta) \mid U(\mathbf{X})\right\}\right\}. \end{aligned}$$

By (3.7.15), if $U(\mathbf{X})$ is an ancillary statistic, $\log g^U(u;\theta)$ is independent of θ. In this case $\frac{\partial}{\partial\theta}\log g^U(u;\theta) = E\left\{\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta) \mid U\right\} = 0$, with probability 1, and

$$I_n(\theta) = E\{I(\theta \mid U)\}.$$

3.7.2 The Kullback–Leibler Information

The Kullback–Leibler (K–L) information function, to discriminate between two distributions $F_\theta(x)$ and $F_\phi(x)$ of $\mathcal{F} = \{F_\theta(x); \theta \in \Theta\}$ is defined as

$$I(\theta;\phi) = E_\theta\left\{\log\frac{f(X;\theta)}{f(X;\phi)}\right\}; \quad \theta, \phi \in \Theta. \tag{3.7.16}$$

The family $\mathcal{F}$ is assumed to be regular. We show now that $I(\theta,\phi) \geq 0$ with equality if and only if $f(X;\theta) = f(X;\phi)$ with probability one. To verify this, we remind that $\log x$ is a concave function of x and by the Jensen inequality (see problem 8,

Section 2.5), $\log(E\{Y\}) \geq E\{\log Y\}$ for every nonnegative random variable Y, having a finite expectation. Accordingly,

$$
\begin{aligned}
-I(\theta, \phi) = E_\theta \left\{ -\log \frac{f(X;\theta)}{f(X;\phi)} \right\} &= \int \log \frac{f(x;\phi)}{f(x;\theta)} dF(x;\theta) \\
&\leq \log \int dF(x;\phi) = 0.
\end{aligned}
\tag{3.7.17}
$$

Thus, multiplying both sides of (3.7.17) by -1, we obtain that $I(\theta, \phi) \geq 0$. Obviously, if $P_\theta\{f(X;\theta) = f(X;\phi)\} = 1$, then $I(\theta, \phi) = 0$. If $X_1, \ldots, X_n$ are i.i.d. random variables, then the information function in the whole sample is

$$
I_n(\theta, \phi) = E_\theta \left\{ \log \frac{f(X;\theta)}{f(X;\phi)} \right\} = E_\theta \left\{ \sum_{i=1}^{n} \log \frac{f(X_i;\theta)}{f(X_i;\phi)} \right\} = nI(\theta, \phi). \tag{3.7.18}
$$

This shows that the K–L information function is additive if the random variables are independent.

If $S(\mathbf{X}) = (S_1(\mathbf{X}), \ldots, S_r(\mathbf{X}))$, $1 \leq r \leq n$, is a statistic having a p.d.f. $g^S(y_1, \ldots, y_r; \theta)$, then the K–L information function based on the information in $S(\mathbf{X})$ is

$$
I^S(\theta, \phi) = E_\theta \left\{ \log \frac{g^S(Y;\theta)}{g^S(Y;\phi)} \right\}. \tag{3.7.19}
$$

We show now that

$$
I^S(\theta, \phi) \leq I_n(\theta, \phi), \tag{3.7.20}
$$

for all $\theta, \phi \in \Theta$ and every statistic $S(\mathbf{X})$ with equality if $S(\mathbf{X})$ is a sufficient statistic. Since the logarithmic function is concave, we obtain from the Jensen inequality

$$
-I_n(\theta, \phi) = E_\theta \left\{ \log \frac{f(X;\phi)}{f(X;\theta)} \right\} \leq E_\theta \left\{ \log E_\theta \left\{ \frac{f(X;\phi)}{f(X;\theta)} \mid S(\mathbf{X}) \right\} \right\}. \tag{3.7.21}
$$

Generally, if S is a statistic,

$$
S : (\mathcal{X}, \mathcal{B}, \mathcal{F}) \to (\mathcal{S}, \Gamma, \mathcal{G}),
$$

then for any $C \in \Gamma$

$$\begin{aligned}\int_C E_\theta \left\{ \frac{f(\mathbf{X};\phi)}{f(\mathbf{X};\theta)} \mid S = s \right\} g^S(s;\theta) d\lambda(s) \\ = \int_{S^{-1}(C)} \frac{f(x;\phi)}{f(x;\theta)} f(x;\theta) d\mu(s) = P_\phi\{S^{-1}(C)\} \\ = \int_C g^S(s;\phi) d\lambda(s) = \int_C \frac{g^S(s;\phi)}{g^S(s;\theta)} g(s;\theta) d\lambda(s).\end{aligned}$$

This proves that

$$\frac{g^S(s;\phi)}{g^S(s;\theta)} = E_\theta \left\{ \frac{f(\mathbf{X};\phi)}{f(\mathbf{X};\theta)} \mid S(\mathbf{X}) = s \right\}. \tag{3.7.22}$$

Substituting this expression for the conditional expectation in (3.7.21) and multiplying both sides of the inequality by -1, we obtain (3.7.20). To show that if $S(\mathbf{X})$ is sufficient then equality holds in (3.7.20), we apply the Factorization Theorem. Accordingly, if $S(\mathbf{X})$ is sufficient for $\mathcal{F}$,

$$\frac{f(\mathbf{x};\phi)}{f(\mathbf{x};\theta)} = \frac{K(\mathbf{x}) g(S(\mathbf{x});\phi)}{K(\mathbf{x}) g(S(\mathbf{x});\theta)} \tag{3.7.23}$$

at all points $\mathbf{x}$ at which $K(\mathbf{x}) > 0$. We recall that this set is independent of θ and has probability 1. Furthermore, the p.d.f. of $S(\mathbf{X})$ is

$$g^S(y;\theta) = A(y) g(y;\theta), \quad \text{all } \theta \in \Theta. \tag{3.7.24}$$

Therefore,

$$\begin{aligned} I_n(\theta,\phi) &= E_\theta \left(\log \frac{A(S(\mathbf{X})) g(S(\mathbf{X});\theta)}{A(S(\mathbf{X})) g(S(\mathbf{X});\phi)} \right) \\ &= E_\theta \left\{ \log \frac{g^S(Y;\theta)}{g^S(Y;\phi)} \right\} \\ &= I^S(\theta,\phi), \qquad \text{for all } \theta, \phi \in \Theta. \end{aligned} \tag{3.7.25}$$

3.8 THE FISHER INFORMATION MATRIX

We generalize here the notion of the information for cases where $f(x;\boldsymbol{\theta})$ depends on a vector of k-parameters. The **score function**, in the multiparameter case, is defined as the random vector

$$\mathbf{S}(\boldsymbol{\theta}; X) = \nabla_\theta \log f(X;\boldsymbol{\theta}). \tag{3.8.1}$$

Under the regularity conditions (3.7.2), which are imposed on each component of θ,

$$E_\theta\{\mathbf{S}(\boldsymbol{\theta}; X)\} = \mathbf{0}. \tag{3.8.2}$$

The covariance matrix of $\mathbf{S}(\theta; X)$ is the **Fisher Information Matrix** (FIM)

$$I(\theta) = \mathfrak{D}_\theta[\mathbf{S}(\boldsymbol{\theta}; X)]. \tag{3.8.3}$$

If the components of (3.8.1) are not linearly dependent, then $I(\boldsymbol{\theta})$ is positive definite.

In the k-parameter canonical exponential type family

$$\log f(X; \psi) = \log A^*(X) + \psi' \mathbf{U}(X) - K(\psi). \tag{3.8.4}$$

The score vector is then

$$\mathbf{S}(\psi; X) = \mathbf{U}(X) - \nabla_\psi K(\psi), \tag{3.8.5}$$

and the FIM is

$$\begin{aligned} I(\psi) &= \mathfrak{D}_\psi[\mathbf{U}(X)] \\ &= \left(\frac{\partial^2}{\partial \psi_i \partial \psi_j} K(\psi); i, j = 1, \ldots, k \right). \end{aligned} \tag{3.8.6}$$

Thus, in the canonical exponential type family, $I(\psi)$ is the Hessian matrix of the cumulant generating function $K(\psi)$.

It is interesting to study the effect of reparametrization on the FIM. Suppose that the original parameter vector is $\boldsymbol{\theta}$. We reparametrize by defining the k functions

$$w_j = w_j(\theta_1, \ldots, \theta_k), \quad j = 1, \ldots, k.$$

Let

$$\theta_j = \psi_j(w_1, \ldots, w_k), \quad j = 1, \ldots, k$$

and

$$D(\mathbf{w}) = \left(\frac{\partial \psi_i(w_1, \ldots, w_k)}{\partial w_j} \right).$$

Then,

$$S(\mathbf{w}; X) = D(\mathbf{w}) \nabla_\psi \log f(x; \psi(\mathbf{w})).$$

It follows that the FIM, in terms of the parameters $\mathbf{w}$, is

$$\begin{aligned} I(\mathbf{w}) &= \mathcal{E}_{\mathbf{w}}[S(\mathbf{w}; X)] \\ &= D(\mathbf{w}) I(\psi(\mathbf{w})) D'(\mathbf{w}). \end{aligned} \tag{3.8.7}$$

Notice that $I(\psi(\mathbf{w}))$ is obtained from $I(\boldsymbol{\theta})$ by substituting $\psi(\mathbf{w})$ for $\boldsymbol{\theta}$.

Partition $\boldsymbol{\theta}$ into subvectors $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(l)}$ $(2 \leq l \leq k)$. We say that $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(l)}$ are **orthogonal** subvectors if the FIM is **block diagonal**, with l blocks, each containing only the parameters in the corresponding subvector.

In Example 3.14, μ and σ^2 are orthogonal parameters, while ψ_1 and ψ_2 are not orthogonal.

3.9 SENSITIVITY TO CHANGES IN PARAMETERS

3.9.1 The Hellinger Distance

There are a variety of distance functions for probability functions. Following Pitman (1979), we apply here the Hellinger distance.

Let $\mathcal{F} = \{F(x;\theta), \theta \in \Theta\}$ be a family of distribution functions, dominated by a σ-finite measure μ, i.e., $dF(x;\theta) = f(x;\theta)d\mu(x)$, for all $\theta \in \Theta$. Let θ_1, θ_2 be two points in Θ. The **Hellinger distance** between $f(x;\theta_1)$ and $f(x;\theta_2)$ is

$$\rho(\theta_1, \theta_2) = \left(\int \left[\sqrt{f(x;\theta_1)} - \sqrt{f(x;\theta_2)} \right]^2 d\mu(x) \right)^{1/2}. \tag{3.9.1}$$

Obviously, $\rho(\theta_1, \theta_2) = 0$ if $\theta_1 = \theta_2$.

Notice that

$$\rho^2(\theta_1, \theta_2) = \int f(x;\theta_1) d\mu(x) + \int f(x;\theta_2) d\mu(x) - 2 \int \sqrt{f(x;\theta_1) f(x;\theta_2)}\, d\mu(x). \tag{3.9.2}$$

Thus, $\rho(\theta_1, \theta_2) \leq \sqrt{2}$, for all $\theta_1, \theta_2 \in \Theta$.

The sensitivity of $\rho(\theta_1, \theta_0)$ at θ_0 is the derivative (if it exists) of $\rho(\theta, \theta_0)$, at $\theta = \theta_0$.

Notice that

$$\frac{\rho^2(\theta, \theta_0)}{(\theta - \theta_0)^2} = \int \frac{(\sqrt{f(x;\theta)} - \sqrt{f(x;\theta_0)})^2}{(\theta - \theta_0)^2} d\mu(x). \tag{3.9.3}$$

If one can introduce the limit, as $\theta \to \theta_0$, under the integral at the r.h.s. of (3.9.3), then

$$\begin{aligned}\lim_{\theta\to\theta_0}\frac{\rho^2(\theta,\theta_0)}{(\theta-\theta_0)^2} &= \int\left(\frac{\partial}{\partial\theta}\sqrt{f(x;\theta)}\big|_{\theta=\theta_0}\right)^2 d\mu(x)\\ &= \int\frac{(\frac{\partial}{\partial\theta}f(x;\theta_0))^2}{4f(x;\theta_0)}d\mu(x)\\ &= \frac{1}{4}I(\theta_0).\end{aligned} \tag{3.9.4}$$

Thus, if the regularity conditions (3.7.2) are satisfied, then

$$\lim_{\theta\to\theta_0}\frac{\rho(\theta,\theta_0)}{|\theta-\theta_0|} = \frac{1}{2}\sqrt{I(\theta_0)}. \tag{3.9.5}$$

Equation (3.9.5) expresses the sensitivity of $\rho(\theta,\theta_0)$, at θ_0, as a function of the Fisher information $I(\theta_0)$.

Families of densities that do not satisfy the regularity conditions (3.7.2) usually will not satisfy (3.9.5). For example, consider the family of rectangular distributions $\mathcal{F} = \{R(0,\theta), 0 < \theta < \infty\}$.

For $\theta > \theta_0 > 0$,

$$\rho(\theta,\theta_0) = \sqrt{2}\left(1-\sqrt{\frac{\theta_0}{\theta}}\right)^{1/2}$$

$$\frac{\partial}{\partial\theta}\rho(\theta,\theta_0) = \frac{1}{2\sqrt{2}}\left(1-\sqrt{\frac{\theta_0}{\theta}}\right)^{-1/2}\left(\frac{\theta_0}{\theta}\right)^{-1/2}\frac{1}{\theta^2}.$$

Thus,

$$\lim_{\theta\downarrow\theta_0}\frac{\partial}{\partial\theta}\rho(\theta,\theta_0) = \infty.$$

On the other hand, according to (3.7.1) with $n = 1$, $\frac{1}{2}\sqrt{I(\theta_0)} = \frac{1}{2\theta_0}$.

The results of this section are generalizable to families depending on k parameters $(\theta_1,\ldots,\theta_k)$. Under similar smoothness conditions, if $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_k)'$ is such that $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$, then

$$\lim_{\nu\to 0}\int\frac{(\sqrt{f(x;\boldsymbol{\theta}_0+\boldsymbol{\lambda}\nu)}-\sqrt{f(x;\boldsymbol{\theta}_0)})^2}{\nu^2}d\mu(x) = \frac{1}{4}\boldsymbol{\lambda}'I(\boldsymbol{\theta}_0)\boldsymbol{\lambda}, \tag{3.9.6}$$

where $I(\boldsymbol{\theta}_0)$ is the FIM.

PART II: EXAMPLES

Example 3.1. Let $X_1, \ldots, X_n$ be i.i.d. random variables having an absolutely continuous distribution with a p.d.f. $f(x)$. Here we consider the family $\mathcal{F}$ of **all** absolutely continuous distributions. Let $T(\mathbf{X}) = (X_{(1)}, \ldots, X_{(n)})$, where $X_{(1)} \leq \cdots \leq X_{(n)}$, be the order statistic. It is immediately shown that

$$h(\mathbf{x} \mid T(\mathbf{X}) = t) = \frac{1}{n!} I\{\mathbf{x}; x_{(1)} = t_1, \ldots, x_{(n)} = t_n\}.$$

Thus, the order statistic is a sufficient statistic. This result is obvious because the order at which the observations are obtained is irrelevant to the model. **The order statistic is always a sufficient statistic, when the random variables are i.i.d.** On the other hand, as will be shown in the sequel, any statistic that further reduces the data is insufficient for $\mathcal{F}$ and causes some loss of information. ■

Example 3.2. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a Poisson distribution, $P(\lambda)$. The family under consideration is $\mathcal{F} = \{P(\lambda); 0 < \lambda < \infty\}$. Let $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$. We know that $T(\mathbf{X}) \sim P(n\lambda)$. Furthermore, the joint p.d.f. of $\mathbf{X}$ and $T(\mathbf{X})$ is

$$p(x_1, \ldots, x_n, t; \lambda) = \frac{e^{-n\lambda}}{\prod_{i-1}^{n} x_i!} \lambda^{\sum_{i=1}^{n} x_i} \cdot I\left\{\mathbf{x} : \sum_{i=1}^{n} x_i = t\right\}.$$

Hence, the conditional p.d.f. of $\mathbf{X}$ given $T(\mathbf{X}) = t$ is

$$h(\mathbf{x} \mid t, \lambda) = \frac{t!}{\prod_{i=1}^{n} x_i! n^t} I\left\{\mathbf{x} : \sum_{i=1}^{n} x_i = t\right\};$$

where $x_1, \ldots, x_n$ are nonnegative integers and $t = 0, 1, \ldots$. We see that the conditional p.d.f. of $\mathbf{X}$ given $T(\mathbf{X}) = t$ is independent of λ. Hence $T(\mathbf{X})$ is a sufficient statistic. Notice that $X_1, \ldots, X_n$ have a conditional multinomial distribution given $\Sigma X_i = t$. ■

Example 3.3. Let $\mathbf{X} = (X_1, \ldots, X_n)'$ have a multinormal distribution $N(\mu \mathbf{1}_n, I_n)$, where $\mathbf{1}_n = (1, 1, \ldots, 1)'$. Let $T = \sum_{i=1}^{n} X_i$. We set $\mathbf{X}^* = (X_2, \ldots, X_n)$ and derive the

joint distribution of $(\mathbf{X}^*, T)$. According to Section 2.9, $(\mathbf{X}^*, T)$ has the multinormal distribution

$$N\left(\mu\begin{pmatrix}\mathbf{1}_{n-1}\\ n\end{pmatrix}, \mathfrak{V}\right)$$

where

$$\mathfrak{V} = \begin{pmatrix} I_{n-1} & \mathbf{1}_{n-1} \\ \mathbf{1}'_{n-1} & n \end{pmatrix}.$$

Hence, the conditional distribution of $\mathbf{X}^*$ given T is the multinormal

$$N(\bar{X}_n \mathbf{1}_{n-1}, V_{n-1}),$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean and $V_{n-1} = I_{n-1} - \frac{1}{n}J_{n-1}$. It is easy to verify that V_{n-1} is nonsingular. This conditional distribution is independent of μ. Finally, the conditional p.d.f. of X_1 given $(\mathbf{X}^*, T)$ is that of a one-point distribution

$$h(x_1 \mid \mathbf{X}^*, T; \mu) = I\{\mathbf{x} : x_1 = T - \mathbf{X}^{*\prime}\mathbf{1}_{n-1}\}.$$

We notice that it is independent of μ. Hence the p.d.f. of $\mathbf{X}$ given T is independent of μ and T is a sufficient statistic. ■

Example 3.4. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. random vectors having a bivariate normal distribution. The joint p.d.f. of the n vectors is

$$f(x, y; \xi, \eta, \rho, \sigma_1, \sigma_2) = \frac{1}{(2\pi)^{n/2}\sigma_1^n\sigma_2^n(1-\rho^2)^{n/2}} \cdot$$

$$\exp\left\{ -\frac{1}{2(1-\rho^2)}\left[\sum_{i=1}^{n}\left(\frac{x_i-\xi}{\sigma_1}\right)^2\right.\right.$$

$$\left.\left. -2\rho\sum_{i=1}^{n}\frac{x_i-\xi}{\sigma_1}\cdot\frac{y_i-\eta}{\sigma_2} + \sum_{i=1}^{n}\left(\frac{y_i-\eta}{\sigma_2}\right)^2\right]\right\}$$

where $-\infty < \xi,\ \eta < \infty;\ 0 < \sigma_1,\ \sigma_2 < \infty;\ -1 \le \rho \le 1$. This joint p.d.f. can be written in the form

$$f(\mathbf{x}, \mathbf{y}; \xi, \eta, \sigma_1, \sigma_2, \rho) = \frac{1}{(2\pi)^{n/2}\sigma_1^n\sigma_2^n(1-\rho^2)^{n/2}} \cdot$$

$$\exp\left\{ -\frac{n}{2(1-\rho^2)}\left[\frac{(\bar{x}-\xi)^2}{\sigma_1^2} - 2\rho\frac{(\bar{x}-\xi)(\bar{y}-\eta)}{\sigma_1\sigma_2} + \frac{(\bar{y}-\eta)^2}{\sigma_2^2}\right]\right.$$

$$\left. -\frac{1}{2(1-\rho^2)} \cdot \left[\frac{Q(\mathbf{x})}{\sigma_1^2} - 2\rho\frac{P(\mathbf{x}, \mathbf{y})}{\sigma_1\sigma_2} + \frac{Q(y)}{\sigma_2^2}\right]\right\},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $Q(\mathbf{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2$, $Q(\mathbf{y}) = \sum_{i=1}^{n}(y_i - \bar{y})^2$, $P(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$.

According to the Factorization Theorem, a sufficient statistic for $\mathcal{F}$ is

$$T(\mathbf{X}, \mathbf{Y}) = (\bar{X}, \bar{Y}, Q(\mathbf{X}), Q(Y), P(\mathbf{X}, \mathbf{Y})).$$

It is interesting that even if σ_1 and σ_2 are known, the sufficient statistic is still $T(\mathbf{X}, \mathbf{Y})$. On the other hand, if $\rho = 0$ then the sufficient statistic is $T^*(\mathbf{X}, \mathbf{Y}) = (\bar{X}, \bar{Y}, Q(\mathbf{X}), Q(\mathbf{Y}))$. ■

Example 3.5.

A. **Binomial Distributions**

$\mathcal{F} = \{B(n, \theta),\ 0 < \theta < 1\}$, n is known. $X_1, \ldots, X_n$ is a sample of i.i.d. random variables. For every point $\mathbf{x}^0$, at which $f(\mathbf{x}^0, \theta) > 0$, we have

$$\frac{f(\mathbf{x}; \theta)}{f(\mathbf{x}^0; \theta)} = \prod_{i=1}^{n} \frac{\binom{n}{x_i}}{\binom{n}{x_i^0}} \cdot \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^{n}(x_i - x_i^0)}.$$

Accordingly, this likelihood ratio can be independent of θ if and only if $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i^0$. Thus, the m.s.s. is $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$.

B. **Hypergeometric Distributions**

$X_i \sim H(N, M, S)$, $i = 1, \ldots, n$. The joint p.d.f. of the sample is

$$p(\mathbf{x}; N, M, S) = \prod_{i=1}^{n} \frac{\binom{M}{x_i}\binom{N-M}{S-x_i}}{\binom{N}{S}}.$$

The unknown parameter here is M, $M = 0, \ldots, N$. N and S are fixed known values. The minimal sufficient statistic is the order statistic $T_n = (X_{(1)}, \ldots, X_{(n)})$. To realize it, we consider the likelihood ratio

$$\frac{p(\mathbf{x}; N, M, S)}{p(\mathbf{x}^0; N, M, S)} = \prod_{i=1}^{n} \frac{\binom{M}{x_i}}{\binom{M}{x_i^0}} \cdot \frac{\binom{N-M}{S-x_i}}{\binom{N-M_0}{S-x_i^0}}.$$

This ratio is independent of (M) if and only if $x_{(i)} = x_{(i)}^0$, for all $i = 1, 2, \ldots, n$.

C. **Negative-Binomial Distributions**

$X_i \sim NB(\psi, \nu)$, $i = 1, \ldots, n$; $0 < \psi < 1$, $0 < \nu < \infty$.

(i) If ν is known, the joint p.d.f. of the sample is

$$p(\mathbf{x}; \psi, \nu) = (1-\psi)^{n\nu} \cdot \psi^{\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \frac{\Gamma(x_i + \nu)}{\Gamma(\nu)\Gamma(x_i + 1)}.$$

Therefore, the m.s.s. is $T_n = \sum_{i=1}^{n} X_i$.

(ii) If ν is unknown, the p.d.f.s ratio is

$$\psi^{\Sigma x_i - \Sigma x_i^0} \prod_{i=1}^{n} \frac{\Gamma(x_i^0 + 1)}{\Gamma(x_i + 1)} \cdot \frac{\Gamma(x_i + \nu)}{\Gamma(x_i^0 + \nu)}.$$

Hence, the minimal sufficient statistic is the order statistic.

D. **Multinomial Distributions**

We have a sample of n i.i.d. random vectors $\mathbf{X}^{(i)} = (X_1^{(i)}, \ldots, X_k^{(i)})$, $i = 1, \ldots, n$. Each $\mathbf{X}^{(i)}$ is distributed like the multinomial $M(s, \theta)$. The joint p.d.f. of the sample is

$$p(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}; s, \theta) = \prod_{i=1}^{n} \frac{s!}{x_1^{(i)}! \ldots x_k^{(i)}!} \prod_{i=1}^{k} \theta_j^{\sum_{i=1}^{n} x_j^{(i)}}.$$

Accordingly, an m.s.s. is $T_n = (T_n^{(1)}, \ldots, T_n^{(k-1)})$, where $T_n^{(j)} = \sum_{i=1}^{n} X_j^{(i)}$, $j = 1, \ldots, k-1$. Notice that $T_n^{(k)} = ns - \sum_{i=1}^{k-1} T_n^{(i)}$.

E. **Beta Distributions**

$$X_i \sim \beta(p, q), \quad i = 1, \ldots, n; \quad 0 < p,\ q < \infty.$$

The joint p.d.f. of the sample is

$$f(\mathbf{x}; p, q) = \frac{1}{B^n(p, q)} \prod_{i=1}^{n} x_i^{p-1}(1 - x_i)^{q-1}.$$

$0 \le x_i \le 1$ for all $i = 1, \ldots, n$. Hence, an m.s.s. is $T_n = \left(\prod_{i=1}^{n} X_i, \prod_{i=1}^{n} (1 - X_i) \right)$. In cases where either p or q are known, the m.s.s. reduces to the component of T_n that corresponds to the unknown parameter.

F. **Gamma Distributions**

$$X_i \sim G(\lambda, \nu), \quad i = 1, \ldots, n; \quad 0 < \lambda,\ \nu < \infty.$$

The joint distribution of the sample is

$$f(\mathbf{x}; \lambda, \nu) = \frac{\lambda^{\nu n}}{\Gamma^n(\nu)} \left(\prod_{i=1}^{n} x_i \right)^{\nu - 1} \exp\left\{ -\lambda \sum_{i=1}^{n} x_i \right\}.$$

Thus, if both λ and ν are unknown, then an m.s.s. is $T_n = \left(\prod_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i\right)$. If only ν is unknown, the m.s.s. is $T_n = \prod_{i=1}^{n} X_i$. If only λ is unknown, the corresponding statistic $\sum_{i=1}^{n} X_i$ is minimal sufficient.

G. **Weibull Distributions**

X has a Weibull distribution if $(X - \xi)^{\alpha} \sim E(\lambda)$. This is a three-parameter family, $\boldsymbol{\theta} = (\xi, \lambda, \alpha)$; where ξ is a location parameter (the density is zero for all $x < \xi$); λ^{-1} is a scale parameter; and α is a shape parameter. We distinguish among three cases.

(i) ξ and α are known.

Let $Y_i = X_i - \xi$, $i = 1, \ldots, n$. Since $Y_i^{\alpha} \sim E(\lambda)$, we immediately obtain from that an m.s.s., which is,

$$T_n = \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} (X_i - \xi)^{\alpha}.$$

(ii) If α and λ are known but ξ is unknown, then a minimal sufficient statistic is the order statistic.

(iii) α is unknown.

The joint p.d.f. of the sample is

$$f(\mathbf{x}; \xi, \lambda) = \lambda^n \alpha^n \prod_{i=1}^{n} (x_i - \xi)^{\alpha} \exp\left\{-\lambda \sum_{i=1}^{n} (x_i - \xi)^{\alpha}\right\}, \quad x_i \geq \xi$$

for all $i = 1, \ldots, n$. By examining this joint p.d.f., we realize that a minimal sufficient statistic is the order statistic, i.e., $T_n = (X_{(1)}, \ldots, X_{(n)})$.

H. **Extreme Value Distributions**

The joint p.d.f. of the sample is

$$f(\mathbf{x}; \lambda, \alpha) = \lambda^n \alpha^n \exp\left\{-\alpha \sum_{i=1}^{n} x_i - \lambda \sum_{i=1}^{n} e^{-\alpha x_i}\right\}.$$

Hence, if α is known then $T_n = \sum_{i=1}^{n} e^{-\alpha X_i}$ is a minimal sufficient statistic; otherwise, a minimal sufficient statistic is the order statistic.

Normal Distributions

(i) Single (Univariate) Distribution Model

$$X_i \sim N(\xi, \sigma^2), \quad i = 1, \ldots, n; \quad -\infty < \xi < \infty, \quad 0 < \sigma < \infty.$$

The m.s.s. is $T_n = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2 \right)$. If ξ is known, then an m.s.s. is $\sum_{i=1}^{n} (X_i - \xi)^2$; if σ is known, then the first component of T_n is sufficient.

(ii) Two Distributions Model

We consider a two-sample model according to which $X_1, \ldots, X_n$ are i.i.d. having a $N(\xi, \sigma_1^2)$ distribution and $Y_1, \ldots, Y_m$ are i.i.d. having a $N(\eta, \sigma_2^2)$ distribution. The X-sample is independent of the Y-sample. In the general case, an m.s.s. is

$$T = \left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j, \sum_{i=1}^{n} X_i^2, \sum_{j=1}^{m} Y_j^2 \right).$$

If $\sigma_1^2 = \sigma_2^2$ then the m.s.s. reduces to $T^* = \left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j, \sum_{i=1}^{n} X_i^2 + \sum_{j=1}^{m} Y_j^2 \right)$. On the other hand, if $\xi = \eta$ but $\sigma_1 \neq \sigma_2$ then the minimal statistic is T. ■

Example 3.6. Let (X, Y) have a bivariate distribution $N\left(\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right)$ with $-\infty < \xi_1, \xi_2 < \infty$; $0 < \sigma_1, \sigma_2 < \infty$. The p.d.f. of (X, Y) is

$$f(x, y; \xi_1, \xi_2, \sigma_1, \sigma_2) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{ \frac{\xi_1}{\sigma_1^2} x + \frac{\xi_2}{\sigma_2^2} y - \frac{1}{2\sigma_1^2} x^2 - \frac{1}{2\sigma_2^2} y^2 - \frac{1}{2}\left(\frac{\xi_1^2}{\sigma_1^2} + \frac{\xi_2^2}{\sigma_2^2} \right) \right\}.$$

This bivariate p.d.f. can be written in the canonical form

$$f(x, y; \psi_1, \ldots, \psi_4) = \frac{1}{2\pi} \exp\{\psi_1 x + \psi_2 y + \psi_3 x^2 + \psi_4 y^2 - K(\psi_1, \ldots, \psi_4)\},$$

where

$$\psi_1 = \frac{\xi_1}{\sigma_1^2}, \quad \psi_2 = \frac{\xi_2}{\sigma_2^2}, \quad \psi_3 = -\frac{1}{2\sigma_1^2}, \quad \psi_4 = -\frac{1}{2\sigma_2^2},$$

and

$$K(\psi_1, \ldots, \psi_4) = -\frac{1}{4}\left(\frac{\psi_1^2}{\psi_3} + \frac{\psi_2^2}{\psi_4}\right) + \frac{1}{2}\left[\log\left(-\frac{1}{2\psi_2}\right) + \log\left(-\frac{1}{2\psi_4}\right)\right].$$

Thus, if ψ_1, ψ_2, ψ_3, and ψ_4 are independent, then an m.s.s. is $T(\mathbf{X}) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i, \sum_{i=1}^n X_i^2, \sum_{i=1}^n Y_i^2\right)$. This is obviously the case when ξ_1, ξ_2, σ_1, σ_2 can assume arbitrary values. Notice that if $\xi_1 = \xi_2$ but $\sigma_1 \neq \sigma_2$ then $\psi_1, \ldots, \psi_4$ are still independent and $T(\mathbf{X})$ is an m.s.s. On the other hand, if $\xi_1 \neq \xi_2$ but $\sigma_1 = \sigma_2$ then an m.s.s. is

$$T^*(\mathbf{X}) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i, \sum_{i=1}^n (X_i^2 + Y_i^2)\right).$$

The case of $\xi_1 = \xi_2, \sigma_1 \neq \sigma_2$, is a case of four-dimensional m.s.s., when the parameter space is three-dimensional. This is a case of a curved exponential family. ■

Example 3.7. Binomial Distributions

$\mathcal{F} = \{B(n, \theta); 0 < \theta < 1\}$, n fixed. Suppose that $E_\theta\{h(X)\} = 0$ for all $0 < \theta < 1$. This implies that

$$\sum_{j=0}^n h(j)\binom{N}{j}\psi^j = 0, \quad \text{all } \psi,$$

$0 < \psi < \infty$, where $\psi = \theta/(1-\theta)$ is the odds ratio. Let $a_{n,j} = h(j)\binom{n}{j}$, $j = 0, \ldots, n$. The expected value of $h(X)$ is a polynomial of order n in ψ. According to the fundamental theorem of algebra, such a polynomial can have at most n roots. However, the hypothesis is that the expected value is zero for **all** ψ in $(0, \infty)$. Hence $a_{n,j} = 0$ for all $j = 0, \ldots, n$, independently of ψ. Or,

$$P_\theta\{h(X) = 0\} = 1, \quad \text{all } \theta.$$

■

Example 3.8. Rectangular Distributions

Suppose that $\mathcal{F} = \{R(0, \theta); 0 < \theta < \infty\}$. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common distribution from $\mathcal{F}$. Let $X_{(n)}$ be the sample maximum. We show that the family of distributions of $X_{(n)}$, $\mathcal{F}_n^*$, is complete. The p.d.f. of $X_{(n)}$ is

$$f_n(t; \theta) = \frac{n}{\theta^n} t^{n-1}, \quad 0 \leq t \leq \theta.$$

Suppose that $E_\theta\{h(X_{(n)})\} = 0$ for all $0 < \theta < \infty$. That is

$$\int_0^\theta h(t)t^{n-1}dt = 0, \quad \text{for all } \theta, \quad 0 < \theta < \infty.$$

Consider this integral as a Lebesque integral. Differentiating with respect to θ yields

$$h(x)x^{n-1} = 0, \quad \text{a.s. } [P_\theta],$$

$\theta \in (0, \infty)$. ■

Example 3.9. In Example 2.15, we considered the Model II of analysis of variance. The complete sufficient statistic for that model is

$$T(\mathbf{X}) = (\bar{\bar{X}}, S_w^2, S_b^2),$$

where $\bar{\bar{X}}$ is the grand mean; $S_w^2 = \dfrac{1}{k(n-1)}\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2$ is the "within" sample variance; and $S_b^2 = \dfrac{n}{k-1}\displaystyle\sum_i(\bar{X}_i - \bar{\bar{X}})^2$ is the "between" sample variance. Employing Basu's Theorem we can immediately conclude that $\bar{\bar{X}}$ is independent of (S_w^2, S_b^2). Indeed, if we consider the subfamily $\mathcal{F}_{\sigma,\rho}$ for a fixed σ and ρ, then $\bar{\bar{X}}$ is a complete sufficient statistic. The distributions of S_w^2 and S_b^2, however, do not depend on μ. Hence, they are independent of $\bar{\bar{X}}$. Since this holds for any σ and ρ, we obtain the result. ■

Example 3.10. This example follows Example 2.23 of Barndorff-Nielsen and Cox (1994, p. 42). Consider the random vector $\mathbf{N} = (N_1, N_2, N_3, N_4)$ having a multinomial distribution $M(n, \mathbf{p})$ where $\mathbf{p} = (p_1, \ldots, p_4)'$ and, for $0 < \theta < 1$,

$$p_1 = \frac{1}{6}(1 - \theta)$$
$$p_2 = \frac{1}{6}(1 + \theta)$$
$$p_3 = \frac{1}{6}(2 - \theta)$$
$$p_4 = \frac{1}{6}(2 + \theta).$$

The distribution of $\mathbf{N}$ is a curved exponential type. $\mathbf{N}$ is an m.s.s., but $\mathbf{N}$ is incomplete. Indeed, $E_\theta\left\{N_1 + N_2 - \dfrac{n}{3}\right\} = 0$ for all $0 < \theta < 1$, but $P_\theta\left\{N_1 + N_2 = \dfrac{n}{3}\right\} < 1$ for

all θ. Consider the statistic $A_1 = N_1 + N_2$. $A_1 \sim B\left(n, \frac{1}{6}\right)$. Thus, A_1 is ancillary. The conditional p.d.f. of $\mathbf{N}$, given $A_1 = a$ is

$$p(\mathbf{n} \mid a, \theta) = \binom{a}{n_1} \left(\frac{1-\theta}{2}\right)^{n_1} \left(\frac{1+\theta}{2}\right)^{a-n_1} \cdot \binom{n-a}{n_3} \left(\frac{2-\theta}{4}\right)^{n_3} \left(\frac{2+\theta}{4}\right)^{n-a-n_3},$$

for $n_1 = 0, 1, \ldots, a; n_3 = 0, 1, \ldots, n-a; n_2 = a - n_1$ and $n_4 = n - a - n_3$. Thus, N_1 is **conditionally** independent of N_3 given $A_1 = a$. ■

Example 3.11. A. Let $X \sim B(n, \theta)$, n known, $0 < \theta < 1$; $f(x;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ satisfies the regularity conditions (3.7.2). Furthermore,

$$\frac{\partial}{\partial\theta} \log f(x;\theta) = \frac{x}{\theta} - \frac{n-x}{1-\theta}, \quad 0 < \theta < 1.$$

Hence, the Fisher information function is

$$I(\theta) = E_\theta \left\{ \left[\frac{X}{\theta} - \frac{n-X}{1-\theta} \right]^2 \right\} = n/\theta(1-\theta), \quad 0 < \theta < 1.$$

B. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a rectangular distribution $R(0, \theta)$, $0 < \theta < \infty$. The joint p.d.f. is

$$f(\mathbf{x};\theta) = \frac{1}{\theta^n} I\{\mathbf{x}_{(n)} \le \theta\},$$

where $\mathbf{x}_{(n)} = \max_{1 \le i \le n} \{\mathbf{x}_i\}$. Accordingly,

$$\frac{\partial}{\partial\theta} \log f(\mathbf{x};\theta) = -\frac{n}{\theta} I\{x_{(n)} \le \theta\}$$

and the Fisher information in the whole sample is

$$I_n(\theta) = \frac{n^2}{\theta^2}, \quad 0 < \theta < \infty.$$

C. Let $X \sim \mu + G(1, 2)$, $-\infty < \mu < \infty$. In this case,

$$f(x;\mu) = (x-\mu)e^{-(x-\mu)} I(x \ge \mu).$$

Thus,

$$\left(\frac{\partial}{\partial \mu} \log f(x;\mu)\right)^2 = \left(-\frac{1}{x-\mu} + 1\right)^2.$$

But

$$E_\mu\left\{\frac{1}{(X-\mu)^2}\right\} = \int_\mu^\infty (x-\mu)^{-1} e^{-(x-\mu)} = \infty.$$

Hence, $I(\mu)$ **does not exist**. ■

Example 3.12. In Example 3.10, we considered a four-nomial distribution with parameters $p_i(\theta)$, $i = 1, \ldots, 4$, which depend on a real parameter θ, $0 < \theta < 1$. We considered two alternative ancillary statistics $A = N_1 + N_2$ and $A' = N_1 + N_4$. The question was, which ancillary statistic should be used for conditional inference. Barndorff-Nielsen and Cox (1994, p. 43) recommend to **use the ancillary statistic which maximizes the variance of the conditional Fisher information**.

A version of the log-likelihood function, conditional on $\{A = a\}$ is

$$\begin{aligned} l(\theta \mid a) &= N_1 \log(1-\theta) + (a - N_1)\log(1+\theta) \\ &\quad + N_3 \log(2-\theta) + (n - a - N_3)\log(2+\theta). \end{aligned}$$

This yields the conditional score function

$$\begin{aligned} S(\theta \mid a) &= \frac{\partial}{\partial \theta} l(\theta \mid a) \\ &= -N_1 \frac{2}{1-\theta^2} - N_3 \frac{4}{4-\theta^2} + \frac{a}{1+\theta} + \frac{n-a}{2+\theta}. \end{aligned}$$

The corresponding conditional Fisher information is

$$I(\theta \mid a) = \frac{n}{4-\theta^2} + a\frac{3}{(1-\theta^2)(4-\theta^2)}.$$

Finally, since $A \sim \mathcal{B}\left(n, \frac{1}{3}\right)$, the Fisher information is

$$\begin{aligned} I_n(\theta) &= E\{I(\theta \mid A)\} \\ &= n\frac{2-\theta^2}{(1-\theta^2)(4-\theta^2)}. \end{aligned}$$

In addition,

$$V\{I(\theta \mid A)\} = \frac{2n}{(1-\theta^2)^2(4-\theta^2)^2}.$$

In a similar fashion, we can show that

$$I(\theta \mid A') = \frac{n}{(2-\theta)(1+\theta)} + \frac{2\theta A'}{(1-\theta^2)(4-\theta^2)}$$

and

$$V\{I(\theta \mid A')\} = \frac{n\theta^2}{(1-\theta^2)^2(4-\theta^2)^2}.$$

Thus, $V\{I(\theta \mid A)\} > V\{I(\theta \mid A')\}$ for all $0 < \theta < 1$. Ancillary A is preferred. ■

Example 3.13. We provide here a few examples of the Kullback–Leibler information function.

A. **Normal Distributions**

Let $\mathcal{F}$ be the class of all the normal distributions $\{N(\mu, \sigma^2); -\infty < \mu < \infty, 0 < \sigma < \infty\}$. Let $\boldsymbol{\theta}_1 = (\mu_1, \sigma_1)$ and $\boldsymbol{\theta}_2 = (\mu_2, \sigma_2)$. We compute $I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. The likelihood ratio is

$$\frac{f(x;\boldsymbol{\theta}_1)}{f(x;\boldsymbol{\theta}_2)} = \frac{\sigma_2}{\sigma_1} \exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \left(\frac{x-\mu_2}{\sigma_2}\right)^2\right]\right\}.$$

Thus,

$$\log \frac{f(x;\boldsymbol{\theta}_1)}{f(x;\boldsymbol{\theta}_2)} = \log\left(\frac{\sigma_2}{\sigma_1}\right) - \frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \left(\frac{x-\mu_2}{\sigma_2}\right)^2\right].$$

Obviously, $E_{\boldsymbol{\theta}_1}\{(\frac{X-\mu_1}{\sigma_1})^2\} = 1$. On the other hand

$$E_{\boldsymbol{\theta}_1}\left\{\left(\frac{X-\mu_2}{\sigma_2}\right)^2\right\} = \left(\frac{\sigma_1}{\sigma_2}\right)^2 E_{\boldsymbol{\theta}_1}\left\{\frac{(\mu_1+\sigma_1 U-\mu_2)^2}{\sigma_1^2}\right\} = \frac{\sigma_1^2}{\sigma_2^2}\left(1 + \frac{(\mu_1-\mu_2)^2}{\sigma_1^2}\right),$$

where $U \sim N(0, 1)$. Hence, we obtain that

$$I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{2}\left[\left(\left(\frac{\sigma_1}{\sigma_2}\right)^2 - 1\right) + \frac{(\mu_1-\mu_2)^2}{\sigma_2^2}\right].$$

We see that the distance between the means contributes to the K–L information function quadratically while the contribution of the variances is through the ratio $\rho = \sigma_2/\sigma_1$.

B. **Gamma Distributions**

Let $\theta_i = (\lambda_i, \nu_i)$, $i = 1, 2$, and consider the ratio

$$\frac{f(x;\theta_1)}{f(x;\theta_2)} = \frac{\Gamma(\nu_2)}{\Gamma(\nu_1)} \cdot \frac{\lambda_1^{\nu_1}}{\lambda_2^{\nu_2}} x^{\nu_1-\nu_2} \exp\{-x(\lambda_1 - \lambda_2)\}.$$

We consider here two cases.

Case I: $\nu_1 = \nu_2 = \nu$. Since the νs are the same, we simplify by setting $\theta_i = \lambda_i$ $(i = 1, 2)$. Accordingly,

$$I(\lambda_1, \lambda_2) = E_{\lambda_1}\left\{\nu \log\left(\frac{\lambda_1}{\lambda_2}\right) + (\lambda_2 - \lambda_1)X\right\}$$

$$= \nu\left[\log\left(\frac{\lambda_1}{\lambda_2}\right) + \left(\frac{\lambda_2}{\lambda_1} - 1\right)\right].$$

This information function depends on the scale parameters $\lambda_i (i = 1, 2)$, through their ratio $\rho = \lambda_2/\lambda_1$.

Case II: $\lambda_1 = \lambda_2 = \lambda$. In this case, we write

$$I(\nu_1, \nu_2) = \log\frac{\Gamma(\nu_2)}{\Gamma(\nu_1)} - (\nu_2 - \nu_1)\log\lambda + E_{\nu_1}\{(\nu_1 - \nu_2)\log X\}.$$

Furthermore,

$$E_{\nu_1}\{\log X\} = \frac{\lambda^{\nu_1}}{\Gamma(\nu_1)}\int_0^\infty (\log x)x^{\nu_1-1}e^{-\lambda x}dx$$

$$= \frac{d}{d\nu_1}\log\Gamma(\nu_1) - \log\lambda.$$

The derivative of the log-gamma function is tabulated (Abramowitz and Stegun, 1968). ■

Example 3.14. Consider the normal distribution $N(\mu, \sigma^2)$; $-\infty < \mu < \infty$, $0 < \sigma^2 < \infty$. The score vector, with respect to $\boldsymbol{\theta} = (\mu, \sigma^2)$, is

$$\mathbf{S}(\boldsymbol{\theta}; X) = \begin{pmatrix} \dfrac{X-\mu}{\sigma^2} \\ -\dfrac{1}{2\sigma^2} + \dfrac{(X-\mu)^2}{2\sigma^4} \end{pmatrix}.$$

Thus, the FIM is

$$I(\mu, \sigma^2) = \begin{bmatrix} \dfrac{1}{\sigma^2} & 0 \\ 0 & \dfrac{1}{2\sigma^4} \end{bmatrix}.$$

We have seen in Example 2.16 that this distribution is a two-parameter exponential type, with canonical parameters $\psi_1 = \dfrac{\mu}{\sigma^2}$ and $\psi_2 = -\dfrac{1}{2\sigma^2}$. Making the reparametrization in terms of ψ_1 and ψ_2, we compute the FIM as a function of ψ_1, ψ_2.

The inverse transformation is

$$\mu = -\frac{\psi_1}{\psi_2},$$
$$\sigma^2 = -\frac{1}{2\psi_2}.$$

Thus,

$$D(\psi) = \begin{pmatrix} -\dfrac{1}{2\psi_2} & \dfrac{\psi_1}{2\psi_2^2} \\ 0 & \dfrac{1}{2\psi_2^2} \end{pmatrix}.$$

Substituting (3.8.9) into (3.8.8) and applying (3.8.7) we obtain

$$I(\psi_1, \psi_2) = D(\psi) \begin{bmatrix} -2\psi_2 & 0 \\ 0 & 2\psi_2^2 \end{bmatrix} (D(\psi))'$$

$$= \begin{bmatrix} \dfrac{\psi_1^2 - 4\psi_2^3}{32\psi_2^6} & \dfrac{\psi_1}{32\psi_2^6} \\ \dfrac{\psi_1}{32\psi_2^6} & \dfrac{1}{32\psi_2^6} \end{bmatrix}.$$

Notice that $\psi_1^2 - 4\psi_2^3 = \dfrac{\mu^2}{\sigma^4} + \dfrac{1}{2\sigma^6} > 0$. ■

Example 3.15. Let (X, Y) have the bivariate normal distribution $N\left(\mathbf{0}, \sigma^2\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$, $0 < \sigma^2 < \infty$, $-1 < \rho < 1$. This is a two-parameter exponential type family with

$$f(x, y) = \frac{1}{2\pi} \exp\{\psi_1 U_1(x, y) + \psi_2 U_2(x, y) - K(\psi_1, \psi_2)\},$$

where $U_1(x, y) = x^2 + y^2$, $U_2(x, y) = xy$ and

$$K(\psi_1, \psi_2) = -\frac{1}{2} \log(4\psi_1^2 - \psi_2^2).$$

The Hessian of $K(\psi_1, \psi_2)$ is the FIM, with respect to the canonical parameters. We obtain

$$I(\psi_1, \psi_2) = \frac{1}{(4\psi_1^2 - \psi_2^2)^2} \begin{bmatrix} 4(4\psi_1^2 + \psi_2^2) & -8\psi_1\psi_2 \\ -8\psi_1\psi_2 & 4\psi_1^2 + \psi_2^2 \end{bmatrix}.$$

Using the reparametrization formula, we get

$$I(\sigma^2, \rho) = \begin{bmatrix} \frac{1}{\sigma^4} & -\frac{\rho}{\sigma^2(1-\rho^2)} \\ -\frac{\rho}{\sigma^2(1-\rho^2)} & \frac{1+\rho^2}{(1-\rho^2)^2} \end{bmatrix}.$$

Notice that in this example neither ψ_1, ψ_2 nor σ^2, ρ are orthogonal parameters. ■

Example 3.16. Let $\mathcal{F} = \{E(\lambda), 0 < \lambda < \infty\}$. $\rho^2(\lambda_1, \lambda_2) = 2\left(1 - \frac{2\sqrt{\lambda_1\lambda_2}}{\lambda_1 + \lambda_2}\right)$. Notice that $\sqrt{\lambda_1, \lambda_2} \leq \frac{\lambda_1 + \lambda_2}{2}$ for all $0 < \lambda_1, \lambda_2 < \infty$. If $\lambda_1 = \lambda_2$ then $\rho(\lambda_1, \lambda_2) = 0$. On the other hand, $0 < \rho^2(\lambda_1, \lambda_2) < 2$ for all $0 < \lambda_1, \lambda_2 < \infty$. However, for λ_1 fixed

$$\lim_{\lambda_2 \to \infty} \rho^2(\lambda_1, \lambda_2) = 2.$$

■

PART III: PROBLEMS

Section 3.2

3.2.1 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common rectangular distribution $R(\theta_1, \theta_2)$, $-\infty < \theta_1 < \theta_2 < \infty$.

(i) Apply the Factorization Theorem to prove that $X_{(1)} = \min\{X_i\}$ and $X_{(n)} = \max\{X_i\}$ are sufficient statistics.

(ii) Derive the conditional p.d.f. of $\mathbf{X} = (X_1, \ldots, X_n)$ given $(X_{(1)}, X_{(n)})$.

3.2.2 Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having a two-parameter exponential distribution, i.e., $X \sim \mu + E(\lambda)$, $-\infty < \mu < \infty$, $0 < \lambda < \infty$. Let $X_{(1)} \leq \cdots \leq X_n$ be the order statistic.

(i) Apply the Factorization Theorem to prove that $X_{(1)}$ and $S = \sum_{i=2}^{n}(X_{(i)} - X_{(1)})$ are sufficient statistics.

(ii) Derive the conditional p.d.f. of $\mathbf{X}$ given $(X_{(1)}, S)$.

(iii) How would you generate an equivalent sample $\mathbf{X}'$ (by simulation) when the value of $(X_{(1)}, S)$ are given?

3.2.3 Consider the linear regression model (Problem 3, Section 2.9). The unknown parameters are (α, β, σ). What is a sufficient statistic for $\mathcal{F}$?

3.2.4 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a Laplace distribution with p.d.f.

$$f(x; \mu, \sigma) = \frac{1}{2\sigma} \exp\left\{-\frac{|x-\mu|}{\sigma}\right\}, \quad -\infty < x < \infty; \quad -\infty < \mu < \infty;$$

$0 < \sigma < \infty$. What is a sufficient statistic for $\mathcal{F}$

(i) when μ is known?

(ii) when μ is unknown?

Section 3.3

3.3.1 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common Cauchy distribution with p.d.f.

$$f(x; \mu, \sigma) = \frac{1}{\sigma\pi} \cdot \left[1 + \left(\frac{x-\mu}{\sigma}\right)^2\right]^{-1}, \quad -\infty < x < \infty;$$

$-\infty < \mu < \infty$, $0 < \sigma < \infty$. What is an m.s.s. for $\mathcal{F}$?

3.3.2 Let $X_1, \ldots, X_n$ be i.i.d. random variables with a distribution belonging to a family $\mathcal{F}$ of contaminated normal distributions, having p.d.f.s,

$$f(x; \alpha, \mu, \sigma) = (1-\alpha)\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} + \alpha \cdot \frac{1}{\pi\sigma}\left[1 + \left(\frac{x-\mu}{\sigma}\right)^2\right]^{-1}, \quad -\infty < x < \infty;$$

$-\infty < \mu < \infty$; $0 < \sigma < \infty$; $0 < \alpha < 10^{-2}$. What is an m.s.s. for $\mathcal{F}$?

3.3.3 Let $X_1, \ldots, X_n$ be i.i.d. having a common distribution belonging to the family $\mathcal{F}$ of all location and scale parameter beta distributions, having the p.d.f.s

$$f(x; \mu, \sigma, p, q) = \frac{1}{\sigma B(p,q)} \left(\frac{x-\mu}{\sigma}\right)^{p-1} \left(1 - \frac{x-\mu}{\sigma}\right)^{q-1},$$

$-\mu \leq x \leq \mu + \sigma; -\infty < \mu < \infty; 0 < \sigma < \infty; 0 < p, q < \infty.$

(i) What is an m.s.s. when all the four parameters are unknown?

(ii) What is an m.s.s. when p, q are known?

(iii) What is an m.s.s. when μ, σ are known?

3.3.4 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a rectangular $R(\theta_1, \theta_2)$, $-\infty < \theta_1 < \theta_2 < \infty$. What is an m.s.s.?

3.3.5 $\mathcal{F}$ is a family of joint distributions of (X, Y) with p.d.f.s

$$g(x, y; \lambda) = \exp\{-\lambda x - y/\lambda\}, \quad 0 < \lambda < \infty.$$

Given a sample of n i.i.d. random vectors (X_i, Y_i), $i = 1, \ldots, n$, what is an m.s.s. for $\mathcal{F}$?

3.3.6 The following is a model in population genetics, called the **Hardy–Weinberg model**. The frequencies $N_1, N_2, N_3, \sum_{i=1}^{3} N_i = n$, of three genotypes among n individuals have a distribution belonging to the family $\mathcal{F}$ of trinomial distributions with parameters $(n, p_1(\theta), p_2(\theta), p_3(\theta))$, where

$$p_1(\theta) = \theta^2, \quad p_2(\theta) = 2\theta(1-\theta), \quad p_3(\theta) = (1-\theta)^2, \quad (3.3.1)$$

$0 < \theta < 1$. What is an m.s.s. for $\mathcal{F}$?

Section 3.4

3.4.1 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common distribution with p.d.f.

$$f(x; \psi) = I\{\psi_1 \leq x \leq \psi_2\} h(x) \exp\left\{\sum_{i=3}^{k} \psi_i U_i(x) - K(\psi)\right\},$$

$-\infty < \psi_1 < \psi_2 < \infty$. Prove that $T(\mathbf{X}) = \left(X_{(1)}, X_{(n)}, \sum_{i=1}^{n} U_3(X_i), \ldots, \sum_{i=1}^{n} U_k(X_i)\right)$ is an m.s.s.

3.4.2 Let $\{(X_k, Y_i), i = 1, \ldots, n\}$ be i.i.d. random vectors having a common bivariate normal distribution

$$N\left(\begin{bmatrix}\xi\\ \eta\end{bmatrix}, \begin{bmatrix}\sigma_x^2 & \rho\sigma_x\sigma_y\\ \rho\sigma_x\sigma_y & \sigma_y^2\end{bmatrix}\right),$$

where $-\infty < \xi, \eta < \infty$; $0 < \sigma_x^2, \sigma_y^2 < \infty$; $-1 < \rho < 1$.

(i) Write the p.d.f. in canonical form.

(ii) What is the m.s.s. for $\mathcal{F}$?

3.4.3 In continuation of the previous problem, what is the m.s.s.

(i) when $\xi = \eta = 0$?

(ii) when $\sigma_x = \sigma_y = 1$?

(iii) when $\xi = \eta = 0$, $\sigma_x = \sigma_y = 1$?

Section 3.5

3.5.1 Let $\mathcal{F} = \{G^{\alpha}(\lambda, 1); 0 < \alpha < \infty, 0 < \lambda < \infty\}$ be the family of Weibull distributions. Is $\mathcal{F}$ complete?

3.5.2 Let $\mathcal{F}$ be the family of extreme-values distributions. Is $\mathcal{F}$ complete?

3.5.3 Let $\mathcal{F} = \{R(\theta_1, \theta_2); -\infty < \theta_1 < \theta_2 < \infty\}$. Let $X_1, X_2, \ldots, X_n$, $n \geq 2$, be a random sample from a distribution of $\mathcal{F}$. Is the m.s.s. complete?

3.5.4 Is the family of trinomial distributions complete?

3.5.5 Show that for the Hardy–Weinberg model the m.s.s. is complete.

Section 3.6

3.6.1 Let $\{(X_i, Y_i),\ i = 1, \ldots, n\}$ be i.i.d. random vectors distributed like $N\left(\mathbf{0}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$, $-1 < \rho < 1$.

(i) Show that the random vectors $\mathbf{X}$ and $\mathbf{Y}$ are ancillary statistics.

(ii) What is an m.s.s. based on the conditional distribution of $\mathbf{Y}$ given $\{\mathbf{X} = \mathbf{x}\}$?

3.6.2 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a normal distribution $N(\mu, \sigma^2)$, where both μ and σ are unknown.

(i) Show that $U(\mathbf{X}) = \dfrac{M_e - \bar{X}}{Q_3 - Q_1}$ is ancillary, where M_e is the sample median; $\bar{X}$ is the sample mean; Q_1 and Q_3 are the sample first and third quartiles.

(ii) Prove that $U(\mathbf{X})$ is independent of $|\bar{X}|/S$, where S is the sample standard deviation.

Section 3.7

3.7.1 Consider the one-parameter exponential family with p.d.f.s

$$f(x;\theta) = h(x)\exp\{U(x)\psi(\theta) - K(\theta)\}.$$

Show that the Fisher information function for θ is

$$I(\theta) = K''(\theta) - \psi''(\theta)\frac{K'(\theta)}{\psi'(\theta)}.$$

Check this result specifically for the Binomial, Poisson, and Negative-Binomial distributions.

3.7.2 Let (X_i, Y_i), $i = 1, \ldots, n$ be i.i.d. vectors having the bivariate standard normal distribution with unknown coefficient of correlation ρ, $-1 \leq \rho \leq 1$. Derive the Fisher information function $I_n(\rho)$.

3.7.3 Let $\phi(x)$ denote the p.d.f. of $N(0, 1)$. Define the family of mixtures

$$f(x;\alpha) = \alpha\phi(x) + (1-\alpha)\phi(x-1), \quad 0 \leq \alpha \leq 1.$$

Derive the Fisher information function $I(\alpha)$.

3.7.4 Let $\mathcal{F} = \{f(x;\psi), -\infty < \psi < \infty\}$ be a one-parameter exponential family, where the canonical p.d.f. is

$$f(x;\psi) = h(x)\exp\{\psi U(x) - K(\psi)\}.$$

(i) Show that the Fisher information function is

$$I(\psi) = K''(\psi).$$

(ii) Derive this Fisher information for the Binomial and Poisson distributions.

3.7.5 Let $X_1, \ldots, X_n$ be i.i.d. $N(0, \sigma^2)$, $0 < \sigma^2 < \infty$.

(i) What is the m.s.s. T?

(ii) Derive the Fisher information $I(\sigma^2)$ from the distribution of T.

(iii) Derive the Fisher information $I^{S^2}(\sigma^2)$, where $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is the sample variance. Show that $I^{S^2}(\sigma^2) < I(\sigma^2)$.

3.7.6 Let (X, Y) have the bivariate standard normal distribution $N\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$, $-1 < \rho < 1$. X is an ancillary statistic. Derive the conditional Fisher information $I(\rho \mid X)$ and then the Fisher information $I(\rho)$.

3.7.7 Consider the model of Problem 6. What is the Kullback–Leibler information function $I(\rho_1, \rho_2)$ for discriminating between ρ_1 and ρ_2 where $-1 \leq \rho_1 < \rho_2 \leq 1$.

3.7.8 Let $X \sim P(\lambda)$. Derive the Kullback–Leibler information $I(\lambda_1, \lambda_2)$ for $0 < \lambda_1, \lambda_2 < \infty$.

3.7.9 Let $X \sim B(n, \theta)$. Derive the Kullback–Liebler information function $I(\theta_1, \theta_2)$, $0 < \theta_1, \theta_2 < 1$.

3.7.10 Let $X \sim G(\lambda, \nu)$, $0 < \lambda < \infty$, ν known.

(i) Express the p.d.f. of X as a one-parameter canonical exponential type density, $g(x; \psi)$.

(ii) Find $\hat{\psi}$ for which $g(x; \psi)$ is maximal.

(iii) Find the Kullback–Leibler information function $I(\hat{\psi}, \psi)$ and show that $\dfrac{\partial^2}{\partial \psi^2} I(\hat{\psi}, \psi) = I(\psi) = \dfrac{\nu}{\psi^2}$.

Section 3.8

3.8.1 Consider the trinomial distribution $M(n, p_1, p_2)$, $0 < p_1, p_2, p_1 + p_2 < 1$.

(i) Show that the FIM is

$$I(p_1, p_2) = \frac{n}{1 - p_1 - p_2} \begin{bmatrix} \dfrac{1 - p_2}{p_1} & 1 \\ 1 & \dfrac{1 - p_1}{p_2} \end{bmatrix}.$$

(ii) For the Hardy–Weinberg model, $p_1(\theta) = \theta^2$, $p_2(\theta) = 2\theta(1 - \theta)$, derive the Fisher information function

$$I(\theta) = \frac{2n}{\theta(1 - \theta)}.$$

3.8.2 Consider the bivariate normal distribution. Derive the FIM $I(\xi, \eta, \sigma_1, \sigma_2, \rho)$.

3.8.3 Consider the gamma distribution $G(\lambda, \nu)$. Derive the FIM $I(\lambda, \nu)$.

3.8.4 Consider the Weibull distribution $W(\lambda, \alpha) \sim (G(\lambda, 1))^{1/2}$; $0 < \alpha, \lambda < \infty$. Derive the Fisher informaton matrix $I(\lambda, \alpha)$.

Section 3.9

3.9.1 Find the Hellinger distance between two Poisson distributions with parameters λ_1 and λ_2.

3.9.2 Find the Hellinger distance between two Binomial distributions with parameters $p_1 \neq p_2$ and the same parameter n.

3.9.3 Show that for the Poisson and the Binomial distributions Equation (3.9.4) holds.

PART IV: SOLUTIONS TO SELECTED PROBLEMS

3.2.1 $X_1, \ldots, X_n$ are i.i.d. $\sim R(\theta_1, \theta_2)$, $0 < \theta_1 < \theta_2 < \infty$.

(i)

$$f(X_1, \ldots, X_n; \boldsymbol{\theta}) = \frac{1}{(\theta_2 - \theta_1)^n} \prod_{i=1}^{n} I(\theta_1 < X_i < \theta_2)$$

$$= \frac{1}{(\theta_2 - \theta_1)^n} I(\theta_1 < X_{(1)} < X_{(n)} < \theta_2).$$

Thus, $f(X_1, \ldots, X_n; \boldsymbol{\theta}) = A(\mathbf{x}) g(T(\mathbf{x}), \boldsymbol{\theta})$, where $A(\mathbf{x}) = 1 \; \forall \; \mathbf{x}$ and

$$g(T(\mathbf{x}); \boldsymbol{\theta}) = \frac{1}{(\theta_2 - \theta_1)^n} I(\theta_1 < X_{(1)} < X_{(n)} < \theta_2).$$

$T(\mathbf{X}) = (X_{(1)}, X_{(n)})$ is a likelihood statistic and thus minimal sufficient.

(ii) The p.d.f. of $(X_{(1)}, X_{(n)})$ is

$$h(x, y) = n(n-1) \frac{(y - x)^{n-2}}{(\theta_2 - \theta_1)^n} I(\theta_1 < x < y < \theta_2).$$

Let $(X_{(1)}, \ldots, X_{(n)})$ be the order statistic. The p.d.f. of $(X_{(1)}, \ldots, X_{(n)})$ is

$$P(X_1, \ldots, X_n; \boldsymbol{\theta}) = \frac{n!}{(\theta_2 - \theta_1)^n} I(\theta_1 < X_1 < \cdots < X_n < \theta_2).$$

The conditional p.d.f. of $(X_{(1)}, \ldots, X_{(n)})$ given $(X_{(1)}, X_{(n)})$ is

$$\frac{p(X_1, \ldots, X_n; \boldsymbol{\theta})}{h(X_1, X_n; \boldsymbol{\theta})} = \frac{(n-2)!}{(X_n - X_1)^{n-2}} I(X_1 < X_2 < \cdots < X_{n-1} < X_n).$$

That is, $(X_{(2)}, \ldots, X_{(n-1)})$ given $(X_{(1)}, X_{(n)})$ are distributed like the $(n-2)$ order statistic of $(n-2)$ i.i.d. from $R(X_{(1)}, X_{(n)})$.

3.3.6 The likelihood function of θ, $0 < \theta < 1$, is

$$L(\theta; n, N_1, N_2) = \theta^{2N_1+N_2}(1-\theta)^{2N_3+N_2}.$$

Since $N_3 = n - N_1 - N_2$, $2N_3 + N_2 = 2n - 2N_1 - N_2$. Hence,

$$L(\theta; n, N_1, N_2) = \left(\frac{\theta}{1-\theta}\right)^{2N_1+N_2}(1-\theta)^{2n}.$$

The sample size n is known. Thus, the m.s.s. is $T_n = 2N_1 + N_2$.

3.4.1

$$f(x;\psi) = I\{\psi_1 \le x \le \psi_2\} \exp\left\{\sum_{i=3}^{k} \psi_i U_i(x) - K(\psi)\right\}.$$

The likelihood function of ψ is

$$L(\psi; \mathbf{X}) = I\{\psi_1 \le X_{(1)} < X_{(n)} \le \psi_2\} \cdot \exp\left\{\sum_{i=3}^{k} \psi_i \sum_{j=1}^{n} U_i(X_j) - nK(\psi)\right\}.$$

Thus $\left(X_{(1)}, X_{(n)}, \sum_{j=1}^{n} U_3(X_j), \ldots, \sum_{j=1}^{n} U_k(X_j)\right)$ is a likelihood statistic, i.e., minimal sufficient.

3.4.2 The joint p.d.f. of (X_i, Y_i), $i = 1, \ldots, n$ is

$$f(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}\sigma_1^n\sigma_2^n(1-\rho^2)^{n/2}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\sum_{i=1}^{n}\left(\frac{X_i-\xi}{\sigma_X}\right)^2 - 2\rho\sum_{i=1}^{n}\frac{X_i-\xi}{\sigma_X}\cdot\frac{Y_i-\eta}{\sigma_Y} + \sum_{i=1}^{n}\left(\frac{Y_i-\eta}{\sigma_Y}\right)^2\right]\right\}.$$

In canonical form, the joint density is

$$f(\mathbf{X}, \mathbf{Y}; \psi) = \frac{1}{(2\pi)^{n/2}} \exp\left(\psi_1 \sum_{i=-1}^{n} X_i + \psi_2 \sum_{i=1}^{n} Y_i + \psi_3 \sum_{i=1}^{n} X_i^2 + \psi_4 \sum_{i=1}^{n} Y_i^2 + \psi_5 \sum_{i=1}^{n} X_i Y_i - n K(\psi) \right),$$

where

$$\psi_1 = \frac{\xi}{\sigma_X^2(1-\rho^2)} - \frac{\eta\rho}{\sigma_X\sigma_Y(1-\rho^2)}$$

$$\psi_2 = \frac{\eta}{\sigma_Y^2(1-\rho^2)} - \frac{\xi\rho}{\sigma_X\sigma_Y(1-\rho^2)}$$

$$\psi_3 = -\frac{1}{2\sigma_X^2(1-\rho^2)}$$

$$\psi_4 = -\frac{1}{2\sigma_Y^2(1-\rho^2)}$$

$$\psi_5 = \frac{\rho}{\sigma_X\sigma_Y(1-\rho^2)}$$

and

$$K(\psi) = \frac{\xi^2\sigma_Y^2 + \eta^2\sigma_X^2 - 2\sigma_X\sigma_Y\rho\xi\eta}{2\sigma_X^2\sigma_Y^2(1-\rho^2)} + \frac{1}{2}\log(\sigma_X^2\sigma_Y^2(1-\rho^2)).$$

The m.s.s. for $\mathcal{F}$ is $T(\mathbf{X}, \mathbf{Y}) = (\Sigma X_i, \Sigma Y_i, \Sigma X_i^2, \Sigma Y_i^2, \Sigma X_i Y_i)$.

3.5.5 We have seen that the likelihood of θ is $L(\theta) \propto \theta^{T(\mathbf{N})}(1-\theta)^{2n-T(\mathbf{N})}$ where $T(\mathbf{N}) = 2N_1 + N_2$. This is the m.s.s. Thus, the distribution of $T(\mathbf{N})$ is $B(2n, \theta)$. Finally $\mathcal{F} = B(2n, \theta), 0 < \theta < 1\}$ is *complete*.

3.6.2

(i)

$$M_e \sim \mu + \sigma M_e(Z)$$

$$\bar{X} \sim \mu + \sigma \bar{Z}$$

$$Q_3 \sim \mu + \sigma Q_3(Z)$$

$$Q_1 \sim \mu + \sigma Q_1(Z)$$

$$U(\mathbf{X}) = \frac{M_e - \bar{X}}{Q_3 - Q_1} \sim \frac{M_e(\mathbf{Z}) - \bar{Z}}{Q_3(\mathbf{Z}) - Q_1(\mathbf{Z})}$$

independent of μ and σ.

(ii) By Basu's Theorem, $U(\mathbf{X})$ is independent of $(\bar{X}, S)$, which is a complete sufficient statistic. Hence, $U(\mathbf{X})$ is independent of $|\bar{X}|/S$.

3.7.1 The score function is

$$S(\theta; X) = U(X)\psi'(\theta) - K'(\theta).$$

Hence, the Fisher information is

$$\begin{aligned} I(\theta) &= V_\theta\{S(\theta; X)\} \\ &= (\psi'(\theta))^2 V_\theta\{U(X)\}. \end{aligned}$$

Consider the equation

$$\int h(x)e^{\psi(\theta)u(x)-K(\theta)}dx = 1.$$

Differentiating both sides of this equation with respect to θ, we obtain that

$$E_\theta\{U(X)\} = \frac{K'(\theta)}{\psi'(\theta)}.$$

Differentiating the above equations twice with respect to θ, yields

$$(\psi'(\theta))^2 E_\theta\{U^2(X)\} = (K'(\theta))^2 - \psi''(\theta)\frac{K'(\theta)}{\psi'(\theta)} + K''(\theta).$$

Thus, we get

$$V_\theta\{U(X)\} = \frac{K''(\theta)}{(\psi'(\theta))^2} - \psi''(\theta)\frac{K'(\theta)}{(\psi'(\theta))^3}.$$

Therefore

$$I(\theta) = K''(\theta) - \psi''(\theta)\frac{K'(\theta)}{\psi'(\theta)}.$$

In the Binomial case,

$$f(x;\theta) = \binom{n}{x} e^{x \log \frac{\theta}{1-\theta} + n \log(1-\theta)}.$$

Thus, $\psi(\theta) = \log \frac{\theta}{1-\theta}$, $K(\theta) = -n \log(1-\theta)$

$$\psi'(\theta) = \frac{1}{\theta(1-\theta)}, \qquad \psi''(\theta) = \frac{2\theta - 1}{\theta^2(1-\theta)^2}$$
$$K'(\theta) = \frac{n}{1-\theta}, \qquad K''(\theta) = \frac{n}{(1-\theta)^2}.$$

Hence, $I(\theta) = \frac{n}{\theta(1-\theta)}$.

In the Poisson case,

$$\psi(\lambda) = \log(\lambda), \qquad K(\lambda) = \lambda$$
$$\psi'(\lambda) = \frac{1}{\lambda}, \qquad \psi''(\lambda) = -\frac{1}{\lambda^2}, \qquad K'(\lambda) = 1, \quad K''(\lambda) = 0$$
$$I(\lambda) = \frac{1}{\lambda}, \qquad I_n(\lambda) = \frac{n}{\lambda}.$$

8 In the Negative-Binomial case, ν known,

$$\psi(p) = \log(1-p), \quad K(p) = -\nu \log(p), \quad I(p) = \frac{\nu}{p^2(1-p)}.$$

3.7.2 Let $Q_Y = \sum_{i=1}^{n} Y_i^2$, $P_{XY} = \sum_{i=1}^{n} X_i Y_i$, $Q_X = \sum_{i=1}^{n} X_i^2$.

Let $l(\rho)$ denote the log-likelihood function of ρ, $-1 < \rho < 1$. This is

$$l(\rho) = -\frac{n}{2} \log(1-\rho^2) - \frac{1}{2(1-\rho^2)}(Q_Y - 2\rho P_{XY} + Q_X).$$

Furthermore,

$$l''(\rho) = \frac{n(1-\rho^4) - (Q_X + Q_Y)(1+3\rho^2) + 2P_{XY}\rho(3+\rho^2)}{(1-\rho^2)^3}.$$

Recall that $I(\rho) = E\{-l''(\rho)\}$. Moreover,

$$E(Q_X + Q_Y) = 2n \quad \text{and} \quad E(P_{XY}) = n\rho.$$

Thus,

$$I(\rho) = \frac{n(1-\rho^4)}{(1-\rho^2)^3} = \frac{n(1+\rho^2)}{(1-\rho^2)^2}.$$

3.7.7

$$(X, Y) \sim N\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \quad -1 < \rho < 1$$

$$f(x, y; \rho) = \frac{1}{(2\pi)} \exp\left(-\frac{1}{2}x^2\right) \frac{1}{(1-\rho^2)^{1/2}} \exp\left(-\frac{1}{2(1-\rho^2)}(y-\rho x)^2\right)$$

$$\frac{f(x, y; \rho_1)}{f(x, y; \rho_2)} = \frac{(1-\rho_2^2)^{1/2}}{(1-\rho_1^2)^{1/2}} \exp\left\{-\frac{1}{2}\left(\frac{(y-\rho_1 x)^2}{1-\rho_1^2} - \frac{(y-\rho_2 x)^2}{1-\rho_2^2}\right)\right\}$$

$$\log \frac{f(x, y; \rho_1)}{f(x, y; \rho_2)} = \frac{1}{2}\log\left(\frac{1-\rho_2^2}{1-\rho_1^2}\right) - \frac{1}{2}\frac{(y-\rho_1 x)^2}{1-\rho_1^2} + \frac{1}{2}\frac{(y-\rho_2 x)^2}{1-\rho_2^2}.$$

Thus, the Kullback–Leibler information is

$$\begin{aligned} I(\rho_1, \rho_2) &= E_{\rho_1}\left\{\log \frac{f(X, Y; \rho_1)}{f(X, Y); \rho_2)}\right\} \\ &= \frac{1}{2}\log\left(\frac{1-\rho_2^2}{1-\rho_1^2}\right). \end{aligned}$$

The formula is good also for $\rho_1 > \rho_2$.

3.7.10 The p.d.f. of $G(\lambda, \nu)$ is

$$f(x; \lambda, \nu) = \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\lambda x}.$$

When ν is known, we can write the p.d.f. as $g(x; \psi) = h(x)\exp(\psi x + \nu \log(-\psi))$, where $h(x) = \dfrac{x^{\nu-1}}{\Gamma(\nu)}$, $\psi = -\lambda$. The value of ψ maximizing $g(x, \psi)$ is $\hat{\psi} = -\dfrac{\nu}{x}$. The K–L information $I(\psi_1, \psi)$ is

$$I(\psi_1, \psi) = \frac{(\psi_1 - \psi)\nu}{\psi_1} + \nu \log\left(\frac{\psi_1}{\psi}\right).$$

Substituting $\psi_1 = \hat{\psi}$, we have

$$I(\hat{\psi}, \psi) = \frac{(\hat{\psi} - \psi)\nu}{\hat{\psi}} + \nu \log\left(\frac{\hat{\psi}}{\psi}\right).$$

Thus,

$$\begin{aligned} \frac{\partial}{\partial \psi} I(\hat{\psi}, \psi) &= -\frac{\nu}{\hat{\psi}} - \frac{\nu}{\psi} \\ \frac{\partial^2}{\partial \psi^2} I(\hat{\psi}, \psi) &= \frac{\nu}{\psi^2} = I(\psi). \end{aligned}$$

3.8.2 The log-likelihood function is

$$l(\xi, \eta, \sigma_1, \sigma_2, \rho) = -\frac{1}{2}\log(1-\rho^2) - \frac{1}{2}\log(\sigma_1^2) - \frac{1}{2}\log(\sigma_2^2) - \frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\xi}{\sigma_1}\right)^2 + \left(\frac{Y-\eta}{\sigma_2}\right)^2 - 2\rho\frac{x-\xi}{\sigma_1}\cdot\frac{Y-\eta}{\sigma_2}\right].$$

The score coefficients are

$$S_1 = \frac{\partial l}{\partial \xi} = \frac{X-\xi}{\sigma_1^2(1-\rho^2)} - \frac{\rho}{1-\rho^2}\frac{Y-\eta}{\sigma_1\sigma_2}$$

$$S_2 = \frac{\partial l}{\partial \eta} = \frac{Y-\eta}{\sigma_2^2(1-\rho^2)} - \frac{\rho}{1-\rho^2}\frac{X-\xi}{\sigma_1\sigma_2}$$

$$S_3 = \frac{\partial l}{\partial \sigma_1^2} = \frac{(X-\xi)^2}{2\sigma_1^4(1-\rho^2)} - \frac{\rho}{2(1-\rho^2)}\frac{X-\xi}{\sigma_1^2}\cdot\frac{Y-\eta}{\sigma_1\sigma_2} - \frac{1}{2\sigma_1^2}$$

$$S_4 = \frac{\partial l}{\partial \sigma_2^2} = \frac{(Y-\eta)^2}{2\sigma_2^4(1-\rho^2)} - \frac{\rho}{2(1-\rho^2)}\frac{X-\xi}{\sigma_1\sigma_2}\frac{Y-\eta}{\sigma_2^2} - \frac{1}{2\sigma_2^2}$$

$$S_5 = \frac{\partial l}{\partial \rho} = \frac{\rho}{1-\rho^2} - \frac{\rho}{(1-\rho^2)^2}\left[\left(\frac{X-\xi}{\sigma_1}\right)^2 + \left(\frac{Y-\eta}{\sigma_2}\right)^2\right] + \frac{1+\rho^2}{(1-\rho^2)^2}\frac{X-\xi}{\sigma_1}\cdot\frac{Y-\eta}{\sigma_2}.$$

The FIM

$$I = (I_{ij}), \quad i, j = 1, \ldots, 5.$$

$$I_{11} = V(S_1) = \frac{1}{\sigma_1^2(1-\rho^2)}$$

$$I_{22} = V(S_2) = \frac{1}{\sigma_2^2(1-\rho^2)}$$

$$I_{12} = cov(S_1, S_2) = \frac{\rho}{\sigma_1\sigma_2(1-\rho^2)}$$

$$I_{33} = V(S_3) = \frac{2-\rho^2}{4\sigma_1^4(1-\rho^2)}$$

$$I_{44} = V(S_4) = \frac{2-\rho^2}{4\sigma_2^4(1-\rho^2)}$$

$$I_{13} = I_{14} = I_{15} = 0$$
$$I_{23} = I_{24} = I_{25} = 0$$
$$I_{34} = cov(S_3, S_4) = -\frac{\rho^2}{4\sigma_1^2\sigma_2^2(1-\rho^2)}$$
$$I_{35} = cov(S_3, S_5) = -\frac{\rho}{2\sigma_1^2(1-\rho^2)}$$
$$I_{45} = cov(S_4, S_5) = -\frac{\rho}{2\sigma_2(1-\rho^2)}$$

and

$$I_{55} = V(S_5) = \frac{1+\rho^2}{(1-\rho^2)^2}.$$

3.8.4 The FIM for the Weibull parameters. The likelihood function is

$$L(\lambda, \alpha) = \lambda\alpha X^{\alpha} e^{-\lambda X^{\alpha}}.$$

Thus,

$$l(\lambda, \alpha) = \log(\lambda) + \log(\alpha) + \log(X^{\alpha}) - \lambda X^{\alpha}.$$

$$S_1 = \frac{\partial l}{\partial \lambda} = \frac{1}{\lambda} - X^{\alpha}$$
$$S_2 = \frac{\partial l}{\partial \alpha} = \frac{1}{\alpha} + \log(X) - \lambda X^{\alpha}\log(X)$$
$$= \frac{1}{\alpha} + \frac{1}{\alpha}\log(X^{\alpha}) - \frac{\lambda}{\alpha}X^{\alpha}\log(X^{\alpha}).$$

Recall that $X^{\alpha} \sim E(\lambda)$. Thus, $E\{X^{\alpha}\} = \frac{1}{\lambda}$ and $E\{S_1\} = 0$. Let $\psi(1)$ denote the di-gamma function at 1 (see Abramowitz and Stegun, 1965, pp. 259). Then,

$$E\{\log(X^{\alpha})\} = -\log(\lambda) + \psi(1)$$
$$E\{X^{\alpha}\log X^{\alpha}\} = \frac{1}{\lambda} - \frac{1}{\lambda}(\log(\lambda) - \psi(1)).$$

Thus,

$$E\{S_2\} = \frac{1}{\alpha} - \frac{1}{\alpha}(\log(\lambda) - \psi(1))$$
$$- \frac{\lambda}{\alpha}\left(\frac{1}{\lambda} - \frac{1}{\lambda}(\log(\lambda) - \psi(1))\right) = 0.$$

$$
\begin{aligned}
I_{11} &= V(S_1) = V\{X^\alpha\} = \frac{1}{\lambda^2} \\
I_{12} &= cov(S_1, S_2) \\
&= cov\left(-X^\alpha, \frac{1}{\alpha}\log X^\alpha - \frac{\lambda}{\alpha}X^\alpha \log X^\alpha\right) \\
&= -\frac{1}{\alpha}cov(Y, \log Y) + \frac{\lambda}{\alpha}cov(Y, Y\log Y),
\end{aligned}
$$

where $Y \sim X^\alpha \sim E(\lambda)$.

$$
\begin{aligned}
E\{\log Y\} &= \psi(1) - \log(\lambda) \\
E\{(\log Y)^2\} &= \psi'(1) + (\psi(1) - \log\lambda)^2 \\
E\{Y\log Y\} &= \frac{1}{\lambda}(1 + \psi(1) - \log\lambda) \\
E\{Y(\log Y)^2\} &= \frac{1}{\lambda}(\psi'(1) + 2(\psi(1) - \log\lambda) + (\psi(1) - \log\lambda)^2) \\
E\{Y^2\log Y\} &= \frac{1}{\lambda^2}(1 + 2(1 + \psi(1) - \log\lambda)) \\
E\{Y^2(\log Y)^2\} &= \frac{2}{\lambda^2}(\psi'(1) + (1 + \psi(1) - \log\lambda)^2 + \psi(1) - \log\lambda).
\end{aligned}
$$

Accordingly,

$$
\begin{aligned}
cov(Y, \log Y) &= \frac{1}{\lambda}, \\
cov(Y, Y\log Y) &= \frac{2 + \psi(1) - \log\lambda}{\lambda^2},
\end{aligned}
$$

and

$$
I_{12} = \frac{1}{\lambda\alpha}(1 + \psi(1) - \log\lambda).
$$

Finally,

$$
\begin{aligned}
I_{22} &= V(S_2) \\
&= V\left(\frac{1}{\alpha}\log Y - \frac{\lambda}{\alpha}Y\log Y\right) \\
&= \frac{1}{\alpha^2}V\{\log Y\} + \frac{\lambda^2}{\alpha^2}V\{Y\log Y\} \\
&- 2\frac{\lambda}{\alpha^2}cov(\log Y, Y\log Y) \\
&= \frac{1}{\alpha^2}(\psi'(1) + 2(\psi'(1) + \psi(1) - \log\lambda) \\
&+ (1 + \psi(1) - \log\lambda)^2 - 2(\psi'(1) + \psi(1) - \log\lambda)).
\end{aligned}
$$

Thus,

$$I_{22} = \frac{1}{\alpha^2}(\psi'(1) + (1 + \psi(1) - \log\lambda)^2).$$

The Fisher Information Matix is

$$I(\lambda, \alpha) = \begin{bmatrix} \frac{1}{\lambda^2} & \frac{1}{\lambda\alpha}(1 + \psi(1) - \log\lambda) \\ \bullet & \frac{1}{\alpha^2}(\psi'(1) + (1 + \psi(1) - \log\lambda)^2) \end{bmatrix}.$$

CHAPTER 4

Testing Statistical Hypotheses

PART I: THEORY

4.1 THE GENERAL FRAMEWORK

Statistical hypotheses are statements about the unknown characteristics of the distributions of observed random variables. The first step in testing statistical hypotheses is to formulate a statistical model that can represent the empirical phenomenon being studied and identify the subfamily of distributions corresponding to the hypothesis under consideration. The statistical model specifies the family of distributions relevant to the problem. Classical tests of significance, of the type that will be presented in the following sections, test whether the deviations of observed sample statistics from the values of the corresponding parameters, as specified by the hypotheses, cannot be ascribed just to randomness. Significant deviations lead to weakening of the hypotheses or to their rejection. This testing of the significance of deviations is generally done by constructing a test statistic based on the sample values, deriving the sampling distribution of the test statistic according to the model and the values of the parameters specified by the hypothesis, and rejecting the hypothesis if the observed value of the test statistic lies in an improbable region under the hypothesis. For example, if deviations from the hypothesis lead to large values of a nonnegative test statistic $T(\mathbf{X})$, we compute the probability that future samples of the type drawn will yield values of $T(\mathbf{X})$ at least as large as the presently observed one. Thus, if we observe the value t_0 of $T(\mathbf{X})$, we compute the tail probability

$$\alpha(t_0) = P_0\{T(\mathbf{X}) \geq t_0\}.$$

This value is called the **observed significance level** or the P-**value** of the test. A small value of the observed significant level means either that an improbable event

Examples and Problems in Mathematical Statistics, First Edition. Shelemyahu Zacks.

has occurred or that the sample data are incompatible with the hypothesis being tested. If $\alpha(t_0)$ is very small, it is customary to reject the hypothesis.

One of the theoretical difficulties with this testing approach is that it does not provide a framework for choosing the test statistic. Generally, our intuition and knowledge of the problem will yield a reasonable test statistic. However, the formulation of one hypothesis is insufficient for answering the question whether the proposed test is a good one and how large should the sample be. In order to construct an optimal test, in a sense that will be discussed later, we have to formulate an alternative hypothesis, against the hypothesis under consideration. For distinguishing between the hypothesis and its alternative (which is also a hypothesis), we call the first one a **null hypothesis** (denoted by H_0) and the other one an **alternative hypothesis** H_1. The alternative hypothesis can also be formulated in terms of a subfamily of distributions according to the specified model. We denote this subfamily by $\mathcal{F}_1$. If the family $\mathcal{F}_0$ or $\mathcal{F}_1$ contains only one element, the corresponding null or alternative hypothesis is called **simple**, otherwise it is called **composite**. The null hypothesis and the alternative one enable us to determine not only the optimal test, but also the sample size required to obtain a test having a certain strength. We distinguish between two kinds of errors. An **error of Type I** is the error due to rejection of the null hypothesis when it is true. An error of **Type II** is the one committed when the null hypothesis is not rejected when it is false. It is generally impossible to guarantee that a test will never commit either one of the two kinds of errors. A trivial test that always accepts the null hypothesis never commits an error of the first kind but commits an error of the second kind whenever the alternative hypothesis is true. Such a test is powerless. The theoretical framework developed here measures the risk in these two kinds of errors by the probabilities that a certain test will commit these errors. Ideally, the probabilities of the two kinds of errors should be kept low. This can be done by choosing the proper test and by observing a sufficiently large sample. In order to further develop these ideas we introduce now the notion of a test function.

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a vector of random variables observable for the purpose of testing the hypothesis H_0 against H_1. A function $\phi(\mathbf{X})$ that assumes values in the interval $[0, 1]$ and is a sample statistic is called a **test function**. Using a test function $\phi(\mathbf{X})$ and observing $\mathbf{X} = \mathbf{x}$, the null hypothesis H_0 is rejected with probability $\phi(\mathbf{x})$. This is actually a conditional probability of rejecting H_0, given $\{\mathbf{X} = \mathbf{x}\}$. For a given value of $\phi(\mathbf{x})$, we draw a value R from a table of random numbers, having a rectangular distribution $R(0, 1)$ and reject H_0 if $R \leq \phi(\mathbf{x})$. Such a procedure is called a **randomized test**. If $\phi(\mathbf{x})$ is either 0 or 1, for all $\mathbf{x}$, we call the procedure a **nonrandomized** test. The set of $\mathbf{x}$ values in the sample space $\mathcal{X}$ for which $\phi(\mathbf{x}) = 1$ is called the **rejection region** corresponding to $\phi(\mathbf{x})$.

We distinguish between test functions according to their **size** and **power**. The size of a test function $\phi(\mathbf{x})$ is the maximal probability of error of the first kind, over all the distribution functions F in $\mathcal{F}_0$, i.e., $\alpha = \sup\{E\{\phi(\mathbf{X}) \mid F\} : F \in \mathcal{F}_0\}$ where $E\{\phi(\mathbf{X}) \mid F\}$ denotes the expected value of $\phi(\mathbf{X})$ (the total probability of rejecting H_0) under the distribution F. We denote the size of the test by α. The **power** of a test is the probability of rejecting H_0 when the parent distribution F belongs to $\mathcal{F}_1$. As we vary F over $\mathcal{F}_1$, we can consider the power of a test as a functional

$\psi(F;\phi)$ over $\mathcal{F}_1$. In parametric cases, where each F can be represented by a real or vector valued parameter θ, we speak about a **power function** $\psi(\theta;\phi)$, $\theta \in \Theta_1$, where Θ_1 is the set of all parameter points corresponding to $\mathcal{F}_1$. A test function $\phi^0(\mathbf{x})$ that maximizes the power, with respect to all test functions $\phi(\mathbf{x})$ having the same size, at every point θ, is called **uniformly most powerful** (UMP) of size α. Such a test function is optimal. As will be shown, uniformly most powerful tests exist only in special situations. Generally we need to seek tests with some other good properties. Notice that if the model specifies a family of distributions $\mathcal{F}$ that admits a (nontrivial) sufficient statistic, $T(\mathbf{X})$, then for any specified test function, $\phi(\mathbf{X})$ say, the test function $\hat{\phi}(T) = E\{\phi(\mathbf{X}) \mid T\}$ is equivalent, in the sense that it has the same size and the same power function. Thus, one can restrict attention only to test functions that depend on minimal sufficient statistics.

The literature on testing statistical hypotheses is so rich that there is no point to try and list here even the important papers. The exposition of the basic theory on various levels of sophistication can be found in almost all the textbooks available on Probability and Mathematical Statistics. For an introduction to the asymptotic (large sample) theory of testing hypotheses, see Cox and Hinkley (1974). More sophisticated discussion of the theory is given in Chapter III of Schmetterer (1974). In the following sections we present an exposition of important techniques. A comprehensive treatment of the theory of optimal tests is given in Lehmann (1997).

4.2 THE NEYMAN–PEARSON FUNDAMENTAL LEMMA

In this section we develop the most powerful test of two **simple** hypotheses. Thus, let $\mathcal{F} = \{F_0, F_1\}$ be a family of two specified distribution functions. Let $f_0(x)$ and $f_1(x)$ be the probability density functions (p.d.f.s) corresponding to the elements of $\mathcal{F}$. The null hypothesis H_0 is that the parent distribution is F_0. The alternative hypothesis H_1 is that the parent distribution is F_1. We exclude the problem of testing H_0 at size $\alpha = 0$ since this is obtained by the trivial test function that accepts H_0 with probability one (according to F_0). The following lemma, which is the basic result of the whole theory, was given by Neyman and Pearson (1933).

Theorem 4.2.1. *(The Neyman–Pearson Lemma) For testing H_0 against H_1,*
(a) Any test function of the form

$$\phi^0(X) = \begin{cases} 1, & \text{if } f_1(X) > kf_0(X) \\ \gamma, & \text{if } f_1(X) = kf_0(X) \\ 0, & \text{otherwise} \end{cases} \tag{4.2.1}$$

for some $0 \le k < \infty$ and $0 \le \gamma \le 1$ is most powerful relative to all tests of its size.

(b) (Existence) For testing H_0 against H_1, at a level of significance α there exist constants k_α, $0 \le k_\alpha < \infty$ and γ_α, $0 \le \gamma_\alpha \le 1$ such that the corresponding test function of the form (4.2.1) is most powerful of size α.

(c) (Uniqueness) If a test ϕ' is most powerful of size α, then it is of the form (4.2.1), except perhaps on the set $\{x; f_1(x) = kf_0(x)\}$, unless there exists a test of size smaller than α and power 1.

Proof. (a) Let α be the size of the test function $\phi^0(X)$ given by (4.2.1). Let $\phi^1(x)$ be any other test function whose size does not exceed α, i.e.,

$$E_0\{\phi^1(X)\} \leq \alpha. \tag{4.2.2}$$

The expectation in (4.2.2) is with respect to the distribution F_0. We show now that the power of $\phi^1(X)$ cannot exceed that of $\phi^0(X)$. Define the sets

$$\begin{aligned} R^- &= \{x; f_1(x) < kf_0(x)\} \\ R^0 &= \{x; f_1(x) = kf_0(x)\} \\ R^+ &= \{x; f_1(x) > kf_0(x)\}. \end{aligned} \tag{4.2.3}$$

We notice that $\{R^-, R^0, R^+\}$ is a partition of χ. We prove now that

$$\int_{-\infty}^{\infty} (\phi^1(x) - \phi^0(x)) f_1(x) d\mu(x) \leq 0. \tag{4.2.4}$$

Indeed,

$$\begin{aligned} &\int_{-\infty}^{\infty} (\phi^1(x) - \phi^0(x))(f_1(x) - kf_0(x)) d\mu(x) \\ &\qquad = \left(\int_{R^-} + \int_{R^0} + \int_{R^+} \right) (\phi^1(x) - \phi^0(x))(f_1(x) - kf_0(x)) d\mu(x). \end{aligned} \tag{4.2.5}$$

Moreover, since on R^- the inequality $f_1(x) - kf_0(x) < 0$ is satisfied and $\phi^0(x) = 0$, we have

$$\int_{R^-} (\phi^1(x) - \phi^0(x))(f_1(x) - kf_0(x)) d\mu(x) \leq 0. \tag{4.2.6}$$

Similarly,

$$\int_{R^0} (\phi^1(x) - \phi^0(x))(f_1(x) - kf_0(x)) d\mu(x) = 0 \tag{4.2.7}$$

and since on R^+ $\phi^0(x) = 1$,

$$\int_{R^+} (\phi^1(x) - \phi^0(x))(f_1(x) - kf_0(x)) d\mu(x) \leq 0. \tag{4.2.8}$$

Hence, from (4.2.6)–(4.2.8) we obtain

$$\int_{-\infty}^{\infty}(\phi^1(x)-\phi^0(x))f_1(x)d\mu(x) \le k\int_{-\infty}^{\infty}(\phi^1(x)-\phi^0(x))f_0(x)d\mu(x) \le 0. \quad (4.2.9)$$

The inequality on the RHS of (4.2.9) follows from the assumption that the size of $\phi^0(x)$ is exactly α and that of $\phi^1(x)$ does not exceed α. Hence, from (4.2.9),

$$\int_{-\infty}^{\infty}\phi^1(x)f_1(x)d\mu(x) \le \int_{-\infty}^{\infty}\phi^0(x)f_1(x)d\mu(x). \quad (4.2.10)$$

This proves (a).

(b) (Existence). Consider the distribution $W(\xi)$ of the random variable $f_1(X)/f_0(X)$, which is induced by the distribution F_0, i.e.,

$$W(\xi) = P_0\left\{\frac{f_1(X)}{f_0(X)} \le \xi\right\}. \quad (4.2.11)$$

We notice that $P_0\{f_0(X)=0\}=0$. Accordingly $W(\xi)$ is a c.d.f. The γ-quantile of $W(\xi)$ is defined as

$$W^{-1}(\gamma) = \inf\{\xi; W(\xi) \ge \gamma\}. \quad (4.2.12)$$

For a given value of α, $0<\alpha<1$, we should determine $0 \le k_\alpha < \infty$ and $0 \le \gamma_\alpha \le 1$ so that, according to (4.2.1),

$$\alpha = E_0\{\phi^0(X)\} = 1 - W(k_\alpha) + \gamma_\alpha[W(k_\alpha) - W(k_\alpha - 0)], \quad (4.2.13)$$

where $W(k_\alpha) - W(k_\alpha - 0)$ is the height of the jump of $W(\xi)$ at k_α. Thus, let

$$k_\alpha = W^{-1}(1-\alpha). \quad (4.2.14)$$

Obviously, $0 < k_\alpha < \infty$, since $W(\xi)$ is a c.d.f. of a nonnegative random variable. Notice that, for a given $0<\alpha<1$, $k_\alpha = 0$ whenever

$$P_0\left\{\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} = 0\right\} \ge 1-\alpha.$$

If $W(k_\alpha) - W(k_\alpha - 0) = 0$ then define $\gamma_\alpha = 0$. Otherwise, let γ_α be the unique solution of (4.2.13), i.e.,

$$\gamma_\alpha = \frac{W(k_\alpha) - (1-\alpha)}{W(k_\alpha) - W(k_\alpha - 0)}. \quad (4.2.15)$$

Obviously, $0 \le \gamma_\alpha \le 1$.

(c) (Uniqueness). For a given α, let $\phi^0(X)$ be a test function of the form (4.2.1) with k_α and γ_α as in (4.2.14)–(4.2.15). Suppose that $\phi^1(X)$ is the most powerful test function of size α. From (4.2.9), we have

$$\int (\phi^0(x) - \phi^1(x))(f_1(x) - k_\alpha f_0(x))d\mu(x) \geq 0. \tag{4.2.16}$$

But,

$$\int \phi^0(x) f_0(x)dx = \int \phi^1(x) f_0(x)d\mu(x) \tag{4.2.17}$$

and since ϕ^0 is most powerful,

$$\int \phi^0(x) f_1(x)dx = \int \phi^1(x) f_1(x)d\mu(x).$$

Hence, (4.2.16) equals to zero. Moreover, the integrand on the LHS of (4.2.16) is nonnegative. Therefore, it must be zero for all x except perhaps on the union of R^0 and a set N of probability zero. It follows that on $(R^+ - N) \cup (R^- - N)$, $\phi^0(x) = \phi^1(x)$. On the other hand, if $\phi^1(x)$ has size less than α and power 1, then the above argument is invalid. QED

An extension of the Neyman–Pearson Fundamental Lemma to cases of testing m hypotheses $H_1, \ldots, H_m$ against an alternative H_{m+1} was provided by Chernoff and Scheffé (1952). This generalization provides a most powerful test of H_{m+1} under the constraint that the Type I error probabilities of $H_1, \ldots, H_m$ do not exceed $\alpha_1, \ldots, \alpha_m$, correspondingly where $0 < \alpha_i < 1, i = 1, \ldots, m$. See also Dantzig and Wald (1951).

4.3 TESTING ONE-SIDED COMPOSITE HYPOTHESES IN MLR MODELS

In this section we show that the most powerful tests, which are derived according to the Neyman–Pearson Lemma, can be uniformly most powerful for testing composite hypotheses in certain models. In the following example we illustrate such a case.

A family of distributions $\mathcal{F} = \{F(x;\theta), \theta \in \Theta\}$, where Θ is an interval on the real line, is said to have the **monotone likelihood ratio** property (MLR) if, for every $\theta_1 < \theta_2$ in Θ, the likelihood ratio

$$f(x;\theta_2)/f(x;\theta_1)$$

is a nondecreasing function of x. We also say that $\mathcal{F}$ is an MLR family with respect to X. For example, consider the one-parameter exponential type family with p.d.f.

$$f(x;\theta) = h(x)\exp\{\theta U(x) - K(\theta)\}, \quad -\infty < \theta < \infty.$$

This family is MLR with respect to $U(X)$.

The following important lemma was proven by Karlin (1956).

Theorem 4.3.1 (Karlin's Lemma). *Suppose that* $\mathcal{F} = \{F(x;\theta); -\infty < \theta < \infty\}$ *is an MLR family w.r.t.* x. *If* $g(x)$ *is a nondecreasing function of* x, *then* $E_\theta\{g(X)\}$ *is a nondecreasing function of* θ. *Furthermore, for any* $\theta < \theta'$, $F(x;\theta) \geq F(x;\theta')$ *for all* x.

Proof. (**i**) Consider two points θ, θ' such that $\theta < \theta'$. Define the sets

$$\begin{aligned} A &= \{x; f(x;\theta') < f(x;\theta)\} \\ B &= \{x; f(x;\theta') > f(x;\theta)\} \end{aligned} \tag{4.3.1}$$

where $f(x;\theta)$ are the corresponding p.d.f.s. Since $f(x;\theta')/f(x;\theta)$ is a nondecreasing function of x, if $x \in A$ and $x' \in B$ then $x < x'$. Therefore,

$$a = \sup_{x \in A} g(x) \leq \inf_{x \in B} g(x) = b. \tag{4.3.2}$$

We wish to show that $E_{\theta'}\{g(X)\} \geq E_\theta\{g(X)\}$. Consider,

$$\begin{aligned} &\int g(x)[f(x;\theta') - f(x;\theta)]d\mu(x) \\ &\quad = \int_A g(x)[f(x;\theta') - f(x;\theta)]d\mu(x) + \int_B g(x)[f(x;\theta') - f(x;\theta)]d\mu(x). \end{aligned} \tag{4.3.3}$$

Furthermore, since on the set A $f(x;\theta') - f(x;\theta) < 0$, we have

$$\int_A g(x)[f(x;\theta') - f(x;\theta)]d\mu(x) \geq a \int_A [f(x;\theta') - f(x;\theta)]d\mu(x). \tag{4.3.4}$$

Hence,

$$\begin{aligned} \int g(x)[f(x;\theta') - f(x;\theta)]d\mu(x) &\geq a \int_A [f(x;\theta') - f(x;\theta)]d\mu(x) \\ &\quad + b \int_B [f(x;\theta') - f(x;\theta)]d\mu(x). \end{aligned} \tag{4.3.5}$$

Moreover, for each $\hat{\theta}$,

$$\int_A f(x;\hat{\theta})d\mu(x) + \int_B f(x;\hat{\theta})d\mu(x) = 1 - P_{\hat{\theta}}[f(x;\theta') = f(x;\theta)].$$

In particular,

$$\int_A f(x;\theta')d\mu(x) = -\int_B f(x;\theta')d\mu(x) + 1 - P_{\theta'}\{f(x;\theta') = f(x;\theta)\},$$
$$-\int_A f(x;\theta)d\mu(x) = \int_B f(x;\theta)d\mu(x) - 1 + P_\theta[f(x;\theta') = f(x;\theta)]. \tag{4.3.6}$$

This implies that

$$\int_A [f(x;\theta') - f(x;\theta)]d\mu(x) = -\int_B [f(x;\theta') - f(x;\theta)]d\mu(x). \tag{4.3.7}$$

Moreover, from (4.3.5) and (4.3.7), we obtain that

$$E_{\theta'}\{g(X)\} - E_\theta\{g(X)\} \geq (b-a)\int_B [f(x;\theta') - f(x;\theta)]d\mu(x) \geq 0. \tag{4.3.8}$$

Indeed, from (4.3.2), $(b-a) \geq 0$ and according to the definition of B, $\int_B [f(x;\theta') - f(x;\theta)]d\mu(x) \geq 0$. This completes the proof of part (i).

(**ii**) For any given x, define $\phi_x(y) = I\{y; y > x\}$. $\phi_x(y)$ is a nondecreasing function of y. According to part (i) if $\theta' > \theta$ then $E_\theta\{\phi_x(Y)\} \leq E_{\theta'}\{\phi_x(Y)\}$. We notice that $E_\theta\{\phi_x(Y)\} = P_\theta\{Y > x\} = 1 - F(x;\theta)$. Thus, if $\theta < \theta'$ then $F(x;\theta) \geq F(x;\theta')$ for all x. QED

Theorem 4.3.2. *If a one-parameter family $\mathcal{F} = \{F_\theta(x); -\infty < \theta < \infty\}$ admits a sufficient statistic $T(\mathbf{X})$ and if the corresponding family of distributions of $T(\mathbf{X})$, $\mathcal{F}^T$, is MLR with respect to $T(\mathbf{X})$, then the test function*

$$\phi^0(T(\mathbf{X})) = \begin{cases} 1, & \text{if } T(\mathbf{X}) > k_\alpha \\ \gamma_\alpha, & \text{if } T(\mathbf{X}) = k_\alpha \\ 0, & \text{otherwise} \end{cases} \tag{4.3.9}$$

has the following properties.

- ***(i)*** *It is UMP of its size for testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$, where $-\infty < \theta_0 < \infty$, provided the size of the test is not zero.*
- ***(ii)*** *For every α, $0 < \alpha < 1$, there exist constants k_α, γ_α; $-\infty < k_\alpha < \infty$, $0 \leq \gamma_\alpha \leq 1$, for which the corresponding test function $\phi^0(T(X))$ is UMP of size α.*
- ***(iii)*** *The power function of $\phi^0(T(X))$ is nondecreasing in θ.*

Proof. For simplicity of notation we let $T(x) = x$ (real).

(**i**) From the Neyman–Pearson Lemma, a most powerful test of $H_0^* : \theta = \theta_0$ against $H_1^* : \theta = \theta_1, \theta_1 > \theta_0$ is of the form

$$\phi^0(X) = \begin{cases} 1, & \text{if } \dfrac{f(X;\theta_1)}{f(X;\theta_0)} > k \\ \gamma, & \text{if } \dfrac{f(X;\theta_1)}{f(X;\theta_0)} = k \\ 0, & \text{otherwise} \end{cases} \tag{4.3.10}$$

provided $0 \le k < \infty$. Hence, since $\mathcal{F}$ is an MLR w.r.t. X, $f(X;\theta_1)/f(X;\theta_0) > k$ implies that $X > x_0$. x_0 is determined from the equation $f(x_0;\theta_1)/f(x_0;\theta_0) = k$. Thus, (4.3.9) is also most powerful for testing H_0^* against H_1^* at the same size as (4.3.10). The constants x_0 and γ are determined so that (4.3.9) and (4.3.10) will have the same size. Thus, if α is the size of (4.3.10) then x_0 and γ should satisfy the equation

$$P_{\theta_0}\{X > x_0\} + \gamma P_{\theta_0}\{X = x_0\} = \alpha. \tag{4.3.11}$$

Hence, x_0 and γ may depend only on θ_0, but are independent of θ_1. Therefore, the test function $\phi^0(X)$ given by (4.3.9) is uniformly most powerful for testing H_0^* against H_0. Moreover, since $\phi^0(X)$ is a nondecreasing function of X, the size of the test ϕ^0 (for testing H_0 against H_1) is α. Indeed, from Karlin's Lemma the power function $\psi(\theta;\phi^0) = E_\theta\{\phi^0(X)\}$ is a nondecreasing function of θ (which proves (iii)). Hence, $\sup_{\theta \le \theta_0} E_\theta\{\phi^0(X)\} = \alpha$. Thus, $\phi^0(X)$ is uniformly most powerful for testing H_0 against H_1.

(**ii**) The proof of this part is simple. Given any α, $0 < \alpha < 1$, we set $x^0 = F^{-1}(1-\alpha;\theta_0)$ where $F^{-1}(\gamma,\theta)$ denotes the γ-quantile of $F(x;\theta)$. If $F(x;\theta_0)$ is continuous at x^0, we set $\gamma = 0$, otherwise

$$\gamma = \frac{F(x_0;\theta_0) - (1-\alpha)}{F(x_0;\theta_0) - F(x_0 - 0;\theta_0)}. \tag{4.3.12}$$

QED

4.4 TESTING TWO-SIDED HYPOTHESES IN ONE-PARAMETER EXPONENTIAL FAMILIES

Consider again the one-parameter exponential type family with p.d.f.s

$$f(x;\theta) = h(x)\exp\{\theta U(x) - K(\theta)\}, \quad -\infty < \theta < \infty.$$

A **two-sided** simple hypothesis is $H_0 : \theta = \theta_0$, $-\infty < \theta_0 < \infty$. We consider H_0 against a composite alternative $H_1 : \theta \ne \theta_0$.

If $\mathbf{X} = (X_1, \ldots, X_n)'$ is a vector of independent and identically distributed (i.i.d.) random variables, then the test is based on the minimal sufficient statistic (m.s.s.) $T(\mathbf{X}) = \sum_{i=1}^{n} U(X_i)$. The distribution of $T(\mathbf{X})$, for any θ, is also a one-parameter exponential type. Hence, without loss of generality, we present the theory of this section under the simplified notation $T(X) = X$. We are seeking a test function $\phi^0(X)$ that will have a power function, which is attaining its minimum at $\theta = \theta_0$ and $E_{\theta_0}\{\phi^0(X)\} = \alpha$, for some preassigned level of significance α, $0 < \alpha < 1$. We consider the class of two-sided test functions

$$\phi^0(x) = \begin{cases} 1, & \text{if } x > c_\alpha^{(2)} \\ \gamma_2, & \text{if } x = c_\alpha^{(2)} \\ 0, & \text{if } c_\alpha^{(1)} < x < c_\alpha^{(2)} \\ \gamma_1, & \text{if } x = c_\alpha^{(1)} \\ 1, & \text{if } x < c_\alpha^{(1)}, \end{cases} \tag{4.4.1}$$

where $c_\alpha^{(1)} < c_\alpha^{(2)}$. Moreover, we determine the values of $c_\alpha^{(1)}, \gamma_1, c_\alpha^{(2)}, \gamma_2$ by considering the requirement

$$\begin{aligned} &\textbf{(i)} \quad E_{\theta_0}\{\phi^0(X)\} = \alpha, \\ &\textbf{(ii)} \quad \frac{\partial}{\partial\theta} E_\theta\{\phi^0(X)\}\big|_{\theta=\theta_0} = 0. \end{aligned} \tag{4.4.2}$$

Assume that $\gamma_1 = \gamma_2 = 0$. Then

$$\frac{\partial}{\partial\theta} E_\theta\{\phi^0(X)\} = -K'(\theta) E_\theta\{\phi^0(X)\} + E_\theta\{X\phi^0(X)\}. \tag{4.4.3}$$

Moreover,

$$K'(\theta) = E_\theta\{X\}. \tag{4.4.4}$$

Thus,

$$\frac{\partial}{\partial\theta} E_\theta\{\phi^0(X)\}\big|_{\theta=\theta_0} = -\alpha E_{\theta_0}\{X\} + E_{\theta_0}\{X\phi^0(X)\}. \tag{4.4.5}$$

It follows that condition (ii) of (4.4.2) is equivalent to

$$E_{\theta_0}\{X\phi^0(X)\} = \alpha E_{\theta_0}\{X\}. \tag{4.4.6}$$

It is easy also to check that

$$\frac{\partial^2}{\partial\theta^2} E_\theta\{\phi(X)\}\big|_{\theta=\theta_0} = \alpha V_{\theta_0}\{X\}. \tag{4.4.7}$$

Since this is a positive quantity, the power function assumes its minimum value at $\theta = \theta_0$, provided $\phi^0(X)$ is determined so that (4.4.2) (i) and (4.4.6) are satisfied. As will be discussed in the next section, the two-sided test functions developed in this section are called **unbiased**.

When the family $\mathcal{F}$ is not of the one-parameter exponential type, UMP unbiased tests may not exist. For examples of such cases, see Jogdio and Bohrer (1973).

4.5 TESTING COMPOSITE HYPOTHESES WITH NUISANCE PARAMETERS—UNBIASED TESTS

In the previous section, we discussed the theory of testing composite hypotheses when the distributions in the family under consideration depend on one real parameter. In this section, we develop the theory of most powerful **unbiased** tests of composite hypotheses. The distributions under consideration depend on several real parameters and the hypotheses state certain conditions on some of the parameters. The theory that is developed in this section is applicable only if the families of distributions under consideration have certain structural properties that are connected with sufficiency. The multiparameter exponential type families possess this property and, therefore, the theory is quite useful. First development of the theory was attained by Neyman and Pearson (1933, 1936a, 1936b). See also Lehmann and Scheffé (1950, 1955) and Sverdrup (1953).

Definition 4.5.1. *Consider a family of distributions, $\mathcal{F} = \{F(x;\theta); \theta \in \Theta\}$, where θ is either real or vector valued. Suppose that the null hypothesis is $H_0 : \theta \in \Theta_0$ and the alternative hypothesis is $H_1 : \theta \in \Theta_1$. A test function $\phi(X)$ is called* ***unbiased of size*** *α if*

$$\sup_{\theta\in\Theta_0} E_\theta\{\phi(X)\} = \alpha$$

and

$$E_\theta\{\phi(X)\} \geq \alpha, \quad \textit{for all } \theta \in \Theta_1. \tag{4.5.1}$$

In other words, a test function of size α is unbiased if the power of the test is not smaller than α whenever the parent distribution belongs to the family corresponding to the alternative hypothesis. Obviously, the trivial test $\phi(X) = \alpha$ with probability one is unbiased, since $E_\theta\{\phi(X)\} = \alpha$ for all $\theta \in \Theta_1$. Thus, unbiasedness in itself

is insufficient. However, under certain conditions we can determine uniformly most powerful tests among the unbiased ones. Let Θ^* be the common boundary of the parametric sets Θ_0 and Θ_1 corresponding to H_0 and H_1 respectively. More formally, if $\bar{\Theta}_0$ is the closure of Θ_0 (the union of the set with its limit points) and $\bar{\Theta}_1$ is the closure of Θ_1, then $\Theta^* = \bar{\Theta}_0 \cap \bar{\Theta}_1$. For example, if $\boldsymbol{\theta} = (\theta_1, \theta_2)$, $\Theta_0 = \{\boldsymbol{\theta}; \theta_1 \leq 0\}$ and $\Theta_1 = \{\boldsymbol{\theta}; \theta_1 > 0\}$, then $\Theta^* = \{\boldsymbol{\theta}; \theta_1 = 0\}$. This is the θ_2-axis. In testing two-sided hypotheses, $H_0 : \theta_1^{(1)} \leq \theta_1 \leq \theta_1^{(2)}$ (θ_2 arbitrary) against $H_1 : \theta_1 < \theta_1^{(1)}$ or $\theta_1 > \theta_1^{(2)}$ (θ_2 arbitrary), the boundary consists of the two parallel lines $\Theta^* = \{\boldsymbol{\theta} : \theta_1 = \theta_1^{(1)}$ or $\theta_1 = \theta_1^{(2)}\}$.

Definition 4.5.2. *For testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$, a test $\phi(x)$ is called* ***α-similar*** *if $E_\theta\{\theta(X)\} = \alpha$ for all $\theta \in \Theta_0$. It is called* ***α-similar on the boundary***[1] *if $E_\theta\{\phi(X)\} = \alpha$ for all $\theta \in \Theta^*$.*

Let $\mathcal{F}^*$ denote the subfamily of $\mathcal{F}$, which consists of all the distributions $F(x;\theta)$ where θ belongs to the boundary Θ^*, between Θ_0 and Θ_1. Suppose that $\mathcal{F}^*$ is such that a nontrivial sufficient statistic $T(X)$ with respect to $\mathcal{F}^*$ exists. In this case, $E\{\phi(\mathbf{X} \mid T(\mathbf{X})\}$ is independent of those θ that belong to the boundary Θ^*. That is, this conditional expectation may depend on the boundary, but does not change its value when θ changes over Θ^*. If a test $\phi(\mathbf{X})$ has the property that

$$E\{\phi(\mathbf{X}) \mid T(\mathbf{X})\} = \alpha \text{ with probability } 1 \text{ all } \theta \in \Theta^*, \tag{4.5.2}$$

then $\phi(X)$ is a boundary α-similar test. If a test $\phi(\mathbf{X})$ satisfies (4.5.2), we say that it has the **Neyman structure**. If the power function of an unbiased test function $\phi(\mathbf{X})$ of size α is a continuous function of θ (θ may be vector valued), then $\phi(\mathbf{X})$ is a boundary α-similar test function. Furthermore, if the family of distribution of $T(\mathbf{X})$ on the boundary is boundedly complete, then every boundary α-similar test function has the Neyman structure. Indeed, since $\mathcal{F}_T^*$ is boundedly complete and since every test function is bounded, $E_\theta\{\phi(X)\} = \alpha$ **for all** $\theta \in \Theta^*$ implies that $E\{\phi(X) \mid T(X)\} = \alpha$ with probability 1 for all θ in Θ^*. It follows that if the power function of every unbiased test is continuous in θ, then the class of all test functions having the Neyman structure with some α, $0 < \alpha < 1$, contains all the unbiased tests of size α. Thus, if we can find a UMP test among those having the Neyman structure and if the test is unbiased, then it is **UMP unbiased**. This result can be applied immediately in cases of the k-parameter exponential type families. Express the joint p.d.f. of $\mathbf{X}$ in the form

$$f(x;\theta,\boldsymbol{\nu}) = h(x)\exp\left\{\theta U(x) + \sum_{i=1}^{k} \nu_i T_i(x) - K(\theta,\boldsymbol{\nu})\right\}, \tag{4.5.3}$$

[1] We also call such a test a boundary α-similar test.

where $\nu = (\nu_1, \ldots, \nu_k)'$ is a vector of nuisance parameters and θ is real valued. We consider the following composite hypotheses.

(i) One-sided hypotheses

$$H_0 : \theta \leq \theta_0, \quad \boldsymbol{\nu} \text{ arbitrary},$$

against

$$H_1 : \theta > \theta_0, \quad \boldsymbol{\nu} \text{ arbitrary}.$$

(ii) Two-sided hypotheses

$$H_0 : \theta_1 \leq \theta \leq \theta_2, \quad \boldsymbol{\nu} \text{ arbitrary}$$

against

$$H_1 : \theta < \theta_1 \text{ or } \theta > \theta_2, \quad \boldsymbol{\nu} \text{ arbitrary}.$$

For the one-sided hypotheses, the boundary is

$$\Theta^* = \{(\theta, \boldsymbol{\nu}); \quad \theta = \theta_0, \quad \boldsymbol{\nu} \text{ arbitrary}\}.$$

For the two-sided hypotheses, the boundary is

$$\Theta^* = \{(\theta, \boldsymbol{\nu}); \quad \theta = \theta_1 \text{ or } \theta = \theta_2, \quad \boldsymbol{\nu} \text{ arbitrary}\}.$$

In both cases, the sufficient statistic w.r.t. $\mathcal{F}^*$ is

$$T(\mathbf{X}) = (T_1(\mathbf{X}), \ldots, T_k(\mathbf{X}))'.$$

We can restrict attention to test functions $\phi(U, T)$ since (U, T) is a sufficient statistic for $\mathcal{F}$. The marginal p.d.f. of T is of the exponential type and is given by

$$g(t; \theta, \boldsymbol{\nu}) = \left[\int_{-\infty}^{\infty} k(u, \mathbf{t}) \exp\{\theta u\} d\lambda(u)\right] \cdot \exp\left\{\sum_{i=1}^{k} \nu_i t_i - K(\theta, \boldsymbol{\nu})\right\}, \tag{4.5.4}$$

where $k(u, t) = \int I\{x : U(x) = u, I(x) = \mathbf{t}\}h(x)d\mu(x)$. Hence, the conditional p.d.f. of U given T is a one-parameter exponential type of the form

$$h(u \mid \mathbf{t}, \theta) = k(u, \mathbf{t}) \exp\{\theta u\} / \int_{-\infty}^{\infty} k(u, \mathbf{t}) \exp\{\theta u\} d\lambda(u). \tag{4.5.5}$$

According to the results of the previous section, we construct uniformly most powerful test functions based on the family of conditional distributions, with p.d.f.s (4.5.5). Accordingly, if the hypotheses are one-sided, we construct the conditional test function

$$\phi^0(u \mid \mathbf{t}) = \begin{cases} 1, & \text{if } u > \xi_\alpha(\mathbf{t}) \\ \gamma_\alpha(\mathbf{t}), & \text{if } u = \xi_\alpha(\mathbf{t}) \\ 0, & \text{otherwise,} \end{cases} \tag{4.5.6}$$

where $\xi_\alpha(\mathbf{t})$ and $\gamma_\alpha(\mathbf{t})$ are determined so that

$$E_{\theta_0, \boldsymbol{\nu}}\{\phi(U \mid \mathbf{t}) \mid T(\mathbf{X}) = \mathbf{t}\} = \alpha \tag{4.5.7}$$

for all $\mathbf{t}$. We notice that since $T(\mathbf{X})$ is sufficient for $\mathcal{F}^*$, $\gamma_\alpha(t)$ and $\xi_\alpha(\mathbf{t})$ can be determined independently of $\boldsymbol{\nu}$. Thus, the test function $\phi^0(U \mid T)$ has the Neyman structure. It is a uniformly most powerful test among all tests having the Neyman structure.

In the two-sided case, we construct the conditional test function

$$\phi^0(U \mid T) = \begin{cases} 1, & \text{if } U < \xi_1(T) \text{ or } U > \xi_2(T) \\ \gamma_i(T), & \text{if } U = \xi_i(T), i = 1, 2 \\ 0, & \text{otherwise} \end{cases} \tag{4.5.8}$$

where $\xi_1(T)$, $\xi_2(T)$, $\gamma_1(T)$, and $\gamma_2(T)$ are determined so that

$$E_\theta\{\phi^0(U \mid T) \mid T(\mathbf{X})\} = \alpha$$

with probability one. As shown in the previous section, if in the two-sided case $\theta_1 = \theta_2 = \theta_0$, then we determine $\gamma_i(T)$ and $\xi_i(T)$ $(i = 1, 2)$ so that

$$\begin{aligned} &\textbf{(i)} \quad E_{\theta_0}\{\phi^0(U \mid T) \mid T\} = \alpha \quad \text{w.p.1,} \\ &\textbf{(ii)} \quad E_{\theta_0}\{U\phi^0(U \mid T) \mid T\} = \alpha\, E_{\theta_0}\{U \mid T\} \quad \text{w.p.1,} \end{aligned} \tag{4.5.9}$$

where w.p.1 means "with probability one." The test functions $\phi^0(U \mid T)$ are uniformly most powerful unbiased ones.

The theory of optimal unbiased test functions is strongly reinforced with the following results. Consider first the one-sided hypotheses $H_0 : \theta < \theta_0$, $\boldsymbol{\nu}$ arbitrary; against $H_1 : \theta > \theta_0$, $\boldsymbol{\nu}$ arbitrary. We show that if there exists function $W(U, T)$ that

is increasing in U for each T (U is real valued) and such that $W(U, T)$ and T are independent under H_0, then the test function

$$\phi^0(W) = \begin{cases} 1, & \text{if } W > C_\alpha \\ \gamma_\alpha, & \text{if } W = C_\alpha \\ 0, & \text{otherwise} \end{cases} \tag{4.5.10}$$

is uniformly most powerful unbiased, where C_α and γ_α are determined so that the size of $\phi^0(W)$ is α. Indeed, the power of $\phi^0(W)$ at $(\theta_0, \boldsymbol{\nu})$ is α by construction. Thus,

$$P_{\theta_0,\boldsymbol{\nu}}\{W(U, T) > C_\alpha\} + \gamma_\alpha P_{\theta_0,\boldsymbol{\nu}}\{W(U, T) = C_\alpha\} = \alpha. \tag{4.5.11}$$

Since $W(U, T)$ is independent of T at (θ_0, ν), C_α and γ_α are independent of T. Furthermore, since $W(U, T)$ is an increasing function of U for each T, the test function ϕ^0 is equivalent to the conditional test function (4.5.6). Similarly, for testing the two-sided hypotheses $H_0 : \theta_1 \le \theta \le \theta_2$, $\boldsymbol{\nu}$ arbitrary, we can employ the equivalent test function

$$\phi^0(W) = \begin{cases} 1, & \text{if } W < C_1 \text{ or } W > C_2 \\ \gamma_i, & \text{if } W = C_i, i = 1, 2 \\ 0, & \text{otherwise.} \end{cases} \tag{4.5.12}$$

Here, we require that $W(U, T)$ is independent of T at all the points (θ_1, ν) and (θ_2, ν). When $\theta_1 = \theta_2 = \theta_0$, we require that $W(U, T) = a(T)U + b(T)$, where $a(T) > 0$ with probability one. This linear function of U for each T implies that condition (4.5.9) and the condition

$$\begin{aligned} E_{\theta_0}\{\phi^0(W) \mid T\} &= \alpha \\ E_{\theta_0}\{W\phi(W) \mid T\} &= \alpha\, E_{\theta_0}\{W \mid T\}, \end{aligned} \tag{4.5.13}$$

are equivalent.

4.6 LIKELIHOOD RATIO TESTS

As defined in Section 3.3, the likelihood function $L(\theta \mid \mathbf{x})$ is a nonnegative function on the parameter space Θ, proportional to the joint p.d.f. $f(\mathbf{x}; \theta)$. We discuss here tests of composite hypotheses analogous to the Neyman–Pearson likelihood ratio tests. If H_0 is a specified null hypothesis, corresponding to the parametric set Θ_0 and if Θ is the whole sample space, we define the likelihood ratio statistic as

$$\Lambda(\mathbf{X}_n) = \frac{\sup_{\theta \in \Theta_0} L(\theta \mid \mathbf{X}_n)}{\sup_{\theta \in \Theta} L(\theta \mid \mathbf{X}_n)}. \tag{4.6.1}$$

Obviously, $0 \leq \Lambda(x_n) \leq 1$. A likelihood ratio test is defined as

$$\phi(\mathbf{X}_n) = \begin{cases} 1, & \text{if } \Lambda(\mathbf{X}_n) \leq C_\alpha \\ 0, & \text{otherwise,} \end{cases} \tag{4.6.2}$$

where C_α is determined so that

$$\sup_{\theta \in \Theta_0} P_\theta\{\Lambda(\mathbf{X}_n) \leq C_\alpha\} \leq \alpha. \tag{4.6.3}$$

Due to the nature of the statistic $\Lambda(\mathbf{X}_n)$, its distribution may be discontinuous at $\Lambda = 1$ even if the distribution of $\mathbf{X}_n$ is continuous. For this reason, the test may not exist for every α.

Generally, even if a generalized likelihood ratio test of size α exists, it is difficult to determine the critical level C_α. In Example 4.14 we demonstrate such a case. Generally, for parametric models, the sampling distribution of $\Lambda(\mathbf{X})$, under H_0, can be approximated by simulation. In addition, under certain regularity conditions, if H_0 is a simple hypotheses and $\boldsymbol{\theta}$ is a k-dimensional vector, then the asymptotic distribution of $-2\log\Lambda(X_n)$ as $n \to \infty$ is like that of $\chi^2[m]$, where $m = \dim(\Theta) - \dim(\Theta_0)$, (Wilks, 1962, Chapter 13, Section 13.4). Thus, if the sample is not too small, the $(1-\alpha)$-quantile of $\chi^2[m]$ can provide a good approximation to $-2\log C_\alpha$. In cases of a composite null hypothesis we have a similar result. However, the asymptotic distribution may not be unique.

4.6.1 Testing in Normal Regression Theory

A normal regression model is one in which n random variables $Y_1, \ldots, Y_n$ are observed at n different experimental setups (treatment combinations). The vector $Y_n = (Y_1, \ldots, Y_n)'$ is assumed to have a multinormal distribution $N(X\beta, \sigma^2 I)$, where X is an $n \times p$ matrix of constants with rank $= p$ and $\beta' = (\beta_1, \ldots, \beta_p)$ is a vector of unknown parameters, $1 \leq p \leq n$. The parameter space is $\Theta = \{(\beta_1, \ldots, \beta_p, \sigma); -\infty < \beta_i < \infty \text{ for all } i = 1, \ldots, p \text{ and } 0 < \sigma < \infty\}$. Consider the null hypothesis

$$H_0 : \beta_{r+1} = \cdots = \beta_p = 0, \quad \beta_1, \ldots, \beta_r, \quad \sigma \text{ arbitrary},$$

where $1 \leq r < p$. Thus, $\Theta_0 = \{(\beta_1, \ldots, \beta_r, 0, \ldots, 0, \sigma); -\infty < \beta_i < \infty \text{ for all } i = 1, \ldots, r; 0 < \sigma < \infty\}$. This is the null hypothesis that tests the significance of the $(p-r)$ β-values β_j $(j = r+1, \ldots, p)$. The likelihood function is

$$L(\beta, \sigma \mid \mathbf{Y}, X) = \frac{1}{\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y} - X\beta)'(\mathbf{Y} - X\beta)\right\}.$$

We determine now the values of β and σ for which the likelihood function is maximized, for the given X and $\mathbf{Y}$. Starting with β, we see that the likelihood function is maximized when $Q(\beta) = (\mathbf{Y} - X\beta)'(\mathbf{Y} - X\beta)$ is minimized irrespective of σ. The

vector β that minimizes $Q(\beta)$ is called the **least-squares estimator** of β. Differentiation of $Q(\beta)$ with respect to the vector β yields

$$\nabla Q(\beta) = -2X'(\mathbf{Y} - X\beta). \tag{4.6.4}$$

Equating this gradient vector to zero yields the vector

$$\hat{\beta} = (X'X)^{-1}X'\mathbf{Y}. \tag{4.6.5}$$

We recall that $X'X$ is nonsingular since X is assumed to be of full rank p. Substituting $Q(\hat{\beta})$ in the likelihood function, we obtain

$$L(\hat{\beta}, \sigma) = \frac{1}{\sigma^n} \exp\left\{-\frac{1}{2\sigma^2} Q(\hat{\beta})\right\}, \tag{4.6.6}$$

where

$$Q(\hat{\beta}) = \mathbf{Y}'(I - X(X'X)^{-1}X')\mathbf{Y}, \tag{4.6.7}$$

and $A = I - X(X'X)^{-1}X'$ is a symmetric idempotent matrix. Differentiating $L(\hat{\beta}, \sigma)$ with respect to σ and equating to zero, we obtain that the value σ^2 that maximizes the likelihood function is

$$\hat{\sigma}^2 = \frac{1}{n}\mathbf{Y}'A\mathbf{Y} = Q(\hat{\beta})/n. \tag{4.6.8}$$

Thus, the denominator of (4.6.1) is

$$\sup_{\sigma} \sup_{\beta} L(\beta, \sigma \mid \mathbf{Y}, X) = \frac{1}{\hat{\sigma}^n} \exp\left\{-\frac{n}{2}\right\}. \tag{4.6.9}$$

We determine now the numerator of (4.6.1). Let $K = (0 : I_{p-r})$ be a $(p - r) \times p$ matrix, which is partitioned to a zero matrix of order $(p - r) \times r$ and the identity matrix of order $(p - r) \times (p - r)$. K is of full rank, and $KK' = I_{p-r}$. The null hypothesis H_0 imposes on the linear model the constraint that $K\beta = \mathbf{0}$. Let β^* and $\tilde{\sigma}^2$ denote the values of β and σ^2, which maximize the likelihood function under the constraint $K\beta = 0$. To determine the value of β^*, we differentiate first the Lagrangian

$$D(\beta, \lambda) = (\mathbf{Y} - X\beta)'(\mathbf{Y} - X\beta) + \beta' K\lambda, \tag{4.6.10}$$

where λ is a $(p-r)\times 1$ vector of constants. Differentiating with respect to β, we obtain the simultaneous equations

$$\begin{aligned} &\textbf{(i)} \quad -2X'(\mathbf{Y}-X\beta)+K'\boldsymbol{\lambda}=0,\\ &\textbf{(ii)} \quad K\beta=0. \end{aligned} \tag{4.6.11}$$

From (i), we obtain that the constrained least-squares estimator β^* is given by

$$\beta^*=(X'X)^{-1}\left[X'Y-\frac{1}{2}K'\boldsymbol{\lambda}\right]=\hat{\beta}-\frac{1}{2}(X'X)^{-1}K'\lambda. \tag{4.6.12}$$

Substituting β^* in (4.6.11) (ii), we obtain

$$K\hat{\beta}=\frac{1}{2}K(X'X)^{-1}K'\boldsymbol{\lambda}. \tag{4.6.13}$$

Since K is of full rank $p-r$, $K(X'X)^{-1}K'$ is nonsingular. Hence,

$$\boldsymbol{\lambda}=2[K(X'X)^{-1}K']^{-1}K\hat{\beta}$$

and the constrained least-squares estimator is

$$\beta^*=[I-(X'X)^{-1}K'[K(X'X)^{-1}K']^{-1}K]\hat{\beta}. \tag{4.6.14}$$

To obtain σ^2, we employ the derivation presented before and find that

$$\begin{aligned} \tilde{\sigma}^2&=Q(\beta^*)/n\\ &=\frac{1}{n}[Y-X\hat{\beta}+X(X'X)^{-1}K'B^{-1}K\hat{\beta}]'\\ &\cdot[\mathbf{Y}-X\beta+X(X'X)^{-1}K'B^{-1}K\hat{\beta}], \end{aligned} \tag{4.6.15}$$

where $B=K(X'X)^{-1}K'$. Simple algebraic manipulations yield that

$$\tilde{\sigma}^2=\hat{\sigma}^2+\frac{1}{n}\hat{\beta}'K'B^{-1}K\hat{\beta}. \tag{4.6.16}$$

Hence, the numerator of (4.6.1) is

$$L(\beta^*,\tilde{\sigma}\mid\mathbf{Y},X)=\frac{\exp\left\{-\dfrac{n}{2}\right\}}{\left[\hat{\sigma}^2+\dfrac{1}{n}\hat{\beta}K'B^{-1}K\hat{\beta}\right]^{n/2}}. \tag{4.6.17}$$

The likelihood ratio is then

$$\Lambda(\mathbf{Y}_n) = \left(1 + \frac{1}{n\hat{\sigma}^2}\hat{\beta}' K' B^{-1} K \hat{\beta}\right)^{-n/2}. \tag{4.6.18}$$

This likelihood ratio is smaller than a constant C_α if

$$F = \frac{n-p}{p-r}\hat{\beta}' K' B^{-1} K \hat{\beta} / Q(\hat{\beta}) \tag{4.6.19}$$

is greater than an appropriate constant k_α . In this case, we can easily find the exact distribution of the F-ratio (4.6.19). Indeed, according to the results of Section 2.10, $Q(\hat{\beta}) = \mathbf{Y}' A \mathbf{Y} \sim \sigma^2 \chi^2[n-p]$ since $A = I - X(X'X)^{-1}X'$ is an idempotent matrix of rank $n-p$ and since the parameter of noncentrality is

$$\lambda = \frac{1}{2\sigma^2}\beta' X'(I - X(X'X)^{-1}X')X\beta = 0.$$

Furthermore,

$$\hat{\beta} K' B^{-1} K \hat{\beta} = \mathbf{Y}' X(X'X)^{-1} K' B^{-1} K (X'X)^{-1} X' \mathbf{Y}. \tag{4.6.20}$$

Let $C = X(X'X)^{-1}K'B^{-1}K(X'X)^{-1}X'$. It is easy to verify that C is an idempotent matrix of rank $p-r$. Hence,

$$\hat{\beta}' K' B^{-1} K \hat{\beta} \sim \sigma^2 \chi^2[p-r; \lambda^*], \tag{4.6.21}$$

where

$$\lambda^* = \frac{1}{2\sigma^2}\beta' K' B^{-1} K \beta. \tag{4.6.22}$$

We notice that $K\beta = (\beta_{r+1}, \ldots, \beta_p)'$, which is equal to zero if the null hypothesis is true. Thus, under H_0, $\lambda^* = 0$ and otherwise, $\lambda^* > 0$. Finally,

$$\begin{aligned} CA &= C - X(X'X)^{-1}K'B^{-1}K(X'X)^{-1}X'X(X'X)^{-1}X' \\ &= 0. \end{aligned} \tag{4.6.23}$$

Hence, the two quadratic forms $\mathbf{Y}'A\mathbf{Y}$ and $\mathbf{Y}'C\mathbf{Y}$ are independent. It follows that under H_0, the F ratio (4.6.19) is distributed like a central $F[p-r, n-p]$ statistic, and the critical level k_α is the $(1-\alpha)$-quantile $F_{1-\alpha}[p-r, n-p]$. The power function of the test is

$$\psi(\lambda^*) = P\{F[p-r, n-p; \lambda^*] \geq F_{1-\alpha}[p-r, n-p]\}. \tag{4.6.24}$$

A special case of testing in normal regression theory is the analysis of variance (ANOVA). We present this analysis in the following section.

4.6.2 Comparison of Normal Means: The Analysis of Variance

Consider an experiment in which r independent samples from normal distributions are observed. The basic assumption is that all the r variances are equal, i.e., $\sigma_1^2 = \cdots = \sigma_r^2 = \sigma^2$ ($r \geq 2$). We test the hypothesis $H_0 : \mu_1 = \cdots = \mu_r$, σ^2 arbitrary. The sample m.s.s. is $(\bar{X}_1, \ldots, \bar{X}_r, S_p^2)$, where $\bar{X}_i$ is the mean of the ith sample and S_p^2 is the pooled "within" variance defined in the following manner. Let n_i be the size of the ith sample, $\nu_i = n_i - 1$; S_i^2, the variance of the ith sample; and let $\nu = \sum_{i=1}^{n} \nu_i$. Then

$$S_p^2 = \frac{1}{\nu} \sum_{i=1}^{r} \nu_i S_i^2. \tag{4.6.25}$$

Since the sample means are independent of the sample variances in normal distributions, S_p^2 is independent of $\bar{X}_1, \ldots, \bar{X}_r$. The variance "between" samples is

$$S_b^2 = \frac{1}{r-1} \sum_{i=1}^{r} n_i (\bar{X}_i - \bar{\bar{X}})^2, \tag{4.6.26}$$

where $\bar{\bar{X}} = \sum_{i=1}^{r} n_i \bar{X}_i / \sum_{i=1}^{r} n_i$. $\bar{\bar{X}}$ is the grand mean. Obviously S_p^2 and S_b^2 are independent. Moreover, under H_0, $S_p^2 \sim \frac{\sigma^2}{\nu} \chi^2[\nu]$ and $S_b^2 \sim \frac{\sigma^2}{r-1} \chi^2[r-1]$. Hence, the variance ratio

$$F = S_b^2 / S_p^2 \tag{4.6.27}$$

is distributed, under H_0, like a central $F[r-1, \nu]$ statistic. The hypothesis H_0 is rejected if $F \geq F_{1-\alpha}[r-1, \nu]$. If the null hypothesis H_0 is not true, the distribution of S_b^2 is like that of $\frac{\sigma^2}{r-1} \chi^2[r-1; \lambda]$, where the noncentrality parameter is given by

$$\lambda = \frac{1}{2\sigma^2} \sum_{i=1}^{r} n_i (\mu_i - \mu)^2 \tag{4.6.28}$$

and $\mu = \sum_{i=1}^{r} n_i \mu_i / \sum_{i=1}^{r} n_i$ is a weighted average of the true means. Accordingly, the power of the test, as a function of λ, is

$$\psi(\lambda) = P\{F[r-1, \nu; \lambda] \geq F_{1-\alpha}[r-1, \nu]\}. \tag{4.6.29}$$

This power function can be expressed according to (2.12.22) as

$$\psi(\lambda) = e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} I_{1-R(\xi)}\left(\frac{\nu}{2}, \frac{r-1}{2} + j\right), \tag{4.6.30}$$

where $\xi = F_{1-\alpha}[r-1, \nu]$ and $R(\xi) = \dfrac{r-1}{\nu}\xi \Big/ \left(1 + \dfrac{r-1}{\nu}\xi\right)$.

4.6.2.1 One-Way Layout Experiments

The F-test given by (4.6.27) is a basic test statistic in the analysis of statistical experiments. The method of analysis is known as a **one-way layout analysis of variance** (ANOVA). Consider an experiment in which $N = n \cdot r$ experimental units are randomly assigned to r groups (blocks). Each group of n units is then subjected to a different treatment. More specifically, one constructs a statistical model assuming that the observed values in the various groups are samples of independent random variables having normal distributions. Furthermore, it is assumed that all the r normal distributions have the same variance σ^2 (unknown). The r means are represented by the linear model

$$\mu_i = \mu + \tau_i, \quad i = 1, \ldots, r \tag{4.6.31}$$

where $\sum_{i=1}^{r} \tau_i = 0$. The parameters $\tau_1, \ldots, \tau_r$ represent the incremental effects of the treatments. μ is the (grand) average yield associated with the experiment. Testing whether the population means are the same is equivalent to testing whether **all** $\tau_i = 0$, $i = 1, \ldots, r$. Thus, the hypotheses are

$$H_0 : \sum_{i=1}^{r} \tau_i^2 = 0$$

against

$$H_1 : \sum_{i=1}^{r} \tau_i^2 > 0.$$

We perform the F-test (4.6.27). The parameter of noncentrality (4.6.28) assumes the value

$$\lambda = \frac{n}{2\sigma^2} \sum_{i=1}^{r} \tau_i^2. \tag{4.6.32}$$

4.6.2.2 Two-Way Layout Experiments

If the experiment is designed to test the incremental effects of two factors (drug A and drug B) and their interaction, and if factor A is observed at r_1 levels and factor B at r_2 levels, there should be $s = r_1 \times r_2$ groups (blocks) of size n. It is assumed that these s samples are mutually independent, and the observations within each sample represent i.i.d. random variables having $N(\mu_{ij}, \sigma^2)$ distributions, $i = 1, \ldots, r_1$; $j = 1, \ldots, r_2$. The variances are all the same. The linear model is expressed in the form

$$\mu_{ij} = \mu + \tau_i^A + \tau_j^B + \tau_{ij}^{AB}, \quad i = 1, \ldots, r_1, \quad j = 1, \ldots, r_2,$$

where $\sum_{i=1}^{r_1} \tau_i^A = 0$, $\sum_{j=1}^{r_2} \tau_j^B = 0$, and $\sum_{j=1}^{r_2} \tau_{ij}^{AB} = 0$ for each $i = 1, \ldots, r_1$ and $\sum_{i=1}^{r_1} \tau_{ij}^{AB} = 0$ for each $j = 1, \ldots, r_2$. The parameters τ_i^A are called the **main effects** of factor A; τ_j^B are called the **main effects** of factors B; and τ_{ij}^{AB} are the **interaction** parameters. The hypotheses that one may wish to test are whether the main effects are significant and whether the interaction is significant. Thus, we set up the null hypotheses:

$$\begin{aligned} H_0^{(1)} &: \sum_i \sum_j (\tau_{ij}^{AB})^2 = 0, \\ H_0^{(2)} &: \sum_i (\tau_i^A)^2 = 0, \\ H_0^{(3)} &: \sum_j (\tau_j^B)^2 = 0. \end{aligned} \tag{4.6.33}$$

These hypotheses are tested by constructing F-tests in the following manner. Let X_{ijk}, $i = 1, \ldots, r_1$; $j = 1, \ldots, r_2$; and $k = 1, \ldots, n$ designate the observed random variable (yield) of the kth unit at the (i, j)th group. Let $\bar{X}_{ij}$ denote the sample mean of the (i, j)th group; $\bar{X}_{i\cdot}$, the overall mean of the groups subject to level i of factor A; $\bar{X}_{\cdot j}$, the overall mean of the groups subject to level j of factor B; and $\bar{\bar{X}}$, the grand mean; i.e.,

$$\begin{aligned} \bar{X}_{i\cdot} &= \frac{1}{r_2} \sum_{j=1}^{r_2} \bar{X}_{ij}, \quad i = 1, \ldots, r_1, \\ \bar{X}_{\cdot j} &= \frac{1}{r_1} \sum_{i=1}^{r_1} \bar{X}_{ij}, \quad j = 1, \ldots, r_2, \end{aligned} \tag{4.6.34}$$

and

$$\bar{\bar{X}} = \frac{1}{r_1 r_2} \sum_i \sum_j \bar{X}_{ij}.$$

The sum of squares of deviations around $\bar{\bar{X}}$ is partitioned into four components in the following manner.

$$\begin{aligned}\sum_{i=1}^{r_1}\sum_{j=1}^{r_2}\sum_{k=1}^{n}(X_{ijk} - \bar{\bar{X}})^2 &= \sum_{i=1}^{r_1}\sum_{j=1}^{r_2}\sum_{k=1}^{n}(X_{ijk} - \bar{X}_{ij})^2 \\ &\quad + n\sum_{i=1}^{r_1}\sum_{j=1}^{r_2}(\bar{X}_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{\bar{X}})^2 \qquad (4.6.35) \\ &\quad + nr_2\sum_{i=1}^{r_1}(\bar{X}_{i\cdot} - \bar{\bar{X}})^2 + nr_1\sum_{j=1}^{r_2}(\bar{X}_{\cdot j} - \bar{\bar{X}})^2.\end{aligned}$$

The four terms on the right-hand side of (4.6.35) are mutually independent quadratic forms having distributions proportional to those of central or noncentral chi-squared random variables. Let us denote by Q_r the quadratic form on the left-hand side of (4.6.35) and the terms on the right-hand side (moving from left to right) by Q_W, Q_{AB}, Q_A, and Q_B, respectively. Then we can show that

$$Q_W \sim \sigma^2 \chi^2[\nu_W], \quad \text{where } \nu_W = N - s. \qquad (4.6.36)$$

Similarly,

$$Q_{AB} \sim \sigma^2 \chi^2[\nu_{AB}; \lambda_{AB}], \quad \text{where } \nu_{AB} = (r_1 - 1) \times (r_2 - 1) \qquad (4.6.37)$$

and the parameter of noncentrality is

$$\lambda_{AB} = \frac{n}{2\sigma^2}\sum_{i=1}^{r_1}\sum_{j=1}^{r_2}(\tau_{ij}^{AB})^2. \qquad (4.6.38)$$

Let $S_W^2 = Q_W/\nu_W$ and $S_{AB}^2 = Q_{AB}/\nu_{AB}$. These are the pooled sample variance within groups and the variance between groups due to interaction. If the null hypothesis $H_0^{(1)}$ of zero interaction is correct, then the F-ratio,

$$F = S_{AB}^2/S_W^2, \qquad (4.6.39)$$

is distributed like a central $F[\nu_{AB}, \nu_W]$. Otherwise, it has a noncentral F-distribution as $F[\nu_{AB}, \nu_W; \lambda_{AB}]$. Notice also that

$$E\{S_W^2\} = \sigma^2 \tag{4.6.40}$$

and

$$E\{S_{AB}^2\} = \sigma^2 + n\sigma_{AB}^2, \tag{4.6.41}$$

where

$$\sigma_{AB}^2 = \frac{1}{\nu_{AB}} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} (\tau_{ij}^{AB})^2. \tag{4.6.42}$$

Formula (4.6.42) can be easily derived from (4.6.38) by employing the mixing relationship (2.8.6), $\chi^2[\nu_{AB}; \lambda_{AB}] \sim \chi^2[\nu_{AB} + 2J]$, where J is a Poisson random variable, $P(\lambda_{AB})$. To test the hypotheses $H_0^{(2)}$ and $H_0^{(3)}$, concerning the main effects of A and B, we construct the F-statistics

$$\begin{aligned} F_A &= \frac{S_A^2}{S_W^2}, \\ F_B &= \frac{S_A^2}{S_W^2}, \end{aligned} \tag{4.6.43}$$

where $S_A^2 = Q_A/\nu_A$, $\nu_A = r_1 - 1$ and $S_B^2 = Q_B/\nu_B$, $\nu_B = r_2 - 1$. Under the null hypotheses these statistics have central $F[\nu_A, \nu_W]$ and $F[\nu_B, \nu_W]$ distributions. Indeed, for each $i = 1, \ldots, r_1$, $\bar{X}_{i\cdot} \sim N(\mu + \tau_i^A, \sigma^2/nr_2)$. Hence,

$$Q_A = nr_2 \sum_{i=1}^{r_1} (\bar{X}_i - \bar{\bar{X}})^2 \sim \sigma^2 \chi^2[\nu_A; \lambda_A] \tag{4.6.44}$$

with

$$\lambda_A = \frac{nr_2}{2\sigma^2} \sum_{i=1}^{r_1} (\tau_i^A)^2. \tag{4.6.45}$$

Similarly,

$$Q_B \sim \sigma^2 \chi^2[\nu_B; \lambda_B] \tag{4.6.46}$$

Table 4.1 A Two-Way Scheme for Analysis of Variance

Source	ν	Sum of Squares	MS	F	$E\{MS\}$
Factor A	$r_1 - 1$	Q_A	S_A^2	$\frac{S_A^2}{S_W^2}$	$\sigma^2 + nr_2\sigma_A^2$
Factor B	$r_2 - 1$	Q_B	S_B^2	$\frac{S_B^2}{S_W^2}$	$\sigma^2 + nr_1\sigma_B^2$
Interaction	$(r_1 - 1)(r_2 - 1)$	Q_{AB}	S_{AB}^2	$\frac{S_{AB}^2}{S_W^2}$	$\sigma^2 + n\sigma_{AB}^2$
Between groups	$r_1 r_2 - 1$	$Q_r - Q_W$	–	–	–
Within groups	$N - r_1 r_2$	Q_W	S_W^2	–	σ^2
Total	$N - 1$	Q_r	–	–	–

with

$$\lambda_B = \frac{nr_1}{2\sigma^2} \sum_{j=1}^{r_2} (\tau_j^B)^2. \tag{4.6.47}$$

Under the null hypotheses $H_0^{(2)}$ and $H_0^{(3)}$ both λ_A and λ_B are zero. Thus, the $(1 - \alpha)$-quantiles of the central F-distributions mentioned above provide critical values of the test statistics F_A and F_B. We also remark that

$$\begin{aligned} E\{S_A^2\} &= \sigma^2 + nr_2\sigma_A^2 \\ E\{S_B^2\} &= \sigma^2 + nr_1\sigma_B^2 \end{aligned} \tag{4.6.48}$$

where

$$\begin{aligned} \sigma_A^2 &= \frac{1}{\nu_A} \sum_{i=1}^{r_1} (\tau_i^A)^2, \\ \sigma_B^2 &= \frac{1}{\nu_B} \sum_{j=1}^{r_2} (\tau_j^B)^2. \end{aligned} \tag{4.6.49}$$

These results are customarily summarized in the following table of ANOVA.

Finally, we would like to remark that the three tests of significance provided by F_{AB}, F_A, and F_B are not independent, since the within variance estimator S_W^2 is used by all the three test statistics. Moreover, if we wish that the level of significance of all the three tests simultaneously will not exceed α, we should reduce that of each test to $\alpha/3$. In other words, suppose that $H_0^{(1)}$, $H_0^{(2)}$, and $H_0^{(3)}$ are true and we wish not to reject either one of these. We accept simultaneously the three hypotheses in the event of $\{F_{AB} \le F_{1-\alpha/3}[\nu_{AB}, \nu_W], F_A \le F_{1-\alpha/3}[\nu_A, \nu_B], F_B \le F_{1-\alpha/3}[\nu_B, \nu_W]\}$.

According to the **Bonferroni inequality**, if E_1, E_2, and E_3 are any three events

$$P\{E_1 \cap E_2 \cap E_3\} = 1 - P\{\bar{E}_1 \cup \bar{E}_2 \cup \bar{E}_3\} \geq 1 - P\{\bar{E}_1\} - P\{\bar{E}_2\} - P\{\bar{E}_3\}, \tag{4.6.50}$$

where $\bar{E}_i$ $(i = 1, 2, 3)$ designates the complement of E_i. Thus, the probability that all the three hypotheses will be simultaneously accepted, given that they are all true, is at least $1 - \alpha$. Generally, a scientist will find the result of the analysis very frustrating if all the null hypotheses are accepted. However, by choosing the overall α as sufficiently small, the rejection of any of these hypotheses becomes very meaningful. For further reading on testing in linear models, see Lehmann (1997, Chapter 7), Anderson (1958), Graybill (1961, 1976), Searle (1971) and others.

4.7 THE ANALYSIS OF CONTINGENCY TABLES

4.7.1 The Structure of Multi-Way Contingency Tables and the Statistical Model

There are several qualitative variables $A_1, \ldots, A_k$. The ith variable assumes m_i levels (categories). A sample of N statistical units are classified according to the $M = \prod_{i=1}^{k} m_i$ combinations of the levels of the k variables. These level combinations will be called cells. Let $f(i_1, \ldots, i_k)$ denote the observed frequency in the $(i_1, \ldots, i_k)$ cell. We distinguish between contingency tables having fixed or random marginal frequencies. In this section we discuss only structures with random margins. **The statistical model** assumes that the vector of M frequencies has a **multinomial** distribution with parameters N and P, where P is the vector of cell probabilities $P(i_1, \ldots, i_k)$. We discuss here some methods of testing the significance of the **association** (dependence) among the categorical variables.

4.7.2 Testing the Significance of Association

We illustrate the test for association in a two-way table that is schematized below.

Table 4.2 A Scheme of a Two-Way Contingency Table

	A_1		A_{m_1}	Σ
B_1	$f(1,1)$	.....	$f(1,m_1)$	$f(1,\cdot)$
B_2	$f(2,1)$	.....	$f(2,m_1)$	$f(2,\cdot)$
$\vdots$				
B_{m_2}	$f(m_2,1)$	.....	$f(m_2,m_1)$	$f(m_2,\cdot)$
Σ	$f(\cdot,1)$	.....	$f(\cdot,m_1)$	N

$f(i, j)$ is the observed frequency of the (i, j)th cell. We further denote the observed marginal frequencies by

$$\begin{aligned} f(i, \cdot) &= \sum_{j=1}^{m_1} f(i, j), \quad i = 1, \ldots, m_2, \\ f(\cdot, j) &= \sum_{i=1}^{m_2} f(i, j), \quad j = 1, \ldots, m_1. \end{aligned} \tag{4.7.1}$$

Let

$$\begin{aligned} P(i, \cdot) &= \sum_{j=1}^{m_1} P(i, j), \\ P(\cdot, j) &= \sum_{i=1}^{m_2} P(i, j), \end{aligned} \tag{4.7.2}$$

denote the marginal probabilities.

The categorical variables A and B are independent if and only if $P(i, j) = P(i, \cdot)P(\cdot, j)$ for **all** (i, j). Thus, if A and B are independent, the expected frequency at (i, j) is

$$E(i, j) = NP(i, \cdot)P(\cdot, j). \tag{4.7.3}$$

Since $P(i, \cdot)$ and $P(\cdot, j)$ are unknown, we estimate $E(i, j)$ by

$$\begin{aligned} e(i, j) &= N \frac{f(i, \cdot)}{N} \cdot \frac{f(\cdot, j)}{N} \\ &= f(i, \cdot) f(\cdot, j)/N. \end{aligned} \tag{4.7.4}$$

The deviations of the observed frequencies from the expected are tested for randomness by

$$X^2 = \sum_{i=1}^{m_2} \sum_{j=1}^{m_1} \frac{(f(i, j) - e(i, j))^2}{e(i, j)}. \tag{4.7.5}$$

Simple algebraic manipulations yield the statistic

$$X^2 = \sum_{i} \sum_{j} \frac{f^2(i, j)}{e(i, j)} - N. \tag{4.7.6}$$

We test the hypothesis of no association by comparing X^2 to the $(1 - \alpha)$th quantile of $\chi^2[\nu]$ with $\nu = (m_1 - 1)(m_2 - 1)$ degrees of freedom. We say that the association is

significant if $X^2 \geq \chi^2_{1-\alpha}[\nu]$. **This is a large sample test**. In small samples it may be invalid. There are appropriate test procedures for small samples, especially for 2×2 tables. For further details, see Lancaster (1969, Chapters XI, XII).

4.7.3 The Analysis of 2 × 2 Tables

Consider the following 2×2 table of cell probabilities

S / R	F	P	Σ
W	$P(1,1)$	$P(1,2)$	$P(1,\cdot)$
NW	$P(2,1)$	$P(2,2)$	$P(2,\cdot)$
Σ	$P(\cdot,1)$	$P(\cdot,2)$	1

S and R are two variables (success in a course and race, for example). **The odds ratio** of F/P for W is defined as $P(1,1)/P(1,2)$ and for NW it is $P(2,1)/P(2,2)$.

These odds ratios are also called the relative risks. We say that there is no **interaction** between the two variables if the odds ratios are the same. Define the **cross product ratio**

$$\rho = \frac{P(1,1)}{P(1,2)} \cdot \frac{P(2,1)}{P(2,2)} = \frac{P(1,1)P(2,2)}{P(1,2)P(2,1)}. \tag{4.7.7}$$

If $\rho = 1$ there is no interaction; otherwise, the interaction is negative or positive according to whether $\rho < 1$ or $\rho > 1$, respectively. Alternatively, we can measure the interaction by

$$\omega = \log \rho = \log P(1,1) - \log P(1,2) - \log P(2,1) + \log P(2,2). \tag{4.7.8}$$

We develop now a test of the significance of the interaction, which is valid for any sample size and is a uniformly most powerful test among the unbiased tests.

Consider first the conditional joint distribution of $X = f(1,1)$ and $Y = f(2,1)$ given the marginal frequency $T = f(1,1) + f(1,2)$. It is easy to prove that conditional on T, X and Y are independent and have conditional binomial distributions $B(T, P(1,1)/P(1,\cdot))$ and $B(N-T, P(2,1)/P(2,\cdot))$, respectively. We consider now the conditional distribution of X given the marginal frequencies $T = f(1,\cdot)$ and $S = f(1,1) + f(2,1) = f(\cdot,1)$. This conditional distribution has the p.d.f.

$$P_\rho[X = x \mid T = t, S = s] = \frac{\binom{t}{x}\binom{N-t}{s-x}\rho^x}{\sum_{j=0}^{t \wedge s}\binom{t}{j}\binom{N-t}{s-j}\rho^j}, \tag{4.7.9}$$

where $t \wedge s = \min(t, s)$ and ρ is the interaction parameter given by (4.7.7). The hypothesis of no interaction is equivalent to $H_0 : \rho = 1$. Notice that for $\rho = 1$ the p.d.f. (4.7.9) is reduced to that of the hypergeometric distribution $H(N, T, S)$. We compare the observed value of X to the $\alpha/2$- and $(1 - \alpha/2)$-quantiles of the hypergeometric distribution, as in the case of comparing two binomial experiments. For a generalization to 2^n contingency tables, see Zelen (1972).

4.7.4 Likelihood Ratio Tests for Categorical Data

Consider a two-way layout contingency table with m_1 levels of factor A and m_2 levels of factor B. The sample is of size N. The likelihood function of the vector $\mathbf{P}$ of $s = m_1 \times m_2$ cell probabilities, $P(i, j)$, is

$$L(\mathbf{P}; N, \mathbf{f}) = \prod_{i=1}^{m_1} \prod_{j=1}^{m_2} (P(i, j))^{f(i,j)}, \tag{4.7.10}$$

where $f(i, j)$ are the cell frequencies. The hypothesis of no association, H_0 imposes the linear restrictions on the cell probabilities

$$P(i, j) = P(i, \cdot)P(\cdot, j), \qquad \text{for all } (i, j). \tag{4.7.11}$$

Thus, Θ_0 is the parameter space restricted by (4.7.11), while Θ is the whole space of P. Thus, the likelihood ratio statistic is

$$\Lambda(\mathbf{f}, N) = \frac{\sup_{\Theta_0} \prod_{i=1}^{m_1} \prod_{j=1}^{m_2} [P(i, \cdot)P(\cdot, j)]^{f(i,j)}}{\sup_{\Theta} \prod_{i=1}^{m_1} \prod_{j=1}^{m_2} (P(i, j))^{f(i,j)}}. \tag{4.7.12}$$

By taking the logarithm of the numerator and imposing the constraint that

$$\sum_{i=1}^{m_1} P(i, \cdot) = 1,$$

$$\sum_{j=1}^{m_2} P(\cdot, j) = 1,$$

we obtain by the usual methods that the values that maximize it are

$$\begin{aligned} P(i, \cdot) &= f(i, \cdot)/N, \quad i = 1, \ldots, m_1 \\ P(\cdot, j) &= f(\cdot, j)/N, \quad j = 1, \ldots, m_2. \end{aligned} \tag{4.7.13}$$

Similarly, the denominator is maximized by substituting for $P(i, j)$ the sample estimate

$$P(i, j) = f(i, j)/N, \quad i = 1, \ldots, m_1 \quad j = 1, \ldots, m_2. \tag{4.7.14}$$

We thus obtain the likelihood ratio statistic

$$\Lambda(\mathbf{f}; N) = \prod_{i=1}^{m_1} \prod_{j=1}^{m_2} \left(\frac{f(i, \cdot) f(\cdot, j)}{N f(i, j)} \right)^{f(i,j)}. \tag{4.7.15}$$

Equivalently, we can consider the test statistic $-\log \Lambda(\mathbf{f}; N)$, which is

$$\Lambda^* = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} f(i, j) \log \frac{N f(i, j)}{f(i, \cdot) f(\cdot, j)}. \tag{4.7.16}$$

Notice that Λ^* is the empirical Kullback–Leibler information number to discriminate between the actual frequency distribution $f(i, j)/N$ and the one corresponding to the null hypothesis $f(i, \cdot) f(\cdot, j)/N^2$. This information discrimination statistic is different from the X^2 statistic given in (4.7.6). In large samples, $2\Lambda^*$ has the same asymptotic $\chi^2[\nu]$ distribution with $\nu = (m_1 - 1)(m_2 - 1)$. In small samples, however, it performs differently.

For further reading and extensive bibliography on the theory and methods of contingency tables analysis, see Haberman (1974), Bishop, Fienberg, and Holland (1975), Fienberg (1980), and Agresti (1990). For the analysis of contingency tables from the point of view of information theory, see Kullback (1959, Chapter 8) and Gokhale and Kullback (1978).

4.8 SEQUENTIAL TESTING OF HYPOTHESES

Testing of hypotheses may become more efficient if we can perform the sampling in a sequential manner. After each observation (group of observations) we evaluate the results obtained so far and decide whether to terminate sampling and accept (or reject) the hypothesis H_0, or whether to continue sampling and observe an additional (group of) observation(s). The main problem of sequential analysis then is to determine the "best" stopping rule. After sampling terminates, the test function applied is generally of the generalized likelihood ratio type, with critical levels associated with the stopping rule, as will be described in the sequel. Early attempts to derive sequential testing procedures can be found in the literature on statistical quality control (sampling inspection schemes) of the early 1930s. The formulation of the general theory was given by Wald (1945). Wald's book on sequential analysis (1947) is the first important monograph on the subject. The method developed by Wald is called the **Wald Sequential Probability Ratio Test** (SPRT). Many papers have been written on the subject since Wald's original work. The reader is referred to

the book of Ghosh (1970) for discussion of the important issues and the significant results, as well as notes on the historical development and important references. See also Siegmund (1985). We provide in Section 4.8.1 a brief exposition of the basic theory of the Wald SPRT for testing two simple hypotheses. Some remarks are given about extension for testing composite hypotheses and about more recent development in the literature. In Section 4.8.2, we discuss sequential tests that can achieve power one.

4.8.1 The Wald Sequential Probability Ratio Test

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables. Consider two simple hypotheses H_0 and H_1, according to which the p.d.f.s of these random variables are $f_0(x)$ or $f_1(x)$, respectively. Let $R(X_i) = f_1(X_i)/f_0(X_i)$ $i = 1, 2, \ldots$ be the likelihood ratio statistics. The SPRT is specified by two boundary points A, B, $-\infty < A < 0 < B < \infty$ and the stopping rule, according to which sampling continues as long as the partial sums $S_n = \sum_{i=1}^{n} \log R(X_i)$, $n = 1, 2, \ldots$, lie between A and B. As soon as $S_n \leq A$ or $S_n \geq B$, sampling terminates. In the first case, H_0 is accepted and in the second case, H_1 is accepted. The sample size N is a random variable that depends on the past observations. More precisely, the event $\{N \leq n\}$ depends on $\{X_1, \ldots, X_n\}$ but is independent of $\{X_{n+1}, X_{n+2}, \ldots\}$ for all $n = 1, 2, \ldots$. Such a nonnegative integer random variable is called a **stopping variable**. Let $\mathcal{B}_n$ denote the σ-field generated by the random variables $Z_i = \log R(X_i)$, $i = 1, \ldots, n$. A stopping variable N defined with respect to $Z_1, Z_2, \ldots$ is an integer valued random variable N, $N \geq 1$, such that the event $\{N \geq n\}$ is determined by $Z_1, \ldots, Z_{n-1}$ $(n \geq 2)$. In this case, we say that $\{N \geq n\} \in \mathcal{B}_{n-1}$ and $I\{N \geq n\}$ is $\mathcal{B}_{n-1}$ measurable. We will show that for any pair (A, B), the stopping variable N is finite with probability one. Such a stopping variable is called regular. We will see then how to choose the boundaries (A, B) so that the error probability α and β will be under control. Finally, formulae for the expected sample size will be derived and some optimal properties will be discussed.

In order to prove that the stopping variable N is finite with probability one, we have to prove that

$$\lim_{n \to \infty} P_\theta\{N > n\} = 0, \quad \text{for } \theta = 0, 1. \tag{4.8.1}$$

Equivalently, for a fixed integer r (as large as we wish)

$$\lim_{m \to \infty} P_\theta\{N > mr\} = 0, \quad \text{for } \theta = 0, 1. \tag{4.8.2}$$

For $\theta = 0$ or 1, let

$$\mu(\theta) = E_\theta\{\log R(X)\}, \tag{4.8.3}$$

and

$$D^2(\theta) = \mathrm{Var}_\theta\{\log R(X)\}. \tag{4.8.4}$$

Assume that

$$0 < D^2(\theta) < \infty, \quad \text{for } \theta = 0, 1. \tag{4.8.5}$$

If $D^2(\theta)$ for some θ, then (4.8.1) holds trivially at that θ.

Thus, for any value of θ, the distribution of $S_n = \sum_{i=1}^{n} \log R(X_i)$ is asymptotically normal. Moreover, for each $m = 1, 2, \ldots$ and a fixed integer r,

$$P_\theta[N > mr] \le P_\theta[A < S_r < B, |S_{2r} - S_r| < C, \ldots, |S_{mr} - S_{r(m-1)}| < C], \tag{4.8.6}$$

where $C = |B - A|$.

The variables $S_r, S_{2r} - S_r, \ldots, S_{mr} - S_{(m-1)r}$ are independent and identically distributed. Moreover, by the Central Limit Theorem, if r is sufficiently large,

$$P_\theta\{|S_r| > c\} \approx 2 - \Phi\left(\frac{c}{\sqrt{r}\, D(\theta)} + \sqrt{r}\,\frac{\mu(\theta)}{D(\theta)}\right) - \Phi\left(\frac{c}{\sqrt{r}\, D(\theta)} - \sqrt{r}\,\frac{\mu(\theta)}{D(\theta)}\right). \tag{4.8.7}$$

The RHS of (4.8.7) approaches 1 as $r \to \infty$. Accordingly for any ρ, $0 < \rho < 1$, if r is sufficiently large, then $P_\theta[|S_r| < c] < \rho$. Finally, since $S_{jr} - S_{(j-1)r}$ is distributed like S_r for all $j = 1, 2, \ldots, r$, if r is sufficiently large, then

$$P_\theta[N > mr] < P_\theta[A < S_r < B]\rho^{m-1}. \tag{4.8.8}$$

This shows that $P_\theta[N > n]$ converges to zero at an exponential rate. This property is called the **exponential boundedness** of the stopping variables (Wijsman, 1971). We prove now a very important result in sequential analysis, which is not restricted only to SPRTs.

Theorem 4.8.1 (Wald Theorem). *If N is a regular stopping variable with finite expectation $E_\theta\{N\}$, and if $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables such that $E_\theta\{|X_1| < \infty$, then*

$$E_\theta\left\{\sum_{i=1}^{N} X_i\right\} = \xi(\theta) E_\theta\{N\}, \tag{4.8.9}$$

where

$$\xi(\theta) = E_\theta\{X_i\}.$$

Proof. Without loss of generality, assume that $X_1, X_2, \ldots$ is a sequence of i.i.d. absolutely continuous random variables. Then,

$$E_\theta\left\{\sum_{i=1}^{N} X_i\right\} = \sum_{n=1}^{\infty} \int I\{N = n\} \sum_{j=1}^{n} x_j f(\mathbf{x}_n; \theta) dx_n, \tag{4.8.10}$$

where $f(\mathbf{x}_n; \theta)$ is the joint p.d.f. of $\mathbf{X}_n = (X_1, \ldots, X_n)$. The integral in (4.8.10) is actually an n-tuple integral. Since $E_\theta\{|X_1|\} < \infty$, we can interchange the order of summation and integration and obtain

$$\begin{aligned} E_\theta\left\{\sum_{i=1}^{N} X_i\right\} &= \sum_{j=1}^{\infty} \int x_j \sum_{n=j}^{\infty} I\{N\} f(\mathbf{x}_n; \theta) d\mathbf{x}_n \\ &= \sum_{j=1}^{\infty} E_\theta\{X_j I\{N \geq j\}\}. \end{aligned} \tag{4.8.11}$$

However, the event $\{N \geq j\}$ is determined by $(X_1, \ldots, X_{j-1})$ and is therefore independent of $X_j, X_{j+1}, \ldots$. Therefore, due to the independence of the Xs,

$$\begin{aligned} E_\theta\left\{\sum_{i=1}^{N} X_i\right\} &= \sum_{j=1}^{\infty} E_\theta\{X_i\} P_\theta\{N \geq j\} \\ &= \xi(\theta) \sum_{j=1}^{\infty} P_\theta\{N \geq j\}. \end{aligned} \tag{4.8.12}$$

Finally, since N is a positive integer random variable with finite expectation,

$$E_\theta\{N\} = \sum_{j=1}^{\infty} P_\theta\{N \geq j\}. \tag{4.8.13}$$

QED

From assumption (4.8.5) and the result (4.8.8), both $\mu(\theta)$ and $E_\theta\{N\}$ exist (finite). Hence, for any SPRT, $E_\theta\{S_N\} = \mu(\theta) E_\theta\{N\}$. Let $\pi(\theta)$ denote the probability of accepting H_0. Thus, if $\mu(\theta) \neq 0$,

$$E_\theta\{N\} = \frac{1}{\mu(\theta)}[\pi(\theta) E_\theta\{S_n \mid S_N \leq A\} + (1 - \pi(\theta)) E_\theta\{S_N \mid S_N \geq B\}]. \tag{4.8.14}$$

An approximation to $E_\theta\{N\}$ can then be obtained by substituting A for $E_\theta\{S_N \mid S_N \leq A\}$ and B for $E_\theta\{S_N \mid S_N \geq B\}$. This approximation neglects the excess over the boundaries by S_N. One obtains

$$E_\theta\{N\} \approx \frac{1}{\mu(\theta)}\{\pi(\theta)A + (1-\pi(\theta))B\}. \tag{4.8.15}$$

Error formulae for (4.8.15) can be found in the literature (Ghosh, 1970).

Let α and β be the error probabilities associated with the boundaries A, B and let $A' = \log \frac{\beta}{1-\alpha}$, $B' = \log \frac{1-\beta}{\alpha}$. Let α' and β' be the error probabilities associated with the boundaries A', B'.

Theorem 4.8.2. *If* $0 < \alpha + \beta < 1$ *then*

(i) $\alpha' + \beta' \leq \alpha + \beta$

and

(i) $A' \leq A$, $B' \geq B$.

Proof. For each $n = 1, 2, \ldots$ define the sets

$$\begin{aligned}
A_n &= \{\mathbf{x}_n; A' < S_1 < B', \ldots, A' < S_{n-1} < B', S_n \leq A'\},\\
R_n &= \{\mathbf{x}_n; A' < S_1 < B', \ldots, A' < S_{n-1} < B', S_n \geq B'\},\\
C_n &= \{\mathbf{x}_n; A' < S_1 < B', \ldots, A' < S_{n-1} < B', A' < S_n < B'\}.
\end{aligned}$$

The error probability α' satisfies the inequality

$$\begin{aligned}
\alpha' &= \sum_{n=1}^{\infty} \int_{R_n} \prod_{j=1}^{n} f_0(x_j) d\mu(x_j)\\
&\leq \frac{\alpha}{1-\beta} \sum_{n=1}^{\infty} \int_{R_n} \prod_{j=1}^{n} f_1(x_j) d\mu(x_j) = \frac{\alpha}{1-\beta}(1-\beta').
\end{aligned} \tag{4.8.16}$$

Similarly,

$$\begin{aligned}
\beta' &= \sum_{n=1}^{\infty} \int_{A_n} \prod_{j=1}^{n} f_1(x_j) d\mu(x_j)\\
&\leq \frac{\beta}{1-\alpha} \sum_{n=1} \int_{A_n} \prod_{j=1}^{n} f_0(x_j) d\mu(x_j) = \frac{\beta}{1-\alpha}(1-\alpha').
\end{aligned} \tag{4.8.17}$$

Thus,

$$\frac{\alpha'}{1-\beta'} \leq \frac{\alpha}{1-\beta}, \quad \frac{\beta'}{1-\alpha'} \leq \frac{\beta}{1-\alpha}. \tag{4.8.18}$$

From these inequalities we obtain the first statement of the theorem. To establish (ii) notice that if $\tilde{R}_n = \{\mathbf{x} : A < S_i < B, i = 1, \ldots, n-1, S_n > B\}$, then

$$\begin{aligned} 1 - \beta &= \sum_{n=1}^{\infty} \int_{\tilde{R}_n} \prod_{i=1}^{n} f_1(x_j) d\mu(x_j) \geq e^B \sum_{n=1}^{\infty} \int_{\tilde{R}_n} \prod_{i=1}^{n} f_0(x_j) d\mu(x_j) \\ &= \alpha e^B. \end{aligned} \tag{4.8.19}$$

Hence, $B' = \log \frac{1-\beta}{\alpha} \geq B$. The other inequality is proven similarly. QED

It is generally difficult to determine the values of A and B to obtain the specified error probabilities α and β. However, according to the theorem, if α and β are small then, by considering the boundaries A' and B', we obtain a procedure with error probabilities α' and β' close to the specified ones and total test size $\alpha' + \beta'$ smaller than $\alpha + \beta$. For this reason A' and B' are generally used in applications. We derive now an approximation to the acceptance probability $\pi(\theta)$. This approximation is based on the following important identity.

Theorem 4.8.3 (Wald Fundamental Identity). *Let N be a stopping variable associated with the Wald SPRT and $M_\theta(t)$ be the moment generating function (m.g.f.) of $Z = \log R(X)$. Then*

$$E_\theta\{e^{tS_N}(M_\theta(t))^{-N}\} = 1 \tag{4.8.20}$$

for all t for which $M_\theta(t)$ exists.

Proof.

$$\begin{aligned} E_\theta\{e^{tS_N}(M_\theta(t))^{-N}\} &= \sum_{n=1}^{\infty} E_\theta\{I\{N = n\}e^{tS_n}(M_\theta(t))^{-n}\} \\ &= \lim_{m\to\infty} \sum_{n=1}^{m} E\{(I\{N \geq n\} - I\{N \geq n+1\})e^{tS_n}(M_\theta(t))^{-n}\}. \end{aligned} \tag{4.8.21}$$

Notice that $I\{N \geq n\}$ is $\mathcal{B}_{n-1}$ measurable and therefore, for $n \geq 2$,

$$E_\theta\{I\{N \geq n\}e^{tS_n}(M_\theta(t))^{-n}\} = E_\theta\{I\{N \geq n\}e^{tS_{n-1}}(M_\theta(t))^{-(n-1)}\} \tag{4.8.22}$$

and $E_\theta\{I\{N \geq 1\}e^{tS_1}(M_\theta(t))^{-1}\} = 1$. Substituting these in (4.8.21), we obtain

$$E_\theta\{e^{tS_N}(M(t))^{-N}\} = 1 - \lim_{m\to\infty} E\{I\{N \geq m+1\}e^{tS_m}(M(t))^{-m}\}.$$

Notice that $E\{e^{tS_m}(M(t))^{-m}\} = 1$ for all $m = 1, 2, \ldots$ and all t in the domain of convergence of $M_\theta(t)$. Thus, $\{e^{tS_m}(M_\theta(t))^{-m}, m \geq 1\}$ is uniformly integrable. Finally, since $\lim_{m\to\infty} P\{N > m\} = 0$,

$$\lim_{m\to\infty} E\{I\{N \geq m+1\}e^{tS_m}(M(t))^{-m}\} = 0.$$

QED

Choose $\epsilon > 0$ so that,

$$P_1 = P_\theta\{Z > \epsilon\} > 0 \tag{4.8.23}$$

and

$$P_2 = P_\theta\{Z < -\epsilon\} > 0.$$

Then for $t > 0$, $M_\theta(t) = E_\theta\{e^{tZ}\} \geq P_1 e^{t\epsilon}$. Similarly, for $t < 0$, $M_\theta(t) \geq P_2 e^{-t\epsilon}$. This proves that $\lim_{|t|\to\infty} M_\theta(t) = \infty$. Moreover, for all t for which $M(t)$ exists,

$$\begin{aligned} \frac{d}{dt}M_\theta(t) &= E_\theta\{Ze^{tZ}\}, \\ \frac{d^2}{dt^2}M_\theta(t) &= E_\theta\{Z^2e^{tZ}\} > 0. \end{aligned} \tag{4.8.24}$$

Thus, we deduce that the m.g.f. $M_\theta(t)$ is a strictly convex function of t. The expectation $\mu(\theta)$ is $M'_\theta(0)$. Hence, if $\mu(\theta) > 0$ then $M_\theta(t)$ attains its unique minimum at a negative value t^* and $M_\theta(t^*) < 1$. Furthermore, there exists a value t_0, $-\infty < t_0 < t^* < 0$, at which $M_\theta(t_0) = 1$. Similarly, if $\mu(\theta) < 0$, there exist positive values t^* and t^0, $0 < t^* < t^0 < \infty$, such that $M_\theta(t^*) < 1$ and $M_\theta(t^0) = 1$. In both cases t^* and t_0 are unique.

The fundamental identity can be applied to obtain an approximation for the acceptance probability $\pi(\theta)$ of the SPRT with boundaries A' and B'. According to the fundamental identity

$$\pi(\theta)E_\theta\{e^{t_0(\theta)S_N} \mid S_N \leq A'\} + (1 - \pi(\theta))E_\theta\{e^{t_0(\theta)S_N} \mid S_n \geq B'\} = 1, \tag{4.8.25}$$

where $t_0(\theta) \neq 0$ is the point at which $M_\theta(t) = 1$. The approximation for $\pi(\theta)$ is obtained by substituting in (4.8.25)

$$E_\theta\{e^{t_0(\theta)S_N} \mid S_N \leq A'\} \cong e^{t_0(\theta)A'} = \left(\frac{\beta}{1-\alpha}\right)^{t_0(\theta)},$$

and

$$E_\theta\{e^{t_0(\theta)S_N} \mid S_N \geq B'\} \cong \left(\frac{1-\beta}{\alpha}\right)^{t_0(\theta)}.$$

This approximation yields the formula

$$\pi(\theta) \cong \frac{\left(\frac{1-\beta}{\alpha}\right)^{t_0(\theta)} - 1}{\left(\frac{1-\beta}{\alpha}\right)^{t_0(\theta)} - \left(\frac{\beta}{1-\alpha}\right)^{t_0(\theta)}}, \tag{4.8.26}$$

for all θ such that $\mu(\theta) \neq 0$. If θ_0 is such that $\mu(\theta_0) = 0$, then

$$\pi(\theta_0) \cong \frac{\log\frac{1-\beta}{\alpha}}{\log\frac{1-\beta}{\alpha} - \log\frac{\beta}{1-\alpha}}. \tag{4.8.27}$$

The approximation for $E_\theta\{N\}$ given by (4.8.15) is inapplicable at θ_0. However, at θ_0, Wald's Theorem yields the result

$$E_{\theta_0}\{S_N^2\} = E_{\theta_0}\{N\}E_{\theta_0}\{Z^2\}. \tag{4.8.28}$$

From this, we obtain for θ_0

$$E_{\theta_0}\{N\} \cong \frac{\pi(\theta_0)\left(\log\frac{\beta}{1-\alpha}\right)^2 + (1-\pi(\theta_0))\left(\log\frac{1-\beta}{\alpha}\right)^2}{E_{\theta_0}\{Z^2\}}. \tag{4.8.29}$$

In Example 4.17, we have illustrated the use of the Wald SPRT for testing two composite hypotheses when the interval Θ_0 corresponding to H_0 is separated from the interval Θ_1 of H_1. We obtained a test procedure with very desirable properties by constructing the SPRT for two simple hypotheses, since the family $\mathcal{F}$ of distribution functions under consideration is MLR. For such families we obtain a monotone $\pi(\theta)$ function, with acceptance probability greater than $1-\alpha$ for all $\theta < \theta_0$ and $\pi(\theta) < \beta$ for all $\theta > \theta_1$ (Ghosh, 1970, pp. 100–103). The function $\pi(\theta)$ is called the **operating characteristic** function O.C. of the SPRT. The expected sample size function $E_\theta\{N\}$

increases to a maximum between θ_0 and θ_1 and then decreases to zero again. At $\theta = \theta_0$ and at $\theta = \theta_1$ the function $E_\theta\{N\}$ assumes the smallest values corresponding to all possible test procedures with error probabilities not exceeding α and β. This is the optimality property of the Wald SPRT. We state this property more precisely in the following theorem.

Theorem 4.8.4 (Wald and Wolfowitz). *Consider any SPRT for testing the two simple hypotheses $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ with boundary points (A, B) and error probabilities α and β. Let $E_{\theta_i}\{N\}$, $i = 0, 1$ be the expected sample size. If s is any sampling procedure for testing H_0 against H_1 with error probabilities $\alpha(s)$ and $\beta(s)$ and finite expected sample size $E_{\theta_i}\{N(s)\}$ $(i = 0, 1)$, then $\alpha(s) \leq \alpha$ and $\beta(s) \leq \beta$ imply that $E_{\theta_i}\{N\} \leq E_{\theta_i}\{N(s)\}$, for $i = 0, 1$.*

For the proof of this important theorem, see Ghosh (1970, pp. 93–98), Siegmund (1985, p. 19). See also Section 8.2.3.

Although the Wald SPRT is optimal at θ_0 and at θ_1 in the above sense, if the actual θ is between θ_0 and θ_1, even in the MLR case, the expected sample size may be quite large. Several papers were written on this subject and more general sequential procedures were investigated, in order to obtain procedures with error probabilities not exceeding α and β at θ_0 and θ_1 and expected sample size at $\theta_0 < \theta < \theta_1$ smaller than that of the SPRT. Kiefer and Weiss (1957) studied the problem of determining a sequential test that, subject to the above constraint on the error probabilities, minimizes the maximal expected sample size. They have shown that such a test is a generalized version of an SPRT. The same problem was studied recently by Lai (1973) for normally distributed random variables using the theory of optimal stopping rules. Lai developed a method of determining the boundaries $\{(A_n, B_n), n \geq 1\}$ of the sequential test that minimizes the maximal expected sample size. The theory required for discussing this method is beyond the scope of this chapter. We remark in conclusion that many of the results of this section can be obtained in a more elegant fashion by using the general theory of optimal stopping rules. The reader is referred in particular to the book of Chow, Robbins, and Siegmund (1971). For a comparison of the asymptotic relative efficiency of sequential and nonsequential tests of composite hypotheses, see Berk (1973, 1975). A comparison of the asymptotic properties of various sequential tests (on the means of normal distributions), which combines both the type I error probability and the expected sample size, has been provided by Berk (1976).

PART II: EXAMPLES

Example 4.1. A new drug is being considered for adoption at a medical center. It is desirable that the probability of success in curing the disease under consideration will be at least $\theta_0 = .75$. A random sample of $n = 30$ patients is subjected to a treatment with the new drug. We assume that all the patients in the sample respond to the treatment independently of each other and have the same probability to be cured,

θ. That is, we adopt a Binomial model $B(30, \theta)$ for the number of successes in the sample. The value $\theta_0 = .75$ is the boundary between undesirable and desirable cure probabilities. We wish to test the hypothesis that $\theta \geq .75$.

If the number of successes is large the data support the hypothesis of large θ value. The question is, how small could be the observed value of X, before we should reject the hypothesis that $\theta \geq .75$. If $X = 18$ and we reject the hypothesis then $\alpha(18) = B(18; 30, .75) = .05066$. This level of significance is generally considered sufficiently small and we reject the hypothesis if $X \leq 18$. ■

Example 4.2. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having a common rectangular distribution $R(0, \theta)$, $0 < \theta < \infty$. We wish to test the hypothesis $H_0 : \theta \leq \theta_0$ against the alternative $H_1 : \theta > \theta_0$. An m.s.s. is the sample maximum $X_{(n)}$. Hence, we construct a test function of size α, for some given α in $(0, 1)$, which depends on $X_{(n)}$. Obviously, if $X_{(n)} \geq \theta_0$ we should reject the null hypothesis. Thus, it is reasonable to construct a test function $\phi(X_{(n)})$ that rejects H_0 whenever $X_{(n)} \geq C_\alpha$. C_α depends on α and θ_0, i.e.,

$$\phi(X_{(n)}) = \begin{cases} 1, & \text{if } X_{(n)} \geq C_\alpha \\ 0, & \text{otherwise.} \end{cases}$$

C_α is determined so that the size of the test will be α. At $\theta = \theta_0$,

$$P_{\theta_0}\{X_{(n)} \geq C_\alpha\} = \frac{n}{\theta_0^n} \int_{C_\alpha}^{\theta_0} t^{n-1} dt = 1 - \left(\frac{C_\alpha}{\theta_0}\right)^n .$$

Hence, we set $C_\alpha = \theta_0(1 - \alpha)^{1/n}$. The power function, for all $\theta > \theta_0$, is

$$\psi(\theta) = P_\theta\{X_{(n)} \geq \theta_0(1 - \alpha)^{1/n}\} = 1 - (1 - \alpha)\left(\frac{\theta_0}{\theta}\right)^n .$$

We see that $\psi(\theta)$ is greater than α for all $\theta > \theta_0$. On the other hand, for $\theta \leq \theta_0$, the probability of rejection is

$$E_\theta\{\phi(X_{(n)})\} = 1 - \min\left\{1, \left(\frac{\theta_0}{\theta}\right)^n (1 - \alpha)\right\}.$$

Accordingly, the maximal probability of rejection, when H_0 is true, is α and if $\theta < \theta_0$, the probability of rejection is smaller than α. Obviously, if $\theta \leq \theta_0(1 - \alpha)^{1/n}$, then the probability of rejection is zero. ■

Example 4.3. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a normal distribution $N(\mu, \sigma^2)$. According to the null hypothesis $H_0 : \mu = \mu_1, \sigma = \sigma_1$. According to the

alternative hypothesis $H_1 : \mu = \mu_2, \sigma = \sigma_2; \sigma_2 > \sigma_1$. The likelihood ratio is

$$\begin{aligned}\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} &= \left(\frac{\sigma_1}{\sigma_2}\right)^n \exp\left\{-\frac{1}{2}\sum_{i=1}^n\left[\left(\frac{x_i-\mu_2}{\sigma_2}\right)^2-\left(\frac{x_i-\mu_1}{\sigma_1}\right)^2\right]\right\}\\ &= \left(\frac{\sigma_1}{\sigma_2}\right)^n \exp\left\{-\frac{1}{2}\cdot\frac{\sigma_1^2-\sigma_2^2}{\sigma_1^2\sigma_2^2}\sum_{i=1}^n\left(x_i-\frac{\sigma_1\mu_2+\sigma_2\mu_1}{\sigma_1+\sigma_2}\right)\right.\\ &\quad\left.\cdot\left(x_i+\frac{\sigma_1\mu_2-\sigma_2\mu_1}{\sigma_2-\sigma_1}\right)\right\}.\end{aligned}$$

We notice that the distribution function of $f_1(\mathbf{X})/f_0(\mathbf{X})$ is continuous and therefore $\gamma_\alpha = 0$. According to the Neyman–Pearson Lemma, a most powerful test of size α is obtained by rejecting H_0 whenever $f_1(\mathbf{X})/f_0(\mathbf{X})$ is greater than some positive constant k_α. But, since $\sigma_2 > \sigma_1$, this is equivalent to the test function that rejects H_0 whenever

$$\sum_{i=1}^n\left(X_i-\frac{\sigma_1\mu_2+\sigma_2\mu_1}{\sigma_1+\sigma_2}\right)\left(X_i+\frac{\sigma_1\mu_2-\sigma_2\mu_1}{\sigma_2-\sigma_1}\right)\geq C_\alpha,$$

where C_α is an appropriate constant. Simple algebraic manipulations yield that H_0 should be rejected whenever

$$\sum_{i=1}^n(X_i-\omega)^2\geq C_\alpha^*,$$

where

$$\omega = (\sigma_2^2\mu_1-\sigma_1^2\mu_2)/(\sigma_2^2-\sigma_1^2).$$

We find C_α^* in the following manner. According to H_0,

$$X_i-\omega\sim N(\delta,\sigma_1^2)$$

with $\delta = \sigma_1^2(\mu_2-\mu_1)/(\sigma_2^2-\sigma_1^2)$. It follows that $\sum_{i=1}^n(X_i-\omega)^2\sim\sigma_1^2\chi^2[n;n\delta^2/2\sigma_1^2]$ and thus,

$$C_\alpha^* = \sigma_1^2\chi^2_{1-\alpha}[n;n\delta^2/2\sigma_1^2],$$

where $\chi^2_{1-\alpha}[\nu;\lambda]$ is the $(1-\alpha)$th quantile of the noncentral χ^2. We notice that if $\mu_1 = \mu_2$ but $\sigma_1\neq\sigma_2$, the two hypotheses reduce to the hypotheses $H_0^*:\mu_1=\mu$, $\sigma^2=\sigma_1^2$ versus $H_1^*:\mu_2=\mu$, $\sigma_2\neq\sigma_1$. In this case, $\delta = 0$ and $C_\alpha^* = \sigma_1^2\chi^2_{1-\alpha}[n]$. If $\sigma_1=\sigma_2$ but $\mu_2>\mu_1$ (or $\mu_2<\mu_1$), the test reduces to the t-test of Example 4.9. ■

Example 4.4. In this example, we present a case of testing the shape parameter of a Weibull distribution. This case is important in reliability life testing. We show that even if the problem is phrased in terms of two simple hypotheses, it is not a simple matter to determine the most powerful test function. This difficulty is due to the fact that if the shape parameter is unknown, the minimal sufficient statistic is the order statistic. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common Weibull distribution $G^{1/\nu}(1, 1)$. We wish to test the null hypothesis $H_0 : \nu = 1$ against the simple alternative $H_1 : \nu = 1 + \delta$; δ is a specified positive number. Notice that under H_0, $X_i \sim E(1)$, i.e., exponentially distributed with mean 1. According to the Neyman–Pearson Lemma, the most powerful test of size α rejects H_0 whenever

$$(1+\delta)^n \left(\prod_{i=1}^{n} X_i\right)^{\delta} \exp\left\{-\sum_{i=1}^{n}(X_i^{1+\delta} - X_i)\right\} \geq k_\alpha,$$

where k_α is determined so that if H_0 is correct, then the probability is exactly α. Equivalently, we have to determine a constant c_α so that, under H_0, the probability of

$$\sum_{i=1}^{n}\left[\log X_i - \frac{1}{\delta}\left(X_i^{1+\delta} - X_i\right)\right] \geq C_\alpha$$

is exactly α. Let $W_i(\delta) = \log X_i - \frac{1}{\delta}(X_i^{1+\delta} - X_i)$ and $S_n(\delta) = \sum_{i=1}^{n} W_i(\delta)$. The problem is to determine the distribution of $S_n(\delta)$ under H_0. If n is large, we can approximate the distribution of $S_n(\delta)$ by a normal distribution. The expected value of $W(\delta)$ is

$$\mu_1(\delta) = \frac{1}{\delta} + \Gamma'(1) - \frac{1}{\delta}\Gamma(2+\delta),$$

where $\Gamma'(1)$ is the derivative of $\Gamma(x)$ at $x = 1$. The second moment of $W(\delta)$ is

$$\mu_2(\delta) = \frac{2}{\delta} + \Gamma''(1) + \frac{2}{\delta}(\Gamma'(2) - \Gamma'(2+\delta)) + \frac{1}{\delta^2}(\Gamma(3+2\delta) - 2\Gamma(3+\delta)).$$

Thus, according to the Central Limit Theorem, if n is sufficiently large, then

$$\lim_{n\to\infty} P\left\{\frac{1}{\sqrt{n}}(S_n(\delta) - n\mu_1(\delta)) \leq x[\mu_2(\delta) - \mu_1^2(\delta)]^{1/2}\right\} = \Phi(x).$$

Accordingly, for large values of n, the critical level C_α is approximately

$$C_\alpha \cong n\mu_1(\delta) + z_{1-\alpha}\sqrt{n}\,\{\mu_2(\delta) - \mu_1^2(\delta)\}^{1/2}.$$

For small values of n, we can determine C_α approximately by simulating many replicas of $S_n(\delta)$ values when $X_1, \ldots, X_n$ are $E(1)$ and determining the $(1-\alpha)$th quantile point of the empirical distribution of $S_n(\delta)$. ■

Example 4.5. Consider an experiment in which n Bernoulli trials are performed. Let K denote the number of successes among these trials and let θ denote the probability of success. Suppose that we wish to test the hypothesis

$$H_0 : \theta \leq \theta_0 \text{ against } H_1 : \theta > \theta_0,$$

at level of significance α. θ_0 and α are specified numbers. The UMP (randomized) test function is

$$\phi(K) = \begin{cases} 1, & \text{if } K > \xi_\alpha(\theta_0) \\ \gamma_\alpha, & \text{if } K = \xi_\alpha(\theta_0) \\ 0, & \text{otherwise} \end{cases}$$

where $\xi_\alpha(\theta_0)$ is the $(1-\alpha)$-quantile of the binomial distribution $B(n, \theta_0)$, i.e.,

$$\xi_\alpha(\theta_0) = \text{ least nonnegative integer, } k,$$
$$\text{such that } \sum_{j=0}^{k} b(j; n, \theta_0) \geq 1 - \alpha.$$

Furthermore,

$$\gamma_\alpha = \frac{B(\xi_\alpha(\theta_0); n, \theta_0) - (1-\alpha)}{b(\xi_\alpha(\theta_0); n, \theta_0)}.$$

Accordingly, if the number of successes K is larger than the $(1-\alpha)$-quantile of $B(n, \theta_0)$, we reject H_0. If K equals $\xi_\alpha(\theta_0)$, the null hypothesis H_0 is rejected only with probability γ_α. That is, a random number R having a $R(0, 1)$ distribution is picked from a table of random numbers. If $K = \xi_\alpha(\theta_0)$ and $R \leq \gamma_\alpha$, H_0 is rejected; if $K = \xi_\alpha(\theta_0)$ and $R > \gamma_\alpha$, then H_0 is accepted. If $K < \xi_\alpha(\theta_0)$, H_0 is accepted. It is easy to verify that if $\theta = \theta_0$ then the probability of rejecting H_0 is exactly α. If $\theta < \theta_0$ this probability is smaller than α and, on the other hand, if $\theta > \theta_0$ the probability of rejection is greater than α. The test of this one-sided hypothesis H_0 can be easily performed with the aid of tables of the cumulative binomial distributions. The exact power of the test can be determined according to the formula

$$\psi(\theta) = 1 - B(\xi_\alpha(\theta_0); n, \theta) + \gamma_\alpha \cdot b(\xi_\alpha(\theta_0); n, \theta),$$

where $\theta > \theta_0$. If the hypotheses are one-sided but to the other direction, i.e., $H_0 : \theta \geq \theta_0$ against $H_1 : \theta < \theta_0$, the UMP test is similar. ■

Example 4.6. Consider the family $\mathcal{F}$ of Poisson distributions $P(\theta)$, $0 < \theta < \infty$. The p.d.f.s are

$$f(x;\theta) = e^{-\theta}\theta^x/x! = \frac{1}{x!}\exp\{x\log\theta - \theta\}, \quad x = 0, 1, \ldots.$$

Thus, if we make the reparametrization $\omega = \log\theta$, then

$$f(x;\omega) = \frac{1}{x!}\exp\{x\omega - e^{\omega}\}, \quad x = 0, 1, \ldots; \quad -\infty < \omega < \infty.$$

This is a one-parameter exponential type family. The hypotheses $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ $(0 < \theta_0 < \infty)$ are equivalent to the hypotheses $H_0 : \omega = \omega_0$ against $H_1 : \omega \neq \omega_0$ where $\omega_0 = \log\theta_0$. The two-sided test $\phi^0(X)$ of size α is obtained by (4.4.1), where the constants are determined according to the conditions (4.4.2) and (4.4.6). Since $\mathcal{F}$ is Poisson, $E_{\theta_0}\{X\} = \theta_0$. Moreover, the p.d.f. of $P(\theta)$ satisfies the relation

$$jp(j;\theta) = \theta p(j-1;\theta) \quad \text{for all } j = 1, 2, \ldots.$$

We thus obtain the equations, for $x_1 = c_\alpha^{(1)}$ and $x_2 = c_\alpha^{(2)}$, γ_1 and γ_2:

(i) $P(x_1 - 1;\theta_0) + \gamma_1 p(x_1;\theta_0) + \gamma_2 p(x_2;\theta_0) + 1 - P(x_2;\theta_0) = \alpha,$

(ii) $P(x_1 - 2;\theta_0) + \gamma_1 p(x_1 - 1;\theta_0) + \gamma_2 p(x_2 - 1;\theta_0) + 1 - P(x_2 - 1;\theta_0) = \alpha.$

Here $P(j;\theta)$ is the Poisson c.d.f. The function is zero whenever the argument j is negative. The determination of x_1, γ_1, x_2, γ_2 can be done numerically. We can start with the initial solution x_1, γ_1 and x_2, γ_2 corresponding to the "equal-tail" test. These initial values are determined from the equations

$$P(x_1 - 1;\theta_0) + \gamma_1 p(x_1;\theta_0) = \alpha/2,$$
$$\gamma_2 p(x_2;\theta_0) + 1 - P(x_2;\theta_0) = \alpha/2.$$

This initial solution can then be modified so that both equations (i) and (ii) will be satisfied simultaneously. ■

Example 4.7. Suppose that $X \sim N(\theta, 1)$. The null hypothesis is $H_0 : \theta = 0$. The alternative is $H_1 : \theta \neq 0$. Thus, x_1 and x_2 should satisfy simultaneously the two equations

(I) $\Phi(x_1) + 1 - \Phi(x_2) = \alpha$

(II) $\displaystyle\int_{-\infty}^{x_1} x\phi(x)dx + \int_{x_2}^{\infty} x\phi(x)dx = 0.$

Notice that $x\phi(x) = -\phi'(x)$. Accordingly, equation (II) can be written as

$$(\mathbf{II}) \quad -\phi(x_1) + \phi(x_2) = 0.$$

If we set $x_1 = z_{1-\alpha/2}$ and $x_2 = -x_1$ where $z_\gamma = \Phi^{-1}(\gamma)$ then, due to the symmetry of the $N(0, 1)$ distribution around $\theta = 0$, we obtain that these x_1 and x_2 satisfy simultaneously the two equations. The "equal-tail" solution is the desired solution in this case. ■

Example 4.8. A. Testing the Significance of the Mean in Normal Samples

The problem studied is that of testing hypotheses about the mean of a normal distribution. More specifically, we have a sample $X_1, \ldots, X_n$ of i.i.d. random variables from a normal distribution $N(\mu, \sigma^2)$. We test the hypothesis

$$H_0 : \mu = \mu_0, \quad \sigma^2 \text{ arbitrary}$$

against

$$H_1 : \mu \neq \mu_0, \quad \sigma^2 \text{ arbitrary}.$$

The m.s.s. is $(\bar{X}_n, Q_n)$, where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $Q_n = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. Consider the t-statistic $t = \sqrt{n}(\bar{X}_n - \mu)/S_n$, where $S_n^2 = Q_n/(n-1)$. The t-test of H_0 against H_1 is given by

$$\phi(\bar{X}_n, S_n) = \begin{cases} 1, & \text{if } \sqrt{n}|\bar{X}_n - \mu_0|/S_n \geq t_{1-\alpha/2}[n-1] \\ 0, & \text{otherwise.} \end{cases}$$

$t_{1-\alpha/2}[n-1]$ is the $(1-\alpha/2)$-quantile of the t-distribution with $n-1$ degrees of freedom. It is easy to verify that this t-test has the size α. Its power function can be determined in the following manner. If $\mu \neq \mu_0$ then

$$\begin{aligned} P_{\mu,\sigma}\left\{\sqrt{n}\frac{|\bar{X}_n - \mu_0|}{S_n} \geq t_{1-\alpha/2}[n-1]\right\} \\ = P\{t[n-1; \delta\sqrt{n}] \leq -t_{1-\alpha/2}[n-1]\} \\ + P\{t[n-1; \delta\sqrt{n}] \geq t_{1-\alpha/2}[n-1]\}, \end{aligned}$$

where $\delta = (\mu - \mu_0)/\sigma$. According to (2.12.22), this power function can be computed according to the formula

$$\psi(\delta^2) = 1 - e^{-\frac{n}{2}\delta^2} \sum_{j=0}^{\infty} \frac{(\frac{n}{2}\delta^2)^j}{j!} I_{R(c)}\left(\frac{1}{2} + j, \frac{\nu}{2}\right),$$

where $\nu = n - 1$, $c = t_{1-\alpha/2}[n-1]$ and $R(c) = c^2/(\nu + c^2)$. We notice that the power function depends on δ^2 and is therefore symmetric around $\delta_0 = 0$. We prove now that the t-test is unbiased. Rewrite the power function as a function of $\lambda = \frac{n\delta^2}{2}$ and a mixture of Poisson $P(\lambda)$ with $I\left(\frac{1}{2} + J, \frac{\nu}{2}\right)$, where $J \sim P(\lambda)$ and $R(c) = c^2/(\nu + c^2)$. The family $P(\lambda)$ is MLR in J. Moreover, $I_{R(c)}\left(\frac{1}{2} + j, \frac{\nu}{2}\right)$ is a decreasing function of j. Hence, by Karlin's Lemma, $\psi(\lambda) = 1 - E_\lambda\left\{I_{R(c)}\left(\frac{1}{2} + J, \frac{\nu}{2}\right)\right\}$ is an increasing function of λ. Moreover, $\psi(0) = \alpha$. This proves that the test is unbiased.

B. Testing the Significance of the Sample Correlation

$(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. random vectors having a bivariate normal distribution. Let r be the sample coefficient of correlation (formula 2.13.1). Consider the problem of testing the hypothesis $H_0 : \rho \leq 0$, $(\mu_1, \mu_2, \sigma_1, \sigma_2)$ arbitrary; against $H_1 : \rho > 0$, $(\mu_1, \mu_2, \sigma_1, \sigma_2)$ arbitrary. Here we have four nuisance parameters. As shown in Section 2.15, the distribution of r is independent of the nuisance parameters $(\mu_1, \mu_2, \sigma_1, \sigma_2)$ and when $\rho = 0$ (on the boundary between Θ_0 and Θ_1), it is independent of all the parameters. Moreover, according to (2.13.11), the following test is boundary α-similar.

$$\phi(r) = \begin{cases} 1, & \text{if } \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \geq t_{1-\alpha}[n-2] \\ 0, & \text{otherwise.} \end{cases}$$

The power function depends only on the parameter ρ and is given by

$$\psi(\rho) = P_\rho\left\{r \geq \left(\frac{t^2_{1-\alpha}[n-2]}{n-2+t^2_{1-\alpha}[n-2]}\right)^{1/2}\right\}.$$

According to (2.13.12), this is equal to

$$\psi(\rho) = \frac{2^{n-4}}{\pi(n-3)!}(1-\rho^2)^{\frac{n-1}{2}} \cdot \sum_{j=0}^{\infty} \Gamma\left(\frac{n+j-1}{2}\right)\Gamma\left(\frac{j+1}{2}\right)\Gamma\left(\frac{n}{2}-1\right)\frac{(2\rho)^j}{j!} I_{R(t)}\left(\frac{j+1}{2}, \frac{n}{2}-1\right),$$

where $R(t) = (n-2)/(n-2+t^2_{1-\alpha}[n-2])$. To show that this power function is a monotone nondecreasing function of ρ, one can prove that the family of densities of r under ρ (2.13.12) is an MLR with respect to r. Therefore, according to Karlin's Lemma, $E_\rho\{\phi(r)\}$ is a nondecreasing function of ρ. Thus, the test function $\phi(r)$ is not only boundary α-similar but also unbiased. ■

Example 4.9. Let X and Y be independent r.v.s having Poisson distributions with means λ_1 and λ_2, respectively. We wish to test the hypotheses $H_0 : \lambda_1 = \lambda_2$ against $H_1 : \lambda_1 \neq \lambda_2$. Let $T = X + Y$. The conditional distribution of X given T is the binomial $B(T, p)$ where $p = \lambda_1/(\lambda_1 + \lambda_2)$. The marginal distribution of T is $P(\nu)$ where $\nu = \lambda_1 + \lambda_2$. We can therefore write the joint p.d.f. of X and T in the form

$$p(x, T; \theta, \tau) = \binom{T}{x} \frac{1}{T!} \exp\{\theta X + \tau T - \nu\},$$

where $\theta = \log(\lambda_1/\lambda_2)$ and $\tau = \log \lambda_2$. Thus, the hypotheses under consideration are equivalent to $H_0 : \theta = 0$, τ arbitrary; against $H_1 : \theta \neq 0$, τ arbitrary.

Accordingly, we consider the two-sided test functions

$$\phi^0(X \mid T) = \begin{cases} 1, & \text{if } X < \xi_1(T) \text{ or } X > \xi_2(T) \\ \gamma_i(T), & \text{if } X = \xi_i(T),\ i = 1, 2 \\ 0, & \text{otherwise.} \end{cases}$$

This test is uniformly most powerful unbiased of size α if the functions $\xi_i(T)$ and $\gamma_i(T)$, $i = 1, 2$, are determined according to the conditional distribution of X given T, under H_0. As mentioned earlier, this conditional distribution is the binomial $B(T, \frac{1}{2})$. This is a symmetric distribution around $X_0 = T/2$. In other words, $b(i; T, \frac{1}{2}) = b(T - i; T, \frac{1}{2})$, for all $i = 0, \ldots, T$. Conditions (4.5.9) are equivalent to

$$\textbf{(i)} \quad \sum_{i=0}^{\xi_1 - 1} b\left(i; T, \frac{1}{2}\right) + \gamma_1 b\left(\xi_1; T, \frac{1}{2}\right) + \gamma_2 b\left(\xi_2; T, \frac{1}{2}\right) + \sum_{i=\xi_2+1}^{T} b\left(i; T, \frac{1}{2}\right) = \alpha,$$

$$\textbf{(ii)} \quad \sum_{i=0}^{\xi_1 - 1} i b\left(i; T, \frac{1}{2}\right) + \gamma_1 \xi_1 b\left(\xi_1; T, \frac{1}{2}\right) + \gamma_2 \xi_2 b\left(\xi_2; T, \frac{1}{2}\right) + \sum_{i=\xi_2+1}^{T} i b\left(i; T, \frac{1}{2}\right) = \alpha \cdot \frac{T}{2}.$$

It is easy to verify that, due to the symmetry of the Binomial $B(T, \frac{1}{2})$, the functions that satisfy (i) and (ii) are

$$\xi_1(T) = B^{-1}\left(\frac{\alpha}{2}; T, \frac{1}{2}\right),$$

$$\xi_2(T) = T - \xi_1(T),$$

$$\gamma_1(T) = \frac{\frac{\alpha}{2} - B(\xi_1(T) - 1; T, \frac{1}{2})}{b(\xi_1(T); T, \frac{1}{2})},$$

$$\text{and} \quad \gamma_2(T) = \gamma_1(T).$$

Here $B^{-1}(\frac{\alpha}{2}; T, \frac{1}{2})$ is the $\frac{\alpha}{2}$-quantile of $B(T, \frac{1}{2})$ and $B(j; T, \frac{1}{2})$ is the c.d.f. of $B(T, \frac{1}{2})$ at $X = j$. ■

Example 4.10. In a clinical trial we test the effect of a certain treatment, in comparison to some standard treatment, at two different stations. The null hypothesis is that the effect of the two treatments relative to the control is the same at the two stations. For this objective, a balanced experiment is conducted in which $2n$ patients are tested at each station, n patients with the new treatment and n with the standard one. The observed random variables, X_{ij} $(i = 1, 2;\ j = 1, 2)$ are the number of successes in each sample of n. There are four **independent** binomial random variables. Let θ_{ij} $(i, j = 1, 2)$ denote the probability of success. $i = 1, 2$ denotes the station index and $j = 1, 2$ denotes the treatment index ($j = 1$ for the standard treatment and $j = 2$ for the new treatment). Thus $X_{ij} \sim B(n, \theta_{ij})$. Let $T_i = X_{i1} + X_{i2}$ $(i = 1, 2)$ and

$$\rho_i = \frac{\theta_{i1}}{\theta_{i2}} \cdot \frac{1 - \theta_{i2}}{1 - \theta_{i1}}, \quad i = 1, 2.$$

Let $Y_i = X_{i1}$ $(i = 1, 2)$. The conditional p.d.f. of Y_i given T_i is the confluent hypergeometric function

$$p(y \mid T_i = t) = \frac{\binom{n}{y}\binom{n}{t-y}\rho_i^y}{\sum_{k=0}^{t}\binom{n}{k}\binom{n}{t-k}\rho_i^k}, \quad y = 0, \ldots, t,$$

where generally $\binom{a}{b} = 0$ if $b > a$. We notice that when $\rho_i = 1$ (i.e., $\theta_{i1} = \theta_{i2}$), then the p.d.f. is the hypergeometric p.d.f. $h(y \mid 2n, n, t)$ as given by (2.3.6). Thus, since Y_1 and Y_2 are independent, the joint conditional p.d.f. of (Y_1, Y_2) given $T_1 = t$ and $T_2 = v$ under (ρ_1, ρ_2) is

$$p(y_1, y_2 \mid T_1 = t, T_2 = v) = \frac{\binom{n}{y_1}\binom{n}{y_2}\binom{n}{t-y_1}\binom{n}{v-y_2}\rho_1^{y_1}\rho_2^{y_2}}{\sum_{k_1=0}^{t}\sum_{k_2=0}^{v}\binom{n}{k_1}\binom{n}{k_2}\binom{n}{t-k_1}\binom{n}{v-k_2}\rho_1^{k_1}\rho_2^{k_2}},$$

$$y_1 = 0, \ldots, t,\ \ y_2 = 0, \ldots, v.$$

We consider the problem of testing the hypotheses:

$$H_0 : \rho_1 = \rho_2 \ \text{ against } \ H_1 : \rho_1 \neq \rho_2.$$

Our hypothesis H_0 means that there is no interaction between the effect of the treatment and that of the station. We notice now that under H_0, $S = Y_1 + Y_2$ is a sufficient statistic for the family of joint conditional distributions given T_1 and T_2.

Furthermore, the conditional p.d.f. of Y_1 given T_1, T_2, and S is

$$p(y \mid T_1 = t, T_2 = v, S = k) = \frac{\binom{n}{y}\binom{n}{k-y}\binom{n}{t-y}\binom{n}{v-k+j}\omega^y}{\sum_{j=0}^{k}\binom{n}{j}\binom{n}{k-j}\binom{n}{t-j}\binom{n}{v-k+j}\omega^j},$$

$$y = 0, \ldots, k$$

where $\omega = \rho_1/\rho_2$. The family of all the conditional distributions of Y_1 given (T_1, T_2, S) is an MLR family w.r.t. Y_1. The hypotheses $H_0 : \rho_1 = \rho_2$ against $H_1 : \rho_1 \neq \rho_2$ are equivalent to the hypotheses $H_0 : \omega = 1$ against $H_1 : \omega \neq 1$. Accordingly, the conditional test function

$$\phi(Y_1 \mid T_1, T_2, S) = \begin{cases} 1, & \text{if } Y_1 < \xi_1(T_1, T_2, S) \text{ or } Y_1 > \xi_2(T_1, T_2, S) \\ \gamma_i, & \text{if } Y_1 = \xi_i(T_1, T_2, S),\ i = 1, 2 \\ 0, & \text{otherwise,} \end{cases}$$

is uniformly most powerful unbiased of size α, if the functions $\xi_i(T_1, T_2, S)$ and $\gamma_i(T_1, T_2, S)$ are determined to satisfy conditions (i) and (ii) of (4.5.9) simultaneously. To prove it, we have to show that the family of conditional joint distributions of S given (T_1, T_2) is complete and that the power function of every test function is continuous in $(\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$. This is left to the reader as an exercise. For the computation of the power function and further investigation, see Zacks and Solomon (1976). ■

Example 4.11. A. In this example we show that the t-test, which was derived in Example 4.9, is uniformly most powerful unbiased. An m.s.s. for the family of normal distributions $\mathcal{F} = \{N(\mu, \sigma^2); -\infty < \mu < \infty, 0 < \sigma < \infty\}$ is $(\Sigma X_i, \Sigma X_i^2)$. Let $U = \frac{1}{n}\Sigma X_i$ and $T = \Sigma X_i^2$. We notice that T is an m.s.s. for $\mathcal{F}^*$ (the family restricted to the boundary, $\mu = 0$). Consider the statistic $W = \sqrt{n}\ U / \left(\frac{1}{n-1}T - nU^2\right)^{1/2}$. We notice that if $\mu = 0$, then $W \sim t[n-1]$ independently of σ^2. On the other hand, $T \sim \sigma^2\chi^2[n]$ when $\mu = 0$. Therefore, according to Basu's Theorem, W and T are independent for each $\theta \in \Theta^*$ (the boundary) since the family $\mathcal{F}^T$ is complete. Furthermore, W is an increasing function of U for each T. Hence, the t-test is uniformly most powerful unbiased.

B. Consider part (B) of Example 4.9. The m.s.s. is $(\Sigma X_i, \Sigma X_i^2, \Sigma Y_i, \Sigma Y_i^2, \Sigma X_i Y_i)$. If we denote by $\mathcal{F}^*$ the family of all bivariate normal distributions with $\rho = 0$ (corresponding to the boundary), then $T = (\Sigma X_i, \Sigma X_i^2, \Sigma Y_i, \Sigma Y_i^2)$ is an m.s.s. for $\mathcal{F}^*$. Let $U = \Sigma X_i Y_i$. The sample correlation coefficient r is given by

$$r = W(U, T) = [nU - (\Sigma X_i)(\Sigma Y_i)]/[n\Sigma X_i^2 - (\Sigma X_i)^2]^{1/2} \cdot [n\Sigma Y_i^2 - (\Sigma Y_i)^2]^{1/2}.$$

This function is increasing in U for each T. We notice that the distribution of r is independent of $\nu = (\mu_1, \mu_2, \sigma_1, \sigma_2)$. Therefore, r is independent of T for each $\boldsymbol{\nu}$ whenever $\rho = 0$. The test function $\phi(r)$ of Example 4.9 is uniformly most powerful unbiased to test $H_0 : \rho \leq 0$, ν arbitrary, against $H_1 : \rho > 0$, $\boldsymbol{\nu}$ arbitrary. ■

Example 4.12. Consider again the components of variance Model II of Analysis of Variance, which is discussed in Example 3.9. Here, we have a three-parameter family of normal distributions with parameters μ, σ^2, and τ^2. We set $\rho = \tau^2/\sigma^2$.

A. For testing the hypotheses

$$H_0 : \mu \leq 0, \quad \boldsymbol{\nu} = (\sigma^2, \rho) \text{ arbitrary},$$

against

$$H_1 : \mu > 0, \quad \boldsymbol{\nu} = (\sigma^2, \rho) \text{ arbitrary},$$

the t-test

$$\phi(W) = \begin{cases} 1, & \text{if } \dfrac{\sqrt{nr}\ \bar{\bar{X}}}{\left(\dfrac{n}{r-1}\displaystyle\sum_{i=1}^{r}(\bar{X}_i - \bar{\bar{X}})^2\right)^{1/2}} \geq t_{1-\alpha}[r-1] \\ 0, & \text{otherwise} \end{cases}$$

is a uniformly most powerful unbiased one. Indeed, if we set $U = T_3(\mathbf{X}) = \bar{\bar{X}}$, $T = (T_1(\mathbf{X}), T_2(X))$, then $W(U, T) = \sqrt{nr}\ U / \left[\dfrac{n}{r-1}\displaystyle\sum_{i=1}^{r}(\bar{X}_i - \bar{\bar{X}})^2\right]^{1/2}$ is distributed when $\mu = 0$, as $t[r-1]$ for all (σ^2, ρ). The exponential family is complete. Hence, $W(U, T)$ and T are independent for each (σ^2, ρ) when $\mu = 0$. Furthermore, $W(U, T)$ is an increasing function of U for each T.

B. For testing the hypotheses

$$H_0 : \rho \leq 1, \quad (\sigma^2, \mu) \text{ arbitrary}$$

against

$$H_1 : \rho > 1, \quad (\sigma^2, \mu) \text{ arbitrary}$$

the test function

$$\phi(W) = \begin{cases} 1, & \text{if } W \geq F_{1-\alpha}[r-1, r(n-1)] \\ 0, & \text{otherwise} \end{cases}$$

is uniformly most powerful unbiased. Here

$$W = nr(n-1)\sum_{i=1}^{r}(\bar{X}_i - \bar{\bar{X}})^2/(r-1)\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2,$$

and $F_{1-\alpha}[r-1, r(n-1)]$ is the $(1-\alpha)$-quantile of the central F-distribution with $(r-1)$ and $r(n-1)$ degrees of freedom. ■

Example 4.13. Let $X \sim N(\theta, 1)$. We consider the two simple hypotheses $H_0 : \theta = 0$ versus $H_1 : \theta = 1$. The statistic $\Lambda(X)$ is

$$\Lambda(X) = \frac{f(X;0)}{\max\{f(X;0),\, f(X;1)\}} = 1/\max(1, e^{X-1/2}).$$

Obviously, $\Lambda(X) = 1$ if and only if $X \leq \frac{1}{2}$. It follows that, under $\theta = 0$, $P_0[\Lambda(X) = 1] = \Phi(\frac{1}{2}) = .691$. Therefore, in this example, the generalized likelihood ratio test can be performed only for $\alpha \leq 1 - .691 = .309$ or for $\alpha = 1$. This is a restriction on the generalized likelihood ratio test. However, generally we are interested in small values of α, for which the test exists. ■

Example 4.14. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common Laplace distribution with p.d.f.

$$f(x;\theta) = \frac{1}{2}\exp\{-|x-\theta|\}, \quad -\infty < x < \infty,$$

where the parameter space is $\Theta = \{-\infty < \theta < \infty\}$. We wish to test

$$H_0 : \theta = 0$$

against

$$H_1 : \theta \neq 0.$$

The sample statistic $\hat{\theta}$, which minimizes $\sum_{i=1}^{n}|x_i - \theta|$, is the sample median

$$M_e = \begin{cases} X_{(m+1)}, & \text{if } n = 2m+1 \\ \frac{1}{2}(X_{(m)} + X_{(m+1)}), & \text{if } n = 2m. \end{cases}$$

Thus, the generalized likelihood statistic is

$$\lambda(\mathbf{X}_n) = \exp\left\{-\left(\sum_{i=1}^{n}|x_i| - \sum_{i=1}^{n}|x_i - M_e|\right)\right\}.$$

$\lambda(\mathbf{X}_n)$ is sufficiently small if

$$T(\mathbf{X}_n) = \sum_{i=1}^{n}|X_i| - \sum_{i=1}^{n}|X_i - M_e|$$

is sufficiently large. To obtain a size α test, we need the $(1-\alpha)$-quantile of the sampling distribution of $T(\mathbf{X}_n)$ under $\theta = 0$. $M = 1000$ simulation runs, using S-PLUS, gave the following estimates of the .95th quantile of $T(\mathbf{X}_n)$.

Notice that $-2\log\Lambda(\mathbf{X}_n) = 2 \cdot T(\mathbf{X}_n)$ and that $\chi^2_{.95}[1] = 3.8415$. Also $2\hat{T}_{.95,n} \cong \chi^2_{.95}[1]$.

n	20	50	100
$\hat{T}_{.95,n}$	1.9815	2.0013	1.9502

■

Example 4.15. Fleiss (1973, p. 131) gave the following 2×2 table of G-6-PD deficiency (A) and type of schizophrenia (B) among $N = 177$ patients.

B / A	Catatonic	Paranoid	Σ
Deficient	15	6	21
Non-deficient	57	99	156
Σ	72	105	177

We test whether the association between the two variables is significant. The X^2 statistic for this table is equal to 9.34. This is greater than $\chi^2_{.95}[1] = 3.84$ and therefore significant at the $\alpha = .05$ level. To perform the conditional test we compute the hypergeometric distribution $H(N, T, S)$ with $N = 177$, $T = 21$, and $S = 72$. In Table 4.3, we present the p.d.f. $h(x; N, T, S)$ and the c.d.f. $H(x; N, T, S)$ of this distribution.

According to this conditional distribution, with $\alpha = .05$, we reject H_0 whenever $X \leq 4$ or $X \geq 14$. If $X = 5$ we reject H_0 only with probability $\gamma_1 = .006$. If $X = 13$ we reject H_0 with probability $\gamma_2 = .699$. In this example, $X = 15$ and therefore we conclude that the association is significant. ■

Table 4.3 The Hypergeometric Distribution $H(177, 21, 72)$

x	$h(x; N, T, S)$	$H(x; N, T, S)$
0	0.000007	0.000007
1	0.000124	0.000131
2	0.001022	0.001153
3	0.005208	0.006361
4	0.018376	0.024736
5	0.047735	0.072471
6	0.094763	0.167234
7	0.147277	0.314511
8	0.182095	0.496607
9	0.181006	0.677614
10	0.145576	0.823190
11	0.095008	0.918198
12	0.050308	0.968506
13	0.021543	0.990049

Example 4.16. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables having a common normal distribution $N(\theta, 1)$, $-\infty < \theta < \infty$. Suppose that for testing the hypothesis $H_0 : \theta \leq 0$ against $H_1 : \theta \geq 1$, we construct the Wald SPRT of the two simply hypotheses $H_0^* : \theta = 0$ against $H_1^* : \theta = 1$ with boundaries A' and B' corresponding to $\alpha = .05$ and $\beta = .05$.

Notice that

$$Z = \log \frac{f_1(X)}{f_0(X)} = -\frac{1}{2}[(X-1)^2 - X^2] = X - 1/2.$$

Accordingly,

$$\mu(\theta) = E_\theta \left\{ X - \frac{1}{2} \right\} = \theta - \frac{1}{2}.$$

The m.g.f. of Z at θ is

$$M_\theta(t) = E_\theta\{e^{t(X-\frac{1}{2})}\} = \exp\left\{ \frac{t^2}{2} + \left(\theta - \frac{1}{2}\right) t \right\}.$$

Thus, $t_0(\theta) = 1 - 2\theta$, and from (4.8.26)–(4.8.27), the acceptance probabilities are

$$\pi(\theta) \cong \begin{cases} \dfrac{19^{1-2\theta} - 1}{19^{1-2\theta} - 19^{2\theta-1}}, & \theta \neq .5 \\ .5 & \theta = .5. \end{cases}$$

In the following table we present some of the $\pi(\theta)$ and $E_\theta\{N\}$ values, determined according to the approximations (4.8.15) and (4.8.26).

θ	-1	$-.5$	0.	.25	.50	.75	1	1.5	2.0
$\pi(\theta)$	.99985	.99724	.95000	.81339	.50000	.18601	.05000	.00276	.00015
$E_\theta\{N\}$	2.0	2.9	5.3	7.4	8.7	7.4	5.3	2.9	2.0

The number of observations required in a **fixed** sample design for testing $H_0^*: \theta = 0$ against $H_1^*: \theta = 1$ with $\alpha = \beta = .05$ is $n = 16$. According to the above table, the expected sample size in a SPRT when $\theta = 0$ or $\theta = 1$ is only one third of that required in a fixed sample testing. ■

PART III: PROBLEMS

Section 4.1

4.1.1 Consider Example 4.1. It was suggested to apply the test statistic $\phi(X) = I\{X \leq 18\}$. What is the power of the test if (i) $\theta = .6$; (ii) $\theta = .5$; (iii) $\theta = .4$? [Hint: Compute the power exactly by applying the proper binomial distributions.]

4.1.2 Consider the testing problem of Example 4.1 but assume that the number of trials is $n = 100$.

- **(i)** Apply the normal approximation to the binomial to develop a large sample test of the hypothesis $H_0: \theta \geq .75$ against the alternative $H_1: \theta < .75$.
- **(ii)** Apply the normal approximation to determine the power of the large sample test when $\theta = .5$.
- **(iii)** Determine the sample size n according to the large sample formulae so that the power of the test, when $\theta = .6$, will not be smaller than 0.9, while the size of the test will not exceed $\alpha = .05$.

4.1.3 Suppose that X has a Poisson distribution with mean λ. Consider the hypotheses $H_0: \lambda = 20$ against $H_1: \lambda \neq 20$.

- **(i)** Apply the normal approximation to the Poisson to develop a test of H_0 against H_1 at level of significance $\alpha = .05$.
- **(ii)** Approximate the power function and determine its value when $\lambda = 25$.

Section 4.2

4.2.1 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common negative-binomial distribution $NB(\psi, \nu)$, where ν is known.

- **(i)** Apply the Neyman–Pearson Lemma to derive the MP test of size α of $H_0: \psi \leq \psi_0$ against $H_1: \psi > \psi_1$, where $0 < \psi_0 < \psi_1 < 1$.

(ii) What is the power function of the test?

(iii) What should be the sample size n so that, when $\psi_0 = .05$ and $\alpha = .10$, the power at $\psi = .15$ will be $1 - \beta = .80$?

4.2.2 Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having a common distribution F_θ belonging to a regular family. Consider the two simple hypotheses $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1;\ \theta_0 \neq \theta_1$. Let $\sigma_i^2 = \text{var}_{\theta_i}\{\log(f(X_1;\theta_1)/f(X_1;\theta_0))\},\ i = 0, 1$, and assume that $0 < \sigma_i^2 < \infty$, $i = 0, 1$. Apply the Central Limit Theorem to approximate the MP test and its power in terms of the Kullback–Leibler information functions $I(\theta_0, \theta_1)$ and $I(\theta_1, \theta_0)$, when the sample size n is sufficiently large.

4.2.3 Consider the one-parameter exponential type family with p.d.f. $f(x;\psi) = h(x)\exp\{\psi x_1 - K(\psi)\}$, where $K(\psi)$ is strictly convex having second derivatives at all $\psi \in \Omega$, i.e., $K''(\psi) > 0$ for all $\psi \in \Omega$, where Ω is an open interval on the real line. For applying the asymptotic results of the previous problem to test $H_0 : \psi \leq \psi_0$ against $H_1 : \psi \geq \psi_1$, where $\psi_0 < \psi_1$, show

(i) $E_{\psi_i}\{\log(f(X;\psi_1)/f(X;\psi_0))\} = K'(\psi_i)\cdot(\psi_1 - \psi_0) - (K(\psi_1) - K(\psi_0));\ i = 0, 1.$

(ii) $\text{Var}_{\psi_i}(\log(f(X;\psi_1)/f(X;\psi_0))\} = (\psi_1 - \psi_0)^2 K''(\psi_i);\ i = 0, 1.$

(iii) If $Z_j = \log(f(X_j;\psi_1)/f(X_j;\psi_0));\ j = 1, 2, \ldots$ where $X_1, X_2, \ldots, X_n$ are i.i.d. and $\bar{Z}_n = \dfrac{1}{n}\sum_{j=1}^{n} Z_j$, then the MP test of size α for $H_0^* : \psi = \psi_0$ against $H_1^* : \psi = \psi_1$ is asymptotically of the form $\phi(\bar{Z}_n) = I\{\bar{Z}_n \geq l_\alpha\}$, where $l_\alpha = K'(\psi_0)(\psi_1 - \psi_0) - (K(\psi_1) - K(\psi_0)) + \dfrac{z_{1-\alpha}}{\sqrt{n}}(\psi_1 - \psi_0)(K''(\psi_0))^{1/2}$ and $z_{1-\alpha} = \Phi^{-1}(1-\alpha)$.

(iv) The power of the asymptotic test at ψ_1 is approximately

$$\Phi\left(\sqrt{n}\frac{1}{(K''(\psi_1))^{1/2}}(K'(\psi_1) - K'(\psi_0)) - z_{1-\alpha}\left(\frac{K''(\psi_0)}{K''(\psi_1)}\right)^{1/2}\right).$$

(v) Show that the power function given in (iv) is monotonically increasing in ψ_1.

4.2.4 Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having a common negative-binomial distribution $NB(p, \nu)$, ν fixed. Apply the results of the previous problem to derive a large sample test of size α of $H_0 : p \leq p_0$ against $H_1 : p \geq p_1, 0 < p_0 < p_1 < 1$.

4.2.5 Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having a common distribution with p.d.f. $f(x;\mu,\theta) = (1-\theta)\phi(x) + \theta\phi(x-\mu),\ -\infty < x < \infty$, where μ is known, $\mu > 0$; $0 \leq \theta \leq 1$; and $\phi(x)$ is the standard normal p.d.f.

(i) Construct the MP test of size α of $H_0 : \theta = 0$ against $H_1 : \theta = \theta_1$, $0 < \theta_1 < 1$.

(ii) What is the critical level and the power of the test?

4.2.6 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common continuous distribution with p.d.f. $f(x;\theta)$. Consider the problem of testing the two simple hypotheses $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, $\theta_0 \neq \theta_1$. The MP test is of the form $\phi(\mathbf{x}) = I\{S_n \geq c\}$, where $S_n = \sum_{i=1}^{n} \log(f(X_i;\theta_1)/f(X_i;\theta_0))$. The two types of error associated with ϕ_c are

$$\epsilon_0(c) = P_0\{S_n \geq c\} \text{ and } \epsilon_1(c) = P_1\{S_n < c\}.$$

A test ϕ_{c^*} is called **minimax** if it minimizes $\max(\epsilon_0(c), \epsilon_1(c))$. Show that ϕ_{c^*} is minimax if there exists a c^* such that $\epsilon_0(c^*) = \epsilon_1(c^*)$.

Section 4.3

4.3.1 Consider the one-parameter exponential type family with p.d.f.s

$$f(x;\theta) = h(x)\exp\{Q(\theta)U(x) + C(\theta)\}, \quad \theta \in \Theta,$$

where $Q'(\theta) > 0$ for all $\theta \in \Theta$; $Q(\theta)$ and $C(\theta)$ have second order derivatives at all $\theta \in \Theta$.

(i) Show that the family $\mathcal{F}$ is MLR in $U(X)$.

(ii) Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables having such a distribution. What is the distribution of the m.s.s. $T(\mathbf{X}) = \sum_{j=1}^{n} U(X_j)$?

(iii) Construct the UMP test of size α of $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$.

(iv) Show that the power function is differentiable and monotone increasing in θ.

4.3.2 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a scale and location parameter exponential distribution with p.d.f.

$$f(x;\mu,\sigma) = \frac{1}{\sigma}\exp\left\{-\frac{1}{\sigma}(x-\mu)\right\} I\{x \geq \mu\};$$

$$0 < \sigma < \infty; \quad -\infty < \mu < \infty.$$

(i) Develop the α-level UMP test of $H_0 : \mu \leq \mu_0$, against $\mu > \mu_0$ when σ is known.

(ii) Consider the hypotheses $H_0 : \mu = \mu_0$, $\sigma = \sigma_0$ against $H_1 : \mu < \mu_0$, $\sigma < \sigma_0$. Show that there exists a UMP test of size α and provide its power function.

4.3.3 Consider n identical systems that operate independently. It is assumed that the time till failure of a system has a $G\left(\frac{1}{\theta}, 1\right)$ distribution. Let $Y_1, Y_2, \ldots, Y_r$ be the failure times until the rth failure.

(i) Show that the total life $T_{n,r} = \sum_{i=1}^{r} Y_i + (n-r)Y_r$ is distributed like $\frac{\theta}{2}\chi^2[2r]$.

(ii) Construct the α-level UMP test of $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$ based on $T_{n,r}$.

(iii) What is the power function of the UMP test?

4.3.4 Consider the linear regression model prescribed in Problem 3, Section 2.9. Assume that α and σ are known.

(i) What is the least-squares estimator of β?

(ii) Show that there exists a UMP test of size α for $H_0 : \beta \leq \beta_0$ against $\beta > \beta_0$.

(iii) Write the power function of the UMP test.

Section 4.4

4.4.1 Let $X_1, \ldots, X_n$ be i.i.d. random variables having an $N(0, \sigma^2)$ distribution. Determine the UMP unbiased test of size α of $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$, where $0 < \sigma_0^2 < \infty$.

4.4.2 Let $X \sim B(20, \theta)$, $0 < \theta < 1$. Construct the UMP unbiased test of size $\alpha = .05$ of $H_0 : \theta = .15$ against $H_1 : \theta \neq .15$. What is the power of the test when $\theta = .05, .15, .20, .25$

4.4.3 Let $X_1, \ldots, X_n$ be i.i.d. having a common exponential distribution $G(\frac{1}{\theta}, 1)$, $0 < \theta < \infty$. Consider the reliability function $\rho = \exp\{-t/\theta\}$, where t is known. Construct the UMP unbiased test of size α for $H_0 : \rho = \rho_0$ against $H_1 : \rho \neq \rho_0$, for some $0 < \rho_0 < 1$.

Section 4.5

4.5.1 Let $X_1, \ldots, X_n$ be i.i.d. random variables where $X_1 \sim \xi + G(\frac{1}{\sigma}, 1)$, $-\infty < \xi < \infty$, $0 < \sigma < \infty$. Construct the UMPU tests of size α and their power function for the hypotheses:

(i) $H_0 : \xi \leq \xi_0$, σ arbitrary; $H_1 : \xi > \xi_0$, σ arbitrary.

(ii) $H_0 : \sigma = \sigma_0$, ξ arbitrary; $H_1 : \sigma \neq \sigma_0$, ξ arbitrary.

4.5.2 Let $X_1, \ldots, X_m$ be i.i.d. random variables distributed like $N(\mu_1, \sigma^2)$ and let $Y_1, \ldots, Y_n$ be i.i.d. random variables distributed like $N(\mu_2, \sigma^2)$; $-\infty < \mu_1$,

$\mu_2 < \infty$; $0 < \sigma^2 < \infty$. Furthermore, the X-sample is independent of the Y-sample. Construct the UMPU test of size α for

(i) $H_0 : \mu_1 = \mu_2, \sigma$ arbitrary; $H_1 : \mu_1 \neq \mu_2, \sigma$ arbitrary.

(ii) What is the power function of the test?

4.5.3 Let $X_1, \ldots, X_n$ be i.i.d. random variables having $N(\mu, \sigma^2)$ distribution. Construct a test of size α for $H_0 : \mu + 2\sigma \geq 0$ against $\mu + 2\sigma < 0$. What is the power function of the test?

4.5.4 In continuation of Problem 3, construct a test of size α for

$$H_0 : \mu_1 + 5\mu_2 \leq 10, \quad \sigma \text{ arbitrary};$$
$$H_1 : \mu_1 + 5\mu_2 > 10, \quad \sigma \text{ arbitrary}.$$

4.5.5 Let (X_1, X_2) have a trinomial distribution with parameters (n, θ_1, θ_2), where $0 < \theta_1, \theta_2 < 1$, and $\theta_1 + \theta_2 \leq 1$. Construct the UMPU test of size α of the hypotheses $H_0 : \theta_1 = \theta_2$; $H_1 : \theta_1 \neq \theta_2$.

4.5.6 Let X_1, X_2, X_3 be independent Poisson random variables with means $\lambda_1, \lambda_2, \lambda_3$, respectively, $0 < \lambda_i < \infty$ $(i = 1, 2, 3)$. Construct the UMPU test of size α of $H_0 : \lambda_1 = \lambda_2 = \lambda_3$; $H_1 : \lambda_1 > \lambda_2 > \lambda_3$.

Section 4.6

4.6.1 Consider the normal regression model of Problem 3, Section 2.9. Develop the likelihood ratio test, of size ϵ, of

(i) $H_0 : \alpha = 0, \beta, \sigma$ arbitrary; $H_1 : \alpha \neq 0$; β, σ arbitrary.

(ii) $H_0 : \beta = 0, \alpha, \sigma$ arbitrary; $H_1 : \beta \neq 0$; α, σ arbitrary.

(iii) $\sigma \geq \sigma_0, \alpha, \beta$ arbitrary; $H_1 : \sigma < \sigma_0$; α, β arbitrary.

4.6.2 Let $(\bar{X}_1, S_1^2), \ldots, (\bar{X}_k, S_k^2)$ be the sample mean and variance of k independent random samples of size $n_1, \ldots, n_k$, respectively, from normal distributions $N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, k$. Develop the likelihood ratio test for testing $H_0 : \sigma_1 = \cdots = \sigma_k$; $\mu_1, \ldots, \mu_k$ arbitrary against the general alternative $H_1 : \sigma_1, \ldots,$

σ_k and $\mu_1, \ldots, \mu_k$ arbitrary. [The test that rejects H_0 when $\sum_{i=1}^{k} n_i \log \frac{S_p^2}{S_i^2} \geq \chi^2_{1-\alpha}[k-1]$, where $S_p^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)S_i^2$ and $N = \sum_{i=1}^{k} n_i$, is known as **the Bartlett test** for the equality of variances (Hald, 1952, p. 290).]

4.6.3 Let $(X_1, \ldots, X_k)$ have a multinomial distribution $MN(n, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, $0 < \theta_i < 1$, $\sum_{i=1}^{k} \theta_i = 1$. Develop the likelihood ratio test of $H_0 : \theta_1 = \cdots = \theta_k = \frac{1}{k}$ against $H_1 : \boldsymbol{\theta}$ arbitrary. Provide a large sample approximation for the critical value.

4.6.4 Let (X_i, Y_i), $i = 1, \ldots, n$, be i.i.d. random vectors having a bivariate normal distribution with zero means and covariance matrix $\mathfrak{V} = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, $0 < \sigma^2 < \infty$, $-1 < \rho < 1$. Develop the likelihood ratio test of $H_0 : \rho = 0$, σ arbitrary against $H_1 : \rho \neq 0, \sigma$ arbitrary.

4.6.5 Let $(x_{11}, Y_{11}), \ldots, (X_{1n}, Y_{1n})$ and $(x_{21}, Y_{21}), \ldots, (x_{2n}, Y_{2n})$ be two sets of independent normal regression points, i.e., $Y_{ij} \sim N(\alpha_1 + \beta_1 x_j, \sigma^2)$, $j = 1, \ldots, n$ and $Y_{2j} \sim N(\alpha_2 + \beta_2 x_j; \sigma^2)$, where $\mathbf{x}^{(1)} = (x_{11}, \ldots, x_{1n})'$ and $\mathbf{x}^{(2)} = (x_{21}, \ldots, x_{2n})'$ are known constants.

(i) Construct the likelihood ratio test of $H_0 : \alpha_1 = \alpha_2$, β_1, β_2, σ arbitrary; against $H_1 : \alpha_1 \neq \alpha_2$; β_1, β_2, σ arbitrary.

(ii) $H_0 : \beta_1 = \beta_2, \alpha_1, \alpha_2$ arbitrary; against $H_1 : \beta_1 \neq \beta_2$; $\alpha_1, \alpha_2, \sigma$ arbitrary.

4.6.6 The one-way analysis of variance (ANOVA) developed in Section 4.6 corresponds to model (4.6.31), which is labelled Model I. In this model, the incremental effects are fixed. Consider now the random effects model of Example 3.9, which is labelled Model II. The analysis of variance tests $H_0 : \tau^2 = 0$, σ^2 arbitrary; against $H_1 : \tau^2 > 0, \sigma^2$ arbitrary; where τ^2 is the variance of the random effects $a_1, \ldots, a_r$. Assume that all the samples are of equal size, i.e., $n_1 = \cdots = n_r = n$.

(i) Show that $S_p^2 = \frac{1}{r} \sum_{i=1}^{r} S_i^2$ and $S_b^2 = \frac{n}{r-1} \sum_{i=1}^{r} (\bar{X}_i - \bar{\bar{X}})^2$ are independent.

(ii) Show that $S_b^2 \sim (\sigma^2 + n\tau^2)\chi^2[r-1]/(r-1)$.

(iii) Show that the F-ratio (4.6.27) is distributed like $(1 + n\frac{\tau^2}{\sigma^2})\, F[r-1, r(n-1)]$.

(iv) What is the ANOVA test of H_0 against H_1?

(v) What is the power function of the ANOVA test? [Express this function in terms of the incomplete beta function and compare the result with (4.6.29)–(4.6.30).]

4.6.7 Consider the two-way layout model of ANOVA (4.6.33) in which the incremental effects of A, $\tau_1^A, \ldots, \tau_{r_1}^A$, are considered fixed, but those of B, $\tau_1^B, \ldots, \tau_{r_2}^B$, are considered i.i.d. random variables having a $N(0, \sigma_B^2)$ distribution. The interaction components τ_{ij}^{AB} are also considered i.i.d. (independent

of τ_j^B) having a $N(0, \sigma_{AB}^2)$ distribution. The model is then called a **mixed effect** model. Develop the ANOVA tests of the null hypotheses

$$H_0^{(1)} : \sigma_{AB}^2 = 0,$$

$$H_0^{(2)} : \sum_{i=1}^{r_1} (\tau_i^A)^2 = 0,$$

$$H_0^{(3)} : \sigma_B^2 = 0.$$

What are the power functions of the various F-tests (see Scheffé, 1959, Chapter 8)?

Section 4.7

4.7.1 Apply the X^2-test to test the significance of the association between the attributes A, B in the following contingency table

	A_1	A_2	A_3	Sum
B_1	150	270	500	920
B_2	550	1750	300	2600
Sum	700	2020	800	3520

At what level of significance, α, would you reject the hypothesis of no association?

4.7.2 The X^2-test statistic (4.7.5) can be applied in large sample to test the equality of the success probabilities of k Bernoulli trials. More specifically, let $f_1, \ldots, f_k$ be independent random variables having binomial distributions $B(n_i, \theta_i)$, $i = 1, \ldots, k$. The hypothesis to test is $H_0 : \theta_1 = \cdots = \theta_k = \theta$, θ arbitrary against H_1 : the θs are not all equal. Notice that if H_0 is correct, then $T = \sum_{i=1}^{k} f_i \sim B(N, \theta)$ where $N = \sum_{i=1}^{k} n_i$. Construct the $2 \times k$ contingency table

	E_1	$\cdots$	E_k	Total
S	f_1	$\cdots$	f_k	T
F	$n_1 - f_1$	$\cdots$	$n_k - f_k$	$N - T$
Total	n_1		n_k	N

This is an example of a contingency table in which one margin is fixed $(n_1, \ldots, n_k)$ and the cell frequencies do not follow a multinomial distribution. The hypothesis H_0 is equivalent to the hypothesis that there is no association between the trial number and the result (success or failure).

(i) Show that the X^2 statistic is equal in the present case to

$$X^2 = \sum_{i=1}^{k} n_i \frac{\left(\frac{f_i}{n_i} - \frac{T}{N}\right)^2}{\frac{T}{N}\left(1 - \frac{T}{N}\right)}.$$

(ii) Show that if $n_i \to \infty$ for all $i = 1, \ldots, k$ so that $\frac{n_i}{N} \to \lambda_i$, $0 < \lambda_i < 1$ for all $i = 1, \ldots, k$, then, under H_0, X^2 is asymptotically distributed like $\chi^2[k-1]$.

4.7.3 The test statistic X^2, as given by (4.7.5) can be applied to test also whether a certain distribution $F_0(x)$ fits the frequency distribution of a certain random variable. More specifically, let Y be a random variable having a distribution over (a, b), where a could assume the value $-\infty$ and/or b could assume the value $+\infty$. Let $\eta_0 < \eta_1 < \cdots < \eta_k$ with $\eta_0 = a$ and $\eta_k = b$, be a given partition of (a, b). Let f_i $(i = 1, \ldots, k)$ be the observed frequency of Y over (η_{i-1}, η_i) among N i.i.d. observations on $Y_1, \ldots, Y_n$, i.e., $f_i = \sum_{j=1}^{n} I\{\eta_{i-1} < Y_j \leq \eta_i\}$, $i = 1, \ldots, k$. We wish to test the hypothesis $H_0 : F_Y(y) \equiv F_0(y)$, where $F_Y(y)$ denotes the c.d.f. of Y. We notice that if H_0 is true, then the expected frequency of Y at $[\eta_{i-1}, \eta_i]$ is $e_i = N\{F_0(\eta_i) - F_0(\eta_{i-1})\}$. Accordingly, the test statistic X^2 assumes the form

$$X^2 = \sum_{i=1}^{k} f_i^2 / N[F_0(\eta_i) - F_0(\eta_{i-1})] - N.$$

The hypothesis H_0 is rejected, in large samples, at level of significance α if $X^2 \geq \chi^2_{1-\alpha}[k-1]$. This is a large sample **test of goodness of fit**, proposed in 1900 by Karl Pearson (see Lancaster, 1969, Chapter VIII; Bickel and Doksum, 1977, Chapter 8, for derivations and proofs concerning the asymptotic distribution of X^2 under H_0).

The following 50 numbers are so-called "random numbers" generated by a desk calculator: 0.9315, 0.2695, 0.3878, 0.9745, 0.9924, 0.7457, 0.8475, 0.6628, 0.8187, 0.8893, 0.8349, 0.7307, 0.0561, 0.2743, 0.0894, 0.8752, 0.6811, 0.2633, 0.2017, 0.9175, 0.9216, 0.6255, 0.4706, 0.6466, 0.1435, 0.3346, 0.8364, 0.3615, 0.1722, 0.2976, 0.7496, 0.2839, 0.4761, 0.9145, 0.2593, 0.6382, 0.2503, 0.3774, 0.2375, 0.8477, 0.8377, 0.5630, 0.2949, 0.6426, 0.9733, 0.4877, 0.4357, 0.6582, 0.6353, 0.2173. Partition the interval (0, 1) to $k = 7$ equal length subintervals and apply the X^2 test statistic to test whether the rectangular distribution $R(0, 1)$ fits the frequency distribution of the above sample. [If any of the seven frequencies is smaller than six, combine two adjacent subintervals until all frequencies are not smaller than six.]

4.7.4 In continuation of the previous problem, if the hypothesis H_0 specifies a distribution $F(x;\boldsymbol{\theta})$ that depends on a parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$, $i \leq r$, but the value of the parameter is unknown, the large sample test of goodness of fit compares

$$X^2 = \sum_{i=1}^{k} f_i^2 / N[F(\eta_i; \hat{\boldsymbol{\theta}}) - F(\eta_{i-1}; \hat{\boldsymbol{\theta}})] - N$$

with $\chi^2_{1-\alpha}[k-1-r]$ (Lancaster, 1969, p. 148), where $\hat{\boldsymbol{\theta}}$ are estimates of $\boldsymbol{\theta}$ obtained by maximizing

$$Q = \sum_{i=1}^{k} f_i \log[F(\eta_i; \boldsymbol{\theta}) - F(\eta_{i-1}; \boldsymbol{\theta})].$$

(i) Suppose that $\eta_0 = 0 < \eta_1 < \cdots < \eta_k = \infty$ and $F(x;\sigma) = 1 - \exp\{-x/\sigma\}$, $0 < \sigma < \infty$. Given $\eta_1, \ldots, \eta_{k-1}$ and $f_1, \ldots, f_k$, N, how would you estimate σ?

(ii) What is the likelihood ratio statistic for testing H_0 against the alternative that the distribution F is arbitrary?

(iii) Under what conditions would the likelihood ratio statistic be asymptotically equivalent, as $N \to \infty$, to X^2 (see Bickel and Doksum, 1977, p. 394)?

4.7.5 Consider Problem 3 of Section 2.9. Let (X_{1i}, X_{2i}), $i = 1, \ldots, n$ be a sample of n i.i.d. such vectors. Construct a test of $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$, at level of significance α.

Section 4.8

4.8.1 Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables having a common binomial distribution $B(1, \theta)$, $0 < \theta < 1$.

(i) Construct the Wald SPRT for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, $0 < \theta_0 < \theta_1 < 1$, aiming at error probabilities α and β, by applying the approximation $A' = \log \beta(1-\alpha)$ and $B' = \log(1-\beta)/\alpha$.

(ii) Compute and graph the OC curve for the case of $\theta_0 = .01$, $\theta_1 = .10$, $\alpha = .05$, $\beta = .05$, using approximations (4.8.26)–(4.8.29).

(iii) What is $E_\theta\{N\}$ for $\theta = .08$?

4.8.2 Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables having a $N(0, \sigma^2)$ distribution. Construct the Wald SPRT to test $H_0 : \sigma^2 = 1$ against $H_1 : \sigma^2 = 2$ with error probabilities $\alpha = .01$ and $\beta = .07$. What is $\pi(\sigma^2)$ and $E_{\sigma^2}\{N\}$ when $\sigma^2 = 1.5$?

PART IV: SOLUTIONS TO SELECTED PROBLEMS

4.1.2 The sample size is $n = 100$.

(i) The large sample test is based on the normal approximation

$$\begin{aligned}\phi(X) &= I\{X \le 100 \times 0.75 - Z_{1-\alpha}\sqrt{100 \times 0.75 \times 0.25}\} \\ &= I\{X \le 75 - Z_{1-\alpha}4.33\}.\end{aligned}$$

(ii) The power of the test when $\theta = 0.50$ is

$$\begin{aligned}\psi(0.5) &= P_{0.5}(X \le 75 - Z_{1-\alpha}4.33) \\ &\doteq \Phi\left(\frac{75 - 4.33Z_{1-\alpha} - 50}{\sqrt{100 \times .5 \times .5}}\right) \\ &= \Phi\left(5 - \frac{4.33}{5}Z_{1-\alpha}\right).\end{aligned}$$

(iii) We have to satisfy two equations:

$$\text{(i) } \Phi\left(\frac{C - n \times 0.75}{\sqrt{n \times 0.75 \times 0.25}}\right) \le 0.05$$

$$\text{(ii) } \Phi\left(\frac{C - n \times 0.6}{\sqrt{n \times 0.6 \times 0.4}}\right) \ge 0.9.$$

The solution of these 2 equations is $n \doteq 82$, $C \doteq 55$. For these values n and C we have $\alpha = 0.066$ and $\psi(0.6) = 0.924$.

4.2.1 Assume that $\nu = 2$. The m.s.s. is $T = \sum_{i=1}^{n} X_i \sim NB(\psi, 2n)$.

(i) This is an MLR family. Hence, the MP test is

$$\phi^0(T) = \begin{cases} 1, & \text{if } T > NB^{-1}(1-\alpha; \psi_0, 2n) \\ \gamma_\alpha, & T = NB^{-1}(1-\alpha; \psi_0, 2n) \\ 0, & T < NB^{-1}(1-\alpha; \psi_0, 2n). \end{cases}$$

Let $T_{1-\alpha}(\psi_0) = NB^{-1}(1-\alpha; 2n, \psi_0)$. Then

$$\gamma_\alpha = \frac{1 - \alpha - P\{T \le T_{1-\alpha}(\psi_0) - 1\}}{P\{T = T_{1-\alpha}(\psi_0)\}}.$$

(ii) The power function is

$$\text{Power}(\psi) = P_\psi\{T > T_{1-\alpha}(\psi_0)\} + \gamma_\alpha P_\psi\{T = T_{1-\alpha}(\psi_0)\}.$$

We can compute these functions with the formula

$$P_\psi\{T \le t\} = I_{1-\psi}(2n, t+1).$$

(iii) We can start with the large sample normal approximation

$$P_\psi(T \le t) \cong \Phi\left(\frac{t - 2n\psi/(1-\psi)}{\sqrt{2n\psi/(1-\psi)^2}}\right) = \Phi\left(\frac{(1-\psi)t - 2n\psi}{\sqrt{2n\psi}}\right).$$

For $\psi_0 = 0.05$, $\alpha = 0.10$, $\psi_1 = 0.15$, $1-\beta = 0.8$, we get the equations

(i) $0.95t - 2n \times 0.05 = 1.2816\sqrt{2n \times 0.05}$.

(ii) $0.85t - 2n \times 0.15 = -0.8416\sqrt{2n \times 0.15}$.

The solution of these equations is $n = 15.303$ and $t = 3.39$. Since n and t are integers we take $n = 16$ and $t = 3$. Indeed, $P_{0.05}(T \le 3) = I_{.95}(32, 4) = 0.904$. Thus,

$$\phi^0(T) = \begin{cases} 1, & \text{if } T > 3 \\ 0.972, & \text{if } T = 3 \\ 0, & \text{if } T < 3. \end{cases}$$

The power at $\psi = 0.15$ is

$$\begin{aligned} \text{Power}(0.15) &= P_{0.15}\{T > 3\} + 0.972 P_{0.15}\{T = 3\} \\ &= 0.8994. \end{aligned}$$

4.2.3

$$f(X;\psi) = h(x)\exp(\psi X - K(\psi)).$$

For $\psi_1 > \psi_0$,

$$\frac{f(X;\psi_1)}{f(X;\psi_0)} = \exp\{(\psi_1 - \psi_0)X - (K(\psi_1) - K(\psi_0))\}.$$

(i)
$$\begin{aligned} E_i\left\{\log\frac{f(X;\psi_1)}{f(X;\psi_0)}\right\} &= E_i\{(\psi_1 - \psi_0)X\} - (K(\psi_1) - K(\psi_0)\} \\ &= (\psi_1 - \psi_0)K'(\psi_i) - (K(\psi_1) - K(\psi_0)). \end{aligned}$$

Since $E_i\{X\} = K'(\psi_i)$, $i = 0, 1$.

(ii)
$$\begin{aligned} V_i\left\{\log\frac{f_1(X)}{f_0(X)}\right\} &= (\psi_1 - \psi_0)^2 V_i\{X\} \\ &= (\psi_1 - \psi_0)^2 K''(\psi_i), \quad i = 0, 1. \end{aligned}$$

(iii) The MP test of H_0^* versus H_1^* is asymptotically (large n)

$$\phi(\bar{Z}_n) = I\{\bar{Z}_n \geq (\psi_1 - \psi_0)K'(\psi_0) - (K(\psi_1) - K(\psi_0)) + Z_{1-\alpha}\frac{1}{\sqrt{n}}(\psi_1 - \psi_0)(K''(\psi_0))^{1/2}\}.$$

(iv)

$$\begin{aligned}\text{Power}(\psi_1) &= P_{\psi_1}\bar{Z}_n \geq l_\alpha\} \\ &= 1 - \Phi\left(\frac{(\psi_1 - \psi_0)(K'(\psi_0) - K'(\psi_1)) + Z_{1-\alpha}\frac{1}{\sqrt{n}}(\psi_1 - \psi_0)\sqrt{K''(\psi_0)}}{\frac{1}{\sqrt{n}}(\psi_1 - \psi_0)\sqrt{K''(\psi_1)}}\right) \\ &= \Phi\left(\sqrt{n}\frac{K'(\psi_1) - K'(\psi_0)}{\sqrt{K''(\psi_1)}} - Z_{1-\alpha}\sqrt{\frac{K''(\psi_0)}{K''(\psi_1)}}\right).\end{aligned}$$

4.3.1 Since $Q'(\theta) > 0$, if $\theta_1 < \theta_2$ then $Q(\theta_2) > Q(\theta_1)$

$$\frac{f(X;\theta_2)}{f(X;\theta_1)} = \exp\{(Q(\theta_2) - Q(\theta_1))U(X) + C(\theta_2) - C(\theta_1)\}.$$

(i) $\mathcal{F} = \{f(X;\theta), \theta \in \Theta\}$ is MLR in $U(X)$ since $\exp\{(Q(\theta_2) - Q(\theta_1))U(X)\} \nearrow U(X)$.

$$T(\mathbf{X}) = \sum_{j=1}^{n} U(X_j)$$

(ii)

$$f(\mathbf{X};\theta) = \prod_{j=1}^{n} h(X_j) \cdot \exp\left\{Q(\theta)\sum_{j=1}^{n} U(X_j) + nC(\theta)\right\}.$$

Thus, the p.d.f. of $T(\mathbf{X})$ is

$$\begin{aligned}f(t;\theta) &= \exp(Q(\theta)t + nC(\theta))A(t;\theta) \\ &= \frac{\exp(Q(\theta)t)}{\int \exp(Q(\theta)t)d\lambda(t)} \\ &= \exp(Q(\theta)t - K(\theta)),\end{aligned}$$

where

$$K(\theta) = \log\left(\int \exp\{Q(\theta)t\}d\lambda(t)\right).$$

This is a one-parameter exponential type density.

(iii) The UMP test of $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ is

$$\phi^0(T) = \begin{cases} 1, & \text{if } T > C_\alpha(\theta_0) \\ \gamma_\alpha, & \text{if } T = C_\alpha(\theta_0) \\ 0, & \text{if } T < C_\alpha(\theta_0), \end{cases}$$

where $C_\alpha(\theta_0)$ is the $(1-\alpha)$ quantile of T under θ_0. If T is discrete then

$$\gamma_\alpha = \frac{1 - \alpha - P_{\theta_0}\{T \leq C_\alpha(\theta_0) - 0\}}{P_{\theta_0}\{T = C_\alpha(\theta_0)\}}.$$

(iv)

$$\begin{aligned} \text{Power}(\theta) &= E_\theta\{\phi^0(T)\} \\ &= \int_{C_\alpha(\theta_0)+}^{\infty} \exp\{Q(\theta)t - K(\theta)\} d\lambda(t) + \gamma_0 P_\theta\{T = C_\alpha(\theta_0)\}. \end{aligned}$$

By Karlin's Lemma, this function is increasing in θ. It is differentiable w.r.t. θ.

4.3.3 **(i)** $T_1, T_2, \ldots, T_n$ are i.i.d. $\theta G(1,1) \sim \frac{\theta}{2}\chi^2[2]$. Define the variables $Y_1, \ldots, Y_r$, where for $T_{(1)} < \cdots < T_{(r)}$

$$\left.\begin{aligned} &Y_i \sim T_{(i)}, \quad i = 1, \ldots, n \\ &Y_1 \sim \frac{\theta}{n} G(1,1) \\ &(Y_2 - Y_1) \sim \frac{\theta}{n-1} G(1,1) \\ &\vdots \\ &(Y_r - Y_{r-1}) \sim \frac{\theta}{n-r+1} G(1,1) \end{aligned}\right\} \text{ independent increments.}$$

Accordingly,

$$nY_1 + (n-1)(Y_2 - Y_1) + \cdots + (n-r+1)(Y_r - Y_{r-1}) \sim \frac{\theta}{2}\chi^2[2r].$$

The left-hand side is equal to $\sum_{i=1}^{r} Y_i + (n-r)Y_r$. Thus $T_{n,r} \sim \frac{\theta}{2}\chi^2[2r]$.

(ii) Since the distribution of $T_{n,r}$ is MLR in $T_{n,r}$, the UMP test of size α is

$$\phi^0(T_{n,r}) = I\left\{T_{n,r} \geq \frac{\theta_0}{2}\chi^2_{1-\alpha}[2r]\right\}.$$

(iii) The power function is

$$\psi(\theta) = P\left\{\frac{\theta}{2}\chi^2[2r] \geq \frac{\theta_0}{2}\chi^2_{1-\alpha}[2r]\right\}$$
$$= P\left\{\chi^2[2r] \geq \frac{\theta_0}{\theta}\chi^2_{1-\alpha}[2r]\right\}.$$

4.4.2 $X \sim B(20, \theta)$. $H_0 : \theta = .15$, $H_1 : \theta \neq .15$, $\alpha = .05$. The UMPU test is

$$\phi^0(X) = \begin{cases} 1, & X > C_2(\theta_0) \\ \gamma_2, & X = C_2(\theta_0) \\ 0, & C_1(\theta) < X < C_2(\theta_0) \\ \gamma_1, & X = C_1(\theta_0) \\ 1, & X < C_1(\theta_0). \end{cases}$$

The test should satisfy the conditions

(i) $E_{.15}\{\phi^0(X)\} = \alpha = 0.05$.

(ii) $E_{.15}\{X\phi^0(X)\} = \alpha E_{.15}\{X\} = .15$.

$$X\binom{20}{X}\theta_0^X(1-\theta_0)^X = \frac{20!}{(X-1)!(20-X)!}\theta_0^X(1-\theta_0)^{20-X}$$
$$= 20\theta_0\binom{19}{X-1}\theta_0^{X-1}(1-\theta_0)^{20-X}$$
$$= 3b(X-1; 19, \theta_0).$$

Thus, we find C_1, γ_1, C_2, γ_2 from the equations

(i) $$\sum_{j<X_1} b(j; 20, .15) + C_1 b(C_1; 20, .15)$$
$$+ \gamma_2 b(C_2; 20, .15) + \sum_{j>X_2} b(j; 20, .15) = .05.$$

(ii) $$\sum_{j<X_1} b(j; 19, .15) + \gamma_1 b(C_1; 19, .15)$$
$$+ \gamma_2 b(C_2; 19, .15) + \sum_{j>X_2} b(j; 19, .15) = 0.15.$$

We start with $C_1 = B^{-1}(.025; 20, .15) = 0$, $C_2 = B^{-1}(.975; 20, .15) = 6$

$$B(0; 20, .15) = 0.3875, \quad \gamma_1 = \frac{.025}{0.03875} = .6452$$
$$B(6; 20, .15) = .9781, \quad \gamma_1 = \frac{.975 - .9327}{.9781 - .9327}$$
$$B(5; 20, .15) = .9327, \quad \gamma_1 = .932.$$

These satisfy (i). We check now (ii).

$$\gamma_1 B(0; 19, .15) + \gamma_2 b(6; 19, .15) + (1 - B(6; 19, .15)) = 0.080576.$$

We keep C_1, γ_1, and C_2 and change γ_2 to satisfy (ii); we get

$$\gamma_2' = 0.1137138.$$

We go back to 1, with γ_2' and recompute γ_1'. We obtain $\gamma_1' = .59096$. With these values of γ_1' and γ_2', Equation (i) yields 0.049996 and Equation (ii) yields 0.0475. This is close enough. The UMPU test is

$$\phi^0(X) = \begin{cases} 1, & \text{if } X > 6 \\ 0.1137, & \text{if } X = 6 \\ 0, & \text{if } 0 < X < 6 \\ 0.5909, & \text{if } X = 0. \end{cases}$$

4.5.1 $X_1, \ldots, X_n$ are i.i.d. $\sim \mu + \sigma G(1, 1)$, $-\infty < \mu < \infty$, $0 < \sigma < \infty$. The m.s.s. is $(nX_{(1)}, U)$, where

$$U = \sum_{i=2}^{n} (X_{(i)} - X_{(1)}) \sim \sigma G(1, n - 1)$$
$$nX_{(1)} \sim n\mu + \sigma G(1, 1)$$
$$X_{(1)} \text{ is independent of } U.$$

(i) Find the UMPU test of size α of

$$H_0 : \mu \leq \mu_0, \quad \sigma \text{ arbitrary}$$

against

$$H_1 : \mu > \mu_0, \quad \sigma \text{ arbitrary.}$$

The m.s.s. for $\mathcal{F}^*$ is $T = \sum_{i=1}^{n} X_i = U + nX_{(1)}$. The conditional distribution of $nX_{(1)}$ given T is found in the following way.

(ii) $T \sim \mu + \sigma G(1, n)$. That is,

$$f_T(t) = \frac{1}{\sigma\Gamma(n)} \left(\frac{t-\mu}{\sigma}\right)^{n-1} \exp\left(-\frac{t-\mu}{\sigma}\right), \quad t \geq \mu.$$

The joint p.d.f. of $nX_{(1)}$ and U is

$$\begin{aligned} f_{nX_{(1)},U}(y,u) &= \frac{1}{\sigma} \exp\left(-\frac{y-\mu}{\sigma}\right) \cdot \frac{1}{\sigma^{n-1}\Gamma(n-1)} \cdot u^{n-2} e^{-u/\sigma} I(y \geq \mu) \\ &= \frac{1}{\sigma^n \Gamma(n-1)} \exp\left(-\frac{y-\mu}{\sigma} - \frac{u}{\sigma}\right) \cdot I(y \geq \mu). \end{aligned}$$

We make the transformation

$$\begin{array}{ll} y = y & y = y \\ t = y + u & u = t - y \end{array} \qquad J = \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = 1.$$

The joint density of $(nX_{(1)}, T)$ is

$$g_{nX_{(1)},T}(y,t) = \frac{1}{\sigma^n \Gamma(n-1)} (t-y)^{n-2} e^{\exp(-(t-\mu)/\sigma)} \cdot I(t \geq \mu).$$

Thus, the conditional density of $nX_{(1)}$, given T, is

$$\begin{aligned} h_{nX_{(1)}|T}(y \mid t) &= \frac{g_{nX_{(1)},T}(y,t)}{f_T(t)} \\ &= \frac{(n-1)(t-y)^{n-2}}{(t-\mu)^{n-1}} I(t \geq \mu) I(y \geq \mu) \\ &= \frac{n-1}{(t-\mu)} \frac{(t-\mu-(y-\mu))^{n-2}}{(t-\mu)^{n-2}} I(t > y > \mu) \\ &= \frac{n-1}{(t-\mu)} \left(1 - \frac{y-\mu}{t-\mu}\right)^{n-2}, \quad \mu < y < t. \end{aligned}$$

The c.d.f. of $nX_{(1)} \mid T$, at $\mu = \mu_0$ is

$$H_0(y \mid t) = \left(1 - \frac{y-\mu_0}{t-\mu_0}\right)^{n-1}, \quad \mu_0 < y < t.$$

The $(1-\alpha)$-quantile of $H_0(y \mid t)$ is the solution y of $\left(1-\frac{y-\mu_0}{t-\mu_0}\right)^{n-1} = 1-\alpha$, which is $C_\alpha(\mu_0, t) = \mu_0 + (t-\mu_0)(1-(1-\alpha)^{1/(n-1)})$. The UMPU test of size α is

$$\phi^0(nX_{(1)} \mid T) = \begin{cases} 1, & \text{if } nX_{(1)} \geq C_\alpha(\mu_0, T) \\ 0, & \text{otherwise.} \end{cases}$$

The power function of this test is

$$\begin{aligned} \psi(\mu) &= P_\mu\{nX_{(1)} \geq C_\alpha(\mu_0, T)\} \\ &= 1 - E_\mu\left\{\left(1 - \frac{C_\alpha(\mu_0, T)-\mu}{T-\mu}\right)^{n-1}\right\}. \end{aligned}$$

$T - \mu \sim \sigma G(1, n)$. Hence, for $\xi = 1-(1-\alpha)^{1/(n-1)}$,

$$\begin{aligned} \psi(\mu) &= 1 - \frac{1}{\Gamma(n)}\int_0^\infty \left(1 - \xi + \frac{\mu-\mu_0}{\sigma}\cdot\frac{1-\xi}{y}\right)^{n-1} \cdot y^{n-1}e^{-y}dy \\ &= 1 - (1-\alpha)e^\delta P(n-1;\delta), \end{aligned}$$

where $\delta = (\mu-\mu_0)/\sigma)$. This is a continuous increasing function of μ.

(iii) Testing

$$\begin{aligned} H_0 &: \sigma = \sigma_0, \quad \mu \text{ arbitrary} \\ H_1 &: \sigma \neq \sigma_0, \quad \mu \text{ arbitrary.} \end{aligned}$$

The m.s.s. for $\mathcal{F}^*$ is $X_{(1)}$. Since U is independent of $X_{(1)}$, the conditional test is based on U only. That is,

$$\phi^0(U) = \begin{cases} 1, & U \geq C_2(\sigma_0) \\ 0, & C_1 < U < C_2 \\ 1, & U \leq C_1(\sigma_0). \end{cases}$$

We can start with $C_1(\sigma_0) = \frac{\sigma_0}{2}\chi^2_{\frac{\alpha}{2}}[2n-2]$ and $C_2(\alpha) = \frac{\sigma_0}{2}\chi^2_{1-\frac{\alpha}{2}}[2n-2]$.

4.5.4 **(i)** $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, $-\infty < \mu < \infty, 0 < \sigma < \infty$. The m.s.s. is $(\bar{X}, Q)$, where $\bar{X} = \frac{1}{n}\sum X_i$ and $Q = \sum_{i=1}^{n}(X_i - \bar{X})^2$. $\bar{X}$ and Q are independent. We have to construct a test of

$$H_0 : \mu + 2\sigma \geq 0$$

against

$$H_1 : \mu + 2\sigma < 0.$$

Obviously, if $\mu \geq 0$ then H_0 is true. A test of $H_0^* : \mu \geq 0$ versus $H_1^* : \mu < 0$ is

$$\phi(\bar{X}, Q) = \begin{cases} 1, & \text{if } \frac{\sqrt{n}\bar{X}}{S} \leq -t_{1-\alpha}[1, n-1] \\ 0, & \text{otherwise.} \end{cases}$$

where $S^2 = Q/(n-1)$ is the sample variance. We should have a more stringent test. Consider the statistic $\frac{n\bar{X}^2}{S^2} \sim F[1, n-1; \lambda]$ where $\lambda = \frac{\mu^2 n}{2\sigma^2}$. Notice that for $\mathcal{F}^*$, $\mu^2 = 4\sigma^2$ or $\lambda_0 = 2n$. Let $\theta = (\mu, \sigma)$. If $\theta \in \Theta_1$, then $\lambda > 2n$ and if $\theta \in \Theta_0$, $\lambda \leq 2n$. Thus, the suggested test is

$$\phi(\bar{X}, S) = \begin{cases} 1, & \text{if } \frac{n\bar{X}^2}{S^2} \geq F_{1-\alpha}[1, n-1; 2n] \\ 0, & \text{otherwise.} \end{cases}$$

(ii) The power function is

$$\psi(\lambda) = P\{F[1, n-1; \lambda] \geq F_{1-\alpha}[1, n-1; 2n]\}.$$

The distribution of a noncentral $F[1, n-1; \lambda]$ is like that of $(1 + 2J)F[1 + 2J, n-1]$, where $J \sim \text{Pois}(\lambda)$ (see Section 2.8). Thus, the power function is

$$\Psi(\lambda) = E_\lambda\{P\{(1 + 2J)F[1 + 2J, n-1] \geq F_{1-\alpha}[1, n-1; 2n] \mid J\}\}.$$

The conditional probability is an increasing function of J. Thus, by Karlin's Lemma, $\Psi(\lambda)$ is an increasing function of λ. Notice that

$$P\{F[1, n-1; \lambda] \leq x\} = e^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^j}{j!}P\{(1 + 2j)F[1 + 2j, n-1] \leq x\}.$$

Thus, $F_{1-\alpha}[1, n-1; \lambda] \nearrow \lambda$.

4.6.2 The GLR statistic is

$$\Lambda(\mathbf{X}) = \frac{\sup_{0<\sigma^2<\infty,\mu_1,\ldots,\mu_k} L(\mu_1,\ldots,\mu_k,\sigma^2)}{\sup_{\sigma^2,\mu} L(\mu_1,\ldots,\mu_k,\sigma_1^2,\ldots,\sigma_k^2)}$$

$$= \frac{\prod_{i=1}^{k}\left(\frac{Q_i}{n_i}\right)^{n_i/2}}{\left(\frac{\sum Q_i}{N}\right)^{N/2}},$$

where $N = \sum_{i=1}^{k} n_i$, $Q_i = (n_i - 1)S_i^2$. $S_p^2 = \sum_{i=1}^{k} Q_i/(N-k)$.

$$-2\log\Lambda(\mathbf{X}) = N\log\left(\frac{(N-k)S_p^2}{N}\right) - \sum_{i=1}^{k} n_i \log\left(\frac{(n_i-1)S_i^2}{n_i}\right)$$

$$= \sum_{i=1}^{k} n_i \log\frac{S_p^2}{S_i^2} + \sum_{i=1}^{k} n_i \log\left(\frac{1-\frac{k}{N}}{1-\frac{1}{n_i}}\right).$$

Notice that if $n_1 = \cdots = n_k = n$, then $\sum_{i=1}^{k} n \log\frac{1-\frac{k}{N}}{1-\frac{1}{n}} = 0$. Thus ,for large samples, as $n \to \infty$, the Bartlett test is: reject H_0 if $\sum_{i=1}^{k} n_i \log\frac{S_p^2}{S_i^2} \geq \chi^2_{1-\alpha}[k-1]$. We have $k-1$ degrees of freedom since σ^2 is unknown.

4.6.4

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\mathbf{0}, \sigma^2\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

Let $Q = \sum_{i=1}^{n}(X_i^2 + Y_i^2)$ and $P = \sum_{i=1}^{n} X_i Y_i$. $H_0 : \rho = 0$, σ^2 arbitrary; $H_1 : \rho \neq 0$, σ^2 arbitrary. The MLE of σ^2 under H_0 is $\hat{\sigma}^2 = \frac{Q}{2n}$. The likelihood of (ρ, σ^2) is

$$L(\rho,\sigma^2) = \frac{1}{(1-\rho^2)^{n/2}(\sigma^2)^n}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{Q}{\sigma^2} - 2\rho\frac{P}{\sigma^2}\right]\right\}.$$

Thus, $\sup_{0<\sigma^2<\infty} L(0,\sigma^2) = \frac{1}{(Q/2n)^n}\exp(-n)$. We find now $\hat{\sigma}^2$ and $\hat{\rho}$ that maximize $L(\rho,\sigma^2)$. Let $l(\rho,\sigma^2) = \log L(\rho,\sigma^2)$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{\sigma^2} + \frac{1}{2\sigma^4(1-\rho^2)}(Q - 2\rho P)$$

$$\frac{\partial l}{\partial \rho} = \frac{n\rho}{1-\rho^2} - \frac{\rho(Q-2\rho P) - P(1-\rho^2)}{(1-\rho^2)^2\sigma^2}.$$

Equating these partial derivatives to zero, we get the two equations

$$\textbf{(I)}\quad \frac{Q-2\rho P}{\sigma^2(1-\rho^2)} = 2n$$

$$\textbf{(II)}\quad n\rho + \frac{P}{\sigma^2} = \frac{\rho}{(1-\rho^2)\sigma^2}(Q-2\rho P).$$

Solution of these equations gives the roots $\hat{\sigma}^2 = \frac{Q}{2n}$ and $\hat{\rho} = \frac{2P}{Q}$. Thus,

$$\sup_{\rho,\sigma} L(\rho,\sigma) = \frac{1}{\left(\frac{Q}{2n}\right)^n \left(1 - \left(\frac{2P}{Q}\right)^2\right)^{n/2}} e^{-n}.$$

It follows that

$$\Lambda(\mathbf{X},\mathbf{Y}) = \frac{1}{\left(1 - \left(\frac{2P}{Q}\right)^2\right)^{n/2}}.$$

$\Lambda(\mathbf{X},\mathbf{Y})$ is small if $\left(\frac{2P}{Q}\right)^2$ is small or P is close to zero.

4.6.6 Model II of ANOVA is

$$X_{ij} = \mu + a_i + e_{ij}, \quad i = 1,\ldots,r \;\; j = 1,\ldots,n$$

where $e_{ij} \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2)$, $a_i \overset{\text{i.i.d.}}{\sim} N(0,\tau^2)$, $\{e_{ij}\}$ independent of $\{a_i\}$.

$$S_i^2 = \sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2/(n-1).$$

Notice that $X_{ij} \mid a_i \sim N(\mu + a_i, \sigma^2)$. Hence, $X_{ij} \sim N(\mu, \sigma^2 + \tau^2)$. Moreover,

$$S_i^2 \mid a_i \sim \sigma^2 \chi^2[n-1]/(n-1), \quad i = 1, \ldots, r$$

and

$$S_p^2 = \frac{1}{r} \sum_{i=1}^{r} S_i^2 \sim \sigma^2 \chi^2[r(n-1)]/r(n-1).$$

(i) $S_b^2 = \dfrac{n}{r-1} \sum_{i=1}^{r} (\bar{X}_i - \bar{\bar{X}})^2$, where $\bar{\bar{X}} = \dfrac{1}{r} \sum_{i=1}^{r} \bar{X}_i$. Given a_i, $\bar{X}_i$ and S_i^2 are conditionally independent. Since the distribution of S_i^2 does not depend on a_i, $\bar{X}_i$ and S_i^2 are independent for all $i = 1, \ldots, r$. Hence S_p^2 is independent of S_b^2.

(ii) $\bar{X}_i \sim N\left(\mu, \dfrac{\sigma^2}{n} + \tau^2\right)$ for all $i = 1, \ldots, r$. Hence,

$$\sum_{i=1}^{r} (\bar{X}_i - \bar{\bar{X}})^2 \sim \left(\frac{\sigma^2}{n} + \tau^2\right) \chi^2[r-1].$$

Thus,

(iii)
$$S_b^2 = \frac{n}{r-1} \sum_{i=1}^{r} (\bar{X}_i - \bar{\bar{X}})^2 \sim (\sigma^2 + n\tau^2) \chi^2[r-1]/(r-1).$$

$$F = \frac{S_b^2}{S_p^2} \sim \frac{\sigma^2 + n\tau^2}{\sigma^2} \cdot \frac{\chi^2[r-1]/(r-1)}{\chi^2[r(n-1)]/r(n-1)}$$

$$\sim \left(1 + n\frac{\tau^2}{\sigma^2}\right) F[r-1, r(n-1)].$$

(iv) The ANOVA test of $H_0 : \tau^2 = 0$, σ arbitrary against $H_1 : \tau^2 > 0$, σ arbitrary is the F test

$$\phi(F) = I\{F \geq F_{1-\alpha}[r-1, r(n-1)]\}.$$

(v) The power function is

$$\psi\left(\frac{\tau^2}{\sigma^2}\right) = P\left\{\left(1 + n\left(\frac{\tau}{\sigma}\right)^2\right) F[r-1, r(n-1)] \geq F_{1-\alpha}[r-1, r(n-1)]\right\}$$

$$= P\left\{F[r-1, r(n-1)] \geq \frac{1}{1 + n\left(\frac{\tau}{\sigma}\right)^2} F_{1-\alpha}[r-1, r(n-1)]\right\}.$$

Let

$$\xi_\alpha\left(\left(\frac{\tau}{\sigma}\right)^2\right) = \frac{1}{1+n\left(\frac{\tau}{\sigma}\right)^2} F_{1-\alpha}[r-1, r(n-1)].$$

Thus,

$$\psi\left(\left(\frac{\tau}{\sigma}\right)^2\right) = 1 - P\left\{F[r-1, r(n-1)] \le \xi_\alpha\left(\left(\frac{\tau}{\sigma}\right)^2\right)\right\}.$$

$$\begin{aligned} F[r-1, r(n-1)] &\sim \frac{r(n-1)}{r-1} \cdot \frac{\chi_2^2[r-1]}{\chi_1^2[r(n-1)]} \\ &\sim \frac{r(n-1)}{r-1} \cdot \frac{G_2\left(1, \frac{r-1}{2}\right)}{G_1\left(1, \frac{r(n-1)}{2}\right)}, \end{aligned}$$

where $G_2\left(1, \dfrac{r-1}{2}\right)$ and $G_1\left(1, \dfrac{r(n-1)}{2}\right)$ are independent.

$$\frac{G_2\left(1, \frac{r-1}{2}\right)}{G_2\left(1, \frac{r-1}{2}\right) + G_1\left(1, \frac{r(n-1)}{2}\right)} \sim \text{Beta}\left(\frac{r-1}{2}, \frac{r(n-1)}{2}\right).$$

Hence,

$$\frac{G_2\left(1, \frac{r-1}{2}\right)/G_1\left(1, \frac{r(n-1)}{2}\right)}{1 + G_2\left(1, \frac{r-1}{2}\right)/G_1\left(1, \frac{r(n-1)}{2}\right)} \sim \text{Beta}\left(\frac{r-1}{2}, \frac{r(n-1)}{2}\right).$$

Let $W = G_2\left(1, \dfrac{r-1}{2}\right)/G_1\left(1, \dfrac{r(n-1)}{2}\right)$. Thus,

$$\begin{aligned} \frac{W}{1+W} &\sim \text{Beta}\left(\frac{r-1}{2}, \frac{r(n-1)}{2}\right) \\ W &\sim \frac{\beta\left(\frac{r-1}{2}, \frac{r(n-1)}{2}\right)}{1 - \beta\left(\frac{r-1}{2}, \frac{r(n-1)}{2}\right)}. \end{aligned}$$

$$\begin{aligned} P\{W \le \xi\} &= P\left\{\beta\left(\frac{r-1}{2}, \frac{r(n-1)}{2}\right) \le \frac{\xi}{1+\xi}\right\} \\ &= I_{\frac{\xi}{1+\xi}}\left(\frac{r-1}{2}, \frac{r(n-1)}{2}\right). \end{aligned}$$

Let $\xi = \dfrac{r-1}{r(n-1)}\xi_\alpha\left(\left(\dfrac{\tau}{\sigma}\right)^2\right)$. Then

$$\psi\left(\left(\frac{\tau}{\sigma}\right)^2\right) = 1 - I_{\frac{\xi}{1+\xi}}\left(\frac{r-1}{2}, \frac{r(n-1)}{2}\right)$$
$$= I_{\frac{1}{1+\xi}}\left(\frac{r(n-1)}{2}, \frac{r-1}{2}\right).$$

CHAPTER 5

Statistical Estimation

PART I: THEORY

5.1 GENERAL DISCUSSION

Point estimators are sample statistics that are designed to yield numerical estimates of certain characteristics of interest of the parent distribution. While in testing hypotheses we are generally interested in drawing general conclusions about the characteristics of the distribution, for example, whether its expected value (mean) is positive or negative, in problems of estimation we are concerned with the actual value of the characteristic. Generally, we can formulate, as in testing of hypotheses, a statistical model that expresses the available information concerning the type of distribution under consideration. In this connection, we distinguish between parametric and nonparametric (or distribution free) models. Parametric models specify parametric families of distributions. It is assumed in these cases that the observations in the sample are generated from a parent distribution that belongs to the prescribed family. The estimators that are applied in parametric models depend in their structure and properties on the specific parametric family under consideration. On the other hand, if we do not wish, for various reasons, to subject the estimation procedure to strong assumptions concerning the family to which the parent distribution belongs, a distribution free procedure may be more reasonable. In Example 5.1, we illustrate some of these ideas.

This chapter is devoted to the theory and applications of these types of estimators: unbiased, maximum likelihood, equivariant, moment equations, pretest, and robust estimators.

Examples and Problems in Mathematical Statistics, First Edition. Shelemyahu Zacks.

5.2 UNBIASED ESTIMATORS

5.2.1 General Definition and Example

Unbiased estimators of a characteristic $\theta(F)$ of F in $\mathcal{F}$ is an estimator $\hat{\theta}(\mathbf{X})$ satisfying

$$E_F\{\hat{\theta}(\mathbf{X})\} = \theta(F), \quad \text{for all } F \in \mathcal{F}, \tag{5.2.1}$$

where $\mathbf{X}$ is a random vector representing the sample random variables. For example, if $\theta(F) = E_F\{X\}$, assuming that $E_F\{|X|\} < \infty$ for all $F \in \mathcal{F}$, then the sample mean $\bar{X}_n = \frac{1}{n}\sum X_i$ is an unbiased estimator of $\theta(F)$. Moreover, if $V_F\{X\} < \infty$ for all $F \in \mathcal{F}$, then the sample variance $S_n^2 = \frac{1}{n-1}\sum(X_i - \bar{X}_n)^2$ is an unbiased estimator of $V_F\{X\}$. We note that all the examples of unbiased estimators given here are distribution free. They are valid for any distribution for which the expectation or the variance exist. For parametric models one can do better by using unbiased estimators which are functions of the minimal sufficient statistics. The comparison of unbiased estimators is in terms of their variances. Of two unbiased estimators, the one having a smaller variance is considered better, or more efficient. One reason for preferring the unbiased estimator with the smaller variance is in the connection between the variance of the estimator and the probability that it belongs to a fixed-width interval centered at the unknown characteristic. In Example 5.2, we illustrate a case in which the distribution-free estimator of the expectation is inefficient.

5.2.2 Minimum Variance Unbiased Estimators

In Example 5.2, one can see a case where an unbiased estimator, which is not a function of the minimal sufficient statistic (m.s.s.), has a larger variance than the one based on the m.s.s. The question is whether this result holds generally. The main theorem of this section establishes that if a family of distribution functions admits a **complete** sufficient statistic then the minimum variance unbiased estimator (MVUE) is unique, with probability one, and is a function of that statistic. The following is the fundamental theorem of the theory of unbiased estimation. It was proven by Rao (1945, 1947, 1949), Blackwell (1947), and Lehmann and Scheffé (1950).

Theorem 5.2.1 (The Rao–Blackwell–Lehmann–Scheffé Theorem). *Let $\mathcal{F} = \{F(\mathbf{x};\theta); \theta \in \Theta\}$ be a parametric family of distributions of a random vector $\mathbf{X} = (X_1, \ldots, X_n)$. Suppose that $\omega = g(\theta)$ has an unbiased estimator $\hat{g}(\mathbf{X})$. If $\mathcal{F}$ admits a (minimal) sufficient statistic $T(\mathbf{X})$ then*

$$\hat{\omega} = E\{\hat{g}(\mathbf{X}) \mid T(\mathbf{X})\} \tag{5.2.2}$$

is an unbiased estimator of ω and

$$\mathrm{Var}_\theta\{\hat{\omega}\} \le \mathrm{Var}_\theta\{\hat{g}(\mathbf{X})\}, \tag{5.2.3}$$

for all $\theta \in \Theta$. Furthermore, if $T(\mathbf{X})$ is a complete sufficient statistic then $\hat{\omega}$ is essentially the unique minimum variance, unbiased (MVU) estimator, for each θ in Θ.

Proof. (**i**) Since $T(\mathbf{X})$ is a sufficient statistic, the conditional expectation $E\{\hat{g}(\mathbf{X}) \mid T(\mathbf{X})\}$ does not depend on θ and is therefore a statistic. Moreover, according to the law of the iterated expectations and since $\hat{g}(\mathbf{X})$ is unbiased, we obtain

$$\begin{aligned} g(\theta) &= E_\theta\{\hat{g}(\mathbf{X})\} \\ &= E_\theta\{E\{\hat{g}(\mathbf{X}) \mid T(\mathbf{X})\}\} \\ &= E_\theta\{\hat{\omega}\}, \qquad \text{for all } \theta \in \Theta. \end{aligned} \tag{5.2.4}$$

Hence, $\hat{\omega}$ is an unbiased estimator of $g(\theta)$. By the law of the total variance,

$$\mathrm{Var}_\theta\{\hat{g}(\mathbf{X})\} = E_\theta\{\mathrm{Var}\{\hat{g}(\mathbf{X}) \mid T(\mathbf{X})\}\} + \mathrm{Var}_\theta\{E\{\hat{g}(\mathbf{X}) \mid T(\mathbf{X})\}\}. \tag{5.2.5}$$

The second term on the RHS of (5.2.5) is the variance of $\hat{\omega}$. Moreover, $\mathrm{Var}\{\hat{g}(\mathbf{X}) \mid T(\mathbf{X})\} \geq 0$ with probability one for each θ in Θ. Hence, the first term on the RHS of (5.2.5) is nonnegative. This establishes (5.2.3).

(**ii**) Let $T(\mathbf{X})$ be a complete sufficient statistic and assume that $\hat{\omega} = \phi_1(T(\mathbf{X}))$. Let $\tilde{\omega}(\mathbf{X})$ be any unbiased estimator of $\omega = g(\theta)$, which depends on $T(\mathbf{X})$, i.e., $\tilde{\omega}(\mathbf{X}) = \phi_2(T(\mathbf{X}))$. Then, $E_\theta\{\hat{\omega}\} = E_\theta\{\tilde{\omega}(\mathbf{X})\}$ for all θ. Or, equivalently

$$E_\theta\{\phi_1(T) - \phi_2(T)\} = 0, \qquad \text{all } \theta \in \Theta. \tag{5.2.6}$$

Hence, from the completeness of $T(\mathbf{X})$, $\phi_1(T) = \phi_2(T)$ with probability one for each $\theta \in \Theta$. This proves that $\hat{\omega} = \phi_1(T)$ is essentially unique and implies also that $\hat{\omega}$ has the minimal variance at each θ. QED

Part (i) of the above theorem provides also a method of constructing MVUEs. One starts with any unbiased estimator, as simple as possible, and then determines its conditional expectation, given $T(\mathbf{X})$. This procedure of deriving MVUEs is called in the literature "Rao–Blackwellization." Example 5.3 illustrates this method.

In the following section, we prove and illustrate an information lower bound for variances of unbiased estimators. This lower bound plays an important role in the theory of statistical inference.

5.2.3 The Cramér–Rao Lower Bound for the One-Parameter Case

The following theorem was first proven by Fréchet (1943) and then by Rao (1945) and Cramér (1946). Although conditions (i)–(iii), (v) of the following theorem coincide with conditions (3.7.8) we restate them. Conditions (i)–(iv) will be labeled the Cramér–Rao **(CR) regularity conditions**.

Theorem 5.2.2. *Let $\mathcal{F}$ be a one-parameter family of distributions of a random vector $\mathbf{X} = (X_1, \ldots, X_n)$, having probability density functions (p.d.f.s) $f(\mathbf{x}; \theta)$, $\theta \in \Theta$. Let $\omega(\theta)$ be a differentiable function of θ and $\hat{\omega}(\mathbf{X})$ an unbiased estimator of $\omega(\theta)$. Assume that the following regularity conditions hold:*

(i) *Θ is an open interval on the real line.*

(ii) *$\frac{\partial}{\partial \theta} f(\mathbf{x}; \theta)$ exists (finite) for every $\mathbf{x}$ and every θ in Θ, and $\{x : f(x; \theta) > 0\}$ does not depend on θ.*

(iii) *For each θ in Θ, there exists a $\delta > 0$ and a positive integrable function $G(\mathbf{x}; \theta)$ such that for all $\phi \in (\theta - \delta, \theta + \delta)$*

$$\left| \frac{f(\mathbf{x}; \phi) - f(\mathbf{x}; \theta)}{\phi - \theta} \right| \leq G(\mathbf{x}; \theta).$$

(iv) *For each θ in Θ, there exists a $\delta' > 0$ and a positive integrable function $H(\mathbf{x}; \theta)$ such that, for all $\phi \in (\theta - \delta'; \theta + \delta')$*

$$\left| \hat{\omega}(\mathbf{x}) \frac{f(\mathbf{x}; \phi) - f(\mathbf{x}; \theta)}{\phi - \theta} \right| \leq H(\mathbf{x}; \theta).$$

(v) $0 < I_n(\theta) = E_\theta \left\{ \left[\frac{\partial}{\partial \theta} \log f(\mathbf{X}; \theta) \right]^2 \right\} < \infty$ *for each $\theta \in \Theta$.*

Then,

$$Var_\theta\{\hat{\omega}(\mathbf{X})\} \geq \frac{(\omega'(\theta))^2}{I_n(\theta)}, \quad \text{for all } \theta \in \Theta. \tag{5.2.7}$$

Proof. Consider the covariance, for a given θ value, between $\frac{\partial}{\partial \theta} \log f(\mathbf{X}; \theta)$ and $\hat{\omega}(\mathbf{X})$. We have shown in (3.7.3) that under the above regularity conditions $E_\theta \left\{ \frac{\partial}{\partial \theta} \log f(\mathbf{X}; \theta) \right\} = 0$. Hence,

$$\begin{aligned} \operatorname{cov}\left(\frac{\partial}{\partial \theta} \log f(\mathbf{X}; \theta), \hat{\omega}(\mathbf{X}) \right) &= E_\theta \left\{ \frac{\partial}{\partial \theta} \log f(\mathbf{X}; \theta) \cdot \hat{\omega}(\mathbf{X}) \right\} \\ &= \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} \log f(x; \theta) \cdot \hat{\omega}(\mathbf{x}) f(\mathbf{x}; \theta) d\mathbf{x} \\ &= \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} \hat{\omega}(\mathbf{x}) f(\mathbf{x}; \theta) d\mathbf{x} = \omega'(\theta). \end{aligned} \tag{5.2.8}$$

The interchange of differentiation and integration is justified by condition (iv). On the other hand, by the Schwarz inequality

$$\left(\text{cov}\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta), \hat{\omega}(\mathbf{X})\right)\right)^2 \leq \text{Var}\{\hat{\omega}(X)\}\cdot I_n(\theta), \tag{5.2.9}$$

since the variance of $\dfrac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)$ is equal to the Fisher information function $I_n(\theta)$, and the square of the coefficient of correlation between $\hat{\omega}(\mathbf{X})$ and $\dfrac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)$ cannot exceed 1. From (5.2.8) and (5.2.9), we obtain the Cramér–Rao inequality (5.2.7). QED

We show that if an unbiased estimator $\hat{\theta}(\mathbf{X})$ has a distribution of the one-parameter exponential type, then the variance of $\hat{\theta}(\mathbf{X})$ attains the Cramér–Rao lower bound. Indeed, let

$$f(\hat{\theta};\theta) = h(\hat{\theta})\exp\{\hat{\theta}\psi(\theta) - K(\theta)\}, \tag{5.2.10}$$

where $\psi(\theta)$ and $K(\theta)$ are differentiable, and $\psi'(\theta) \neq 0$ for all θ then

$$E_\theta\{\hat{\theta}\} = +\frac{K'(\theta)}{\psi'(\theta)} \tag{5.2.11}$$

and

$$V_\theta\{\hat{\theta}\} = \frac{-\psi''(\theta)K'(\theta) + \psi'(\theta)K''(\theta)}{(\psi'(\theta))^3}. \tag{5.2.12}$$

Since $\hat{\theta}(\mathbf{X})$ is a sufficient statistic, $I_n(\theta)$ is equal to

$$I_n(\theta) = (\psi'(\theta))^2 V_\theta\{\hat{\theta}(\mathbf{X})\}. \tag{5.2.13}$$

Moreover, $\hat{\theta}(\mathbf{X})$ is an unbiased estimator of $g(\theta) = +K'(\theta)/\psi'(\theta)$. Hence, we readily obtain that

$$V_\theta\{\hat{\theta}(\mathbf{X})\} = \frac{[\psi''(\theta)K'(\theta) - \psi'(\theta)K''(\theta)]^2}{(\psi'(\theta))^6 V_\theta\{\hat{\theta}(\mathbf{X})\}} = (g'(\theta))^2/I_n(\theta). \tag{5.2.14}$$

We ask now the question: if the variance of an unbiased estimator $\hat{\theta}(\mathbf{X})$ attains the Cramér–Rao lower bound, can we infer that its distribution is of the one-parameter exponential type? Joshi (1976) provided a counter example. However, under the right regularity conditions the above implication can be made. These conditions were given first by Wijsman (1973) and then generalized by Joshi (1976).

Bhattacharyya (1946) generalized the Cramér–Rao lower bound to (regular) cases where $\omega(\theta)$ is k-times differentiable at all θ. This generalization shows that, under

further regularity conditions, if $\omega^{(i)}(\theta)$ is the ith derivative of $\omega(\theta)$ and V is a $k \times k$ positive definite matrix, for all θ, with elements

$$V_{ij} = E_\theta \left\{ \frac{1}{f(\mathbf{X};\theta)} \frac{\partial^i}{\partial \theta^i} f(\mathbf{X};\theta) \cdot \frac{1}{f(\mathbf{X};\theta)} \frac{\partial^j}{\partial \theta^j} f(\mathbf{X};\theta) \right\},$$

then

$$\mathrm{Var}_\theta\{\hat{\omega}(\mathbf{X})\} \geq (\omega^{(1)}(\theta), \ldots, \omega^{(k)}(\theta)) V^{-1} (\omega^{(1)}(\theta), \ldots, \omega^{(k)}(\theta))'. \quad (5.2.15)$$

Fend (1959) has proven that if the distribution of X belongs to the one-parameter exponential family, and if the variance of an unbiased estimator of $\omega(\theta)$, $\hat{\omega}(X)$, attains the kth order Bhattacharyya lower bound (BLB) for all θ, but does not attain the $(k-1)$st lower bound, then $\hat{\omega}(X)$ is a polynomial of degree k in $U(X)$.

5.2.4 Extension of the Cramér–Rao Inequality to Multiparameter Cases

The Cramér–Rao inequality can be generalized to estimation problems in k-parameter models in the following manner. Suppose that $\mathcal{F}$ is a family of distribution functions having density functions (or probability functions) $f(x;\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ is a k-dimensional vector. Let $I(\boldsymbol{\theta})$ denote a $k \times k$ **Fisher information matrix**, with elements

$$I_{ij}(\boldsymbol{\theta}) = E_\theta \left\{ \frac{\partial}{\partial \theta_i} \log f(X;\boldsymbol{\theta}) \cdot \frac{\partial}{\partial \theta_j} \log f(X;\boldsymbol{\theta}) \right\}$$

$i, j = 1, \ldots, k$. We obviously assume that for each $\boldsymbol{\theta}$ in the parameter space Θ, $I_{ij}(\boldsymbol{\theta})$ is finite. It is easy to show that the matrix $I(\boldsymbol{\theta})$ is nonnegative definite. We will assume, however, that the Fisher information matrix is **positive definite**. Furthermore, let $g_1(\boldsymbol{\theta}), \ldots, g_r(\boldsymbol{\theta})$ be r parametric functions $r = 1, 2, \ldots, k$. Define the matrix of partial derivatives

$$D(\boldsymbol{\theta}) = \{D_{ij}(\boldsymbol{\theta}); \quad i = 1, \ldots, r; \quad j = 1, \ldots, k\}, \quad (5.2.16)$$

where $D_{ij}(\boldsymbol{\theta}) = \dfrac{\partial}{\partial \theta_j} g_i(\boldsymbol{\theta})$. Let $\hat{\mathbf{g}}(X)$ be an r-dimensional vector of unbiased estimators of $g_1(\boldsymbol{\theta}), \ldots, g_r(\boldsymbol{\theta})$, i.e., $\hat{\mathbf{g}}(X) = (\hat{g}_1(X), \ldots, \hat{g}_r(X))$. Let $\mathbb{V}(\hat{\mathbf{g}})$ denote the variance–covariance matrix of $\hat{\mathbf{g}}(X)$. The Cramér–Rao inequality can then be generalized, under regularity conditions similar to those of the theorem, to yield the inequality

$$\mathbb{V}(\hat{\mathbf{g}}) \geq D(\boldsymbol{\theta})(I(\boldsymbol{\theta}))^{-1} D'(\boldsymbol{\theta}), \quad (5.2.17)$$

in the sense that $\mathfrak{V}(\hat{\mathbf{g}}) - D(\boldsymbol{\theta})(I(\boldsymbol{\theta}))^{-1}D'(\boldsymbol{\theta})$ is a nonnegative definite matrix. In the special case of one parameter function $g(\boldsymbol{\theta})$, if $\hat{\mathbf{g}}(X)$ is an unbiased estimator of $g(\boldsymbol{\theta})$ then

$$\mathrm{Var}_\theta\{\hat{g}(X)\} \geq (\triangledown g(\boldsymbol{\theta}))'(I(\boldsymbol{\theta}))^{-1} \triangledown g(\boldsymbol{\theta}), \tag{5.2.18}$$

where $\triangledown g(\boldsymbol{\theta}) = \left(\frac{\partial}{\partial\theta_1}g(\boldsymbol{\theta}), \ldots, \frac{\partial}{\partial\theta_k}g(\boldsymbol{\theta})\right)'$.

5.2.5 General Inequalities of the Cramér–Rao Type

The Cramér–Rao inequality is based on four stringent assumptions concerning the family of distributions under consideration. These assumptions may not be fulfilled in cases of practical interest. In order to overcome this difficulty, several studies were performed and various different general inequalities were suggested. Blyth and Roberts (1972) provided a general theoretical framework for these generalizations. We present here the essential results.

Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables having a common distribution F that belongs to a one-parameter family $\mathcal{F}$, having p.d.f. $f(x;\theta)$, $\theta \in \Theta$. Suppose that $g(\theta)$ is a parametric function considered for estimation. Let $T(\mathbf{X})$ be a sufficient statistic for $\mathcal{F}$ and let $\hat{g}(T)$ be an unbiased estimator of $g(\theta)$. Let $W(T;\theta)$ be a real-valued random variable such that $\mathrm{Var}_\theta\{W(T;\theta)\} > 0$ and finite for every θ. We also assume that $0 < \mathrm{Var}_\theta\{\hat{g}(T)\} < \infty$ for each θ in Θ. Then, from the Schwarz inequality, we obtain

$$\mathrm{Var}_\theta\{\hat{g}(T)\} \geq \frac{(\mathrm{cov}_\theta(\hat{g}(T), W(T,\theta))^2}{\mathrm{Var}_\theta\{W(T,\theta)\}}, \tag{5.2.19}$$

for every $\theta \in \Theta$. We recall that for the Cramér–Rao inequality, we have used

$$W(T;\theta) = \frac{\partial}{\partial\theta}\log f(X;\theta) = \frac{\partial}{\partial\theta}\log h(T;\theta), \tag{5.2.20}$$

where $h(t;\theta)$ is the p.d.f. of T at θ.

Chapman and Robbins (1951) and Kiefer (1952) considered a family of random variables $W_\phi(T;\theta)$, where ϕ ranges over Θ and is given by the likelihood ratio $W_\phi(T;\theta) = \dfrac{h(T;\phi)}{h(T;\theta)}$. The inequality (5.2.19) then becomes

$$\mathrm{Var}_\theta\{\hat{g}(T)\} \geq \frac{(g(\phi) - g(\theta))^2}{\mathrm{Var}_\theta\{W(T,\theta)\}}. \tag{5.2.21}$$

One obtains then that (5.2.21) holds for each ϕ in Θ. Hence, considering the supremum of the RHS of (5.2.21) over all values of ϕ, we obtain

$$\mathrm{Var}_\theta\{\hat{g}(T)\} \geq \sup_{\phi\in\Theta} \frac{(g(\theta) - g(\phi))^2}{A(\phi, \theta)}, \tag{5.2.22}$$

where $A(\theta, \phi) = \mathrm{Var}_\theta\{W_\phi(T; \theta)\}$. Indeed,

$$\begin{aligned} \mathrm{cov}(\hat{g}(T), W_\phi(T, \theta)) &= E_\phi\{\hat{g}(T)\} - E_\theta\{\hat{g}(T)\} \cdot E_\theta\{W_\phi(T; \theta)\} \\ &= g(\phi) - g(\theta). \end{aligned} \tag{5.2.23}$$

This inequality requires that all the p.d.f.s of T, i.e., $h(t; \theta)$, $\theta \in \Theta$, will be positive on the same set, which is independent of any unknown parameter. Such a condition restricts the application of the Chapman–Robbins inequality. We cannot consider it, for example, in the case of a life-testing model in which the family $\mathcal{F}$ is that of location-parameter exponential distributions, i.e., $f(x; \theta) = I\{x \geq \theta\} \exp\{-(x - \theta)\}$, with $0 < \theta < \infty$. However, one can consider the variable $W_\phi(T; \theta)$ for all ϕ values such that $h(t; \phi) = 0$ on the set $N_\theta = \{t : h(t; \theta) = 0\}$. In the above location-parameter example, we can restrict attention to the set of ϕ values that are greater than Θ. If we denote this set by $C(\theta)$ then we have the Chapman–Robbins inequality as follow:

$$\mathrm{Var}_\theta\{\hat{g}(T)\} \geq \sup_{\phi\in C(\theta)} \frac{(g(\theta) - g(\phi))^2}{A(\theta, \phi)}. \tag{5.2.24}$$

The Chapman–Robbins inequality is applicable, as we have seen in the previous example, in cases where the Cramér–Rao inequality is inapplicable. On the other hand, we can apply the Chapman–Robbins inequality also in cases satisfying the Cramér–Rao regularity conditions. The question is then, what is the relationship between the Chapman–Robbins lower bound and Cramér–Rao lower bound. Chapman and Robbins (1951) have shown that their lower bound is greater than or equal to the Cramér–Rao lower bound for all θ.

5.3 THE EFFICIENCY OF UNBIASED ESTIMATORS IN REGULAR CASES

Let $\hat{g}_1(\mathbf{X})$ and $\hat{g}_2(\mathbf{X})$ be two **unbiased** estimators of $g(\theta)$. Assume that the density functions and the estimators satisfy the Cramér–Rao regularity conditions. The relative efficiency of $\hat{g}_1(\mathbf{X})$ to $\hat{g}_2(\mathbf{X})$ is defined as the ratio of their variances,

$$\mathcal{E}_\theta(\hat{g}_1, \hat{g}_2) = \frac{\sigma^2_{\hat{g}_2}(\theta)}{\sigma^2_{\hat{g}_1}(\theta)}, \tag{5.3.1}$$

where $\sigma^2_{\hat{g}_i}(\theta)$ $(i = 1, 2)$ is the variance of $\hat{g}_i(\mathbf{X})$ at θ. In order to compare all the unbiased estimators of $g(\theta)$ on the same basis, we replace $\sigma^2_{\hat{g}_2}(\theta)$ by the Cramér–Rao lower bound (5.2.7). In this manner, we obtain the efficiency function

$$\mathcal{E}_\theta(\hat{g}) = \frac{(g'(\theta))^2}{I_n(\theta)\sigma^2_{\hat{g}}(\theta)}, \tag{5.3.2}$$

for all $\theta \in \Theta$. This function assumes values between zero and one. It is equal to one, for all θ, if and only if $\sigma^2_{\hat{g}}(\theta)$ attains the Cramér–Rao lower bound, or equivalently, if the distribution of $\hat{g}(\mathbf{X})$ is of the exponential type.

Consider the covariance between $\hat{g}(\mathbf{X})$ and the score function $S(\mathbf{X};\theta) = \frac{\partial}{\partial\theta}\log f(\mathbf{x};\theta)$. As we have shown in the proof of the Cramér–Rao inequality that

$$(g'(\theta))^2 = \rho^2_\theta(\hat{g}, S) I_n(\theta)\sigma^2_{\hat{g}}(\theta), \tag{5.3.3}$$

where $\rho_\theta(\hat{g}, S)$ is the coefficient of correlation between the estimator $\hat{g}$ and the score function, $S(\mathbf{X};\theta)$, at θ. Hence, the efficiency function is

$$\mathcal{E}_\theta(\hat{g}) = \rho^2_\theta(\hat{g}, S). \tag{5.3.4}$$

Moreover, the relative efficiency of two unbiased estimators $\hat{g}_1$ and $\hat{g}_2$ is given by

$$\mathcal{E}_\theta(\hat{g}_1, \hat{g}_2) = \rho^2_\theta(\hat{g}_1, S)/\rho^2_\theta(\hat{g}_2, S). \tag{5.3.5}$$

This relative efficiency can be expressed also in terms of the ratio of the Fisher information functions obtained from the corresponding distributions of the estimators. That is, if $h(\hat{g}_i;\theta)$, $i = 1, 2$, is the p.d.f. of $\hat{g}_i$ and $I^{\hat{g}_i}(\theta) = E_\theta\{[\frac{\partial}{\partial\theta}\log h(\hat{g}_i;\theta)]^2\}$ then

$$\mathcal{E}_\theta(\hat{g}_1, \hat{g}_2) = \frac{I^{\hat{g}_1}(\theta)}{I^{\hat{g}_2}(\theta)}, \quad \theta \in \Theta. \tag{5.3.6}$$

It is a straightforward matter to show that for every unbiased estimator $\hat{g}$ of $g(\theta)$ and under the Cramér–Rao regularity conditions

$$I^{\hat{g}}(\theta) = (g'(\theta))^2/\sigma^2_{\hat{g}}(\theta), \quad \text{for all } \theta \in \Theta. \tag{5.3.7}$$

Thus, the relative efficiency function (5.3.6) can be written, for cases satisfying the Cramér–Rao regularity condition, in the form

$$\mathcal{E}_\theta(\hat{g}_1, \hat{g}_2) = \frac{(g_1'(\theta))^2}{(g_2'(\theta))^2} \cdot \frac{\sigma^2_{\hat{g}_2}(\theta)}{\sigma^2_{\hat{g}_1}(\theta)}, \tag{5.3.8}$$

where $\hat{g}_1(\mathbf{X})$ and $\hat{g}_2(\mathbf{X})$ are unbiased estimators of $g_1(\theta)$ and $g_2(\theta)$, respectively. If the two estimators are unbiased estimators of the same function $g(\theta)$ then (5.3.8) is reduced to (5.3.1). The relative efficiency function (5.3.8) is known as the **Pitman relative efficiency**. It relates both the variances and the derivatives of the bias functions of the two estimators (see Pitman, 1948).

The information function of an estimator can be generalized to the multiparameter regular case (see Bhapkar, 1972). Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ be a vector of k-parameters and $I(\boldsymbol{\theta})$ be the Fisher information matrix (corresponding to one observation). If $g_1(\boldsymbol{\theta}), \ldots, g_r(\boldsymbol{\theta})$, $1 \leq r \leq k$, are functions satisfying the required differentiability conditions and $\hat{g}_1(\mathbf{X}), \ldots, \hat{g}_r(\mathbf{X})$ are the corresponding unbiased estimators then, from (5.2.18),

$$|\Sigma_{\boldsymbol{\theta}}(\hat{\mathbf{g}})| \geq \frac{1}{n}|D(\boldsymbol{\theta})I^{-1}(\boldsymbol{\theta})D'(\boldsymbol{\theta})|, \tag{5.3.9}$$

where n is the sample size. Note that if $r = k$ then $D(\boldsymbol{\theta})$ is nonsingular (the parametric functions $g_1(\boldsymbol{\theta}), \ldots, g_k(\boldsymbol{\theta})$ are linearly independent), and we can express the above inequality in the form

$$|I(\boldsymbol{\theta})| \geq \frac{|D(\boldsymbol{\theta})|^2}{n|\Sigma_\theta(\hat{g})|}, \quad \text{for all } \theta \in \Theta. \tag{5.3.10}$$

Accordingly, and in analogy to (5.3.7), we define the amount of information in the vector estimator $\hat{\mathbf{g}}$ as

$$\mathcal{I}_g(\boldsymbol{\theta}) = \frac{|D(\boldsymbol{\theta})|^2}{|\Sigma_{\boldsymbol{\theta}}(\hat{g})|}. \tag{5.3.11}$$

If $1 \leq r < k$ but $D(\boldsymbol{\theta})$ is of full rank r, then

$$\mathcal{I}_g(\boldsymbol{\theta}) = \frac{|D(\boldsymbol{\theta})D'(\boldsymbol{\theta})|}{|\Sigma_{\boldsymbol{\theta}}(\hat{g})|}. \tag{5.3.12}$$

The efficiency function of a multiparameter estimator is thus defined by DeGroot and Raghavachari (1970) as

$$\mathcal{E}_\theta(\hat{g}_n) = \frac{\mathcal{I}_g(\boldsymbol{\theta})}{\mathcal{I}_n(\boldsymbol{\theta})}. \tag{5.3.13}$$

In Example 5.9, we illustrate the computation needed to determine this efficiency function.

5.4 BEST LINEAR UNBIASED AND LEAST-SQUARES ESTIMATORS

Best linear unbiased estimators (BLUEs) are linear combinations of the observations that yield unbiased estimates of the unknown parameters with minimal variance. As we have seen in Section 5.3, the uniformly minimum variance unbiased (UMVU) estimators (if they exist) are in many cases nonlinear functions of the observations. Accordingly, if we confine attention to linear estimators, the variance of the BLUE will not be smaller than that of the UMVU. On the other hand, BLUEs may exist when UMVU estimators do not exist. For example, if $X_1, \ldots, X_n$ and i.i.d. random variables having a Weibull distribution $G^{1/\beta}(\lambda, 1)$ and both λ and β are **unknown** $0 < \lambda, \beta < \infty$, the m.s.s. is the order statistic $(X_{(1)}, \ldots, X_{(n)})$. Suppose that we wish to estimate the parametric functions $\mu = \frac{1}{\beta} \log \lambda$ and $\sigma = \frac{1}{\beta}$. There are no UMVU estimators of μ and σ. However, there are BLUEs of these parameters.

5.4.1 BLUEs of the Mean

We start with the case where the n random variables have the same unknown mean, μ and the covariance matrix is known. Thus, let $\mathbf{X} = (X_1, \ldots, X_n)'$ be a random vector; $E\{\mathbf{X}\} = \mu\mathbf{1}$, $\mathbf{1}' = (1, 1, \ldots, 1)$; μ is unknown (real). The covariance of $\mathbf{X}$ is $\mathbb{\Sigma}$. We assume that $\mathbb{\Sigma}$ is finite and nonsingular. A linear estimator of μ is a linear function $\hat{\mu} = \boldsymbol{\lambda}'\mathbf{X}$, where $\boldsymbol{\lambda}$ is a vector of known constants. The expected value of $\hat{\mu}$ is μ if, and only if, $\boldsymbol{\lambda}'\mathbf{1} = 1$. We thus consider the class of all such unbiased estimators and look for the one with the smallest variance. Such an estimator is called **best linear unbiased** (BLUE). The variance of $\hat{\mu}$ is $V\{\boldsymbol{\lambda}'\mathbf{X}\} = \boldsymbol{\lambda}'\mathbb{\Sigma}\boldsymbol{\lambda}$. We, therefore, determine $\boldsymbol{\lambda}^0$ that minimizes this variance and satisfies the condition of unbiasedness. Thus, we have to minimize the Lagrangian

$$L(\boldsymbol{\lambda}, \tau) = \boldsymbol{\lambda}'\mathbb{\Sigma}\boldsymbol{\lambda} + \tau(1 - \boldsymbol{\lambda}'\mathbf{1}). \tag{5.4.1}$$

It is simple to show that the minimizing vector is unique and is given by

$$\boldsymbol{\lambda}^0 = \frac{\mathbb{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}'\mathbb{\Sigma}^{-1}\mathbf{1}}. \tag{5.4.2}$$

Correspondingly, the BLUE is

$$\hat{\mu} = \mathbf{1}'\mathbb{\Sigma}^{-1}\mathbf{X}/\mathbf{1}'\mathbb{\Sigma}^{-1}\mathbf{1}. \tag{5.4.3}$$

Note that this BLUE can be obtained also by minimizing the quadratic form

$$Q(\mu) = (\mathbf{X} - \mu\mathbf{1})'\mathbb{\Sigma}^{-1}(\mathbf{X} - \mu\mathbf{1}). \tag{5.4.4}$$

In Example 5.12, we illustrate a BLUE of the form (5.4.3).

5.4.2 Least-Squares and BLUEs in Linear Models

Consider the problem of estimating a vector of parameters in cases where the means of the observations are linear combinations of the unknown parameters. Such models are called **linear models**. The literature on estimating parameters in linear models is so vast that it would be impractical to try listing here all the major studies. We mention, however, the books of Rao (1973), Graybill (1961, 1976), Anderson (1958), Searle (1971), Seber (1977), Draper and Smith (1966), and Sen and Srivastava (1990). We provide here a short exposition of the **least-squares** theory for cases of full linear rank.

Linear models of full rank. Suppose that the random vector $\mathbf{X}$ has expectation

$$E\{\mathbf{X}\} = A\beta, \tag{5.4.5}$$

where $\mathbf{X}$ is an $n \times 1$ vector, A is an $n \times p$ matrix of **known** constants, and β a $p \times 1$ vector of **unknown parameters**. We furthermore assume that $1 \leq p \leq n$ and A is a matrix of full rank, p. The covariance matrix of $\mathbf{X}$ is $\mathbb{V} = \sigma^2 I$, where σ^2 is unknown, $0 < \sigma^2 < \infty$. An estimator of β that minimizes the quadratic form

$$Q(\beta) = (\mathbf{X} - A\beta)'(\mathbf{X} - A\beta) \tag{5.4.6}$$

is called the **least-squares estimator** (LSE). This estimator was discussed in Example 2.13 and in Section 4.6 in connection with testing in normal regression models. The notation here is different from that of Section 4.6 in order to keep it in agreement with the previous notation of the present section. As given by (4.6.5), the LSE of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (A'A)^{-1}A'\mathbf{X}. \tag{5.4.7}$$

Note that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$. To verify it, substitute $A\boldsymbol{\beta}$ in (5.3.7) instead of $\mathbf{X}$. Furthermore, if $B\mathbf{X}$ is an arbitrary unbiased estimator of $\boldsymbol{\beta}$ (B a $p \times n$ matrix of specified constants) then B should satisfy the condition $BA = I$. Moreover, the covariance matrix of $B\mathbf{X}$ can be expressed in the following manner. Write $B = B - S^{-1}A' + S^{-1}A'$, where $S = A'A$. Accordingly, the covariance matrix of BX is

$$\mathbb{V}(B\mathbf{X}) = \mathbb{V}(C\mathbf{X}) + \mathbb{V}(\hat{\boldsymbol{\beta}}) + 2\mathbb{C}(C\mathbf{X}, \hat{\boldsymbol{\beta}}), \tag{5.4.8}$$

where $C = B - S^{-1}A'$, $\hat{\boldsymbol{\beta}}$ is the LSE and $\mathbb{C}(C\mathbf{X}, \hat{\boldsymbol{\beta}})$ is the covariance matrix of $C\mathbf{X}$ and $\hat{\boldsymbol{\beta}}$. This covariance matrix is

$$\begin{aligned} \mathbb{C}(C\mathbf{X}, \hat{\boldsymbol{\beta}}) &= \sigma^2(B - S^{-1}A')AS^{-1} \\ &= \sigma^2(BAS^{-1} - S^{-1}) = 0, \end{aligned} \tag{5.4.9}$$

since $BA = I$. Thus, the covariance matrix of an arbitrary unbiased estimator of $\boldsymbol{\beta}$ can be expressed as the sum of two covariance matrices, one of the LSE, $\hat{\boldsymbol{\beta}}$, and one of $C\mathbf{X}$. $\mathfrak{V}(C\mathbf{X})$ is a nonnegative definite matrix. Obviously, when $B = S^{-1}A'$ the covariance matrix of $C\mathbf{X}$ is 0. Otherwise, all the components of $\hat{\boldsymbol{\beta}}$ have variances which are smaller than or equal to that of $B\mathbf{X}$. Moreover, any linear combination of the components of $\hat{\boldsymbol{\beta}}$ has a variance not exceeding that of $B\mathbf{X}$. It means that the LSE, $\hat{\boldsymbol{\beta}}$, is also BLUE. We have thus proven the celebrated following theorem.

Gauss–Markov Theorem. *If $\mathbf{X} = A\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where A is a matrix of full rank, $E\{\boldsymbol{\epsilon}\} = 0$ and $\mathfrak{V}(\epsilon) = \sigma^2 I$, then the BLUE of any linear combination $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$, where $\boldsymbol{\lambda}$ is a vector of constants and $\hat{\boldsymbol{\beta}}$ is the LSE of $\boldsymbol{\beta}$. Moreover,*

$$\mathrm{Var}\{\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}\} = \sigma^2\boldsymbol{\lambda}'S^{-1}\boldsymbol{\lambda}, \tag{5.4.10}$$

where $S = A'A$.

Note that an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-p}\mathbf{X}'(I - AS^{-1}A')\mathbf{X}. \tag{5.4.11}$$

If the covariance of $\mathbf{X}$ is $\sigma^2 V$, where V is a **known** symmetric positive definite matrix then, after making the factorization $V = DD'$ and the transformation $\mathbf{Y} = D^{-1}\mathbf{X}$ the problem is reduced to the one with covariance matrix proportional to I. Substituting $D^{-1}\mathbf{X}$ for $\mathbf{X}$ and $D^{-1}A$ for A in (5.3.7), we obtain the general formula

$$\hat{\boldsymbol{\beta}} = (A'V^{-1}A)^{-1}A'V^{-1}\mathbf{X}. \tag{5.4.12}$$

The estimator (5.4.12) is the BLUE of $\boldsymbol{\beta}$ and can be considered as the multidimensional generalization of (5.4.3).

As is illustrated in Example 5.10, when V is an arbitrary positive definite matrix, the BLUE (5.3.12) is not necessarily equivalent to the LSE (5.3.7). The conditions under which the two estimators are equivalent were studied by Watson (1967) and Zyskind (1967). The main result is that the BLUE and the LSE coincide when the rank of A is p, $1 \le p \le n$, if and only if there exist p eigenvectors of V which form a basis in the linear space spanned by the columns of A. Haberman (1974) proved the following interesting inequality. Let $\theta = \sum_{i=1}^{p} c_i\beta_i$, where $(c_1, \ldots, c_p)$ are given constants. Let $\hat{\theta}$ and θ^* be, correspondingly, the BLUE and LSE of θ. If τ is the ratio of the largest to the smallest eigenvalues of V then

$$1 \ge \frac{\mathrm{Var}\{\hat{\theta}\}}{\mathrm{Var}\{\theta^*\}} \ge \frac{4\tau}{(1+\tau)^2}. \tag{5.4.13}$$

5.4.3 Best Linear Combinations of Order Statistics

Best linear combinations of order statistics are particularly attractive estimates when the family of distributions under consideration depends on location and scale parameters and the sample is relatively small. More specifically, suppose that $\mathcal{F}$ is a location- and scale-parameter family, with p.d.f.s

$$f(x;\mu,\sigma) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right),$$

where $-\infty < \mu < \infty$ and $0 < \sigma < \infty$. Let $U = (X-\mu)/\sigma$ be the standardized random variable corresponding to X. Suppose that $X_1, \ldots, X_n$ are i.i.d. and let $\mathbf{X}^* = (X_{(1)}, \ldots, X_{(n)})'$ be the corresponding order statistic. Note that

$$X_{(i)} \sim \mu + \sigma U_{(i)}, \quad i = 1, \ldots, n,$$

where $U_1, \ldots, U_n$ are i.i.d. standard variables and $(U_{(1)}, \ldots, U_{(n)})$ the corresponding order statistic. The p.d.f. of U is $\phi(u)$. If the covariance matrix, V, of the order statistic $(U_{(1)}, \ldots, U_{(n)})$ exists, and if $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)'$ denotes the vector of expectations of this order statistic, i.e., $\alpha_i = E\{U_{(i)}\}$, $i = 1, \ldots, n$, then we have the linear model

$$\mathbf{X}^* = [\mathbf{1}, \boldsymbol{\alpha}]\begin{pmatrix}\mu\\ \sigma\end{pmatrix} + \boldsymbol{\epsilon}^*, \tag{5.4.14}$$

where $E\{\boldsymbol{\epsilon}^*\} = \mathbf{0}$ and $\mathbb{V}(\boldsymbol{\epsilon}^*) = V$. This covariance matrix is known. Hence, according to (5.3.12), the BLUE of (μ, σ) is

$$\begin{pmatrix}\hat{\mu}\\ \hat{\sigma}\end{pmatrix} = \begin{pmatrix}\mathbf{1}'V^{-1}\mathbf{1} & \mathbf{1}'V^{-1}\boldsymbol{\alpha}\\ \mathbf{1}'V^{-1}\boldsymbol{\alpha} & \boldsymbol{\alpha}'V^{-1}\boldsymbol{\alpha}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{1}'V^{-1}\mathbf{X}^*\\ \boldsymbol{\alpha}'V^{-1}\mathbf{X}^*\end{pmatrix}. \tag{5.4.15}$$

Let

$$\lambda = (\mathbf{1}'V^{-1}\mathbf{1})(\boldsymbol{\alpha}'V^{-1}\boldsymbol{\alpha}) - (\mathbf{1}'V^{-1}\boldsymbol{\alpha})^2$$

and

$$C = V^{-1}(\mathbf{1}\boldsymbol{\alpha}' - \boldsymbol{\alpha}\mathbf{1}')V^{-1}/\lambda,$$

then the BLUE can be written as

$$\begin{aligned}\hat{\mu} &= -\boldsymbol{\alpha}'C\mathbf{X}^*,\\ \hat{\sigma} &= \mathbf{1}'C\mathbf{X}^*.\end{aligned} \tag{5.4.16}$$

The variances and covariances of these BLUEs are

$$V\{\hat{\mu}\} = \frac{\sigma^2}{\lambda}(\boldsymbol{\alpha}' C^{-1}\boldsymbol{\alpha}),$$
$$V\{\hat{\sigma}\} = \frac{\sigma^2}{\lambda}(\mathbf{1}' C^{-1}\mathbf{1}), \tag{5.4.17}$$

and

$$\text{cov}(\hat{\mu}, \hat{\sigma}) = -\frac{\sigma^2}{\lambda}(\mathbf{1}' V^{-1}\boldsymbol{\alpha}).$$

As will be illustrated in the following example the proposed BLUE, based on all the n order statistics, becomes impractical in certain situations.

Example 5.11 illustrates an estimation problem for which the BLUE based on all the n order statistics can be determined only numerically, provided the sample is not too large. Various methods have been developed to approximate the BLUEs by linear combinations of a small number of selected order statistics. Asymptotic (large sample) theory has been applied in the theory leading to the optimal choice of selected set of k, $k < n$, order statistics. This choice of order statistics is also called **spacing**. For the theories and methods used for the determination of the optimal spacing see the book of Sarhan and Greenberg (1962).

5.5 STABILIZING THE LSE: RIDGE REGRESSIONS

The method of **ridge regression** was introduced by Hoerl (1962) and by Hoerl and Kennard (1970). A considerable number of papers have been written on the subject since then. In particular see the papers of Marquardt (1970), Stone and Conniffe (1973), and others. The main objective of the ridge regression method is to overcome a phenomenon of possible instability of least-squares estimates, when the matrix of coefficients $S = A'A$ has a large spread of the eigenvalues. To be more specific, consider again the linear model of full rank: $\mathbf{X} = A\boldsymbol{\beta} + \epsilon$, where $E\{\boldsymbol{\epsilon}\} = 0$ and $\mathbb{V}(\boldsymbol{\epsilon}) = \sigma^2 I$. We have seen that the LSE of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}} = S^{-1}A'\mathbf{X}$, minimizes the squared distance between the observed random vector X and the estimate of its expectation $A\boldsymbol{\beta}$, i.e., $||\mathbf{X} - A\mathbf{B}||^2$. $||\mathbf{a}||$ denotes the Euclidean length of the vector $\mathbf{a}$, i.e., $||\mathbf{a}|| = \left(\sum_{i=1}^{n} a_i^2\right)^{1/2}$. As we have shown in Section 5.3.2, the LSE in the present model is BLUE of $\boldsymbol{\beta}$. However, if A is ill-conditioned, in the sense that the positive definite matrix $S = A'A$ has large spread of the eigenvalues, with some being close to zero, then the LSE $\hat{\boldsymbol{\beta}}$ may be with high probability very far from β. Indeed, if $L^2 = ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2$ then

$$E\{L^2\} = \sigma^2 tr.\{S^{-1}\}. \tag{5.5.1}$$

Let P be an orthogonal matrix that diagonalizes S, i.e., $PSP' = \Lambda$, where Λ is a diagonal matrix consisting of the eigenvalues $(\lambda_1, \ldots, \lambda_p)$ of S (all positive). Accordingly

$$E\{L^2\} = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}. \tag{5.5.2}$$

We see that $E\{L^2\} \geq \sigma^2 \frac{1}{\lambda_{\min}}$, where $\lambda_{\min}$ is the smallest eigenvalue. A very large value of $E\{L^2\}$ means that at least one of the components of $\boldsymbol{\beta}$ has a large variance. This implies that the corresponding value of β_i may with high probability be far from the true value. The matrix A in experimental situations often represents the levels of certain factors and is generally under control of the experimenter. A good design will set the levels of the factors so that the columns of A will be orthogonal. In this case $S = I$, $\lambda_1 = \ldots = \lambda_p = 1$ and $E\{L^2\}$ attains the minimum possible value $p\sigma^2$ for the LSE. In many practical cases, however, $\mathbf{X}$ is observed with an ill-conditioned coefficient matrix A. In this case, all the unbiased estimators of $\boldsymbol{\beta}$ are expected to have large values of L^2. The way to overcome this deficiency is to consider **biased** estimators of $\boldsymbol{\beta}$ which are not affected strongly by small eigenvalues. Hoerl (1962) suggested the class of biased estimators

$$\hat{\boldsymbol{\beta}}^*(k) = [A'A + kI]^{-1} A'\mathbf{X} \tag{5.5.3}$$

with $k \geq 0$, called the **ridge regression** estimators. It can be shown for every $k > 0$, $\hat{\boldsymbol{\beta}}^*(k)$ has smaller length than the LSE $\hat{\boldsymbol{\beta}}$, i.e., $||\hat{\boldsymbol{\beta}}^*(k)|| < ||\hat{\boldsymbol{\beta}}||$. The ridge estimator is compared to the LSE. If we graph the values of $\beta_i^*(k)$ as functions of k we often see that the estimates are very sensitive to changes in the values of k close to zero, while eventually as k grows the estimates stabilize. The graphs of $\beta_i^*(k)$ for $i = 1, \ldots, k$ are called the **ridge trace**. It is recommended by Hoerl and Kennard (1970) to choose the value of k at which the estimates start to stabilize.

Among all (biased) estimators $\mathbf{B}$ of $\boldsymbol{\beta}$ that lie at a fixed distance from the origin the ridge estimator $\boldsymbol{\beta}^*(k)$, for a proper choice of k, minimizes the residual sum of squares $||\mathbf{X} - A\mathbf{B}||^2$. For proofs of these geometrical properties, see Hoerl and Kennard (1970). The sum of mean-squared errors (MSEs) of the components of $\hat{\boldsymbol{\beta}}^*(k)$ is

$$E\{L^2(k)\} = E\{||\hat{\boldsymbol{\beta}}^*(k) - \boldsymbol{\beta}||^2\} = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{p} \frac{\gamma_i^2}{(\lambda_i + k)^2}, \tag{5.5.4}$$

where $\gamma = H\boldsymbol{\beta}$ and H is the orthogonal matrix diagonalizing $A'A$. $E\{L^2(k)\}$ is a differentiable function of k, having a unique minimum $k^{(0)}(\gamma)$. Moreover, $E\{L^2(k^0(\beta))\} < E\{L^2(0)\}$, where $E\{L^2(0)\}$ is the sum of variances of the LSE components, as in (5.4.2). The problem is that the value of $k^0(\gamma)$ depends on γ and if k is chosen too far from $k^0(\gamma)$, $E\{L^2(k)\}$ may be greater than $E\{L^2(0)\}$. Thus, a crucial problem in applying the ridge-regression method is the choice of a flattening factor

k. Hoerl, Kennard, and Baldwin (1975) studied the characteristics of the estimator obtained by substituting in (5.4.3) an estimate of the optimal $k^0(\gamma)$. They considered the estimator

$$\hat{k} = \frac{p\hat{\sigma}^2}{||\hat{\boldsymbol{\beta}}||^2}, \tag{5.5.5}$$

where $\hat{\boldsymbol{\beta}}$ is the LSE and $\hat{\sigma}^2$ is the estimate of the variance around the regression line, as in (5.4.11). The estimator $\hat{\boldsymbol{\beta}}^*(\hat{k})$ is not linear in X, since k is a nonlinear function of $\mathbf{X}$. Most of the results proven for a fixed value of k do not necessarily hold when k is random, as in (5.5.5). For this reason Hoerl, Kennard, and Baldwin performed extensive simulation experiments to obtain estimates of the important characteristics of $\hat{\boldsymbol{\beta}}^*(\hat{k})$. They found that with probability greater than 0.5 the ridge-type estimator $\tilde{\hat{\boldsymbol{\beta}}}^*(\hat{k})$ is closer (has smaller distance norm) to the true $\boldsymbol{\beta}$ than the LSE. Moreover, this probability increases as the dimension p of the factor space increases and as the spread of the eigenvalues of S increases. The ridge type estimator $\hat{\boldsymbol{\beta}}^*(\hat{k})$ are similar to other types of nonlinear estimators (James–Stein, Bayes, and other types) designed to reduce the MSE. These are discussed in Chapter 8.

A more general class of ridge-type estimators called the **generalized ridge regression estimators** is given by

$$\hat{\boldsymbol{\beta}} = (A'A + C)^{-1}A'\mathbf{X}, \tag{5.5.6}$$

where C is a positive definite matrix chosen so that $A'A + C$ is nonsingular. [The class is actually defined also for $A'A + C$ singular with a Moore–Penrose generalized inverse replacing $(A'A + C)^{-1}$; see Marquardt (1970).]

5.6 MAXIMUM LIKELIHOOD ESTIMATORS

5.6.1 Definition and Examples

In Section 3.3, we introduced the notion of the likelihood function, $L(\theta; x)$ defined over a parameter space Θ, and studied some of its properties. We develop here an estimation theory based on the likelihood function.

The **maximum likelihood** estimator (MLE) of θ is a value of θ at which the likelihood function $L(\theta; x)$ attains its supremum (or maximum). We remark that if the family $\mathcal{F}$ admits a nontrivial sufficient statistic $T(\mathbf{X})$ then the MLE is a function of $T(\mathbf{X})$. This is implied immediately from the Neyman–Fisher Factorization Theorem. Indeed, in this case,

$$f(\mathbf{x}; \theta) = h(\mathbf{x})g(T(\mathbf{x}); \theta),$$

where $h(\mathbf{x}) > 0$ with probability one. Hence, the kernel of the likelihood function can be written as $L^*(\theta; \mathbf{x}) = g(T(\mathbf{x}); \theta)$. Accordingly, the value θ that maximizes it

depends on $T(\mathbf{x})$. We also notice that although the MLE is a function of the sufficient statistic, the converse is not always true. An MLE is not necessarily a sufficient statistic.

5.6.2 MLEs in Exponential Type Families

Let $X_1, \ldots, X_n$ be i.i.d. random variables having a k-parameter exponential type family, with a p.d.f. of the form (2.16.2). The likelihood function of the natural parameters is

$$L(\boldsymbol{\psi}; \mathbf{X}) = \exp\left\{\sum_{i=1}^{k} \psi_i T_i(\mathbf{X}) - nK(\boldsymbol{\psi})\right\}, \tag{5.6.1}$$

where

$$T_i(\mathbf{X}) = \sum_{j=1}^{n} U_i(X_j), \quad i = 1, \ldots, k.$$

The MLEs of $\psi_1, \ldots, \psi_k$ are obtained by solving the system of k equations

$$\begin{aligned} \frac{\partial}{\partial \psi_1} K(\boldsymbol{\psi}) &= \frac{1}{n}\sum_{j=1}^{n} U_1(X_j), \\ &\vdots \\ \frac{\partial}{\partial \psi_k} K(\boldsymbol{\psi}) &= \frac{1}{n}\sum_{j=1}^{n} U_k(X_j). \end{aligned} \tag{5.6.2}$$

Note that whenever the expectations exist, $E_{\boldsymbol{\psi}}\{U_i(X)\} = \partial K(\boldsymbol{\psi})/\partial \psi_i$ for each $i = 1, \ldots, k$. Hence, if $X_1, \ldots, X_n$ are i.i.d. $E_{\boldsymbol{\psi}}\left\{\frac{\partial}{\partial \psi_i} K(\hat{\boldsymbol{\psi}})\right\} = \partial K(\boldsymbol{\psi})/\partial \psi_i$, for each $i = 1, \ldots, k$, where $\hat{\boldsymbol{\psi}}$ is the vector of MLEs. For all points $\boldsymbol{\psi}$ in the interior of the parameter space n, the matrix $\left(-\frac{\partial^2}{\partial \psi_i \partial \psi_j} K(\boldsymbol{\psi}); i, j = 1, \ldots, k\right)$ exists and is positive definite for all $\boldsymbol{\psi}$ since $K(\boldsymbol{\psi})$ is convex. Thus, the root $\hat{\boldsymbol{\psi}}$ of (5.6.2) is unique and is a m.s.s.

5.6.3 The Invariance Principle

If the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is reparametrized by a one-to-one transformation $\psi_1 = g_1(\boldsymbol{\theta}), \ldots, \psi_k = g_k(\boldsymbol{\theta})$ then the MLEs of ψ_i are obtained by substituting in the g-functions the MLEs of $\boldsymbol{\theta}$. This is obviously true when the transformation $\boldsymbol{\theta} \to \boldsymbol{\psi}$ is one-to-one. Indeed, if $\theta_1 = g_1^{-1}(\boldsymbol{\psi}), \ldots, \theta_k = g_k^{-1}(\boldsymbol{\psi})$ then the likelihood function $L(\theta; \mathbf{x})$ can be expressed as a function of $\boldsymbol{\psi}$, $L(g_1^{-1}(\boldsymbol{\psi}), \ldots, g_k^{-1}(\boldsymbol{\psi}); \mathbf{x})$. If $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$

is a point at which $L(\boldsymbol{\theta}, \mathbf{x})$ attains its supremum, and if $\hat{\boldsymbol{\psi}} = (g_1(\hat{\boldsymbol{\theta}}), \ldots, g_k(\hat{\boldsymbol{\theta}}))$ then, since the transformation is one-to-one,

$$\sup_{\theta} L(\boldsymbol{\theta}; \mathbf{x}) = L(\hat{\boldsymbol{\theta}}; \mathbf{x}) = L(g_1^{-1}(\hat{\boldsymbol{\psi}}), \ldots, g_k^{-1}(\hat{\boldsymbol{\psi}}); \mathbf{x}) = L^*(\hat{\boldsymbol{\psi}}) = \sup_{\psi} L^*(\boldsymbol{\psi}; \mathbf{x}), \tag{5.6.3}$$

where $L^*(\boldsymbol{\psi}; \mathbf{x})$ is the likelihood, as a function of $\boldsymbol{\psi}$. This result can be extended to general transformations, not necessarily one-to-one, by a proper redefinition of the concept of MLE over the space of the $\boldsymbol{\psi}$-values. Let $\boldsymbol{\psi} = g(\boldsymbol{\theta})$ be a vector valued function of $\boldsymbol{\theta}$; i.e., $\boldsymbol{\psi} = g(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), \ldots, g_k(\boldsymbol{\theta}))$ where the dimension of $g(\boldsymbol{\theta})$, r, does not exceed that of $\boldsymbol{\theta}$, k.

Following Zehna (1966), we introduce the notion of the **profile** likelihood function of $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_r)$. Define the cosets of $\boldsymbol{\theta}$-values

$$G(\boldsymbol{\psi}) = \{\boldsymbol{\theta}; g(\boldsymbol{\theta}) = \boldsymbol{\psi}\}, \tag{5.6.4}$$

and let $L(\boldsymbol{\theta}; x)$ be the likelihood function of $\boldsymbol{\theta}$ given $\mathbf{x}$. The profile likelihood of $\boldsymbol{\psi}$ given $\mathbf{x}$ is defined as

$$L^*(\boldsymbol{\psi}; \mathbf{x}) = \sup_{\theta \in G(\boldsymbol{\psi})} L(\boldsymbol{\theta}; \mathbf{x}). \tag{5.6.5}$$

Obviously, in the one-to-one case $L^*(\boldsymbol{\theta}; \mathbf{x}) = L(g_1^{-1}(\boldsymbol{\theta}), \ldots, g_k^{-1}(\boldsymbol{\theta}); \mathbf{x}_n)$. Generally, we define the MLE of $\boldsymbol{\psi}$ to be the value at which $L^*(\boldsymbol{\psi}; \mathbf{x})$ attains its supremum. It is easy then to prove that **if $\hat{\boldsymbol{\theta}}$ is an MLE of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\psi}} = g(\hat{\boldsymbol{\theta}})$, then $\hat{\boldsymbol{\psi}}$ is an MLE of $\boldsymbol{\psi}$**, i.e.,

$$\sup_{\psi} L^*(\boldsymbol{\psi}; \mathbf{x}) = L^*(\hat{\boldsymbol{\psi}}; \mathbf{x}). \tag{5.6.6}$$

5.6.4 MLE of the Parameters of Tolerance Distributions

Suppose that k-independent experiments are performed at controllable real-valued experimental levels (dosages) $-\infty < x_1 < \ldots < x_k < \infty$. At each of these levels n_j Bernoulli trials are performed ($j = 1, \ldots, k$). The success probabilities of these Bernoulli trials are increasing functions $F(x)$ of x. These functions, called **tolerance distributions**, are the expected proportion of (individuals) units in a population whose tolerance against the applied dosage does not exceed the level x. The model thus consists of k-independent random variables $J_1, \ldots, J_k$ such that $J_i \sim B(n_i, F(x_i; \boldsymbol{\theta}))$, $i = 1, \ldots, k$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$, $1 \leq r < k$, is a vector of unknown parameters. The problem is to estimate $\boldsymbol{\theta}$. Frequently applied models are

$$F(\mathbf{x}; \theta) = \begin{cases} \Phi(\alpha + \beta x), & \text{normal distributions;} \\ (1 + \exp\{-(\alpha + \beta x)\})^{-1}, & \text{logistic distributions;} \\ \exp\{-\exp\{-(\alpha + \beta x)\}\}, & \text{extreme-value distribution.} \end{cases} \tag{5.6.7}$$

We remark that in some of the modern literature the tolerance distributions are called **link functions** (see Lindsey, 1996). Generally, if $F(\alpha + \beta x_i)$ is the success probability at level x_i, the likelihood function of (α, β), given $J_1, \ldots, J_k$ and $x_1, \ldots, x_k$, $n_1, \ldots, n_k$, is

$$L(\alpha, \beta \mid \mathbf{J}, \mathbf{x}, \mathbf{n}) = \prod_{i=1}^{k} \left[\frac{F(\alpha + \beta x_i)}{1 - F(\alpha + \beta x_i)} \right]^{J_i} \cdot \prod_{i=1}^{k} [1 - F(\alpha + \beta x_i)]^{n_i}, \qquad (5.6.8)$$

and the log-likelihood function is

$$\log L(\alpha, \beta \mid \mathbf{J}, \mathbf{x}, \mathbf{n}) = \sum_{i=1}^{k} J_i \log \frac{F(\alpha + \beta x_i)}{1 - F(\alpha + \beta x_i)} + \sum_{i=1}^{k} n_i \log(1 - F(\alpha + \beta x_i)).$$

The MLE of α and β are the roots of the nonlinear equations

$$\begin{aligned} \sum_{i=1}^{k} J_j \frac{f(\alpha + \beta x_j)}{F(\alpha + \beta x_j)\bar{F}(\alpha + \beta x_j)} &= \sum_{j=1}^{k} n_j \frac{f(\alpha + \beta x_j)}{\bar{F}(\alpha + \beta x_j)}, \\ \sum_{j=1}^{k} x_j J_j \frac{f(\alpha + \beta x_j)}{F(\alpha + \beta x_j)\bar{F}(\alpha + \beta x_j)} &= \sum_{j=1}^{k} n_j x_j \frac{f(\alpha + \beta x_j)}{\bar{F}(\alpha + \beta x_j)}, \end{aligned} \qquad (5.6.9)$$

where $f(z) = F'(z)$ is the p.d.f. of the standardized distribution $F(z)$ and $\bar{F}(z) = 1 - F(z)$.

Let $\hat{p}_i = J_i/n_i$, $i = 1, \ldots, k$, and define the function

$$G(z; \hat{p}) = \frac{f(z)(\hat{p} - F(z))}{F(z)\bar{F}(z)}, \quad -\infty < z < \infty. \qquad (5.6.10)$$

Accordingly, the MLEs of α and β are the roots $\hat{\alpha}$ and $\hat{\beta}$ of the equations

$$\sum_{i=1}^{k} n_i G(\hat{\alpha} + \hat{\beta} x_i; \hat{p}_i) = 0 \qquad (5.6.11)$$

and

$$\sum_{i=1}^{k} x_i n_i G(\hat{\alpha} + \hat{\beta} x_i; \hat{p}_i) = 0.$$

The solution of this system of (generally nonlinear) equations according to the Newton–Raphson method proceeds as follows. Let $\hat{\alpha}_0$ and $\hat{\beta}_0$ be an initial solution.

The adjustment after the jth iteration ($j = 0, 1, \ldots$) is $\hat{\alpha}_{j+1} = \hat{\alpha}_j + \delta\alpha_j$ and $\hat{\beta}_{j+1} = \hat{\beta}_j + \delta\beta_j$, where $\delta\alpha_j$ and $\delta\beta_j$ are solutions of the linear equations

$$\begin{pmatrix} \sum_{i=1}^{k} W_i^{(j)} & \sum_{i=1}^{k} x_i W_i^{(j)} \\ \sum_{i=1}^{k} x_i W_i^{(j)} & \sum_{i=1}^{k} x_i^2 W_i^{(j)} \end{pmatrix} \begin{pmatrix} \delta\alpha_j \\ \delta\beta_j \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{k} Y_i^{(j)} \\ \sum_{i=1}^{k} x_i Y_i^{(j)} \end{pmatrix}, \tag{5.6.12}$$

where

$$W_i^{(j)} = n_i G'(\hat{\alpha}_j + \hat{\beta}_j x_i; \hat{p}_i) \tag{5.6.13}$$

and

$$Y_i^{(j)} = -n_i G(\hat{\alpha}_j + \hat{\beta}_j x_i; \hat{p}_i)$$

and $G'(z; \hat{p}) = \frac{\partial}{\partial z} G(z; \hat{p})$. The linear equations (5.6.12) resemble the normal equations in weighted least-squares estimation. However, in the present problems the weights depend on the unknown parameters α and β. In each iteration, the current estimates of α and β are substituted. For applications of this procedure in statistical reliability and bioassay quantal response analysis, see Finney (1964), Gross and Clark (1975), and Zacks (1997).

5.7 EQUIVARIANT ESTIMATORS

5.7.1 The Structure of Equivariant Estimators

Certain families of distributions have structural properties that are preserved under transformations of the random variables. For example, if X has an absolutely continuous distribution belonging to a family $\mathcal{F}$ which depends on location and scale parameters, i.e., its p.d.f. is $f(x; \mu, \sigma) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)$, where $-\infty < \mu < \infty$ and $0 < \sigma < \infty$, then any real-affine transformation of X, given by

$$[\alpha, \beta]X = \alpha + \beta X, \quad -\infty < \alpha < \infty, \quad 0 < \beta < \infty$$

yields a random variable $Y = \alpha + \beta X$ with p.d.f. $f(y; \mu, \sigma) = \frac{1}{\bar{\sigma}}\phi\left(\frac{y-\bar{\mu}}{\bar{\sigma}}\right)$, where $\bar{\mu} = \alpha + \beta\mu$ and $\bar{\sigma} = \beta\sigma$. Thus, the distribution of Y belongs to the same family $\mathcal{F}$. The family $\mathcal{F}$ is preserved under transformations belonging to the **group** $\mathcal{G} = \{[\alpha, \beta]; -\infty < \alpha < \infty, 0 < \beta < \infty\}$ of real-affine transformations.

In this section, we present the elements of the theory of families of distributions and corresponding estimators having structural properties that are preserved under

certain groups of transformations. For a comprehensive treatment of the theory and its geometrical interpretation, see the book of Fraser (1968). Advanced treatment of the subject can be found in Berk (1967), Hall, Wijsman, and Ghosh (1965), Wijsman (1990), and Eaton (1989). We require that every element g of $\mathcal{G}$ be a one-to-one transformation of $\mathcal{X}$ onto $\mathcal{X}$. Accordingly, the sample space structure does not change under these transformations. Moreover, if $\mathcal{B}$ is the Borel σ-field on $\mathcal{X}$ then, for all $g \in \mathcal{G}$, we require that $P_\theta[gB]$ will be well defined for all $B \in \mathcal{B}$ and $\theta \in \Theta$. Furthermore, as seen in the above example of the location and scale parameter distributions, if θ is a parameter of the distribution of X the parameter of $Y = gX$ is $\bar{g}\theta$, where $\bar{g}$ is a transformation on the parameter space Θ defined by the relationship

$$P_\theta[B] = P_{\bar{g}\theta}[gB], \quad \text{for every } B \in \mathcal{B}. \tag{5.7.1}$$

In the example of real-affine transformations, if $g = [\alpha, \beta]$ and $\theta = (\mu, \sigma)$, then $\bar{g}(\mu, \sigma) = (\alpha + \beta\mu, \beta\sigma)$. We note that $\bar{g}\Theta = \Theta$ for every $\bar{g}$ corresponding to g in $\mathcal{G}$. Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables whose distribution F belongs to a family $\mathcal{F}$ that is preserved under transformations belonging to a group $\mathcal{G}$. If $T(X_1, \ldots, X_n)$ is a statistic, then we define the transformations $\tilde{g}$ on the range $\mathcal{T}$ of $T(X_1, \ldots, X_n)$, corresponding to transformations g of $\mathcal{G}$, by

$$\tilde{g}T(x_1, \ldots, x_n) = T(gx_1, \ldots, gx_n). \tag{5.7.2}$$

A statistic $S(X_1, \ldots, X_n)$ is called **invariant** with respect to $\mathcal{G}$ if

$$\tilde{g}S(X_1, \ldots, X_n) = S(X_1, \ldots, X_n) \quad \text{for every } g \in \mathcal{G}. \tag{5.7.3}$$

A coset of x^0 with respect to $\mathcal{G}$ is the set of all points that can be obtained as images of x^0, i.e.,

$$C(x^0) = \{x : x = gx^0, \text{ for all } g \in \mathcal{G}\}.$$

Such a coset is called also an **orbit** of $\mathcal{G}$ in $\mathcal{X}$ through $\mathbf{x}^0$. If $\mathbf{x}^0 = (x_1^0, \ldots, x_n^0)$ is a given vector, the orbit of $\mathcal{G}$ in $\mathcal{X}^{(n)}$ through $\mathbf{x}^0$ is the coset

$$C(\mathbf{x}^0) = \{\mathbf{x} : \mathbf{x} = (gx_1^0, \ldots, gx_n^0), \text{ for all } g \in \mathcal{G}\}.$$

If $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ belong to the same orbit and $S(\mathbf{x}) = S(x_1, \ldots, x_n)$ is invariant with respect to $\mathcal{G}$ then $S(\mathbf{x}^{(1)}) = S(\mathbf{x}^{(2)})$. A statistic $U(\mathbf{X}) = U(X_1, \ldots, X_n)$ is called **maximal invariant** if it is invariant and if $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ belong to two different orbits then $U(\mathbf{X}^{(1)}) \neq U(\mathbf{X}^{(2)})$. **Every invariant statistic is a function of a maximal invariant statistic**.

If $\hat{\theta}(X_1, \ldots, X_n)$ is an estimator of θ, it would be often desirable to have the property that the estimator reacts to transformations of $\mathcal{G}$ in the same manner as the parameters θ do, i.e.,

$$\hat{\theta}(\tilde{g}\mathbf{x}) = \bar{g}\hat{\theta}(x), \quad \text{for every } g \in \mathcal{G}. \tag{5.7.4}$$

5.7.2 Minimum MSE Equivariant Estimators

Estimators satisfying (5.7.4) are called **equivariant**. The objective is to derive an equivariant estimator having a **minimum MSE** or another optimal property. The algebraic structure of the problem allows us often to search for such optimal estimators in a systematic manner.

5.7.3 Minimum Risk Equivariant Estimators

A loss function $L(\hat{\theta}(\mathbf{X}), \theta)$ is called **invariant** under $\mathcal{G}$ if

$$L(\bar{g}\hat{\theta}(\mathbf{X}), \bar{g}\theta) = L(\hat{\theta}(\mathbf{X}), \theta), \tag{5.7.5}$$

for all $\theta \in \Theta$ and all $g \in \mathcal{G}$.

The coset $C(\theta_0) = \{\theta; \theta = \bar{g}\theta_0, g \in \mathcal{G}\}$ is called an **orbit** of $\mathcal{G}$ through θ_0 in Θ. We show now that if $\hat{\theta}(\mathbf{X})$ is an equivariant estimator and $L(\hat{\theta}(\mathbf{X}), \theta)$ is an invariant loss function then the *risk function* $R(\hat{\theta}, \theta) = E\{L(\hat{\theta}(\mathbf{X}), \theta)\}$ is constant on each orbit of $\mathcal{G}$ in Θ. Indeed, for any $g \in \mathcal{G}$, if the distribution of X is $F(x; \theta)$ and the distribution of $Y = gX$ is $F(y; \bar{g}\theta)$, then if $\hat{\theta}$ is equivariant

$$\begin{aligned} R(\hat{\theta}, \theta) &= \int L(\hat{\theta}(x), \theta) dF(x; \theta) \\ &= \int L(\bar{g}\hat{\theta}(x), \bar{g}\theta) dF(x; \theta) \\ &= \int L(\hat{\theta}(gx), \bar{g}\theta) dF(x; \theta) \\ &= \int L(\hat{\theta}(y), \bar{g}\theta) dF(y; \bar{g}\theta) \\ &= R(\hat{\theta}, \bar{g}\theta), \quad \text{for all } g \in \mathcal{G}. \end{aligned} \tag{5.7.6}$$

Thus, whenever the structure of the model is such that Θ contains only one orbit with respect to $\mathcal{G}$, and there exist equivariant estimators with finite risk, then each such equivariant estimator has a constant risk function. In Example 5.23, we illustrate such cases. We consider there the location and scale parameter family of the normal distributions $N(\mu, \sigma)$. This family has a parameter space Θ, which has only one orbit with respect to the group $\mathcal{G}$ of real-affine transformations. If the parameter space has

various orbits, as in the case of Example 5.24, there is no global uniformly minimum risk equivariant estimator, but only locally for each orbit. In Example 5.26, we construct uniformly minimum risk equivariant estimators of the scale and shape parameters of Weibull distributions for a group of transformations and a corresponding invariant loss function.

5.7.4 The Pitman Estimators

We develop here the minimum MSE equivariant estimators for the special models of **location parameters** and **location** and **scale parameters**. These estimators are called the **Pitman estimators**.

Consider first the family $\mathcal{F}$ of location parameters distributions, i.e., every p.d.f. of $\mathcal{F}$ is given by $f(x;\theta) = \phi(x-\theta)$, $-\infty < \theta < \infty$. $\phi(x)$ is the standard p.d.f. According to our previous discussion, we consider the group $\mathcal{G}$ of real translations. Let $\hat{\theta}(\mathbf{X})$ be an equivariant estimator of θ. Then, writing $T = (\hat{\theta}, X_{(1)} - \hat{\theta}, \ldots, X_{(n)} - \hat{\theta})$, where $X_{(1)} \leq \ldots \leq X_{(n)}$, for any equivariant estimator, $d(\mathbf{X})$, of θ, we have

$$d(\mathbf{X}) = \hat{\theta} + \psi(X_{(1)} - \hat{\theta}, \ldots, X_{(n)} - \hat{\theta}).$$

Note that $\mathbf{U} = (X_{(1)} - \hat{\theta}, \ldots, X_{(n)} - \hat{\theta})$ has a distribution that does not depend on θ. Moreover, since $\hat{\theta}(\mathbf{X})$ is an equivariant estimator, we can write

$$\hat{\theta}(\mathbf{X}) = \theta + F(\mathbf{Y}), \quad \text{where } \mathbf{Y} = \mathbf{X} - \theta\mathbf{1}.$$

Thus, the MSE of $d(\mathbf{X})$ is

$$\text{MSE}\{d\} = E\{[F(\mathbf{Y}) + \psi(X_{(1)} - \hat{\theta}, \ldots, X_{(n)} - \hat{\theta})]^2\}. \tag{5.7.7}$$

It follows immediately that the function $\psi(\mathbf{U})$ which minimizes the MSE is the conditional expectation

$$\psi^0(\mathbf{U}) = -E\{F(\mathbf{Y}) \mid \mathbf{U}\}. \tag{5.7.8}$$

Thus, the minimum MSE equivariant estimator is

$$d^0(\mathbf{X}) = \hat{\theta}(\mathbf{X}) - E\{F(\mathbf{Y}) \mid \mathbf{U}\}. \tag{5.7.9}$$

This is a generalized form of the Pitman estimator. The well-known specific form of the Pitman estimator is obtained by starting with $\hat{\theta}(\mathbf{X}) = X_{(1)}$. In this case,

$F(\mathbf{Y}) = Y_{(1)}$, where $Y_{(1)}$ is the minimum of a sample from a standard distribution. Formula (5.7.9) is then reduced to the special form

$$\begin{aligned} d^0(\mathbf{X}) &= X_{(1)} - E\{Y_{(1)} \mid X_{(2)} - X_{(1)}, \ldots, X_{(n)} - X_{(1)}\} \\ &= X_{(1)} - \frac{\int_{-\infty}^{\infty} u\phi(u)\prod_{i=2}^{n}\phi(Y_{(i)} + u)du}{\int_{-\infty}^{\infty} \phi(u)\prod_{i=2}^{n}\phi(Y_{(i)} + u)du}, \end{aligned} \tag{5.7.10}$$

where $Y_{(i)} = X_{(i)} - X_{(1)}, i = 2, \ldots, n$. In the derivation of (5.7.9), we have assumed that the MSE of $d(\mathbf{X})$ exists. A minimum risk equivariant estimator may not exist. Finally, we mentioned that the minimum MSE equivariant estimators are **unbiased**. Indeed

$$E_\theta\{d^0(\mathbf{x})\} = \theta + E\{F(\mathbf{Y}) - E\{F(\mathbf{Y}) \mid \mathbf{u}\}\} = \theta, \quad -\infty < \theta < \infty. \tag{5.7.11}$$

If $\mathcal{F}$ is a scale and location family of distribution, with p.d.f.s of the form

$$f(x; \mu, \sigma) = \frac{1}{\sigma}\phi\left(\frac{x - \mu}{\sigma}\right), \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad 0 < \sigma < \infty,$$

where $\phi(u)$ is a p.d.f., then every equivariant estimator of μ with respect to the group $\mathcal{G}$ of real-affine transformations can be expressed in the form

$$\hat{\mu}(\mathbf{X}) = X_{(1)} + (X_{(2)} - X_{(1)})\psi(\mathbf{Z}), \tag{5.7.12}$$

where $X_{(1)} \leq \ldots \leq X_{(n)}$ is the order statistic, $X_{(2)} - X_{(1)} > 0$ and $\mathbf{Z} = (Z_3, \ldots, Z_n)'$, with $Z_i = (X_{(i)} - X_{(1)})/(X_{(2)} - X_{(1)})$. The MSE of $\hat{\mu}(\mathbf{X})$ is given by

$$\begin{aligned} \mathrm{MSE}\{\hat{\mu}(\mathbf{X}); \mu, \sigma\} &= \sigma^2 E_0\{[X_{(1)} + (X_{(2)} - X_{(1)})\psi(\mathbf{Z})]^2\} \\ &= \sigma^2 E_0\{E_0\{[X_{(1)} + (X_{(2)} - X_{(1)})\psi(\mathbf{Z})]^2 \mid \mathbf{Z}\}\}, \end{aligned} \tag{5.7.13}$$

where $E_0\{\cdot\}$ designates an expectation with respect to the standard distribution ($\mu = 0$, $\sigma = 1$). An optimal choice of $\psi(\mathbf{Z})$ is such for which $E_0\{[X_{(1)} + (X_{(2)} - X_{(1)})\psi(\mathbf{Z})]^2 \mid \mathbf{Z}\}$ is minimal. Thus, the minimum MSE equivariant estimator of μ is

$$\hat{\mu}(\mathbf{X}) = X_{(1)} + (X_{(2)} - X_{(1)})\psi^0(\mathbf{Z}), \tag{5.7.14}$$

where

$$\psi^0(\mathbf{Z}) = -\frac{E_0\{X_{(1)}(X_{(2)} - X_{(1)}) \mid \mathbf{Z}\}}{E_0\{(X_{(2)} - X_{(1)})^2 \mid \mathbf{Z}\}}. \tag{5.7.15}$$

Equivalently, the Pitman estimator of the location parameter is expressed as

$$\hat{\mu}(\mathbf{X}) = X_{(1)} - (X_{(2)} - X_{(1)}) \cdot \cdot \frac{\int_{-\infty}^{\infty} u\phi(u) \int_0^{\infty} v^{n-1} \phi(u+v) \prod_{i=3}^{n} \phi(u + vZ_i) dv du}{\int_{-\infty}^{\infty} \phi(u) \int_0^{\infty} v^{n} \phi(u+v) \prod_{i=3}^{n} \phi(u + vZ_i) dv du}. \tag{5.7.16}$$

In a similar manner, we show that the minimum MSE equivariant estimator for σ is $\hat{\sigma}_0(\mathbf{X}_n) = (X_{(2)} - X_{(1)})\psi^0(Z_3, \ldots, Z_n)$, where

$$\psi^0(Z_3, \ldots, Z_n) = \frac{E\{U_2 \mid Z_3, \ldots, Z_n\}}{E\{U_2^2 \mid Z_3, \ldots, Z_n\}}. \tag{5.7.17}$$

Indeed, $\psi^0(\mathbf{Z})$ minimizes $E_0\{(U_2\psi(\mathbf{Z}) - 1)^2 \mid \mathbf{Z}\}$. Accordingly, the Pitman estimator of the scale parameter, σ, is

$$\hat{\sigma}_0(\mathbf{X}_n) = (X_{(2)} - X_{(1)}) \cdot \frac{\int_{\infty}^{\infty} \phi(u_1) \int_0^{\infty} u_2^{n-1} \phi(u_1 + u_2) \prod_{i=3}^{n} \phi(u_1 + u_2 Z_i) du_2 du_1}{\int_{\infty}^{\infty} \phi(u_1) \int_0^{\infty} u_2^{n} \phi(u_1 + u_2) \prod_{i=3}^{n} \phi(u_1 + u_2 Z_i) du_2 du_1}. \tag{5.7.18}$$

5.8 ESTIMATING EQUATIONS

5.8.1 Moment-Equations Estimators

Suppose that $\mathcal{F}$ is a family of distributions depending on k real parameters, $\theta_1, \ldots, \theta_k$, $1 \leq k$. Suppose that the moments μ_r, $1 \leq r \leq k$, exist and are given by some specified functions

$$\mu_r = M_r(\theta_1, \ldots, \theta_k), \quad 1 \leq r \leq k.$$

If $X_1, \ldots, X_n$ are i.i.d. random variables having a distribution in $\mathcal{F}$, the sample moments $M_r = \frac{1}{n}\Sigma X_j^r$ are unbiased estimators of $\mu_r (1 \leq r \leq k)$ and by the laws of large numbers (see Section 1.11) they converge almost surely to μ_r as $n \to \infty$. The roots of the system of equations

$$M_r = M_r(\hat{\theta}_1, \ldots, \hat{\theta}_k), \quad 1 \leq r \leq k, \tag{5.8.1}$$

are called the **moment-equations estimators** (MEEs) of $\theta_1, \ldots, \theta_k$.

In Examples 5.28–5.29, we discuss cases where both the MLE and the MEE can be easily determined, but the MLE exhibiting better characteristics. The question is then, why should we consider the MEEs at all? The reasons for considering MEEs are as follows:

1. By using the method of moment equations one can often easily determine consistent estimators having asymptotically normal distributions. These notions of consistency and asymptotic normality are defined and discussed in Chapter 7.
2. There are cases in which it is difficult to determine the MLEs, while the MEEs can be readily determined, and can be used as a first approximation in an iterative algorithm.
3. There are cases in which MLEs do not exist, while MEEs do exist.

5.8.2 General Theory of Estimating Functions

Both the MLE and the MME are special cases of a class of estimators called estimating functions estimator. A function $g(\mathbf{X};\theta)$, $\mathbf{X} \in \mathcal{X}^{(n)}$ and $\theta \in \Theta$, is called an **estimating function**, if the root $\hat{\theta}(\mathbf{X})$ of the equation

$$g(\mathbf{X}, \theta) = 0 \tag{5.8.2}$$

belongs to Θ; i.e., $\hat{\theta}(\mathbf{X})$ is an estimator of θ. Note that if θ is a k-dimensional vector then (5.8.2) is a system of k-independent equations in θ. In other words, $g(\mathbf{X}, \theta)$ is a k-dimensional vector function, i.e.,

$$\mathbf{g}(\mathbf{X}, \boldsymbol{\theta}) = (g_1(\mathbf{X}, \boldsymbol{\theta}), \ldots, g_k(\mathbf{X}, \boldsymbol{\theta}))'.$$

$\hat{\theta}(\mathbf{X})$ is the simultaneous solution of

$$\begin{aligned} g_1(\mathbf{X}, \boldsymbol{\theta}) &= 0, \\ g_2(\mathbf{X}, \boldsymbol{\theta}) &= 0, \\ &\vdots \\ g_k(\mathbf{X}, \boldsymbol{\theta}) &= 0. \end{aligned} \tag{5.8.3}$$

In the MEE case, $g_i(\mathbf{X}, \boldsymbol{\theta}) = M_i(\theta_1, \ldots, \theta_k) - m_i$ $(i = 1, \ldots, k)$. In the MLE case,

$$g_i(\mathbf{X}, \boldsymbol{\theta}) = \frac{\partial}{\partial \theta_i} \log f(\mathbf{X}; \boldsymbol{\theta}), \quad i = 1, \ldots, k.$$

In both cases, $E_\theta\{\mathbf{g}(\mathbf{X}, \boldsymbol{\theta})\} = 0$ for all θ, under the CR regularity conditions (see Theorem 5.2.2).

An estimating function $g(\mathbf{X}, \theta)$ is called **unbiased** if $E_\theta\{g(\mathbf{X}; \theta)\} = 0$ for all θ. The **information** in an estimating function $g(\mathbf{X}, \theta)$ is defined as

$$I_g(\theta) = \frac{\left[E_\theta\left\{\frac{\partial}{\partial\theta} g(\mathbf{X}, \theta)\right\}\right]^2}{E_\theta\{g^2(\mathbf{X}, \theta)\}}. \tag{5.8.4}$$

For example, if $g(\mathbf{X}, \theta)$ is the score function $S(\mathbf{X}, \theta)$, then under the regularity conditions (3.7.2), $E\left\{\frac{\partial}{\partial\theta} g(\mathbf{X}; \theta)\right\} = -I(\theta)$ and $E_\theta\{S^2(\mathbf{X}; \theta)\} = I(\theta)$, where $I(\theta)$ is the Fisher information function. A basic result of is that $I_g(\theta) \leq I(\theta)$ for all unbiased estimating functions.

The CR regularity conditions are now generalized for estimating functions. The regularity conditions for estimating functions are as follows:

(i) $\frac{\partial g(x, \theta)}{\partial\theta}$ exists for all θ, for almost all x (with probability one).

(ii) $\int g(x, \theta) dF(x, \theta)$ is differentiable with respect to θ under the integral sign, for all θ.

(iii) $E_\theta\left\{\frac{\partial}{\partial\theta} g(X, \theta)\right\} \neq 0$ and exists for all θ.

(iv) $E_\theta\{g^2(X, \theta)\} < \infty$ for all θ.

Let T be a sufficient statistic for a parametric family $\mathcal{F}$. Bhapkar (1972) proved that, for any unbiased estimating function g, if

$$g^*(T, \theta) = E\{g(X, \theta) \mid T\}$$

then $I_g(\theta) \leq I_{g^*}(\theta)$ for all θ with equality if and only if $g^* \in \mathcal{F}^T$. This is a generalization of the Blackwell–Rao Theorem to unbiased estimating functions. Under the regularity conditions, the score function $S(X, \theta) = \frac{\partial}{\partial\theta} \log f(X, \theta)$ depends on X only through the likelihood statistic $T(X)$, which is minimal sufficient. Thus, the score function is most informative among the unbiased estimating functions that satisfy the regularity conditions. If $\boldsymbol{\theta}$ is a vector parameter, then the information in g is

$$I_g(\boldsymbol{\theta}) = G^T(\boldsymbol{\theta}) \Sigma_g^{-1}(\boldsymbol{\theta}) G(\boldsymbol{\theta}), \tag{5.8.5}$$

where

$$G(\boldsymbol{\theta}) = \left(E_\theta\left\{\frac{\partial g_i(\mathbf{X}, \boldsymbol{\theta})}{\partial\theta_j}\right\}, \quad i, j = 1, \ldots, k\right) \tag{5.8.6}$$

and

$$\mathfrak{I}_g(\boldsymbol{\theta}) = E_\theta\{\mathbf{g}(\mathbf{X}, \boldsymbol{\theta})\mathbf{g}^T(\mathbf{X}, \theta)\}, \tag{5.8.7}$$

where $\mathbf{g}(\mathbf{X}, \boldsymbol{\theta}) = (g_1(\mathbf{X}, \boldsymbol{\theta}), \ldots, g_k(\mathbf{X}, \boldsymbol{\theta}))'$ is a vector of k estimating functions, for estimating the k components of $\boldsymbol{\theta}$.

We can show that $I(\theta) = I_g(\theta)$ is a nonnegative definite matrix, and $I(\theta)$ is the Fisher information matrix.

Various applications of the theory of estimating functions can be found in Godambe (1991).

5.9 PRETEST ESTIMATORS

Pretest estimators (PTEs) are estimators of the parameters, or functions of the parameters of a distribution, which combine testing of some hypothesis (es) and estimation for the purpose of reducing the MSE of the estimator. The idea of preliminary testing has been employed informally in statistical methodology in many different ways and forms. Statistical inference is often based on some model, which assumes a certain set of assumptions. If the model is correct, or adequately fits the empirical data, the statistician may approach the problem of estimating the parameters of interest in a certain manner. However, if the model is rejectable by the data the estimation of the parameter of interest may have to follow a different procedure. An estimation procedure that assumes one of two alternative forms, according to the result of a test of some hypothesis, is called a pretest estimation procedure.

PTEs have been studied in various estimation problems, in particular in various least-squares estimation problems for linear models. As we have seen in Section 4.6, if some of the parameters of a linear model can be assumed to be zero (or negligible), the LSE should be modified, according to formula (4.6.14). Accordingly, if $\hat{\boldsymbol{\beta}}$ denotes the unconstrained LSE of a full-rank model and $\boldsymbol{\beta}^*$ the constrained LSE (4.6.14), the PRE of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_A = \hat{\boldsymbol{\beta}} I\{\bar{A}\} + \boldsymbol{\beta}^* I\{A\}, \tag{5.9.1}$$

where A denotes the acceptance set of the hypothesis $H_0 : \beta_{r+1} = \beta_{r+2} = \ldots = \beta_p = 0$; and $\bar{A}$ the complement of A. An extensive study of PREs for linear models, of the form (5.8.5), is presented in the book of Judge and Bock (1978). The reader is referred also to the review paper of Billah and Saleh (1998).

5.10 ROBUST ESTIMATION OF THE LOCATION AND SCALE PARAMETERS OF SYMMETRIC DISTRIBUTIONS

In this section, we provide some new developments concerning the estimation of the location parameter, μ, and the scale parameter, σ, in a parametric family, $\mathcal{F}$,

whose p.d.f.s are of the form $f(x;\mu,\sigma)=\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$, and $f(-x)=f(x)$ for all $-\infty<x<\infty$. We have seen in various examples before that an estimator of μ, or of σ, which has small MSE for one family may not be as good for another. We provide below some variance comparisons of the sample mean, $\bar{X}$, and the sample median, M_e, for the following families: normal, mixture of normal and rectangular, $t[\nu]$, Laplace and Cauchy. The mixtures of normal and rectangular distributions will be denoted by $(1-\alpha)N+\alpha R(-3\sigma,3\sigma)$. Such a family of mixtures has the standard density function

$$f(x)=\frac{1-\alpha}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}x^2\right\}+\frac{\alpha}{6\sigma}I\{-3\sigma<x<3\sigma\},\quad -\infty<x<\infty.$$

The $t[\nu]$ distributions have a standard p.d.f. as given in (2.13.5). The asymptotic (large sample) variance of the sample median, M_e, is given by the formula (7.9.3)

$$AV\{M_e\}=\frac{\sigma^2}{4f^2(0)n}, \tag{5.10.1}$$

provided $f(0)>0$, and $f(x)$ is continuous at $x=0$.

In Table 5.1, we provide the asymptotic variances of $\bar{X}$ and M_e and their ratio $E=AV\{\bar{X}\}/AV\{M_e\}$, for the families mentioned above. We see that the sample mean $\bar{X}$ which is a very good estimator of the location parameter, μ, when $\mathcal{F}$ is the family of normal distributions loses its efficiency when $\mathcal{F}$ deviates from normality. The reason is that the sample mean is very sensitive to deviations in the sample of the extreme values. The sample mean performs badly when the sample is drawn from a distribution having heavy tails (relatively high probabilities of large deviations from the median of the distribution). This phenomenon becomes very pronounced in the case of the Cauchy family. One can verify (Fisz, 1963, p. 156) that if $X_1,\ldots,X_n$ are i.i.d. random variables having a common Cauchy distribution than the sample mean $\bar{X}$ has the same Cauchy distribution, irrespective of the sample size. Furthermore, the Cauchy distribution does not have moments, or we can say that the variance of $\bar{X}$ is infinite. In order to avoid such possibly severe consequences due to the use of $\bar{X}$ as an estimator of μ, when the statistician specifies the model erroneously, several

Table 5.1 Asymptotic Variances of $\bar{X}$ and M_e

Family	$\bar{X}$	M_e	E
Normal	σ^2/n	$\pi\sigma^2/2n$	0.6366
$0.9N+0.1R(-3\sigma,3\sigma)$	$1.2\sigma^2/n$	$1.77\sigma^2/n$	0.6776
$0.5N+0.5R(-3\sigma,3\sigma)$	$1.5\sigma^2/n$	$3.1258\sigma^2/n$	0.4799
$t[4]$	$2\sigma^2/n$	$16\sigma^2/9n$	1.125
Laplace	$2\sigma^2/n$	σ^2/n	2.000
Cauchy	–	$\sigma^2\pi^2/4$	∞

types of less sensitive estimators of μ and σ were developed. These estimators are called **robust** in the sense that their performance is similar, in terms of the sampling variances and other characteristics, over a wide range of families of distributions. We provide now a few such robust estimators of the location parameter:

1. α *-Trimmed Means:* The sample is ordered to obtain $X_{(1)} \leq \ldots \leq X_{(n)}$. A proportion α of the smallest and largest values are removed and the mean of the remaining $(1-\alpha)n$ of the values is determined. If $[n\alpha]$ denotes the largest integer not exceeding $n\alpha$ and if $p = 1 + [n\alpha] - n\alpha$ then the α-trimmed mean is

$$\hat{\mu}_\alpha = \frac{pX_{([n\alpha]+1)} + X_{([n\alpha]+2)} + \cdots + pX_{(n-[\alpha n])}}{n(1-2\alpha)}. \tag{5.10.2}$$

   The median, M_e is a special case, when $\alpha \to 0.5$.

2. *Linear Combinations of Selected Order Statistics:* This is a class of estimates which are linear combinations, with some specified weights of some selected order statistics. Gastwirth (1977) suggested the estimator

$$LG = .3X_{([\frac{n}{3}]+1])} + .4M_e + .3X_{(n-[\frac{n}{3}])}. \tag{5.10.3}$$

   Another such estimator is called the **trimean** and is given by

$$TRM = 0.25X_{([\frac{n}{4}]+1)} + 0.5M_e + 0.25X_{(n-[\frac{n}{4}])}.$$

3. M *-Estimates:* The MLE estimates of μ and σ are the simultaneous solutions of the equations

$$\text{(I)} \quad \sum_{i=1}^{n} \frac{f'\left(\frac{X_i-\mu}{\sigma}\right)}{f\left(\frac{X_i-\mu}{\sigma}\right)} = 0 \tag{5.10.4}$$

   and

$$\text{(II)} \quad \sum_{i=1}^{n} \left[\left(\frac{X_i-\mu}{\sigma}\right) \frac{f'\left(\frac{X_i-\mu}{\sigma}\right)}{f\left(\frac{X_i-\mu}{\sigma}\right)} - 1 \right] = 0.$$

In analogy to the MLE solution and, in order to avoid strong dependence on a particular form of $f(x)$, a general class of M-estimators is defined as the simultaneous solution of

$$\sum_{i=1}^{n} \psi\left(\frac{X_i-\mu}{\sigma}\right) = 0 \tag{5.10.5}$$

and

$$\sum_{i=1}^{n} \chi\left(\frac{X_i - \mu}{\sigma}\right) = 0$$

for suitably chosen $\psi(\cdot)$ and $\chi(\cdot)$ functions. Huber (1964) proposed the M-estimators for which

$$\psi_k(z) = \begin{cases} -k, & z \le -k, \\ z, & -k < z < k, \\ k, & z \ge k, \end{cases} \tag{5.10.6}$$

and

$$\chi(z) = \psi_k^2(z) - \beta(k), \tag{5.10.7}$$

where

$$\beta(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \psi_k^2(z) e^{-\frac{1}{2}z^2} dz.$$

The determination of Huber's M-estimators requires numerical iterative solutions. It is customary to start with the initial solution of $\mu = M_e$ and $\sigma = (Q_3 - Q_1)/1.35$, where $Q_3 - Q_1$ is the interquartile range, or $X_{(n-[\frac{n}{4}])} - X_{([\frac{n}{4}]+1)}$. Values of k are usually taken in the interval $[1, 2]$.

Other M-estimators were introduced by considering a different kind of $\psi(\cdot)$ function. Having estimated the value of γ by $\hat{\gamma}$, use the estimator

$$\hat{\mu}(\hat{\gamma}) = \begin{cases} \text{outer-mean}, & \text{if } \hat{\gamma} < 2, \\ \bar{X}, & \text{if } 2 \le \hat{\gamma} \le 4, \\ \hat{\mu}_{.125}, & \text{if } 4 < \hat{\gamma} \le 4.5, \\ LG, & \text{if } \hat{\gamma} < 4.5, \end{cases}$$

where the "outer-mean" is the mean of the extreme values in the sample. The reader is referred to the Princeton Study (Andrews et al., 1972) for a comprehensive examination of these and many other robust estimators of the location parameter. Another important article on the subject is that of Huber (1964, 1967).

Robust estimators of the scale parameter, σ, are not as well developed as those of the location parameter. The estimators that are used are

$$\begin{aligned} \hat{\sigma}_1 &= (Q_3 - Q_1)/1.35, \\ \hat{\sigma}_2 &= \text{Median } (|X_{(i)} - M_e|, \quad i = 1, \ldots, n)/0.6754, \\ \hat{\sigma}_3 &= \frac{\sqrt{2}}{n} \sum_{i=1}^{n} |X_i - M_e|. \end{aligned}$$

Further developments have been recently attained in the area of robust estimation of regression coefficients in multiple regression problems.

PART II: EXAMPLES

Example 5.1. In the production of concrete, it is required that the proportion of concrete cubes (of specified dimensions) having compressive strength not smaller than ξ_0 be at least 0.95. In other words, if X is a random variable representing the compressive strength of a concrete cube, we require that $P\{X \geq \xi_0\} = 0.95$. This probability is a numerical characteristic of the distribution of X. Let $X_1, \ldots, X_n$ be a sample of i.i.d. random variables representing the compressive strength of n randomly chosen cubes from the production process under consideration. If we do not wish to subject the estimation of $p_0 = P\{X \geq \xi_0\}$ to strong assumptions concerning the distribution of X we can estimate this probability by the proportion of cubes in the sample whose strength is at least ξ_0; i.e.,

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} I\{X_i \geq \xi_0\}.$$

We note that $n\hat{p}$ has the binomial distribution $B(n, p_0)$. Thus, properties of the estimator $\hat{p}$ can be deduced from this binomial distribution.

A commonly accepted model for the compressive strength is the family of log-normal distributions. If we are willing to commit the estimation procedure to this model we can obtain estimators of p_0 which are more efficient than $\hat{p}$, provided the model is correct. Let $Y_i = \log X_i$, $i = 1, \ldots,$ and let $\bar{Y}_n = \frac{1}{n}\sum_i Y_i$, $S_n^2 = \sum_i (Y_i - \bar{Y}_n)^2/(n-1)$. Let $\eta_0 = \log \xi_0$. Then, an estimator of p_0 can be

$$\tilde{p} = \Phi\left(\frac{\bar{Y}_n - \eta_0}{S_n}\right),$$

where $\Phi(u)$ is the standard normal c.d.f. Note that $\bar{Y}_n$ and S_n are the sample statistics that are substituted to estimate the unknown parameters (ξ, σ). Moreover, $(\bar{Y}_n, S_n)$ is a m.s.s. for the family of log-normal distributions. The estimator we have exhibited depends on the sample values only through the m.s.s. As will be shown later the estimator $\tilde{p}$ has certain optimal properties in large samples, and even in small samples it is a reasonable estimator to use, provided the statistical model used is adequate for the real phenomenon at hand. ■

Example 5.2. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a rectangular distribution $R(0, \theta)$, $0 < \theta < \infty$. Suppose that the characteristic of interest is the expectation $\mu = \theta/2$. The unbiased estimator $\tilde{\mu} = \bar{X}_n$ has a variance

$$V_\theta\{\bar{X}_n\} = \frac{\theta^2}{12n}.$$

On the other hand, consider the m.s.s. $X_{(n)} = \max_{1 \le i \le n} \{X_i\}$. The expected value of $X_{(n)}$ is

$$E_\theta\{X_{(n)}\} = \frac{n}{\theta^n} \int_0^\theta t^n dt = \frac{n}{n+1}\theta.$$

Hence, the estimator $\hat{\mu} = \frac{n+1}{2n} X_{(n)}$ is also an unbiased estimator of μ. The variance of $\hat{\mu}$ is

$$V_\theta\{\hat{\mu}\} = \frac{\theta^2}{4n(n+2)}.$$

Thus, $V_\theta\{\hat{\mu}\} < V_\theta\{\bar{X}_n\}$ for all $n \ge 2$, and $\hat{\mu}$ is a better estimator than $\bar{X}_n$. We note that $\hat{\mu}$ depends on the m.s.s. $X_{(n)}$, while $\bar{X}_n$ is not a sufficient statistic. This is the main reason for the superiority of $\hat{\mu}$ over $\bar{X}_n$. The theoretical justification is provided in the Rao–Blackwell Theorem. ■

Example 5.3. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common normal distribution, i.e., $\mathcal{F} = \{N(\xi, \sigma^2); -\infty < \xi < \infty, 0 < \sigma < \infty\}$. Both the mean ξ and the variance σ^2 are unknown. We wish to estimate unbiasedly the probability $g(\xi, \sigma) = P_{\xi,\sigma}\{X \le \xi_0\}$. Without loss of generality, assume that $\xi_0 = 0$, which implies that $g(\xi, \sigma) = \Phi(\xi/\sigma)$. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ be the sample mean and variance. $(\bar{X}, S^2)$ is a complete sufficient statistic. According to the Rao–Blackwell Theorem, there exists an essentially unique unbiased estimator of $\Phi(\xi/\sigma)$ that is a function of the complete sufficient statistic. We prove now that this UMVU estimator is

$$\hat{g}(\bar{X}, S) = \begin{cases} 0, & \text{if } w(\bar{X}, S) \le 0, \\ I_{w(\bar{X},S)}(\frac{n}{2} - 1, \frac{n}{2} - 1), & \text{if } 0 < w(\bar{X}, S) < 1, \\ 1, & \text{if } w(\bar{X}, S) \ge 1, \end{cases}$$

where

$$w(\bar{X}, S) = \frac{1}{2}\left[\frac{\bar{X}\sqrt{n}}{\sqrt{(n-1)}S} + 1\right].$$

The proof is based on the following result (Ellison, 1964). If U and V are independent random variables $U \sim \beta(\frac{\nu-1}{2}, \frac{\nu-1}{2})$ and $V \sim (\chi^2[\nu])^{1/2}$ then $(2U-1)V \sim N(0,1)$. Let $\nu = n-1$ and $V = \sqrt{n-1}\, S/\sigma$. Accordingly

$$\hat{g}(\bar{X}, S) = P\left\{\beta\left(\frac{n}{2} - 1, \frac{n}{2} - 1\right) \le w(\bar{X}, S) \mid \bar{X}, S)\right\},$$

where $\beta(\frac{n}{2} - 1, \frac{n}{2} - 1)$ is independent of $(\bar{X}, S)$. Thus, by substituting in the expression for $w(\bar{X}, S)$, we obtain

$$E_{\xi,\sigma}\{\hat{g}(\bar{X}, S)\} = P\left\{\sigma\left(2\beta\left(\frac{n}{2} - 1, \frac{n}{2} - 1\right) - 1\right) V \leq \frac{\sqrt{n}}{\sqrt{n-1}}\bar{X}\right\}$$
$$= P\left\{\sigma N_1(0, 1) - \frac{\sigma}{\sqrt{n-1}} N_2(0, 1) \leq \frac{\sqrt{n}}{\sqrt{n-1}}\xi\right\},$$

with $N_1(0, 1)$ and $N_2(0, 1)$ independent standard normal random variables. Thus,

$$E_{\xi,\sigma}\{\hat{g}(\bar{X}, S)\} = P\left\{N(0, 1) \leq \frac{\xi}{\sigma}\right\}$$
$$= \Phi(\xi/\sigma), \quad \text{for all } (\xi, \sigma).$$

We provide an additional example that illustrates the Rao–Blackwellization method. ■

Example 5.4. Let $X_1, \ldots, X_n$ be i.i.d. random variables, having a common Poisson distribution, $P(\lambda)$, $0 < \lambda < \infty$. We wish to estimate unbiasedly the Poisson probability $p(k; \lambda) = e^{-\lambda}\lambda^k/k!$ An unbiased estimator of $p(k; \lambda)$ based on one observation is

$$\tilde{p}(k; X_1) = I\{X_1 = k\}, \quad k = 0, 1, \ldots.$$

Obviously, this estimator is inefficient. According to the Rao–Blackwell Theorem the MVUE of $p(k; \lambda)$ is

$$\hat{p}(k; T_n) = E\{I\{X_1 = k\} \mid T_n\}$$
$$= P[X_1 = k \mid T_n],$$

where $T_n = \sum X_i$ is the complete sufficient statistic. If $T_n > 0$ the conditional distribution of X_1, given T_n is the binomial $B\left(T_n, \frac{1}{n}\right)$. Accordingly, the MVUE of $p(k; \lambda)$ is

$$\hat{p}(k; T_n) = \begin{cases} I\{k = 0\}, & \text{if } T_n = 0, \\ b\left(k \mid T_n, \frac{1}{n}\right), & \text{if } T_n > 0, \end{cases}$$

where $b\left(k \mid T_n, \frac{1}{n}\right)$ is the p.d.f. of the Binomial distribution $B\left(T_n, \frac{1}{n}\right)$. ■

Example 5.5. We have seen in Section 3.6 that if the m.s.s. $S(\mathbf{X})$ is incomplete, there is reason to find an ancillary statistic $A(\mathbf{X})$ and base the inference on the conditional

distribution of $S(\mathbf{X})$, given $A(\mathbf{X})$. We illustrate in the following example a case where such an analysis does not improve.

Let $X_1, \ldots, X_n$ be i.i.d. random variables having a rectangular distribution in

$$\mathcal{F} = \{R(\theta - 1, \theta + 1); -\infty < \theta < \infty\}.$$

A likelihood function for θ is

$$L(\theta; X) = \frac{1}{2^n} I\{X_{(n)} - 1 < \theta < X_{(1)} + 1\},$$

where $X_{(1)} < \cdots < X_{(n)}$ is the order statistic. A m.s.s. is $(X_{(1)}, X_{(n)})$. This statistic, however, is incomplete. Indeed, $E_\theta \left\{ X_{(n)} - X_{(1)} - 2\frac{n-2}{n+1} \right\} = 0$, but $P_\theta \left\{ X_{(n)} - X_{(1)} = 2\frac{n-1}{n+1} \right\} = 0$ for each θ.

Writing $R(\theta - 1, \theta + 1) \sim \theta - 1 + 2R(0, 1)$ we have $X_{(1)} \sim \theta - 1 + 2U_{(1)}$ and $X_{(n)} \sim \theta - 1 + 2U_{(n)}$, where $U_{(1)}$ and $U_{(n)}$ are the order statistics from $R(0, 1)$. Moreover, $E\{U_{(1)}\} = \frac{1}{n+1}$ and $E\{U_{(n)}\} = \frac{n}{n+1}$. It follows immediately that $\hat{\theta} = \frac{1}{2}(X_{(1)} + X_{(n)})$ is unbiased. By the Blackwell–Rao Theorem it cannot be improved by conditioning on the sufficient statistic.

We develop now the conditional distribution of $\hat{\theta}$, given the ancillary statistic $W = X_{(n)} - X_{(1)}$. The p.d.f. of W is

$$f_W(r) = \frac{n(n-1)}{2^n} r^{n-2}(2 - r), \quad 0 \le r \le 2.$$

The transformation $(X_{(1)}, X_{(n)}) \to (\hat{\theta}, W)$ is one to one. The joint p.d.f. of $(\hat{\theta}, W)$ is

$$f_{\hat{\theta}, W}(t, r; \theta) = \frac{n(n-1)}{2^n} r^{n-2} I\left\{ \theta + \frac{r}{2} - 1 \le t \le \theta - \frac{r}{2} + 1 \right\}.$$

Accordingly,

$$p_{\hat{\theta}|W}(t \mid r) = \frac{1}{2 - r} I\left\{ \theta + \frac{r}{2} - 1 \le t \le \theta - \frac{r}{2} + 1 \right\}.$$

That is, $\hat{\theta} \mid W \sim R\left(\theta + \frac{W}{2} - 1, \theta - \frac{W}{2} + 1 \right)$. Thus,

$$E\{\hat{\theta} \mid W\} = \theta, \quad \text{for all } -\infty < \theta < \infty,$$

and

$$V\{\hat{\theta} \mid W\} = \frac{(2 - W)^2}{12},$$

We have seen already that $\hat{\theta}$ is an unbiased estimator. From the law of total variance, we get

$$V_\theta\{\hat{\theta}\} = \frac{2}{(n+1)(n+2)}$$

for all $-\infty < \theta < \infty$. Thus, the variance of $\hat{\theta}$ was obtained from this conditional analysis. One can obtain the same result by computing $V\{U_{(1)} + U_{(n)}\}$. ■

Example 5.6. Consider the MVUE of the Poisson probabilities $p(k;\lambda)$, derived in Example 5.4. We derive here the Cramér–Rao lower bound for the variance of this estimator. We first note that the Fisher information for a sample of n i.i.d. Poisson random variables is $I_n(\lambda) = n/\lambda$. Furthermore, differentiating $p(k;\lambda)$ with respect to λ we obtain that $\frac{\partial}{\partial\lambda} p(k;\lambda) = -(p(k;\lambda) - p(k-1;\lambda))$, where $p(-1;\lambda) \equiv 0$. If $\hat{p}(k;T_n)$ is the MVUE of $p(k;\lambda)$, then according to the Cramér–Rao inequality

$$\mathrm{Var}_\lambda\{\hat{p}(k;T_n)\} > \begin{cases} \frac{\lambda}{n} p^2(k-1;\lambda)\left(\frac{\lambda}{k} - 1\right)^2, & k \geq 1, \\ \frac{\lambda}{n} e^{-2\lambda}, & k = 0. \end{cases}$$

Strict inequality holds for all values of λ, $0 < \lambda < \infty$, since the distribution of $\hat{p}(k;T_n)$ is not of the exponential type, although the distribution of T_n is Poisson. The Poisson family satisfies all the conditions of Joshi (1976) and therefore since the distribution of $\hat{p}(k;T_n)$ is not of the exponential type, the inequality is strict. Note that $V\{\hat{p}(k;T_n\} = E\{(b(k;T_n\tfrac{1}{n}))^2\} - p^2(k;\lambda)$. We can compute this variance numerically. ■

Example 5.7. Consider again the estimation problem of Examples 5.4 and 5.5, with $k = 0$. The MVUE of $\omega(\lambda) = e^{-\lambda}$ is $\hat{\omega}(T_n) = \left(1 - \frac{1}{n}\right)^{T_n}$. The variance of $\hat{\omega}(T_n)$ can be obtained by considering the probability generating function of $T_n \sim P(n\lambda)$ at $t = \left(1 - \frac{1}{n}\right)$. We thus obtain

$$\mathrm{Var}_\lambda\{\hat{\omega}(T_n)\} = e^{-2\lambda}(e^{\lambda/n} - 1).$$

Since $\omega(\lambda)$ is an analytic function, we can bound the variance of $\hat{\omega}(T_n)$ from below by using BLB of order $k = 2$ (see (5.2.15)). We obtain, $V_{11} = \frac{n}{\lambda}$, $V_{12} = 0$, $V_{22} = \frac{2n^2}{\lambda^2}$. Hence, the lower bound for $k = 2$ is

$$L_2(\lambda) = \frac{\lambda}{n} e^{-2\lambda}\left(1 + \frac{\lambda}{2n}\right), \quad 0 < \lambda < \infty.$$

This lower bound is larger than the Cramér–Rao lower bound for all $0 < \lambda < \infty$. ■

Example 5.8. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. vectors having a common bivariate normal distribution $N(\mathbf{0}, \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})$, $-1 \leq \rho \leq 1, 0 < \sigma^2 < \infty$. The complete sufficient statistic for this family of bivariate normal distributions is $T_1(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n}(X_i^2 + Y_i^2)$ and $T_2(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n} X_i Y_i$. We wish to estimate the coefficient of correlation ρ.

An unbiased estimator of ρ is given by $\hat{\rho} = \sum_{i=1}^{n} X_i Y_i / \sum_{i=1}^{n} X_i^2$. Indeed

$$E\{\hat{\rho} \mid \mathbf{X}\} = \frac{1}{\sum_i X_i^2} \sum_{i=1}^{n} X_i E\{Y_i \mid \mathbf{X}\}.$$

But $E\{Y_i \mid \mathbf{X}\} = \rho X_i$ for all $i = 1, \ldots, n$. Hence, $E\{\hat{\rho} \mid \mathbf{X}\} = \rho$ w.p.1. The unbiased estimator is, however, **not** an MVUE. Indeed, $\hat{\rho}$ is **not a** function of $(T_1(\mathbf{X}, \mathbf{Y}), T_2(\mathbf{X}, \mathbf{Y}))$. The MVUE can be obtained, according to the Rao–Blackwell Theorem by determining the conditional expectation $E\{\hat{\rho} \mid T_1, T_2\}$.

The variance of $\hat{\rho}$ is

$$V\{\hat{\rho}\} = \frac{1 - \rho^2}{n - 2}.$$

The Fisher information matrix in the present case is

$$I(\sigma^2, \rho) = n \begin{pmatrix} \frac{1}{\sigma^4} & -\frac{\rho}{\sigma^2(1-\rho^2)} \\ \frac{-\rho}{\sigma^2(1-\rho^2)} & \frac{1+\rho^2}{(1-\rho^2)^2} \end{pmatrix}.$$

The inverse of the Fisher information matrix is

$$(I(\sigma^2, \rho))^{-1} = \frac{1}{n} \begin{pmatrix} \sigma^4(1+\rho^2) & \sigma^2\rho(1-\rho^2) \\ \sigma^2\rho(1-\rho^2) & (1-\rho^2)^2 \end{pmatrix}.$$

The lower bound on the variances of unbiased estimators of ρ is, therefore, $(1 - \rho^2)^2/n$. The ratio of the lower bound of the variance of $\hat{\rho}$ to the actual variance is $\frac{(1-\rho^2)(n-2)}{n} \approx 1 - \rho^2$ for large n. Thus, $\hat{\rho}$ is a good unbiased estimator only if ρ^2 is close to zero. ■

Example 5.9.

A. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common location-parameter exponential distribution with a p.d.f.

$$f(x;\theta) = I\{x \geq \theta\} \exp\{-(x-\theta)\}, \quad -\infty < \theta < \infty.$$

The sample minimum $X_{(1)}$ is a complete sufficient statistic. $X_{(1)}$ is distributed like $\theta + G(n, 1)$. Hence, $E\{X_{(1)}\} = \theta + \frac{1}{n}$ and the MVUE of θ is $\hat{\theta}(X_{(1)}) = X_{(1)} - \frac{1}{n}$. The variance of this estimator is

$$\mathrm{Var}_\theta\{\hat{\theta}(X_{(1)})\} = \frac{1}{n^2}, \quad \text{for all } -\infty < \theta < \infty.$$

In the present case, the Fisher information $I(\theta)$ does not exist. We derive now the modified Chapman–Robbins lower bound for the variance of an unbiased estimator of θ. Notice first that $W_\phi(X_{(1)};\theta) = I\{X_{(1)} \geq \phi\} e^{n(\phi-\theta)}$, where $T = X_{(1)}$, for all $\phi \geq \theta$. It is then easy to prove that

$$A(\theta, \phi) = \exp\{n(\phi-\theta)\}^{-1}, \quad \phi > \theta.$$

Accordingly,

$$\mathrm{Var}_\theta\{\hat{\theta}(X_{(1)})\} \geq \sup_{\phi>\theta} \frac{(\phi-\theta)^2}{\exp\{n(\phi-\theta)\} - 1}.$$

The function $x^2/(e^{nx} - 1)$ assumes a unique maximum over $(0, \infty)$ at the root of the equation $e^{nx}(2 - nx) = 2$. This root is approximately $x_0 = \frac{1.592}{n}$. This approximation yields

$$\mathrm{Var}_\theta\{\hat{\theta}(X_{(1)})\} \geq \frac{0.6476}{n^2}.$$

B. Consider the case of a random sample from $R(0, \theta)$, $0 < \theta < \infty$. As shown in Example 3.11 A, $I_n(\theta) = \frac{n^2}{\theta^2}$. The UMVU estimator of θ is $\hat{\theta}_n = \frac{n+1}{n} X_{(n)}$. The variance of $\hat{\theta}_n$ is $V_\theta\{\hat{\theta}_n\} = \frac{\theta^2}{n(n+2)}$. Thus, in this nonregular case

$$V_\theta\{\hat{\theta}_n\} < \frac{1}{I_n(\theta)} \quad \text{for all } 0 < \theta < \infty.$$

However,

$$\lim_{n\to\infty} V_\theta\{\hat{\theta}_n\} I_n(\theta) = 1 \quad \text{for all } \theta.$$

■

Example 5.10. Let $X_1, \ldots, X_n$ be i.i.d. random variables having the normal distribution $N(\theta, \sigma^2)$ and $Y_1, \ldots, Y_n$ i.i.d. random variables having the normal distribution $N(\gamma\theta^2, \sigma^2)$, where $-\infty < \theta, \ \gamma < \infty$, and $0 < \sigma < \infty$. The vector $\mathbf{X} = (X_1, \ldots, X_n)'$ is independent of $\mathbf{Y} = (Y_1, \ldots, Y_n)'$. A m.s.s. is $(\bar{X}_n, \bar{Y}_n, Q_n)$, where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$, and $Q_n = \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2$. The Fisher information matrix can be obtained from the likelihood function

$$L(\theta, \gamma, \sigma^2 \mid \bar{X}_n, \bar{Y}_n, Q_n) = \frac{1}{(\sigma^2)^n} \exp\left\{-\frac{n}{2\sigma^2}\left[(\bar{X}_n - \theta)^2 + (\bar{Y}_n - \gamma\theta^2)^2 + \frac{Q_n}{n}\right]\right\}.$$

The covariance matrix of the score functions is

$$nI(\theta, \gamma, \sigma^2) = \frac{n}{\sigma^2}\begin{pmatrix} 1 + 4\gamma^2\theta^2 & 2\gamma\theta^3 & 0 \\ 2\gamma\theta^3 & \theta^4 & 0 \\ 0 & 0 & \dfrac{1}{\sigma^2} \end{pmatrix}.$$

Thus,

$$\mathcal{I}_n(\theta, \gamma, \sigma^2) = |nI(\theta, \gamma, \sigma^2)| = \frac{n^3\theta^4}{\sigma^8}.$$

Consider the reparametrization $g_1(\theta, \gamma, \sigma) = \theta$, $g_2(\theta, \gamma, \sigma^2) = \gamma\theta^2$ and $g_3(\theta, \gamma, \sigma) = \sigma^2$. The UMVU estimator is $\hat{\mathbf{g}} = (\bar{X}_n, \bar{Y}_n, Q_n/2(n-1))$. The variance covariance matrix of $\hat{\mathbf{g}}$ is

$$\Sigma(\hat{g}) = \frac{\sigma^2}{n}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sigma^2\dfrac{n}{n-2} \end{pmatrix},$$

and

$$D(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 2\gamma\theta & \theta^2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus, $\mathcal{I}_g = \dfrac{|D(\theta)D'(\theta)|}{|\mathbb{\Sigma}(\hat{\mathbf{g}})|} = \dfrac{\theta^4 n^2(n-2)}{\sigma^8}$. The efficiency coefficient is $\mathcal{E}_\theta(\hat{\mathbf{g}}) = \dfrac{n-2}{n} = 1 - \dfrac{2}{n}$. ■

Example 5.11. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of n i.i.d. vectors having a joint bivariate normal distribution

$$N\left(\begin{bmatrix}\mu\\ \mu\end{bmatrix}, \begin{pmatrix}\sigma^2 & \rho\sigma\tau\\ \rho\sigma\tau & \tau^2\end{pmatrix}\right),$$

where $-\infty < \mu < \infty, 0 < \tau < \infty, 0 < \sigma < \infty$, and $-1 < \rho < 1$. Assume that σ^2, τ^2, and ρ are known. The problem is to estimate the common mean μ. We develop the formula of the BLUE of μ. In the present case,

$$\mathbb{\Sigma} = \frac{1}{n}\begin{pmatrix}\sigma^2 & \rho\sigma\tau\\ \rho\sigma\tau & \tau^2\end{pmatrix}$$

and

$$\mathbb{\Sigma}^{-1} = \frac{n}{\sigma^2\tau^2(1-\rho^2)}\begin{pmatrix}\tau^2 & -\rho\sigma\tau\\ -\rho\sigma\tau & \sigma^2\end{pmatrix}.$$

The BLUE of the common mean μ is according to (5.3.3)

$$\hat{\mu} = \frac{\omega\bar{X}_n + \bar{Y}_n}{\omega + 1},$$

where $\bar{X}_n$ and $\bar{Y}_n$ are the sample means and

$$\omega = \frac{\tau^2 - \rho\sigma\tau}{\sigma^2 - \rho\sigma\tau}, \quad \text{provided } \rho\tau \neq \sigma.$$

Since $\mathbb{\Sigma}$ is known, $\hat{\mu}$ is UMVU estimator. ■

Example 5.12. Let $X_1, \ldots, X_n$ be i.i.d. Weibull variables, i.e., $X \sim G^{1/\beta}(\lambda, 1)$, where $0 < \lambda, \beta < \infty$. Both λ and β are unknown. The m.s.s. is $(X_{(1)}, \ldots, X_{(n)})$. Let $Y_i = \log X_i$, $i = 1, \ldots, n$, and $Y_{(i)} = \log X_{(i)}$. Obviously, $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$. We obtain the linear model

$$Y_{(i)} = \mu + \sigma \log G_{(i)}, \quad i = 1, \ldots, n,$$

where $\mu = \frac{1}{\beta}\log\lambda$ and $\sigma = \frac{1}{\beta}$; $G_{(i)}$ is the ith order statistic of n i.i.d. variables distributed like $G(1, 1)$. BLUEs of μ and σ are given by (5.4.16), where $\boldsymbol{\alpha}$ is the vector of $E\{\log G_{(i)}\}$ and V is the covariance matrix of $\log G_{(i)}$.

The p.d.f. of $G_{(i)}$ is

$$f_{(i)}(x) = \frac{n!}{(i-1)!(n-i)!}(1-e^{-x})^{i-1}e^{-x(n-i+1)},$$

$0 \le x \le \infty$. Hence,

$$\alpha_i = E\{\log G_{(i)}\} = \binom{n}{i}\sum_{j=0}^{i-1}(-1)^j \frac{i!}{j!(i-1-j)!}\int_{-\infty}^{\infty} ue^{-u-(n-i+1+j)e^{-u}}du.$$

The integral on the RHS is proportional to the expected value of the extreme value distribution. Thus,

$$\alpha_i = \binom{n}{i}\sum_{j=0}^{i-1}(-1)^j \frac{i!}{j!(i-1-j)!}\cdot\frac{\log(n-i+1+j)+\gamma}{n-i+1+j},$$

where $\gamma = 0.577216\ldots$ is the Euler constant. The values of α_i can be determined numerically for any n and $i = 1, \ldots, n$. Similar calculations yield formulae for the elements of the covariance matrix V. The point is that, from the obtained formulae of α_i and V_{ij}, we can determine the estimates only numerically. Moreover, the matrix V is of order $n \times n$. Thus, if the sample involves a few hundreds observation the numerical inversion of V becomes difficult, if at all possible. ■

Example 5.13. Consider the multiple regression problem with $p = 3$, $\sigma^2 = 1$, for which the normal equations are

$$\begin{pmatrix} 1.07 & 0.27 & 0.66 \\ 0.27 & 1.07 & 0.66 \\ 0.66 & 0.66 & 0.68 \end{pmatrix}\begin{pmatrix}\beta_1 \\ \beta_2 \\ \beta_3\end{pmatrix} = \begin{pmatrix} 1.05 \\ -0.06 \\ 0.83 \end{pmatrix}.$$

By employing the orthogonal (Helmert) transformation

$$H = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} \end{pmatrix},$$

we obtain that

$$H(A'A)H' = \begin{pmatrix} 2.0 & 0 & 0 \\ 0 & 0.8 & 0 \\ 0 & 0 & 0.02 \end{pmatrix}.$$

That is, the eigenvalues of $A'A$ are $\lambda_1 = 2$, $\lambda_2 = 0.8$ and $\lambda_3 = 0.02$. The LSEs of $\boldsymbol{\beta}$ are $\hat{\beta}_1 = -4.58625$, $\hat{\beta}_2 = -5.97375$, and $\hat{\beta}_3 = 11.47$. The variance covariance matrix of the LSE is

$$\mathfrak{V}(\hat{\boldsymbol{\beta}}) = (A'A)^{-1} = \begin{pmatrix} 9.125 & 7.875 & -16.5 \\ 7.875 & 9.125 & -16.5 \\ -16.5 & -16.5 & 33.5 \end{pmatrix},$$

having a trace $E\{L^2(0)\} = 51.75 = \sum \lambda_i^{-1}$. In order to illustrate numerically the effect of the ridge regression, assume that the true value of $\boldsymbol{\beta}$ is $(1.5, -6.5, 0.5)$. Let $\gamma = H\boldsymbol{\beta}$. The numerical value of γ is $(-2.59809, 5.65685, -2.44949)$. According to (5.4.4), we can write the sum of the MSEs of the components of $\hat{\boldsymbol{\beta}}(k)$ by

$$E\{L^2(k)\} = \sum_{i=1}^{3} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^{3} \frac{\gamma_i^2}{(\lambda_i + k)^2}.$$

The estimate of k^0 is $\hat{k} = 0.249$. In the following table, we provide some numerical results.

k	0	0.05	0.075	0.10	0.125	0.15
$\hat{\beta}_1(k)$	−4.58625	0.64636	−0.24878	−0.02500	0.11538	0.209518
$\hat{\beta}_2(k)$	−5.97375	−1.95224	−1.51735	−1.25833	−1.08462	−0.958900
$\hat{\beta}_3(k)$	11.47000	3.48641	2.64325	2.15000	1.82572	1.59589
$E\{L^2(k)\}$	51.75000	8.84077	7.70901	7.40709	7.39584	7.51305

We see that the minimal $E\{L^2(k)\}$ is minimized for k^0 around 0.125. At this value of k, $\hat{\boldsymbol{\beta}}(k)$ is substantially different from the LSE $\hat{\boldsymbol{\beta}}(0)$. ■

Example 5.14.

A. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a rectangular distribution $R(0, \theta)$, $0 < \theta < \infty$. A m.s.s. is the sample maximum $X_{(n)}$. The likelihood function is $L(\theta; X_{(n)}) = \theta^{-n} I\{\theta \geq X_{(n)}\}$. Accordingly, the MLE of θ is $\hat{\theta} = X_{(n)}$.

B. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a rectangular distribution $R(\theta, 3\theta)$, where $0 < \theta < \infty$. The likelihood function is

$$\begin{aligned} L(\theta; \mathbf{X}) &= (2\theta)^{-n} I\{\theta \leq X_{(1)}, X_{(n)} \leq 3\theta\} \\ &= (2\theta)^{-n} I\left\{\frac{1}{3} X_{(n)} \leq \theta \leq X_{(1)}\right\}, \end{aligned}$$

where $X_{(1)} = \min\{X_i\}$ and $X_{(n)} = \max\{X_i\}$. The m.s.s. is $(X_{(1)}, X_{(n)})$. We note that according to the present model $X_{(n)} \leq 3X_{(1)}$. If this inequality is not

satisfied then the model is incompatible with the data. It is easy to check that the MLE of θ is $\hat{\theta} = \frac{1}{3}X_{(n)}$. The MLE is not a sufficient statistic.

C. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a rectangular distribution $R(\theta, \theta + 1)$, $-\infty < \theta < \infty$. The likelihood function in this case is

$$L(\theta; \mathbf{X}) = I\{\theta \leq X_{(1)} \leq X_{(n)} \leq \theta + 1\}.$$

Note that this likelihood function assumes a constant value 1 over the θ interval $[X_{(n)} - 1, X_{(1)}]$. Accordingly, any value of θ in this interval is an MLE. In the present case, the MLE is not unique. ■

Example 5.15. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common Laplace (double-exponential) distribution with p.d.f.

$$f(x; \mu, \beta) = \frac{1}{2\beta} \exp\left\{-\frac{|x - \mu|}{\beta}\right\}, \quad -\infty < x < \infty,$$

$-\infty < \mu < \infty$, $0 < \beta < \infty$.

A m.s.s. in the present case is the order statistic $X_{(1)} \leq \ldots \leq X_{(n)}$. The likelihood function of (μ, β), given $\mathbf{T} = (X_{(1)}, \ldots, X_{(n)})$, is

$$L(\mu, \beta; \mathbf{T}) = \frac{1}{\beta^n} \exp\left\{-\frac{1}{\beta}\sum_{i=1}^{n} |X_{(i)} - \mu|\right\}.$$

The value of μ which minimizes $\sum_{i=1}^{n} |X_{(i)} - \mu|$ is the sample median M_e. Hence,

$$\begin{aligned} \sup_{\mu} L(\mu, \beta; \mathbf{T}) &= L(M_e, \beta; T_n) \\ &= \frac{1}{\beta^n} \exp\left\{-\frac{1}{\beta}\textstyle\sum_{i=1}^{n} |X_i - M_e|\right\}. \end{aligned}$$

Finally, by differentiating $\log L(M_e, \beta; T)$ with respect to β, we find that the value of β that maximizes $L(M_e, \beta; \mathbf{T})$ is

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} |X_i - M_e|.$$

In the present case, the sample median M_e and the sample mean absolute deviation from M_e are the MLEs of μ and β, respectively. The MLE is not a sufficient statistic. ■

Example 5.16. Consider the normal case in which $X_1, \ldots, X_n$ are i.i.d. random variables distributed like $N(\mu, \sigma^2)$; $-\infty < \mu < \infty$, $0 < \sigma^2 < \infty$. Both parameters

are unknown. The m.s.s. is $(\bar{X}, Q)$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $Q = \sum_{i=1}^{n}(X_i - \bar{X})^2$. The likelihood function can be written as

$$L(\mu, \sigma^2; \bar{X}, Q) = \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{Q}{2\sigma^2} - \frac{n}{2\sigma^2}(\bar{X} - \mu)^2\right\}.$$

Whatever the value of σ^2 is, the likelihood function is maximized by $\hat{\mu} = \bar{X}$. It is easy to verify that the value of σ^2 maximizing (5.5.9) is $\sigma^2 = Q/n$.

The normal distributions under consideration can be written as a two-parameter exponential type, with p.d.f.s

$$f(x; \psi_1, \psi_2) = \frac{1}{(2\pi)^{n/2}} \exp\{\psi_1 T_1 + \psi_2 T_2 - nK(\psi_1, \psi_2)\},$$

where

$$T_1 = \sum X_i, \quad T_2 = \sum X_i^2, \quad \psi_1 = \mu/\sigma^2, \quad \psi_2 = -1/2\sigma^2,$$

and $K(\psi_1, \psi_2) = -\psi_1^2/4\psi_2 + \frac{1}{2}\log(-1/2\psi_2)$. Differentiating the log-likelihood partially with respect to ψ_1 and ψ_2, we obtain that the MLEs of these (natural) parameters should satisfy the system of equations

$$\begin{aligned} \frac{1}{2} \cdot \frac{\hat{\psi}_1}{\hat{\psi}_2} &= -\frac{T_1}{n} \\ + \frac{\hat{\psi}_1^2}{4\hat{\psi}_2^2} - \frac{1}{2\hat{\psi}_2} &= +\frac{T_2}{n}. \end{aligned}$$

We note that $T_1/n = \hat{\mu}$ and $T_2/n = \hat{\sigma}^2 + \hat{\mu}^2$ where $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = Q/n$ are the MLEs of μ and σ^2, respectively. Substituting of μ and $\sigma^2 + \mu^2$, we obtain $\hat{\psi}_1 = \hat{\mu}/\hat{\sigma}^2$, $\hat{\psi}_2 = -1/2\hat{\sigma}^2$. In other words, the relationship between the MLEs $\hat{\psi}_1$ and $\hat{\psi}_2$ to the MLEs $\hat{\mu}$ and $\hat{\sigma}^2$ is exactly like that of ψ_1 and ψ_2 to μ and σ^2. ■

Example 5.17. Consider again the model of Example 5.9. Differentiating the log-likelihood

$$l(\theta, \gamma, \sigma^2 \mid \bar{X}, \bar{Y}, Q) = -n\log(\sigma^2) - \frac{n}{2\sigma^2}\left[(\bar{X} - \theta)^2 + (\bar{Y} - \gamma\theta^2)^2 + \frac{Q}{n}\right],$$

with respect to the parameters, we obtain the equations

$$\begin{aligned} \frac{n}{\hat{\sigma}^2}(\bar{X} - \hat{\theta}) + \frac{2n}{\hat{\sigma}^2}(\bar{Y} - \hat{\gamma}\hat{\theta}^2)\hat{\theta} &= 0 \\ \frac{n}{\hat{\sigma}^2}(\bar{Y} - \gamma\hat{\theta}^2)\hat{\theta}^2 &= 0 \end{aligned}$$

and

$$-\frac{n}{\hat{\sigma}^2}+\frac{n}{2\hat{\sigma}^2}\left[(\bar{X}-\hat{\theta})^2+(\bar{Y}-\hat{\gamma}\hat{\theta}^2)^2+\frac{Q}{n}\right]=0.$$

The unique solution of these equations is

$$\begin{aligned}\hat{\theta}&=\bar{X},\\ \hat{\gamma}&=\bar{Y}/\bar{X}^2,\end{aligned}$$

and

$$\hat{\sigma}^2=\frac{Q}{2n}.$$

It is interesting to realize that $E\{\hat{\gamma}\}$ does not exist, and obviously $\hat{\gamma}$ does not have a finite variance. By the delta method one can find the asymptotic mean and variance of $\hat{\gamma}$. ■

Example 5.18. Let $X_1,\ldots,X_n$ be i.i.d. random variables having a log-normal distribution $LN(\mu,\sigma^2)$. The expected value of X and its variance are

$$\xi=\exp\{\mu+\sigma^2/2\}$$

and

$$D^2=\xi^2(e^{\sigma^2}-1).$$

We have previously shown that the MLEs of μ and σ^2 are $\hat{\mu}=\frac{1}{n}\sum Y_i=\bar{Y}$ and $\hat{\sigma}^2=\frac{1}{n}\sum(Y_i-\bar{Y})^2$, where $Y_i=\log X_i$, $i=1,\ldots,n$. Thus, the MLEs of ξ and D^2 are

$$\hat{\xi}=\exp\{\hat{\mu}+\hat{\sigma}^2/2\}$$

and

$$\hat{D}^2=\hat{\sigma}^2(e^{\hat{\sigma}^2}-1).$$

■

Example 5.19. Let $X_1,X_2,\ldots,X_n$ be i.i.d. random variables having a normal distribution $N(\mu,\sigma^2)$, $-\infty<\mu<\infty$, $0<\sigma^2<\infty$. The MLEs of μ and σ^2 are $\hat{\mu}=\bar{X}_n$ and $\hat{\sigma}^2=\frac{Q}{n}$, where $Q=\sum_{i=1}^{n}(X_i-\bar{X}_n)^2$. By the invariance principle, the MLE of $\theta=\Phi(\frac{\mu}{\sigma})$ is $\hat{\theta}=\Phi(\frac{\bar{X}_n}{\hat{\sigma}})$. ■

Example 5.20. Consider the Weibull distributions, $G^{1/\beta}(\lambda, 1)$, where $0 < \lambda, \beta < \infty$ are unknown. The likelihood function of (λ, β) is

$$L(\lambda, \beta; \mathbf{X}) = (\lambda\beta)^n \left(\prod_{i=1}^{n} X_i\right)^{\beta} \exp\left\{-\lambda \sum_{i=1}^{n} X_i^{\beta}\right\}.$$

Note that the likelihood is equal to the joint p.d.f. of $\mathbf{X}$ multiplied by $\prod_{i=1}^{n} X_i$, which is positive with probability one. To obtain the MLEs of λ and β, we differentiate the log-likelihood partially with respect to these variables and set the derivatives equal to zero. We obtain the system of equations:

$$\hat{\lambda} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^{\hat{\beta}}\right)^{-1},$$

$$\hat{\beta} = \left(\frac{\frac{1}{n}\sum_{i=1}^{n} X_i^{\hat{\beta}} \log X_i}{\frac{1}{n}\sum_{i=1}^{n} X_i^{\hat{\beta}}} - \frac{1}{n}\sum_{i=1}^{n} \log X_i\right)^{-1}.$$

We show now that $\hat{\beta}$ is always positive and that a unique solution exists. Let $\mathbf{x} = (x_1, \ldots, x_n)$, where $0 < x_i < \infty$, $i = 1, \ldots, n$, and let $F(\beta; \mathbf{x}) = \sum_{i=1}^{n} x_i^{\beta} \log x_i / \sum_{i=1}^{n} x_i^{\beta}$. Note that, for every $\mathbf{x}$,

$$\frac{\partial}{\partial\beta} F(\beta; \mathbf{x}) = \frac{\sum_{i=1}^{n} x_i^{\beta} (\log x_i)^2 \cdot \sum_{i=1}^{n} x_i^{\beta} - \left(\sum_{i=1}^{n} x_i^{\beta} \log x_i\right)^2}{\left(\sum_{i=1}^{n} x_i^{\beta}\right)^2} \geq 0$$

with a strict inequality if the x_i values are not all the same. Indeed, if $\omega_i = x_i^{\beta}$ and $\bar{\eta} = \sum_{i=1}^{n} \omega_i \log x_i / \sum_{i=1}^{n} \omega_i$ then $\frac{\partial}{\partial\beta} F(\beta; \mathbf{x}) = \sum_{i=1}^{n} \omega_i (\log x_i - \bar{\eta})^2 / \sum_{i=1}^{n} \omega_i$. Hence, $F(\beta; \mathbf{x})$ is strictly increasing in β, with probability one. Furthermore, $\lim_{\beta \to 0} F(\beta; \mathbf{x}) = \frac{1}{n} \sum \log x_i$ and $\lim_{\beta \to \infty} F(\beta; \mathbf{x}) = \log x_{(n)}$. Thus, the RHS of the β-equation is positive, decreasing function of β, approaching ∞ as $\beta \to 0$ and approaching $(\log x_{(n)} - \frac{1}{n}\sum_{i=1}^{n} \log x_i)^{-1}$ as $\beta \to \infty$. This proves that the solution $\hat{\beta}$ is unique.

The solution for β can be obtained iteratively from the recursive equation

$$\hat{\beta}_{j+1} = \left[\frac{1}{n} \sum_{i=1}^{n} x_i^{\hat{\beta}_j} \log x_i \Big/ \left(\frac{1}{n} \sum_{i=1}^{n} x_i^{\hat{\beta}_j} \right) - \frac{1}{n} \sum_{i=1}^{n} \log x_i \right]^{-1},$$

starting with $\hat{\beta}_0 = 1$. ■

Example 5.21. The present example was given by Stein (1962) in order to illustrate a possible anomalous property of the MLE.

Let $\mathcal{F}$ be a scale-parameter family of distributions, with p.d.f.

$$f(x;\theta) = \frac{1}{\theta} \phi\left(\frac{x}{\theta}\right), \quad 0 < \theta < \infty,$$

where

$$\phi(x) = \begin{cases} B \frac{1}{x} \exp\left\{ -50 \left(1 - \frac{1}{x}\right)^2 \right\}, & \text{if } 0 \le x \le b, \\ 0, & \text{otherwise,} \end{cases}$$

where

$$0 \le b < \infty \text{ and } B^{-1} = \int_0^b \frac{1}{x} \exp\left\{ -50 \left(1 - \frac{1}{x}\right)^2 \right\} dx.$$

Note that $\int_0^\infty \frac{1}{x} \exp\left\{ -50 \left(1 - \frac{1}{x}\right)^2 \right\} dx = \infty$. Accordingly, we choose b sufficiently large so that $\int_{10}^{b} \phi(x) dx = 0.99$. The likelihood function of θ corresponding to one observation is thus

$$L(\theta; x) = \begin{cases} \exp\{-50(\theta - x)^2 / x^2\}, & \text{if } \frac{x}{b} < \theta < \infty, \\ 0, & \text{if } 0 < \theta \le \frac{x}{b}. \end{cases}$$

The MLE of θ is $\hat{\theta} = X$. However, according to the construction of $\phi(x)$,

$$P_\theta\{\hat{\theta} \ge 10\theta\} = \int_{10\theta}^{b\theta} f(x;\theta) dx = \int_{10}^{b} \phi(x) dx = 0.99, \quad \text{for all } \theta.$$

The MLE here is a bad estimator for all θ. ■

Example 5.22. Another source for anomality of the MLE is in the effect of nuisance parameters. A very well-known example of the bad effect of nuisance parameters is due to Neyman and Scott (1948). Their example is presented here.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be n i.i.d. random vectors having the distributions $N(\mu_i \mathbf{1}_2, \sigma^2 I_2)$, $i = 1, \ldots, n$. In other words, each pair (X_i, Y_i) can be considered as representing two independent random variables having a normal distribution with mean μ_i and variance σ^2. The variance is common to all the vectors. We note that $D_i = X_i - Y_i \sim N(0, 2\sigma^2)$ for all $i = 1, \ldots, n$. Hence, $\hat{\sigma}_n^2 = \frac{1}{2n}\sum^n D_i^2$ is an unbiased estimator of σ^2. The variance of $\hat{\sigma}_n^2$ is $\mathrm{Var}\{\hat{\sigma}_n^2\} = 2\sigma^4/n$. Thus, $\hat{\sigma}_n^2$ approaches the value of σ^2 with probability 1 for all (μ_i, σ). We turn now to the MLE of σ^2. The parameter space is $\Theta = \{\mu_1, \ldots, \mu_n, \sigma^2 : -\infty < \mu_i < \infty, i = 1, \ldots, n; 0 < \sigma^2 < \infty\}$. We have to determine a point $(\mu_1, \ldots, \mu_n, \sigma^2)$ that maximizes the likelihood function

$$L(\mu_1, \ldots, \mu_n, \sigma^2; \mathbf{x}, \mathbf{y}) = \frac{1}{\sigma^{2n}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[(x_i - \mu_i)^2 + (y_i - \mu_i)^2]\right\}.$$

We note that $(x_i - \mu_i)^2 + (y_i - \mu_i)^2$ is minimized by $\hat{\mu}_i = (x_i + y_i)/2$. Thus,

$$L(\hat{\mu}_1, \ldots, \hat{\mu}_n, \sigma^2; \mathbf{x}, \mathbf{y}) = \frac{1}{\sigma^{2n}} \exp\left\{-\frac{1}{4\sigma^2}\sum_{i=1}^{n} D_i^2\right\}.$$

The value of σ^2 that maximizes the likelihood is $\tilde{\sigma}^2 = \frac{1}{4n}\sum_{i=1}^{n} D_i^2$. Note that $E_\theta\{\tilde{\sigma}^2\} = \sigma^2/2$ and that by the strong law of large numbers, $\tilde{\sigma}^2 \to \sigma^2/2$ with probability one for each σ^2.

Thus, the more information we have on σ^2 (the larger the sample is) the worse the MLE becomes. It is interesting that if we do not use all the information available then the MLE may become a reasonable estimator. Note that at each given value of σ^2, $M_i = (X_i + Y_i)/2$ is a sufficient statistic for μ_i. Accordingly, the conditional distribution of $(\mathbf{X}, \mathbf{Y})$ given $\mathbf{M} = (M_1, \ldots, M_n)'$ is independent of μ. If we consider the semi-likelihood function, which is proportional to the conditional p.d.f. of $(\mathbf{X}, \mathbf{Y})$, given $\mathbf{M}$ and σ^2, then the value of σ^2 that maximizes this semi-likelihood function coincides with the unbiased estimator $\hat{\sigma}_n^2$. ■

Example 5.23. Consider the standard logistic tolerance distribution, i.e.,

$$F(z) = \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z}, \quad -\infty < z < \infty.$$

The corresponding p.d.f. is

$$f(z) = \frac{e^z}{(1+e^z)^2}.$$

The corresponding function $G(z; \hat{p})$ given by (5.6.10) is

$$G(z; \hat{p}) = \hat{p} - F(z), \quad -\infty < z < \infty.$$

The *logit*, $F^{-1}(z)$, is given by

$$z = \log\left(\frac{F(z)}{1-F(z)}\right).$$

Let $\hat{p}_i$ be the observed proportion of response at dosage x_i. Define $\hat{\zeta}_i = \log(\frac{\hat{p}}{1-\hat{p}_i})$, if $0 < \hat{p}_i < 1$.

According to the model

$$\log\left(\frac{F(\alpha+\beta x_i)}{1-F(\alpha+\beta x_i)}\right) = \alpha + \beta x_i, \quad i = 1, \ldots, k.$$

We, therefore, fit by least squares the line

$$\hat{z}_i = \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\alpha} + \hat{\beta} x_i, \quad i = 1, \ldots, k$$

to obtain the initial estimates of α and β. After that we use the iterative procedure (5.6.12) to correct the initial estimates. For example suppose that the dosages (log dilution) are $x_1 = -2.5, x_2 = -2.25, x_3 = -2, x_4 = -1.75$, and $x_5 = -1.5$. At each dosage a sample of size $n = 20$ is observed, and the results are

x_i	−2.5	−2.25	−2	−1.75	−1.5
$\hat{p}_i$	0.05	0.10	0.15	0.45	0.50
$\hat{z}_i$	−2.9444	−2.1972	−1.7346	−0.2007	0

Least-squares fitting of the regression line $\hat{z}_i = \hat{\alpha} + \hat{\beta} x_i$ yields the initial estimates $\hat{\alpha} = 4.893$ and $\hat{\beta} = 3.154$. Since $G'(z; \hat{p}) = -f(z)$, we define the weights

$$W_i^{(j)} = n_i \frac{\exp(\hat{\alpha}^{(j)} + \hat{\beta}^{(j)} x_i)}{(1+\exp(\hat{\alpha}^{(j)} + \hat{\beta}^{(j)} x_i))^2}$$

and

$$Y_i^{(j)} = n_i \left(\hat{p}_i - \frac{\exp(\hat{\alpha}^{(j)} + \beta^{(j)} x_i)}{1+\exp(\hat{\alpha}^{(j)} + \hat{\beta}^{(j)} x_i)}\right).$$

We solve then equations (5.6.12) to obtain the corrections to the initial estimates. The first five iterations gave the following results:

j	$\hat{\alpha}^{(j)}$	$\hat{\beta}^{(j)}$
0	4.89286	3.15412
1	4.93512	3.16438
2	4.93547	3.16404
3	4.93547	3.16404
4	4.93547	3.16404

■

Example 5.24. $X_1, \ldots, X_n$ are i.i.d. random variables distributed like $N(\mu, \sigma^2)$, where $-\infty < \mu < \infty, 0 < \sigma < \infty$. The group $\mathcal{G}$ considered is that of the real-affine transformations. A m.s.s. is $(\bar{X}, Q)$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $Q = \sum_{i=1}^{n}(X_i - \bar{X})^2$. If $[\alpha, \beta]$ is an element of $\mathcal{G}$ then

$$
\begin{aligned}
[\alpha, \beta]x_i &= \alpha + \beta x_i, \quad i = 1, \ldots, n, \\
[\overline{\alpha, \beta}](\mu, \sigma) &= (\alpha + \beta\mu, \beta\sigma),
\end{aligned}
$$

and

$$
[\widetilde{\alpha, \beta}](\bar{X}, Q) = (\alpha + \beta\bar{X}, \beta^2 Q).
$$

If $\hat{\mu}(\bar{X}, Q)$ is an equivariant estimator of μ then

$$
\hat{\mu}(\alpha + \beta\bar{X}, \beta^2 Q) = \alpha + \beta\hat{\mu}(\bar{X}, Q) = [\alpha, \beta]\hat{\mu}(\bar{X}, Q)
$$

for all $[\alpha, \beta] \in \mathcal{G}$. Similarly, every equivariant estimator of σ^2 should satisfy the relationship

$$
[\overline{\alpha, \beta}]\hat{\sigma}^2(\bar{X}, Q) = \beta^2\hat{\sigma}^2(\bar{X}, Q),
$$

for all $[\alpha, \beta] \in \mathcal{G}$. The m.s.s. $(\bar{X}, Q)$ is reduced by the transformation $[-\bar{X}, 1]$ to $(0, Q)$. This transformation is a maximal invariant reduction of $(\bar{X}, Q)$ with respect to the subgroup of translations $\mathcal{G}_1 = \{[\alpha, 1], -\infty < \alpha < \infty\}$. The difference $D(\bar{X}, Q) = \hat{\mu}(\bar{X}, Q) - \bar{X}$ is translation invariant, i.e., $[\alpha, 1]D(\bar{X}, Q) = D(\bar{X}, Q)$ for all $[\alpha, 1] \in \mathcal{G}_1$. Hence, $D(\bar{X}, Q)$ is a function of the maximal invariant with respect to $\mathcal{G}_1$. Accordingly, every equivariant estimator can be expressed as

$$
\hat{\mu}(\bar{X}, Q) = \bar{X} + f(Q),
$$

where $f(Q)$ is a statistic depending only on Q. Similarly, we can show that every equivariant estimator of σ^2 should be of the form

$$\hat{\sigma}^2(\bar{X}, Q) = \lambda Q,$$

where λ is a positive constant. We can also determine the equivariant estimators of μ and σ^2 having the minimal MSE. We apply the result that $\bar{X}$ and Q are independent. The MSE of $\bar{X} + f(Q)$ for any statistic $f(Q)$ is

$$\begin{aligned} E\{[\bar{X} + f(Q) - \mu]^2\} &= E\{E\{[\bar{X} - \mu + f(Q)]^2 \mid Q\}\} \\ &= \frac{\sigma^2}{n} + E\{f^2(Q)\}. \end{aligned}$$

Hence, the MSE is minimized, by choosing $f(Q) = 0$. Accordingly, the sample mean, $\bar{X}$ is the minimal MSE equivariant estimator of μ. Similarly, one can verify that the equivariant estimator of σ^2, which has the minimal MSE, is $\hat{\sigma}^2 = Q/(n+1)$. Note that this estimator is biased. The UMVU estimator is $Q/(n-1)$ and the MLE is Q/n. ■

Example 5.25. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common $N(\mu, \sigma_1^2)$ distribution. Let $Y_1, \ldots, Y_n$ be i.i.d. random variables distributed as $N(\mu, \sigma_2^2)$. The $\mathbf{X}$ and the $\mathbf{Y}$ vectors are independent. The two distributions have a common mean μ, $-\infty < \mu < \infty$, and possibly different variances. The variance ratio $\rho = \sigma_2^2/\sigma_1^2$ is unknown. A m.s.s. is $(\bar{X}, Q(\mathbf{X}), \bar{Y}, Q(\mathbf{Y}))$, where $\bar{X}$ and $\bar{Y}$ are the sample means and $Q(\mathbf{X})$ and $Q(\mathbf{Y})$ are the sample sums of squares of deviations around the means. $\bar{X}$, $Q(\mathbf{X})$, $\bar{Y}$, and $Q(\mathbf{Y})$ are mutually independent. Consider the group $\mathcal{G}$ of affine transformations $\mathcal{G} = \{[\alpha, \beta] : -\infty < \alpha < \infty, -\infty < \beta < \infty\}$. A maximal invariant statistic is $V = \left(\frac{Q(X)}{(\bar{X} - \bar{Y})^2}, \frac{Q(Y)}{(\bar{X} - \bar{Y})^2}\right)$. Let $W = (\bar{X}, \bar{Y} - \bar{X})$. The vector (W, V) is also a m.s.s. Note that

$$[\alpha, \beta](W, V) = (\alpha + \beta\bar{X}, \beta(\bar{Y} - \bar{X}), V),$$

for all $[\alpha, \beta] \in \mathcal{G}$. Hence, if $\hat{\mu}(W, V)$ is an equivariant estimator of the common mean μ it should be of the form

$$\hat{\mu}(W, V) = \bar{X} + (\bar{Y} - \bar{X})\psi(V),$$

where $\psi(V)$ is a function of the maximal invariant statistic V. Indeed, $\bar{Y} \neq \bar{X}$ with probability one, and $(\hat{\mu}(W, V) - \bar{X})/(\bar{Y} - \bar{X})$ is an invariant statistic, with respect to $\mathcal{G}$. We derive now the MSE of $\hat{\mu}(W, V)$. We prove first that every such equivariant estimator is unbiased. Indeed, for every $\theta = (\mu, \sigma_1^2, \rho)$

$$E_\theta\{\hat{\mu}(W, V)\} = E_\theta\{\bar{X} + (\bar{Y} - \bar{X})\psi(V)\} = \mu + E_\theta\{(\bar{Y} - \bar{X})\psi(V)\}.$$

Moreover, by Basu's Theorem (3.6.1), V is independent of $(\bar{X}, \bar{Y})$. Hence,

$$E_\theta\{(\bar{Y} - \bar{X})\psi(V) \mid |\bar{Y} - \bar{X}|\} = E\{\psi(V)\}E_\theta\{(\bar{Y} - \bar{X}) \mid |\bar{Y} - \bar{X}|\} = 0,$$

with probability one, since the distribution of $\bar{Y} - \bar{X}$ is symmetric around zero. This implies the unbiasedness of $\hat{\mu}(W, V)$. The variance of this estimator is

$$\begin{aligned} V_\theta\{\bar{X} + (\bar{Y} - \bar{X})\psi(V)\} = \frac{\sigma_1^2}{n} &+ 2\mathrm{cov}_\theta(\bar{X}, (\bar{Y} - \bar{X})\psi(V)) \\ &+ V_\theta\{(\bar{Y} - \bar{X})\psi(V)\}. \end{aligned}$$

Since $E_\theta\{(\bar{Y} - \bar{X})\psi(V)\} = 0$, we obtain that

$$\mathrm{cov}_\theta(\bar{X}, (\bar{Y} - \bar{X})\psi(V)) = E_\theta\{(\bar{X} - \mu)(\bar{Y} - \bar{X})\psi(V)\}.$$

The distribution of $\bar{X} - \mu$ depends only on σ_1^2. The maximal invariant statistic V is independent of μ and σ_1^2. It follows from Basu's Theorem that $(\bar{X} - \mu)$ and $\psi(V)$ are independent. Moreover, the conditional distribution of $\bar{X} - \mu$ given $\bar{Y} - \bar{X}$ is the normal distribution $N\left(-\frac{1}{1-\rho}(\bar{Y} - \bar{X}), \frac{\sigma^2}{n}\frac{\rho}{1+\rho}\right)$. Thus,

$$\begin{aligned} \mathrm{cov}_\theta(\bar{X}, (\bar{Y} - \bar{X})\psi(V)) &= E_\theta\{\psi(V)(\bar{Y} - \bar{X})E_\theta\{(\bar{X} - \mu) \mid \bar{Y} - \bar{X}\}\} \\ &= -\frac{1}{1+\rho}E_\theta\{\psi(V)(\bar{Y} - \bar{X})^2\}. \end{aligned}$$

The conditional distribution of $(\bar{Y} - \bar{X})^2$ given V is the gamma distribution $G(\lambda, \nu)$ with

$$\lambda = \frac{1}{2\sigma_1^2}\left(\frac{n}{1+\rho} + Z_1 + \frac{Z_2}{\rho}\right) \quad \text{and} \quad \nu = n - \frac{1}{2},$$

where $Z_1 = Q(\mathbf{X})/(\bar{Y} - \bar{X})^2$ and $Z_2 = Q(\mathbf{Y})/(\bar{Y} - \bar{X})^2$. We thus obtain the expression

$$\begin{aligned} V_\theta\{\hat{\mu}(W, V)\} = \frac{\sigma_1^2}{n}\Bigg(1 + (2n-1)E_\rho\Bigg(\Bigg[&\psi^2(Z_1, Z_2) \\ &- \frac{2}{1+\rho}\psi(Z_1, Z_2)\Bigg]\frac{1+\rho}{1 + (1+\rho)\dfrac{Z_1}{n} + \dfrac{1+\rho}{\rho}\cdot\dfrac{Z_2}{n}}\Bigg)\Bigg). \end{aligned}$$

We see that in the present example the variance divided by σ_1^2/n depends not only on the particular function $\psi(Z_1, Z_2)$ but also on the (nuisance) parameter $\rho = \sigma_1^2/\sigma_1^2$. This is due to the fact that ρ is invariant with respect to $\mathcal{G}$. Thus, if ρ is unknown there is no equivariant estimator having minimum variance for all θ values. There are several papers in which this problem is studied (Brown and Cohen, 1974; Cohen and Sackrowitz, 1974; Kubokawa, 1987; Zacks, 1970a). ■

Example 5.26. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common Weibull distribution $G^{1/\beta}(\lambda^{-\beta}, 1)$, where $0 < \lambda, \beta < \infty$. Note that the parametrization here is different from that of Example 5.20. The present parametrization yields the desired structural properties. The m.s.s. is the order statistic, $T(X) = (X_{(1)}, \ldots, X_{(n)})$, where $X_{(1)} \leq \ldots \leq X_{(n)}$. Let $\hat{\lambda}(T)$ and $\hat{\beta}(T)$ be the MLEs of λ and β, respectively. We obtain the values of these estimators as in Example 5.20, with proper modification of the likelihood function. Define the group of transformations

$$\mathcal{G} = \{[a, b]; \quad 0 < a, \ b < \infty\},$$

where

$$[a, b]X_i = aX_i^{1/b}, \quad i = 1, \ldots, n.$$

Note that the distribution of $[a, b]X$ is as that of $a\lambda^{1/b}G^{1/\beta b}(1, 1)$ or as that of $G^{1/\beta b}((a\lambda^{1/b})^{-b\beta}, 1)$. Accordingly, if $X \to [a, b]X$ then the parametric point (λ, β) is transformed to

$$[\overline{a, b}](\lambda, \beta) = (a\lambda^{1/b}, b\beta).$$

It is easy to verify that

$$[a, b][c, d] = [ac^{1/b}, bd]$$

and

$$[a, b]^{-1} = \left[\frac{1}{a^b}, \frac{1}{b}\right].$$

The reduction of the m.s.s. T by the transformation $[\hat{\lambda}, \hat{\beta}]^{-1}$ yields the maximal invariant $U(T)$

$$U(X_{(1)}, \ldots, X_{(n)}) = \left(\left(\frac{\lambda}{\hat{\lambda}}\right)^{\hat{\beta}} G_{(1)}^{\hat{\beta}/\beta}, \ldots, \left(\frac{\lambda}{\hat{\lambda}}\right)^{\hat{\beta}} G_{(n)}^{\hat{\beta}/\beta}\right),$$

where $G_{(1)} \leq \ldots \leq G_{(n)}$ is the order statistic of n i.i.d. $E(1)$ random variables. The distribution of $U(T)$ does not depend on (λ, β). Thus, $(\frac{\lambda}{\hat{\lambda}})^{\hat{\beta}}$ is distributed independently of (λ, β) and so is that of $\hat{\beta}/\beta$.

Let $\tilde{\lambda} = F(\hat{\lambda}, \hat{\beta}, U(T))$ and $\tilde{\beta} = G(\hat{\lambda}, \hat{\beta}, U(T))$ be equivariant estimators of λ and β respectively. According to the definition of equivariance

$$\begin{aligned}
[\hat{\lambda}, \hat{\beta}]^{-1}F(\hat{\lambda}, \hat{\beta}, U(T)) &= (F(\hat{\lambda}, \hat{\beta}, U(\mathbf{T}))^{\hat{\beta}}/\hat{\lambda}^{\hat{\beta}} \\
&= F([\hat{\lambda}, \hat{\beta}]^{-1}\hat{\lambda}, [\hat{\lambda}, \hat{\beta}]^{-1}\hat{\beta}, U(T)) \\
&= F(1, 1, U(\mathbf{T})) = \psi(U(\mathbf{T})).
\end{aligned}$$

Accordingly, every equivariant estimator of λ is of the form

$$\tilde{\lambda} = \hat{\lambda}(\psi(U(\mathbf{T})))^{1/\hat{\beta}}.$$

Similarly, every equivariant estimator β is of the form

$$\tilde{\beta} = \hat{\beta} H(U(\mathbf{T})).$$

Note that the relationship between the class of all equivariant estimators $(\tilde{\lambda}, \tilde{\beta})$ and the MLEs $(\hat{\lambda}, \hat{\beta})$. In particular, if we choose $\psi(U(T)) = 1$ w.p.1 and $H(U(T)) = 1$ w.p.1 we obtain that the MLEs $\hat{\lambda}$ and $\hat{\beta}$ are equivariant. This property also follows from the fact that the MLE of $[\overline{a, b}](\lambda, \beta)$ is $[\overline{a, b}](\hat{\lambda}, \hat{\beta})$ for all $[a, b]$ in $\mathcal{G}$. We will consider now the problem of choosing the functions $H(U(T))$ and $\psi(U(T))$ to minimize the risk of the equivariant estimator. For this purpose we consider a quadratic loss function in the logarithms, i.e.,

$$L((\tilde{\lambda}, \tilde{\beta}), (\lambda, \beta)) = \left(\log\left(\frac{\tilde{\lambda}}{\lambda}\right)^{\tilde{\beta}}\right)^2 + \left(\log\frac{\tilde{\beta}}{\beta}\right)^2.$$

It is easy to check that this loss function is invariant with respect to $\mathcal{G}$. Furthermore, the risk function does not depend on (λ, β). We can, therefore, choose ψ and H to minimize the risk. The **conditional** risk function, given $U(T)$, when $\psi(U(T)) = \psi$ and $H(U(T)) = H$, is

$$\begin{aligned} R(\psi, H) &= E\left\{\left(\log\left(\frac{(\hat{\lambda}\psi^{1/\hat{\beta}})}{\lambda}\right)\right)^2 \mid U\right\} + E\left\{\left(\log\frac{H\hat{\beta}}{\beta}\right)^2 \mid U\right\} \\ &= H^2 E\left\{\left[\log\left(\frac{\hat{\lambda}}{\lambda}\right)^{\hat{\beta}} + \log\psi\right]^2 \mid U\right\} + E\left\{\left[\log\frac{\hat{\beta}}{\beta} + \log H\right]^2 \mid U\right\}. \end{aligned}$$

Since $(\frac{\hat{\lambda}}{\lambda})^{\hat{\beta}}$ and $\frac{\hat{\beta}}{\beta}$ are ancillary statistics, and since $\mathbf{T}$ is a complete sufficient statistic, we infer from Basu's Theorem that $(\frac{\hat{\lambda}}{\beta})^{\beta}$ and $\frac{\hat{\beta}}{\beta}$ are **independent** of $U(\mathbf{T})$. Hence, the conditional expectations are equal to the total expectations. Partial differentiation with respect to H and ψ yields the system of equations:

$$\textbf{(I)} \quad H^0 E\left\{\left[\log\left(\frac{\hat{\lambda}}{\lambda}\right)^{\hat{\beta}} + \log\psi\right]^2\right\} + \frac{1}{H^0} E\left\{\left[\log\left(\frac{\hat{\beta}}{\beta}\right) + \log H^0\right]\right\} = 0.$$

$$\textbf{(II)} \quad E\left\{\log\left(\frac{\hat{\lambda}}{\lambda}\right)^{\hat{\beta}} + \log\psi^0\right\} = 0.$$

From equation (II), we immediately obtain the expression

$$\psi^0 = \exp\left\{-E\left\{\log\left(\frac{\hat{\lambda}}{\lambda}\right)^{\hat{\beta}}\right\}\right\}.$$

Substituting this ψ^0 in (I), we obtain the equation

$$(H^0)^2 V\left\{\log\left(\frac{\hat{\lambda}}{\lambda}\right)^{\hat{\beta}}\right\} + \log H^0 + E\left\{\log\frac{\hat{\beta}}{\beta}\right\} = 0.$$

This equation can be solved numerically to obtain the optimal constant H^0. Thus, by choosing the functions $\psi(U)$ and $H(U)$ equal (with probability one) to the constants ψ^0 and H^0, respectively, we obtain the minimum MSE equivariant estimators. We can estimate ψ^0 and H^0 by simulation, using the special values of $\lambda = 1$ and $\beta = 1$. ■

Example 5.27. As in Example 5.15, let $X_1, \ldots, X_n$ be i.i.d random variables having a Laplace distribution with a location parameter μ and scale parameter β, where $-\infty < \mu < \infty$ and $0 < \beta < \infty$. The two moments of this distribution are

$$\mu_1 = \mu \quad \mu_2 = 2\beta^2 + \mu^2.$$

The sample moments are $M_1 = \bar{X}$ and $M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$. Accordingly, the MEEs of μ and β are

$$\hat{\mu} = \bar{X} \quad \hat{\beta} = \hat{\sigma}/\sqrt{2},$$

where $\hat{\sigma}^2 = M_2 - M_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. It is interesting to compare these MEEs to the MLEs of μ and β that were derived in Example 5.15. The MLE of μ is the sample median M_e, while the MEE of μ is the sample mean $\bar{X}$. The MEE is an unbiased estimator of μ, with variance $V\{\bar{X}\} = 2\beta^2/n$. The median is also an unbiased estimator of μ. Indeed, let $n = 2m + 1$ then $M_e \sim \mu + \beta Y_{(m+1)}$, where $Y_{(m+1)}$ is the $(m+1)$st order statistic of a sample of $n = 2m + 1$ i.i.d. random variables having a standard Laplace distribution ($\mu = 0$, $\beta = 1$). The p.d.f. of $Y_{(m+1)}$ is

$$g(y) = \frac{(2m+1)!}{(m!)^2} f(y) F^m(y)[1 - F(y)]^m, \quad -\infty < y < \infty,$$

where

$$f(y) = \frac{1}{2}\exp\{-|y|\}, \quad -\infty < y < \infty$$

and

$$F(y) = \begin{cases} \frac{1}{2}e^{y}, & \text{if } y < 0, \\ 1 - \frac{1}{2}e^{-y}, & \text{if } y \geq 0. \end{cases}$$

It is easy to verify that $g(-y) = g(y)$ for all $-\infty < y < \infty$. Hence, $E\{Y_{(m+1)}\} = 0$ and $E\{M_e\} = \mu$. The variance of M_e, for $m \geq 1$, is

$$\begin{aligned} V\{M_e\} &= \sigma^2 V\{Y_{(m+1)}\} \\ &= \sigma^2 \frac{(2m+1)!}{2^m (m!)^2} \int_0^\infty y^2 e^{-(m+1)y} \left(1 - \frac{1}{2}e^{-y}\right)^m dy \\ &= \sigma^2 \frac{(2m+1)!}{2^m (m!)^2} \sum_{j=0}^{m} \left(-\frac{1}{2}\right)^j \binom{m}{j} \int_0^\infty y^2 e^{-(m+j+1)y} dy \\ &= \sigma^2 \frac{(2m+1)!}{2^m (m!)^2} \sum_{j=0}^{m} \left(-\frac{1}{2}\right)^j \binom{m}{j} \frac{1}{(1+j+m)^3}. \end{aligned}$$

Thus, for $\beta = 1$, one obtains the following values for the variances of the estimators:

Est.	$m = 1$	$m = 2$	$m = 3$	$m = 10$	$m = 20$
M_e	0.3194	0.1756	0.1178	0.0327	0.0154
$\bar{X}_n$	0.6666	0.4000	0.2857	0.0952	0.0488

We see that the variance of M_e in small samples is about half the variance of $\bar{X}_n$. As will be shown in Section 5.10, as $n \to \infty$, the ratio of the asymptotic variances approaches 1/2. It is also interesting to compare the expectations and MSE of the MLE and MEE of the scale parameter β. ■

Example 5.28. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common log-normal distribution $LN(\mu, \sigma^2)$, $-\infty < \mu < \infty$, and $0 < \sigma^2 < \infty$. Let $Y_i = \log X_i$, $i = 1, \ldots, n$; $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. $\bar{Y}_n$ and $\hat{\sigma}_n^2$ are the MLEs of μ and σ^2, respectively. We derive now the MEEs of μ and σ^2. The first two moments of $LN(\mu, \sigma^2)$ are

$$\mu_1 = \exp\{\mu + \sigma^2/2\} \qquad \mu_2 = \exp\{2\mu + 2\theta^2\}.$$

Accordingly, the MEEs of μ and σ^2 are

$$\tilde{\mu} = 2\log M_1 - \frac{1}{2}\log M_2 \quad \tilde{\sigma}^2 = \log M_2 - 2\log M_1,$$

where $M_1 = \bar{X}_n$ and $M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ are the sample moments. Note that the MEEs $\tilde{\mu}$ and $\tilde{\sigma}^2$ are **not** functions of the minimal sufficient statistics $\bar{Y}_n$ and $\hat{\sigma}^2$, and therefore are expected to have larger MSEs than those of the MLEs. ■

Example 5.29. In Example 5.20, we discussed the problem of determining the values of the MLEs of the parameters λ and β of the Weibull distribution, where $X_1, \ldots, X_n$ are i.i.d. like $G^{1/\beta}(\lambda, 1)$ where $0 < \beta, \lambda < \infty$. The MEEs are obtained in the following manner. According to Table 2.2, the first two moments of $G^{1/\beta}(\lambda, 1)$ are

$$\mu_1 = \Gamma(1 + 1/\beta)/\lambda^{1/\beta} \qquad \mu_2 = \Gamma(1 + 2/\beta)/\lambda^{2/\beta}.$$

Thus, we set the moment equations

$$M_1 = \Gamma(1 + 1/\hat{\beta})/\hat{\lambda}^{1/\hat{\beta}} \quad M_2 = \Gamma(1 + 2/\hat{\beta})/\hat{\lambda}^{2/\hat{\beta}}.$$

Accordingly, the MEE $\hat{\beta}$ is the root of the equation

$$\frac{2\hat{\beta}}{B(\frac{1}{\hat{\beta}}, \frac{1}{\hat{\beta}})} = \frac{M_2}{M_1}.$$

The solution of this equation can be obtained numerically. After solving for $\hat{\beta}$, one obtains $\hat{\lambda}$ as follows:

$$\hat{\lambda} = \left(\frac{\Gamma(1 + 1/\hat{\beta})}{M_1}\right)^{\hat{\beta}}.$$

We illustrate this solution with the numbers in the sample of Example 5.14. In that sample, $n = 50$, $\sum_{i=1}^{n} X_i = 46.6897$, and $\sum_{i=1}^{n} X_i^2 = 50.9335$. Thus, $M_1 = .9338$ and $M_2 = 1.0187$. Equation (5.8.9) becomes

$$1.71195\hat{\beta} = B\left(\frac{1}{\hat{\beta}}, \frac{1}{\hat{\beta}}\right).$$

The solution should be in the neighborhood of $\beta = 2$, since $2 \times 1.71195 = 3.4239$ and $B(\frac{1}{2}, \frac{1}{2}) = \pi = 3.14195\ldots$. In the following table, we approximate the solution:

β	$\Gamma(\frac{1}{\beta})$	$B(\frac{1}{\beta}, \frac{1}{\beta})$	1.71195β
2.5	2.21815	4.22613	4.2798
2.6	2.30936	4.44082	4.4510
2.7	2.40354	4.66284	4.6222

Accordingly, the MEE of β is approximately $\hat{\beta} = 2.67$ and that of λ is approximately $\hat{\lambda} = 0.877$. The values of the MLE of β and λ, obtained in Example 5.20, are 1.875 and 0.839, respectively. The MLEs are closer to the true values, but are more difficult to obtain. ■

Example 5.30.

A. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. random vectors having a bivariate normal distribution $N(\mathbf{0}, R)$, where $R = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ $-1 < \rho < 1$. Accordingly, an estimator of ρ is the sample mixed moment $M_{11} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$. This is also an unbiased estimator of ρ. There is no UMVU estimator of ρ, since the family of all such distributions is incomplete.

The likelihood function of ρ is

$$L(\rho; \mathbf{X}, \mathbf{Y}) = \frac{1}{(1-\rho^2)^{n/2}} \exp\left\{-\frac{1}{2(1-\rho^2)}[Q_X + Q_Y - 2\rho P_{XY}]\right\},$$

where $Q_X = \Sigma X_i^2$, $Q_Y = \Sigma Y_i^2$, and $P_{XY} = \Sigma X_i Y_i$. Note that the m.s.s. is $\mathbf{T} = (Q_X + Q_Y, P_{XY})$. The maximal likelihood estimator of ρ is a real solution of the cubic equation

$$n\rho^3 - \rho^2 P_{XY} + (S - n)\rho - P_{XY} = 0,$$

where $S = Q_x + Q_y$. In the present example, the MEE is a very simple estimator. There are many different unbiased estimators of ρ. The MEE is one such unbiased estimator. Another one is

$$\tilde{\rho} = 1 - \frac{1}{2n}(S - 2P_{XY}).$$

B. Consider the model of Example 5.10. The likelihood function is

$$L(\theta, \gamma, \sigma^2 \mid \bar{X}, \bar{Y}, Q) = \frac{1}{(\sigma^2)^n} \exp\left\{-\frac{n}{2\sigma^2}\left[(\bar{X} - \theta)^2 + (\bar{Y} - \gamma\theta^2)^2 + \frac{Q}{n}\right]\right\},$$

$-\infty < \theta, \gamma < \infty$, $0 < \sigma^2 < \infty$. The MEE of σ^2 is $\hat{\sigma}_n^2 = \frac{Q}{2n}$. Similarly, we find that the MEEs of θ and γ are

$$\hat{\theta} = \bar{X} \quad \hat{\gamma} = \frac{\bar{Y}}{\bar{X}^2}.$$

The MLEs are the same. ■

Example 5.31. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common $N(\mu, \sigma^2)$ distribution. The problem is to estimate the variance σ^2. If $\mu = 0$ then the minimum MSE equivariant estimator of σ^2 is $\hat{\sigma}_0^2 = \frac{1}{n+2}\sum_{i=1}^{n} X_i^2$. On the other hand, if μ is unknown the minimum MSE equivariant estimator of σ^2 is $\hat{\sigma}_1^2 = \frac{1}{n+1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$, where $\bar{X}_n = \frac{1}{n}\Sigma X_i$. One could suggest to test first the hypothesis $H_0 : \mu = 0$, σ arbitrary, against $H_1 : \mu \neq 0$, σ arbitrary, at some level of significance α. If H_0 is accepted the estimator is $\hat{\sigma}_0^2$, otherwise, it is $\hat{\sigma}_1^2$. Suppose that the preliminary test is the t-test. Thus, the estimator of σ^2 assumes the form:

$$\hat{\sigma}^2 = \hat{\sigma}_0^2 I\{\bar{X}, S^2 : \frac{\sqrt{n}|\bar{X}_n|}{S_n} \leq t_{1-\alpha/2}[n-1]\}$$

$$+ \hat{\sigma}_1^2 I\{\bar{X}, S^2; \frac{\sqrt{n}|X_n|}{S_n} > t_{1-\alpha/2}[n-1]\},$$

where S_n^2 is the sample variance. Note that this PTE is not translation invariant, since neither the t-test of H_0 is translation invariant, nor is $\hat{\sigma}_0^2$. The estimator σ^2 may have smaller MSE values than those of $\hat{\sigma}_0^2$ or of $\hat{\sigma}_1^2$, on some intervals of (μ, σ^2) values. Actually, $\hat{\sigma}^2$ has smaller MSE than that of $\hat{\sigma}_1^2$ for all (μ, σ^2) if $t_{1-\alpha/2}[n-1] = \sqrt{\frac{n-1}{n+1}} \approx 1$. This corresponds to (when n is large) a value of α approximately equal to $\alpha = 0.3173$. ■

Example 5.32. Let $X_1, \ldots, X_n$ be a sample of i.i.d. random variables from $N(\mu, \sigma_1^2)$ and let $Y_1, \ldots, Y_n$ be a sample of i.i.d. random variables from $N(\mu, \sigma_2^2)$. The $\mathbf{X}$ and $\mathbf{Y}$ vectors are independent. The problem is to estimate the common mean μ. In Example 5.24, we studied the MSE of equivariant estimators of the common mean μ. In Chapter 8, we will discuss the problem of determining an optimal equivariant estimator of μ in a Bayesian framework. We present here a PTE of μ. Let $\rho = \sigma_2^2/\sigma_1^2$. If $\rho = 1$ then the UMVU estimator of μ is $\hat{\mu}_1 = (\bar{X} + \bar{Y})/2$, where $\bar{X}$ and $\bar{Y}$ are the sample means. When ρ is unknown then a reasonably good unbiased estimator of μ is $\hat{\mu}_R = (\bar{X}R + \bar{Y})/(R+1)$, where $R = S_Y^2/S_X^2$ is the ratio of the sample variances S_Y^2 to S_X^2. A PTE of μ can be based on a preliminary test of $H_0 : \rho = 1$, μ, σ_1, σ_2

arbitrary against $H_1 : \rho \neq 1, \mu, \sigma_1, \sigma_2$ arbitrary. If we apply the F-test, we obtain the PTE

$$\hat{\mu} = \hat{\mu}_1 I\{R \leq F_{1-\alpha}[n-1, n-1]\} + \hat{\mu}_R I\{R > F_{1-\alpha}[n-1, n-1]\}.$$

This estimator is unbiased, since $\bar{X}$ and $\bar{Y}$ are independent of R. Furthermore,

$$V\{\hat{\mu} \mid R\} = \begin{cases} \frac{\sigma_1^2}{n} \cdot \frac{1+\rho}{4}, & \text{if } R \leq F_{1-\alpha}[n-1, n-1], \\ \frac{\sigma_1^2}{n} \cdot \frac{\rho + R^2}{(1+R)^2}, & \text{if } R > F_{1-\alpha}[n-1, n-1]. \end{cases}$$

Hence, since $E\{\hat{\mu} \mid R\} = \mu$ for all R, we obtain from the law of total variance that the variance of the PTE is

$$V\{\hat{\mu}\} = \frac{\sigma_1^2}{n}\left(\frac{1+\rho}{4} P\{F[n-1, n-1] \leq \frac{1}{\rho} F_{1-\alpha}[n-1, n-1]\} + \int_{R^*}^{\infty} \frac{\rho + R^2}{(1+R)^2} f_\rho(R) dR\right),$$

where $R^* = F_{1-\alpha}[n-1, n-1]$, and $f_\rho(R)$ is the p.d.f. of $\rho F[n-1, n-1]$ at R. Closed formulae in cases of small n were given by Zacks (1966). ■

PART III: PROBLEMS

Section 5.2

5.2.1 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a rectangular distribution $R(\theta_1, \theta_2)$, $-\infty < \theta_1 < \theta_2 < \infty$.

(i) Determine the UMVU estimators of θ_1 and θ_2.

(ii) Determine the covariance matrix of these UMVU estimators.

5.2.2 Let $X_1, \ldots, X_n$ be i.i.d. random variables having an exponential distribution, $E(\lambda)$, $0 < \lambda < \infty$.

(i) Derive the UMVU estimator of λ and its variance.

(ii) Show that the UMVU estimator of $\rho = e^{-\lambda}$ is

$$\hat{\rho} = \left(\left(1 - \frac{1}{T}\right)^+\right)^{n-1},$$

where $T = \sum_{i=1}^{n} X_i$ and $a^+ = \max(a, 0)$.

(iii) Prove that the variance of $\hat{\rho}$ is

$$V\{\hat{\rho}\} = \frac{1}{\Gamma(n)}\left(\sum_{i=0}^{n-1}(-\lambda)^i\binom{2n-2}{i}(n-i-1)!\mathrm{P}(n-i-1;\lambda) + \sum_{i=n}^{2n-2}(-\lambda)^i\binom{2n-2}{i}H(i-n+1\mid\lambda)\right) - e^{-2\lambda},$$

where $\mathrm{P}(j;\lambda)$ is the c.d.f. of $P(\lambda)$ and $H(k\mid x) = \int_x^\infty e^{-u}/u^k du$. [$H(k\mid x)$ can be determined recursively by the relation

$$H(k\mid x) = \frac{1}{k-1}\left(\frac{e^{-x}}{x^{k-1}} - H(k-1\mid x)\right), \quad k \geq 2$$

and $H(1\,|\,x)$ is the exponential integral (Abramowitz and Stegun, 1968).

5.2.3 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a two-parameter exponential distribution, $X_1 \sim \mu + G(\lambda, 1)$. Derive the UMVU estimators of μ and λ and their covariance matrix.

5.2.4 Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, 1)$ random variables.

(i) Find a $\lambda(n)$ such that $\Phi(\lambda(n)\bar{X})$ is the UMVU estimator of $\Phi(\mu)$.

(ii) Derive the variance of this UMVU estimator.

5.2.5 Consider Example 5.4. Find the variances of the UMVU estimators of $p(0;\lambda)$ and of $p(1;\lambda)$. [Hint: Use the formula of the p.g.f. of a $P(n\lambda)$.]

5.2.6 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a $NB(\psi, \nu)$ distribution; $0 < \psi < \infty$ (ν known). Prove that the UMVU estimator of ψ is

$$\hat{\psi} = \frac{T}{n\nu + T - 1}, \quad \text{where } T = \sum_{i=1}^{n} X_i.$$

5.2.7 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a binomial distribution $B(N, \theta)$, $0 < \theta < 1$.

(i) Derive the UMVU estimator of θ and its variance.

(ii) Derive the UMVU estimator of $\sigma^2(\theta) = \theta(1-\theta)$ and its variance.

(iii) Derive the UMVU estimator of $b(j; N, \theta)$.

5.2.8 Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, 1)$ random variables. Find a constant $b(n)$ so that

$$f(\xi; \bar{X}) = \frac{1}{\sqrt{2\pi}\,\sqrt{b(n)}}\exp\left\{-\frac{1}{2b(n)}(\xi - \bar{X})^2\right\}$$

is a UMVU estimator of the p.d.f. of X at ξ, i.e., $\frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(\xi - \mu)^2\}$. [Hint: Apply the m.g.f. of $(\bar{X} - \xi)^2$.]

5.2.9 Let $J_1, \ldots, J_n$ be i.i.d. random variables having a binomial distribution $B(1, e^{-\Delta/\theta})$, $0 < \theta < 1$ (Δ known). Let $\hat{p}_n = \left(\sum_{i=1}^{n} J_i + \frac{1}{2}\right)/(n+1)$. Consider the estimator of θ

$$\hat{\theta}_n = -\Delta / \log(\hat{p}_n).$$

Determine the bias of $\hat{\theta}_n$ as a power-series in $1/n$.

5.2.10 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a binomial distribution $B(N, \theta)$, $0 < \theta < 1$. What is the Cramér–Rao lower bound to the variance of the UMVU estimator of $\omega = \theta(1 - \theta)$?

5.2.11 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a negative-binomial distribution $NB(\psi, \nu)$. What is the Cramér–Rao lower bound to the variance of the UMVU estimator of ψ? [See Problem 6.]

5.2.12 Derive the Cramér–Rao lower bound to the variance of the UMVU estimator of $\delta = e^{-\lambda}$ in Problem 2.

5.2.13 Derive the Cramér–Rao lower bound to the variance of the UMVU estimator of $\Phi(\mu)$ in Problem 4.

5.2.14 Derive the BLBs of the second and third order for the UMVU estimator of $\Phi(\mu)$ is Problem 4.

5.2.15 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common $N(\mu, \sigma^2)$ distribution, $-\infty < \mu < \infty$, $0 < \sigma^2 < \infty$.

(i) Show that $\hat{\omega} = \exp\{\bar{X}\}$ is the UMVU estimator of $\omega = \exp\{\mu + \sigma^2/2n\}$.

(ii) What is the variance of $\hat{\omega}$?

(iii) Show that the Cramér–Rao lower bound for the variance of $\hat{\omega}$ is $\frac{\sigma^2}{n} e^{2\mu + \sigma^2/n} \left(1 + \frac{\sigma^2}{2n^2}\right)$.

5.2.16 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common $N(\mu, \sigma^2)$ distribution, $-\infty < \mu < \infty$, $0 < \sigma < \infty$. Determine the Cramér–Rao lower bound for the variance of the UMVU estimator of $\omega = \mu + z_\gamma \sigma$, where $z_\gamma = \Phi^{-1}(\gamma)$, $0 < \gamma < 1$.

5.2.17 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a $G(\lambda, \nu)$ distribution, $0 < \lambda < \infty$, $\nu \geq 3$ fixed.

(i) Determine the UMVU estimator of λ^2.

(ii) Determine the variance of this UMVU.

(iii) What is the Cramér–Rao lower bound for the variance of the UMVU estimator?

(iv) Derive the BLBs of orders 2, 3, and 4.

5.2.18 Consider Example 5.8. Show that the Cramér–Rao lower bound for the variance of the MVU estimator of $\text{cov}(X, Y) = \rho\sigma^2$ is $\frac{\sigma^4}{n}(1 + \rho^2)$.

5.2.19 Let $X_1, \ldots, X_n$ be i.i.d. random variables from $N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_n$ i.i.d. from $N(\mu_2, \sigma_2^2)$. The random vectors $\mathbf{X}$ and $\mathbf{Y}$ are independent and $n \geq 3$. Let $\delta = \sigma_2^2/\sigma_1^2$.

(i) What is the UMVU estimator of δ and what is its variance?

(ii) Derive the Cramér–Rao lower bound to the variance of the UMVU estimator of δ.

5.2.20 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a rectangular distribution $R(0, \theta)$, $0 < \theta < \infty$. Derive the Chapman–Robbins inequality for the UMVU of θ.

5.2.21 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a Laplace distribution $L(\mu, \sigma)$, $-\infty < \mu < \infty$, $0 < \sigma < \infty$. Derive the Chapman–Robbins inequality for the variances of unbiased estimators of μ.

Section 5.3

5.3.1 Show that if $\hat{\theta}(\mathbf{X})$ is a biased estimator of θ, having a differentiable bias function $B(\theta)$, then the efficiency of $\hat{\theta}(\mathbf{X})$, when the regularity conditions hold, is

$$\mathcal{E}_\theta(\hat{\theta}) = \frac{(1 + B'(\theta))^2}{I_n(\theta) V_\theta\{\hat{\theta}\}}.$$

5.3.2 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a negative exponential distribution $G(\lambda, 1)$, $0 < \lambda < \infty$.

(i) Derive the efficiency function $\mathcal{E}(\lambda)$ of the UMVU estimator of λ.

(ii) Derive the efficiency function of the MLE of λ.

5.3.3 Consider Example 5.8.

(i) What are the efficiency functions of the unbiased estimators of σ^2 and ρ, where $\hat{\rho} = \Sigma X_i Y_i / \Sigma X_i^2$ and $\hat{\sigma}^2 = \frac{1}{2n}\sum_{i=1}^{n}(X_i^2 + Y_i^2)$.

(ii) What is the combined efficiency function (5.3.13) for the two estimators simultaneously?

Section 5.4

5.4.1 Let $X_1, \ldots, X_n$ be equicorrelated random variables having a common unknown mean μ. The variance of each variable is σ^2 and the correlation between any two variables is $\rho = 0.7$.

(i) Show that the covariance matrix of $\mathbf{X} = (X_1, \ldots, X_n)'$ is $\mathfrak{V} = \sigma^2(0.3I_n + 0.7J_n) = 0.3\sigma^2(I_n + \frac{7}{3}J_n)$, where I_n is the identity matrix of order n and J_n is an $n \times n$ matrix of 1s.

(ii) Determine the BLUE of μ.

(iii) What is the variance of the BLUE of μ?

(iv) How would you estimate σ^2?

5.4.2 Let X_1, X_2, X_3 be i.i.d. random variables from a rectangular distribution $R(\mu - \sigma, \mu + \sigma), -\infty < \mu < \infty, 0 < \sigma < \infty$. What is the best linear combination of the order statistics $X_{(i)}$, $i = 1, 2, 3$, for estimating μ, and what is its variance?

5.4.3 Suppose that $X_1, \ldots, X_n$ are i.i.d. from a Laplace distribution with p.d.f. $f(x; \mu, \sigma) = \frac{1}{\sigma}\psi\left(\frac{x-\mu}{\sigma}\right), -\infty < x < \infty$, where $\psi(z) = \frac{1}{2}e^{-|z|}; -\infty < \mu < \infty, 0 < \sigma < \infty$. What is the best linear unbiased combination of $X_{(1)}$, M_e, and $X_{(n)}$ for estimating μ, when $n = 5$?

5.4.4 Let $\psi_k(T) = \sum_{t=1}^{T} t^k$.

(i) Show that

$$\sum_{k=0}^{p} \binom{p+1}{k} \psi_k(T) = (T+1)^{p+1} - 1.$$

(ii) Apply (i) to derive the following formulae:

$$\sum_{t=1}^{T} t = \frac{1}{2}T(T+1),$$

$$\sum_{t=1}^{T} t^2 = \frac{1}{6}T(T+1)(2T+1),$$

$$\sum_{t=1}^{T} t^3 = \frac{1}{4}T^2(T+1)^2,$$

$$\sum_{t=1}^{T} t^4 = \frac{1}{30}T(T+1)(2T+1)(3T^2+3T-1),$$

$$\sum_{t=1}^{T} t^5 = \frac{1}{12} T^2(T+1)^2(2T^2+2T-1),$$

$$\sum_{t=1}^{T} t^6 = \frac{1}{42} T(T+1)(2T+1)(3T^4+6T^3-3T+1).$$

[Hint: To prove (i), show that both sides are $\sum_{t=1}^{T}(t+1)^{p+1} - \sum_{t=1}^{T} t^{p+1}$ (Anderson, 1971, p. 83).]

5.4.5 Let $X_t = f(t) + e_t$, where $t = 1, \ldots, T$, where

$$f(t) = \sum_{i=0}^{p} \beta_i t^i, \quad t = 1, \ldots, T;$$

e_t are uncorrelated random variables, with $E\{e_t\} = 0$, $V\{e_t\} = \sigma^2$ for all $t = 1, \ldots, T$.

(i) Write the normal equations for the least-squares estimation of the polynomial coefficients β_i $(i = 0, \ldots, p)$.

(ii) Develop explicit formula for the coefficients β_i in the case of $p = 2$.

(iii) Develop explicit formula for $V\{\beta_i\}$ and σ^2 for the case of $p = 2$. [The above results can be applied for a polynomial trend fitting in time series analysis when the errors are uncorrelated.]

5.4.6 The annual consumption of meat per capita in the United States during the years 1919–1941 (in pounds) is (Anderson, 1971, p. 44)

t	19	20	21	22	23	24	25	26	27
X_t	171.5	167.0	164.5	169.3	179.4	179.2	172.6	170.5	168.6

t	28	29	30	31	32	33	34	35	36
X_t	164.7	163.6	162.1	160.2	161.2	165.8	163.5	146.7	160.2

t	37	38	39	40	41
X_t	156.8	156.8	165.4	174.7	178.7

(i) Fit a cubic trend to the data by the method of least squares.

(ii) Estimate the error variance σ^2 and test the significance of the polynomial coefficients, assuming that the errors are i.i.d. $N(0, \sigma^2)$.

5.4.7 Let (x_{1i}, Y_{1i}), $i = 1, \ldots, n_1$, and (x_{2i}, Y_{2i}), $i = 1, \ldots, n_2$, be two independent sets of regression points. It is assumed that

$$Y_{ji} = \beta_{0j} + \beta_1 x_{ji} + e_{ji} \qquad j = 1, 2 \quad i = 1, \ldots, n_j,$$

where x_{ji} are constants and e_{ji} are i.i.d. $N(0, \sigma^2)$. Let

$$SDX_j = \sum_{i=1}^{n_j} (x_{ji} - \bar{x}_j)^2,$$

$$SPD_j = \sum_{i=1}^{n_j} (x_{ji} - \bar{x}_j)(Y_{ji} - \bar{Y}_j), \quad j = 1, 2,$$

$$SDY_j = \sum_{i=1}^{n_j} (Y_{ji} - \bar{Y}_j)^2,$$

where $\bar{x}_j$ and $\bar{Y}_j$ are the respective sample means.

(i) Show that the LSE of β_1 is

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{2} SPD_j}{\sum_{j=1}^{2} SDX_j},$$

and that the LSEs of β_{0j} $(j = 1, 2)$ are

$$\hat{\beta}_{0j} = \bar{Y}_j - \hat{\beta}_1 \bar{x}_j.$$

(ii) Show that an unbiased estimator of σ^2 is

$$s^2_{y/x} = \frac{1}{N-3} \left\{ \sum_{j=1}^{2} SDY_j - \hat{\beta}_1 \sum_{j=1}^{2} SPD_j \right\},$$

where $N = n_1 + n_2$.

(iii) Show that

$$V\{\hat{\beta}_1\} = \sigma^2 / \sum_{j=1}^{2} SDX_j; \quad V\{\hat{\beta}_{0j}\} = \frac{\sigma^2}{n_j} \left\{ 1 + \frac{n_j \bar{x}_j^2}{\sum_{j=1}^{2} SDX_j} \right\}.$$

Section 5.5

5.5.1 Consider the following raw data (Draper and Smith, 1966, p. 178).

Y	78.5	74.3	104.3	87.6	95.9	109.2	102.7	72.5	93.1	115.9	83.8	113.3	109.4
X_1	7	1	11	11	7	11	3	1	2	21	1	11	10
X_2	26	29	56	31	52	55	71	31	54	47	40	66	68
X_3	6	15	8	8	6	9	17	22	18	4	23	9	8
X_4	60	52	20	47	33	22	6	44	22	26	34	12	12

(i) Determine the LSE of $\beta_0, \ldots, \beta_4$ and of σ^2 in the model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_4 X_4 + e$, where $e \sim N(0, \sigma^2)$.

(ii) Determine the ridge-regression estimates $\beta_i(k)$, $i = 0, \ldots, 4$ for $k = 0.1, 0.2, 0.3$.

(iii) What value of k would you use?

Section 5.6

5.6.1 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a binomial distribution $B(1, \theta)$, $0 < \theta < 1$. Find the MLE of

(i) $\sigma^2 = \theta(1 - \theta)$;

(ii) $\rho = e^{-\theta}$;

(iii) $\omega = e^{-\theta}/(1 + e^{-\theta})$;

(iv) $\phi = \log(1 + \theta)$.

5.6.2 Let $X_1, \ldots, X_n$ be i.i.d. $P(\lambda)$, $0 < \lambda < \infty$. What is the MLE of $p(j; \lambda) = e^{-\lambda} \cdot \lambda^j / j!$, $j = 0, 1, \ldots$?

5.6.3 Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$, $-\infty < \mu < \infty, 0 < \sigma < \infty$. Determine the MLEs of

(i) $\mu + Z_\gamma \sigma$, where $Z_\gamma = \Phi^{-1}(\gamma)$, $0 < \gamma < 1$;

(ii) $\omega(\mu, \sigma) = \Phi(\mu/\sigma) \cdot [1 - \Phi(\mu/\sigma)]$.

5.6.4 Using the delta method (see Section 1.13.4), determine the large sample approximation to the expectations and variances of the MLEs of Problems 1, 2, and 3.

5.6.5 Consider the normal regression model

$$Y_i = \beta_0 + \beta_1 x_i + e_i \quad (i = 1, \ldots, n),$$

where $x_1, \ldots, x_n$ are constants such that $\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0$, and $e_1, \ldots, e_n$ are i.i.d. $N(0, \sigma^2)$.

(i) Show that the MLEs of β_0 and β_1 coincide with the LSEs.

(ii) What is the MLE of σ^2?

5.6.6 Let (x_i, T_i), $i = 1, \ldots, n$ be specified in the following manner. $x_1, \ldots, x_n$ are constants such that $\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $T_1, \ldots, T_n$ are independent random variables and $T_i \sim G\left(\frac{1}{\alpha + \beta x_i}, 1\right)$, $i = 1 = 1, ..., n$.

(i) Determine the maximum likelihood equations for α and β.

(ii) Set up the Newton–Raphson iterative procedure for determining the MLE, starting with the LSE of α and β as initial solutions.

5.6.7 Consider the MLE of the parameters of the normal, logistic, and extreme-value tolerance distributions (Section 5.6.6). Let $x_1 < \ldots < x_k$ be controlled experimental levels, $n_1, \ldots, n_k$ the sample sizes and $J_1, \ldots, J_k$ the number of response cases from those samples. Let $p_i = (J_i + 1/2)/(n_i + 1)$. The following transformations:

1. **Normit**: $Y_i = \Phi^{-1}(p_i)$, $i = 1, \ldots, k$;
2. **Logit**: $Y_i = \log(p_i/(1 - p_i))$, $i = 1, \ldots, k$;
3. **Extremit**: $Y_i = -\log(-\log p_i)$, $i = 1, \ldots, k$; are applied first to determine the initial solutions. For the normal, logistic, and extreme-value models, determine the following:

   (i) The LSEs of θ_1 and θ_2 based on the linear model $Y_i = \theta_1 + \theta_2 x_i + e_i$ $(i = 1, \ldots, k)$.

   (ii) The MLE of θ_1 and θ_2, using the LSEs as initial solutions.

   (iii) Apply (i) and (ii) to fit the normal, logistic, and extreme-value models to the following set of data in which $k = 3$; $n_i = 50$ $(i = 1, 2, 3)$; $x_1 = -1$, $x_2 = 0$, $x_3 = 1$; $J_1 = 15$, $J_2 = 34$, $J_3 = 48$.

   (iv) We could say that one of the above three models fits the data better than the other two if the corresponding statistic

$$W^2 = \sum_{i=1}^{k} n_i p_i^2 / F(x; \theta)$$

   is minimal; or

$$D^2 = \sum_{i=1}^{k} n_i p_i \log F(x; \theta)$$

   is maximal. Determine W^2 and D^2 to each one of the above models, according to the data in (iii), and infer which one of the three models better fits the data.

5.6.8 Consider a trinomial according to which (J_1, J_2) has the trinomial distribution $M(n, \mathbf{P}(\theta))$, where

$$P_1(\theta) = \theta^2, \quad P_2(\theta) = 2\theta(1-\theta), \quad P_3(\theta) = (1-\theta)^2.$$

This is the Hardy–Weinberg model.

(i) Show that MLE of θ is

$$\hat{\theta}_n = \frac{2J_1 + J_2}{2n}.$$

(ii) Find the Fisher information function $I_n(\theta)$.

(iii) What is the efficiency in small samples $\mathcal{E}(\hat{\theta}_n)$ of the MLE?

5.6.9 A **minimum chi-squared estimator** (MCE) of θ in a multinomial model $M(n, \mathbf{P}(\theta))$ is an estimator $\tilde{\theta}_n$ minimizing

$$X^2 = \sum_{i=1}^{k} (J_i - nP_i(\theta))^2 / nP_i(\theta).$$

For the model of Problem 8, show that the MCE of θ is the real root of the equation

$$2(2J_1^2 - J_2^2 + 2J_3^2)\theta^3 - 3(4J_1^2 - J_2^2)\theta^2 + (12J_1^2 - J_2^2)\theta - 4J_1^2 = 0.$$

Section 5.7

5.7.1 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common rectangular distribution $R(0, \theta)$, $0 < \theta < \infty$.

(i) Show that this model is preserved under the group of transformations of scale changing, i.e., $\mathcal{G} = \{g_\beta : g_\beta X = \beta X, 0 < \beta < \infty\}$.

(ii) Show that the minimum MSE equivariant estimator of θ is $\dfrac{n+2}{n+1} X_{(n)}$.

5.7.2 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common location-parameter Cauchy distribution, i.e., $f(x; \mu) = \dfrac{1}{\pi}(1 + (x-\mu)^2)^{-1}$, $-\infty < x < \infty$; $-\infty < \mu < \infty$. Show that the Pitman estimator of μ is

$$\hat{\mu} = X_{(1)} - \left\{ \int_{-\infty}^{\infty} u(1+u^2)^{-1} \prod_{i=2}^{n} (1 + (Y_{(i)} + u)^2)^{-1} du \right\} \Bigg/ \left\{ \int_{-\infty}^{\infty} (1+u^2)^{-1} \prod_{i=2}^{n} (1 + (Y_{(i)} + u)^2)^{-1} du \right\},$$

where $Y_{(i)} = X_{(i)} - X_{(1)}$, $i = 2, \ldots, n$. Or, by making the transformation $\omega = (1 + u^2)^{-1}$ one obtains the expression

$$\hat{\mu} = X_{(1)} + \frac{\int_0^1 \frac{1}{\omega} \left\{ \prod_{i=2}^{n} \left[1 + \left(Y_{(i)} - \sqrt{\frac{1-\omega}{\omega}} \right)^2 \right]^{-1} + \prod_{i=2}^{n} \left[1 + \left(Y_{(i)} + \sqrt{\frac{1-\omega}{\omega}} \right)^2 \right]^{-1} \right\} d\omega}{\int_0^1 \frac{1}{\sqrt{\omega(1-\omega)}} \left\{ \prod_{i=2}^{n} \left[1 + \left(Y_{(i)} - \sqrt{\frac{1-\omega}{\omega}} \right)^2 \right]^{-1} + \prod_{i=2}^{n} \left[1 + \left(Y_{(i)} + \sqrt{\frac{1-\omega}{\omega}} \right)^2 \right]^{-1} \right\} d\omega}.$$

This estimator can be evaluated by numerical integration.

5.7.3 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a $N(\mu, \sigma^2)$ distribution. Determine the Pitman estimators of μ and σ, respectively.

5.7.4 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a location and scale parameter p.d.f. $f(x; \mu, \sigma) = \frac{1}{\sigma} \psi\left(\frac{x - \mu}{\sigma} \right)$, where $-\infty < \mu < \infty, 0 < \sigma < \infty$ and $\psi(z)$ is of the form

(i) $\psi(z) = \frac{1}{2} \exp\{-|z|\}, \quad -\infty < z < \infty$ (Laplace);

(ii) $\psi(z) = 6z(1 - z), 0 \le z \le 1$. ($\beta(2, 2)$).

Determine the Pitman estimators of μ and σ for (i) and (ii).

Section 5.8

5.8.1 Let $X_1, \ldots, X_n$ be i.i.d. random variables. What are the MEEs of the parameters of

(i) $NB(\psi, \nu)$; $0 < \psi < 1, 0 < \nu < \infty$;

(ii) $G(\lambda, \nu)$; $0 < \lambda < \infty, 0 < \nu < \infty$;

(iii) $LN(\mu, \sigma^2)$; $-\infty < \mu < \infty, 0 < \sigma^2 < \infty$;

(iv) $G^{1/\beta}(\lambda, 1)$; $0 < \lambda < \infty, 0 < \beta < \infty$ (Weibull);

(v) Location and scale parameter distributions with p.d.f. $f(x; \mu, \sigma) = \frac{1}{\sigma} \psi\left(\frac{x - \mu}{\sigma} \right)$; $-\infty < \mu < \infty, 0 < \sigma < \infty$; with

(a) $\psi(z) = \frac{1}{2} \exp\{-|z|\}$, (Laplace),

(b) $\psi(z) = \frac{1}{B\left(\frac{1}{2}, \frac{\nu}{2}\right)\sqrt{\nu}} \left(1 + \frac{z^2}{\nu}\right)^{-\frac{\nu+1}{2}}$, ν known,

(c) $\psi(\bar{z}) = \frac{1}{B(\nu_1, \nu_2)} z^{\nu_1 - 1}(1 - z)^{\nu_2 - 1}$, $0 \le z \le 1$, ν_1 and ν_2 known.

5.8.2 It is a common practice to express the degree and **skewness** and **kurtosis** (peakness) of a p.d.f. by the coefficients

$$\beta_1 = \mu_3^*/(\mu_2^*)^{3/2} \quad \text{(skewness)}$$

and

$$\beta_2 = \mu_4^*/(\mu_2^*)^2 \quad \text{(kurtosis)}.$$

Provide MEEs of $\sqrt{\beta_1}$ and β_2 based on samples of n i.i.d. random variables $X_1, \ldots, X_n$.

5.8.3 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common distribution which is a mixture $\alpha G(\lambda, \nu_1) + (1-\alpha)G(\lambda, \nu_2)$, $0 < \alpha < 1$, $0 < \lambda$, $\nu_1, \nu_2 < \infty$. Construct the MEEs of α, λ, ν_1, and ν_2.

5.8.4 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common truncated normal distribution with p.d.f.

$$f(x; \mu, \sigma, \xi) = \left[n(x \mid \mu, \sigma^2) / \left(1 - \Phi\left(\left(\frac{\xi - \mu}{\sigma}\right)\right)\right)\right] \cdot I(x \geq \xi\},$$

where $\quad n(x \mid \mu, \sigma^2) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2\right\}, \quad -\infty < x < \infty.$

Determine the MEEs of (μ, σ, ξ).

Section 5.9

5.9.1 Consider the fixed-effects two-way ANOVA (Section 4.6.2). Accordingly, $X_{ijk}, i = 1, \ldots, r_1; j = 1, \ldots r_2, k = 1, \ldots, n$, are independent normal random variables, $N(\mu_{ij}, \sigma^2)$, where

$$\mu_{ij} = \mu + \tau_i^A + \tau_j^B + \tau_{ij}^{AB}; \quad i = 1, \ldots, r_1, \quad j = 1, \ldots, r_2.$$

Construct PTEs of the interaction parameters τ_{ij}^{AB} and the main-effects τ_i^A, τ_j^B $(i = 1, \ldots, r_1; j = 1, \ldots, r_2)$. [The estimation is preceded by a test of significance. If the test indicates nonsignificant effects, the estimates are zero; otherwise they are given by the value of the contrasts.]

5.9.2 Consider the linear model $\mathbf{Y} = A\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{Y}$ is an $N \times 1$ vector, A is an $N \times p$ matrix $(p < N)$ and $\boldsymbol{\beta}$ a $p \times 1$ vector. Suppose that rank $(A) = p$. Let $\boldsymbol{\beta}' = (\boldsymbol{\beta}_{(1)}, \boldsymbol{\beta}'_{(2)})$, where $\boldsymbol{\beta}_{(1)}$ is a $k \times 1$ vector, $1 \leq k < p$. Construct the LSE PTE of $\boldsymbol{\beta}(1)$. What is the expectation and the covariance of this estimator?

5.9.3 Let S_1^2 be the sample variance of n_1 i.i.d. random variables having a $N(\mu_1, \sigma_1^2)$ distribution and S_2^2 the sample variance of n_2 i.i.d. random variables having a $N(\mu_2, \sigma_2^2)$ distribution. Furthermore, S_1^2 and S_2^2 are independent. Construct PTEs of σ_1^2 and σ_2^2. What are the expectations and variances of these estimators? For which level of significance, α, these PTEs have a smaller MSE than S_1^2 and S_2^2 separately.

Section 5.10

5.10.1 What is the asymptotic distribution of the sample median M_e when the i.i.d. random variables have a distribution which is the mixture

$$0.9N(\mu, \sigma^2) + 0.1L(\mu, \sigma),$$

$L(\mu, \sigma)$ designates the Laplace distribution with location parameter μ and scale parameter σ.

5.10.2 Suppose that $X_{(1)} \leq \ldots \leq X_{(9)}$ is the order statistic of a random sample of size $n = 9$ from a rectangular distribution $R(\mu - \sigma, \mu + \sigma)$. What is the expectation and variance of

(i) the tri-mean estimator of μ;

(ii) the Gastwirth estimator of μ?

5.10.3 Simulate $N = 1000$ random samples of size $n = 20$ from the distribution of $X \sim 10 + 5t[10]$. Estimate in each sample the location parameter $\mu = 10$ by $\bar{X}$, M_e, GL, $\hat{\mu}_{.10}$ and compare the means and MSEs of these estimators over the 1000 samples.

PART IV: SOLUTIONS OF SELECTED PROBLEMS

5.2.1

(i) The m.s.s. is $(X_{(1)}, X_{(n)})$, where $X_{(1)} < \cdots < X_{(n)}$. The m.s.s. is complete, and

$$X_{(1)} \sim \theta_1 + (\theta_2 - \theta_1)U_{(1)},$$
$$X_{(n)} \sim \theta_1 + (\theta_2 - \theta_1)U_{(n)},$$

where $U_{(1)} < \cdots < U_{(n)}$ are the order statistics of n i.i.d. $R(0, 1)$ random variables. The p.d.f. of $U_{(i)}$, $i = 1, \ldots, n$ is

$$f_{U_{(i)}}(u) = \frac{n!}{(i-1)!(n-i)!} u^{i-1}(1-u)^{n-i}.$$

Thus, $U_i \sim \text{Beta}(i, n-i+1)$. Accordingly

$$E(X_{(1)}) = \theta_1 + (\theta_2 - \theta_1)\frac{1}{n+1},$$
$$E(X_{(n)}) = \theta_1 + (\theta_2 - \theta_1)\frac{n}{n+1}.$$

Solving the equations

$$\hat{\theta}_1 + (\hat{\theta}_2 - \hat{\theta}_1)\frac{1}{n+1} = X_{(1)}$$
$$\hat{\theta}_1 + (\hat{\theta}_2 - \theta_1)\frac{n}{n+1} = X_{(n)}$$

for $\hat{\theta}_1$ and $\hat{\theta}_2$, we get UMVU estimators

$$\hat{\theta}_1 = \frac{n}{n-1}X_{(1)} - \frac{1}{n-1}X_{(n)}$$

and

$$\hat{\theta}_2 = -\frac{1}{n-1}X_{(1)} + \frac{n}{n-1}X_{(n)}.$$

(ii) The covariance matrix of $\hat{\theta}_1, \hat{\theta}_2$ is

$$\frac{(\theta_2 - \theta_1)^2}{(n+1)^2(n+2)}\begin{pmatrix} n & 1 \\ 1 & n \end{pmatrix}.$$

5.2.3 The m.s.s. is $nX_{(1)}$ and $U = \sum_{i=2}^{n}(X_{(i)} - X_{(1)})$. $U \sim \frac{1}{\lambda}G(1, n-1)$. $\hat{\lambda} = \frac{n-2}{U}$;

$$E(\hat{\lambda}) = (n-2)\frac{\lambda^{n-1)}}{\Gamma(n-1)}\int_0^\infty x^{n-3}e^{-\lambda x}dx$$
$$= \lambda.$$

Thus $\hat{\lambda}$ is UMVU of λ, provided $n > 2$. $X_{(1)} \sim \mu + G(n\lambda, 1)$. Thus, $E\left\{X_{(1)} - \frac{U}{n(n-1)}\right\} = \mu$, since $E\{U\} = \frac{n-1}{\lambda}$. Thus, $\hat{\mu} = X_{(1)} -$

$\frac{U}{n(n-1)}$ is a UMVU.

$$V\{\hat{\lambda}\} = E\left\{\frac{(n-2)^2}{U^2}\right\} - \lambda^2$$

$$E\left\{\frac{1}{U^2}\right\} = \frac{\lambda^{n-1}}{\Gamma(n-1)}\int_0^\infty x^{n-4}e^{-\lambda x}dx = \frac{\lambda^2}{(n-2)(n-3)}.$$

Thus,

$$V\{\hat{\lambda}\} = \lambda^2\left(\frac{n-2}{n-3} - 1\right) = \frac{\lambda^2}{n-3},$$

provided $n > 3$. Since $X_{(1)}$ and U are independent,

$$\begin{aligned} V\{\hat{\mu}\} &= V\left\{X_{(1)} - \frac{U}{n(n-1)}\right\} \\ &= V\{X_{(1)}\} + \frac{1}{n^2(n-1)^2}V\{U\} \\ &= \frac{1}{n^2\lambda^2} + \frac{1}{n^2(n-1)^2}\frac{n-1}{\lambda^2} \\ &= \frac{1}{n^2\lambda^2}\left(1 + \frac{1}{n-1}\right) = \frac{1}{n(n-1)\lambda^2} \end{aligned}$$

for $n > 1$.

$$\begin{aligned} \mathrm{cov}(\hat{\lambda}, \hat{\mu}) &= \mathrm{cov}\left(\frac{n-2}{U}, X_{(1)} - \frac{U}{n(n-1)}\right) \\ &= -\mathrm{cov}\left(\frac{n-2}{U}, \frac{U}{n(n-1)}\right) \\ &= \frac{1}{n(n-1)}. \end{aligned}$$

5.2.4

(i) $X_1, \ldots, X_n$ i.i.d. $N(\mu, 1)$. $\bar{X}_n \sim N\left(\mu, \frac{1}{n}\right)$. Thus,

$$E\{\Phi(\lambda(n)\bar{X})\} = \Phi\left(\frac{\lambda(n)\mu}{\sqrt{1 + \frac{\lambda^2(n)}{n}}}\right).$$

Set $\frac{\lambda(n)}{\sqrt{1+\frac{\lambda^2(n)}{n}}} = 1$; $\lambda^2(n) = 1 + \frac{\lambda^2(n)}{n}$; $\lambda^2(n) = \frac{1}{1-\frac{1}{n}}$; $\lambda(n) = \frac{1}{\sqrt{1-\frac{1}{n}}}$. The UMVU estimator of $\Phi(\mu)$ is $\Phi\left(\frac{\bar{X}}{\sqrt{1-\frac{1}{n}}}\right)$.

(ii) The variance of $\Phi\left(\frac{\bar{X}}{\sqrt{1-\frac{1}{n}}}\right)$ is

$$E\left\{\Phi^2\left(\frac{\bar{X}}{\sqrt{1-\frac{1}{n}}}\right)\right\} - \Phi^2(\mu).$$

If $Y \sim N(\eta, \tau^2)$ then

$$E\{\Phi^2(Y)\} = \Phi_2\left(\frac{\mu}{\sqrt{1+\tau^2}}, \frac{\mu}{\sqrt{1+\tau^2}}; \frac{\tau^2}{1+\tau^2}\right).$$

In our case,

$$V\left\{\Phi\left(\frac{\bar{X}}{\sqrt{1-\frac{1}{n}}}\right)\right\} = \Phi_2\left(\mu, \mu; \frac{1}{n}\right) - \Phi^2(\mu).$$

5.2.6 $X_1, \ldots, X_n \sim$ i.i.d. $NB(\psi, \nu)$, $0 < \psi < \infty$ ν is *known*. The m.s.s. is $T = \sum_{i=1}^{n} X_i \sim NB(\psi, n\nu)$.

$$\begin{aligned} E\left\{\frac{T}{n\nu+T-1}\right\} &= \sum_{t=1}^{\infty} \frac{\Gamma(n\nu+t)}{t!\Gamma(n\nu)} \cdot \frac{t}{n\nu+t-1} \psi^t (1-\psi)^{n\nu} \\ &= \psi \sum_{t=1}^{\infty} \frac{\Gamma(n\nu+t-1)}{(t-1)!\Gamma(n\nu)} \psi^{t-1}(1-\psi)^{n\nu} \\ &= \psi \sum_{t=0}^{\infty} \frac{\Gamma(n\nu+t)}{t!\Gamma(n\nu)} \psi^t (1-\psi)^{n\nu} = \psi. \end{aligned}$$

5.2.8 $(\bar{X} - \xi) \sim N\left(\mu - \xi, \frac{1}{n}\right)$. Hence, $(\bar{X} - \xi)^2 \sim \frac{1}{n}\chi^2[1;\lambda]$, where $\lambda = \frac{n(\xi - \mu)^2}{2}$. Therefore,

$$E\{e^{t\chi[1;\lambda]/n}\} = e^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^j}{j!}\left(1 - \frac{2t}{n}\right)^{-\frac{1}{2}-j}$$
$$= \left(1 - 2\frac{t}{n}\right)^{-1/2}\exp\left\{-\lambda\left(1 - \frac{1}{1 - 2\frac{t}{n}}\right)\right\}.$$

Set $t = -\frac{1}{2b(n)}$, we get

$$\exp\left(-\frac{(\mu - \xi)^2}{2b(n)} \cdot \frac{1}{1 + \frac{1}{nb(n)}}\right) = e^{-\frac{1}{2}(\mu-\xi)^2}.$$

Thus, $b(n) + \frac{1}{n} = 1$ or $b(n) = 1 - \frac{1}{n} = \frac{n-1}{n}$.

5.2.9 The probability of success is $p = e^{-\Delta/\theta}$. Accordingly, $\theta = -\Delta/\log(p)$. Let $\tilde{p}_n = \sum_{i=1}^{n} J_i/n$. Thus, $\hat{p}_n = \frac{n\tilde{p}_n + 1/2}{n+1}$. The estimator of θ is $\hat{\theta}_n = -\Delta/\log(\hat{p}_n)$. The bias of $\hat{\theta}_n$ is $B(\hat{\theta}_n) = -\Delta\left(E\left\{\frac{1}{\log \hat{p}_n}\right\} - \frac{1}{\log(p)}\right)$. Let $g(\hat{p}_n) = \frac{1}{\log(\hat{p}_n)}$. Taylor expansion around p yields

$$E\{g(\hat{p}_n)\} = g(p) + g'(p)E\{\hat{p}_n - p\} + \frac{1}{2}g''(p) \cdot$$
$$\cdot\ E\{(\hat{p}_n - p)^2\} + \frac{1}{6}g^{(3)}(p)E\{(p^* - p)^3\},$$

where $|p^* - p| < |\hat{p}_n - p|$. Moreover,

$$g'(p) = -\frac{1}{p\log^2(p)},$$
$$g''(p) = \frac{2 + \log(p)}{p^2\log^3(p)},$$
$$g^{(3)}(p) = -2\frac{3\log(p) + 3 + \log^2(p)}{p^3\log 4(p)}.$$

Furthermore,

$$E\{\hat{p}_n - p) = -\frac{2p-1}{2(n+1)},$$
$$E\{(\hat{p}_n - p)^2\} = \frac{(n-1)p(1-p)+1/4}{(n+1)^2}.$$

Hence,

$$B(\hat{\theta}_n) = \frac{2p-1}{2(n+1)p\log^2(p)} + \frac{1}{2}\frac{(n-1)p(1-p)+1/4}{(n+1)^2} \cdot$$
$$\cdot \quad \frac{2+\log(p)}{p^2\log^3(p)} + o\left(\frac{1}{n}\right), \quad \text{as } n \to \infty.$$

5.2.17 $X_1, \ldots, X_n \sim G(\lambda, \nu)$, $\nu \geq 3$ fixed.

(i) $\bar{X} \sim \frac{1}{n}G(\lambda, n\nu)$. Hence

$$E\left\{\frac{1}{\bar{X}^2}\right\} = \frac{n^2\lambda^{n\nu}}{\Gamma(n\nu)}\int_0^\infty x^{n\nu-3}e^{-\lambda x}dx$$
$$= \frac{\lambda^2 n^2}{(n\nu-1)(n\nu-2)}.$$

The UMVU of λ^2 is $\hat{\lambda}^2 = \frac{(n\nu-1)(n\nu-2)}{n^2}\cdot\frac{1}{\bar{X}^2}$.

(ii) $V\{\hat{\lambda}^2\} = \lambda^4\frac{2(2n\nu-5)}{(n\nu-3)(n\nu-4)}$.

(iii) $I_n(\lambda) = \frac{n\nu}{\lambda^2}$. The Cramér–Rao lower bound for the variance of the UMVU of λ^2 is

$$\text{CRLB} = \frac{4\lambda^2\cdot\lambda^2}{n\nu} = \frac{4\lambda^4}{n\nu}.$$

(iv) $l'(\lambda) = \frac{n\nu}{\lambda} - T$, $l''(\lambda) = -\frac{n\nu}{\lambda^2}$.

$$V = \frac{n\nu}{\lambda^2}\begin{bmatrix} 1 & 0 \\ 0 & \frac{n\nu}{\lambda^2} \end{bmatrix}.$$

$w(\lambda) = \lambda^2$, $w'(\lambda) = 2\lambda$, $w''(\lambda) = 2$.

$$(2\lambda, 2)V^{-1}\begin{pmatrix} 2\lambda \\ 2 \end{pmatrix} = 4(\lambda, 1)V^{-1}\begin{pmatrix} \lambda \\ 1 \end{pmatrix}.$$

The second order BLB is

$$\frac{4\lambda^4(n\nu+1)}{(n\nu)^2},$$

and third and fourth order BLB do not exist.

5.3.3 This is continuation of Example 5.8. We have a sample of size n of vectors (X, Y), where $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\mathbf{0}, \sigma^2\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$. We have seen that $V\{\hat{\rho}\} = \frac{1-\rho^2}{n-2}$. We derive now the variance of $\hat{\sigma}^2 = (Q_X + Q_Y)/2n$, where $Q_X = \sum_{i=1}^{n} X_i^2$ and $Q_Y = \sum_{i=1}^{n} Y_i^2$. Note that $Q_Y \mid Q_X \sim \sigma^2(1-\rho^2)\chi^2\left[n; \frac{\rho^2}{2(1-\rho^2)}Q_Y\right]$. Hence,

$$E(Q_Y \mid Q_X) = \sigma^2 n(1-\rho^2) + \sigma^2\rho^2 Q_X$$

and

$$V(Q_Y \mid Q_X) = \sigma^4(1-\rho^2)^2\left(2n + 4\frac{\rho^2 Q_X}{1-\rho^2}\right).$$

It follows that

$$V(\hat{\sigma}^2) = \frac{2\sigma^4(1-\rho^2)}{n}.$$

Finally, since $(Q_X + Q_Y, P_{XY})$ is a *complete* sufficient statistic, and since $\hat{\rho}$ is invariant with respect to translations and change of scale, Basu's Theorem implies that $\hat{\sigma}^2$ and $\hat{\rho}$ are independent. Hence, the variance–covariance matrix of $(\hat{\sigma}^2, \hat{\rho})$ is

$$V = \frac{1}{n}\begin{pmatrix} 2\sigma^4(1-\rho^2) & 0 \\ 0 & \frac{1-\rho^2}{1-\frac{2}{n}} \end{pmatrix}.$$

Thus, the efficiency of $(\hat{\sigma}^2, \hat{\rho})$, according to (5.3.13), is

$$\begin{aligned} \text{eff.} &= \frac{\sigma^4(1+\rho^2)(1-\rho^2)^2 - \sigma^4\rho^2(1-\rho^2)^2}{2\sigma^4(1-\rho^2)^2/(1-\frac{1}{n})} \\ &= \frac{1}{2}\left(1-\frac{1}{n}\right) = \frac{1}{2} + O\left(\frac{1}{n}\right), \quad \text{as } n \to \infty. \end{aligned}$$

5.4.3 $X_1, \ldots, X_5$ are i.i.d. having a Laplace distribution with p.d.f. $\frac{1}{\sigma}\phi\left(\frac{X-\mu}{\sigma}\right)$, where $-\infty < \mu < \infty$, $0 < \sigma < \infty$, and $\phi(x) = \frac{1}{2}e^{-|x|}$, $-\infty < x < \infty$. The standard c.d.f. ($\mu = 0, \sigma = 1$) is

$$F(x) = \begin{cases} \frac{1}{2}e^{x}, & -\infty < x < 0, \\ 1 - \frac{1}{2}e^{-x}, & 0 \le x < \infty. \end{cases}$$

Let $X_{(1)} < X_{(2)} < X_{(3)} < X_{(4)} < X_{(5)}$ be the order statistic. Note that $M_e = X_{(3)}$. Also $X_{(i)} = \mu + \sigma U_{(i)}$, $i = 1, \ldots, 5$, where $U_{(i)}$ are the order statistic from a standard distribution.

The densities of $U_{(1)}, U_{(3)}, U_{(5)}$ are

$$p_{(1)}(u) = \frac{5}{2}\exp(-|u|)(1 - F(u))^4, \quad -\infty < u < \infty,$$

$$p_{(3)}(u) = 15\exp(-|u|)(F(u))^2(1 - F(u))^2, \quad -\infty < u < \infty,$$

$$p_{(5)}(u) = \frac{5}{2}\exp(-|u|)(F(u))^4, \quad -\infty < u < \infty.$$

Since $\phi(u)$ is symmetric around $u = 0$, $F(-u) = 1 - F(u)$. Thus, $p_{(3)}(u)$ is symmetric around $u = 0$, and $U_{(1)} \sim -U_{(5)}$. It follows that $\alpha_1 = E\{U_{(1)}\} = -\alpha_5$ and $\alpha_3 = 0$. Moreover,

$$\begin{aligned} \alpha_1 &= \frac{5}{2}\int_{-\infty}^{0} ue^{u}\left(1 - \frac{1}{2}e^{u}\right)^4 du + \frac{5}{2}\int_{0}^{\infty} ue^{-u}\left(\frac{1}{2}e^{-u}\right)^4 du \\ &= -1.58854. \end{aligned}$$

Accordingly, $\alpha' = (-1.58854, 0, 1.58854)$, $V\{U_{(1)}\} = V\{U_{(5)}\} = 1.470256$, $V\{U_{(3)}\} = 0.35118$.

$$\begin{aligned} \mathrm{cov}(U_{(1)}, U_{(3)}) &= E\{U_{(1)}U_{(3)}\} \\ &= \frac{5!}{8}\int_{-\infty}^{\infty} xe^{-|x|}\int_{x}^{\infty} y\exp(-|y|)(F(y) - F(x)) \cdot \\ &\quad \cdot \ (1 - F(y))^2 dy\, dx = 0.264028 \\ \mathrm{cov}(U_{(1)}, U_{(5)}) &= \mathrm{cov}(U_{(1)}, -U_{(1)}) = -V\{U_{(1)}\} = -1.470256 \\ \mathrm{cov}(U_{(3)}, U_{(5)}) &= E\{U_{(3)}, -U_{(1)}\} = -0.264028. \end{aligned}$$

Thus,

$$V = \begin{bmatrix} 1.47026 & 0.26403 & -1.47026 \\ 0.26403 & 0.35118 & -0.26403 \\ -1.47026 & -0.26403 & 1.47026 \end{bmatrix}.$$

Let

$$A = \begin{bmatrix} 1 & -1.58854 \\ 1 & 0 \\ 1 & 1.58854 \end{bmatrix}.$$

The matrix V is singular, since $U_{(5)} = -U_{(1)}$. We take the generalized inverse

$$V^- = \begin{bmatrix} 0.7863 & -0.5912 & 0 \\ -0.5912 & 3.2920 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

We then compute

$$(A'V^-A)^{-1}A'V^{-1} = \begin{pmatrix} 0 & 1 & 0 \\ -0.6295 & 0.6295 & 0 \end{pmatrix}.$$

According to this, the estimator of μ is $\hat{\mu} = X_{(3)}$, and that of σ is $\hat{\sigma} = 0.63X_{(3)} - 0.63X_{(1)}$. These estimators are not BLUE. Take the ordinary LSE given by

$$\begin{pmatrix} \hat{\mu} \\ \hat{\sigma} \end{pmatrix} = (A'A)^{-1}A' \begin{pmatrix} X_{(1)} \\ X_{(3)} \\ X_{(5)} \end{pmatrix} = \begin{pmatrix} \frac{1}{3}(X_{(1)} + X_{(3)} + X_{(5)}) \\ -0.31775X_{(1)} + 2.31475X_{(5)} \end{pmatrix}$$

then the variance covariance matrix of these estimators is

$$\begin{pmatrix} 0.0391 & -0.0554 \\ -0.0554 & 0.5827 \end{pmatrix}.$$

Thus, $V\{\hat{\mu}\} = 0.0391 < V\{\hat{\mu}\} = 0.3512$, and $V\{\hat{\sigma}\} = 0.5827 > V\{\hat{\sigma}\} = 0.5125$. Due to the fact that V is singular, $\hat{\mu}$ is a better estimator than $\hat{\mu}$.

5.6.6 $(x_i, T_i), i = 1, \ldots, n$. $T_i \sim (\alpha + \beta x_i)G(1, 1), i = 1, \ldots, n$.

$$L(\alpha, \beta; \mathbf{x}, \mathbf{T}) = \prod_{i=1}^{n} \frac{1}{\alpha + \beta x_i} \exp\left\{-\frac{1}{\alpha + \beta X_i} T_i\right\}$$

$$l(\alpha, \beta) = \log L(\alpha, \beta; \mathbf{X}, \mathbf{T}) = -\sum_{i=1}^{n} \left(\log(\alpha + \beta X_i) + \frac{1}{\alpha + \beta X_i} T_i\right)$$

$$-\frac{\partial}{\partial \alpha} l(\alpha, \beta) = \sum_{i=1}^{n} \frac{1}{\alpha + \beta X_i} - \sum_{i=1}^{n} \frac{1}{(\alpha + \beta X_i)^2} T_i$$

$$-\frac{\partial}{\partial \beta} l(\alpha, \beta) = \sum_{i=1}^{n} \frac{X_i}{\alpha + \beta X_i} - \sum \frac{X_i T_i}{(\alpha + \beta X_i)^2}.$$

The MLEs of α and β are the roots of the equations:

(I) $\displaystyle\sum_{i=1}^{n}\frac{1}{\alpha+\beta X_i}=\sum_{i=1}^{n}\frac{T_i}{(\alpha+\beta X_i)^2}$

(II) $\displaystyle\sum_{i=1}^{n}\frac{X_i}{\alpha+\beta X_i}=\sum_{i=1}^{n}\frac{T_iX_i}{(\alpha+\beta X_i)^2}.$

The Newton–Raphson method for approximating numerically (α, β) is

$$G_1(\alpha,\beta)=\sum_{i=1}^{n}\frac{T_i-(\alpha+\beta X_i)}{(\alpha+\beta X_i)^2},$$

$$G_2(\alpha,\beta)=\sum_{i=1}^{n}\frac{T_iX_i-X_i(\alpha+\beta X_i)}{(\alpha+\beta X_i)^2}.$$

The matrix

$$D(\alpha,\beta)=\begin{pmatrix}\dfrac{\partial G_1}{\partial\alpha} & \dfrac{\partial G_1}{\partial\beta}\\[2ex] \dfrac{\partial G_2}{\partial\alpha} & \dfrac{\partial G_2}{\partial\beta}\end{pmatrix}=\begin{pmatrix}\Sigma w_i & \Sigma w_iX_i\\ \Sigma w_iX_i & \Sigma w_iX_i^2\end{pmatrix},$$

where $w_i=\dfrac{\alpha+\beta X_i-2T_i}{(\alpha+\beta X_i)^3}$.

$$\begin{aligned}|D(\alpha,\beta)| &= (\Sigma w_i)(\Sigma w_iX_i^2)-(\Sigma w_iX_i)^2\\ &= (\Sigma w_i)\left(\sum_{i=1}^{n}w_i\left(X_i-\frac{\Sigma w_iX_i}{\Sigma w_i}\right)^2\right).\end{aligned}$$

We assume that $|D(\alpha,\beta)|>0$ in each iteration. Starting with (α_1,β_1), we get the following after the lth iteration:

$$\begin{pmatrix}\alpha_{l+1}\\ \beta_{l+1}\end{pmatrix}=\begin{pmatrix}\alpha_l\\ \beta_l\end{pmatrix}-(D(\alpha_l,\beta_l))^{-1}\begin{pmatrix}G_1(\alpha_l,\beta_l)\\ G_2(\alpha_l,\beta_l)\end{pmatrix}.$$

The LSE initial solution is

$$\hat\beta_1=\frac{\Sigma T_i(X_i-\bar X)}{\Sigma(X_i-\bar X)^2},$$

$$\hat\alpha_1=\bar T-\hat\beta_1\bar X.$$

5.6.9 $X^2 = \sum_{i=1}^{k} \frac{(J_i - nP_i(\theta))^2}{nP_i(\theta)}$. For the Hardy–Weinberg model,

$$X^2(\theta) = \sum_{i=1}^{3} \frac{(J_i - nP_i(\theta))^2}{nP_i(\theta)}$$
$$= \frac{(J_1 - n\theta^2)^2}{n\theta^2} + \frac{(J_2 - 2n\theta(1-\theta))^2}{2n\theta(1-\theta)} + \frac{(J_3 - n(1-\theta)^2)^2}{n(1-\theta)^2}.$$

$$\frac{d}{d\theta} X^2(\theta) = \frac{N(\theta)}{2n\theta^3(1-\theta)^3},$$

where

$$N(\theta) = 2(2J_1^2 + 2(n - J_1 - J_2)^2 - J_2^2)\theta^3$$
$$- 3(4J_1^2 - J_2^2)\theta^2 + (12J_1^2 - J_2^2)\theta - 4J_1^2.$$

Note that $J_3 = n - J_1 - J_2$. Thus, the MCE of θ is the root of $N(\theta) \equiv 0$.

5.7.1 $X_1, \ldots, X_n$ are i.i.d. $R(\theta, \theta)$, $0 < \theta < \infty$.

(i) $cX \sim R(0, c\theta)$ for all $0 < c < \infty$. Thus, the model is preserved under the scale transformation $\mathcal{G}$.

(ii) The m.s.s. is $X_{(n)}$. Thus, an equivariant estimator of θ is

$$\hat{\theta}(X_{(n)}) = X_{(n)}\psi(1).$$

Consider the following invariant loss function:

$$L(\hat{\theta}, \theta) = \frac{(\hat{\theta} - \theta)^2}{\theta^2}.$$

There is only one orbit of $\bar{\mathcal{G}}$ in Θ. Thus, find ψ to minimize

$$Q(\psi) = E\{(\psi X_{(n)} - 1)^2\} = \psi^2 E(X_{(n)}^2) - 2\psi E(X_n) + 1,$$
$$Q'(\psi) = 2\psi E(X_{(n)}^2) - 2E(X_{(n)}).$$

Thus, $\psi^0 = \frac{E(X_{(n)})}{E(X_{(n)}^2)}$ computed at $\theta = 1$. Note that under $\theta = 1$, $X_{(n)} \sim$ Beta$(n, 1)$ $E(X_{(n)}) = \frac{n}{n+1}$, $E(X_{(n)}^2) = \frac{n}{n+2}$, $\psi^0 = \frac{n+2}{n+1}$.

5.7.3

(i) The Pitman estimator of the location parameter in the normal case, according to Equation (5.7.33), is

$$\hat{\mu} = X_{(1)} - \frac{\int_{-\infty}^{\infty} u e^{-\frac{1}{2}(u^2+\sum_{i=2}^{n}(Y_{(i)}+u)^2)} du}{\int_{-\infty}^{\infty} e^{-\frac{1}{2}(u^2+\sum_{i=2}^{n}(Y_{(i)}+u)^2)} du}$$

$$u^2 + \sum_{i=2}^{n}(Y_{(i)}+u)^2 = u^2 + \sum_{i=2}^{n} u^2 + 2u\sum_{i=2}^{n} Y_{(i)}$$

$$+ \sum_{i=2}^{n} Y_{(i)}^2 = n\left(u^2 + 2u\frac{1}{n}\sum_{i=2}^{n} Y_{(i)} + \left(\frac{1}{n}\sum_{i=2}^{n} Y_{(i)}\right)^2\right)$$

$$-\frac{1}{n}\left(\sum_{i=2}^{n} Y_{(i)}\right)^2 + \sum_{i=2}^{n} Y_{(i)}^2.$$

Moreover,

$$\sqrt{n}\int_{-\infty}^{\infty} \exp\left(-\frac{n}{2}\left(u + \frac{1}{n}\sum_{i=2}^{n} Y_{(i)}\right)^2\right) du = \sqrt{2\pi}$$

and

$$\sqrt{n}\int_{-\infty}^{\infty} u \exp\left(-\frac{n}{2}\left(u + \frac{1}{n}\sum_{i=2}^{n} Y_{(i)}\right)^2\right) du = -\sqrt{2\pi}\frac{1}{n}\sum_{i=2}^{n} Y_{(i)}.$$

The other terms are cancelled. Thus,

$$\hat{\mu} = X_{(1)} + \frac{1}{n}\sum_{i=2}^{n}(X_{(i)} - X_{(1)})$$

$$= \frac{1}{n}X_{(1)} + \frac{1}{n}\sum_{i=2}^{n} X_{(i)} = \bar{X}_n.$$

This is the best equivariant estimator for the squared error loss.

(ii) For estimating σ in the normal case, let $(\bar{X}, S)$ be the m.s.s., where $S = \left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{1/2}$. An equivariant estimator of σ, for the translation-scale group, is

$$\hat{\sigma} = S\psi\left(\frac{X_1 - \bar{X}}{S}, \ldots, \frac{X_n - \bar{X}}{S}\right).$$

$\mathbf{u} = \left(\frac{X_1 - \bar{X}}{S}, \ldots, \frac{X_n - \bar{X}}{S}\right)$. By Basu's Theorem, S is independent of $\mathbf{u}$. We find $\psi(\mathbf{u})$ by minimizing $E\{(S\psi - 1)^2\}$ under $\sigma = 1$. $E(S^2 \mid \psi) = E_1\{S^2\} = 1$ and $E_1\{S \mid \psi\} = E_1\{S\}$. Here, $\psi^0 = E_1\{S\}/E_1\{S^2\}$.

CHAPTER 6

Confidence and Tolerance Intervals

PART I: THEORY

6.1 GENERAL INTRODUCTION

When θ is an unknown parameter and an estimator $\hat{\theta}$ is applied, the precision of the estimator $\hat{\theta}$ can be stated in terms of its sampling distribution. With the aid of the sampling distribution of an estimator we can determine the probability that the estimator θ lies within a prescribed interval around the true value of the parameter θ. Such a probability is called **confidence (or coverage) probability**. Conversely, for a preassigned **confidence level**, we can determine an interval whose limits depend on the observed sample values, and whose coverage probability is not smaller than the prescribed confidence level, for all θ. Such an interval is called a **confidence interval**. In the simple example of estimating the parameters of a normal distribution $N(\mu, \sigma^2)$, a minimal sufficient statistic for a sample of size n is $(\bar{X}_n, S_n^2)$. We wish to determine an interval $(\underline{\mu}(\bar{X}_n, S_n^2), \bar{\mu}(\bar{X}_n, S_n^2))$ such that

$$P_{\mu,\sigma}\{\underline{\mu}(\bar{X}_n, S_n^2) \leq \mu \leq \bar{\mu}(\bar{X}_n, S_n^2)\} \geq 1 - \alpha, \tag{6.1.1}$$

for all μ, σ. The prescribed confidence level is $1 - \alpha$ and the confidence interval is $(\underline{\mu}, \bar{\mu})$. It is easy to prove that if we choose the functions

$$\begin{aligned} \underline{\mu}(\bar{X}_n, S_n^2) &= \bar{X}_n - t_{1-\alpha/2}[n-1]\frac{S_n}{\sqrt{n}}, \\ \bar{\mu}(\bar{X}_n, S_n^2) &= \bar{X}_n + t_{1-\alpha/2}[n-1]\frac{S_n}{\sqrt{n}}, \end{aligned} \tag{6.1.2}$$

then (6.1.1) is satisfied. The two limits of the confidence interval $(\underline{\mu}, \bar{\mu})$ are called the **lower** and **upper confidence limits**. Confidence limits for the variance σ^2 in

Examples and Problems in Mathematical Statistics, First Edition. Shelemyahu Zacks.

the normal case can be obtained from the sampling distribution of S_n^2. Indeed, since $S_n^2 \sim \frac{\sigma^2}{n-1}\chi^2[n-1]$. The lower and upper confidence limits for σ^2 are given by

$$\begin{aligned} \underline{\sigma}^2 &= \frac{(n-1)S_n^2}{\chi^2_{1-\alpha/2}[n-1]}, \\ \bar{\sigma}^2 &= \frac{(n-1)S_n^2}{\chi^2_{\alpha/2}[n-1]}. \end{aligned} \tag{6.1.3}$$

A general method to derive confidence intervals in parametric cases is given in Section 6.2. The theory of optimal confidence intervals is developed in Section 6.3 in parallel to the theory of optimal testing of hypotheses. The theory of **tolerance intervals** and **regions** is discussed in Section 6.4. Tolerance intervals are estimated intervals of a prescribed probability content according to the unknown parent distribution. One sided tolerance intervals are often applied in engineering designs and screening processes as illustrated in Example 6.1.

Distribution free methods, based on the properties of order statistics, are developed in Section 6.5. These methods yield tolerance intervals for all distribution functions having some general properties (log-convex for example). Section 6.6 is devoted to the problem of determining simultaneous confidence intervals for several parameters. In Section 6.7, we discuss two-stage and sequential sampling to obtain fixed-width confidence intervals.

6.2 THE CONSTRUCTION OF CONFIDENCE INTERVALS

We discuss here a more systematic method of constructing confidence intervals.

Let $\mathcal{F} = \{F(x;\theta), \theta \in \Theta\}$ be a parametric family of d.f.s. The parameter θ is real or vector valued. Given the observed value of X, we construct a set $S(X)$ in Θ such that

$$P_\theta\{\theta \in S(X)\} \geq 1-\alpha, \quad \text{for all } \theta. \tag{6.2.1}$$

$S(X)$ is called a **confidence region** for θ at level of confidence $1-\alpha$. Note that the set $S(X)$ is a random set, since it is a function of X. For example, consider the multinormal $N(\boldsymbol{\theta}, I)$ case. We know that $(\mathbf{X}-\boldsymbol{\theta})'(\mathbf{X}-\boldsymbol{\theta})$ is distributed like $\chi^2[k]$, where k is the dimension of $\mathbf{X}$. Thus, define

$$S(\mathbf{X}) = \{\boldsymbol{\theta} : (\mathbf{X}-\boldsymbol{\theta})'(\mathbf{X}-\boldsymbol{\theta}) \leq \chi^2_{1-\alpha}[k]\}. \tag{6.2.2}$$

It follows that, for all $\boldsymbol{\theta}$,

$$P_\theta\{\boldsymbol{\theta} \in S(\mathbf{X})\} = P\{(\mathbf{X}-\boldsymbol{\theta})'(\mathbf{X}-\boldsymbol{\theta}) \leq \chi^2_{1-\alpha}[k]\} = 1-\alpha. \tag{6.2.3}$$

Accordingly, $S(X)$ is a confidence region. Note that if the problem, in this multinormal case, is to test the simple hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against the composite alternative $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ we would apply the test statistic

$$T(\boldsymbol{\theta}_0) = (\mathbf{X} - \boldsymbol{\theta}_0)'(\mathbf{X} - \boldsymbol{\theta}_0), \tag{6.2.4}$$

and reject H_0 whenever $T(\theta_0) \geq \chi^2_{1-\alpha}[k]$. This test has size α. If we define the acceptance region for H_0 as the set

$$A(\boldsymbol{\theta}_0) = \{\mathbf{X}; (\mathbf{X} - \boldsymbol{\theta}_0)'(\mathbf{X} - \boldsymbol{\theta}_0) \leq \chi^2_{1-\alpha}[k]\}, \tag{6.2.5}$$

then H_0 is accepted if $\mathbf{X} \in A(\boldsymbol{\theta}_0)$. The structures of $A(\boldsymbol{\theta}_0)$ and $S(\mathbf{X})$ are similar. In $A(\theta_0)$, we fix θ at θ_0 and vary $\mathbf{X}$, while in $S(X)$ we fix $\mathbf{X}$ and vary $\boldsymbol{\theta}$. Thus, let $\mathcal{A} = \{A(\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$ be a family of acceptance regions for the above testing problem, when θ varies over all the points in Θ. Such a family induces a family of confidence sets $\mathcal{S} = \{S(\mathbf{X}) : \mathbf{X} \in \mathcal{X}\}$ according to the relation

$$S(\mathbf{X}) = \{\boldsymbol{\theta} : \mathbf{X} \in A(\boldsymbol{\theta}); A(\boldsymbol{\theta}) \in \mathcal{A}\}. \tag{6.2.6}$$

In such a manner, we construct generally confidence regions (or intervals). We first construct a family of acceptance regions, $\mathcal{A}$ for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ at level of significance α. From this family, we construct the dual family $\mathcal{S}$ of confidence regions. We remark here that in cases of a real parameter θ we can consider one-sided hypotheses $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$; or $H_0 : \theta \geq \theta_0$ against $H_1 : \theta < \theta_0$. The corresponding families of acceptance regions will induce families of one-sided confidence intervals $(-\infty, \bar{\theta}(\mathbf{X}))$ or $(\underline{\theta}(\mathbf{X}), \infty)$, respectively.

6.3 OPTIMAL CONFIDENCE INTERVALS

In the previous example, we have seen two different families of lower confidence intervals, one of which was obviously inefficient. We introduce now the theory of uniformly most accurate (UMA) confidence intervals. According to this theory, the family of lower confidence intervals $\underline{\theta}_\alpha$ in the above example is optimal.

Definition. *A lower confidence limit for θ, $\underline{\theta}(\mathbf{X})$ is called UMA if, given any other lower confidence limit $\underline{\theta}^*(\mathbf{X})$,*

$$P_\theta\{\underline{\theta}(\mathbf{X}) \leq \theta'\} \leq P_\theta\{\theta^*(\mathbf{X}) \leq \theta'\} \tag{6.3.1}$$

for all $\theta' < \theta$, and all θ.

That is, although both the $\underline{\theta}(\mathbf{X})$ and $\underline{\theta}^*(\mathbf{X})$ are smaller than θ with confidence probability $(1 - \alpha)$, the probability is larger that the UMA limit $\underline{\theta}(\mathbf{X})$ is closer to the true value θ than that of $\underline{\theta}^*(\mathbf{X})$. Whenever a size α uniformly most powerful

(UMP) test exists for testing the hypothesis $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$, then a UMA $(1 - \alpha)$-lower confidence limit exists. Moreover, one can obtain the UMA lower confidence limit from the UMP test function according to relationship (6.2.6). The proof of this is very simple and left to the reader. Thus, as proven in Section 4.3, if the family of d.f.s $\mathcal{F}$ is a one-parameter MLR family, the UMP test of size α, of $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_1$ is of the form

$$\phi^0(T_n) = \begin{cases} 1, & \text{if } T_n > C_\alpha(\theta_0), \\ \gamma_\alpha, & \text{if } T_n = C_\alpha(\theta_0), \\ 0, & \text{if otherwise,} \end{cases} \tag{6.3.2}$$

where T_n is the minimal sufficient statistic. Accordingly, if T_n is a continuous random variable, the family of acceptance intervals is

$$\mathcal{A} = \{(-\infty, C_\alpha(\theta)), \theta \in \Theta\}. \tag{6.3.3}$$

The corresponding family of $(1 - \alpha)$-lower confidence limits is

$$\mathcal{S} = \{(\underline{\theta}_\alpha, \infty); T_n = C_\alpha(\underline{\theta}_\alpha), \quad \theta \in \Theta\}. \tag{6.3.4}$$

In the **discrete monotone likelihood ratio (MLR)** case, we reduce the problem to that of a continuous MLR by randomization, as specified in (6.3.2). Let T_n be the minimal sufficient statistic and, without loss of generality, assume that T_n assumes only the nonnegative integers. Let $H_n(t; \theta)$ be the cumulative distribution function (c.d.f.) of T_n under θ. We have seen in Chapter 4 that the critical level of the test (6.3.2) is

$$C_\alpha(\theta_0) = \text{least nonnegative integer } t \text{ such that } H_n(t; \theta_0) \geq 1 - \alpha. \tag{6.3.5}$$

Moreover, since the distributions are MLR, $C_\alpha(\theta)$ is a nondecreasing function of θ. In the continuous case, we determined the lower confidence limit $\underline{\theta}_\alpha$ as the root, θ, of the equation $T_n = C_\alpha(\theta)$. In the discrete case, we determine $\underline{\theta}_\alpha$ as the root, θ, of the equation

$$H_n(T_n - 1; \theta) + R[H_n(T_n; \theta) - H_n(T_n - 1; \theta)] = 1 - \alpha, \tag{6.3.6}$$

where R is a random variable independent of T_n and having a rectangular distribution $R(0, 1)$. We can express Equation (6.3.6) in the form

$$RH_n(T_n; \underline{\theta}_\alpha) + (1 - R)H_n(T_n - 1; \underline{\theta}_\alpha) = 1 - \alpha. \tag{6.3.7}$$

If UMP tests do not exist we cannot construct UMA confidence limits. However, we can define UMA-unbiased or UMA-invariant confidence limits and apply the

theory of testing hypotheses to construct such limits. Two-sided confidence intervals $(\underline{\theta}_\alpha(\mathbf{X}), \bar{\theta}_\alpha(\mathbf{X}))$ should satisfy the requirement

$$P_\theta\{\underline{\theta}_\alpha(\mathbf{X}) \leq \theta \leq \bar{\theta}_\alpha(\mathbf{X})\} \geq 1-\alpha, \quad \text{for all } \theta. \tag{6.3.8}$$

A two-sided $(1-\alpha)$ confidence interval $(\underline{\theta}_\alpha(\mathbf{X}), \bar{\theta}_\alpha(\mathbf{X}))$ is called UMA if, subject to (6.3.8), it minimizes the coverage probabilities

$$P_\theta\{\underline{\theta}(\mathbf{X}) \leq \theta_1 \leq \bar{\theta}_\alpha(\mathbf{X})\}, \quad \text{for all } \theta_1 \neq \theta. \tag{6.3.9}$$

In order to obtain UMA two-sided confidence intervals, we should construct a UMP test of size α of the hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Such a test generally does not exist. However, we can construct a UMP-unbiased (UMPU) test of such hypotheses (in cases of exponential families) and derive then the corresponding confidence intervals.

A confidence interval of level $1-\alpha$ is called **unbiased** if, subject to (6.3.8), it satisfies

$$P_\theta\{\underline{\theta}_\alpha(\mathbf{X}) \leq \theta_1 \leq \bar{\theta}_\alpha(\mathbf{X})\} \leq 1-\alpha, \quad \text{for all } \theta_1 \neq \theta. \tag{6.3.10}$$

Confidence intervals constructed on the basis of UMPU tests are UMAU (uniformly most accurate unbiased) ones.

6.4 TOLERANCE INTERVALS

Tolerance intervals can be described in general terms as estimated prediction intervals for future realization(s) of the observed random variables. In Example 6.1, we discuss such an estimation problem and illustrate a possible solution. Consider a sequence $X_1, X_2, \ldots$ of independent and identically distributed (i.i.d.) random variables having a common distribution $F(x;\theta)$, $\theta \in \Theta$. A p-**content prediction interval** for a possible realization of X, when θ is known, is an interval $(l_p(\theta), u_p(\theta))$ such that $P_\theta[X \in (l_p(\theta), u_p(\theta))] \geq p$. Such two-sided prediction intervals are not uniquely defined. Indeed, if $F^{-1}(p;\theta)$ is the pth quantile of $F(x;\theta)$ then for every $0 \leq \epsilon \leq 1$, $l_p = F^{-1}(\epsilon(1-p);\theta)$ and $u_p = F^{-1}(1-(1-\epsilon)(1-p);\theta)$ are lower and upper limits of a p-content prediction interval. Thus, p-content two-sided prediction intervals should be defined more definitely, by imposing further requirement on the location of the interval. This is, generally, done according to the specific problem under consideration. We will restrict attention here to one-sided prediction intervals of the form $(-\infty, F^{-1}(p;\theta)]$ or $[F^{-1}(1-p;\theta), \infty)$.

When θ is unknown the limits of the prediction intervals are estimated. In this section, we develop the theory of such parametric estimation. The estimated prediction intervals are called **tolerance intervals**. Two types of tolerance intervals are discussed in the literature: p-**content tolerance intervals** (see Guenther, 1971), which are called also **mean tolerance predictors** (see Aitchison and Dunsmore, 1975);

and $(1-\alpha)$ **level p-content intervals**, also called **guaranteed coverage tolerance intervals** (Aitchison and Dunsmore, 1975; Guttman, 1970). p-Content one-sided tolerance intervals, say $(-\infty, L_p(\mathbf{X}_n))$, are determined on the basis of n sample values $\mathbf{X}_n = (X_1, \ldots, X_n)$ so that, if Y has the $F(x;\theta)$ distribution then

$$P_\theta[Y \leq L_p(\mathbf{X}_n)] \geq p, \quad \text{for all } \theta. \tag{6.4.1}$$

Note that

$$P_\theta[Y \leq L_p(\mathbf{X}_n)] = E_\theta\{P_\theta[Y \leq L_p(\mathbf{X}_n) \mid \mathbf{X}_n]\}. \tag{6.4.2}$$

Thus, given the value of $\mathbf{X}_n$, the upper tolerance limit $L_p(\mathbf{X}_n)$ is determined so that the expected probability content of the interval $(-\infty, L_p(\mathbf{X}_n)]$ will be p. The $(p, 1-\alpha)$ guaranteed coverage one-sided tolerance interval $(-\infty, L_{\alpha,p}(\mathbf{X}_n))$ are determined so that

$$P_\theta[F^{-1}(p;\theta) \leq L_{\alpha,p}(\mathbf{X}_n)] \geq 1-\alpha, \tag{6.4.3}$$

for all θ. In other words, $L_{\alpha,p}(\mathbf{X}_n)$ is a $(1-\alpha)$-upper confidence limit for the pth quantile of the distribution $F(x;\theta)$. Or, with confidence level $(1-\alpha)$, we can state that the expected proportion of future observations not exceeding $L_{\alpha,p}(X_n)$ is at least p. $(p, 1-\alpha)$-upper tolerance limits can be obtained in cases of MLR parametric families by substituting the $(1-\alpha)$-upper confidence limit $\bar{\theta}_\alpha$ of θ in the formula of $F^{-1}(p;\theta)$. Indeed, if $\mathcal{F} = \{F(x;\theta); \theta \in \Theta\}$ is a family depending on a real parameter θ, and $\mathcal{F}$ is MLR with respect to X, then the pth quantile, $F^{-1}(p;\theta)$, is an increasing function of θ, for each $0 < p < 1$. Thus, a one-sided p-content, $(1-\alpha)$-level tolerance interval is given by

$$L_{\alpha,p}(\mathbf{X}_n) = F^{-1}(p; \bar{\theta}_\alpha(\mathbf{X}_n)). \tag{6.4.4}$$

Moreover, if the upper confidence limit $\bar{\theta}_\alpha(\mathbf{X}_n)$ is UMA then the corresponding tolerance limit is a UMA upper confidence limit of $F^{-1}(p;\theta)$. For this reason such a tolerance interval is called UMA. For more details, see Zacks (1971, p. 519).

In Example 6.1, we derive the $(\beta, 1-\alpha)$ guaranteed lower tolerance limit for the log-normal distribution. It is very simple in that case to determine the β-content lower tolerance interval. Indeed, if $(\bar{Y}_n, S_n^2)$ are the sample mean and variance of the corresponding normal variables $Y_i = \log X_i$ $(i = 1, \ldots, n)$ then

$$l(\bar{Y}_n, S_n) = \bar{Y}_n - t_\beta[n-1]S_n\sqrt{1+\frac{1}{n}} \tag{6.4.5}$$

is such a β-content lower tolerance limit. Indeed, if a $N(\mu, \sigma)$ random variable Y is independent of $(\bar{Y}_n, S_n^2)$ then

$$P_{\mu,\sigma}\{Y \geq l_\beta(\bar{Y}_n, S_n)\} = P_{\mu,\sigma}\{(Y - \bar{Y}_n)/\left(S \cdot \left(1 + \frac{1}{n}\right)^{1/2}\right) \geq -t_\beta[n-1]\} = \beta, \tag{6.4.6}$$

since $Y - \bar{Y}_n \sim N\left(0, \sigma^2\left(1 + \frac{1}{n}\right)\right)$ and $S_n \sim \sigma\left(\frac{\chi^2[n-1]}{n-1}\right)^{1/2}$. It is interesting to compare the β-content lower tolerance limit (6.4.5) with the $(1-\alpha, \beta)$ guaranteed coverage lower tolerance limit (6.4.6). We can show that if $\beta = 1 - \alpha$ then the two limits are approximately the same in large samples.

6.5 DISTRIBUTION FREE CONFIDENCE AND TOLERANCE INTERVALS

Let $\mathcal{F}$ be the class of all absolutely continuous distributions. Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables having a distribution $F(x)$ in $\mathcal{F}$. Let $X_{(1)} \leq \cdots \leq X_{(n)}$ be the order statistics. This statistic is minimal sufficient. The transformed random variable $Y = F(X)$ has a rectangular distribution on $(0, 1)$. Let x_p be the pth quantile of $F(x)$, i.e., $x_p = F^{-1}(p)$, $0 < p < 1$. We show now that the order statistics $X_{(i)}$ can be used as (p, γ) tolerance limits, irrespective of the functional form of $F(x)$. Indeed, the transformed random variables $Y_{(i)} = F(X_{(i)})$ have the beta distributions $\beta(i, n - i + 1)$, $i = 1, \ldots, n$. Accordingly,

$$P[Y_{(i)} \geq p] = P[X_{(i)} \geq F^{-1}(p)] = I_{1-p}(n - i + 1, i), \quad i = 1, \ldots, n. \tag{6.5.1}$$

Therefore, a distribution free (p, γ) upper tolerance limit is the smallest $X_{(j)}$ satisfying condition (6.5.1). In other words, for any continuous distribution $F(x)$, define

$$i^0 = \text{least } j \geq 1 \text{ such that } I_{1-p}(n - j + 1, j) \geq \gamma. \tag{6.5.2}$$

Then, the order statistic $X_{(i^0)}$ is a (p, γ)-**upper** tolerance limit. We denote this by $L_{p,\gamma}(\mathbf{X})$. Similarly, a distribution free (p, γ)-**lower** tolerance limit is given by

$$L_{p,\gamma}(x) = X_{(i^0)} \text{ where } i^0 = \text{largest } j \geq 1 \text{ such that } I_{1-p}(j, n - j + 1) \geq \gamma. \tag{6.5.3}$$

The upper and lower tolerance intervals given in (6.5.2) and (6.5.3) might not exist if n is too small. They could be applied to obtain distribution free confidence intervals for the mean, μ, of a **symmetric** continuous distribution. The method is based on the fact that the expected value, μ, and the median, $F^{-1}(0.5)$, of continuous symmetric

distributions coincide. Since $I_{0.5}(a, b) = 1 - I_{0.5}(b, a)$ for all $0 < a,\ b < \infty$, we obtain from (6.5.2) and (6.5.3) by substituting $p = 0.5$ that the $(1 - \alpha)$ upper and lower distribution free confidence limits for μ are $\bar{\mu}_\alpha$ and $\underline{\mu}_\alpha$ where, for sufficiently large n,

$$\bar{\mu}_\alpha = X_{(j)},\ \text{where } j = \text{least positive integer,} k,\ \text{such that } I_{0.5}(n - k + 1, k + 1) \geq 1 - \alpha/2; \tag{6.5.4}$$

and

$$\underline{\mu}_\alpha = X_{(i)},\ \text{where } i = \text{largest positive integer, } k,\ \text{such that } I_{0.5}(n - k + 1, k + 1) \leq \alpha/2. \tag{6.5.5}$$

Let F be a log-convex distribution function. Then for any positive real numbers $a_1, \ldots, a_r$,

$$-\log\left(1 - F\left(\sum_{i=1}^{r} a_i X_{(i)}\right)\right) \leq -\sum_{i=1}^{r} a_i \log(1 - F(X_{(i)})), \tag{6.5.6}$$

or equivalently

$$F\left(\sum_{i=1}^{r} a_i X_{(i)}\right) \leq 1 - \exp\left\{\sum_{i=1}^{r} a_i \log(1 - F(X_{(i)}))\right\}. \tag{6.5.7}$$

Let

$$G(X_{(i)}) = -\log(1 - F(X_{(i)})), \quad i = 1, \ldots, r. \tag{6.5.8}$$

Since $F(X) \sim R(0, 1)$ and $-\log(1 - R(0, 1)) \sim G(1, 1)$. The statistic $G(X_{(i)})$ is distributed like the ith order statistic from a standard exponential distribution. Substitute in (6.5.7)

$$\sum_{i=1}^{r} a_i X_{(i)} = \sum_{i=1}^{r} A_i (X_{(i)} - X_{(i-1)})$$

and

$$\sum_{i=1}^{r} a_i G(X_{(i)}) = \sum_{i=1}^{r} A_i (G(X_{(i)}) - G(X_{(i-1)})),$$

where $A_i = \sum_{j=i}^{r} a_i$, $i = 1, \ldots, r$ and $X_{(0)} \equiv 0$. Moreover,

$$G(X_{(i)}) - G(X_{(i-1)}) \sim \frac{1}{n-i+1} G(1,1), \quad i = 1, \ldots, r. \tag{6.5.9}$$

Hence, if we define

$$A_i = \frac{2\log(1-p)}{\chi^2_{1-\alpha}[2r]}(n-i+1), \quad i = 1, \ldots, r,$$

then, from (6.5.7)

$$\begin{aligned} P\{L_{p,1-\alpha}(T_{n,r}) \le F^{-1}(p)\} &= P\left\{F\left(\sum_{i=1}^{r} A_i(X_{(i)} - X_{(i-1)})\right) \le p\right\} \\ &\ge P\left\{1 - \exp\left\{-\sum_{i=1}^{r} A_i(G(X_{(i)}) - G(X_{(i-1)}))\right\} \le p\right\} \\ &= P\left\{\sum_{i=1}^{r} A_i(G(X_{(i)}) - G(X_{(i-1)})) \le -\log(1-p)\right\} \\ &= P\{\chi^2[2r] \le \chi^2_{1-\alpha}[2r]\} = 1-\alpha, \end{aligned} \tag{6.5.10}$$

since $2\sum_{i=1}^{r}(n-i+1)(G(X_{(i)}) - G(X_{(i-1)})) \sim \chi^2[2r]$. This result was published first by Barlow and Proschan (1966).

6.6 SIMULTANEOUS CONFIDENCE INTERVALS

It is often the case that we estimate simultaneously several parameters on the basis of the same sample values. One could determine for each parameter a confidence interval at level $(1-\alpha)$ irrespectively of the confidence intervals of the other parameters. The result is that the overall confidence level is generally smaller than $(1-\alpha)$. For example, suppose that $(X_1, \ldots, X_n)$ is a sample of n i.i.d. random variables from $N(\mu, \sigma^2)$. The sample mean $\bar{X}$ and the sample variance S^2 are independent statistics. Confidence intervals for μ and for σ, determined separately for each parameter, are

$$I_1(\bar{X}, S) = \left(\bar{X} - t_{1-\alpha/2}[n-1]\frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2}[n-1]\frac{S}{\sqrt{n}}\right)$$

and

$$I_2(S) = \left(S\left(\frac{n-1}{\chi^2_{1-\alpha/2}[n-1]}\right)^{1/2}, \quad S\left(\frac{n-1}{\chi^2_{1-\alpha}[n-1]}\right)^{1/2} \right),$$

respectively. These intervals are not independent. We can state that the probability for μ to be in $I_1(\bar{X}, S)$ is $(1-\alpha)$ and that of σ to be in $I_2(S)$ is $(1-\alpha)$. But, what is the probability that both statements are simultaneously true? According to the **Bonferroni inequality** (4.6.50)

$$\begin{aligned} P_{\mu,\sigma}\{\mu \in I_1(\bar{X}, S), \sigma \in I_2(S)\} &\geq 1 - P_{\mu,\sigma}\{\mu \notin I_1(\bar{X}, S)\} - P_{\mu,r}\{\sigma \notin I_2(S)\} \\ &= 1 - 2\alpha, \quad \text{for all } \mu, \sigma. \end{aligned} \tag{6.6.1}$$

We see that a lower bound to the simultaneous coverage probability of (μ, σ) is according to (6.6.1), $1-2\alpha$. The actual simultaneous coverage probability of $I_1(\bar{X}, S)$ and $I_2(S)$ can be determined by evaluating the integral

$$P(\sigma) = 2\int_{\chi^2_{\alpha/2}[n-1]}^{\chi^2_{1-\alpha/2}[n-1]} \Phi\left(t_{1-\alpha/2}[n-1]\frac{\sigma x}{\sqrt{n(n-1)}}\right) g_n(x)dx - (1-\alpha), \tag{6.6.2}$$

where $g_n(x)$ is the probability density function (p.d.f.) of $\chi^2[n-1]$ and $\Phi(\cdot)$ is the standard normal integral. The value of $P(\sigma)$ is smaller than $(1-\alpha)$. In order to make it at least $(1-\alpha)$, we can modify the individual confidence probabilities of $I_1(\bar{X}, S)$ and of $I_2(S)$ to be $1-\alpha/2$. Then the simultaneous coverage probability will be between $(1-\alpha)$ and $(1-\alpha/2)$. This is a simple procedure that is somewhat conservative. It guarantees a simultaneous confidence level not smaller than the nominal $(1-\alpha)$. This method of constructing simultaneous confidence intervals, called the Bonferroni method, has many applications. We have shown in Chapter 4 an application of this method in a two-way analysis of variance problem. Miller (1966, p. 67) discussed an application of the Bonferroni method in a case of simultaneous estimation of k normal means.

Consider again the linear model of full rank discussed in Section 5.3.2, in which the vector $\mathbf{X}$ has a multinormal distribution $N(A\boldsymbol{\beta}, \sigma^2 I)$. A is an $n \times p$ matrix of full rank and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters. The least-squares estimator (LSE) of a specific linear combination of $\boldsymbol{\beta}$, say $\lambda = \boldsymbol{\alpha}'\boldsymbol{\beta}$, is $\hat{\lambda} = \boldsymbol{\alpha}'\hat{\boldsymbol{\beta}} = \boldsymbol{\alpha}'(A'A)^{-1}A'\mathbf{X}$. We proved that $\hat{\lambda} \sim N(\boldsymbol{\alpha}'\boldsymbol{\beta}, \sigma^2\boldsymbol{\alpha}'(A'A)^{-1}\alpha)$. Moreover, an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-p}\mathbf{X}'(I - A(A'A)^{-1}A')\mathbf{X},$$

where $\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p}\chi^2[n-p]$. Hence, a $(1-\alpha)$ confidence interval for the particular parameter λ is

$$\hat{\lambda} \pm t_{1-\alpha/2}[n-p]\hat{\sigma}(\boldsymbol{\alpha}'(A'A)^{-1}\boldsymbol{\alpha})^{1/2}. \tag{6.6.3}$$

Suppose that we are interested in the simultaneous estimation of all (many) linear combinations belonging to a certain r-dimensional linear subspace $1 \le r \le p$. For example, if we are interested in **contrasts** of the $\boldsymbol{\beta}$-component, then $\lambda = \sum_{i=1}^{p} \alpha_i \beta_i$ where $\Sigma\alpha_i = 0$. In this case, the linear subspace of all such contrasts is of dimension $r = p-1$. Let L be an $r \times p$ matrix with r row vectors that constitute a basis for the linear subspace under consideration. For example, in the case of all contrasts, the matrix L can be taken as the $(p-1) \times p$ matrix:

$$L = \begin{pmatrix} 1 & -1 & & & & \\ & 1 & -1 & & 0 & \\ & & \ddots & \ddots & & \\ & 0 & & & & \\ & & & & 1 & -1 \end{pmatrix}.$$

Every vector $\boldsymbol{\alpha}$ belonging to the specified subspace is given by some linear combination $\boldsymbol{\alpha}' = \boldsymbol{\gamma}'L$. Thus, $\boldsymbol{\alpha}'(A'A)^{-1}\boldsymbol{\alpha} = \boldsymbol{\gamma}'L(A'A)^{-1}L'\boldsymbol{\gamma}$. Moreover,

$$L\hat{\boldsymbol{\beta}} \sim N(L\boldsymbol{\beta}, \sigma^2 L(A'A)^{-1}L') \tag{6.6.4}$$

and

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'L'(L(A'A)^{-1}L')^{-1}L(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \sim \sigma^2\chi^2[r], \tag{6.6.5}$$

where r is the rank of L. Accordingly,

$$\frac{1}{\hat{\sigma}^2}(\hat{\beta} - \beta)'L'(L(A'A)^{-1}L')^{-1}L(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim rF[r, n-p] \tag{6.6.6}$$

and the probability is $(1-\alpha)$ that $\boldsymbol{\beta}$ belongs to the ellipsoid

$$E_\alpha(\boldsymbol{\beta}, \hat{\sigma}^2, L) = \{\boldsymbol{\beta} : (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'L'(L(A'A)^{-1}L')^{-1}L(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \le r\hat{\sigma}^2 F_{1-\alpha}[r, n-p]\}. \tag{6.6.7}$$

$E_\alpha(\boldsymbol{\beta}, \sigma^2, L)$ is a simultaneous confidence region for all $\boldsymbol{\alpha}'\boldsymbol{\beta}$ at level $(1-\alpha)$. Consider any linear combination $\lambda = \boldsymbol{\alpha}'\boldsymbol{\beta} = \gamma' L\boldsymbol{\beta}$. The simultaneous confidence interval for

λ can be obtained by the orthogonal projection of the ellipsoid $E_\alpha(\boldsymbol{\beta}, \hat{\sigma}^2, L)$ on the line l spanned by the vector $\boldsymbol{\gamma}$. We obtain the following formula for the confidence limits of this interval

$$\hat{\lambda} \pm (r F_{1-\alpha}[r, n-p])^{1/2} \hat{\sigma} (\gamma' L (A'A)^{-1} L' \gamma)^{1/2}, \tag{6.6.8}$$

where $\hat{\lambda} = \boldsymbol{\gamma}' L \hat{\boldsymbol{\beta}} = \boldsymbol{\alpha}' \hat{\boldsymbol{\beta}}$. We see that in case of $r = 1$ formula (6.6.8) reduces to (6.6.3), otherwise the coefficient $(r F_{1-\alpha}[r, n-p])^{1/2}$ is greater than $t_{1-\alpha/2}[n-p]$. This coefficient is called Scheffés S-coefficient. Various applications and modifications of the S-method have been proposed in the literature. For applications often used in statistical practice, see Miller (1966, p. 54). Scheffé (1970) suggested some modifications for increasing the efficiency of the S-method for simultaneous confidence intervals.

6.7 TWO-STAGE AND SEQUENTIAL SAMPLING FOR FIXED WIDTH CONFIDENCE INTERVALS

We start the discussion with the problem of determining fixed-width confidence intervals for the mean μ of a normal distribution when the variance σ^2 is unknown and can be arbitrarily large. We saw previously that if the sample consists of n i.i.d. random variables $X_1, \ldots, X_n$, where n is fixed before the sampling, then a UMAU confidence limit for μ are given, in correspondence to the t-test, by $\bar{X} \pm t_{1-\alpha/2}[n-1]\frac{S}{\sqrt{n}}$, where $\bar{X}$ and S are the sample mean and standard deviation, respectively. The width of this confidence interval is

$$\Delta^* = 2 t_{1-\alpha/2}[n-1] \frac{S}{\sqrt{n}}. \tag{6.7.1}$$

Although the width of the interval is converging to zero, as $n \to \infty$, for each fixed n, it can be arbitrarily large with positive probability. The question is whether there exists another confidence interval with bounded width. We show now that there is no fixed-width confidence interval in the present normal case if the sample is of fixed size. Let $I_\delta(\bar{X}, S)$ be any fixed width interval centered at $\hat{\mu}(\bar{X}, S)$, i.e.,

$$I_\delta(\bar{X}, S) = (\hat{\mu}(\bar{X}, S) - \delta, \hat{\mu}(\bar{X}, S) + \delta). \tag{6.7.2}$$

We show that the maximal possible confidence level is

$$\sup_{\hat{\mu}} \inf_{\mu, \sigma} P_{\mu,\sigma}\{\mu \in I_\delta(\bar{X}, S)\} = 0. \tag{6.7.3}$$

This means that there is no statistic $\hat{\mu}(\bar{X}, S)$ for which $I_\delta(\bar{X}, S)$ is a confidence interval. Indeed,

$$\sup_{\hat{\mu}} \inf_{\mu,\sigma} P_{\mu,\sigma}\{\mu \in I_\delta(\bar{X}, S)\} \leq \lim_{\sigma\to\infty} \inf_{\mu} \sup_{\hat{\mu}} P_{\mu,\sigma}\{\mu \in I_\delta(\bar{X}, S)\}. \tag{6.7.4}$$

In Example 9.2, we show that $\hat{\mu}(\bar{X}, S) = \bar{X}$ is a minimax estimator, which maximizes the minimum coverage. Accordingly,

$$\inf_{\mu} \sup_{\hat{\mu}} P_{\mu,\sigma}\{\mu \in I_\delta\{(\bar{X}, S)\} = P_\sigma\{\bar{X} - \delta \leq \mu \leq \bar{X} + \delta\} = 2\Phi\left(\frac{\delta}{\sigma}\sqrt{n}\right) - 1. \tag{6.7.5}$$

Substituting this result in (6.7.4), we readily obtain (6.7.3), by letting $\sigma \to \infty$.

Stein's two-stage procedure. Stein (1945) provided a two-stage solution to this problem of determining a fixed-width confidence interval for the mean μ. According to Stein's procedure the sampling is performed in two stages:

Stage I:

(i) Observe a sample of n_1 i.i.d. random variables from $N(\mu, \sigma^2)$.

(ii) Compute the sample mean $\bar{X}_{n_1}$ and standard deviation S_{n_1}.

(iii) Determine

$$N = 1 + \left[t^2_{1-\alpha/2}[n_1 - 1]\frac{S^2}{\delta^2}\right], \tag{6.7.6}$$

where $[x]$ designates the integer part of x.

(iv) If $N > n_1$ go to Stage II; else set the interval

$$I_\delta(\bar{X}_{n_1}) = (\bar{X}_{n_1} - \delta, \bar{X}_{n_1} + \delta).$$

Stage II:

(i) Observe $N_2 = N - n_1$ additional i.i.d. random variables from $N(\mu, \sigma^2)$; $Y_1, \ldots, Y_{N_2}$.

(ii) Compute the overall mean $\bar{X}_N = (n_1\bar{X}_{n_1} + N_2\bar{Y}_{N_2})/N$.

(iii) Determine the interval $I_\delta(\bar{X}_N) = (\bar{X}_N - \delta, \bar{X}_N + \delta)$.

The size of the second stage sample $N_2 = (N - n_1)^+$ is a random variable, which is a function of the first stage sample variance $S^2_{n_1}$. Since $\bar{X}_{n_1}$ and $S^2_{n_1}$ are independent,

$\bar{X}_{n_1}$ and N_2 are independent. Moreover, $\bar{Y}_{N_2}$ is conditionally independent of $S^2_{n_1}$, given N_2. Hence,

$$
\begin{aligned}
P_{\mu,\sigma}\{|\bar{X}_N - \mu| < \delta\} &= E\{P_{\mu,\sigma}\{|\bar{X}_N - \mu| < \delta \mid N\}\} \\
&= E\left\{2\Phi\left(\frac{\delta}{\sigma}\sqrt{N}\right) - 1\right\} \\
&\geq 2E\left\{\Phi\left(\frac{\delta}{\sigma} \cdot \frac{S_{n_1}}{\delta} t_{1-\alpha/2}[n_1 - 1]\right)\right\} - 1 \qquad (6.7.7) \\
&= 2P\left\{\frac{N(0,1)}{\sqrt{\chi^2[n_1 - 1]/(n_1 - 1)}} \leq t_{1-\alpha/2}[n_1 - 1]\right\} - 1 = 1 - \alpha.
\end{aligned}
$$

This proves that the fixed width interval $I_\delta(\bar{X}_N)$ based on the prescribed two-stage sampling procedure is a confidence interval. The Stein two-stage procedure is not an efficient one, unless one has good knowledge of how large n_1 should be. If σ^2 is known there exists a UMAU confidence interval of fixed size, i.e., $I_\delta(\bar{X}_{n^0(\delta)})$ where

$$
n^0(\delta) = 1 + \left[\frac{\chi^2_{1-\alpha}[1]}{\delta^2}\sigma^2\right]. \qquad (6.7.8)
$$

If n_1 is close to $n^0(\delta)$ the procedure is expected to be efficient. $n^0(\delta)$ is, however, unknown. Various approaches have been suggested to obtain efficient procedures of sampling. We discuss here a sequential procedure that is asymptotically efficient. Note that the optimal sample size $n^0(\delta)$ increases to infinity like $1/\delta^2$ as $\delta \to 0$. Accordingly, a sampling procedure, with possibly random sample size, N, which yields a fixed-width confidence interval $I_\delta(\bar{X}_N)$ is called **asymptotically efficient** if

$$
\overline{\lim_{\delta \to 0}} \frac{E_\delta\{N\}}{n^0(\delta)} = 1. \qquad (6.7.9)
$$

Sequential fixed-width interval estimation. Let $\{a_n\}$ be a sequence of positive numbers such that $a_n \to \chi^2_{1-\alpha}[1]$ as $n \to \infty$. We can set, for example, $a_n = F_{1-\alpha}[1, n]$ for all $n \geq n_1$ and $a_n = \infty$ for $n < n_1$. Consider now the following sequential procedure:

1. Starting with $n = n_1$ i.i.d. observations compute $\bar{X}_n$ and S^2_n.
2. If $n > a_n S^2_n/\delta^2$ stop sampling and, estimate μ by $I_\delta(\bar{X}_n)$; else take an additional independent observation and return to (i). Let

$$
N(\delta) = \text{least } n \geq n_1, \text{ such that } n > a_n S^2_n/\delta^2. \qquad (6.7.10)
$$

According to the specified procedure, the sample size at termination is $N(\delta)$. $N(\delta)$ is called a **stopping variable**. We have to show first that $N(\delta)$ is finite with probability one, i.e.,

$$\lim_{n\to\infty} P_{\mu,\sigma}\{N(\delta) > n\} = 0, \tag{6.7.11}$$

for each $\delta > 0$. Indeed, for any given n,

$$\begin{aligned} P_{\mu,\sigma}\{N(\delta) > n\} &= P_{\mu,\sigma}\left\{\bigcap_{j=n_1}^{n}\left\{S_j^2 > \frac{j\delta^2}{a_j}\right\}\right\} \\ &\le P_{\mu,\sigma}\left\{S_n^2 > \frac{n\delta^2}{a_n}\right\}. \end{aligned} \tag{6.7.12}$$

But

$$P\left\{S_n^2 > \frac{n\delta^2}{a_n}\right\} = P\left\{\frac{S_n^2}{n} > \frac{\delta^2}{a_n}\right\}. \tag{6.7.13}$$

$\frac{S_n^2}{n} \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$, therefore

$$\lim_{n\to\infty} P\left\{\frac{S_n^2}{n} > \frac{\delta^2}{a_n}\right\} = \lim_{n\to\infty} P\left\{\frac{\chi^2[n-1]}{n-1} > \frac{n\delta^2}{\sigma^2\chi^2_{1-\alpha}[1]}\right\} = 0 \tag{6.7.14}$$

as $n \to \infty$. Thus, (6.7.11) is satisfied and $N(\delta)$ is a finite random variable. The present sequential procedure attains in large samples the required confidence level and is also an efficient one. One can prove in addition the following optimal properties:

(i) If $a_n = a$ for all $n \ge n_1$ then $E_\sigma\{N(\delta)\} \le n_0(\delta) + n_1 + 1$, for all σ^2. (6.7.15)

This obviously implies the asymptotic efficiency (6.7.9). It is, however, a much stronger property. One does not have to pay, on the average, more than the equivalent of $n_1 + 1$ observations. The question is whether we do not tend to stop too soon and thus lose confidence probability. Simons (1968) proved that if we follow the above procedure, $n_1 \ge 3$ and $a_n = a$ for all $n \ge 3$, then there exists a finite integer k such that

$$P_{\mu,\sigma}\{|\bar{X}_{N+k} - \mu| \le \delta\} \ge 1 - \alpha, \tag{6.7.16}$$

for all μ, σ and δ. This means that the possible loss of confidence probability is not more than the one associated with a finite number of observations. In other words, if the sample is large we generally attain the required confidence level.

We have not provided here proofs of these interesting results. The reader is referred to Zacks (1971, p. 560). The results were also extended to general classes of distributions originally by Chow and Robbins (1965), followed by studies of Starr (1966), Khan (1969), Srivastava (1971), Ghosh, Mukhopadhyay, and Sen (1997), and Mukhopadhyay and de Silva (2009).

PART II: EXAMPLES

Example 6.1. It is assumed that the compressive strength of concrete cubes follows a log-normal distribution, $LN(\mu, \sigma^2)$, with unknown parameters (μ, σ). It is desired that in a given production process the compressive strength, X, will not be smaller than ξ_0 in $(1-\beta)\times 100\%$ of the concrete cubes. In other words, the β-quantile of the parent log-normal distribution should not be smaller than ξ_0, where the β-quantile of $LN(\mu, \sigma^2)$ is $x_\beta = \exp\{\mu + z_\beta\sigma\}$, and z_β is the β-quantile of $N(0, 1)$. We observe a sample of n i.i.d. random variables $X_1, \ldots, X_n$ and should decide on the basis of the observed sample values whether the strength requirement is satisfied. Let $Y_i = \log X_i$ $(i = 1, \ldots, n)$. The sample mean and variance $(\bar{Y}_n, S_n^2)$, where $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$, constitute a minimal sufficient statistic. On the basis of $(\bar{Y}_n, S_n^2)$, we wish to determine a $(1-\alpha)$-lower confidence limit, $\underline{x}_{\alpha,\beta}$ to the unknown β-quantile x_β. Accordingly, $\underline{x}_{\alpha,\beta}$ should satisfy the relationship

$$P_{\mu,\sigma}\{\underline{x}_{\alpha,\beta} \le x_\beta\} \ge 1-\alpha, \quad \text{for all } (\mu, \sigma).$$

$\underline{x}_{\alpha,\beta}$ is called a **lower** $(1-\alpha, 1-\beta)$ **guaranteed coverage tolerance limit**. If $\underline{x}_{\alpha,\beta} \ge \xi_0$, we say that the production process is satisfactory (meets the specified standard). Note that the problem of determining $\underline{x}_{\alpha,\beta}$ is equivalent to the problem of determining a $(1-\alpha)$-lower confidence limit to $\mu + z_\beta\sigma$. This lower confidence limit is constructed in the following manner. We note first that if $U \sim N(0, 1)$, then

$$\sqrt{n}[\bar{Y}_n - (\mu + z_\beta\sigma)]/S_n \sim \frac{U + \sqrt{n}\, z_{1-\beta}}{(\chi^2[n-1]/(n-1))^{1/2}} \sim t[n-1; \sqrt{n}\, z_{1-\beta}],$$

where $t[\nu; \delta]$ is the noncentral t-distribution. Thus, a $(1-\alpha)$-lower confidence limit for $\mu + z_\beta\sigma$ is

$$\underline{\eta}_{\alpha,\beta} = \bar{Y}_n - t_{1-\alpha}[n-1; \sqrt{n}\, z_{1-\beta}]\frac{S_n}{\sqrt{n}}$$

and $\underline{x}_{\alpha,\beta} = \exp\{\underline{\eta}_{\alpha,\beta}\}$ is a lower $(1-\alpha, 1-\beta)$-tolerance limit. ■

Example 6.2. Let $X_1, \ldots, X_n$ be i.i.d. random variables representing the life length of electronic systems and distributed like $G\left(\frac{1}{\theta}, 1\right)$. We construct two different $(1-\alpha)$-lower confidence limits for θ.

(**i**) The minimal sufficient statistic is $T_n = \Sigma X_i$. This statistic is distributed like $\frac{\theta}{2}\chi^2[2n]$. Thus, for testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$ at level of significance α, the acceptance regions are of the form

$$A(\theta_0) = \left\{T_n : T_n \leq \frac{\theta_0}{2}\chi^2_{1-\alpha}[2n]\right\}, \quad 0 < \theta_0 < \infty.$$

The corresponding confidence intervals are

$$S(T_n) = \left(\theta : \theta \geq \frac{2T_n}{\chi^2_{1-\alpha}[2n]}\right).$$

The lower confidence limit for θ is, accordingly,

$$\underline{\theta}_\alpha = 2T_n/\chi^2_{1-\alpha}[2n].$$

(**ii**) Let $X_{(1)} = \min_{1\leq i\leq n}\{X_i\}$. $X_{(1)}$ is distributed like $\frac{\theta}{2n}\chi^2[2]$. Hence, the hypotheses $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$ can be tested at level α by the acceptance regions

$$A'(\theta_0) = \left\{X_{(1)} : X_{(1)} \leq \frac{\theta_0}{2n}\chi^2_{1-\alpha}[2]\right\}, \quad 0 < \theta_0 < \infty.$$

These regions yield the confidence intervals

$$S'(X_{(1)}) = \left(\theta : \theta \geq \frac{2nX_{(1)}}{\chi^2_{1-\alpha}[2]}\right).$$

The corresponding lower confidence limit is $\underline{\theta}'_\alpha = 2nX_{(1)}/\chi^2_{1-\alpha}[2]$. Both families of confidence intervals provide lower confidence limits for the mean-time between failures, θ, at the same confidence level $1-\alpha$. The question is which family is more efficient. Note that $\underline{\theta}_\alpha$ is a function of the minimal sufficient statistic, while $\underline{\theta}'_\alpha$ is not. The expected value of $\underline{\theta}_\alpha$ is $E_\theta\{\underline{\theta}_\alpha\} = \frac{2n\theta}{\chi^2_{1-\alpha}[2n]}$. This expected value is approximately, as $n \to \infty$,

$$E_\theta\{\underline{\theta}_\alpha\} = \theta/\left(1 + \frac{z_{1-\alpha}}{\sqrt{n}}\right) \approx \theta\left(1 - \frac{z_{1-\alpha}}{\sqrt{n}} + 0\left(\frac{1}{n}\right)\right), \quad \text{as } n \to \infty.$$

Thus, $E\{\underline{\theta}_\alpha\}$ is always smaller than θ, and approaches θ as n grows. On the other hand, the expected value of $\underline{\theta}'_\alpha$ is

$$E_\theta\left(\frac{2nX_{(1)}}{\chi^2_{1-\alpha}[2]}\right) = \frac{2\theta}{\chi^2_{1-\alpha}[2]} = \frac{\theta}{-\log\alpha}.$$

This expectation is about $\theta/3$ when $\alpha = 0.05$ and $\theta/4.6$ when $\alpha = 0.01$. It does not converge to θ as n increases. Thus, $\underline{\theta}'_\alpha$ is an inefficient lower confidence limit of θ. ■

Example 6.3.

A. Let $X_1, \ldots, X_n$ be i.i.d. $N(0, \sigma^2)$ random variables. We would like to construct the UMA $(1-\alpha)$-lower confidence limit of σ^2. The minimal sufficient statistic is $T_n = \Sigma X_i^2$, which is distributed like $\sigma^2\chi^2[n]$. The UMP test of size α of $H_0 : \sigma^2 \le \sigma_0^2$ against $H_1 : \sigma^2 > \sigma_0^2$ is

$$\phi^0(T_n) = I\{T_n \ge \sigma_0^2\chi^2_{1-\alpha}[n]\}.$$

Accordingly, the UMA $(1-\alpha)$-lower confidence limit σ^2_α is

$$\underline{\sigma}^2_\alpha = T_n/\chi^2_{1-\alpha}[n].$$

B. Let $X \sim B(n, \theta)$, $0 < \theta < 1$. We determine the UMA $(1-\alpha)$-lower confidence limit of the success probability θ. In (2.2.4), we expressed the c.d.f. of $B(n, \theta)$ in terms of the incomplete beta function ratio. Let R be a random number in (0, 1), independent of X, then $\underline{\theta}_\alpha$ is the root of the equation

$$RI_{1-\underline{\theta}_\alpha}(n-X, X+1) + (1-R)I_{1-\underline{\theta}_\alpha}(n-X+1, X) = 1-\alpha,$$

provided $1 \le X \le n-1$. If $X = 0$, the lower confidence limit is $\underline{\theta}_\alpha(0) = 0$. When $X = n$ the lower confidence limit is $\underline{\theta}_\alpha(n) = \alpha^{1/n}$. By employing the relationship between the central F-distribution and the beta distribution (see Section 2.14), we obtain the following for $X \ge 1$ and $R = 1$:

$$\underline{\theta}_\alpha = \frac{X}{(n-X+1) + XF_{1-\alpha}[2X, 2(n-X+1)]}.$$

If $X \ge 1$ and $R = 0$ the lower limit, $\underline{\theta}'_\alpha$ is obtained from (6.3.11) by substituting $(X-1)$ for X. Generally, the lower limit can be obtained as the average $R\underline{\theta}_\alpha + (1-R)\underline{\theta}'_\alpha$. In practice, the nonrandomized solution (6.3.11) is often applied.

■

Example 6.4. Let X and Y be independent random variables having the normal distribution $N(0, \sigma^2)$ and $N(0, \rho\sigma^2)$, respectively. We can readily prove that

$$\psi(\sigma^2, \rho) = P_{\sigma^2,\rho}\{X^2 + Y^2 \leq 1\} = 1 - E\left\{P\left(J; \frac{1}{2\sigma^2}\right)\right\},$$

where J has the negative binomial distribution $NB\left(1 - \frac{1}{\rho}, \frac{1}{2}\right)$. $P(j; \lambda)$ designates the c.d.f. of the Poisson distributions with mean λ. $\psi(\sigma^2, \rho)$ is the coverage probability of a circle of radius one. We wish to determine a $(1 - \alpha)$-lower confidence limit for $\psi(\sigma^2, \rho)$, on the basis of n independent vectors, $(X_1, Y_1), \ldots, (X_n, Y_n)$, when ρ is known. The minimal sufficient statistic is $T_{2n} = \Sigma X_i^2 + \frac{1}{\rho}\Sigma Y_i^2$. This statistic is distributed like $\sigma^2\chi^2[2n]$. Thus, the UMA $(1 - \alpha)$-upper confidence limit for σ^2

$$\bar{\sigma}_\alpha^2 = T_{2n}/\chi_\alpha^2[2n].$$

The Poisson family is an MLR one. Hence, by Karlin's Lemma, the c.d.f. $P(j; 1/2\sigma^2)$ is an increasing function of σ^2 for each $j = 0, 1, \ldots$. Accordingly, if $\sigma^2 \leq \bar{\sigma}_\alpha^2$ then $P(j; 1/2\sigma^2) \leq P(j; 1/2\bar{\sigma}_\alpha^2)$. It follows that $E\left\{P\left(J; \frac{1}{2\sigma^2}\right)\right\} \leq E\left\{P\left(J; \frac{1}{2\bar{\sigma}_\alpha^2}\right)\right\}$. From this relationship we infer that

$$\psi(\bar{\sigma}_\alpha^2, \rho) = 1 - E\left\{P\left(J; \frac{1}{2\bar{\sigma}_\alpha^2}\right)\right\}$$

is a $(1 - \alpha)$-lower confidence limit for $\psi(\sigma^2, \rho)$. We show now that $\psi(\bar{\alpha}_\alpha^2, \rho)$ is a UMA lower confidence limit. By negation, if $\psi(\bar{\sigma}_\alpha^2, \rho)$ is not a UMA, there exists another $(1 - \alpha)$ lower confidence limit, $\underline{\hat{\psi}}_\alpha$ say, and some $0 < \psi' < \psi(\sigma^2, \rho)$ such that

$$P\{\psi(\bar{\sigma}_\alpha^2, \rho) \leq \psi'\} > P\{\underline{\hat{\psi}}_\alpha \leq \psi'\}.$$

The function $P\left(j; \frac{1}{2\sigma^2}\right)$ is a strictly increasing function of σ^2. Hence, for each ρ there is a unique inverse $\sigma_\rho^2(\psi)$ for $\psi(\sigma^2, \rho)$. Thus, we obtain that

$$P_{\sigma^2}\{\bar{\sigma}_\alpha^2 \geq \sigma_\rho^2(\psi')\} > P_{\sigma^2}\{\sigma_\rho^2(\underline{\hat{\psi}}_\alpha) \geq \sigma_\rho^2(\psi')\},$$

where $\sigma_\rho^2(\psi') < \sigma^2$. Accordingly, $\sigma_\rho^2(\underline{\hat{\psi}}_\alpha)$ is a $(1 - \alpha)$-upper confidence limit for σ^2. But then the above inequality contradicts the assumption that $\bar{\sigma}_\alpha^2$ is UMA. ■

Example 6.5. Let $X_1, \ldots, X_n$ be i.i.d. random variables distributed like $N(\mu, \sigma^2)$. The UMP-unbiased test of the hypotheses

$$H_0 : \mu = \mu_0, \quad \sigma^2 \text{ arbitrary}$$

against

$$H_1 : \mu \neq \mu_0, \quad \sigma^2 \text{ arbitrary}$$

is the t-test

$$\phi^0(\bar{X}, S) = \begin{cases} 1, & \text{if } \dfrac{|\bar{X} - \mu_0|\sqrt{n}}{S} > t_{1-\frac{\alpha}{2}}[n-1], \\ 0, & \text{otherwise,} \end{cases}$$

where $\bar{X}$ and S are the sample mean and standard deviation, respectively. Correspondingly, the confidence interval

$$\left(\bar{X} - t_{1-\alpha/2}[n-1]\frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2}[n-1]\frac{S}{\sqrt{n}}\right)$$

is a UMAU at level $(1-\alpha)$. ■

Example 6.6. In Example 4.11, we discussed the problem of comparing the binomial experiments in two clinics at which standard treatment is compared with a new (test) treatment. If X_{ij} designates the number of successes in the jth sample at the ith clinic $(i = 1, 2; j = 1, 2)$, we assumed that X_{ij} are independent and $X_{ij} \sim B(n, \theta_{ij})$. We consider the cross-product ratio

$$\rho = \frac{\theta_{11}(1-\theta_{12})}{(1-\theta_{11})\theta_{12}} \Big/ \frac{\theta_{21}(1-\theta_{22})}{(1-\theta_{21})\theta_{22}}.$$

In Example 4.11, we developed the UMPU test of the hypothesis $H_0 : \rho = 1$ against $H_1 : \rho \neq 1$. On the basis of this UMPU test, we can construct the UMAU confidence limits of ρ.

Let $Y = X_{11}$, $T_1 = X_{11} + X_{12}$, $T = X_{21} + X_{22}$, and $S = X_{11} + X_{21}$. The conditional p.d.f. of Y given (T_1, T_2, S) under ρ was given in Example 4.11. Let $H(y \mid T_1, T_2, S)$ denote the corresponding conditional c.d.f. This family of conditional distributions is MLR in Y. Thus, the quantiles of the distributions are increasing functions of ρ. Similarly, $H(y \mid T_1, T_2, S)$ are strictly decreasing functions of ρ for each $y = 0, 1, \ldots, \min(T_1, S)$ and each (T_1, T_2, S).

As shown earlier one-sided UMA confidence limits require in discrete cases further randomization. Thus, we have to draw at random two numbers R_1 and R_2

Table 6.1 0.95—Confidence Limits for the Cross-Product Ratio

n_1	N_1	n_2	N_2	Y	T_1	T_2	S	$\underline{\rho}$	$\bar{\rho}$
32	112	78	154	5	15	17	18	.1103	2.4057
20	40	20	40	5	20	30	20	.0303	1.2787
25	50	25	50	15	25	27	22	5.8407	169.4280
20	50	20	50	15	25	27	22	5.6688	164.2365
40	75	30	80	33	43	25	48	.9049	16.2156

independently from a rectangular distribution $R(0, 1)$ and solve simultaneously the equations

$$R_1 H(Y; T_1, T_2, S, \underline{\rho}) + (1 - R_1) H(Y - 1; T_1, T_2, S, \underline{\rho}) = 1 - \epsilon_1,$$
$$R_2 H(Y - 1; T_1, T_2, S, \bar{\rho}) + (1 - R_2) H(Y; T_1, T_2, S, \bar{\rho}) = \epsilon_2,$$

where $\epsilon_1 + \epsilon_2 = \alpha$. Moreover, in order to obtain UMA unbiased intervals we have to determine $\underline{\rho}$, $\bar{\rho}$, ϵ_1 and ϵ_2 so that the two conditions of (4.4.2) will be satisfied simultaneously. One can write a computer algorithm to obtain this objective. However, the computations may be lengthy and tedious. If T_1, T_2 and S are not too small we can approximate the UMAU limits by the roots of the equations

$$H(Y; T_1, T_2, S, \underline{\rho}) = 1 - \alpha/2,$$
$$H(Y; T_1, T_2, S, \bar{\rho}) = \alpha/2.$$

These equations have unique roots since the c.d.f. $H(Y; T_1, T_2, S, \rho)$ is a strictly decreasing function of ρ for each (Y, T_1, T_2, S) having a continuous partial derivative with respect to ρ. The roots $\underline{\rho}$ and $\bar{\rho}$ of the above equations are generally the ones used in applications. However, they are **not** UMAU. In Table 6.1, we present a few cases numerically. The confidence limits in Table 6.1 were computed by determining first the large sample approximate confidence limits (see Section 7.4) and then correcting the limits by employing the monotonicity of the conditional c.d.f. $H(Y; T_1, T_2, S, \rho)$ in ρ. The limits are determined by a numerical search technique on a computer. ■

Example 6.7. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having a negative-binomial distribution $NB(\psi, \nu)$; ν is known and $0 < \psi < 1$. A minimal sufficient statistic is $T_n = \sum_{i=1}^{n} X_i$, which has the negative-binomial distribution $NB(\psi, n\nu)$. Consider the β-content one-sided prediction interval $[0, G^{-1}(\beta; \psi, \nu)]$, where $G^{-1}(p; \psi, \nu)$ is the pth quantile of $NB(\psi, \nu)$. The c.d.f. of the negative-binomial

distribution is related to the incomplete beta function ratio according to formula (2.2.12), i.e.,

$$G(x; \psi, \nu) = I_{1-\psi}(\nu, x+1), \quad x = 0, 1, \ldots.$$

The pth quantile of the $NB(\psi, \nu)$ can thus be defined as

$$G^{-1}(p; \psi, \nu) = \text{least nonnegative integer } j \text{ such that } I_{1-\psi}(\nu, j+1) \geq p.$$

This function is nondecreasing in ψ for each p and ν. Indeed, $\mathcal{F} = \{NB(\psi, \nu); 0 < \psi < 1\}$ is an MLR family. Furthermore, since $T_n \sim NB(\psi, n\nu)$, we can obtain a UMA upper confidence limit for ψ, $\bar{\psi}_\alpha$ at confidence level $\gamma = 1 - \alpha$. A nonrandomized upper confidence limit is the root ψ_α of the equation

$$I_{1-\bar{\psi}_\alpha}(n\nu, T_n + 1) = 1 - \alpha.$$

If we denote by $\beta^{-1}(p; a, b)$ the pth quantile of the beta distribution $\beta(a, b)$ then $\bar{\psi}_\alpha$ is given accordingly by

$$\bar{\psi}_\alpha = 1 - \beta^{-1}(\alpha; n\nu, T_n + 1).$$

The p-content $(1 - \alpha)$-level tolerance interval is, therefore, $[0, G^{-1}(p; \bar{\psi}_\alpha, \nu)]$. ■

Example 6.8. In statistical life testing families of increasing failure rate (IFR) are often considered. The **hazard** or **failure rate function** $h(x)$ corresponding to an absolutely continuous distribution $F(x)$ is defined as

$$h(x) = f(x)/(1 - F(x)),$$

where $f(x)$ is the p.d.f. A distribution function $F(x)$ is IFR if $h(x)$ is a nondecreasing function of x. The function $F(x)$ is differentiable almost everywhere. Hence, the failure rate function $h(x)$ can be written (for almost all x) as

$$h(x) = -\frac{d}{dx} \log(1 - F(x)).$$

Thus, if $F(x)$ is an IFR distribution, $-\log(1 - F(x))$ is a convex function of x. A distribution function $F(x)$ is called **log-convex** if its logarithm is a convex function of x. The tolerance limits that will be developed in the present example will be applicable for any log-convex distribution function.

Let $X_{(1)} \leq \cdots \leq X_{(n)}$ be the order statistic. It is instructive to derive first a $(p, 1-\alpha)$-lower tolerance limit for the simple case of the exponential distribution $G\left(\frac{1}{\theta}, 1\right)$, $0 < \theta < \infty$. The pth quantile of $G\left(\frac{1}{\theta}, 1\right)$ is

$$F^{-1}(p;\theta) = -\theta \log(1-p), \quad 0 < p < 1.$$

Let $T_{n,r} = \sum_{i=1}^{r}(n-i+1)(X_{(i)} - X_{(i-1)})$ be the **total life** until the rth failure. $T_{n,r}$ is distributed like $\frac{\theta}{2}\chi^2[2r]$. Hence, the UMA-$(1-\alpha)$-lower confidence limit for θ is

$$\underline{\theta}_\alpha = \frac{2T_{n,r}}{\chi^2_{1-\alpha}[2r]}.$$

The corresponding $(p, 1-\alpha)$ lower tolerance limit is

$$\underline{L}_{p,\alpha}(T_{n,r}) = -\log(1-p)\frac{2T_{n,r}}{\chi^2_{1-\alpha}[2r]}.$$

■

Example 6.9. The MLE of σ in samples from normal distributions is asymptotically normal with mean σ and variance $\sigma^2/2n$. Therefore, in large samples,

$$P_{\mu,\sigma}\left\{n\frac{(\bar{X}-\mu)^2}{\sigma^2} + 2n\frac{(S-\sigma)^2}{\sigma^2} \leq \chi^2_{1-\alpha}[2]\right\} \approx 1-\alpha,$$

for all μ, σ. The region given by

$$C_\alpha(\bar{X}, S) = \left\{(\mu,\sigma); n\left(\frac{\bar{X}-\mu}{\sigma}\right)^2 + 2n\left(\frac{S-\sigma}{\sigma}\right)^2 \leq \chi^2_{1-\alpha}[2]\right\}$$

is a simultaneous confidence region with coverage probability approximately $(1-\alpha)$. The points in the region $C_\alpha(\bar{X}, S)$ satisfy the inequality

$$|\bar{X}-\mu| \leq \left[\frac{\sigma^2\chi^2_{1-\alpha}[2]}{n} - 2(S-\sigma)^2\right]^{1/2}.$$

Hence, the values of σ in the region are only those for which the square root on the RHS of the above is real. Or, for all $n > \chi^2_{1-\alpha}[2]/2$,

$$\frac{S}{1+\left(\frac{\chi^2_{1-\alpha}[2]}{2n}\right)^{1/2}} \leq \sigma \leq \frac{S}{1-\left(\frac{\chi^2_{1-\alpha}[2]}{2n}\right)^{1/2}}.$$

Note that this interval is not symmetric around S. Let $\underline{\sigma}_n$ and $\bar{\sigma}_n$ denote the lower and upper limits of the σ interval. For each σ within this interval we determine a μ interval symmetrically around $\bar{X}$, as specified above. Consider the linear combination $\lambda = a_1\mu + a_2\sigma$, where $a_1 + a_2 = 1$. We can obtain a $(1-\alpha)$-level confidence interval for λ from the region $C_\alpha(\bar{X}, S)$ by determining two lines parallel to $a_1\mu + a_2\sigma = 0$ and tangential to the confidence region $C_\alpha(\bar{X}, S)$. These lines are given by the formula $a_1\mu + a_2\sigma = \underline{\lambda}_\alpha$ and $a_1\mu = a_2\sigma = \bar{\lambda}_\alpha$. The confidence interval is $(\underline{\lambda}_\alpha, \bar{\lambda}_\alpha)$. This interval can be obtained geometrically by projecting $C_\alpha(\bar{X}, S)$ onto the line l spanned by (a_1, a_2)); i.e., $l = \{(\rho a_1, \rho a_2); -\infty < \rho < \infty\}$. ■

PART III: PROBLEMS

Section 6.2

6.2.1 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common exponential distribution, $G(\frac{1}{\theta}, 1)$, $0 < \theta < \infty$. Determine a $(1-\alpha)$-upper confidence limit for $\delta = e^{-\theta}$.

6.2.2 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common Poisson distribution $P(\lambda)$, $0 < \lambda < \infty$. Determine a two-sided confidence interval for λ, at level $1-\alpha$. [Hint: Let $T_n = \Sigma X_i$. Apply the relationship $P_\lambda\{T_n \leq t\} = P\{\chi^2[2t+2] \geq 2n\lambda\}$, $t = 0, 1, \ldots$ to show that $(\underline{\lambda}_\alpha, \bar{\lambda}_\alpha)$ is a $(1-\alpha)$-level confidence interval, where $\underline{\lambda}_\alpha = \frac{1}{2n}\chi^2_{\alpha/2}[2T_n + 2]$ and $\bar{\lambda}_\alpha = \frac{1}{2n}\chi^2_{1-\alpha/2}[2T_n + 2]$.

6.2.3 Let $X_1, \ldots, X_n$ be i.i.d. random variables distributed like $G(\lambda, 1)$, $0 < \lambda < \infty$; and let $Y_1, \ldots, Y_m$ be i.i.d. random variables distributed like $G(\eta, 1)$, $0 < \eta < \infty$. The X-variables and the Y-variables are independent. Determine a $(1-\alpha)$-upper confidence limit for $\omega = (1 + \eta/\lambda)^{-1}$ based on the statistic $\sum_{i=1}^{n} X_i / \sum_{i=1}^{m} Y_i$.

6.2.4 Consider a vector $\mathbf{X}$ of n equicorrelated normal random variables, having zero mean, $\mu = 0$, and variance σ^2 [Problem 1, Section 5.3]; i.e., $\mathbf{X} \sim N(0, \Sigma)$,

where $\mathfrak{V} = \sigma^2(1-\rho)I + \sigma^2\rho J; 0 < \sigma^2 < \infty, \frac{-1}{n-1} < \rho < 1$. Construct a $(1-\alpha)$-level confidence interval for ρ. [Hint:

(i) Make the transformation $\mathbf{Y} = H\mathbf{X}$, where H is a Helmert orthogonal matrix;

(ii) Consider the distribution of $Y_1^2 / \sum_{i=2}^{n} Y_i^2$.

6.2.5 Consider the linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, \ldots, n,$$

where $e_1, \ldots, e_n$ are i.i.d. $N(0, \sigma^2)$, $x_1, \ldots, x_n$ specified constants such that $\Sigma(x_i - \bar{x})^2 > 0$. Determine the formulas of $(1-\alpha)$-level confidence limits for β_0, β_1, and σ^2. To what tests of significance do these confidence intervals correspond?

6.2.6 Let X and Y be independent, normally distributed random variables, $X \sim N(\xi, \sigma_1^2)$ and $Y \sim N(\eta, \sigma_2^2)$; $-\infty < \xi < \infty, 0 < \eta < \infty, \sigma_1$ and σ_2 known. Let $\delta = \xi/\eta$. Construct a $(1-\alpha)$-level confidence interval for δ.

Section 6.3

6.3.1 Prove that if an upper (lower) confidence limit for a real parameter θ is based on a UMP test of $H_0 : \theta \geq \theta_0$ ($\theta \leq \theta_0$) against $H_1 : \theta < \theta_0$ ($\theta > \theta_0$) then the confidence limit is UMA.

6.3.2 Let $X_1, \ldots, X_n$ be i.i.d. having a common two parameter exponential distribution, i.e., $X \sim \mu + G(\frac{1}{\beta}, 1); -\infty < \mu < \infty, 0 < \beta < \infty$.

(i) Determine the $(1-\alpha)$-level UMAU lower confidence limit for μ.

(ii) Determine the $(1-\alpha)$-level UMAU lower confidence limit for β.

[Hint: See Problem 1, Section 4.5.]

6.3.3 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common rectangular distribution $R(0, \theta)$; $0 < \theta < \infty$. Determine the $(1-\alpha)$-level UMA lower confidence limit for θ.

6.3.4 Consider the random effect model, Model II, of ANOVA (Example 3.9). Derive the $(1-\alpha)$-level confidence limits for σ^2 and τ^2. Does this system of confidence intervals have optimal properties?

Section 6.4

6.4.1 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a Poisson distribution $P(\lambda)$, $0 < \lambda < \infty$. Determine a $(p, 1-\alpha)$ guaranteed coverage upper tolerance limit for X.

6.4.2 Consider the normal simple regression model (Problem 7, Section 5.4). Let ξ be a point in the range of controlled experimental levels $x_1, \ldots, x_n$ (regressors). A p-content prediction limit at ξ is the point $\eta_p = \beta_0 + \beta_1 \xi + z_p \sigma$.

(i) Determine a $(p, 1-\alpha)$ guaranteed upper tolerance limit at ξ, i.e., determine $l_{p,\alpha}(\xi)$ so that

$$P_{\underline{\theta}}\left\{\beta_0 + \beta_1 \xi + z_p \sigma \leq \hat{\beta}_0 + \hat{\beta}_1 \xi + l_{p,\alpha}(\xi)\hat{\sigma}\left(\frac{1}{n} + \frac{(\xi - \bar{x})^2}{SDX}\right)^{1/2}\right\} = 1 - \alpha, \quad \text{for all } \underline{\theta} = (\beta_0, \beta_1, \sigma).$$

(ii) What is the form of the asymptotic $(p, 1-\alpha)$-level upper tolerance limit?

Section 6.5

6.5.1 Consider a symmetric continuous distribution $F(x - \mu)$, $-\infty < \mu < \infty$. How large should the sample size n be so that $(X_{(i)}, X_{(n-i+1)})$ is a distribution-free confidence interval for μ, at level $1 - \alpha = 0.95$, when
(i) $i = 1$, **(ii)** $i = 2$, and **(iii)** $i = 3$.

6.5.2 Apply the large sample normal approximation to the binomial distribution to show that for large size random samples from symmetric distribution the $(1-\alpha)$-level distribution free confidence interval for the median is given by $(X_{(i)}, X_{(n-i+1)})$, where $i = [\frac{n}{2} - \frac{1}{2}\sqrt{n}\, z_{1-\alpha}]$ (David, 1970, p. 14).

6.5.3 How large should the sample size n be so that a (p, γ) upper tolerance limit will exist with $p = 0.95$ and $\gamma = 0.95$?

6.5.4 Let $F(x)$ be a continuous c.d.f. and $X_{(1)} \leq \cdots \leq X_{(n)}$ the order statistic of a random sample from such a distribution. Let $F^{-1}(p)$ and $F^{-1}(q)$, with $0 < p < q < 1$, be the pth and qth quantiles of this distribution. Consider the interval $E_{p,q} = (F^{-1}(p), F^{-1}(q))$. Let $p \leq r < s \leq n$. Show that

$$\begin{aligned}\gamma &= P\{E_{p,q} \subset (X_{(r)}, X_{(s)})\} \\ &= \frac{n!}{(r-1)!} \sum_{j=0}^{s-r-1} (-1)^j \frac{p^{r+j}}{(n-r-j)!\, j!} I_{1-q}(n-s+1, s-r-j).\end{aligned}$$

If $q = 1 - \beta/2$ and $p = \beta/2$ then $(X_{(r)}, X_{(s)})$ is a $(1-\beta, \gamma)$ tolerance interval, where γ is given by the above formula.

Section 6.6

6.6.1 In a one-way ANOVA $k = 10$ samples were compared. Each of the samples consisted of $n = 10$ observations. The sample means in order of magnitude were: 15.5, 17.5, 20.2, 23.3, 24.1, 25.5, 28.8, 28.9, 30.1, 30.5. The pooled variance estimate is $s_p^2 = 105.5$. Perform the Scheffé simultaneous testing to determine which differences are significant at level $\alpha = 0.05$.

6.6.2 $n = 10$ observations Y_{ij} $(i = 1, \ldots, 3; j = 1, \ldots, n)$ were performed at three values of x. The sample statistics are:

	$x_1 = 0$	$x_2 = 1.5$	$x_3 = 3.0$
$\bar{Y}$	5.5	9.7	17.3
SDY	13.7	15.8	14.5

(i) Determine the LSEs of β_0, β_1, and σ^2 for the model: $Y_{ij} = \beta_0 + \beta_1 x_i + e_{ij}$, where $\{e_{ij}\}$ are i.i.d. $N(0, \sigma^2)$.

(ii) Determine simultaneous confidence intervals for $E\{Y\} = \beta_0 + \beta_1 x$, for **all** $0 \leq x \leq 3$, using the Scheffé's s-method.

Section 6.7

6.7.1 Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables having a common log-normal distribution, $LN(\mu, \sigma^2)$. Consider the problem of estimating $\xi = \exp\{\mu + \sigma^2/2\}$. The **proportional-closeness** of an estimator, $\hat{\xi}$, is defined as $P_\theta\{|\hat{\xi} - \xi| < \lambda\xi\}$, where λ is a specified positive real.

(i) Show that with a fixed sample procedure, there exists no estimator, $\hat{\xi}$, such that the proportional-closeness for a specified λ is at least γ, $0 < \gamma < 1$.

(ii) Develop a two-stage procedure so that the estimator $\hat{\xi}_N$ will have the prescribed proportional-closeness.

6.7.2 Show that if $\mathcal{F}$ is a family of distribution function depending on a location parameter of the translation type, i.e., $F(x; \theta) = F_0(x - \theta)$, $-\infty < \theta < \infty$, then there exists a fixed width confidence interval estimator for θ.

6.7.3 Let $X_1, \ldots, X_n$ be i.i.d. having a rectangular distribution $R(0, \theta)$, $0 < \theta < 2$. Let $X_{(n)}$ be the sample maximum, and consider the fixed-width interval estimator $I_\delta(X_{(n)}) = (X_{(n)}, X_{(n)} + \delta)$, $0 < \delta < 1$. How large should n be so that $P_\theta\{\theta \in I_\delta(X_{(n)})\} \geq 1 - \alpha$, for all $\theta \leq 2$?

6.7.4 Consider the following three-stage sampling procedure for estimating the mean of a normal distribution. Specify a value of δ, $0 < \delta < \infty$.

(i) Take a random sample of size n_1. Compute the sample variance $S^2_{n_1}$. If $n_1 > (a^2/\delta^2)S^2_{n_1}$, where $a^2 = \chi^2_{1-\alpha}[1]$, terminate sampling. Otherwise, add an independent sample of size

$$N_2 = \left\lfloor \frac{a^2}{\delta^2} S^2_{n_1} \right\rfloor + 1 - n_1.$$

(ii) Compute the pooled sample variance, $S^2_{n_1+N_2}$. If $n_1 + N_2 \geq (a^2/\delta^2)S^2_{n_1+N_2}$ terminate sampling; otherwise, add

$$N_3 = \left[\frac{a^2}{\delta^2} S^2_{n_1+N_2} \right] + 1 - (n_1 + N_2)$$

independent observations. Let $N = n_1 + N_2 + N_3$. Let $\bar{X}_N$ be the average of the sample of size N and $I_\delta(\bar{X}_n) = (\bar{X}_N - \delta, \bar{X}_N + \delta)$.

(i) Compute $P_\theta\{\mu \in I_\delta(\bar{X}_N)\}$ for $\theta = (\mu, \sigma)$.

(ii) Compute $E_\theta\{N\}$.

PART IV: SOLUTION TO SELECTED PROBLEMS

6.2.2 $X_1, \ldots, X_n$ are i.i.d. $P(\lambda)$. $T_n = \sum_{i=1}^{n} X_i \sim P(n\lambda)$

$$\begin{aligned} P\{T_n \leq t\} &= P(t; n\lambda) \\ &= 1 - P(G(1, t+1) < n\lambda) \\ &= P\{\chi^2[2t+2] \geq 2n\lambda\}. \end{aligned}$$

The UMP test of $H_0 : \lambda \geq \lambda_0$ against $H_1 : \lambda < \lambda_0$ is $\phi(T_n) = I(T_n < t_\alpha)$. Note that $P(\chi^2[2t_\alpha + 2] \geq 2n\lambda_0) = \alpha$ if $2n\lambda_0 = \chi^2_{1-\alpha}[2t_\alpha + 2]$. For two-sided confidence limits, we have $\underline{\lambda}_\alpha = \dfrac{\chi^2_{\alpha/2}[2T_n + 2]}{2n}$ and $\bar{\lambda}_\alpha = \dfrac{\chi^2_{1-\alpha/2}[2T_n + 2]}{2n}$.

6.2.4 Without loss of generality assume that $\sigma^2 = 1$

$$\mathbf{X} \sim N(\mathbf{0}, (1-\rho)I + \rho J),$$

where $-\dfrac{1}{n-1} < \rho < 1$, n is the dimension of $\mathbf{X}$, and $J = 1_n 1_n'$.

(i) The Helmert transformation yields $\mathbf{Y} = H\mathbf{X}$, where

$$\mathbf{Y} \sim N(\mathbf{0}, H(1-\rho)I + \rho J)H').$$

Note that $H((1-\rho)I + \rho J)H' = \text{diag}((1-\rho) + n\rho, (1-\rho), \ldots, (1-\rho))$.

(ii) $Y_1^2 \sim ((1-\rho) + n\rho)\chi_1^2[1]$ and $\sum_{i=1}^{n} Y_j^2 \sim (1-\rho)\chi_2^2[n-1]$, where χ_1^2 and $\chi_2^2[n-1]$ are independent. Thus,

$$W = \frac{Y_1^2(n-1)}{\sum_{j=2}^{n} Y_j^2} \sim \left(1 + \frac{n\rho}{1-\rho}\right) F[1, n-1].$$

Hence, for a given $0 < \alpha < 1$,

$$P\left\{\left(1 + \frac{n\rho}{1-\rho}\right) F_{\alpha/2}[1, n-1] \le W \le \left(1 + \frac{n\rho}{1-\rho}\right) F_{1-\alpha/2}[1, n-1]\right\} = 1 - \alpha.$$

Recall that $F_{\alpha/2}[1, n-1] = \dfrac{1}{F_{1-\alpha/2}[n-1, 1]}$. Let

$$R_{n,\alpha} = \frac{1}{n}(WF_{1-\alpha/2}[n-1, 1] - 1)$$

$$R_{n,\alpha}^{-} = \frac{1}{n}\left(\frac{W - F_{1-\alpha/2}[1, n-1]}{F_{1-\alpha/2}[1, n-1]}\right).$$

Since $\rho/(1-\rho)$ is a strictly increasing function of ρ, the confidence limits for ρ are

$$\text{Lower limit} = \max\left(-\frac{1}{n-1}, \frac{R_{n,\alpha}^{-}}{1 + R_{n,\alpha}^{-}}\right).$$

$$\text{Upper limit} = \frac{R_{n,\alpha}}{1 + R_{n,\alpha}}.$$

6.2.6 The method used here is known as Fieller's method. Let $U = X - \delta Y$. Accordingly, $U \sim N(0, \sigma_1^2 + \delta^2\sigma_2^2)$ and

$$\frac{(X - \delta Y)^2}{\sigma_1^2 + \delta^2\sigma_2^2} \sim \chi^2[1].$$

It follows that there are two real roots (if they exist) of the quadratic equation in δ,

$$\delta^2(Y^2 - \chi^2_{1-\alpha}[1]\sigma_2^2) - 2\delta XY + (X^2 - \chi^2_{1-\alpha}[1]\sigma_1^2) = 0.$$

These roots are given by

$$\delta_{1,2} = \frac{XY}{Y^2 - \sigma_2^2\chi^2_{1-\alpha}[1]} \pm z_{1-\alpha/2}\frac{|Y|\sigma_1}{Y^2 - \sigma_2^2\chi^2_{1-\alpha}[1]} \cdot$$
$$\cdot \left[1 + \frac{\sigma_2^2}{\sigma_1^2}\left(\frac{X}{Y}\right)^2 - \frac{\sigma_2^2\chi^2_{1-\alpha}[1]}{Y^2}\right]^{1/2}.$$

It follows that if $Y^2 \geq \sigma_2^2\chi^2_{1-\alpha}[1]$ the two real roots exist. These are the confidence limits for δ.

6.4.1 The m.s.s. for λ is $T_n = \sum_{i=1}^{n} X_i$. A p-quantile of $\mathcal{P}(\lambda)$ is

$$\begin{aligned} X_p(\lambda) &= \min\{j \geq 0 : P(j;\lambda) \geq p\} \\ &= \min\{j \geq 0 : \chi^2_{1-p}[2j+2] \geq 2\lambda\}. \end{aligned}$$

Since $\mathcal{P}(\lambda)$ is an MLR family in X, the $(p, 1-\alpha)$ guaranteed upper tolerance limit is

$$L_{\alpha,p}(T_n) = X_p(\bar{\lambda}_\alpha(T_n)),$$

where $\bar{\lambda}_\alpha(T_n) = \frac{1}{2n}\chi^2_{1-\alpha}[2T_n + 2]$ is the upper confidence limit for λ. Accordingly,

$$L_{\alpha,\beta}(T_n) = \min\left\{j \geq 0 : \frac{1}{2}\chi^2_{1-p}[2j+2] \geq \frac{1}{2n}\chi^2_{1-\alpha}[2T_n + 2]\right\}.$$

6.5.1 **(i)** Since F is symmetric, if $i = 1$, then $j = n$. Thus,

$$I_{0.5}(1, n+1) \geq 0.975,$$
$$\begin{aligned} I_{0.5}(1, n+1) &= \sum_{j=1}^{n+1} \binom{n+1}{j} / 2^{n+1} \\ &= 1 - \frac{1}{2^{n+1}} \geq 0.975. \end{aligned}$$

Or,

$$-(n+1)\log 2 \le \log(0.025),$$
$$n \ge \frac{-\log(0.025)}{\log(2)} - 1, \quad n = 5.$$

(ii)
$$I_{0.5}(2, n) = 1 - \frac{n-2}{2^{n+1}} \ge 0.975,$$
$$\frac{n+2}{2^{n+1}} \le 0.025.$$

For $n = 8$, $\frac{10}{2^9} = 0.0195$. For $n = 7$, $\frac{9}{2^8} = 0.0352$. Thus, $n = 8$.

(iii)
$$I_{0.5}(3, n-1) = 1 - \frac{1 + n + 1 + n(n+1)/2}{2^{n+1}} \ge 0.975$$
$$= 1 - \frac{1 + \frac{(n+1)(n+2)}{2}}{2^{n+1}} \ge 0.975.$$

Or

$$\frac{2 + (n+1)(n+2)}{2^{n+2}} \le 0.025.$$

For $n = 10$, we get $\frac{2 + 11 \times 12}{2^{12}} = 0.0327$. For $n = 11$, we get $\frac{2 + 12 \times 13}{2^{13}} = 0.0193$. Thus, $n = 11$.

6.5.2
$$I_{0.5}(i, n+2-i) = \sum_{j=i}^{n+1} \frac{\binom{n+1}{j}}{2^{n+1}} \ge 1 - \alpha/2.$$

For large n, by Central Limit Theorem,

$$I_{0.5}(i, n+2-i) = \sum_{j=i}^{n+1} b\left(j; n+1, \frac{1}{2}\right)$$
$$\cong 1 - \Phi\left(\frac{i - \frac{1}{2} - (n+1)/2}{\frac{1}{2}\sqrt{n+1}}\right) \ge 1 - \alpha/2.$$

Thus,

$$i \cong 1 + \frac{n}{2} - \frac{1}{2}\sqrt{n-1}\, z_{1-\alpha/2} \cong \frac{n}{2} - \frac{1}{2}\sqrt{n}\, z_{1-\alpha},$$

$$j = n - i \cong \frac{n}{2} + \frac{1}{2}\sqrt{n+1}\, z_{1-\alpha/2},$$

$$\cong \frac{n}{2} + \frac{1}{2}\sqrt{n}\, z_{1-\alpha/2}.$$

6.7.1 Let $Y_i = \log X_i$, $i = 1, 2, \ldots$. If we have a random sample of fixed size n, then the MLE of ξ is $\hat{\xi}_n = \exp\left\{\bar{Y}_n + \frac{1}{2}\hat{\sigma}^2\right\}$, where $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$.

(i) For a given λ, $0 < \lambda < 1$, the proportional closeness of $\hat{\xi}_n$ is

$$\begin{aligned}\text{PC} &= P\left\{(1-\lambda)\exp\left\{\mu + \frac{1}{2}\sigma^2\right\} \le \exp\left\{\bar{Y}_n + \frac{1}{2}\hat{\sigma}_n^2\right\}\right.\\ &\quad \left.\le (1+\lambda)\exp\left\{\mu + \frac{1}{2}\sigma^2\right\}\right\}\\ &= P\left\{\log(1-\lambda) \le (\bar{Y}_n - \mu) + \frac{1}{2}(\hat{\sigma}_n^2 - \sigma^2) \le \log(1+\lambda)\right\}.\end{aligned}$$

For large values of n, the distribution of $W_n = (\bar{Y}_n - \mu) + \frac{1}{2}(\hat{\sigma}_n^2 - \sigma^2)$ is approximately, by CLT, $N\left(0, \frac{\sigma^2}{n}\left(1 + \frac{\sigma^2}{2}\right)\right)$.

$$\begin{aligned}&\lim_{\sigma^2\to\infty} P\left\{\log(1-\lambda) \le N\left(0, \frac{\sigma^2}{n}\left(1+\frac{\sigma^2}{2}\right)\right) \le \log(1+\lambda)\right\}\\ &\quad = \lim_{\sigma^2\to\infty}\left(\Phi\left(\frac{\log(1+\lambda)\sqrt{n}}{\sigma\sqrt{1+\sigma^2/2}}\right) - \Phi\left(\frac{\log(1-\lambda)\sqrt{n}}{\sigma\sqrt{1+\sigma^2/2}}\right)\right) = 0.\end{aligned}$$

Hence, there exists no fixed sample procedure with PC $\ge \gamma > 0$.

(ii) Consider the following two-stage procedure.

Stage I. Take a random sample of size m. Compute $\bar{Y}_m$ and $\hat{\sigma}_m^2$. Let $\delta = \log(1+\lambda)$. Note that $\delta < -\log(1-\lambda)$. If σ were known, then the proportional closeness would be at least γ if

$$n \ge \frac{\chi_\gamma^2[1]\sigma^2(1+\frac{\sigma}{2})^2}{\delta^2}.$$

Accordingly, we define the stopping variable

$$N = \left\lfloor \frac{\chi^2_\gamma[1]\hat{\sigma}^2(1+\frac{\hat{\sigma}^2}{2})}{\delta^2} \right\rfloor + 1.$$

If $\{N \le m\}$ stop sampling and use $\hat{\xi}_m = \exp\left\{\bar{Y}_m + \frac{1}{2}\hat{\sigma}^2_m\right\}$. On the other hand, if $\{N > m\}$ go to Stage II.

Stage II. Let $N_2 = N - m$. Draw additional N_2 observations, conditionally independent of the initial sample. Combine the two samples and compute $\bar{Y}_N$ and $\hat{\sigma}^2_N$. Stop sampling with $\hat{\xi}_N = \exp\left\{\bar{Y}_n + \frac{1}{2}\hat{\sigma}^2_N\right\}$. The distribution of the total sample size $N_m = \max\{m, N\}$ can be determined in the following manner.

(i)

$$P_{\sigma^2}\{N_m = m\} = P_{\sigma^2}\left\{\hat{\sigma}^2(1+\hat{\sigma}^2/2) \le \frac{m\delta^2}{\chi^2_\gamma[1]}\right\}$$

$$= P\left\{\frac{\sigma^2\chi^2[m-1]}{m-1} + \frac{\sigma^4(\chi^2[m-1])^2}{2(m-1)^2} - \frac{m\delta^2}{\chi^2_\gamma[1]} \le 0\right\}$$

$$= P\left\{\chi^2[m-1] \le \frac{(m-1)(\sqrt{1+2m\delta^2/\chi^2_\gamma[1]}-1)}{\sigma^2}\right\}.$$

For $l = m+1, m+2, \ldots$ let

$$\lambda_m(l) = \frac{(m-1)(\sqrt{1+2l\delta^2/\chi^2_\gamma[1]}-1)}{\sigma^2},$$

then

$$P\{N_m = l\} = P\{\lambda_m(l-1) \le \chi^2[m-1] \le \lambda_m(l)\}.$$

CHAPTER 7

Large Sample Theory for Estimation and Testing

PART I: THEORY

We have seen in the previous chapters several examples in which the exact sampling distribution of an estimator or of a test statistic is difficult to obtain analytically. Large samples yield approximations, called **asymptotic approximations**, which are easy to derive, and whose error decreases to zero as the sample size grows. In this chapter, we discuss asymptotic properties of estimators and of test statistics, such as **consistency**, **asymptotic normality,** and **asymptotic efficiency**. In Chapter 1, we presented results from probability theory, which are necessary for the development of the theory of asymptotic inference. Section 7.1 is devoted to the concept of **consistency** of estimators and test statistics. Section 7.2 presents conditions for the **strong consistency** of the maximum likelihood estimator (MLE). Section 7.3 is devoted to the asymptotic normality of MLEs and discusses the notion of best asymptotically normal (BAN) estimators. In Section 7.4, we discuss second and higher order efficiency. In Section 7.5, we present asymptotic confidence intervals. Section 7.6 is devoted to Edgeworth and saddlepoint approximations to the distribution of the MLE, in the one-parameter exponential case. Section 7.7 is devoted to the theory of asymptotically efficient test statistics. Section 7.8 discusses the Pitman's asymptotic efficiency of tests.

7.1 CONSISTENCY OF ESTIMATORS AND TESTS

Consistency of an estimator is a property, which guarantees that in large samples, the estimator yields values close to the true value of the parameter, with probability close to one. More formally, we define consistency as follows.

Examples and Problems in Mathematical Statistics, First Edition. Shelemyahu Zacks.

Definition 7.1.1. *Let $\{\hat{\theta}_n; n = n_0, n_0 + 1, \ldots\}$ be a sequence of estimators of a parameter θ. $\hat{\theta}_n$ is called consistent if $\hat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$. The sequence is called strongly consistent if $\theta_n \to \theta$ almost surely (a.s.) as $n \to \infty$ for all θ.*

Different estimators of a parameter θ might be consistent. Among the consistent estimators, we would prefer those having asymptotically, smallest mean squared error (MSE). This is illustrated in Example 7.2.

As we shall see later, the MLE is asymptotically most efficient estimator under general regularity conditions.

We conclude this section by defining the consistency property for test functions.

Definition 7.1.2. *Let $\{\phi_n\}$ be a sequence of test functions, for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$. The sequence $\{\phi_n\}$ is called consistent if*

(i) $\overline{\lim_{n\to\infty}} \sup_{\theta\in\Theta_0} E_\theta\{\phi_n(\mathbf{X}_n)\} \leq \alpha$, $0 < \alpha < 1$
and

(ii) $\lim_{n\to\infty} E_\theta\{\phi_n(\mathbf{X}_n)\} = 1$, *for all* $\theta \in \Theta_1$.

A test function ϕ_n satisfying property (i) is called **asymptotically size α test**.

All test functions discussed in Chapter 4 are consistent. We illustrate in Example 7.3 a test which is not based on an explicit parametric model of the distribution $F(x)$. Such a test is called a **distribution free** test, or a **nonparametric test**. We show that the test is consistent.

As in the case of estimation, it is not sufficient to have consistent test functions. One should consider asymptotically efficient tests, in a sense that will be defined later.

7.2 CONSISTENCY OF THE MLE

The question we address here is whether the MLE is consistent. We have seen in Example 5.22 a case where the MLE is not consistent; thus, one needs conditions for consistency of the MLE. Often we can prove the consistency of the MLE immediately, as in the case of the MLE of $\boldsymbol{\theta} = (\mu, \sigma)$ in the normal case, or in the Binomial and Poisson distributions.

Let $X_1, X_2, \ldots, X_n, \ldots$ be independent identically distributed (i.i.d.) random variables having a p.d.f. $f(x; \theta)$, $\theta \in \Theta$. Let

$$l(\theta; \mathbf{X}_n) = \sum_{i=1}^{n} \log f(X_i; \theta).$$

If θ_0 is the parameter value of the distribution of the Xs, then from the strong law of large numbers (SLLN)

$$\frac{1}{n}(l(\theta;\mathbf{X}_n) - l(\theta_0;\mathbf{X}_n) \xrightarrow{\text{a.s.}} -I(\theta_0,\theta), \tag{7.2.1}$$

as $n \to \infty$, where $I(\theta_0,\theta)$ is the Kullback–Leibler information. Assume that $I(\theta_0,\theta') > 0$ for all $\theta' \neq \theta_0$. Since the MLE, $\hat{\theta}_n$, maximizes the left-hand side of (7.2.1) and since $I(\theta_0,\theta_0) = 0$, we can immediately conclude that if Θ contains only a finite number of points, then the MLE is strongly consistent. This result is generalized in the following theorem.

Theorem 7.2.1. *Let $X_1,\ldots,X_n$ be i.i.d. random variables having a p.d.f. $f(x;\theta)$, $\theta \in \Theta$, and let θ_0 be the true value of θ. If*

- ***(i)*** *Θ is compact;*
- ***(ii)*** *$f(x;\theta)$ is upper semi-continuous in θ, for all x;*
- ***(iii)*** *there exists a function $K(x)$, such that $E_{\theta_0}\{|K(X)|\} < \infty$ and $\log f(x;\theta) - \log f(x;\theta_0) \le K(x)$, for all x and θ;*
- ***(iv)*** *for all $\theta \in \Theta$ and sufficiently small $\delta > 0$, $\sup_{|\phi-\theta_0|<\delta} f(x;\phi)$ is measurable in x;*
- ***(v)*** *$f(x;\theta) = f(x;\theta_0)$ for almost all x, implies that $\theta = \theta_0$ (identifiability);*

then the MLE $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$, as $n \to \infty$.

The proof is outlined only. For $\delta > 0$, let $\Theta_\delta = \{\theta : |\theta - \theta_0| \ge \delta\}$. Since Θ is compact so is Θ_δ. Let $U(X;\theta) = \log f(X;\theta) - \log f(X;\theta_0)$. The conditions of the theorem imply (see Ferguson, 1996, p. 109) that

$$P_{\theta_0}\left\{\overline{\lim_{n\to\infty}} \sup_{\theta\in\Theta_\delta} \frac{1}{n}\sum_{i=1}^{n} U(X_i;\theta) \le \sup_{\theta\in\Theta_\delta} \mu(\theta)\right\} = 1,$$

where $\mu(\theta) = -I(\theta_0,\theta) < 0$, for all $\theta \in \Theta_\delta$. Thus, with probability one, for n sufficiently large,

$$\sup_{\theta\in\Theta_\delta} \frac{1}{n}\sum_{i=1}^{n} U(X_i;\theta) \le \sup_{\theta\in\Theta_\delta} \mu(\theta) < 0.$$

But,

$$\frac{1}{n}\sum_{i=1}^{n} U(X_i;\hat{\theta}_n) = \sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} U(X_i;\theta) \ge 0.$$

Thus, with probability one, for n sufficiently large, $|\hat{\theta}_n - \theta_0| < \delta$. This demonstrates the consistency of the MLE.

For consistency theorems that require weaker conditions, see Pitman (1979, Ch. 8). For additional reading, see Huber (1967), Le Cam (1986), and Schervish (1995, p. 415).

7.3 ASYMPTOTIC NORMALITY AND EFFICIENCY OF CONSISTENT ESTIMATORS

The presentation of concepts and theory is done in terms of real parameter θ. The results are generalized to k-parameters cases in a straightforward manner.

A consistent estimator $\hat{\theta}(\mathbf{X}_n)$ of θ is called **asymptotically normal** if, there exists an increasing sequence $\{c_n\}$, $c_n \nearrow \infty$ as $n \to \infty$, so that

$$c_n(\hat{\theta}(\mathbf{X}_n) - \theta) \xrightarrow{d} N(0, v^2(\theta)), \quad \text{as } n \to \infty. \tag{7.3.1}$$

The function $AV\{\hat{\theta}_n\} = v^2(\theta)/c_n^2$ is called the **asymptotic variance** of $\hat{\theta}(\mathbf{X}_n)$. Let $S(\mathbf{X}_n; \theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i; \theta)$ be the score function and $I(\theta)$ the Fisher information. An estimator $\hat{\theta}_n$ that, under the Cramér–Rao (CR) regularity conditions, satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{I(\theta)\sqrt{n}} S(\mathbf{X}_n; \theta) + o_p(1), \tag{7.3.2}$$

as $n \to \infty$, is called **asymptotically efficient**. Recall that, by the Central Limit Theorem (CLT), $\frac{1}{\sqrt{n}} S(\mathbf{X}_n; \theta) \xrightarrow{d} N(0, I(\theta))$. Thus, efficient estimators satisfying (7.3.2) have the asymptotic property that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right), \quad \text{as } n \to \infty. \tag{7.3.3}$$

For this reason, such asymptotically efficient estimators are also called **BAN** estimators.

We show now a set of conditions under which the MLE $\hat{\theta}_n$ is a BAN estimator.

In Example 1.24, we considered a sequence $\{X_n\}$ of i.i.d. random variables, with $X_1 \sim B(1, \theta)$, $0 < \theta < 1$. In this case, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a strongly consistent estimator of θ. The **variance stabilizing** transformation $g(\bar{X}_n) = 2\sin^{-1}\sqrt{\bar{X}_n}$ is a (strongly) consistent estimator of $\omega = 2\sin^{-1}\sqrt{\theta}$. This estimator is asymptotically normal with an asymptotic variance $AV\{g(\bar{X}_n)\} = \frac{1}{n}$ for all ω.

Although consistent estimators satisfying (7.3.3) are called BAN, one can construct sometimes asymptotically normal consistent estimators, which at some θ values have an asymptotic variance, with $v^2(\theta) < \frac{1}{I(\theta)}$. Such estimators are called **super efficient**. In Example 7.5, we illustrate such an estimator.

Le Cam (1953) proved that the set of point on which $v^2(\theta) < \frac{1}{I(\theta)}$ has a Lebesgue measure zero, as in Example 7.5.

The following are sufficient conditions for a **consistent** MLE to be a BAN estimator.

C.1. The CR regularity conditions hold (see Theorem 5.2.2);

C.2. $\frac{\partial}{\partial\theta} S(X_n;\theta)$ is continuous in θ, a.s.;

C.3. $\hat{\theta}_n$ exists, and $S(\mathbf{X}_n;\hat{\theta}_n) = 0$ with probability greater than $1-\delta$, $0 < \delta$ arbitrary, for n sufficiently large.

C.4. $\frac{1}{n} \cdot \frac{\partial}{\partial\theta} S(\mathbf{X}_n;\hat{\theta}_n) \xrightarrow{p} -I(\theta), \quad$ as $n \to \infty$.

Theorem 7.3.1 (Asymptotic efficiency of MLE). *Let $\hat{\theta}_n$ be an MLE of θ then, under conditions C.1.–C.4.*

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, 1/I(\theta)), \quad \text{as } n \to \infty.$$

Sketch of the Proof. Let $B_{\delta,\theta,n}$ be a Borel set in $\mathcal{B}^n$ such that, for all $\mathbf{X}_n \in B_{\delta,\theta,n}$, $\hat{\theta}_n$ exists and $S(\mathbf{X}_n;\hat{\theta}_n) = 0$. Moreover, $P_\theta(B_{\delta,\theta,n}) \geq 1-\delta$. For $\mathbf{X}_n \in B_{\delta,\theta,n}$, consider the expansion

$$S(\mathbf{X}_n;\hat{\theta}_n) = S(\mathbf{X}_n;\theta) + (\hat{\theta}_n - \theta) \cdot \frac{\partial}{\partial\theta} S(\mathbf{X}_n;\theta_n^*),$$

where $|\theta_n^* - \theta| \leq |\hat{\theta}_n - \theta|$.

According to conditions (iii)–(v) in Theorem 7.2.1, and Slutzky's Theorem,

$$\sqrt{n}(\hat{\theta}_n - \theta) = -\frac{\frac{1}{\sqrt{n}} S(\mathbf{X};\theta)}{\frac{1}{n} \cdot \frac{\partial}{\partial\theta} S(\mathbf{X}_n;\theta_n^*)} \xrightarrow{d} N(0, 1/I(\theta)),$$

as $n \to \infty$, since by the CLT

$$\frac{1}{\sqrt{n}} S(\mathbf{X}_n;\theta) \xrightarrow{d} N(0, I(\theta)),$$

as $n \to \infty$.

7.4 SECOND-ORDER EFFICIENCY OF BAN ESTIMATORS

Often BAN estimators $\hat{\theta}_n$ are biased, with

$$E_\theta\{\hat{\theta}_n\} = \theta + \frac{b}{n} + o\left(\frac{1}{n}\right), \quad \text{as } n \to \infty. \tag{7.4.1}$$

The problem then is how to compare two different BAN estimators of the same parameter. Due to the bias, the asymptotic variance may not present their precision correctly, when the sample size is not extremely large. Rao (1963) suggested to adjust first an estimator $\hat{\theta}_n$ to reduce its bias to an order of magnitude of $1/n^2$. Let $\hat{\theta}_n^*$ be the adjusted estimator, and let

$$V_\theta\{\hat{\theta}_n^*\} = \frac{1}{nI(\theta)} + \frac{D}{n^2} + o\left(\frac{1}{n^2}\right), \quad \text{as } n \to \infty. \tag{7.4.2}$$

The coefficient D of $1/n^2$ is called the **second-order deficiency** coefficient. Among two BAN estimators, we prefer the one having a smaller second-order deficiency coefficient.

Efron (1975) analyzed the structure of the second-order coefficient D in exponential families in terms of their curvature, the Bhattacharyya second-order lower bound, and the bias of the estimators. Akahira and Takeuchi (1981) and Pfanzagl (1985) established the structure of the distributions of asymptotically high orders most efficient estimators. They have shown that under the CR regularity conditions, the distribution of the most efficient second-order estimator θ_n^* is

$$P\{\sqrt{nI(\theta)}\,(\theta_n^* - \theta) \le t\} = \Phi(t) + \frac{3J_{1,2}(\theta) + 2J_3(\theta)}{6\sqrt{n}\, I^{3/2}(\theta)} t^2 \phi(t) + o\left(\frac{1}{\sqrt{n}}\right), \tag{7.4.3}$$

where

$$J_{1,2}(\theta) = E_\theta\left\{S(X;\theta) \cdot \frac{\partial^2}{\partial\theta^2} \log f(X;\theta)\right\}, \tag{7.4.4}$$

and

$$J_3(\theta) = E_\theta\{S^3(X;\theta)\}. \tag{7.4.5}$$

For additional reading, see also Barndorff-Nielsen and Cox (1994).

7.5 LARGE SAMPLE CONFIDENCE INTERVALS

Generally, the large sample approximations to confidence limits are based on the MLEs of the parameter(s) under consideration. This approach is meaningful in cases where the MLEs are known. Moreover, under the regularity conditions given in the theorem of Section 7.3, the MLEs are BAN estimators. Accordingly, if the sample size is large, one can in regular cases employ the BAN property of MLE to construct confidence intervals around the MLE. This is done by using the quantiles of the standard normal distribution, and the square root of the inverse of the Fisher information function as the standard deviation of the (asymptotic) sampling distribution. In many situations, the inverse of the Fisher information function depends on the unknown parameters. The practice is to substitute for the unknown parameters their respective MLEs. If the samples are very large this approach may be satisfactory. However, as will be shown later, if the samples are not very large it may be useful to apply first a **variance stabilizing transformation** $g(\theta)$ and derive the confidence limits of $g(\theta)$.

A transformation $g(\theta)$ is called **variance stabilizing** if $g'(\theta) = \sqrt{I(\theta)}$. If $\hat{\theta}_n$ is an MLE of θ then $g(\hat{\theta}_n)$ is an MLE of $g(\theta)$. The asymptotic variance of $g(\hat{\theta}_n)$ under the regularity conditions is $(g'(\theta))^2/nI(\theta)$. Accordingly, if $g'(\theta) = \sqrt{I(\theta)}$ then the asymptotic variance of $g(\hat{\theta}_n)$ is $\frac{1}{n}$. For example, suppose that $X_1, \ldots, X_n$ is a sample of n i.i.d. binomial random variables, $B(1, \theta)$. Then, the MLE of θ is $\bar{X}_n$. The Fisher information function is $I_n(\theta) = n/\theta(1-\theta)$. If $g(\theta) = 2\sin^{-1}\sqrt{\theta}$ then $g'(\theta) = 1/\sqrt{\theta(1-\theta)}$. Hence, the asymptotic variance of $g(\bar{X}_n) = 2\sin^{-1}\sqrt{\bar{X}_n}$ is $\frac{1}{n}$. Transformations stabilizing whole covariance matrices are discussed in the paper of Holland (1973).

Let $\theta = t(g)$ be the inverse of a variance stabilizing transformation $g(\theta)$, and suppose (without loss of generality) that $t(g)$ is strictly increasing. For cases satisfying the BAN regularity conditions, if $\hat{\theta}_n$ is the MLE of θ,

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \xrightarrow{d} N(0, 1), \quad \text{as } n \to \infty. \tag{7.5.1}$$

A $(1-\alpha)$ confidence interval for $g(\theta)$ is given asymptotically by $(g(\hat{\theta}_n) - z_{1-\alpha/2}/\sqrt{n}, g(\hat{\theta}_n) + z_{1-\alpha/2}/\sqrt{n})$, where $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$. Let g_L and g_U denote these lower and upper confidence intervals. We assume that both limits are within the range of the function $g(\theta)$; otherwise, we can always truncate it in an appropriate manner. After obtaining the limits g_L and g_U we make the inverse transformation on these limits and thus obtain the limits $\theta_L = t(g_L)$ and $\theta_U = t(g_U)$. Indeed, since $t(g)$ is a one-to-one increasing transformation,

$$\begin{aligned} P_\theta\{\theta_L \le \theta \le \theta_U\} &= P_\theta\{g_L \le g(\theta) \le g_U\} \\ &= P_\theta\left\{g(\theta) - \frac{z_{1-\alpha/2}}{\sqrt{n}} \le g(\hat{\theta}_n) \le g(\theta) + \frac{z_{1-\alpha/2}}{\sqrt{n}}\right\} \approx 1-\alpha. \end{aligned} \tag{7.5.2}$$

Thus, (θ_L, θ_U) is an asymptotically $(1-\alpha)$-confidence interval.

7.6 EDGEWORTH AND SADDLEPOINT APPROXIMATIONS TO THE DISTRIBUTION OF THE MLE: ONE-PARAMETER CANONICAL EXPONENTIAL FAMILIES

The asymptotically normal distributions for the MLE require often large samples to be effective. If the samples are not very large one could try to modify or correct the approximation by the Edgeworth expansion. We restrict attention in this section to the one-parameter exponential type families in canonical form.

According to (5.6.2), the MLE, $\hat{\psi}_n$, of the canonical parameter ψ satisfies the equation

$$K'(\hat{\psi}_n) = \frac{1}{n}\sum_{i=1}^{n} U(X_i) = \bar{U}_n. \tag{7.6.1}$$

The cumulant generating function $K(\psi)$ is analytic. Let $G(x)$ be the inverse function of $K'(\psi)$. $G(x)$ is also analytic and one can write, for large samples,

$$\begin{aligned}\hat{\psi}_n &= G(\bar{U}_n) \\ &= G(K'(\psi)) + (\bar{U}_n - K'(\psi))G'(K'(\psi)) + o_p\left(\frac{1}{\sqrt{n}}\right) \\ &= \psi + (\bar{U}_n - K'(\psi))/K''(\psi) + o_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned} \tag{7.6.2}$$

Recall that $K''(\psi) = I(\psi)$ is the Fisher information function, and for large samples,

$$\sqrt{nI(\psi)}\,(\hat{\psi}_n - \psi) = \sqrt{n}\,\frac{\bar{U}_n - K'(\psi)}{\sqrt{I(\psi)}} + o_p(1). \tag{7.6.3}$$

Moreover, $E\{\bar{U}_n\} = K'(\psi)$ and $V\{\sqrt{n}\,\bar{U}_n\} = I(\psi)$. Thus, by the CLT,

$$\sqrt{n}\,\frac{\bar{U}_n - K'(\psi)}{\sqrt{I(\psi)}} \xrightarrow{d} N(0, 1), \quad \text{as } n \to \infty. \tag{7.6.4}$$

Equivalently,

$$\sqrt{n}\,(\hat{\psi}_n - \psi) \xrightarrow{d} N\left(0, \frac{1}{I(\psi)}\right), \quad \text{as } n \to \infty. \tag{7.6.5}$$

This is a version of Theorem 7.3.1, in the present special case.

If the sample is not very large, we can add terms to the distribution of $\sqrt{nI(\psi)}\,(\hat{\psi}_n - \psi)$ according to the Edgeworth expansion. We obtain

$$P\{\sqrt{nI(\psi)}\,(\hat{\psi}_n - \psi) \le x\} \cong \Phi(x) - \frac{\beta_1}{6\sqrt{n}}(x^2 - 1)\phi(x) - \frac{x}{n}\left[\frac{\beta_2 - 3}{24}(x^2 - 3) + \frac{\beta_1^2}{72}(x^4 - 10x^2 + 15)\right]\phi(x), \tag{7.6.6}$$

where

$$\beta_1 = \frac{K^{(3)}(\psi)}{(K^{(2)}(\psi))^{3/2}}, \tag{7.6.7}$$

and

$$\beta_2 - 3 = \frac{K^{(4)}(\psi)}{(K^{(2)}(\psi))^2}. \tag{7.6.8}$$

Let $T_n = \sum_{i=1}^{n} U(X_i)$. T_n is the likelihood statistic. As shown in Reid (1988) the saddlepoint approximation to the p.d.f. of the MLE, $\hat{\psi}_n$, is

$$g_{\hat{\psi}_n}(x;\psi) = c_n(K^{(2)}(x))^{1/2}\exp\{-(x-\psi)T_n - n(K(\psi) - K(x))\}(1 + O(n^{-3/2})), \tag{7.6.9}$$

where c_n is a factor of proportionality, such that $\int g_{\hat{\psi}_n}(x;\psi)d\mu(x) = 1$.

Let $L(\theta;\mathbf{X}_n)$ and $l(\theta;\mathbf{X}_n)$ denote the likelihood and log-likelihood functions. Let $\hat{\theta}_n$ denotes the MLE of θ, and

$$J_n(\hat{\theta}_n) = -\frac{1}{n}\cdot\frac{\partial^2}{\partial\theta^2}l(\theta;\mathbf{X}_n)\big|_{\theta=\hat{\theta}_n}. \tag{7.6.10}$$

We have seen that $E_\theta\{J_n(\theta)\} = I(\theta)$. Thus, $J_n(\theta) \xrightarrow{\text{a.s.}} I(\theta)$, as $n \to \infty$ (the Fisher information function). $J_n(\hat{\theta}_n)$ is an MLE estimator of $J_n(\theta)$. Thus, if $\hat{\theta}_n \xrightarrow{p} \theta$, as $n \to \infty$, then, as in condition C.4. of Theorem 7.3.1, $J_n(\hat{\theta}_n) \xrightarrow{p} I(\theta)$, as $n \to \infty$. The saddlepoint approximation to $g_{\hat{\theta}_n}(x;\theta)$ in the general regular case is

$$g_{\hat{\theta}_n}(x;\theta) = c_n(J_n(\hat{\theta}))^{1/2}\frac{L(\theta;\mathbf{X}_n)}{L(\hat{\theta}_n;\mathbf{X}_n)}. \tag{7.6.11}$$

Formula (7.6.11) is called the **Barndorff-Nielsen p^*-formula**. The order of magnitude of its error, in large samples, is $O(n^{-3/2})$.

7.7 LARGE SAMPLE TESTS

For testing two simple hypotheses there exists a most powerful test of size α. We have seen examples in which it is difficult to determine the exact critical level k_α of the test. Such a case was demonstrated in Example 4.4. In that example, we have used the asymptotic distribution of the test statistic to approximate k_α. Generally, if $X_1, \ldots, X_n$ are i.i.d. with common p.d.f. $f(x;\boldsymbol{\theta})$ let

$$R(X) = \frac{f(X;\boldsymbol{\theta}_1)}{f(X;\boldsymbol{\theta}_0)}, \tag{7.7.1}$$

where the two sample hypotheses are $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$. The most powerful test of size α can be written as

$$\phi(\mathbf{X}_n) = \begin{cases} 1, & \text{if } \sum_{i=1}^{n} \log R(X_i) > k_\alpha, \\ \gamma_\alpha, & \text{if } \sum_{i=1}^{n} \log R(X_i) = k_\alpha, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, in large samples we can consider the test function $\phi(S_n) = I\{S_n \geq k_\alpha\}$, where $S_n = \sum_{i=1}^{n} \log R(X_i)$. Note that under H_0, $E_{\boldsymbol{\theta}_0}\{S_n\} = -nI(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ while under H_1, $E_{\boldsymbol{\theta}_1}\{S_n\} = nI(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0)$, where $I(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is the Kullback–Leibler information.

Let $\sigma_0^2 = V_{\theta_0}\{\log R(X_1)\}$. Assume that $0 < \sigma_0^2 < \infty$. Then, by the CLT, $\lim_{n\to\infty} P_{\boldsymbol{\theta}_0}\left\{\frac{S_n + nI(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)}{\sqrt{n}\,\sigma_0} \leq x\right\} = \Phi(x)$. Hence, a large sample approximation to k_α is

$$k_\alpha = -nI(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) + Z_{1-\alpha}\sqrt{n}\,\sigma_0. \tag{7.7.2}$$

The large sample approximation to the power of the test is

$$\psi(\boldsymbol{\theta}_1, \sigma_1) = \Phi\left(\sqrt{n}\,\frac{I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) + I(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0)}{\sigma_1} - Z_{1-\alpha}\frac{\sigma_0}{\sigma_1}\right), \tag{7.7.3}$$

where $\sigma_1^2 = V_{\boldsymbol{\theta}_1}\{\log R(X_1)\}$. Generally, for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}$ is a k-parameter vector, the following three test statistics are in common use, in cases satisfying the CR regularity condition:

1. The Wald Statistic

$$Q_w = n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)' J(\hat{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0), \tag{7.7.4}$$

where $\hat{\boldsymbol{\theta}}_n$ is the MLE of $\boldsymbol{\theta}$, and

$$J(\hat{\theta}_n) = -H(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n}. \tag{7.7.5}$$

Here, $H(\boldsymbol{\theta})$ is the matrix of partial derivatives

$$H(\boldsymbol{\theta}) = \frac{1}{n}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\sum_{l=1}^{n}\log f(X_l;\boldsymbol{\theta}); i, j = 1, \ldots, k\right).$$

An alternative statistic, which is asymptotically equivalent to Q_w, is

$$Q_w^* = n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)' J(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0). \tag{7.7.6}$$

One could also use the FIM, $I(\boldsymbol{\theta}_0)$, instead of $J(\boldsymbol{\theta}_0)$.

2. **The Wilks' Likelihood Ratio Statistic**:

$$Q_L = 2\{l(\hat{\boldsymbol{\theta}}_n; \mathbf{X}_n) - l(\boldsymbol{\theta}_0; \mathbf{X}_n)\}. \tag{7.7.7}$$

3. **Rao's Efficient Score Statistic**:

$$Q_R = \frac{1}{n}\mathbf{S}(\mathbf{X}_n;\boldsymbol{\theta}_0)'(J(\boldsymbol{\theta}_0))^{-1}\mathbf{S}(\mathbf{X}_n;\boldsymbol{\theta}_0), \tag{7.7.8}$$

where $\mathbf{S}(\mathbf{X}_n;\boldsymbol{\theta})$ is the score function, namely, the gradient vector $\nabla_\theta \sum_{i=1}^{n}\log f(X_i;\boldsymbol{\theta})$. Q_R does not require the computation of the MLE $\hat{\boldsymbol{\theta}}_n$.

On the basis of the multivariate asymptotic normality of $\hat{\boldsymbol{\theta}}_n$, we can show that all these three test statistics have in the regular cases, under H_0, an asymptotic $\chi^2[k]$ distribution. The asymptotic power function can be computed on the basis of the non-central $\chi^2[k;\lambda]$ distribution.

7.8 PITMAN'S ASYMPTOTIC EFFICIENCY OF TESTS

The Pitman's asymptotic efficiency is an index of the relative performance of test statistics in large samples. This index is called the **Pitman's asymptotic relative efficiency** (ARE). It was introduced by Pitman in 1948.

Let $X_1, \ldots, X_n$ be i.i.d. random variables, having a common distribution $F(x;\theta)$, $\theta \in \Theta$. Let T_n be a statistic. Suppose that there exist functions $\mu(\theta)$ and $\sigma_n(\theta)$ so that, for each $\theta \in \Theta$, $Z_n = (T_n - \mu(\theta))/\sigma_n(\theta) \xrightarrow{d} N(0,1)$, as $n \to \infty$. Often $\sigma_n(\theta) = c(\theta)w(n)$, where $w(n) = n^{-\alpha}$ for some $\alpha > 0$.

Consider the problem of testing the hypotheses $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$, at level $\alpha_n \to \alpha$, as $n \to \infty$. Let the sequence of test functions be

$$\phi_n(T_n) = I\{(T_n - \mu(\theta_0))/\sigma_n(\theta_0) \geq k_n\}, \tag{7.8.1}$$

where $k_n \to Z_{1-\alpha}$. The corresponding power functions are

$$\psi_n(\theta; T_n) \approx \Phi\left(\frac{\mu(\theta) - \mu(\theta_0)}{w(n)c(\theta_0)} \cdot \frac{c(\theta_0)}{c(\theta)} - Z_{1-\alpha}\frac{c(\theta_0)}{c(\theta)}\right). \tag{7.8.2}$$

We assume that

1. $\mu(\theta)$ is continuously differentiable in the neighborhood of θ_0, and $\mu'(\theta_0) > 0$;
2. $c(\theta)$ is continuous in the neighborhood of θ_0, and $c(\theta_0) > 0$.

Under these assumptions, if $\theta_n = \theta_0 + \delta w(n)$ then, with $\delta > 0$,

$$\lim_{n\to\infty} \psi_n(\theta_n; T_n) = \Phi\left(\frac{\delta\mu'(\theta_0)}{c(\theta_0)} - Z_{1-\alpha}\right) = \psi^*. \tag{7.8.3}$$

The function

$$J(\theta; T) = \frac{(\mu'(\theta))^2}{c^2(\theta)} \tag{7.8.4}$$

is called the **asymptotic efficacy** of T_n.

Let V_n be an alternative test statistic, and $W_n = (V_n - \eta(\theta))/(v(\theta)w(n)) \xrightarrow{d} N(0, 1)$, as $n \to \infty$. The asymptotic efficacy of V_n is $J(\theta; V) = (\eta'(\theta))^2/v^2(\theta)$. Consider the case of $w(n) = n^{-1/2}$. Let $\theta_n = \theta_0 + \dfrac{\delta}{\sqrt{n}}$, $\delta > 0$, be a sequence of local alternatives. Let $\psi_n(\theta_n; V_n)$ be the sequence of power functions at $\theta_n = \theta_0 + \delta/\sqrt{n}$ and sample size $n'(n)$ so that $\lim_{n\to\infty} \psi_{n'}(\theta_n; V_{n'}) = \psi^* = \lim_{n\to\infty} \psi_n(\theta_n; T_n)$. For this

$$n'(n) = \frac{nJ(\theta_0; T)}{J(\theta_0; V)} \tag{7.8.5}$$

and

$$\lim_{n\to\infty} \frac{n}{n'(n)} = \frac{J(\theta_0; V)}{J(\theta_0; T)} = \left(\frac{\eta'(\theta_0)}{\mu'(\theta_0)}\right)^2 \frac{c^2(\theta_0)}{v^2(\theta_0)}. \tag{7.8.6}$$

This limit (7.8.6) is the Pitman ARE of V_n relative to T_n.

We remark that the asymptotic distributions of Z_n and W_n do not have to be $N(0, 1)$, but they should be the same. If Z_n and W_n converge to two different distributions, the Pitman ARE is not defined.

7.9 ASYMPTOTIC PROPERTIES OF SAMPLE QUANTILES

Give a random sample of n i.i.d. random variables, the empirical distribution of the sample is

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \le x\}, \quad -\infty < x < \infty. \tag{7.9.1}$$

This is a step function, with jumps of size $1/n$ at the location of the sample random variables $\{X_i, i = 1, \ldots, n\}$. The pth quantile of a distribution F is defined as

$$\xi_p = F^{-1}(p) = \inf\{x : F(x) \ge p\} \tag{7.9.2}$$

according to this definition the quanitles are unique. Similarly, the pth sample quantile are defined as $\xi_{n,p} = F_n^{-1}(p)$.

Theorem 7.9.1. *Let $0 < p < 1$. Suppose that F is differentiable at the pth quantile ξ_p, and $F'(\xi_p) > 0$, then $\xi_{n,p} \to \xi_p$ a.s. as $n \to \infty$.*

Proof. Let $\epsilon > 0$ then

$$F(\xi_p - \epsilon) < p < F(\xi_p + \epsilon).$$

By SLLN

$$F_n(\xi_p - \epsilon) \to F(\xi_p - \epsilon) \text{ a.s., as } n \to \infty$$

and

$$F_n(\xi_p + \epsilon) \to F(\xi_p + \epsilon) \text{ a.s., as } n \to \infty.$$

Hence,

$$P\{F_m(\xi_p - \epsilon) < p < F_m(\xi_p + \epsilon),\ \forall\, m \ge n\} \to 1$$

as $n \to \infty$. Thus,

$$P\{\xi_p - \epsilon < F_m^{-1}(p) < \xi_p + \epsilon,\ \forall\, m \ge n\} \to 1$$

as $n \to \infty$. That is,

$$P\{\sup_{m \ge n} |\xi_{m,p} - \xi_p| > \epsilon\} \to 0, \text{ as } n \to \infty.$$

QED

Note that if $0 < F(\xi) < 1$ then, by CLT,

$$\lim_{n\to\infty} P\left\{\frac{\sqrt{n}(F_n(\xi) - F(\xi))}{\sqrt{F(\xi)(1-F(\xi))}} \le t\right\} = \Phi(t), \tag{7.9.3}$$

for all $-\infty < t < \infty$. We show now that, under certain conditions, $\xi_{n,p}$ is asymptotically normal.

Theorem 7.9.2. *Let $0 < p < 1$. Suppose that F is continuous at $\xi_p = F^{-1}(p)$. Then,*

(i) *If $F'(\xi_p-) > 0$ then, for all $t < 0$,*

$$\lim_{n\to\infty} P\left\{\frac{n^{1/2}(\xi_{p,n} - \xi_p)}{(p(1-p))^{1/2}/F'(\xi_p-)} \le t\right\} = \Phi(t). \tag{7.9.4}$$

(ii) *If $F'(\xi_p+) > 0$ then, for all $t > 0$,*

$$\lim_{n\to\infty} P\left\{\frac{n^{1/2}(\xi_{n,p} - \xi_p)}{\sqrt{p(1-p)}/F'(\xi_p+)} \le t\right\} = \Phi(t). \tag{7.9.5}$$

Proof. Fix t. Let $A > 0$ and define

$$G_n(t) = P\left\{\frac{n^{1/2}(\xi_{n,p} - \xi_p)}{A} \le t\right\}. \tag{7.9.6}$$

Thus,

$$\begin{aligned} G_n(t) &= P\{\xi_{n,p} \le \xi_p + tAn^{-1/2}\} \\ &= P\{p \le F_n(\xi_p + tAn^{-1/2})\}. \end{aligned} \tag{7.9.7}$$

Moreover, since $nF_n(\xi_p) \sim B(n, F(\xi_p))$,

$$G_n(t) = P\left\{p \le \frac{1}{n}B(n, F(\xi_p + tAn^{-1/2})\right\}. \tag{7.9.8}$$

By CLT,

$$P\left\{\frac{B(n, F(\xi_p)) - nF(\xi_p)}{(nF(\xi_p)\bar{F}(\xi_p))^{1/2}} \le Z\right\} \to \Phi(z), \tag{7.9.9}$$

as $n \to \infty$, where $\bar{F}(\xi_p) = 1 - F(\xi_p)$. Let

$$\Delta_n(t) = F(\xi_p + tAn^{-1/2}), \tag{7.9.10}$$

and

$$Z_n^*(\Delta) = \frac{B(n, \Delta) - n\Delta}{\sqrt{n\Delta(1-\Delta)}}. \tag{7.9.11}$$

Then

$$G_n(t) = P\{Z_n^*(\Delta_n(t)) \geq -C_n(t)\}, \tag{7.9.12}$$

where

$$C_n(t) = \frac{n^{1/2}(\Delta_n(t) - p)}{\sqrt{\Delta_n(t)(1-\Delta_n(t))}}. \tag{7.9.13}$$

Since F is continuous at ξ_p,

$$\Delta_n(t)(1 - \Delta_n(t)) \to p(1-p), \quad \text{as } n \to \infty.$$

Moreover, if $t > 0$, $F(\xi_p + tAn^{-1/2}) - F(\xi_p) = \frac{tA}{\sqrt{n}} F'(\xi_p+) + o\left(\frac{1}{\sqrt{n}}\right)$. Hence, if $t > 0$

$$C_n(t) \to \frac{tA}{\sqrt{p(1-p)}} F'(\xi_p+), \quad \text{as } n \to \infty. \tag{7.9.14}$$

Similarly, if $t < 0$

$$C_n(t) \to \frac{tA}{\sqrt{p(1-p)}} F'(\xi_p-), \quad \text{as } n \to \infty. \tag{7.9.15}$$

Thus, let

$$A = \begin{cases} \dfrac{\sqrt{p(1-p)}}{F'(\xi_p-)}, & \text{if } t < 0, \\ \dfrac{\sqrt{p(1-p)}}{F'(\xi_p+)}, & \text{if } t > 0. \end{cases}$$

Then, $\lim_{n\to\infty} C_n(t) = t$. Hence, from (7.9.12),

$$\lim_{n\to\infty} G_n(t) = \Phi(t).$$

QED

Corollary. *If F is differentiable at ξ_p, and $f(\xi_p) = \frac{d}{dx}F(x)\big|_{x=\xi_p} > 0$, then $\xi_{n,p}$ is asymptotically $N\left(\xi_p, \frac{p(1-p)}{nf^2(\xi_p)}\right)$.*

PART II: EXAMPLES

Example 7.1. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables, such that $E\{|X_1|\} < \infty$. By the SLLN, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mu$, as $n \to \infty$, where $\mu = E\{X_1\}$. Thus, the sample mean $\bar{X}_n$ is a strongly consistent estimator of μ. Similarly, if $E\{|X_1|^r\} < \infty$, $r \geq 1$, then the rth sample moment $M_{n,r}$ is strongly consistent estimator of $\mu_r = E\{X_1^r\}$, i.e.,

$$M_{n,r} = \frac{1}{n}\sum_{i=1}^{n} X_i^r \xrightarrow{\text{a.s.}} \mu_r, \quad \text{as } n \to \infty.$$

Thus, if $\sigma^2 = V\{X_1\}$, and $0 < \sigma^2 < \infty$,

$$\hat{\sigma}_n^2 = M_{n,2} - (M_{n,1})^2 \xrightarrow[n\to\infty]{\text{a.s.}} \sigma^2.$$

That is, $\hat{\sigma}_n^2$ is a strongly consistent estimator of σ^2. It follows that $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ is also a strongly consistent estimator of σ^2. Note that, since $M_{n,r} \xrightarrow{\text{a.s.}} \mu_r$, as $n \to \infty$ whenever $E\{|X_1|^r\} < \infty$, then for any continuous function $g(\cdot)$, $g(M_{n,r}) \xrightarrow{\text{a.s.}} g(\mu_r)$, as $n \to \infty$. Thus, if

$$\beta_1 = \frac{\mu_3^*}{(\mu_2^*)^{3/2}}$$

is the coefficient of skewness, the sample coefficient of skewness is strongly consistent estimator of β_1, i.e.,

$$\hat{\beta}_{1,n} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^3}{(\hat{\sigma}_n)^3} \xrightarrow{\text{a.s.}} \beta_1, \quad \text{as } n \to \infty.$$

■

Example 7.2. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables having a rectangular distribution $R(0, \theta)$, $0 < \theta < \infty$. Since $\mu_1 = \theta/2$, $\hat{\theta}_{1,n} = 2\bar{X}_n$ is a **strongly**

consistent estimator of θ. The MLE $\hat{\theta}_{2,n} = X_{(n)}$ is also **strongly consistent estimator** of θ. Actually, since for any $0 < \epsilon < \theta$,

$$P\{\hat{\theta}_{2,n} \leq \theta - \epsilon\} = \left(1 - \frac{\epsilon}{\theta}\right)^n, \quad n \geq 1.$$

Hence, by Borel–Cantelli Lemma, $P\{\hat{\theta}_{2,n} \leq \theta - \epsilon, i.o.\} = 0$. This implies that $\hat{\theta}_{2,n} \xrightarrow{\text{a.s.}} \theta$, as $n \to \infty$. The MLE is **strongly consistent**. The expected value of the MLE is $E\{\hat{\theta}_{2,n}\} = \frac{n}{n+1}\theta$. The variance of the MLE is

$$V\{\hat{\theta}_{2,n}\} = \frac{n\theta^2}{(n+1)^2(n+2)}.$$

The MSE of $\{\hat{\theta}_{2,n}\}$ is $V\{\hat{\theta}_{2,n}\} + \text{Bias}^2\{\hat{\theta}_{2,n}\}$, i.e.,

$$\text{MSE}\{\hat{\theta}_{2,n}\} = \frac{2\theta^2}{(n+1)(n+2)}.$$

The variance of $\hat{\theta}_{1,n}$ is $V\{\hat{\theta}_{1,n}\} = \frac{\theta^2}{3n}$. The relative efficiency of $\hat{\theta}_{1,n}$ against $\hat{\theta}_{2,n}$ is

$$\text{Rel. eff.} = \frac{\text{MSE}\{\hat{\theta}_{2,n}\}}{V\{\hat{\theta}_{1,n}\}} = \frac{6n}{(n+1)(n+2)} \to 0,$$

as $n \to \infty$. Thus, in large samples, $2\bar{X}_n$ is very inefficient estimator relative to the MLE. ■

Example 7.3. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having a **continuous** distribution $F(x)$, **symmetric** around a point θ. θ is obviously the median of the distribution, i.e., $\theta = F^{-1}\left(\frac{1}{2}\right)$. We index these distributions by θ and consider the location family $\mathcal{F}_s = \{F_\theta : F_\theta(x) = F(x - \theta)$, **and** $F(-z) = 1 - F(z); -\infty < \theta < \infty\}$. The functional form of F is not specified in this model. Thus, $\mathcal{F}_s$ is the family of **all** symmetric, continuous distributions. We wish to test the hypotheses

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta > \theta_0.$$

The following test is the **Wilcoxon signed-rank test**:

Let $Y_i = X_i - \theta_0, i = 1, \ldots, n$. Let $S(Y_i) = I\{Y_i > 0\}, i = 1, \ldots, n$. We consider now the ordered absolute values of Y_i, i.e.,

$$|Y|_{(1)} < |Y|_{(2)} < \cdots < |Y|_{(n)}$$

and let $R(Y_i)$ be the index (j) denoting the place of Y_i in the ordered absolute values, i.e., $R(Y_i) = j$, $j = 1, \ldots, n$ if, and only if, $|Y_i| = |Y|_{(j)}$. Define the test statistic

$$T_n = \sum_{i=1}^{n} S(Y_i) R(Y_i).$$

The test of H_0 versus H_1 based on T_n, which rejects H_0 if T_n is sufficiently large is called the Wilcoxon signed-rank test. We show that this test is consistent. Note that under H_0, $P_0\{S(Y_i) = 1\} = \frac{1}{2}$. Moreover, for each $i = 1, \ldots, n$, under H_0

$$\begin{aligned} P_0\{S(Y_i) = 1, |Y_i| \le y\} &= P\{0 \le Y_i \le y\} \\ &= F(y) - \frac{1}{2} \\ &= \frac{1}{2}(2F(y) - 1) \\ &= P_0\{S(Y_i) = 1\} P_0\{|Y_i| \le y\}. \end{aligned}$$

Thus, $S(Y_i)$ and $|Y_i|$ are independent. This implies that, under H_0, $S(Y_1), \ldots, S(Y_n)$ are independent of $R(Y_1), \ldots, R(Y_n)$, and the distribution of T_n, under H_0, is like that of $T_n = \sum_{j=1}^{n} j W_j \sim \sum_{j=1}^{n} j B\left(1, \frac{1}{2}\right)$. It follows that, under H_0,

$$\mu_0 = E_0\{T_n\} = \frac{1}{2} \sum_{j=1}^{n} j = \frac{n(n+1)}{4}.$$

Similarly, under H_0,

$$\begin{aligned} V_0\{T_n\} &= \frac{1}{4} \sum_{j=1}^{n} j^2 \\ &= \frac{n(n+1)(2n+1)}{24}. \end{aligned}$$

According to Problem 3 of Section 1.12, the CLT holds, and

$$P_0\left\{\frac{T_n - \mu_0}{\sqrt{V_0\{T_n\}}} \le x\right\} \to \Phi(x).$$

Thus, the test function

$$\phi(T_n) = \begin{cases} 1, & \text{if } T_n \geq \dfrac{n(n+1)}{4} + Z_{1-\alpha}\sqrt{\dfrac{n(n+1)(2n+1)}{24}} \\ 0, & \text{otherwise} \end{cases}$$

has, asymptotically size α, $0 < \alpha < 1$. This establishes part (i) of the definition of consistency.

When $\theta > \theta_0$ (under H_1) the distribution of T_n is more complicated. We can consider the test statistic $V_n = \dfrac{T_n}{n(n+1)}$. One can show (see Hettmansperger, 1984, p. 47) that the asymptotic mean of V_n, as $n \to \infty$, is $\dfrac{1}{2}p_2(\theta)$, and the asymptotic variance of V_n is

$$AV = \frac{1}{n}(p_4(\theta) - p_2^2(\theta)),$$

where

$$p_2(\theta) = P_\theta\{Y_1 + Y_2 > 0\},$$
$$p_4(\theta) = P_\theta\{Y_1 + Y_2 > 0, Y_1 + Y_3 > 0\}.$$

In addition, one can show that the asymptotic distribution of V_n (under H_1) is normal (see Hettmansperger, 1984).

$$P_\theta\left\{T_n > \frac{n(n+1)}{4} + Z_{1-\alpha}\sqrt{\frac{n(n+1)(2n+1)}{24}}\right\}$$
$$= P_\theta\left\{V_n > \frac{1}{4} + Z_{1-\alpha}\sqrt{\frac{2n+1}{24n(n+1)}}\right\}.$$

Finally, when $\theta > 0$, $p_2(\theta) > \dfrac{1}{2}$ and

$$\lim_{n\to\infty} P_\theta\left\{T_n > \frac{n(n+1)}{4} + Z_{1-\alpha}\sqrt{\frac{n(n+1)(2n+1)}{24}}\right\}$$
$$= 1 - \lim_{n\to\infty} \Phi\left(\frac{\sqrt{n}\left(\dfrac{1}{4} - \dfrac{1}{2}p_2(\theta)\right)}{\sqrt{p_4(\theta) - p_2^2(\theta)}}\right) = 1,$$

for all $\theta > 0$. Thus, the Wilcoxon signed-rank test is consistent. ■

Example 7.4. Let $T_1, T_2, \ldots, T_n$ be i.i.d. random variables having an exponential distribution with mean β, $0 < \beta < \infty$. The observable random variables are $X_i = \min(T_i, t^*)$, $i = 1, \ldots, n$; $0 < t^* < \infty$. This is the case of Type I censoring of the random variables $T_1, \ldots, T_n$.

The likelihood function of β, $0 < \beta < \infty$, is

$$L(\beta; \mathbf{X}_n) = \frac{1}{\beta^{K_n}} \exp\left\{-\frac{1}{\beta}\sum_{i=1}^{n} X_i\right\},$$

where $K_n = \sum_{i=1}^{n} I\{X_i < t^*\}$. Note that the MLE of β does not exist if $K_n = 0$. However, $P\{K_n = 0\} = e^{-nt^*/\beta} \to 0$ as $n \to \infty$. Thus, for sufficiently large n, the MLE of β is

$$\hat{\beta}_n = \frac{\sum_{i=1}^{n} X_i}{K_n}.$$

Note that by the SLLN,

$$\frac{1}{n} K_n \xrightarrow{\text{a.s.}} 1 - e^{-t^*/\beta}, \quad \text{as } n \to \infty,$$

and

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} E\{X_1\}, \quad \text{as } n \to \infty.$$

Moreover,

$$\begin{aligned} E\{X_1\} &= \frac{1}{\beta}\int_0^{t^*} te^{-t/\beta}dt + t^*e^{-t^*/\beta} \\ &= \beta\left(1 - P\left(1; \frac{t^*}{\beta}\right)\right) + t^*e^{-t^*/\beta} \\ &= \beta(1 - e^{-t^*/\beta}). \end{aligned}$$

Thus, $\hat{\beta}_n \xrightarrow{\text{a.s.}} \beta$, as $n \to \infty$. This establishes the strong consistency of $\hat{\beta}_n$. ■

Example 7.5. Let $\{X_n\}$ be a sequence of i.i.d. random variables, $X_1 \sim N(\theta, 1)$, $-\infty < \theta < \infty$. Given a sample of n observations, the minimal sufficient statistic

is $\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$. The Fisher information function is $I(\theta) = 1$, and $\bar{X}_n$ is a BAN estimator. Consider the estimator,

$$\hat{\theta}_n = \begin{cases} \frac{1}{2}\bar{X}_n, & \text{if } |\bar{X}_n| \le \frac{\log n}{\sqrt{n}} \\ \bar{X}_n, & \text{otherwise.} \end{cases}$$

Let

$$\Lambda_n = \left\{ \mathbf{X}_n : |\bar{X}_n| \le \frac{\log n}{\sqrt{n}} \right\}.$$

Now,

$$P_\theta\{\Lambda_n\} = \begin{cases} 2\Phi(\log n) - 1, & \text{if } \theta = 0, \\ \Phi(\log n - \sqrt{n}\,\theta) + \Phi(\log n + \sqrt{n}\,\theta) - 1, & \text{if } \theta \neq 0. \end{cases}$$

Thus,

$$\lim_{n\to\infty} P_\theta\{\Lambda_n\} = \begin{cases} 1, & \text{if } \theta = 0, \\ 0, & \text{if } \theta \neq 0. \end{cases}$$

We show now that $\hat{\theta}_n$ is consistent. Indeed, for any $\delta > 0$,

$$P_\theta\{|\hat{\theta}_n - \theta| > \delta\} = P_\theta\{|\hat{\theta}_n - \theta| > \delta, I_{\Lambda_n}(\bar{X}_n) = 1\} + P_\theta\{|\hat{\theta}_n - \theta| > \delta, I_{\Lambda_n}(\bar{X}_n) = 0\}.$$

If $\theta = 0$ then

$$\begin{aligned} P_\theta\{|\hat{\theta}_n| > \delta\} &= P_\theta\{|\bar{X}_n| > 2\delta, I_{\Lambda_n}(\bar{X}_n) = 1\} + P_\theta\{|\bar{X}_n| > \delta, I_{\Lambda_n}(\bar{X}_n) = 0\} \\ &\le P_\theta\{|\bar{X}_n| > 2\delta\} + P_\theta\{I_{\Lambda_n}(\bar{X}_n) = 0\} \to 0, \quad \text{as } n \to \infty, \end{aligned}$$

since $\bar{X}_n$ is consistent. Similarly, if $\theta \neq 0$,

$$P_\theta\{|\hat{\theta}_n| > \delta\} \le P_\theta\{I_{\Lambda_n}(\bar{X}_n) = 1\} + P_\theta\{|\bar{X}_n - \theta| > \delta\} \to 0, \quad \text{as } n \to \infty.$$

Thus, $\hat{\theta}_n$ is consistent. Furthermore,

$$\begin{aligned} \hat{\theta}_n &= \frac{1}{2}\bar{X}_n I_{\Lambda_n}(\bar{X}_n) + \bar{X}_n(1 - I_{\Lambda_n}(\bar{X}_n)) \\ &= \bar{X}_n - \frac{1}{2}\bar{X}_n I_{\Lambda_n}(\bar{X}_n). \end{aligned}$$

Hence,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \begin{cases} N\left(0, \frac{1}{4}\right), & \text{if } \theta = 0, \\ N(0,1), & \text{if } \theta \neq 0, \end{cases}$$

as $n \to \infty$. This shows that $\hat{\theta}_n$ is asymptotically normal, with asymptotic variance

$$AV_\theta\{\hat{\theta}_n\} = \begin{cases} \frac{1}{4n}, & \text{if } \theta = 0, \\ \frac{1}{n}, & \text{otherwise.} \end{cases}$$

$\hat{\theta}_n$ is **super efficient**. ■

Example 7.6. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables, $X_1 \sim B(1, e^{-\theta})$, $0 < \theta < \infty$. The MLE of θ after n observations is

$$\hat{\theta}_n = -\log \frac{\sum_{i=1}^{n} X_i}{n}.$$

$\hat{\theta}_n$ does not exist if $\sum_{i=1}^{n} X_i = 0$. The probability of this event is $(1 - e^{-\theta})^n$. Thus, if $n \geq N(\delta, \theta) = \dfrac{\log \delta}{\log(1 - e^{-\theta})}$, then $P_\theta \left\{ \sum_{i=1}^{n} X_i = 0 \right\} < \delta$. For $n \geq N(\delta, \theta)$, let

$$B_n = \left\{ \mathbf{X}_n : \sum_{i=1}^{n} X_i \geq 1 \right\},$$

then $P_\theta\{B_n\} > 1 - \delta$. On the set B_n, $-\frac{1}{n} \frac{\partial}{\partial \theta} S(\mathbf{X}_n; \hat{\theta}_n) = \dfrac{\hat{p}_n}{1 - \hat{p}_n}$, where $\hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ is the MLE of $p = e^{-\theta}$. Finally, the Fisher information function is $I(\theta) = e^{-\theta}/(1 - e^{-\theta})$, and

$$-\frac{1}{N} \frac{\partial}{\partial \theta} S(\mathbf{X}_n; \hat{\theta}_n) \xrightarrow[n \to \infty]{\text{a.s.}} I(\theta).$$

All the conditions of Theorem 7.3.1 hold, and

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\left(0, \frac{1 - e^{-\theta}}{e^{-\theta}}\right), \quad \text{as } n \to \infty.$$

■

Example 7.7. Consider again the MLEs of the parameters of a Weibull distribution$G^{1/\beta}(\lambda, 1)$; $0 < \beta, \lambda < \infty$, which have been developed in Example 5.19. The likelihood function $L(\lambda, \beta; \mathbf{X}_n)$ is specified there. We derive here the asymptotic covariance matrix of the MLEs λ and β. Note that the Weibull distributions satisfy all the required regularity conditions.

Let I_{ij}, $i = 1, 2$, $j = 1, 2$ denote the elements of the Fisher information matrix. These elements are defined as

$$I_{11} = E\left\{\left[\frac{\partial}{\partial\lambda}\log L(\lambda, \beta; \mathbf{X})\right]^2\right\},$$

$$I_{12} = E\left\{\left[\frac{\partial}{\partial\lambda}\log L(\lambda, \beta; \mathbf{X})\right]\left[\frac{\partial}{\partial\beta}\log L(\lambda, \beta; \mathbf{X})\right]\right\},$$

$$I_{22} = E\left\{\left[\frac{\partial}{\partial\beta}\log L(\lambda, \beta; \mathbf{X})\right]^2\right\}.$$

We will derive the formulae for these elements under the assumption of $n = 1$ observation. The resulting information matrix can then be multiplied by n to yield that of a random sample of size n. This is due to the fact that the random variables are i.i.d.

The partial derivatives of the log-likelihood are

$$\frac{\partial}{\partial\lambda}\log L(\lambda, \beta; X) = \frac{1}{\lambda} - X^\beta,$$

$$\frac{\partial}{\partial\beta}\log L(\lambda, \beta; X) = \frac{1}{\beta} + \log X - \lambda X^\beta \log X.$$

Thus,

$$I_{11} = E\left\{\left(\frac{1}{\lambda} - X^\beta\right)^2\right\} = \frac{1}{\lambda^2},$$

since $X^\beta \sim E(\lambda)$. It is much more complicated to derive the other elements of $I(\theta)$. For this purpose, we introduce first a few auxiliary results. Let $M(t)$ be the moment generating function of the extreme-value distribution. We note that

$$M'(t) = \int_{-\infty}^{\infty} z e^{(t-1)z - e^{-z}} dz, \quad t < 1,$$

$$M''(t) = \int_{\infty}^{\infty} z^2 e^{(t-1)z - e^{-z}} dz, \quad t < 1.$$

Accordingly,

$$\frac{d}{dt}\Gamma(1+t) = \int_0^\infty (\log x)x^t e^{-x}dx$$
$$= \int_{-\infty}^\infty z e^{-z(1+t)e^{-z}}dz = M'(-t), \quad t > -1,$$

similarly,

$$\frac{d^2}{dt^2}\Gamma(1+t) = M''(-t), \quad t > -1.$$

These identities are used in the following derivations:

$$I_{12} = E\left\{\left(\frac{1}{\beta} + \log X - \lambda X^\beta \log X\right)\left(\frac{1}{\lambda} - X^\beta\right)\right\}$$
$$= \frac{1}{\lambda\beta}[\Gamma'(3) - 2\Gamma'(2) + \gamma - \log\lambda],$$

where $\gamma = 0.577216\ldots$ is the Euler constant. Moreover, as compiled from the tables of Abramowitz and Stegun (1968, p. 253)

$$\Gamma'(2) = 0.42278\ldots \text{ and } \Gamma'(3) = 1.84557\ldots$$

We also obtain

$$I_{22} = \frac{1}{\beta^2}\Bigg[1 + \frac{\pi^2}{6} + (\gamma - \log\lambda)^2 + \Gamma''(3) - 2\Gamma''(2)$$
$$- 2\log\lambda(\Gamma'(3) - 2\Gamma'(2) - 1) + \frac{2}{\lambda}(\gamma - \log\lambda)\Bigg],$$

where $\Gamma''(2) = 0.82367$ and $\Gamma''(3) = 2.49293$. The derivations of formulae for I_{12} and I_{22} are lengthy and tedious. We provide here, for example, the derivation of one expectation:

$$\lambda E\{X^\beta(\log X)^2\} = \frac{2\lambda}{\beta^2}E\{X^\beta(\log X^\beta)^2\}.$$

However, $X^\beta \sim E(\lambda) \sim \frac{1}{\lambda}U$, where $U \sim E(1)$. Therefore,

$$\begin{aligned}\lambda E\{X^\beta(\log X)^2\} &= 2E\left\{U\left(\log\frac{U}{\lambda}\right)^2\right\}/\beta^2 \\ &= \frac{2}{\beta^2}\left\{\int_{-\infty}^{\infty} z^2 e^{-2z-e^{-z}}dz - 2\log\lambda\int_{-\infty}^{\infty} z e^{-2z-e^{-z}}dz + (\log\lambda)^2\right\} \\ &= \frac{2}{\beta^2}[\Gamma''(2) - 2(\log\lambda)\Gamma'(2) + (\log\lambda)^2].\end{aligned}$$

The reader can derive other expressions similarly.

For each value of λ and β, we evaluate I_{11}, I_{12}, and I_{22}. The asymptotic variances and covariances of the MLEs, designated by AV and AC, are determined from the inverse of the Fisher information matrix by

$$AV\{\hat{\lambda}\} = \frac{I_{22}}{n[I_{11}I_{22} - I_{12}^2]},$$

$$AV\{\hat{\beta}\} = \frac{I_{11}}{n[I_{11}I_{22} - I_{12}^2]},$$

and

$$AC(\hat{\lambda}, \hat{\beta}) = \frac{-I_{12}}{n[I_{11}I_{22} - I_{12}^2]}.$$

Applying these formulae to determine the asymptotic variances and asymptotic covariance of $\hat{\lambda}$ and $\hat{\beta}$ of Example 5.20, we obtain, for $\lambda = 1$ and $\beta = 1.75$, the numerical results $I_{11} = 1$, $I_{12} = 0.901272$, and $I_{22} = 1.625513$. Thus, for $n = 50$, we have $AV\{\hat{\lambda}\} = 0.0246217$, $AV\{\hat{\beta}\} = 0.0275935$ and $AC(\hat{\lambda}, \hat{\beta}) = -0.0221655$. The asymptotic standard errors (square roots of AV) of $\hat{\lambda}$ and $\hat{\beta}$ are, 0.1569 and 0.1568, respectively. Thus, the estimates $\hat{\lambda} = 0.839$ and $\hat{\beta} = 1.875$ are not significantly different from the true values $\lambda = 1$ and $\beta = 1.75$. ■

Example 7.8. Let $X_1, \ldots, X_n$ be i.i.d. random variables with $X_1 \sim E\left(\frac{1}{\xi}\right)$, $0 < \xi < \infty$. Let $Y_1, \ldots, Y_n$ be i.i.d. random variables, $Y_1 \sim G\left(\frac{1}{\eta}, 1\right)$, $0 < \eta < \infty$, and assume that the Y-sample is independent of the X-sample.

The parameter to estimate is $\theta = \frac{\xi}{\eta}$. The MLE of θ is $\hat{\theta}_n = \frac{\bar{X}_n}{\bar{Y}_n}$, where $\bar{X}_n$ and $\bar{Y}_n$ are the corresponding sample means. For each $n \geq 1$, $\hat{\theta}_n \sim \theta F[2n, 2n]$. The

asymptotic distribution of $\hat{\theta}_n$ is $N\left(\theta, \frac{2\theta^2}{n}\right)$, $0 < \theta < \infty$. $\hat{\theta}_n$ is a BAN estimator. To find the asymptotic bias of $\hat{\theta}_n$, verify that

$$\begin{aligned} E\{\hat{\theta}_n\} &= \theta \frac{2n}{2n-2} \\ &= \theta\left(1 + \frac{1}{n-1}\right). \end{aligned}$$

The bias of the MLE is $B(\hat{\theta}_n) = \frac{\theta}{n-1}$, which is of $O\left(\frac{1}{n}\right)$. Thus, we adjust $\hat{\theta}_n$ by $\hat{\theta}_n^* = \left(1 - \frac{1}{n}\right)\hat{\theta}_n$. The bias of $\hat{\theta}_n^*$ is $B(\hat{\theta}_n^*) = 0$. The variance of $\hat{\theta}_n^*$ is

$$\begin{aligned} V\{\hat{\theta}_n^*\} &= \frac{2\theta^2}{n} \frac{\left(1 - \frac{1}{2n}\right)}{\left(1 - \frac{2}{n}\right)} \\ &= \frac{2\theta^2}{n}\left(1 + \frac{3}{2n} + \frac{3}{n^2} + o\left(\frac{1}{n^2}\right)\right). \end{aligned}$$

Thus, the second-order deficiency coefficients of $\hat{\theta}_n^*$ is $D = 3\theta^2$. Note that $\hat{\theta}_n^*$ is the UMVU estimator of θ. ■

Example 7.9. Let $X_1, X_2, \ldots, X_n$ be i.i.d. Poisson random variables, with mean λ, $0 < \lambda < \infty$. We consider $\theta = e^{-\lambda}$, $0 < \theta < 1$.

The UMVU of θ is

$$\tilde{\theta}_n = \left(1 - \frac{1}{n}\right)^{T_n}, \quad n \geq 1,$$

where $T_n = \sum_{i=1}^{n} X_i$. The MLE of θ is

$$\hat{\theta}_n = \exp\{-\bar{X}_n\}, \quad n \geq 1,$$

where $\bar{X}_n = T_n/n$. Note that $\tilde{\theta}_n - \hat{\theta}_n \xrightarrow{\text{a.s.}} 0$. The two estimators are asymptotically equivalent. Using moment generating functions, we prove that

$$\begin{aligned} E\{\hat{\theta}_n\} &= \exp\{-n\lambda(1 - e^{-1/n})\} \\ &= e^{-\lambda} + \frac{\lambda}{2n} e^{-\lambda} + O\left(\frac{1}{n^2}\right). \end{aligned}$$

Thus, adjusting the MLE for the bias, let

$$\hat{\theta}_n^* = e^{-\bar{X}_n} - \frac{\bar{X}_n}{2n} \exp\{-\bar{X}_n\}.$$

Note that, by the delta method,

$$E\{\bar{X}_n \exp\{-\bar{X}_n\}\} = \frac{\lambda e^{-\lambda}}{2n} + \frac{\lambda^2}{4n^2} e^{-2\lambda} + o\left(\frac{1}{n^2}\right).$$

Thus,

$$E\{\hat{\theta}_n^*\} = e^{-\lambda} + O\left(\frac{1}{n^2}\right), \quad \text{as } n \to \infty.$$

The variance of the bias adjusted estimator $\hat{\theta}_n^*$ is

$$V\{\hat{\theta}_n^*\} = V\{e^{-\bar{X}_n}\} - \frac{1}{n}\text{cov}(e^{-\bar{X}_n}, \bar{X}_n e^{-\bar{X}_n}) + O\left(\frac{1}{n^3}\right), \quad \text{as } n \to \infty.$$

Continuing the computations, we find

$$V\{e^{-\bar{X}_n}\} = e^{-2\lambda}\left(\frac{\lambda}{n} + \frac{3\lambda^2 - 2\lambda}{2n^2} + o\left(\frac{1}{n^2}\right)\right).$$

Similarly,

$$\frac{1}{n}\text{cov}(e^{-\bar{X}_n}, \bar{X}_n e^{-\bar{X}_n}) = \frac{\lambda e^{-2\lambda}(3\lambda - 2)}{2n^2} + o\left(\frac{1}{n^2}\right).$$

It follows that

$$V\{\hat{\theta}_n^*\} = \frac{\lambda e^{-2\lambda}}{n} + o\left(\frac{1}{n^2}\right).$$

In the present example, the bias adjusted MLE is most efficient second-order estimator. The variance of the UMVU $\tilde{\theta}_n$ is

$$V\{\tilde{\theta}_n\} = \frac{\lambda e^{-2\lambda}}{n} + \frac{\lambda^2 e^{-2\lambda}}{2n^2} + o\left(\frac{1}{n^2}\right).$$

Thus, $\tilde{\theta}_n$ has the deficiency coefficient $D = \lambda^2 e^{-2\lambda}/2$. ■

Example 7.10.

(a) In $n = 100$ Bernoulli trials, we observe 56 successes. The model is $X \sim B(100, \theta)$. The MLE of θ is $\hat{\theta} = 0.56$ and that of $g(\theta) = 2\sin^{-1}(\sqrt{\theta})$ is $g(0.56) = 1.69109$. The 0.95-confidence limits for $g(\theta)$ are $g_L = g(0.56) - z_{.975}/10 = 1.49509$ and $g_U = 1.88709$. The function $g(\theta)$ varies in the range $[0, \pi]$. Thus, let $g_L^* = \max(0, g_L)$ and $g_U^* = \min(g_U, \pi)$. In the present case, both g_L and g_U are in $(0, \pi)$. The inverse transformation is

$$\theta_L = \sin^2(g_L/2),$$
$$\theta_U = \sin^2(g_U/2).$$

In the present case, $\theta_L = 0.462$ and $\theta_U = 0.656$. We can also, as mentioned earlier, determine the approximate confidence limits directly on θ by estimating the variance of θ. In this case, we obtain the limits

$$\theta_L = \hat{\theta} - \frac{z_{1-\alpha/2}}{\sqrt{100}}\sqrt{\hat{\theta}(1-\hat{\theta})} = 0.463$$
$$\theta_U = \hat{\theta} + \frac{z_{1-\alpha/2}}{\sqrt{100}}\sqrt{\hat{\theta}(1-\hat{\theta})} = 0.657.$$

Both approaches yield here close results, since the sample is sufficiently large.

(b) Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. vectors having the bivariate normal distribution, with expectation vector (ξ, η) and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix}; \quad -\infty < \xi, \quad \eta < \infty, \quad 0 < \sigma, \quad \tau < \infty, \quad -1 < \rho < 1.$$

The MLE of ρ is the sample coefficient of correlation $r = \Sigma(X_i - \bar{X})(Y_i - \bar{Y})/[\Sigma(X_i - \bar{X})^2 \cdot \Sigma(Y_i - \bar{Y})^2]^{1/2}$. By determining the inverse of the Fisher information matrix one obtains that the asymptotic variance of r is $AV\{r\} = \frac{1}{n}(1-\rho^2)^2$. Thus, if we make the transformation $g(\rho) = \frac{1}{2}\log\frac{1+\rho}{1-\rho}$ then $g'(\rho) = \frac{1}{1-\rho^2}$. Thus, $g(r) = \frac{1}{2}\log((1+r)/(1-r))$ is a variance stabilizing transformation for r, with an asymptotic variance of $1/n$. Suppose that in a sample of $n = 100$ we find a coefficient of correlation $r = 0.79$. Make the transformation

$$g(0.79) = \frac{1}{2}\log\frac{1.79}{0.21} = 1.0714.$$

We obtain on the basis of this transformation the asymptotic limits

$$g_L = 1.0714 - 0.196 = 0.8754$$
$$g_U = 1.0714 + 0.196 = 1.2674.$$

The inverse transformation is $\rho = (e^{2g} - 1)/(e^{2g} + 1)$. Thus, the confidence interval of ρ has the limits $\rho_L = 0.704$ and $\rho_U = 0.853$. On the other hand, if we use the formula

$$r \pm \frac{z_{1-\alpha/2}}{\sqrt{n}}(1 - r^2).$$

We obtain the limits $\rho_L = 0.716$ and $\rho_U = 0.864$. The two methods yield confidence intervals which are close, but not the same. A sample of size 100 is not large enough. ■

Example 7.11. In Example 6.6, we determined the confidence limits for the cross-ratio productρ. We develop here the large sample approximation, according to the two approaches discussed above. Let

$$\omega = \log \rho = \log \frac{\theta_{11}}{1 - \theta_{11}} - \log \frac{\theta_{12}}{1 - \theta_{12}} - \log \frac{\theta_{21}}{1 - \theta_{21}} + \log \frac{\theta_{22}}{1 - \theta_{22}}.$$

Let $\hat{\theta}_{ij} = X_{ij}/n_{ij}$ $(i, j = 1, 2)$. $\hat{\theta}_{ij}$ is the MLE of θ_{ij}. Let $\psi_{ij} = \log(\theta_{ij}/(1 - \theta_{ij}))$. The MLE of ψ_{ij} is $\hat{\psi}_{ij} = \log(\hat{\theta}_{ij}/(1 - \hat{\theta}_{ij}))$. The asymptotic distribution of $\hat{\psi}_{ij}$ is normal with mean ψ_{ij} and

$$AV\left\{\log \frac{\hat{\theta}_{ij}}{1 - \hat{\theta}_{ij}}\right\} = [n_{ij}\theta_{ij}(1 - \theta_{ij})]^{-1}.$$

Furthermore, the MLE of ω is

$$\hat{\omega} = \hat{\psi}_{11} - \hat{\psi}_{12} - \hat{\psi}_{21} + \hat{\psi}_{22}.$$

Since X_{ij}, $(i, j) = 1, 2$, are mutually independent so are the terms on the RHS of $\hat{\omega}$. Accordingly, the asymptotic distribution of $\hat{\omega}$ is normal with expectation ω and asymptotic variance

$$AV\{\hat{\omega}\} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{1}{n_{ij}\theta_{ij}(1 - \theta_{ij})}.$$

Since the values θ_{ij} are unknown we substitute their MLEs. We thus define the standard error of $\hat{\omega}$ as

$$SE\{\hat{\omega}\} = \left\{ \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{1}{n_{ij}\hat{\theta}_{ij}(1-\hat{\theta}_{ij})} \right\}^{1/2}.$$

According to the asymptotic normal distribution of $\hat{\omega}$, the asymptotic confidence limits for ρ are

$$\hat{\rho}^{(1)} = \hat{\rho} \exp\{-z_{1-\alpha/2} \cdot SE\{\hat{\omega}\}\},$$
$$\hat{\rho}^{(2)} = \hat{\rho} \exp\{z_{1-\alpha/2} SE\{\hat{\omega}\}\},$$

where $\hat{\rho}$ is the MLE of $\rho = e^{\omega}$. These limits can be easily computed. For a numerical example, consider the following table (Fleiss, 1973, p. 126) in which we present the proportions of patients diagnosed as schizophrenic in two studies both performed in New York and London.

	New York		London	
Study	n	$\hat{\theta}$	n	$\hat{\theta}$
1	105	0.771	105	0.324
2	192	0.615	174	0.394

These samples yield the MLE $\hat{\rho} = 2.9$. The asymptotic confidence limits at level $1 - \alpha = 0.95$ are $\hat{\rho}^{(1)} = 1.38$ and $\hat{\rho}^{(2)} = 6.08$. This result indicates that the interaction parameter ρ is significantly greater than 1. We show now the other approach, using the variance stabilizing transformation $2\sin^{-1}(\sqrt{\theta})$. Let $\hat{\hat{\theta}}_{ij} = (X_{ij} + 0.5)/(n_{ij} + 1)$ and $Y_{ij} = 2\sin^{-1}(\hat{\hat{\theta}}_{ij})$. On the basis of these variables, we set the $1 - \alpha$ confidence limits for $\eta_{ij} = 2\sin^{-1}(\sqrt{\theta_{ij}})$. These are

$$\eta_{ij}^{(1)} = Y_{ij} - z_{1-\alpha/2}/\sqrt{n_{ij}} \quad \text{and} \quad \eta_{ij}^{(2)} = Y_{ij} + z_{1-\alpha/2}/\sqrt{n_{ij}}.$$

For these limits, we directly obtain the asymptotic confidence limits for ψ_{ij} that are

$$\psi_{ij}^{(k)} = 2\log\tan(\hat{\eta}_{ij}^{(k)}/2), \quad k = 1, 2,$$

where

$$\hat{\eta}_{ij}^{(1)} = \max(0, \eta_{ij}^{(1)}), \quad i, j = 1, 2,$$

and

$$\hat{\eta}_{ij}^{(2)} = \min(\pi, \eta_{ij}^{(2)}), \quad i, j = 1, 2.$$

We show now how to construct asymptotic confidence limits for ρ from these asymptotic confidence limits for ψ_{ij}.

Define

$$\hat{\hat{\omega}} = \frac{1}{2} \sum_{k=1}^{2} (\psi_{11}^{(k)} - \psi_{12}^{(k)} - \psi_{21}^{(k)} + \psi_{22}^{(k)}),$$

and

$$D = \frac{1}{4} \sum_{i=1}^{2} \sum_{j=1}^{2} (\psi_{ij}^{(2)} - \psi_{ij}^{(1)})^2.$$

D is approximately equal to $z_{1-\alpha/2}^2 AV\{\hat{\hat{\omega}}\}$. Indeed, from the asymptotic theory of MLEs, $D_{ij} = (\psi_{ij}^{(2)} - \psi_{ij}^{(1)})/4$ is approximately the asymptotic variance of ψ_{ij} times $z_{1-\alpha/2}^2$. Accordingly,

$$\sqrt{D} = z_{1-\alpha/2} SE\{\hat{\hat{\omega}}\},$$

and by employing the normal approximation, the asymptotic confidence limits for ρ are

$$\rho^{(k)} = \exp\{\hat{\hat{\omega}} + (-1)^k \sqrt{D}\}, \quad k = 1, 2.$$

Thus, we obtain the approximate confidence limits for Fleiss' example, $\rho^{(1)} = 1.40$ and $\rho^{(2)} = 6.25$. These limits are close to the ones obtained by the other approach. For further details, see Zacks and Solomon (1976). ■

Example 7.12. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having the gamma distribution $G(1, \nu)$, $0 < \nu < \infty$. This is a one-parameter exponential type family, with canonical p.d.f.

$$f(x; \nu) = \frac{1}{x} \exp\{\nu \log x - \log \Gamma(\nu)\}.$$

Here, $K(\nu) = \log \Gamma(\nu)$.

The MLE of ν is the root of the equation

$$\begin{aligned} K'(\nu) &= \frac{\Gamma'(\nu)}{\Gamma(\nu)} \\ &= \bar{U}_n, \end{aligned}$$

where $\bar{U}_n = \frac{1}{n}\sum_{i=1}^{n} \log(X_i)$. The function $\Gamma'(\nu)/\Gamma(\nu)$ is known as the di-gamma, or psi function, $\psi(\nu)$ (see Abramowitz and Stegun, 1968, p. 259). $\psi(\nu)$ is tabulated for $1 \le \nu \le 2$ in increments of $\Delta = 0.05$. For ν values smaller than 1 or greater than 2 use the recursive equation

$$\psi(1+\nu) = \psi(\nu) + \frac{1}{\nu}.$$

The values of $\hat{\nu}_n$ can be determined by numerical interpolation.

The function $\psi(\nu)$ is analytic on the complex plane, excluding the points $\nu = 0, -1, -2, \ldots$. The nth order derivative of $\psi(\nu)$ is

$$\begin{aligned} \psi^{(n)}(\nu) &= (-1)^n \int_0^\infty \frac{t^n e^{-\nu t}}{1 - e^{-t}} dt \\ &= (-1)^n n! \sum_{j=0}^{\infty} \frac{1}{(j+\nu)^{n+1}}. \end{aligned}$$

Accordingly,

$$\begin{aligned} K'(\nu) &= \psi(\nu), \\ I(\nu) &= \psi'(\nu), \\ \beta_1 &= \frac{\psi^{(2)}(\nu)}{(\psi'(\nu))^{3/2}}, \\ \beta_2 &= \frac{\psi^{(3)}(\nu)}{(\psi'(\nu))^2}. \end{aligned}$$

To assess the normal and the Edgeworth approximations to the distribution of $\hat{\nu}_n$, we have simulated 1000 independent random samples of size $n = 20$ from the gamma distribution with $\nu = 1$. In this case $I(1) = 1.64493$, $\beta_1 = -1.1395$ and $\beta_2 - 3 = 2.4$. In Table 7.1, we present some empirical quantiles of the simulations. We see that the Edgeworth approximation is better than the normal for all standardized values of $\hat{\nu}_n$ between the 0.2th and 0.8th quantiles. In the tails of the distribution, one could get better results by the saddlepoint approximation. ■

Table 7.1 Normal and Edgeworth Approximations to the Distribution of $\hat{\nu}_{20}$, $n = 20$, $\nu = 1$

$\hat{Z}$	Exact	Normal	Edgeworth
−1.698	0.01	0.045	0.054
−1.418	0.05	0.078	0.083
−1.148	0.10	0.126	0.127
−0.687	0.20	0.246	0.238
−0.433	0.30	0.333	0.320
−0.208	0.40	0.417	0.401
−0.012	0.50	0.495	0.478
0.306	0.60	0.620	0.606
0.555	0.70	0.711	0.701
0.887	0.80	0.812	0.811
1.395	0.90	0.919	0.926
1.932	0.95	0.973	0.981
2.855	0.99	0.999	0.999

Example 7.13. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having the exponential distribution $X_1 \sim E(\psi)$, $0 < \psi < \infty$. This is a one-parameter exponential family with canonical p.d.f.

$$f(x;\psi) = \exp\{\psi U(x) - K(\psi)\}, \quad 0 < x < \infty,$$

where $U(x) = -x$ and $K(\psi) = -\log(\psi)$.

The MLE of ψ is $\hat{\psi}_n = 1/\bar{X}_n$. The p.d.f. of $\hat{\psi}_n$ is obtained from the density of $\bar{X}_n$ and is

$$g_{\hat{\psi}_n}(x;\psi) = \frac{(n\psi)^n}{\Gamma(n)} \cdot \frac{1}{x^{n+1}} e^{-n\psi/x},$$

for $0 < x < \infty$.

The approximation to the p.d.f. according to (7.6.9) yields

$$\begin{aligned} g_{\hat{\psi}_n}(x;\psi) &= c_n \exp\{-(x-\psi)T_n - n(-\log\psi + \log x)\} \\ &= c_n \frac{\psi^n}{x^{n+1}} \exp\{n(x-\psi)/x\} \\ &= c_n \frac{\psi^n e^n}{x^{n+1}} e^{-n\psi/x}. \end{aligned}$$

Substituting $c_n = n^n e^{-n}/\Gamma(n)$ we get the exact equation. ■

Example 7.14. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables, having a common normal distribution $N(\theta, 1)$. Consider the problem of testing the hypothesis $H_0 : \theta \leq 0$

against $H_1 : \theta > 0$. We have seen that the uniformly most powerful (UMP) test of size α is

$$\phi_n^0 = I\left\{\bar{X}_n \geq \frac{Z_{1-\alpha}}{\sqrt{n}}\right\},$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and and $Z_{1-\alpha} = \Phi^{-1}(1-\alpha)$. The power function of this UMP test is

$$\psi_n(\theta) = \Phi(\sqrt{n}\,\theta - Z_{1-\alpha}), \quad \theta \geq 0.$$

Let $\theta_1 > 0$ be specified. The number of observations required so that $\psi_n(\theta_1) \geq \gamma$ is

$$N(\alpha, \theta_1, \gamma) = \textbf{ least integer } n \textbf{ greater than } (Z_\gamma + Z_{1-\alpha})^2/\theta_1^2.$$

Note that

(i) $\lim_{n\to\infty} \psi_n(\theta_1) = 1$ for each $\theta_1 > 0$
and

(ii) if $\delta > 0$ and $\theta_1 = \frac{\delta}{\sqrt{n}}$ then

$$\lim_{n\to\infty} \psi_n\left(\frac{\delta}{\sqrt{n}}\right) = \Phi(\delta - Z_{1-\alpha}) \equiv \psi_\infty,$$

where $0 < \alpha < \psi_\infty < 1$.

Suppose that one wishes to consider a more general model, in which the p.d.f. of X_1 is $f(x-\theta)$, $-\infty < \theta < \infty$, where $f(x)$ is symmetric about θ but not necessarily equal to $\phi(x)$, and $V_\theta\{X\} = \sigma^2$ for all $-\infty < \theta < \infty$. We consider the hypotheses $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$.

Due to the CLT, one can consider the sequence of test statistics

$$\phi_n^{(1)}(\mathbf{X}_n) = I\left\{\bar{X}_n \geq \frac{a_n\sigma}{\sqrt{n}}\right\},$$

where $a_n \downarrow Z_{1-\alpha}$ as $n \to \infty$, and the alternative sequence

$$\phi_n^{(2)}(\mathbf{X}_n) = I\left\{M_e \geq \frac{a_n}{2f(0)\sqrt{n}}\right\},$$

where M_e is the sample median

$$M_e = \begin{cases} X_{(m+1)}, & \text{if } n = 2m+1, \\ \frac{1}{2}(X_{(m)} + X_{(m+1)}), & \text{if } n = 2m. \end{cases}$$

According to Theorem 1.13.7, $\sqrt{n}(M_e - \theta) \xrightarrow{d} N\left(0, \frac{1}{4f^2(\theta)}\right)$ as $n \to \infty$. Thus, the asymptotic power functions of these tests are

$$\psi_n^{(1)}(\theta) \approx \Phi\left(\frac{\theta}{\sigma}\sqrt{n} - Z_{1-\alpha}\right), \quad \theta > 0,$$

and

$$\psi_n^{(2)}(\theta) \approx \Phi\left(2\theta f(0)\sqrt{n} - Z_{1-\alpha}\right), \quad \theta > 0.$$

Both $\psi_n^{(1)}(\theta_1)$ and $\psi_n^{(2)}(\theta_1)$ converge to 1 as $n \to \infty$, for any $\theta_1 > 0$, which shows their consistency. We wish, however, to compare the behavior of the sequences of power functions for $\theta_n = \frac{\delta}{\sqrt{n}}$. Note that each hypothesis, with $\theta_{1,n} = \frac{\delta}{\sqrt{n}}$ is an alternative one. But, since $\theta_{1,n} \to 0$ as $n \to \infty$, these alternative hypotheses are called **local hypotheses**. Here we get

$$\psi_n^{(1)}\left(\frac{\delta}{\sqrt{n}}\right) \approx \Phi\left(\frac{\delta}{\sigma} - Z_{1-\alpha}\right) = \psi^*$$

and

$$\psi_n^{(2)}\left(\frac{\delta}{\sqrt{n}}\right) \approx \Phi\left(2f(0)\delta - Z_{1-\alpha}\right) = \psi^{**}.$$

To insure that $\psi^* = \psi^{**}$ one has to consider for $\psi^{(2)}$ a sequence of alternatives $\frac{\delta}{\sqrt{n}}$ with sample size $n' = \frac{n}{4f^2(0)}$ so that

$$\psi_{n'}\left(\frac{\delta}{\sqrt{n}}\right) \approx \Phi\left(2\frac{\delta}{\sqrt{n}}f(0)\sqrt{n'} - Z_{1-\alpha}\right) = \psi^*.$$

The Pitman ARE of $\phi_n^{(2)}$ to $\phi_n^{(1)}$ is defined as the limit of $n/n'(n)$ as $n \to \infty$. In the present example,

$$\text{ARE}(\phi^{(2)}, \phi^{(1)}) = 4f^2(0).$$

If the original model of $X \sim N(\theta, 1)$, $f(0) = \dfrac{1}{\sqrt{2\pi}}$ and ARE of $\psi^{(2)}$ to $\psi^{(1)}$ is 0.637. On the other hand, if $f(x) = \dfrac{1}{2}e^{-|x|}$, which is the Laplace distribution, then the ARE of $\psi^{(2)}$ to $\psi^{(1)}$ is 1. ■

Example 7.15. In Example 7.3, we discussed the Wilcoxon signed-rank test of $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, when the distribution function F is absolutely continuous and symmetric around θ. We show here the Pitman's asymptotic efficiency of this test relative to the t-test. The t-test is valid only in cases where $\sigma_f^2 = V\{X\}$ and $0 < \sigma_f^2 < \infty$. The t-statistic is $t_n = \sqrt{n}\dfrac{(\bar{X}_n - \theta_0)}{S_n}$, where S_n^2 is the sample variance. Since $S_n \xrightarrow{\text{a.s.}} \sigma_f$, as $n \to \infty$, we consider

$$\phi_n(t_n) = I\left\{\frac{\sqrt{n}(\bar{X}_n - \theta_0)}{S_n} \geq t_{1-\alpha}[n-1]\right\}.$$

The asymptotic efficacy of the t-test is

$$J(\theta; t_n) = \frac{1}{\sigma_f^2},$$

where σ_f^2 is the variance of X, under the p.d.f. $f(x)$. Indeed, $\mu(\theta) = \theta$.

Consider the Wilcoxon signed-rank statistic T_n, given by (7.1.3). The test function, for large n, is given by (7.1.8). For this test

$$\mu(\theta) = P_\theta\{Y_1 > 0\} + \frac{(n-1)}{2}p_2(\theta),$$

where $p_2(\theta)$ is given in Example 7.3. Thus,

$$\mu(\theta) = (1 - F(-\theta)) + \frac{(n-1)}{2}\int_{-\infty}^{\infty}(1 - F(-x-\theta))f(x-\theta)dx.$$

Hence,

$$\mu'(0) = f(0) + \frac{(n-1)}{2}\int_{-\infty}^{\infty} f^2(x)dx - \frac{1}{2}.$$

Using $\sigma^2(0) = \dfrac{(n+1)(2n+1)}{24}$, we obtain the asymptotic efficacy of

$$J(\theta; T_n) = 12\left(\int_{-\infty}^{\infty} f^2(x)dx\right)^2 + O\left(\frac{1}{n}\right),$$

as $n \to \infty$. Thus, the Pitman ARE of T_n versus t_n is

$$\text{ARE}(T_n, t_n) = 12\sigma_f^2 \left(\int_{-\infty}^{\infty} f^2(x)dx \right)^2. \tag{7.9.16}$$

Thus, if $f(x) = \phi(x)$ (standard normal) $\text{ARE}(T_n, t_n) = 0.9549$. On the other hand, if $f(x) = \frac{1}{2}\exp\{-|x|\}$ (standard Laplace) then $\text{ARE}(T_n, t_n) = 1.5$. These results deem the Wilcoxon signed-rank test to be asymptotically very efficient nonparametric test. ■

Example 7.16. Let $X_1, \ldots, X_n$ be i.i.d. random variables, having a common Cauchy distribution, with a location parameter θ, $-\infty < \theta < \infty$, i.e.,

$$f(x;\theta) = \frac{1}{\pi(1 + (x-\theta)^2)}, \quad -\infty < x < \infty.$$

We derive a confidence interval for θ, for large n (asymptotic). Let $\hat{\theta}_n$ be the sample median, i.e.,

$$\hat{\theta}_n = F_n^{-1}\left(\frac{1}{2}\right).$$

Note that, due to the symmetry of $f(x;\theta)$ around θ, $\theta = F^{-1}\left(\frac{1}{2}\right)$. Moreover,

$$f(\theta;\theta) = \frac{1}{\pi}.$$

Hence, the $(1-\alpha)$ confidence limits for θ are

$$\hat{\theta}_n \pm Z_{1-\alpha/2}\frac{\pi}{2\sqrt{n}}.$$

■

PART III: PROBLEMS

Section 7.1

7.1.1 Let $X_i = \alpha + \beta z_i + \epsilon_i$, $i = 1, \ldots, n$, be a simple linear regression model, where $z_1, \ldots, z_n$ are prescribed constants, and $\epsilon_1, \ldots, \epsilon_n$ are independent random variables with $E\{e_i\} = 0$ and $V\{\epsilon_i\} = \sigma^2$, $0 < \sigma^2 < \infty$, for all $i = 1, \ldots, n$. Let $\hat{\alpha}_n$ and $\hat{\beta}_n$ be the LSE of α and β.

(i) Show that if $\sum_{i=1}^{n}(z_i - \bar{z}_n)^2 \to \infty$, as $n \to \infty$, then $\hat{\beta}_n \xrightarrow{p} \beta$, i.e., $\hat{\beta}_n$ is consistent.

(ii) What is a sufficient condition for the consistency of $\hat{\alpha}_n$?

7.1.2 Suppose that $X_1, X_2, \ldots, X_n, \ldots$ are i.i.d. random variables and $0 < E\{X_1^4\} < \infty$. Give a strongly consistent estimator of the kurtosis coefficient $\beta_2 = \frac{\mu_4^*}{(\mu_2^*)^2}$.

7.1.3 Let $X_1, \ldots, X_k$ be independent random variables having binomial distributions $B(n, \theta_i)$, $i = 1, \ldots, k$. Consider the null hypothesis $H_0 : \theta_1 = \cdots = \theta_k$ against the alternative $H_1 : \sum_{i=1}^{k}(\theta_i - \bar{\theta})^2 > 0$, where $\bar{\theta} = \frac{1}{k}\sum_{i=1}^{k}\theta_i$. Let $p_i = X_i/n$ and $\bar{p} = \frac{1}{k}\sum_{i=1}^{k}p_i$. Show that the test function

$$\phi(\mathbf{X}) = \begin{cases} 1, & \text{if } \frac{n}{\bar{p}(1-\bar{p})}\sum_{i=1}^{k}(p_i - \bar{p})^2 > \chi^2_{1-\alpha}[k-1], \\ 0, & \text{otherwise,} \end{cases}$$

has a size α_n converging to α as $n \to \infty$. Show that this test is consistent.

7.1.4 In continuation of Problem 3, define $Y_i = 2\sin^{-1}\sqrt{p_i}$, $i = 1, \ldots, k$.

(i) Show that the asymptotic distribution of Y_i, as $n \to \infty$, is $N(\eta_i, \frac{1}{n})$, where $\eta_i = 2\sin^{-1}\sqrt{\theta_i}$.

(ii) Show that $Q = n\sum_{i=1}^{k}(Y_i - \bar{Y})^2$, where $\bar{Y} = \frac{1}{k}\sum_{i=1}^{k}Y_i$, is distributed asymptotically (as $n \to \infty$) like $\chi^2[k-1; \lambda(\boldsymbol{\theta})]$, where $\lambda(\boldsymbol{\theta}) = \frac{n}{2}\sum_{i=1}^{k}(\eta_i - \bar{\eta})^2$; $\bar{\eta} = \frac{1}{k}\sum_{i=1}^{k}\eta_i$. Thus, prove the consistency of the test.

(iii) Derive the formula for computing the asymptotic power of the test $\phi(\mathbf{X}) = I\{Q \geq \chi^2_{1-\alpha}[k-1]\}$.

(iv) Assuming that $\sum_{i=1}^{k}(\eta_i - \bar{\eta})^2$ is independent of n, how large should n be so that the probability of rejecting H_0 when $\sum_{i=1}^{k}(\eta_i - \bar{\eta})^2 \geq 10^{-1}$ will not be smaller than 0.9?

Section 7.2

7.2.1 Let $X_1, X_2, \ldots, X_n, \ldots$ be i.i.d. random variables, $X_1 \sim G(1, \nu)$, $0 < \nu \leq \nu^* < \infty$. Show that all conditions of Theorem 7.2.1 are satisfied, and hence the MLE, $\hat{\nu}_n \xrightarrow{\text{a.s.}} \nu$ as $n \to \infty$ (strongly consistent).

7.2.2 Let $X_1, X_2, \ldots, X_n, \ldots$ be i.i.d. random variables, $X_1 \sim \beta(\nu, 1), 0 < \nu < \infty$. Show that the MLE, $\hat{\nu}_n$, is strongly consistent.

7.2.3 Consider the Hardy–Weinberg genetic model, in which $(J_1, J_2) \sim MN(n, (p_1(\theta), p_2(\theta)))$, where $p_1(\theta) = \theta^2$ and $p_2(\theta) = 2\theta(1-\theta)$, $0 < \theta < 1$. Show that the MLE of θ, $\hat{\theta}_n$, is strongly consistent.

7.2.4 Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables from $G(\lambda, 1)$, $0 < \lambda < \infty$. Show that the following estimators $\hat{\omega}(\bar{X}_n)$ are consistent estimators of $\omega(\lambda)$:

(i) $\hat{\omega}(\bar{X}_n) = -\log \bar{X}_n$, $\omega(\lambda) = \log \lambda$;

(ii) $\hat{\omega}(\bar{X}_n) = \bar{X}_n^2$, $\omega(\lambda) = 1/\lambda^2$;

(iii) $\hat{\omega}(\bar{X}_n) = \exp\{-1/\bar{X}_n\}$, $\omega(\lambda) = \exp\{-\lambda\}$.

7.2.5 Let $X_1, \ldots, X_n$ be i.i.d. from $N(\mu, \sigma^2)$, $-\infty < \mu < \infty$, $0 < \sigma < \infty$. Show that

(i) $\log(1 + \bar{X}_n^2)$ is a consistent estimator of $\log(1 + \mu^2)$;

(ii) $\phi(\bar{X}_n/S)$ is a consistent estimator of $\phi(\mu/\sigma)$, where S^2 is the sample variance.

Section 7.3

7.3.1 Let (X_i, Y_i), $i = 1, \ldots, n$ be i.i.d. random vectors, where

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{bmatrix} \xi \\ \eta \end{bmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right),$$

$-\infty < \xi < \infty, 0 < \eta < \infty, 0 < \sigma_1, \sigma_2 < \infty, -1 < \rho < 1$. Find the asymptotic distribution of $W_n = \bar{X}_n/\bar{Y}_n$, where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

7.3.2 Let $X_1, X_2, \ldots, X_n, \ldots$ be i.i.d. random variables having a Cauchy distribution with location parameter θ, i.e.,

$$f(x; \theta) = \frac{1}{\pi} \cdot \frac{1}{1 + (x - \theta)^2}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty.$$

Let M_e be the sample median, or $M_e = F_n^{-1}\left(\frac{1}{2}\right)$. Is M_e a BAN estimator?

7.3.3 Derive the asymptotic variances of the MLEs of Problems 1–3 of Section 5.6 and compare the results with the large sample approximations of Problem 4 of Section 5.6.

7.3.4 Let $X_1, \ldots, X_n$ be i.i.d. random variables. The distribution of X_1 as that of $N(\mu, \sigma^2)$. Derive the asymptotic variance of the MLE of $\Phi(\mu/\sigma)$.

7.3.5 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a log-normal distribution $LN(\mu, \sigma^2)$. What is the asymptotic covariance matrix of the MLE of $\xi = \exp\{\mu + \sigma^2/2\}$ and $D^2 = \xi^2 \exp\{\sigma^2 - 1\}$?

Section 7.4

7.4.1 Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having a normal distribution $N(\mu, \sigma^2)$, $-\infty < \mu < \infty$, $0 < \sigma < \infty$. Let $\theta = e^{\mu}$.

(i) What is the bias of the MLE $\hat{\theta}_n$?

(ii) Let $\hat{\hat{\theta}}_n$ be the bias adjusted MLE. What is $\hat{\hat{\theta}}_n$, and what is the order of its bias, in terms of n?

(iii) What is the second order deficiency coefficient of $\hat{\hat{\theta}}_n$?

7.4.2 Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables, $X_1 \sim G\left(\frac{1}{\beta}, 1\right)$, $0 < \beta < \infty$. Let $\theta = e^{-1/\beta}$, $0 < \theta < 1$.

(i) What is the MLE of θ?

(ii) Use the delta method to find the bias of the MLE, $\hat{\theta}_n$, up to $O\left(\frac{1}{n^2}\right)$.

(iii) What is the second-order deficiency coefficient of the bias adjusted MLE?

7.4.3 Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having a one-parameter canonical exponential type p.d.f. Show that the first order bias term of the MLE $\hat{\psi}_n$ is

$$B_n(\psi) = -\frac{1}{2n} \cdot \frac{K^{(3)}(\psi)}{I(\psi)}.$$

Section 7.5

7.5.1 In a random sample of size $n = 50$ of random vectors (X, Y) from a bivariate normal distribution, $-\infty < \mu, \eta < \infty$, $0 < \sigma_1, \sigma_2 < \infty$, $-1 < \rho < 1$, the MLE of ρ is $\hat{\rho} = 0.85$. Apply the variance stabilizing transformation to determine asymptotic confidence limits of $\phi = \sin^{-1}(\rho)$; $-\frac{\pi}{2} < \phi < \frac{\pi}{2}$.

7.5.2 Let S_n^2 be the sample variance in a random sample from a normal distribution $N(\mu, \sigma^2)$. Show that the asymptotic variance of

$$W_n = \frac{1}{\sqrt{2}} \log(S_n^2) \text{ is } AV\{W_n\} = \frac{1}{n}.$$

Suppose that $n = 250$ and $S_n^2 = 17.39$. Apply the above transformation to determine asymptotic confidence limits, at level $1 - \alpha = 0.95$, for σ^2.

7.5.3 Let $X_1, \ldots, X_n$ be a random sample (i.i.d.) from $N(\mu, \sigma^2)$; $-\infty < \mu < \infty$, $0 < \sigma^2 < \infty$.

(i) Show that the asymptotic variance of the MLE of σ is $\sigma^2/2n$.

(ii) Determine asymptotic confidence intervals at level $(1 - \alpha)$ for $\omega = \mu + Z_\gamma \sigma$.

(iii) Determine asymptotic confidence intervals at level $(1 - \alpha)$ for μ/σ and for $\Phi(\mu/\sigma)$.

7.5.4 Let $X_1, \ldots, X_n$ be a random sample from a location parameter Laplace distribution; $-\infty < \mu < \infty$. Determine a $(1 - \alpha)$-level asymptotic confidence interval for μ.

Section 7.6

7.6.1 Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having a one-parameter Beta(ν, ν) distribution.

(i) Write the common p.d.f. $f(x; \nu)$ in a canonical exponential type form.

(ii) What is the MLE, $\hat{\nu}_n$?

(iii) Write the Edgeworth expansion approximation to the distribution of the MLE $\hat{\nu}_n$.

7.6.2 In continuation of the previous problem, derive the p^*-formula of the density of the MLE, $\hat{\nu}_n$?

PART IV: SOLUTION OF SELECTED PROBLEMS

7.1.3 Let $\boldsymbol{\theta}' = (\theta_1, \ldots, \theta_k)$ and $\mathbf{p}' = (p_1, \ldots, p_k)$. Let $D = (\text{diag}(\theta_i(1 - \theta_i))$, $i = 1, \ldots, k)$ be a $k \times k$ diagonal matrix. Generally, we denote $\mathbf{X}_n \sim AN\left(\boldsymbol{\xi}, \frac{1}{n} V\right)$ if $\sqrt{n}(\mathbf{X}_n - \boldsymbol{\xi}) \xrightarrow{d} N(\mathbf{0}, V)$ as $n \to \infty$. [$AN(\cdot, \cdot)$ stands for 'asymptotically normal']. In the present case, $\mathbf{p} \sim AN\left(\boldsymbol{\theta}, \frac{1}{n} D\right)$.

Let $H_0 : \theta_1 = \theta_2 = \cdots = \theta_k = \theta$. Then, if H_0 is true,

$$\mathbf{p} \overset{H_0}{\sim} AN\left(\theta 1_k, \frac{\theta(1-\theta)}{n} I_k\right).$$

Now, $\sum_{i=1}^{k}(p_i - \bar{p}_k)^2 = \mathbf{p}'\left(I_k - \frac{1}{k}J_k\right)\mathbf{p}$, where $J_k = 1_k 1_k'$. Since $\left(I_k - \frac{1}{k}J_k\right)$ is idempotent, of rank $(k-1)$, $n\mathbf{p}'\left(I_k - \frac{1}{k}J_k\right)\mathbf{p} \xrightarrow[n\to\infty]{d} \theta(1-\theta)\chi^2[k-1]$. Moreover, $\bar{p}_k = \frac{1}{k}\sum_{i=1}^{k} p_i \to \theta$ a.s., as $n \to \infty$. Thus, by Slutsky's Theorem

$$\frac{n\sum_{i=1}^{k}(p_i - \bar{p}_k)^2}{\bar{p}_k(1-\bar{p}_k)} \xrightarrow{d} \chi^2[k-1],$$

and

$$\lim_{n\to\infty} P\left\{\frac{n\sum_{i=1}^{k}(p_i - \bar{p}_k)^2}{\bar{p}_k(1-\bar{p}_k)} \geq \chi^2_{1-\alpha}[k-1]\right\} = \alpha.$$

If H_0 is not true, $\sum_{i=1}^{k}(\theta_i - \bar{\theta}_k)^2 > 0$. Also,

$$\lim_{n\to\infty}\sum_{i=1}^{k}(p_i - \bar{p}_k)^2 = \sum_{i=1}^{k}(\theta_i - \bar{\theta}_k)^2 > 0 \quad \text{a.s.}$$

Thus, under H_1,

$$\lim_{n\to\infty} P\left\{\frac{n\sum_{i=1}^{k}(p_i - \bar{p}_k)^2}{\bar{p}_k(1-\bar{p}_k)} \leq \chi^2_{1-\alpha}[k-1]\right\} = 0.$$

Thus, the test is consistent.

7.2.3 The MLE of θ is $\hat{\theta}_n = \frac{2J_1 + J_2}{2n}$. $J_1 \sim B(n, \theta^2)$. Hence, $\frac{J_1}{n} \xrightarrow{\text{a.s.}} \theta^2$ and $\frac{J_2}{n} \xrightarrow{\text{a.s.}} 2\theta(1-\theta)$. Thus, $\lim_{n\to\infty}\hat{\theta}_n = \theta$ a.s.

7.3.1 By SLLN, $\lim_{n\to\infty} \frac{\bar{X}_n}{\bar{Y}_n} = \frac{\xi}{\eta}$ a.s.

$$\sqrt{n}\left(\frac{\bar{X}_n}{\bar{Y}_n} - \frac{\xi}{\eta}\right) \xrightarrow{d} N(0, D^2) \text{ as } n \to \infty,$$

where

$$D^2 = \sigma_1^2 \frac{1}{\eta^2} + \sigma_2^2 \frac{\xi^2}{\eta^4} - 2\frac{\xi}{\eta^3}\rho\sigma_1\sigma_2.$$

Thus, $\frac{\bar{X}_n}{\bar{Y}_n} \sim AN\left(\frac{\xi}{\eta}, \frac{1}{n}D^2\right)$.

7.3.2 As shown in Example 7.16,

$$f(X;\theta) = \frac{1}{\pi(1+(x-\theta)^2)}, \quad -\infty < x < \infty.$$

Also,

$$\hat{\theta}_n = F_n^{-1}\left(\frac{1}{2}\right) \sim AN\left(\theta, \frac{\pi^2}{4n}\right).$$

On the other hand, the Fisher information is $I_n(\theta) = \frac{n}{2}$. Thus, $AV(\hat{\theta}_n) = \frac{\pi^2}{4n} > \frac{1}{I_n(\theta)} = \frac{2}{n}$. Thus, $\hat{\theta}_n$ is *not* a BAN estimator.

7.4.2

(i) $X \sim \beta G(1, 1)$. Hence, the MLE of β is $\bar{X}_n$. It follows that the MLE of $\theta = e^{-1/\beta}$ is $\hat{\theta}_n = e^{-1/\bar{X}_n}$.

(ii)

$$\hat{\theta}_n - \theta = (\bar{X}_n - \beta)\frac{1}{\beta^2}e^{-1/\beta} + \frac{1}{2}(\bar{X}_n - \beta)^2 e^{-1/\beta} \cdot \frac{1}{\beta^3}\left(\frac{1}{\beta} - 2\right) + o_p\left(\frac{1}{n}\right).$$

Hence,

$$\begin{aligned}\text{Bias}(\hat{\theta}_n) &= E\{\hat{\theta}_n - \theta\} \\ &= \frac{\beta^2 e^{-1/\beta}}{2n\beta^3}\left(-2 + \frac{1}{\beta}\right) + o\left(\frac{1}{n}\right) \\ &= \frac{e^{-1/\beta}}{2n\beta^2}(1 - 2\beta) + o\left(\frac{1}{n}\right).\end{aligned}$$

The bias adjusted estimator is

$$\begin{aligned}\hat{\hat{\theta}}_n &= e^{-1/\bar{X}_n}\left(1 - \frac{1}{2n\bar{X}_n}\left(\frac{1}{\bar{X}_n} - 2\right)\right) \\ &= e^{-1/\bar{X}_n}\left(1 + \frac{1}{n\bar{X}_n} - \frac{1}{2n\bar{X}_n^2}\right).\end{aligned}$$

(iii) Let $f(x) = e^{-1/x}\left(1 + \frac{1}{nx} - \frac{1}{2nx^2}\right)$. Then

$$\begin{aligned}f'(x) &= \frac{1}{x^2}e^{-1/x}\left(1 + \frac{1}{nx} - \frac{1}{2nx^2}\right) \\ &\quad + e^{-1/x}\left(-\frac{1}{nx^2} + \frac{1}{nx^3}\right) \\ &= \frac{1}{x^2}e^{-1/x}\left(1 - \frac{1}{n} + \frac{2}{nx} - \frac{1}{2nx^2}\right).\end{aligned}$$

It follows that

$$AV\{\hat{\hat{\theta}}_n\} = \frac{e^{-2/\beta}}{n\beta^2} - \frac{2e^{-2/\beta}}{n^2\beta^2}\left(1 - \frac{2}{\beta} + \frac{1}{2\beta^2}\right) + O\left(\frac{1}{n^3}\right).$$

Accordingly, the second-order deficiency coefficient of $\hat{\hat{\theta}}_n$ is

$$D = -\frac{2e^{-2/\beta}}{\beta^2}\left(1 - \frac{2}{\beta} + \frac{1}{2\beta^2}\right).$$

7.6.1 $X_1, \ldots, X_n$ are i.i.d. like Beta(ν, ν), $0 < \nu < \infty$.

(i)
$$\begin{aligned} f(x;\nu) &= \frac{1}{B(\nu,\nu)} x^{\nu-1}(1-x)^{\nu-1}, \quad 0 < x < 1 \\ &= \frac{1}{B(\nu,\nu)x(1-x)} e^{\nu \log(x(1-x))} \\ &= \frac{1}{x(1-x)} \exp\{\nu \log(x(1-x)) - K(\nu)\}, \end{aligned}$$

where $K(\nu) = \log B(\nu, \nu)$.

(ii) The likelihood function is equivalent to

$$L(\nu \mid \mathbf{X}) = \exp\left(\nu \sum_{i=1}^{n} \log(X_i(1-X_i)) - nK(\nu)\right).$$

The log likelihood is

$$l(\nu \mid \mathbf{X}) = \nu \sum_{i=1}^{n} \log(X_i(1-X_i)) - nK(\nu).$$

Note that the derivative of $K(\nu)$ is

$$K'(\nu) = 2\left(\frac{\Gamma'(\nu)}{\Gamma(\nu)} - \frac{\Gamma'(2\nu)}{\Gamma(2\nu)}\right).$$

It follows that the MLE of ν is the root of

$$-\frac{1}{2n}\sum_{i=1}^{n} \log(X_i(1-X_i)) = -\frac{\Gamma'(\nu)}{\Gamma(\nu)} + \frac{\Gamma'(2\nu)}{\Gamma(2\nu)}.$$

The function $\frac{d}{d\nu}\log\Gamma(\nu)$ is also called the psi function, i.e., $\psi(\nu) = \frac{d}{d\nu}\log\Gamma(\nu)$. As shown in Abramowitz and Stegun (1968, p. 259), $\psi(2\nu) - \psi(\nu) = \frac{1}{2}(\psi\left(\nu + \frac{1}{2}\right) - \psi(\nu)) + \log 2$. Also, $-\frac{1}{n}\sum_{i=1}^{n} \log(X_i(1-X_i)) > \log 4$. Thus, the MLE is the value of ν for which

$$\psi\left(\nu + \frac{1}{2}\right) - \psi(\nu) = -\frac{1}{n}\sum_{i=1}^{n} \log(X_i(1-X_i)) - \frac{1}{2}\log(2).$$

(iii) Since Beta(ν, ν) is symmetric distribution around $x = \frac{1}{2}$, $X \sim (1 - X)$ and hence

$$V\left(\sum_{i=1}^{n} \log(X_i(1 - X_i))\right) = 4nV\{\log X\}.$$

Thus, since $X_1, \ldots, X_n$ are i.i.d., the Fisher information is $I(\nu) = 4V\{\log X\}$. The first four central moments of Beta(ν, ν) are $\mu_i^* = 0$; $\mu_2^* = \dfrac{1}{4(2\nu + 1)}$; $\mu_3^* = 0$ and $\mu_4^* = \dfrac{2\nu^2 + 8\nu + 5}{16(2\nu + 1)}$. Thus,

$$\beta_1 = 0,$$

and

$$\beta_2 = (2\nu^2 + 8\nu + 5)(2\nu + 1).$$

It follows that the Edgeworth asymptotic approximation to the distribution of the MLE, $\hat{\nu}_n$, is

$$P\{\sqrt{nI(\nu)}\,(\hat{\nu}_n - \nu) \leq x\} \cong \Phi(x) - \frac{x(x^2 - 3)}{24n}\left(\frac{2\nu^2 + 8\nu + 5}{16(2\nu + 1)} - 3\right).$$

CHAPTER 8

Bayesian Analysis in Testing and Estimation

PART I: THEORY

This chapter is devoted to some topics of estimation and testing hypotheses from the point of view of statistical decision theory. The decision theoretic approach provides a general framework for both estimation of parameters and testing hypotheses. The objective is to study classes of procedures in terms of certain associated risk functions and determine the existence of optimal procedures. The results that we have presented in the previous chapters on minimum mean-squared-error (MSE) estimators and on most powerful tests can be considered as part of the general statistical decision theory. We have seen that uniformly minimum MSE estimators and uniformly most powerful tests exist only in special cases. One could overcome this difficulty by considering procedures that yield minimum average risk, where the risk is defined as the expected loss due to erroneous decision, according to the particular distribution F_θ. The MSE in estimation and the error probabilities in testing are special risk functions. The risk functions depend on the parameters θ of the parent distribution. The average risk can be defined as an expected risk according to some probability distribution on the parameter space. Statistical inference that considers the parameter(s) as random variables is called a **Bayesian inference**. The expected risk with respect to the distribution of $\boldsymbol{\theta}$ is called in Bayesian theory the **prior risk**, and the probability measure on the parameter space is called a **prior distribution**. The estimators or test functions that minimize the prior risk, with respect to some prior distribution, are called **Bayes procedures** for the specified prior distribution. Bayes procedures have certain desirable properties. This chapter is devoted, therefore, to the study of the structure of optimal decision rules in the framework of Bayesian theory. We start Section 8.1 with a general discussion of the basic Bayesian tools and information functions. We outline the decision theory and provide an example of an optimal

Examples and Problems in Mathematical Statistics, First Edition. Shelemyahu Zacks.

statistical decision procedure. In Section 8.2, we discuss testing of hypotheses from the Bayesian point of view, and in Section 8.3, we present Bayes credibility intervals. The Bayesian theory of point estimation is discussed in Section 8.4. Section 8.5 discusses analytical and numerical techniques for evaluating posterior distributions on complex cases. Section 8.6 is devoted to empirical Bayes procedures.

8.1 THE BAYESIAN FRAMEWORK

8.1.1 Prior, Posterior, and Predictive Distributions

In the previous chapters, we discussed problems of statistical inference, testing hypotheses, and estimation, considering the parameters of the statistical models as fixed unknown constants. This is the so-called classical approach to the problems of statistical inference. In the Bayesian approach, the unknown parameters are considered as values determined at random according to some specified distribution, called the **prior distribution**. This prior distribution can be conceived as a normalized nonnegative weight function that the statistician assigns to the various possible parameter values. It can express his degree of belief in the various parameter values or the amount of prior information available on the parameters. For the philosophical foundations of the Bayesian theory, see the books of DeFinneti (1974), Barnett (1973), Hacking (1965), Savage (1962), and Schervish (1995). We discuss here only the basic mathematical structure.

Let $\mathcal{F} = \{F(x;\theta); \theta \in \Theta\}$ be a family of distribution functions specified by the statistical model. The parameters θ of the elements of $\mathcal{F}$ are real or vector valued parameters. The parameter space Θ is specified by the model. Let $\mathcal{H}$ be a family of distribution functions defined on the parameter space Θ. The statistician chooses an element $H(\theta)$ of $\mathcal{H}$ and assigns it the role of a **prior** distribution. The actual parameter value θ_0 of the distribution of the observable random variable X is considered to be a realization of a random variable having the distribution $H(\theta)$. After observing the value of X the statistician adjusts his prior information on the value of the parameter θ by converting $H(\theta)$ to the **posterior** distribution $H(\theta \mid X)$. This is done by Bayes Theorem according to which if $h(\theta)$ is the prior probability density function (p.d.f.) of θ and $f(x;\theta)$ the p.d.f. of X under θ, then the posterior p.d.f. of θ is

$$h(\theta \mid x) = h(\theta) f(x;\theta) / \int_{\Theta} f(x;\theta) dH(\theta). \tag{8.1.1}$$

If we are given a sample of n observations or random variables $X_1, X_2, \ldots, X_n$, whose distributions belong to a family $\mathcal{F}$, the question is whether these random variables are independent identically distributed (i.i.d.) given θ, or whether θ might be randomly chosen from $H(\theta)$ for each observation.

At the beginning, we study the case that $X_1, \ldots, X_n$ are conditionally i.i.d., given θ. This is the classical Bayesian model. In Section 8.6, we study the so-called *empirical Bayes* model, in which θ is randomly chosen from $H(\theta)$ for each observation. In

the classical model, if the family $\mathcal{F}$ admits a sufficient statistic $T(X)$, then for any prior distribution $H(\theta)$, the posterior distribution is a function of $T(X)$, and can be determined from the distribution of $T(X)$ under θ. Indeed, by the Neyman–Fisher Factorization Theorem, if $T(X)$ is sufficient for $\mathcal{F}$ then $f(x;\theta) = k(x)g(T(x);\theta)$. Hence,

$$h(\theta \mid x) = h(\theta)g(T(x);\theta) / \int_{\Theta} g(T(x);\theta)dH(\theta). \tag{8.1.2}$$

Thus, the posterior p.d.f. is a function of $T(X)$. Moreover, the p.d.f. of $T(X)$ is $g^*(t;\theta) = k^*(t)g(t;\theta)$, where $k^*(t)$ is independent of θ. It follows that the conditional p.d.f. of θ given $\{T(X) = t\}$ coincides with $h(\theta \mid x)$ on the sets $\{x; T(x) = t\}$ for all t.

Bayes **predictive distributions** are the marginal distributions of the observed random variables, according to the model. More specifically, if a random vector $\mathbf{X}$ has a joint distribution $F(\mathbf{x};\theta)$ and the prior distribution of θ is $H(\theta)$ then the joint predictive distribution of $\mathbf{X}$ under H is

$$F_H(\mathbf{x}) = \int_{\Theta} F(\mathbf{x};\theta)dH(\theta). \tag{8.1.3}$$

A most important question in Bayesian analysis is what prior distribution to choose. The answer is, generally, that the prior distribution should reflect possible prior knowledge available on possible values of the parameter. In many situations, the prior information on the parameters is vague. In such cases, we may use formal prior distributions, which are discussed in Section 8.1.3. On the other hand, in certain scientific or technological experiments much is known about possible values of the parameters. This may guide in selecting a prior distribution, as illustrated in the examples.

There are many examples of posterior distribution that belong to the same parametric family of the prior distribution. Generally, if the family of prior distributions $\mathcal{H}$ relative to a specific family $\mathcal{F}$ yields posteriors in $\mathcal{H}$, we say that $\mathcal{F}$ and $\mathcal{H}$ are **conjugate** families. For more discussion on conjugate prior distributions, see Raiffa and Schlaifer (1961). In Example 8.2, we illustrate a few conjugate prior families.

The situation when conjugate prior structure exists is relatively simple and generally leads to analytic expression of the posterior distribution. In research, however, we often encounter much more difficult problems, as illustrated in Example 8.3. In such cases, we cannot often express the posterior distribution in analytic form, and have to resort to numerical evaluations to be discussed in Section 8.5.

8.1.2 Noninformative and Improper Prior Distributions

It is sometimes tempting to obtain posterior densities by multiplying the likelihood function by a function $h(\theta)$, which is not a proper p.d.f. For example, suppose that

$X \mid \theta \sim N(\theta, 1)$. In this case $L(\theta; X) = \exp\left\{-\frac{1}{2}(\theta - x)^2\right\}$. This likelihood function is integrable with respect to $d\theta$. Indeed,

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}(\theta - X)^2\right\} d\theta = \sqrt{2\pi}.$$

Thus, if we consider formally the function $h(\theta)d\theta = cd\theta$ or $h(\theta) = c$ then

$$h(\theta \mid X) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(\theta - X)^2\right\}, \tag{8.1.4}$$

which is the p.d.f. of $N(X, 1)$. The function $h(\theta) = c$, $c > 0$ for all θ is called an **improper prior density** since $\int_{-\infty}^{\infty} cd\theta = \infty$. Another example is when $X \mid \lambda \sim \mathbf{P}(\lambda)$, i.e., $L(\lambda \mid X) = e^{-\lambda}\lambda^X$. If we use the improper prior density $h(\lambda) = c > 0$ for all $\lambda > 0$ then the posterior p.d.f. is

$$h(\lambda \mid X) = \frac{1}{X!}\lambda^X e^{-\lambda}, \quad 0 < \lambda < \infty. \tag{8.1.5}$$

This is a proper p.d.f. of $G(1, X + 1)$ despite the fact that $h(\lambda)$ is an improper prior density. Some people justify the use of an improper prior by arguing that it provides a "diffused" prior, yielding an equal weight to all points in the parameter space. For example, the improper priors that lead to the proper posterior densities (8.1.4) and (8.1.5) may reflect a state of ignorance, in which all points θ in $(-\infty, \infty)$ or λ in $(0, \infty)$ are "equally" likely.

Lindley (1956) defines a prior density $h(\theta)$ to be **noninformative**, if it maximizes the predictive gain in information on θ when a random sample of size n is observed. He shows then that, in large samples, if the family $\mathcal{F}$ satisfies the Cramer–Rao regularity conditions, and the maximum likelihood estimator (MLE) $\hat{\theta}_n$ is minimal sufficient for $\mathcal{F}$, then the noninformative prior density is proportional to $|I(\theta)|^{1/2}$, where $|I(\boldsymbol{\theta})|$ is the determinant of the Fisher information matrix. As will be shown in Example 8.4, $h(\boldsymbol{\theta}) \propto |I(\theta)|^{1/2}$ is sometimes a proper p.d.f. and sometimes an improper one.

Jeffreys (1961) justified the use of the noninformative prior $|I(\theta)|^{1/2}$ on the basis of invariance. He argued that if a statistical model $\mathcal{F} = \{f(x; \boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$ is reparametrized to $\mathcal{F}^* = \{f^*(x; \omega); \omega \in \Omega\}$, where $\boldsymbol{\omega} = \boldsymbol{\phi}(\boldsymbol{\theta})$ then the prior density $h(\theta)$ should be chosen so that $h(\theta \mid X) = h(\omega \mid X)$.

Let $\boldsymbol{\theta} = \boldsymbol{\phi}^{-1}(\boldsymbol{\omega})$ and let $J(\boldsymbol{\omega})$ be the Jacobian of the transformation, then the posterior p.d.f. of $\boldsymbol{\omega}$ is

$$h^*(\boldsymbol{\omega} \mid X) \propto h(\phi^{-1}(\boldsymbol{\omega}))f(x; \phi^{-1}(\boldsymbol{\omega}))|J(\boldsymbol{\omega})|. \tag{8.1.6}$$

Recall that the Fisher information matrix of $\boldsymbol{\omega}$ is

$$I^*(\boldsymbol{\omega}) = J^{-1}(\boldsymbol{\omega})I(\phi^{-1}(\boldsymbol{\omega}))J^{-1}(\boldsymbol{\omega}). \tag{8.1.7}$$

Thus, if $h(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{1/2}$ then from (8.1.7) and (8.1.8), since

$$|I^*(\boldsymbol{\omega})|^{1/2} = |I(\phi^{-1}(\boldsymbol{\omega}))|^{1/2}/|J(\boldsymbol{\omega})|, \tag{8.1.8}$$

we obtain

$$h^*(\omega \mid X) \propto f(x; \phi^{-1}(\boldsymbol{\omega}))|I^*(\omega)|^{1/2}. \tag{8.1.9}$$

The structure of $h(\theta \mid X)$ and of $h^*(\omega \mid X)$ is similar. This is the "invariance" property of the posterior, with respect to transformations of the parameter.

A prior density proportional to $|I(\boldsymbol{\theta})|^{1/2}$ is called a **Jeffreys prior density**.

8.1.3 Risk Functions and Bayes Procedures

In statistical decision theory, we consider the problems of inference in terms of a specified set of actions, $\mathcal{A}$, and their outcomes. The outcome of the decision is expressed in terms of some **utility function**, which provides numerical quantities associated with actions of $\mathcal{A}$ and the given parameters, θ, characterizing the elements of the family $\mathcal{F}$ specified by the model. Instead of discussing utility functions, we discuss here **loss functions**, $L(a, \theta)$, $a \in \mathcal{A}$, $\theta \in \Theta$, associated with actions and parameters. The loss functions are nonnegative functions that assume the value zero if the action chosen does not imply some utility loss when θ is the true state of Nature. One of the important questions is what type of loss function to consider. The answer to this question depends on the decision problem and on the structure of the model. In the classical approach to testing hypotheses, the loss function assumes the value zero if no error is committed and the value one if an error of either kind is done. In a decision theoretic approach, testing hypotheses can be performed with more general loss functions, as will be shown in Section 8.2. In estimation theory, the squared-error loss function $(\hat{\theta}(x) - \theta)^2$ is frequently applied, when $\hat{\theta}(x)$ is an estimator of θ. A generalization of this type of loss function, which is of theoretical importance, is the general class of quadratic loss function, given by

$$L(\hat{\theta}(x), \theta) = Q(\theta)(\hat{\theta}(x) - \theta)^2, \tag{8.1.10}$$

where $Q(\theta) > 0$ is an appropriate function of θ. For example, $(\hat{\theta}(x) - \theta)^2/\theta^2$ is a quadratic loss function. Another type of loss function used in estimation theory is the type of function that depends on $\hat{\theta}(x)$ and θ only through the absolute value of their difference. That is, $L(\hat{\theta}(x), \theta) = W(|\hat{\theta}(x) - \theta|)$. For example, $|\hat{\theta}(x) - \theta|^\nu$ where $\nu > 0$, or $\log(1 + |\hat{\theta}(x) - \theta|)$. Bilinear convex functions of the form

$$L(\hat{\theta}, \theta) = a_1(\hat{\theta} - \theta)^- + a_2(\hat{\theta} - \theta)^+ \tag{8.1.11}$$

are also in use, where a_1, a_2 are positive constants; $(\hat{\theta} - \theta)^- = -\min(\hat{\theta} - \theta, 0)$ and $(\hat{\theta} - \theta)^+ = \max(\hat{\theta} - \theta, 0)$. If the value of θ is known one can always choose a proper action to insure no loss. The essence of statistical decision problems is that the true parameter θ is unknown and decisions are made under uncertainty. The random vector $\mathbf{X} = (X_1, \ldots, X_n)$ provides information about the unknown value of θ. A function from the sample space $\mathcal{X}$ of $\mathbf{X}$ into the action space $\mathcal{A}$ is called a **decision function**. We denote it by $d(\mathbf{X})$ and require that it should be a statistic. Let $\mathcal{D}$ denotes a specified set or class of proper decision functions. Using a decision function $d(\mathbf{X})$ the associated loss $L(d(\mathbf{X}), \theta)$ is a random variable, for each θ. The expected loss under θ, associated with a decision function $d(\mathbf{X})$, is called the **risk function** and is denoted by $R(d, \theta) = E_\theta\{L(d(\mathbf{X}), \theta)\}$. Given the structure of a statistical decision problem, the objective is to select an optimal decision function from $\mathcal{D}$. Ideally, we would like to choose a decision function $d^0(\mathbf{X})$ that minimizes the associated risk function $R(d, \theta)$ uniformly in θ. Such a uniformly optimal decision function may not exist, since the function d^0 for which $R(d^0, \theta) = \inf_{\mathcal{D}} R(d, \theta)$ generally depends on the particular value of θ under consideration. There are several ways to overcome this difficulty. One approach is to restrict attention to a subclass of decision functions, like unbiased or invariant decision functions. Another approach for determining optimal decision functions is the **Bayesian approach**. We define here the notion of Bayes decision function in a general context.

Consider a specified prior distribution, $H(\theta)$, defined over the parameter space Θ. With respect to this prior distribution, we define the prior risk, $\rho(d, H)$, as the expected risk value when θ varies over Θ, i.e.,

$$\rho(d, H) = \int_\Theta R(d, \theta) h(\theta) d\theta, \tag{8.1.12}$$

where $h(\theta)$ is the corresponding p.d.f. A **Bayes decision function**, with respect to a prior distribution H, is a decision function $d_H(x)$ that minimizes the prior risk $\rho(d, H)$, i.e.,

$$\rho(d_H, H) = \inf_{\mathcal{D}} \rho(d, H). \tag{8.1.13}$$

Under some general conditions, a Bayes decision function $d_H(x)$ exists. The Bayes decision function can be generally determined by minimizing the posterior expectation of the loss function for a given value x of the random variable X. Indeed, since $L(d, \theta) \geq 0$ one can interchange the integration operations below and write

$$\begin{aligned} \rho(d, H) &= \int_\Theta \left\{ \int_\chi L(d(x), \theta) f(x; \theta) dx \right\} h(\theta) d\theta \\ &= \int_\chi f_H(x) \left\{ \int_\Theta L(d(x), \theta) \frac{f(x; \theta) h(\theta)}{f_H(x)} d\theta \right\} dx, \end{aligned} \tag{8.1.14}$$

where $f_H(x) = \int f(x;\tau)h(\tau)d\tau$ is the predictive p.d.f. The conditional p.d.f. $h(\theta \mid x) = f(x;\theta)h(\theta)/f_H(x)$ is the **posterior p.d.f. of** θ, **given** $X = x$. Similarly, the conditional expectation

$$R(d(x), H) = \int_{\Theta} L(d(x), \theta)h(\theta \mid x)d\theta \tag{8.1.15}$$

is called the **posterior risk** of $d(x)$ under H. Thus, for a given $X = x$, we can choose $d(x)$ to minimize $R(d(x), H)$. Since $L(d(x), \theta) \geq 0$ for all $\theta \in \Theta$ and $d \in \mathcal{D}$, the minimization of the posterior risk minimizes also the prior risk $\rho(d, H)$. Thus, $d_H(X)$ is a Bayes decision function.

8.2 BAYESIAN TESTING OF HYPOTHESIS

8.2.1 Testing Simple Hypothesis

We start with the problem of testing two **simple** hypotheses H_0 and H_1. Let $F_0(x)$ and $F_1(x)$ be two specified distribution functions. The hypothesis H_0 specifies the parent distribution of X as $F_0(x)$, H_1 specified it as $F_1(x)$. Let $f_0(x)$ and $f_1(x)$ be the p.d.f.s corresponding to $F_0(x)$ and $F_1(x)$, respectively. Let π, $0 \leq \pi \leq 1$, be the prior probability that H_0 is true. In the special case of two simple hypotheses, the loss function can assign 1 unit to the case of rejecting H_0 when it is true and b units to the case of rejecting H_1 when it is true. The prior risks associated with accepting H_0 and H_1 are, respectively, $\rho_0(\pi) = (1 - \pi)b$ and $\rho_1(\pi) = \pi$. For a given value of π, we accept hypothesis H_i $(i = 0, 1)$ if $\rho_i(\pi)$ is the minimal prior risk. Thus, a Bayes rule, prior to making observations is

$$d = \begin{cases} 0, & \text{if } \pi \geq b/(1 + b), \\ 1, & \text{otherwise,} \end{cases} \tag{8.2.1}$$

where $d = i$ is the decision to accept H_i $(i = 0, 1)$.

Suppose that a sample of n i.i.d. random variables $X_1, \ldots, X_n$ has been observed. After observing the sample, we determine the posterior probability $\pi(\mathbf{X}_n)$ that H_0 is true. This posterior probability is given by

$$\pi(\mathbf{X}_n) = \pi \prod_{j=1}^{n} f_0(X_j) \Big/ \left[\pi \cdot \prod_{j=1}^{n} f_0(X_j) + (1 - \pi) \prod_{j=1}^{n} f_1(X_j) \right]. \tag{8.2.2}$$

We use the decision rule (8.2.1) with π replaced by $\pi(\mathbf{X}_n)$. Thus, the Bayes decision function is

$$d(\mathbf{X}_n) = \begin{cases} 0, & \text{if } \pi(\mathbf{X}_n) \geq \dfrac{b}{1+b}, \\ 1, & \text{otherwise.} \end{cases} \tag{8.2.3}$$

The Bayes decision function can be written in terms of the test function discussed in Chapter 4 as

$$\phi_\pi(\mathbf{X}_n) = \begin{cases} 1, & \text{if } \prod_{j=1}^{n} \dfrac{f_1(X_j)}{f_0(X_j)} \geq \dfrac{\pi}{b(1-\pi)}, \\ 0, & \text{otherwise.} \end{cases} \tag{8.2.4}$$

The Bayes test function $\phi_\pi(\mathbf{X}_n)$ is similar to the Neyman–Pearson most powerful test, except that the Bayes test is not necessarily randomized even if the distributions $F_i(x)$ are discrete. Moreover, the likelihood ratio $\prod_{j=1}^{n} f_1(X_j)/f_0(X_j)$ is compared to the ratio of the prior risks.

We discuss now some of the important optimality characteristics of Bayes tests of two simple hypotheses. Let $R_0(\phi)$ and $R_1(\phi)$ denote the risks associated with an arbitrary test statistic ϕ, when H_0 or H_1 are true, respectively. Let $R_0(\pi)$ and $R_1(\pi)$ denote the corresponding risk values of a Bayes test function, with respect to a prior probability π. Generally

$$R_0(\phi) = c_1 \epsilon_0(\phi); \quad 0 < c_1 < \infty$$

and

$$R_1(\phi) = c_2 \epsilon_1(\phi); \quad 0 < c_2 < \infty,$$

where $\epsilon_0(\phi)$ and $\epsilon_1(\phi)$ are the error probabilities of the test statistic ϕ, c_1 and c_2 are costs of erroneous decisions. The set $R = \{R_0(\phi), R_1(\phi)); \text{ all test functions } \phi\}$ is called the risk set. Since for every $0 \leq \alpha \leq 1$ and any functions $\phi^{(1)}$ and $\phi^{(2)}$, $\alpha\phi^{(1)} + (1-\alpha)\phi^{(2)}$ is also a test function, and since

$$R_i(\alpha\phi^{(1)} + (1-\alpha)\phi^{(2)}) = \alpha R_i(\phi^{(1)}) + (1-\alpha) R_i(\phi^{(2)}), \quad i = 0, 1 \tag{8.2.5}$$

the risk set R is convex. Moreover, the set

$$S = \{(R_0(\pi), R_1(\pi)); \ 0 \leq \pi \leq 1\} \tag{8.2.6}$$

of all risk points corresponding to the Bayes tests is the lower boundary for R. Indeed, according to (8.2.4) and the Neyman–Pearson Lemma, $R_1(\pi)$ is the smallest possible

risk of all test functions ϕ with $R_0(\phi) = R_0(\pi)$. Accordingly, all the Bayes tests constitute a **complete class** in the sense that, for any test function outside the class, there exists a corresponding Bayes test with a risk point having component smaller or equal to those of that particular test and at least one component is strictly smaller (Ferguson, 1967, Ch. 2). From the decision theoretic point of view there is no sense in considering test functions that do not belong to the complete class. These results can be generalized to the case of testing k simple hypotheses (Blackwell and Girshick, 1954; Ferguson, 1967).

8.2.2 Testing Composite Hypotheses

Let Θ_0 and Θ_1 be the sets of θ-points corresponding to the (composite) hypotheses H_0 and H_1, respectively. These sets contain finite or infinite number of points. Let $H(\theta)$ be a prior distribution function specified over $\Theta = \Theta_0 \cup \Theta_1$. The posterior probability of H_0, given n i.i.d. random variables $X_1, \ldots, X_n$, is

$$\pi(\mathbf{X}_n) = \frac{\int_{\Theta_0} \prod_{i=1}^{n} f(X_i;\theta) dH(\theta)}{\int_{\Theta} \prod_{i=1}^{n\theta} f(X_i;\theta) dH(\theta)}, \tag{8.2.7}$$

where $f(x;\theta)$ is the p.d.f. of X under θ. The notation in (8.2.7) signifies that if the sets are discrete the corresponding integrals are sums and $dH(\theta)$ are prior probabilities, otherwise $dH(\theta) = h(\theta)d\theta$, where $h(\theta)$ is a p.d.f. The Bayes decision rule is obtained by computing the **posterior** risk associated with accepting H_0 or with accepting H_1 and making the decision associated with the minimal posterior risk. The form of the Bayes test depends, therefore, on the loss function employed.

If the loss functions associated with accepting H_0 or H_1 are

$$L_0(\theta) = c_0 I\{\theta \in \Theta_1\} \text{ and } L_1(\theta) = c_1 I\{\theta \in \Theta_0\}$$

then the associated posterior risk functions are

$$R_0(\mathbf{X}) = c_0 \int_{\Theta_1} f(\mathbf{X};\theta) dH(\theta) / \int_{\Theta} f(\mathbf{X};\theta) dH(\theta) \tag{8.2.8}$$

and

$$R_1(\mathbf{X}) = c_1 \int_{\Theta_0} f(\mathbf{X};\theta) dH(\theta) / \int_{\Theta} f(\mathbf{X};\theta) dH(\theta).$$

In this case, the Bayes test function is

$$\phi_H(\mathbf{X}) = \begin{cases} 1, & \text{if } c_1 \int_{\Theta_0} f(\mathbf{X};\theta)dH(\theta) < c_0 \int_{\Theta_1} f(\mathbf{X};\theta)dH(\theta), \\ 0, & \text{otherwise.} \end{cases} \tag{8.2.9}$$

In other words, the hypothesis H_0 is rejected if the **predictive** likelihood ratio

$$\Lambda_H(\mathbf{X}) = \int_{\Theta_1} f(\mathbf{X};\theta)dH(\theta) \Big/ \int_{\Theta_0} f(\mathbf{X};\theta)dH(\theta) \tag{8.2.10}$$

is greater than the loss ratio c_1/c_0. This can be considered as a generalization of (8.2.4). The predictive likelihood ratio $\Lambda_H(\mathbf{X})$ is called also the **Bayes Factor** in favor of H_1 against H_0 (Good, 1965, 1967).

Cornfield (1969) suggested as a test function the ratio of the posterior odds in favor of H_0, i.e., $P[H_0 \mid \mathbf{X}]/(1 - P[H_0 \mid \mathbf{X}])$, to the prior odds $\pi/(1-\pi)$ where $\pi = P[H_0]$ is the prior probability of H_0. The rule is to reject H_0 when this ratio is smaller than a suitable constant. Cornfield called this statistic the **relative betting odds**. Note that this relative betting odds is $[\Lambda_H(\mathbf{X})\pi/(1-\pi)]^{-1}$. We see that Cornfield's test function is equivalent to (8.2.9) for suitably chosen cost factors.

Karlin (l956) and Karlin and Rubin (1956) proved that in monotone likelihood ratio families the Bayes test function is monotone in the sufficient statistic $T(\mathbf{X})$. For testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$, the Bayes procedure rejects H_0 whenever $T(\mathbf{X}) \geq \xi_0$. The result can be further generalized to the problem of testing multiple hypotheses (Zacks, 1971; Ch. 10).

The problem of testing the composite hypothesis that all the probabilities in a multinomial distribution have the same value has drawn considerable attention in the statistical literature; see in particular the papers of Good (1967), Good and Crook (1974), and Good (1975). The Bayes test procedure proposed by Good (1967) is based on the symmetric Dirichlet prior distribution. More specifically if $\mathbf{X} = (X_1, \ldots, X_k)'$ is a random vector having the multinomial distribution $M(n, \boldsymbol{\theta})$ then the parameter vector $\boldsymbol{\theta}$ is ascribed the prior distribution with p.d.f.

$$h(\theta_1, \ldots, \theta_k) = \frac{\Gamma(k\nu)}{\Gamma^k(\nu)} \prod_{i=1}^{k} \theta_i^{\nu-1}, \tag{8.2.11}$$

$0 < \theta_1, \ldots, \theta_k < 1$ and $\sum_{i=1}^{k} \theta_i = 1$. The Bayes factor for testing $h_0 : \boldsymbol{\theta} = \frac{1}{k}\mathbf{1}$ against the composite alternative hypothesis $H_1 : \boldsymbol{\theta} \neq \frac{1}{k}\mathbf{1}$, where $\mathbf{1} = (1, \ldots, 1)'$, according to (8.2.10) is

$$\Lambda(\nu; \mathbf{X}) = \frac{k^n \Gamma(k\nu) \prod_{i=1}^{k} \Gamma(\nu + X_i)}{\Gamma^k(\nu)\Gamma(\nu k + n)}. \tag{8.2.12}$$

From the purely Bayesian point of view, the statistician should be able to choose an appropriate value of ν and some cost ratio c_1/c_0 for erroneous decisions, according to subjective judgment, and reject H_0 if $\Lambda(\nu; \mathbf{X}) \geq c_1/c_0$. In practice, it is generally not so simple to judge what are the appropriate values of ν and c_1/c_0. Good and Crook (1974) suggested two alternative ways to solve this problem. One suggestion is to consider an integrated Bayes factor

$$\Lambda(\mathbf{X}) = \int_0^\infty \phi(\nu)\Lambda(\nu; \mathbf{X})d\nu, \tag{8.2.13}$$

where $\phi(\nu)$ is the p.d.f. of a log-Cauchy distribution, i.e.,

$$\phi(\nu) = \frac{1}{\nu\pi} \cdot \frac{1}{1 + (\log \nu)^2}, \quad 0 < \nu < \infty. \tag{8.2.14}$$

The second suggestion is to find the value ν^0 for which $\Lambda(\nu; \mathbf{X})$ is maximized and reject H_0 if $\Lambda^* = (2 \log \Lambda(\nu^0; \mathbf{X}))^{1/2}$ exceeds the $(1 - \alpha)$-quantile of the asymptotic distribution of Λ^* under H_0. We see that non-Bayesian (frequentists) considerations are introduced in order to arrive at an appropriate critical level for Λ^*. Good and Crook call this approach a "Bayes/Non-Bayes compromise." We have presented this problem and the approaches suggested for its solution to show that in practical work a nondogmatic approach is needed. It may be reasonable to derive a test statistic in a Bayesian framework and apply it in a non-Bayesian manner.

8.2.3 Bayes Sequential Testing of Hypotheses

We consider in the present section an application of the general theory of Section 8.1.5 to the case of testing two simple hypotheses. We have seen in Section 8.2.1 that the Bayes decision test function, after observing $\mathbf{X}_n$, is to reject H_0 if the posterior probability, $\pi(\mathbf{X}_n)$, that H_0 is true is less than or equal to a constant π^*. The associated Bayes risk is $\rho^{(0)}(\pi(\mathbf{X}_n)) = \pi(\mathbf{X}_n)I\{\pi(\mathbf{X}_n) \leq \pi^*\} + b(1 - \pi(\mathbf{X}_n))I\{\pi(\mathbf{X}_n) > \pi^*\}$, where $\pi^* = b/(1 + b)$. If $\pi(\mathbf{X}_n) = \pi$ then the posterior probability of H_0 after the $(n + 1)$st observation is $\psi(\pi, \mathbf{X}_{n+1}) = \left(1 + \frac{1 - \pi}{\pi} R(X_{n+1})\right)^{-1}$, where $R(x) = \frac{f_1(x)}{f_0(x)}$ is the likelihood ratio. The predictive risk associated with an additional observation is

$$\bar{\rho}_1(\pi) = c + E\{\rho^{(0)}(\psi(\pi, X))\}, \tag{8.2.15}$$

where c is the cost of one observation, and the expectation is with respect to the predictive distribution of X given π. We can show that the function $\bar{\rho}_1(\pi)$ is concave on $[0, 1]$ and thus continuous on $(0, 1)$. Moreover, $\bar{\rho}_1(0) \geq c$ and $\bar{\rho}_1(1) \geq c$. Note that the function $\psi(\pi, X) \to 0$ w.p.1 if $\pi \to 0$ and $\psi(\pi, X) \to 1$ w.p.1 if $\pi \to 1$. Since $\rho^{(0)}(\pi)$ is bounded by π^*, we obtain by the Lebesgue Dominated Convergence

Theorem that $E\{\rho^0(\psi(\pi, X))\} \to 0$ as $\pi \to 0$ or as $\pi \to 1$. The Bayes risk associated with an additional observation is

$$\rho^{(1)}(\pi) = \min\{\rho^{(0)}(\pi), \bar{\rho}_1(\pi)\}. \tag{8.2.16}$$

Thus, if $c \geq b/(1+b)$ it is not optimal to make any observation. On the other hand, if $c < b/(1+b)$ there exist two points $\pi_1^{(1)}$ and $\pi_2^{(1)}$, such that $0 < \pi_1^{(1)} < \pi^* < \pi_2^{(1)} < 1$, and

$$\rho^{(1)}(\pi) = \begin{cases} \rho^{(0)}(\pi), & \text{if } \pi \leq \pi_1^{(1)} \text{ or } \pi \geq \pi_2^{(1)}, \\ \bar{\rho}_1(\pi), & \text{otherwise.} \end{cases} \tag{8.2.17}$$

Let

$$\bar{\rho}_2(\pi) = c + E\{\rho^{(1)}(\psi(\pi, X))\}, \quad 0 \leq \pi \leq 1, \tag{8.2.18}$$

and let

$$\rho^{(2)}(\pi) = \min\{\rho^{(0)}(\pi), \bar{\rho}_2(\pi)\}, \quad 0 \leq \pi \leq 1. \tag{8.2.19}$$

Since $\rho^{(1)}(\psi(\pi, X)) \leq \rho^0(\psi(\pi, X))$ for each π with probability one, we obtain that $\bar{\rho}_2(\pi) \leq \bar{\rho}_1(\pi)$ for all $0 \leq \pi \leq 1$. Thus, $\rho^{(2)}(\pi) \leq \rho^{(1)}(\pi)$ for all $\pi, 0 \leq \pi \leq 1$. $\bar{\rho}_2(\pi)$ is also a concave function of π on $[0, 1]$ and $\bar{\rho}_2(0) = \bar{\rho}_2(1) = c$. Thus, there exists $\pi_1^{(2)} \leq \pi_1^{(1)}$ and $\pi_2^{(2)} \geq \pi_2^{(1)}$ such that

$$\rho^{(2)}(\pi) = \begin{cases} \rho^{(0)}(\pi), & \text{if } \pi \leq \pi_1^{(2)} \text{ or } \pi \geq \pi_2^{(2)}, \\ \bar{\rho}_2(\pi), & \text{otherwise.} \end{cases} \tag{8.2.20}$$

We define now recursively, for each π on $[0, 1]$,

$$\bar{\rho}_n(\pi) = c + E\{\rho^{(n-1)}(\psi(\pi, X))\}, \quad n \geq 1; \tag{8.2.21}$$

and

$$\rho^{(n)}(\pi) = \min\{\rho^{(0)}(\pi), \bar{\rho}_n(\pi)\}. \tag{8.2.22}$$

These functions constitute for each π monotone sequences $\bar{\rho}_n(\pi) \leq \bar{\rho}_{n-1}$ and $\rho^{(n)}(\pi) \leq \rho^{(n-1)}(\pi)$ for every $n \geq 1$. Moreover, for each n there exist $0 < \pi_1^{(n)} \leq \pi_1^{(n-1)} < \pi_2^{(n-1)} \leq \pi_2^{(n)} < 1$ such that

$$\rho^{(n)}(\pi) = \begin{cases} \rho^{(0)}(\pi), & \text{if } \pi \leq \pi_1^{(n)} \text{ or } \pi \geq \pi_2^{(n)}, \\ \bar{\rho}_n(\pi), & \text{otherwise.} \end{cases} \tag{8.2.23}$$

Let $\rho(\pi) = \lim_{n\to\infty} \rho^{(n)}(\pi)$ for each π in $[0, 1]$ and $\bar{\rho}(\pi) = E\{\rho(\psi(\pi, X))\}$. By the Lebesgue Monotone Convergence Theorem, we prove that $\bar{\rho}(\pi) = \lim_{n\to\infty} \bar{\rho}_n(\pi)$ for each $\pi \in [0, 1]$. The boundary points $\pi_1^{(n)}$ and $\pi_2^{(n)}$ converge to π_1 and π_2, respectively, where $0 < \pi_1 < \pi_2 < 1$. Consider now a nontruncated Bayes sequential procedure, with the stopping variable

$$N = \min\{n \geq 0 : \rho^{(0)}(\pi(\mathbf{X}_n)) = \rho(\pi(\mathbf{X}_n))\}, \tag{8.2.24}$$

where $X_0 \equiv 0$ and $\pi(X_0) \equiv \pi$. Since under H_0, $\pi(\mathbf{X}_n) \to 1$ with probability one and under H_1, $\pi(\mathbf{X}_n) \to 0$ with probability 1, the stopping variable (8.2.24) is finite with probability one.

It is generally very difficult to determine the exact Bayes risk function $\rho(\pi)$ and the exact boundary points π_1 and π_2. One can prove, however, that the Wald sequential probability ratio test (SPRT) (see Section 4.8.1) is a Bayes sequential procedure in the class of all stopping variables for which $N \geq 1$, corresponding to some prior probability π and cost parameter b. For a proof of this result, see Ghosh (1970, p. 93) or Zacks (1971, p. 456). A large sample approximation to the risk function $\rho(\pi)$ was given by Chernoff (1959). Chernoff has shown that in the SPRT given by the boundaries (A, B) if $A \to -\infty$ and $B \to \infty$, we have

$$\begin{aligned} A &\approx \log c - \log \frac{I(0, 1)b(1 - \pi)}{\pi}, \\ B &\approx \log \frac{1}{c} + \log \frac{I(1, 0)\pi}{1 - \pi}, \end{aligned} \tag{8.2.25}$$

where the cost of observations $c \to 0$ and $I(0, 1)$, $I(1, 0)$ are the Kullback–Leibler information numbers. Moreover, as $c \to 0$

$$\rho(\pi) \approx (-c \log c)\left(\frac{\pi}{I(0, 1)} + \frac{1 - \pi}{I(1, 0)}\right). \tag{8.2.26}$$

Shiryayev (1973, p. 127) derived an expression for the Bayes risk $\rho(\pi)$ associated with a continuous version of the Bayes sequential procedure related to a Wiener process. Reduction of the testing problem for the mean of a normal distribution to a free boundary problem related to the Wiener process was done also by Chernoff (1961, 1965, 1968); see also the book of Dynkin and Yushkevich (1969).

A simpler sequential stopping rule for testing two simple hypotheses is

$$N_\epsilon = \min\{n \geq 1 : \pi(\mathbf{X}_n) \leq \epsilon \text{ or } \pi(\mathbf{X}_n) \geq 1 - \epsilon\}. \tag{8.2.27}$$

If $\pi(\mathbf{X}_N) \leq \epsilon$ then H_0 is rejected, and if $\pi(\mathbf{X}_N) \geq 1 - \epsilon$ then H_0 is accepted. This stopping rule is equivalent to a Wald SPRT (A, B) with the limits

$$A = \frac{\epsilon\pi}{(1-\epsilon)(1-\pi)} \quad \text{and} \quad B = \frac{(1-\epsilon)\pi}{\epsilon(1-\pi)}.$$

If $\pi = \frac{1}{2}$ then, according to the results of Section 4.8.1, the average error probability is less than or equal to ϵ. This result can be extended to the problem of testing k simple hypotheses ($k \geq 2$), as shown in the following.

Let $H_1, \ldots, H_k$ be k hypotheses ($k \geq 2$) concerning the distribution of a random variable (vector) X. According to H_j, the p.d.f. of X is $f_j(x; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_j$, $j = 1, \ldots, k$. The parameter $\boldsymbol{\theta}$ is a nuisance parameter, whose parameter space Θ_j may depend on H_j. Let $G_j(\boldsymbol{\theta})$, $j = 1, \ldots, k$, be a prior distribution on Θ_j, and let π_j be the prior probability that H_j is the true hypothesis, $\sum_{j=1}^{k} \pi_j = 1$. Given n observations on $X_1, \ldots, X_n$, which are assumed to be conditionally i.i.d., we compute the **predictive likelihood** of H_j, namely,

$$L_j(\mathbf{X}_n) = \int_{\Theta_j} \prod_{i=1}^{n} f(X_i; \boldsymbol{\theta}) dG_j(\boldsymbol{\theta}), \tag{8.2.28}$$

$j = 1, \ldots, k$. Finally, the posterior probability of H_j, after n observations, is

$$\pi_j(\mathbf{X}_n) = \frac{\pi_j L_j(\mathbf{X}_n)}{\sum_{i=1}^{k} \pi_j L_i(\mathbf{X}_n)}, \quad j = 1, \ldots, k. \tag{8.2.29}$$

We consider the following Bayesian stopping variable, for some $0 < \epsilon < 1$.

$$N_\epsilon = \min\{n,\ n \geq 1 : \max_{1 \leq j \leq k} \pi_j(\mathbf{X}_n) \geq 1 - \epsilon\}. \tag{8.2.30}$$

Obviously, one considers small values of ϵ, $0 < \epsilon < 1/2$, and for such ϵ, there is a unique value j_ϵ^0 such that $\pi_{j_\epsilon^0}(\mathbf{X}_{N_\epsilon}) \geq 1 - \epsilon$. At stopping, hypothesis $H_{j_\epsilon^0}$ is accepted.

For each $n \geq 1$, partition the sample space $\mathcal{X}^{(n)}$ of $\mathbf{X}_n$ to $(k+1)$ disjoint sets

$$D_j^{(n)} = \{\mathbf{x}_n : \pi_j(\mathbf{x}_n) \geq 1 - \epsilon\}, \quad j = 1, \ldots, k$$

and $D_0^{(n)} = \mathcal{X}^n - \bigcup_{j=1}^{k} D_j^{(n)}$. As long as $\mathbf{x}_n \in D_0^{(n)}$ we continue sampling. Thus, $N_\epsilon = \min\left\{ n \geq 1 : \mathbf{x}_n \in \bigcup_{j=1}^{k} D_j^{(n)} \right\}$. In this sequential testing procedure, decision errors occur at stopping, when the wrong hypothesis is accepted. Thus, let δ_{ij} denote the predictive probability of accepting H_i when H_j is the correct hypothesis. That is,

$$\delta_{ij} = \sum_{n=1}^{\infty} \int_{\{\mathbf{x}_n \in D_i^{(n)}\}} L_j(\mathbf{x}_n) d\mu(\mathbf{x}_n). \tag{8.2.31}$$

Note that, for $\pi^* = 1 - \epsilon$, $\pi_j(\mathbf{x}_n) \geq \pi^*$ if, and only if,

$$\sum_{i \neq j} \pi_i L_i(\mathbf{x}_n) \leq \frac{1 - \pi^*}{\pi^*} \pi_j L_j(\mathbf{x}_n). \tag{8.2.32}$$

Let α_j denote the predictive error probability of rejecting H_j when it is true, i.e., $\alpha_j = \sum_{i \neq j} \delta_{ij}$.

The average predictive error probability is $\bar{\alpha}_\pi = \sum_{j=1}^{k} \pi_j \alpha_j$.

Theorem 8.2.1. *For the stopping variable N_ϵ, the average predictive error probability is $\bar{\alpha}_\pi \leq \epsilon$.*

Proof. From the inequality (8.2.32), we obtain

$$\begin{aligned} \delta_{ij} &\leq \sum_{n=1}^{\infty} \int_{D_i^{(n)}} \left[\frac{1 - \pi^*}{\pi^*} \cdot \frac{\pi_i}{\pi_j} L_i(\mathbf{x}_n) - \sum_{l \neq i \neq j} \frac{\pi_l}{\pi_j} L_l(\mathbf{x}_n) \right] d\mu(\mathbf{x}_n) \\ &= \frac{1 - \pi^*}{\pi^*} \cdot \frac{\pi_i}{\pi_j} (1 - \alpha_i) - \sum_{l \neq i \neq j} \frac{\pi_l}{\pi_j} \delta_{il}. \end{aligned} \tag{8.2.33}$$

Summing over i, we get

$$\alpha_j = \sum_{i \neq j} \delta_{ij} \leq \frac{1 - \pi^*}{\pi^*} \frac{1}{\pi_j} \sum_{i \neq j} \pi_i (1 - \alpha) - \frac{1}{\pi_j} \sum_{i \neq j} \sum_{l \neq i \neq j} \pi_l \delta_{il}$$

or

$$\pi_j \alpha_j \leq \frac{1-\pi^*}{\pi^*} \sum_{i \neq j} \pi_i (1-\alpha_i) - \sum_{i \neq j} \sum_{l \neq i \neq j} \pi_l \delta_{il}.$$

Summing over j, we obtain

$$\bar{\alpha}_\pi \leq \frac{1-\pi^*}{\pi^*} \sum_j \sum_{i \neq j} = \pi_i (1-\alpha_i) - \sum_j \sum_{i \neq j} \sum_{l \neq i \neq j} \pi_l \delta_{il}. \tag{8.2.34}$$

The first term on the RHS of (8.2.34) is

$$\begin{aligned} \frac{1-\pi^*}{\pi^*} \sum_j \sum_{i \neq j} \pi_i (1-\alpha_i) &= \frac{1-\pi^*}{\pi^*} \sum_j (1 - \bar{\alpha}_\pi - \pi_j (1-\alpha_j)) \\ &= \frac{1-\pi^*}{\pi^*} (k-1)(1-\bar{\alpha}_\pi). \end{aligned} \tag{8.2.35}$$

The second term on the RHS of (8.2.34) is

$$\begin{aligned} -\sum_j \sum_{i \neq j} \sum_{l \neq i \neq j} \pi_l \delta_{il} &= -\sum_j \sum_{l \neq j} \pi_l (\alpha_l - \delta_{jl}) \\ &= -\sum_j (\bar{\alpha}_\pi - \pi_j \alpha_j) + \sum_j \sum_{l \neq j} \pi_l \delta_{jl} \\ &= -(k-1)\bar{\alpha}_\pi + \bar{\alpha}_\pi = -(k-2)\bar{\alpha}_\pi. \end{aligned} \tag{8.2.36}$$

Substitution of (8.2.35) and (8.2.36) into (8.2.34) yields

$$\bar{\alpha}_\pi \leq \frac{1-\pi^*}{\pi^*} (1-\bar{\alpha}_\pi)$$

or

$$\bar{\alpha}_\pi \leq \epsilon.$$

QED

Thus, the Bayes sequential procedure given by the stopping variable N_ϵ and the associated decision rule can provide an excellent testing procedure when the number of hypothesis k is large. Rogatko and Zacks (1993) applied this procedure for testing the correct gene order. In this problem, if one wishes to order m gene loci on a chromosome, the number of hypotheses to test is $k = m!/2$.

8.3 BAYESIAN CREDIBILITY AND PREDICTION INTERVALS

8.3.1 Credibility Intervals

Let $\mathcal{F} = \{F(x;\theta); \theta \in \Theta\}$ be a parametric family of distribution functions. Let $H(\theta)$ be a specified prior distribution of θ and $H(\theta \mid \mathbf{X})$ be the corresponding posterior distribution, given $\mathbf{X}$. If θ is real then an interval $(\underline{\mathbf{L}}_\alpha(\mathbf{X}), \bar{L}_\alpha(\mathbf{X}))$ is called a **Bayes credibility interval** of level $1-\alpha$ if for all $\mathbf{X}$ (with probability 1)

$$P_H\{\underline{L}_\alpha(\mathbf{X}) \leq \theta \leq \bar{L}_\alpha(\mathbf{X}) \mid \mathbf{X}\} \geq 1-\alpha. \tag{8.3.1}$$

In multiparameter cases, we can speak of Bayes credibility regions. Bayes tolerance intervals are defined similarly.

Box and Tiao (1973) discuss Bayes intervals, called **highest posterior density** (HPD) intervals. These intervals are defined as θ intervals for which the posterior coverage probability is at least $(1-\alpha)$ and every θ-point within the interval has a posterior density not smaller than that of any θ-point outside the interval. More generally, a region $R_H(\mathbf{X})$ is called a $(1-\alpha)$ HPD region if

(i) $P_H(\theta \in R_H(\mathbf{X}) \mid \mathbf{X}] \geq 1-\alpha$, for all $\mathbf{X}$; and
(ii) $h(\theta \mid \mathbf{x}) \geq h(\phi \mid \mathbf{x})$, for every $\theta \in R_H(\mathbf{x})$ and $\phi \notin R_H(\mathbf{x})$.

The HPD intervals in cases of unimodal posterior distributions provide in nonsymmetric cases Bayes credibility intervals that are not equal tail ones. For various interesting examples, see Box and Tiao (1973).

8.3.2 Prediction Intervals

Suppose X is a random variable (vector) having a p.d.f. $f(x;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$. If $\boldsymbol{\theta}$ is known, an interval $I_\alpha(\boldsymbol{\theta})$ is called a prediction interval for X, at level $(1-\alpha)$ if

$$P_{\boldsymbol{\theta}}\{X \in I_\alpha(\boldsymbol{\theta})\} = 1-\alpha. \tag{8.3.2}$$

When $\boldsymbol{\theta}$ is unknown, one can use a Bayesian predictive distribution to determine an interval $I_\alpha(H)$ such that the predictive probability of $\{X \in I_\alpha(H)\}$ is at least $1-\alpha$. This predictive interval depends on the prior distribution $H(\boldsymbol{\theta})$. After observing $X_1, \ldots, X_n$, one can determine prediction interval (region) for $(X_{n+1}, \ldots, X_{n+m})$ by using the posterior distribution $H(\boldsymbol{\theta} \mid \mathbf{X}_n)$ for the predictive distribution $f_H(\mathbf{x} \mid \mathbf{x}_n) = \int_\Theta f(\mathbf{x};\boldsymbol{\theta}) dH(\boldsymbol{\theta} \mid \mathbf{x}_n)$. In Example 8.12, we illustrate such prediction intervals. For additional theory and examples, see Geisser (1993).

8.4 BAYESIAN ESTIMATION

8.4.1 General Discussion and Examples

When the objective is to provide a point estimate of the parameter θ or a function $\omega = g(\theta)$ we identify the action space with the parameter space. The decision function $d(\mathbf{X})$ is an estimator with domain χ and range Θ, or $\Omega = g(\Theta)$. For various loss functions the Bayes decision is an estimator $\hat{\theta}_H(\mathbf{X})$ that minimizes the posterior risk. In the following table, we present some loss functions and the corresponding Bayes estimators.

In the examples, we derived Bayesian estimators for several models of interest, and show the dependence of the resulting estimators on the loss function and on the prior distributions.

Loss Function	Bayes Estimator
$(\hat{\theta} - \theta)^2$	$\hat{\theta}(\mathbf{X}) = E_H\{\theta \mid \mathbf{X}\}$ (The posterior expectation)
$Q(\theta)(\hat{\theta}^2 - \theta)^2$	$E_H\{\theta Q(\theta) \mid \mathbf{X}\}/E_H\{Q(\theta) \mid \mathbf{X}\}$
$\lvert\hat{\theta} - \theta\rvert$	$\hat{\theta}(\mathbf{X})$ = median of the posterior distribution, i.e., $H^{-1}(.5 \mid \mathbf{X})$.
$a(\hat{\theta} - \theta)^- + b(\hat{\theta} - \theta)^+$	The $\frac{a}{a+b}$ quantile of $H(\theta \mid \mathbf{X})$; i.e., $H^{-1}(\frac{a}{a+b} \mid \mathbf{X})$.

8.4.2 Hierarchical Models

Lindley and Smith (1972) and Smith (1973a, b) advocated a somewhat more complicated methodology. They argue that the choice of a proper prior should be based on the notion of exchangeability. Random variables $W_1, W_2, \ldots, W_k$ are called **exchangeable** if the joint distribution of $(W_1, \ldots, W_k)$ is the same as that of $(W_{i_1}, \ldots, W_{i_k})$, where $(i_1, \ldots, i_k)$ is any permutation of $(1, 2, \ldots, k)$. The joint p.d.f. of exchangeable random variables can be represented as a mixture of appropriate p.d.f.s of i.i.d. random variables. More specifically, if, conditional on w, $W_1, \ldots, W_k$ are i.i.d. with p.d.f. $f(W_1, \ldots, W_k; w) = \prod_{i=1}^{k} g(W_i, w)$, and if w is given a probability distribution $P(w)$ then the p.d.f.

$$f^*(W_1, \ldots, W_r) = \int \prod_{i=1}^{k} g(W_i; w) dP(w) \tag{8.4.1}$$

represents a distribution of exchangeable random variables. If the vector $\mathbf{X}$ represents the means of k independent samples the present model coincides with the Model II

of ANOVA, with known variance components and an unknown grand mean μ. This model is a special case of a Bayesian linear model called by Lindley and Smith a three-stage linear model or **hierarchical models**. The general formulation of such a model is

$$X \sim N(A_1\boldsymbol{\theta}_1, V),$$
$$\boldsymbol{\theta}_1 \sim N(A_2\boldsymbol{\theta}_2, \Sigma),$$

and

$$\boldsymbol{\theta}_2 \sim N(A_3\boldsymbol{\theta}_3, C),$$

where $\mathbf{X}$ is an $n \times 1$ vector, $\boldsymbol{\theta}_i$ are $p_i \times 1$ $(i = 1, 2, 3)$, A_1, A_2, A_3 are known constant matrices, and V, Σ, C are **known** covariance matrices. Lindley and Smith (1972) have shown that for a noninformative prior for $\boldsymbol{\theta}_2$ obtained by letting $C^{-1} \to 0$, the Bayes estimator of $\boldsymbol{\theta}$, for the loss function $L(\hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}) = ||\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1||^2$, is given by

$$\hat{\boldsymbol{\theta}}_1 = B^{-1}A_1'\mathbf{X}, \tag{8.4.2}$$

where

$$B = A_1'V^{-1}A_1 + \Sigma^{-1} - \Sigma^{-1}A_2(A_2'\Sigma^{-1}A_2)^{-1}A_2'\Sigma^{-1}. \tag{8.4.3}$$

We see that this Bayes estimator coincides with the LSE, $(A'A)^{-1}A'X$, when $V = I$ and $\Sigma^{-1} \to 0$. This result depends very strongly on the knowledge of the covariance matrix V. Lindley and Smith (1972) suggested an iterative solution for a Bayesian analysis when V is unknown. Interesting special results for models of one way and two-way ANOVA can be found in Smith (1973b).

A comprehensive Bayesian analysis of the hierarchical Model II of ANOVA is given in Chapter 5 of Box and Tiao (1973).

In Gelman et al. (1995, pp. 129–134), we find an interesting example of a hierarchical model in which

$$J_i \mid \theta_i \sim B(n_i, \theta_i), \quad i = 1, \ldots, k.$$

$\theta_1, \ldots, \theta_k$ are conditionally i.i.d., with

$$\theta_i \mid \alpha, \beta \sim Beta(\alpha, \beta), \quad i = 1, \ldots, k$$

and (α, β) have an improper prior p.d.f.

$$h(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}.$$

According to this model, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is a vector of priorly exchangeable (not independent) parameters. We can easily show that the posterior joint p.d.f. of $\boldsymbol{\theta}$, given $\mathbf{J} = (J_1, \ldots, J_k)$ and (α, β) is

$$h(\boldsymbol{\theta} \mid \mathbf{J}, \alpha, \beta) = \prod_{j=1}^{k} \frac{1}{B(\alpha + J_j, \beta + n_j - J_j)} \theta_j^{\alpha + J_j - 1} (1 - \theta_j)^{\beta + n_j - J_j - 1}. \quad (8.4.4)$$

In addition, the posterior p.d.f. of (α, β) is

$$g(\alpha, \beta \mid \mathbf{J}) \propto g(\alpha, \beta) \prod_{j=1}^{k} \frac{B(\alpha + J_i, \beta + n_j - J_j)}{B(\alpha, \beta)}. \quad (8.4.5)$$

The objective is to obtain the joint posterior p.d.f.

$$\begin{aligned} h(\boldsymbol{\theta} \mid \mathbf{J}) &= \int_0^\infty \int_0^\infty h(\boldsymbol{\theta} \mid \mathbf{J}, \alpha, \beta) g(\alpha, \beta \mid \mathbf{J}) d_\alpha d_\beta \\ &= \int_0^\infty \int_0^\infty g(\alpha, \beta) \prod_{j=1}^{k} \frac{1}{B(\alpha, \beta)} \theta_j^{\alpha + J_j - 1} (1 - \theta_j)^{\beta + n_j - J_j - 1} d_\alpha d_\beta. \end{aligned}$$

From $h(\boldsymbol{\theta} \mid \mathbf{J})$ one can derive a credibility region for $\boldsymbol{\theta}$, etc.

8.4.3 The Normal Dynamic Linear Model

In time-series analysis for econometrics, signal processing in engineering and other areas of applications, one often encounters series of random vectors that are related according to the following linear dynamic model

$$\begin{aligned} \mathbf{Y}_n &= A\boldsymbol{\theta}_n + \boldsymbol{\epsilon}_n, \\ \boldsymbol{\theta}_n &= G\boldsymbol{\theta}_{n-1} + \boldsymbol{\omega}_n, \quad n \geq 1, \end{aligned} \quad (8.4.6)$$

where A and G are known matrices, which are (for simplicity) fixed. $\{\boldsymbol{\epsilon}_n\}$ is a sequence of i.i.d. random vectors; $\{\boldsymbol{\omega}_n\}$ is a sequence of i.i.d. random vectors; $\{\boldsymbol{\epsilon}_n\}$ and $\{\boldsymbol{\omega}_n\}$ are independent sequences, and

$$\begin{aligned} \boldsymbol{\epsilon}_n &\sim N(\mathbf{0}, V), \\ \boldsymbol{\omega}_n &\sim N(\mathbf{0}, \Omega). \end{aligned} \quad (8.4.7)$$

We further assume that $\boldsymbol{\theta}_0$ has a prior normal distribution, i.e.,

$$\boldsymbol{\theta}_0 \sim N(\eta_0, C_0), \quad (8.4.8)$$

and that $\boldsymbol{\theta}_0$ is independent of $\{\epsilon_t\}$ and $\{\omega_t\}$. This model is called the **normal random walk** model.

We compute now the posterior distribution of $\boldsymbol{\theta}_1$, given $\mathbf{Y}_1$. From multivariate normal theory, since

$$\mathbf{Y}_1 \mid \boldsymbol{\theta}_1 \sim N(A\boldsymbol{\theta}_1, V),$$
$$\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_0 \sim N(G\boldsymbol{\theta}_0, \Omega),$$

and

$$\boldsymbol{\theta}_0 \sim N(\boldsymbol{\eta}_0, C_0),$$

we obtain

$$\boldsymbol{\theta}_1 \sim N(G\boldsymbol{\eta}_0, \Omega + GC_0G').$$

Let $F_1 = \Omega + GC_0G'$. Then, we obtain after some manipulations

$$\boldsymbol{\theta}_1 \mid \mathbf{Y}_1 \sim N(\boldsymbol{\eta}_1, C_1),$$

where

$$\eta_1 = G\boldsymbol{\eta}_0 + F_1A'[V + AF_1A']^{-1}(\mathbf{Y}_1 - AG\boldsymbol{\eta}_0), \tag{8.4.9}$$

and

$$C_1 = F_1 - F_jA'[V + AF_1A']^{-1}AF_1. \tag{8.4.10}$$

Define, recursively for $j \geq 1$

$$F_j = \Omega + GC_{j-1}G',$$

and

$$\begin{aligned} C_j &= F_j - F_jA'[V + AF_jA']^{-1}AF_j \\ \eta_j &= G\boldsymbol{\eta}_{j-1} + F_jA'[V + AF_jA']^{-1}(\mathbf{Y}_j - AG\boldsymbol{\eta}_{j-1}). \end{aligned} \tag{8.4.11}$$

The recursive equations (8.4.11) are called the **Kalman filter**. Note that, for each $n \geq 1$, $\boldsymbol{\eta}_n$ depends on $\mathcal{D}_n = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$. Moreover, we can prove by induction on n, that

$$\boldsymbol{\theta}_n \mid \mathcal{D}_n \sim N(\boldsymbol{\eta}_n, C_n), \tag{8.4.12}$$

for all $n \geq 1$. For additional theory and applications in Bayesian forecasting and smoothing, see Harrison and Stevens (1976), West, Harrison, and Migon (1985), and the book of West and Harrison (1997). We illustrate this sequential Bayesian process in Example 8.19.

8.5 APPROXIMATION METHODS

In this section, we discuss two types of methods to approximate posterior distributions and posterior expectations. The first type is analytical, which is usually effective in large samples. The second type of approximation is numerical. The numerical approximations are based either on numerical integration or on simulations. Approximations are required when an exact functional form for the factor of proportionality in the posterior density is not available. We have seen such examples earlier, like the posterior p.d.f. (8.1.4).

8.5.1 Analytical Approximations

The analytic approximations are saddle-point approximations, based on variations of the **Laplace method**, which is explained now.

Consider the problem of evaluating the integral

$$I = \int \cdots \int f(\boldsymbol{\theta}) \exp\{-nk(\boldsymbol{\theta})\} d\boldsymbol{\theta}, \tag{8.5.1}$$

where $\boldsymbol{\theta}$ is m-dimensional, and $k(\boldsymbol{\theta})$ has sufficiently high-order continuous partial derivatives. Consider first the case of $m = 1$. Let $\hat{\theta}$ be an argument maximizing $-k(\theta)$. Make a Taylor expansion of $k(\theta)$ around $\hat{\theta}$, i.e.,

$$k(\theta) = k(\hat{\theta}) + (\theta - \hat{\theta})k'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 k''(\hat{\theta}) + o(\theta - \hat{\theta})^2, \quad \text{as } \theta \to \hat{\theta}. \tag{8.5.2}$$

$k'(\hat{\theta}) = 0$ and $k''(\hat{\theta}) > 0$. Thus, substituting (8.5.2) in (8.5.1), the integral I is approximated by

$$\begin{aligned} I &\cong \int f(\theta) \exp\left\{-nk(\hat{\theta}) - \frac{n}{2}k''(\hat{\theta})(\theta - \hat{\theta})^2\right\} d\theta \\ &= \exp\{-nk(\hat{\theta})\}\sqrt{\frac{2\pi}{nk''(\hat{\theta})}} E_N\{f(\theta)\}, \end{aligned} \tag{8.5.3}$$

where $E_N\{f(\theta)\}$ is the expected value of $f(\theta)$, with respect to the normal distribution with mean $\hat{\theta}$ and variance $\hat{\sigma}_n^2 = \dfrac{1}{nk''(\hat{\theta})}$. The expectation $E_N\{f(\theta)\}$ can

be sometimes computed exactly, or one can apply the delta method to obtain the approximation

$$E_N\{f(\theta)\} = f(\hat{\theta}) + \frac{1}{2n}\frac{f''(\hat{\theta})}{k''(\hat{\theta})} + O(n^{-3/2}). \tag{8.5.4}$$

Often we see the simpler approximation, in which $f(\hat{\theta})$ is used for $E_N\{f(\hat{\theta})\}$. In this case, the approximation error is $O(n^{-1})$. If we use $f(\hat{\theta})$ for $E_N\{f(\theta)\}$, we obtain the approximation

$$I \cong \exp\{-nk(\hat{\theta})\} f(\hat{\theta}) \sqrt{\frac{2\pi}{nk''(\theta)}}. \tag{8.5.5}$$

In the $m > 1$ case, the approximating formula becomes

$$\hat{I} = \left(\frac{2\pi}{n}\right)^{m/2} |\hat{\Sigma}(\hat{\boldsymbol{\theta}})|^{1/2} f(\hat{\boldsymbol{\theta}}) \exp\{-nk(\hat{\boldsymbol{\theta}})\}, \tag{8.5.6}$$

where

$$\hat{\Sigma}^{-1}(\hat{\boldsymbol{\theta}}) = \left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} k(\boldsymbol{\theta}), i, j = 1, \ldots, m\right)\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{8.5.7}$$

These approximating formulae can be applied in Bayesian analysis, by letting $-nk(\boldsymbol{\theta})$ be the log-likelihood function, $l(\boldsymbol{\theta}^* \mid \mathbf{X}_n)$; $\hat{\boldsymbol{\theta}}$ be the MLE, $\hat{\boldsymbol{\theta}}_n$, and $\hat{\Sigma}^{-1}(\hat{\boldsymbol{\theta}})$ be $\mathbf{J}(\hat{\boldsymbol{\theta}}_n)$ given in (7.7.15). Accordingly, the posterior p.d.f., when the prior p.d.f. is $h(\boldsymbol{\theta})$, is approximated by

$$h(\boldsymbol{\theta} \mid \mathbf{X}_n) \cong C_n(\mathbf{X}_n) h(\boldsymbol{\theta}) \exp\left\{-\frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)' J(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)\right\}. \tag{8.5.8}$$

In this formula, $\hat{\theta}_n$ is the MLE of $\boldsymbol{\theta}$ and

$$\begin{aligned} C_n(\mathbf{X}_n) &= \left[\int \cdots \int h(\boldsymbol{\theta}) \exp\left\{-\frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)' \mathbf{J}(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)\right\} d\boldsymbol{\theta}\right]^{-1} \\ &= [(2\pi)^{\frac{m}{2}} \mid \mathbf{J}(\hat{\boldsymbol{\theta}}_n)|^{-1/2} E_N\{h(\boldsymbol{\theta})\}]^{-1}. \end{aligned} \tag{8.5.9}$$

If we approximate $E_N\{h(\boldsymbol{\theta})\}$ by $h(\hat{\boldsymbol{\theta}}_n)$, then the approximating formula reduces to

$$h(\theta \mid \mathbf{X}_n) \cong \frac{n^{m/2}}{(2\pi)^{m/2}} |J(\hat{\boldsymbol{\theta}}_n)|^{1/2} \exp\left\{-\frac{n}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)' \mathbf{J}(\hat{\boldsymbol{\theta}}_n) \cdot (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)\right\}. \tag{8.5.10}$$

This is a large sample normal approximation to the posterior density of $\boldsymbol{\theta}$. We can write this, for large samples, as

$$\boldsymbol{\theta} \mid \mathbf{X}_n \approx N\left(\hat{\boldsymbol{\theta}}_n, \frac{1}{n}(\mathbf{J}_n(\hat{\theta}_n))^{-1}\right). \tag{8.5.11}$$

Note that Equation (8.5.11) does not depend on the prior distribution, and is not expected therefore to yield good approximation to $h(\boldsymbol{\theta} \mid \mathbf{X}_n)$ if the samples are not very large.

One can improve upon the normal approximation (8.5.11) by combining the likelihood function and the prior density $h(\boldsymbol{\theta})$ in the definition of $k(\boldsymbol{\theta})$. Thus, let

$$\tilde{k}(\boldsymbol{\theta}) = -\frac{1}{n}(l(\boldsymbol{\theta} \mid \mathbf{X}_n) + \log h(\boldsymbol{\theta})). \tag{8.5.12}$$

Let $\tilde{\boldsymbol{\theta}}$ be a value of $\boldsymbol{\theta}$ maximizing $-n\tilde{k}(\boldsymbol{\theta})$, or $\tilde{\boldsymbol{\theta}}_n$ the root of

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta} \mid \mathbf{X}_n) + \frac{1}{h(\boldsymbol{\theta})} \nabla_{\theta} h(\boldsymbol{\theta}) = \mathbf{0}. \tag{8.5.13}$$

Let

$$\tilde{\mathbf{J}}(\boldsymbol{\theta}) = \mathbf{J}(\boldsymbol{\theta}) - \frac{1}{n}\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log h(\boldsymbol{\theta})\right). \tag{8.5.14}$$

Then, the saddle-point approximation to the posterior p.d.f. $h(\boldsymbol{\theta} \mid \mathbf{X}_n)$ is

$$h^*(\boldsymbol{\theta} \mid \mathbf{X}_n) = \left(\frac{n}{2\pi}\right)^{m/2} |\tilde{\mathbf{J}}(\tilde{\boldsymbol{\theta}}_n)|^{1/2} \cdot \frac{h(\boldsymbol{\theta})L(\boldsymbol{\theta} \mid \mathbf{X}_n)}{h(\tilde{\boldsymbol{\theta}}_n)L(\tilde{\boldsymbol{\theta}}_n \mid \mathbf{X}_n)}. \tag{8.5.15}$$

This formula is similar to Barndorff-Nielsen p^*-formula (7.7.15) and reduces to the p^*-formula if $h(\boldsymbol{\theta})d\boldsymbol{\theta} \propto d\boldsymbol{\theta}$. The normal approximation is given by (8.5.11), in which $\hat{\boldsymbol{\theta}}_n$ is replaced by $\tilde{\boldsymbol{\theta}}_n$ and $\mathbf{J}(\hat{\boldsymbol{\theta}}_n)$ is replaced by $\tilde{\mathbf{J}}(\tilde{\boldsymbol{\theta}}_n)$.

For additional reading on analytic approximation for large samples, see Gamerman (1997, Ch. 3), Reid (1995, pp. 351–368), and Tierney and Kadane (1986).

8.5.2 Numerical Approximations

In this section, we discuss two types of numerical approximations: numerical integrations and simulations. The reader is referred to Evans and Swartz (2001).

I. Numerical Integrations

We have seen in the previous sections that, in order to evaluate posterior p.d.f., one has to evaluate integrals of the form

$$I = \int_{-\infty}^{\infty} L(\theta \mid \mathbf{X}_n) h(\theta) d\theta. \tag{8.5.16}$$

Sometimes these integrals are quite complicated, like that of the RHS of Equation (8.1.4).

Suppose that, as in (8.5.16), the range of integration is from $-\infty$ to ∞ and $I < \infty$. Consider first the case where θ is real. Making the one-to-one transformation $\omega = e^{\theta}/(1 + e^{\theta})$, the integral of (8.5.16) is reduced to

$$I = \int_0^1 q\left(\log \frac{\omega}{1-\omega}\right) \frac{1}{\omega(1-\omega)} d\omega, \tag{8.5.17}$$

where $q(\theta) = L(\theta \mid \mathbf{X}_n) h(\theta)$. There are many different methods of numerical integration. A summary of various methods and their accuracy is given in Abramowitz and Stegun (1968, p. 885). The reader is referred also to the book of Davis and Rabinowitz (1984).

If we define $f(\omega)$ so that

$$f(\omega) = q\left(\log \frac{\omega}{1-\omega}\right) \frac{1}{\omega^{3/2}(1-\omega)^{1/2}} \tag{8.5.18}$$

then, an n-point approximation to I is given by

$$\hat{I}_n = \sum_{i=1}^{n} p_i f(\omega_i), \tag{8.5.19}$$

where

$$\begin{aligned} \omega_i &= \cos^2\left(\frac{\pi}{2} \cdot \frac{2i-1}{2n+1}\right), \quad i = 1, \ldots, n \\ p_i &= \frac{2\pi}{2n+1} \omega_i, \quad i = 1, \ldots, n. \end{aligned} \tag{8.5.20}$$

The error in this approximation is

$$R_n = \frac{\pi}{(2n)! 2^{4n+1}} f^{(2n)}(\xi), \quad 0 < \xi < 1. \tag{8.5.21}$$

Integrals of the form

$$I = \int_{-1}^{1} q(u)du = \int_{0}^{1} q(u)du + \int_{0}^{1} q(-u)du. \tag{8.5.22}$$

Thus, (8.5.22) can be computed according to (8.5.19). Another method is to use an n-points Gaussian quadrature formula:

$$I_n^* = \sum_{i=1}^{n} \omega_i q(u_i), \tag{8.5.23}$$

where u_i and w_i are tabulated in Table 25.4 of Abramowitz and Stegun (1968, p. 916). Often it suffices to use $n = 8$ or $n = 12$ points in (8.5.23).

II. Simulation

The basic theorem applied in simulations to compute an integral $I = \int f(\theta)dH(\theta)$ is the strong law of large numbers (SLLN). We have seen in Chapter 1 that if $X_1, X_2, \ldots$ is a sequence of i.i.d. random variables having a distribution $F_X(x)$, and if $\int_{-\infty}^{\infty} |g(x)|dF(x) < \infty$ then

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow[n\to\infty]{a.s.} \int_{-\infty}^{\infty} g(x)dF(x).$$

This important result is applied to approximate an integral $\int_{-\infty}^{\infty} f(\theta)dH(\theta)$ by a sequence $\theta_1, \theta_2, \ldots$ of i.i.d. random variables, generated from the prior distribution $H(\theta)$. Thus, for large n,

$$\int_{-\infty}^{\infty} f(\theta)dH(\theta) \cong \frac{1}{n}\sum_{i=1}^{n} f(\theta_i). \tag{8.5.24}$$

Computer programs are available in all statistical packages that simulate realizations of a sequence of i.i.d. random variables, having specified distributions. All programs use **linear congruential generators** to generate "pseudo" random numbers that have approximately uniform distribution on (0, 1). For discussion of these generators, see Bratley, Fox, and Schrage (1983).

Having generated i.i.d. uniform $R(0, 1)$ random variables $U_1, U_2, \ldots, U_n$, one can obtain a simulation of i.i.d. random variables having a specific c.d.f. F, by the transformation

$$X_i = F^{-1}(U_i), \quad i = 1, \ldots, n. \tag{8.5.25}$$

In some special cases, one can use different transformations. For example, if U_1, U_2 are independent $R(0, 1)$ random variables then the **Box-Muller transformation**

$$\begin{aligned} X_1 &= (-2 \log U_1)^{1/2} \cos(2\pi U_2), \\ X_2 &= (-2 \log U_1)^{1/2} \sin(2\pi U_2), \end{aligned} \tag{8.5.26}$$

yields two independent random variables having a standard normal distribution. It is easier to simulate a $N(0, 1)$ random variable according to (8.5.26) than according to $X = \Phi^{-1}(U)$. In today's technology, one could choose from a rich menu of simulation procedures for many of the common distributions.

If a prior distribution $H(\boldsymbol{\theta})$ is not in a simulation menu, or if $h(\theta)d\theta$ is not proper, one can approximate $\int_{-\infty}^{\infty} f(\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta}$ by generating $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ from another convenient distribution, $\lambda(\boldsymbol{\theta})d\boldsymbol{\theta}$ say, and using the formula

$$\int_{-\infty}^{\infty} f(\boldsymbol{\theta})h(\boldsymbol{\theta})d\boldsymbol{\theta} \cong \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{\theta}_i) \frac{h(\boldsymbol{\theta}_i)}{\lambda(\boldsymbol{\theta}_i)}. \tag{8.5.27}$$

The method of simulating from a substitute p.d.f. $\lambda(\boldsymbol{\theta})$ is called **importance sampling**, and $\lambda(\boldsymbol{\theta})$ is called an **importance density**. The choice of $\lambda(\boldsymbol{\theta})$ should follow the following guidelines:

(i) The support of $\lambda(\boldsymbol{\theta})$ should be the same as that of $h(\boldsymbol{\theta})$;

(ii) $\lambda(\boldsymbol{\theta})$ should be similar in shape, as much as possible, to $h(\boldsymbol{\theta})$; i.e., $\lambda(\boldsymbol{\theta})$ should have the same means, standard deviations and other features, as those of $h(\boldsymbol{\theta})$.

The second guideline is sometimes complicated. For example, if $h(\theta)d(\theta)$ is the improper prior $d\theta$ and $I = \int_{-\infty}^{\infty} f(\theta)d\theta$, where $\int_{-\infty}^{\infty} |f(\theta)|d\theta < \infty$, one could use first the monotone transformation $x = e^{\theta}/(1 + e^{\theta})$ to reduce I to $I = \int_0^1 f\left(\log \frac{x}{1-x}\right) \frac{dx}{x(1-x)}$. One can use then a beta, $\beta(p, q)$, importance density to simulate from, and approximate I by

$$\tilde{I} = \frac{1}{n} \sum_{i=1}^{n} f\left(\log \frac{X_i}{1 - X_i}\right) \frac{B(p, q)}{X_i^p (1 - X_i)^q}.$$

It would be simpler to use $\beta(1, 1)$, which is the uniform $R(0, 1)$.

An important question is, how large should the simulation sample be, so that the approximation will be sufficiently precise. For large values of n, the approximation

$\hat{I}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f(\boldsymbol{\theta}_i) h(\boldsymbol{\theta}_i)$ is, by Central Limit Theorem, approximately distributed like $N\left(I, \frac{\tau^2}{n}\right)$, where

$$\tau^2 = V_S\{f(\boldsymbol{\theta}) h(\boldsymbol{\theta})\}.$$

$V_S(\cdot)$ is the variance according to the simulation density. Thus, n could be chosen sufficiently large, so that $Z_{1-\alpha/2} \cdot \frac{\tau}{\sqrt{n}} < \delta$. This will guarantee that with confidence probability close to $(1-\alpha)$ the true value of I is within $\hat{I} \pm \delta$. The problem, however, is that generally τ^2 is not simple or is unknown. To overcome this problem one could use a sequential sampling procedure, which attains asymptotically the fixed width confidence interval. Such a procedure was discussed in Section 6.7.

We should remark in this connection that simulation results are less accurate than those of numerical integration. **One should use, as far as possible, numerical integration rather than simulation**.

To illustrate this point, suppose that we wish to compute numerically

$$I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}x^2\right\} dx = 1.$$

Reduce I, as in (8.5.17), to

$$I = \frac{1}{\sqrt{2\pi}} \int_0^1 \exp\left\{-\frac{1}{2}\left(\log \frac{u}{1-u}\right)^2\right\} \frac{du}{u(1-u)}.$$

Simulation of $N = 10,000$ random variables $U_i \sim R(0,1)$ yields the approximation

$$\begin{aligned}\hat{I}_{10,000} &= \frac{1}{10,000\sqrt{2\pi}} \sum_{i=1}^{10,000} \exp\left\{-\frac{1}{2}\left(\log \frac{U_i}{1-u_i}\right)^2\right\} \frac{1}{u_i(1-u_i)}. \\ &= 1.001595.\end{aligned}$$

On the other hand, a 10-point numerical integration, according to (8.5.29), yields

$$\hat{I}_{10} = 1.$$

When $\boldsymbol{\theta}$ is m-dimensional, $m \geq 2$, numerical integration might become too difficult. In such cases, simulations might be the answer.

8.6 EMPIRICAL BAYES ESTIMATORS

Empirical Bayes estimators were introduced by Robbins (1956) for cases of repetitive estimation under similar conditions, when Bayes estimators are desired but the statistician does not wish to make specific assumptions about the prior distribution. The following example illustrates this approach. Suppose that X has a Poisson distribution $P(\lambda)$, and λ has some prior distribution $H(\lambda)$, $0 < \lambda < \infty$. The Bayes estimator of λ for the squared-error loss function is

$$E\{\lambda \mid X\} = \frac{\int_0^\infty \lambda p(X;\lambda) dH(\lambda)}{\int_0^\infty p(X;\lambda) dH(\lambda)},$$

where $p(x;\lambda)$ denotes the p.d.f. of $P(\lambda)$ at the point x. Since $\lambda p(x;\lambda) = (x+1) \cdot p(x+1;\lambda)$ for every λ and each $x = 0, 1, \ldots$ we can express the above Bayes estimator in the form

$$\begin{aligned} E_H\{\lambda \mid X\} &= (X+1)\frac{\int_0^\infty p(X+1;\lambda) dH(\lambda)}{\int_0^\infty p(X;\lambda) dH(\lambda)} \\ &= (X+1) p_H(X+1)/p_H(X), \end{aligned} \tag{8.6.1}$$

where $p_H(x)$ is the predictive p.d.f. at x. Obviously, in order to determine the posterior expectation we have to know the prior distribution $H(\lambda)$. On the other hand, if the problem is repetitive in the sense that a sequence $(X_1, \lambda_1), (X_2, \lambda_2), \ldots, (X_n, \lambda_n), \ldots$, is generated independently so that $\lambda_1, \lambda_2, \ldots$ are i.i.d. having the same prior distribution $H(\lambda)$, and $X_1, \ldots, X_n$ are conditionally independent, given $\lambda_1, \ldots, \lambda_n$, then we consider the sequence of observable random variables $X_1, \ldots, X_n, \ldots$ as i.i.d. from the mixture of Poisson distribution with p.d.f. $p_H(j)$, $j = 0, 1, 2, \ldots$. Thus, if on the nth epoch, we observe $X_n = i^0$ we estimate, on the basis of all the data, the value of $p_H(i^0+1)/p_H(i^0)$. A consistent estimator of $p_H(j)$, for any $j = 0, 1, \ldots$ is $\frac{1}{n}\sum_{i=1}^n I\{X_i = j\}$, where $I\{X_i = j\}$ is the indicator function of $\{X_i = j\}$. This follows from the SLLN. Thus, a consistent estimator of the Bayes estimator $E_H\{\lambda \mid X_n\}$ is

$$\hat{\lambda}_n = (X_n+1)\frac{\sum_{i=1}^n I\{X_i = X_n + 1\}}{\sum_{i=1}^n I\{X_i = X_n\}}. \tag{8.6.2}$$

This estimator is independent of the unknown $H(\lambda)$, and for large values of n is approximately equal to $E_H\{\lambda \mid X_n\}$. The estimator $\hat{\lambda}_n$ is called an **empirical Bayes** estimator. The question is whether the prior risks, under the true $H(\lambda)$, of the estimators λ_n converge, as $n \to \infty$, to the Bayes risk under $H(\lambda)$. A general discussion of this issue with sufficient conditions for such convergence of the associated prior risks is given in the paper of Robbins (1964).

Many papers were written on the application of the empirical Bayes estimation method to repetitive estimation problems in which it is difficult or impossible to specify the prior distribution exactly. We have to remark in this connection that the empirical Bayes estimators are only asymptotically optimal. We have an adaptive decision process which corrects itself and approaches the optimal decisions only when n grows. How fast does it approach the optimal decisions? It depends on the amount of a priori knowledge of the true prior distribution. The initial estimators may be far from the true Bayes estimators. A few studies have been conducted to estimate the rate of approach of the prior risks associated with the empirical Bayes decisions to the true Bayes risk. Lin (1974) considered the one parameter exponential family and the estimation of a function $\lambda(\theta)$ under squared-error loss. The true Bayes estimator is

$$\hat{\lambda}(x) = \int \lambda(\theta) f(x;\theta) dH(\theta) / \int f(x;\theta) dH(\theta),$$

and it is assumed that $\hat{\lambda}(x) f_H(x)$ can be expressed in the form $\sum_{i=0}^{m} \omega_t(x) f_H^{(i)}(x)$, where $f_H^{(i)}(x)$ is the ith order derivative of $f_H(x)$ with respect to x. The empirical Bayes estimators considered are based on consistent estimators of the p.d.f. $f_H(x)$ and its derivatives. For the particular estimators suggested it is shown that the rate of approach is of the order $0(n^{-\alpha})$ with $0 < \alpha \leq 1/3$, where n is the number of observations.

In Example 8.26, we show that if the form of the prior is known, the rate of approach becomes considerably faster. When the form of the prior distribution is known the estimators are called semi-empirical Bayes, or **parametric empirical Bayes**.

For further reading on the empirical Bayes method, see the book of Maritz (1970) and the papers of Casella (1985), Efron and Morris (1971, l972a, 1972b), and Susarla (1982).

The E-M algorithm discussed in Example 8.27 is a very important procedure for estimation and overcoming problems of missing values. The book by McLachlan and Krishnan (1997) provides the theory and many interesting examples.

PART II: EXAMPLES

Example 8.1. The experiment under consideration is to produce concrete under certain conditions of mixing the ingredients, temperature of the air, humidity, etc. Prior experience shows that concrete cubes manufactured in that manner will have a

compressive strength X after 3 days of hardening, which has a log-normal distribution $LN(\mu, \sigma^2)$. Furthermore, it is expected that 95% of such concrete cubes will have compressive strength in the range of 216–264 (kg/cm^2).

According to our model, $Y = \log X \sim N(\mu, \sigma^2)$. Taking the (natural) logarithms of the range limits, we expect most Y values to be within the interval (5.375,5.580).

The conditional distribution of Y given (μ, σ^2) is

$$Y \mid \mu, \sigma^2 \sim N(\mu, \sigma^2).$$

Suppose that σ^2 is fixed at $\sigma^2 = 0.001$, and μ has a prior normal distribution $\mu \sim N(\mu_0, \tau^2)$, then the predictive distribution of Y is $N(\mu_0, \sigma^2 + \tau^2)$. Substituting $\mu_0 = 5.475$, the predictive probability that $Y \in (5.375, 5.580)$, if $\sqrt{\sigma^2 + \tau^2} = 0.051$ is 0.95. Thus, we choose $\tau^2 = 0.0015$ for the prior distribution of μ.

From this model of $Y \mid \mu, \sigma^2 \sim N(\mu, \sigma^2)$ and $\mu \sim N(\mu_0, \tau^2)$. The bivariate distribution of (Y, μ) is

$$\begin{pmatrix} Y \\ \mu \end{pmatrix} \sim N\left(\begin{bmatrix} \mu_0 \\ \mu_0 \end{bmatrix}, \begin{bmatrix} \sigma^2 + \tau^2 & \tau^2 \\ \tau^2 & \tau^2 \end{bmatrix} \right).$$

Hence, the conditional distribution of μ given $\{Y = y\}$ is, as shown in Section 2.9,

$$\mu \mid Y = y \sim N\left(\mu_0 + \frac{\tau^2}{\sigma^2 + \tau^2}(y - \mu_0), \tau^2 \frac{\sigma^2}{\sigma^2 + \tau^2} \right) \sim N(219 + 0.6y, 0.0006).$$

The posterior distribution of μ, given $\{Y = y\}$ is normal. ■

Example 8.2. (a) $X_1, X_2, \ldots, X_n$ given λ are conditionally i.i.d., having a Poisson distribution $P(\lambda)$, i.e., $\mathcal{F} = \{P(\lambda), 0 < \lambda < \infty\}$.

Let $\mathcal{H} = \{G(\Lambda, \alpha), 0 < \alpha, \Lambda < \infty\}$, i.e., $\mathcal{H}$ is a family of prior gamma distributions for λ. The minimal sufficient statistics, given λ, is $T_n = \sum_{i=1}^{n} X_i$. $T_n \mid \lambda \sim P(\lambda n)$. Thus, the posterior p.d.f. of λ, given T_n, is

$$\begin{aligned} h(\lambda \mid T_n) &\propto \lambda^{T_n} e^{-n\lambda} \cdot \lambda^{\alpha - 1} e^{-\lambda \Lambda} \\ &= \lambda^{T_n + \alpha - 1} e^{-\lambda(n + \Lambda)}. \end{aligned}$$

Hence, $\lambda \mid T_n \sim G(n + \Lambda, T_n + \alpha)$. The posterior distribution belongs to $\mathcal{H}$.

(b) $\mathcal{F} = \{G(\lambda, \alpha), 0 < \lambda < \infty\}$, α fixed. $\mathcal{H} = \{G(\Lambda, \nu), 0 < \nu, \Lambda < \infty\}$.

$$\begin{aligned} h(\lambda \mid X) &\propto \lambda^{\alpha} e^{-\lambda X} \cdot \lambda^{\nu - 1} e^{-\lambda \Lambda} \\ &= \lambda^{\nu + \alpha - 1} e^{-\lambda(X + \Lambda)} \end{aligned}$$

Thus, $\lambda \mid X \sim G(X + \Lambda, \nu + \alpha)$. ■

Example 8.3. The following problem is often encountered in high technology industry.

The number of soldering points on a typical printed circut board (PCB) is often very large. There is an automated soldering technology, called "wave soldering," which involves a large number of different factors (conditions) represented by variables $X_1, X_2, \cdots, X_k$. Let J denote the number of faults in the soldering points on a PCB. One can model J as having conditional Poisson distribution with mean λ, which depends on the manufacturing conditions $X_1, \ldots, X_k$ according to a log-linear relationship

$$\log \lambda = \beta_0 + \sum_{i=1}^{k} \beta_i x_i = \boldsymbol{\beta}' \mathbf{x},$$

where $\boldsymbol{\beta}' = (\beta_0, \ldots, \beta_k)$ and $\mathbf{x} = (1, x_1, \ldots, x_k)$. $\boldsymbol{\beta}$ is generally an unknown parametric vector. In order to estimate $\boldsymbol{\beta}$, one can design an experiment in which the values of the control variables $X_1, \ldots, X_k$ are changed.

Let J_i be the number of observed faulty soldering points on a PCB, under control conditions given by $\mathbf{x}_i$ $(i = 1, \ldots, N)$. The likelihood function of $\boldsymbol{\beta}$, given $J_1, \ldots, J_N$ and $\mathbf{x}_1, \ldots, \mathbf{x}_N$, is

$$\begin{aligned} L(\beta \mid J_1, \ldots, J_N, \mathbf{x}_1, \ldots, \mathbf{x}_N) &= \prod_{i=1}^{N} \exp\{J_i \mathbf{x}_i' \boldsymbol{\beta} - e^{\boldsymbol{\beta}' \mathbf{x}_i}\} \\ &= \exp\{\mathbf{w}_N' \boldsymbol{\beta} - \sum_{i=1}^{N} e^{\mathbf{x}_i' \boldsymbol{\beta}}\}, \end{aligned}$$

where $\mathbf{w}_N = \sum_{i=1}^{N} J_i \mathbf{x}_i$. If we ascribe $\boldsymbol{\beta}$ a prior multinormal distribution, i.e., $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, V)$ then the posterior p.d.f. of $\boldsymbol{\beta}$, given $\mathcal{D}_N = (J_1, \ldots, J_N, \mathbf{x}_1, \ldots, \mathbf{x}_N)$, is

$$h(\boldsymbol{\beta} \mid \mathcal{D}_N) \propto \exp\left\{ \mathbf{w}_N' \boldsymbol{\beta} - \sum_{i=1}^{N} e^{\mathbf{x}_i' \boldsymbol{\beta}} - \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' V^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right\}.$$

It is very difficult to express analytically the proportionality factor, even in special cases, to make the RHS of $h(\boldsymbol{\beta} \mid \mathcal{D}_N)$ a p.d.f. ■

Example 8.4. In this example, we derive the Jeffreys prior density for several models.

A. $\mathcal{F} = \{b(x; n, \theta), 0 < \theta < 1\}$.

This is the family of binomial probability distributions. The Fisher information function is

$$I(\theta) = \frac{1}{\theta(1-\theta)}, \quad 0 < \theta < 1.$$

Thus, the Jeffreys prior for θ is

$$h(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}, \quad 0 < \theta < 1.$$

In this case, the prior density is

$$h(\theta) = \frac{1}{\pi}\theta^{-1/2}(1-\theta)^{-1/2}, \quad 0 < \theta < 1.$$

This is a proper prior density. The posterior distribution of θ, given X, under the above prior is $\text{Beta}\left(X + \frac{1}{2}, n - X + \frac{1}{2}\right)$.

B. $\mathcal{F} = \{N(\mu, \sigma^2); -\infty < \mu < \infty, 0 < \sigma < \infty\}$.

The Fisher information matrix is given in (3.8.8). The determinant of this matrix is $|I(\mu, \sigma^2)| = 1/2\sigma^6$. Thus, the Jeffreys prior for this model is

$$h(\mu, \sigma^2) \propto d\mu \frac{1}{(\sigma^2)^{3/2}}.$$

Using this improper prior density the posterior p.d.f. of (μ, σ^2), given $X_1, \ldots, X_n$, is

$$h(\mu, \sigma^2 \mid \bar{X}_n, Q) = \frac{\sqrt{n}\, Q^{\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(n(\mu - \bar{X}_n)^2 + Q)}}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}\right) 2^{\frac{n+1}{2}} (\sigma^2)^{\frac{n+3}{2}}},$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, $Q = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. The parameter $\phi = \frac{1}{\sigma^2}$ is called the **precision parameter**. In terms of μ and ϕ, the improper prior density is

$$h(\mu, \phi) \propto \phi^{-1/2}.$$

The posterior density of (μ, ϕ) correspondingly is

$$h(\mu, \phi \mid \bar{X}_n, Q) = \frac{\sqrt{n}\, Q^{\frac{n}{2}} e^{-\frac{\phi}{2}[n(\mu - \bar{X}_n)^2 + Q]} \phi^{\frac{n-1}{2}}}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}\right) 2^{\frac{n+1}{2}}}.$$

■

Example 8.5. Consider a simple inventory system in which a certain commodity is stocked at the beginning of every day, according to a policy determined by the following considerations. The daily demand (in number of units) is a random variable X whose distribution belongs to a specified parametric family $\mathcal{F}$. Let $X_1, X_2, \ldots$ denote a sequence of i.i.d. random variables, whose common distribution $F(x;\theta)$ belongs to $\mathcal{F}$ and which represent the observed demand on consecutive days. The stock level at the beginning of each day, $S_n, n = 1, 2, \ldots$ can be adjusted by increasing or decreasing the available stock at the end of the previous day. We consider the following inventory cost function

$$K(s, x) = c(s - x)^+ - h(s - x)^-,$$

where c, $0 < c < \infty$, is the daily cost of holding a unit in stock and h, $0 < h < \infty$ is the cost (or penalty) for a shortage of one unit. Here $(s - x)^+ = \max(0, s - x)$ and $(s - x)^- = -\min(0, s - x)$. If the distribution of X, $F(x;\theta)$ is known, then the expected cost $R(S, \theta) = E_\theta\{K(S, X)\}$ is minimized by

$$S^0(\theta) = F^{-1}\left(\frac{h}{c+h};\theta\right),$$

where $F^{-1}(\gamma;\theta)$ is the γ-quantile of $F(x;\theta)$. If θ is unknown we cannot determine $S^0(\theta)$. We show now a Bayesian approach to the determination of the stock levels. Let $H(\theta)$ be a specific prior distribution of θ. The prior expected daily cost is

$$\rho(S, H) = \int_\Theta R(S, \theta)h(\theta)d\theta,$$

or, since all the terms are nonnegative

$$\begin{aligned}\rho(S, H) &= \int_\Theta h(\theta)\left\{\sum_{x=0}^{\infty} f(x;\theta)K(S, x)\right\}d\theta \\ &= \sum_{x=0}^{\infty} K(S, x)\int_\Theta f(x;\theta)h(\theta)d\theta.\end{aligned}$$

The value of S which minimizes $\rho(S, H)$ is similar to (8.1.27),

$$S^0(H) = F_H^{-1}\left(\frac{h}{c+h}\right),$$

i.e., the $h/(c + h)$th-quantile of the predictive distribution $F_H(x)$.

After observing the value x_1 of X_1, we convert the prior distribution $H(\theta)$ to a posterior distribution $H_1(\theta \mid x_1)$ and determine the predictive p.d.f. for the second day, namely

$$f_{H_1}(y) = \int_\Theta f(y;\theta) h_1(\theta \mid x_1) d\theta.$$

The expected cost for the second day is

$$\rho(S_2, H) = \sum_{y=0}^{\infty} K(S_2, y) f_H(y).$$

Moreover, by the law of the iterated expectations

$$f_H(y) = \sum_{x=0}^{\infty} f_{H_1}(y \mid x) f_H(x).$$

Hence,

$$\rho(S_2, H) = \sum_{x=0}^{\infty} f_H(x) \sum_{y=0}^{\infty} K(S_2, y) f_{H_1}(y \mid x).$$

The conditional expectation $\sum_{y=0}^{\infty} K(S_2, y) f_{H_1}(y \mid x)$ is the posterior expected cost given $X_1 = x$; or the predictive cost for the second day. The optimal choice of S_2 given $X_1 = x$ is, therefore, the $h/(c+h)$-quantile of the predictive distribution $F_{H_1}(y \mid x)$ i.e., $S_2^0(H) = F_{H_1}^{-1}\left(\frac{h}{c+h} \mid x\right)$. Since this function minimizes the predictive risk for every x, it minimizes $\rho(S_2, H)$. In the same manner, we prove that after n days, given $\mathbf{X}_n = (x_1, \ldots, x_n)$ the optimal stock level for the beginning of the $(n+1)$st day is the $\frac{h}{c+h}$-quantile of the predictive distribution of X_{n+1}, given $\mathbf{X}_n = \mathbf{x}_n$, i.e., $S_{n+1}^0(\mathbf{x}_n) = F_{H_n}^{-1}\left(\frac{h}{c+h} \mid \mathbf{x}_n\right)$, where the predictive p.d.f. of X_{n+1}, given $\mathbf{X}_n = \mathbf{x}$ is

$$f_{H_n}(y \mid \mathbf{x}) = \int_\Theta f(y;\theta) h(\theta \mid \mathbf{x}) d\theta,$$

and $h(\theta \mid \mathbf{x})$ is the posterior p.d.f. of θ given $\mathbf{X}_n = x$. The optimal stock levels are determined sequentially for each day on the basis of the demand of the previous days. Such a procedure is called an **adaptive procedure**. In particular, if $X_1, X_2, \ldots$ is a sequence of i.i.d. Poisson random variables (r.v.s), $P(\theta)$ and if the prior distribution

$H(\theta)$ is the gamma distribution, $G\left(\frac{1}{\tau}, \nu\right)$, the posterior distribution of θ after n observations is the gamma distribution $G\left(\frac{1}{\tau} + n, \nu + T_n\right)$, where $T_n = \sum_{i=1}^{n} X_i$. Let $g(\theta \mid n + \frac{1}{\tau}, \nu + T_n)$ denote the p.d.f. of this posterior distribution. The predictive distribution of X_{n+1} given $\mathbf{X}_n$, which actually depends only on T_n, is

$$
\begin{aligned}
f_{H_n}(y \mid T_n) &= \frac{1}{y!} \int_0^\infty e^{-\theta} \theta^y g\left(\theta \mid \frac{1}{\tau} + n, \nu + T_n\right) d\theta \\
&= \frac{\Gamma(\nu + y + T_n)}{\Gamma(y+1)\Gamma(\nu + T_n)} \psi_n^y (1 - \psi_n)^{\nu + T_n},
\end{aligned}
$$

where $\psi_n = \tau/(1 + (n+1)\tau)$. This is the p.d.f. of the negative binomial $NB(\psi_n, \nu + T_n)$. It is interesting that in the present case the predictive distribution belongs to the family of the negative-binomial distributions for all $n = 1, 2, \ldots$. We can also include the case of $n = 0$ by defining $T_0 = 0$. What changes from one day to another are the parameters $(\psi_n, \nu + T_n)$. Thus, the optimal stock level at the beginning of the $(n+1)$st day is the $h/(c+h)$-quantile of the $NB(\psi_n, \nu + T_n)$. ■

Example 8.6. Consider the testing problem connected with the problem of detecting disturbances in a manufacturing process. Suppose that the quality of a product is presented by a random variable X having a normal distribution $N(\theta, 1)$. When the manufacturing process is under control the value of θ should be θ_0. Every hour an observation is taken on a product chosen at random from the process. Consider the situation after n hours. Let $X_1, \ldots, X_n$ be independent random variables representing the n observations. It is suspected that after k hours of operation $1 < k < n$ a malfunctioning occurred and the expected value θ shifted to a value θ_1 greater than θ_0. The loss due to such a shift is $(\theta_1 - \theta_0)$ [\$] per hour. If a shift really occurred the process should be stopped and rectified. On the other hand, if a shift has not occurred and the process is stopped a loss of K [\$] is charged. The prior probability that the shift occurred is ψ. We present here the Bayes test of the two hypotheses

$$
H_0 : X_1, \ldots, X_n \text{ are i.i.d. like } N(\theta_0, 1)
$$

against

$$
\begin{aligned}
H_1 : &X_1, \ldots, X_k \text{ are i.i.d. like } N(\theta_0, 1) \text{ and} \\
&X_{k+1}, \ldots, X_n \text{ are i.i.d. like } N(\theta_1, 1),
\end{aligned}
$$

for a specified k, $1 \le k \le n-1$; which is performed after the nth observation.

The likelihood functions under H_0 and under H_1 are, respectively, when $\mathbf{X}_n = \mathbf{x}_n$

$$L_0(\mathbf{x}_n) = \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta_0)^2\right\},$$

and

$$L_1(\mathbf{x}_n) = \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{k}(x_i - \theta_0)^2 + \sum_{i=k+1}^{n}(x_i - \theta_1)^2\right]\right\}.$$

Thus, the posterior probability that H_0 is true is

$$\pi(\mathbf{x}_n) = \pi L_0(\mathbf{x}_n)/\{\pi L_0(\mathbf{x}_n) + (1-\pi)L_1(\mathbf{x}_n)\},$$

where $\pi = 1 - \psi$. The ratio of prior risks is in the present case $K\pi/((1-\pi)(n-k)(\theta_1 - \theta_0))$. The Bayes test implies that H_0 should be rejected if

$$\bar{X}^*_{n-k} \geq \frac{\theta_0 + \theta_1}{2} + \frac{1}{(n-k)(\theta_1 - \theta_0)}\log\frac{K\pi}{(1-\pi)(n-k)(\theta_1 - \theta_0)},$$

where $\bar{X}^*_{n-k} = \dfrac{1}{n-k}\displaystyle\sum_{j=k+1}^{n} X_j$.

The Bayes (minimal prior) risk associated with this test is

$$\rho(\pi) = \pi K \epsilon_0(\pi) + (1-\pi)(n-k)(\theta_1 - \theta_0)\epsilon_1(\pi),$$

where $\epsilon_0(\pi)$ and $\epsilon_1(\pi)$ are the error probabilities of rejecting H_0 or H_1 when they are true. These error probabilities are given by

$$\begin{aligned}\epsilon_0(\pi) &= P_{\theta_0}\left[\bar{X}^*_{n-k} \geq \frac{\theta_0 + \theta_1}{2} + A_{n-k}(\pi)\right] \\ &= 1 - \Phi\left(\sqrt{n-k}\left(\frac{\theta_1 - \theta_0}{2} + A_{n-k}(\pi)\right)\right),\end{aligned}$$

where $\Phi(z)$ is the standard normal integral and

$$A_{n-k}(\pi) = \frac{1}{(n-k)(\theta_1 - \theta_0)}\left(\log\frac{K}{(n-k)(\theta_1 - \theta_0)} + \log\frac{\pi}{1-\pi}\right).$$

Similarly,

$$\epsilon_1(\pi) = 1 - \Phi\left(\sqrt{n-k}\left(\frac{\theta_1 - \theta_0}{2} - A_{n-k}(\pi)\right)\right).$$

The function $A_{n-k}(\pi)$ is monotone increasing in π and $\lim_{\pi\to 0} A_{n-k}(\pi) = -\infty$, $\lim_{\pi\to 1} A_{n-k}(\pi) = \infty$. Accordingly, $\epsilon_0(0) = 1$, $\epsilon_1(0) = 0$ and $\epsilon_0(1) = 0$, $\epsilon_1(1) = 1$. ■

Example 8.7. Consider the detection problem of Example 8.6 but now the point of shift k is unknown. If θ_0 and θ_1 are known then we have a problem of testing the simple hypothesis H_0 (of Example 8.6) against the composite hypothesis

$$H_1 : X_1, \ldots, X_k \sim N(\theta_0, 1), X_{k+1}, \ldots, X_n \sim N(\theta_1, 1) \text{ for } k = 1, \ldots, n-1.$$

Let π_0 be the prior probability of H_0 and π_j, $j = 1, \ldots, n-1$, the prior probabilities under H_1 that $\{k = j\}$. The posterior probability of H_0 is then

$$\Pi_0(\mathbf{X}_n) = \left(1 + \frac{1}{\sqrt{n}} \sum_{j=1}^{n-1} \frac{\pi_j}{\pi_0} \sqrt{j(n-j)} \cdot \right.$$
$$\left. \cdot \exp\left\{ -\frac{n}{2}\left[\frac{j}{n}(\bar{X}_j - \theta_0)^2 + \left(1 - \frac{j}{n}\right)(\bar{X}^*_{n-j} - \theta_1)^2 - (\bar{X}_n - \theta_0)^2\right]\right\}\right)^{-1},$$

where $\bar{X}_j = \frac{1}{j}\sum_{i=1}^{j} X_i$, $j = 1, \ldots, n$ and $\bar{X}^*_{n-j} = \frac{1}{n-j} \sum_{i=n-j+1}^{n} X_i$. The posterior probability of $\{k = j\}$ is, for $j = 1, \ldots, n-1$,

$$\Pi_j(\mathbf{X}_n) = \Pi_0(\mathbf{X}_n) \frac{\pi_j}{\pi_0} \frac{1}{\sqrt{n}} \sqrt{j(n-j)} \cdot$$
$$\cdot \exp\left\{ -\frac{n}{2}\left[\frac{j}{n}(\bar{X}_j - \theta_0)^2 + \left(1 - \frac{j}{n}\right)(\bar{X}^*_{n-j} - \theta_j) - (\bar{X}_n - \theta_0)^2\right]\right\}.$$

Let $R_i(\mathbf{X}_n)$ $(i = 0, 1)$ denote the posterior risk associated with accepting H_i. These functions are given by

$$R_0(\mathbf{X}_n) = |\theta_1 - \theta_0| \sum_{j=1}^{n-1} (n-j)\Pi_j(X_n),$$

and

$$R_1(\mathbf{X}_n) = K\,\Pi_0(\mathbf{X}_n),$$

H_0 is rejected if $R_1(\mathbf{X}_n) \le R_0(\mathbf{X}_n)$, or when

$$\sum_{j=1}^{n-1}(n-j)\pi_j\sqrt{j(n-j)}\exp\left\{-\frac{i}{2}(\bar{X}_j-\theta_0)^2-\frac{n}{2}\left(1-\frac{j}{n}\right)(\bar{X}^*_{n-j}-\theta_1)^2\right\}$$
$$\ge \frac{K\pi_0\sqrt{n}}{|\theta_1-\theta_0|}\exp\left\{-\frac{n}{2}(\bar{X}_n-\theta_0)^2\right\}.$$

■

Example 8.8. We consider here the problem of testing whether the mean of a normal distribution is negative or positive. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a $N(\theta, 1)$ distribution. The null hypothesis is $H_0 : \theta \le 0$ and the alternative hypothesis is $H_1 : \theta > 0$. We assign the unknown θ a prior normal distribution, i.e., $\theta \sim N(0, \tau^2)$. Thus, the prior probability of H_0 is $\pi = \frac{1}{2}$. The loss function $L_0(\theta)$ of accepting H_0 and that of accepting H_1, $L_1(\theta)$, are of the form

$$L_0(\theta) = \begin{cases} 0, & \text{if } \theta \le 0, \\ \theta^2, & \text{if } \theta > 0, \end{cases} \qquad L_1(\theta) = \begin{cases} \theta^2, & \text{if } \theta < 0, \\ 0, & \text{if } \theta \ge 0. \end{cases}$$

For the determination of the posterior risk functions, we have to determine first the posterior distribution of θ given $\mathbf{X}_n$. Since $\bar{X}_n$ is a minimal sufficient statistic the conditional distribution of θ given $\bar{X}_n$ is the normal

$$N\left(\bar{X}_n\frac{\tau_n^2}{1+\tau_n^2}, \frac{1}{n}\frac{\tau_n^2}{1+\tau_n^2}\right).$$

(See Example 8.9 for elaboration.) It follows that the posterior risk associated with accepting H_0 is

$$R_0(\bar{X}_n) = \frac{\sqrt{n}(1+1/\tau^2 n)^{1/2}}{\sqrt{2\pi}}\int_0^\infty \theta^2\exp\left\{-\frac{n}{2}\left(1+\frac{1}{\tau^2 n}\right)(\theta-\hat{\theta}(\bar{X}_n))^2\right\}d\theta,$$

where $\hat{\theta}(\bar{X}_n)$ is the posterior mean. Generally, if $X \sim N(\xi, D^2)$ then

$$\frac{1}{\sqrt{2\pi}\,D}\int_0^\infty x^2\exp\left\{-\frac{1}{2D^2}(x-\xi)^2\right\}dx = (\xi^2+D^2)\Phi\left(\frac{\xi}{D}\right)+\xi D\phi\left(\frac{\xi}{D}\right).$$

Substituting the expressions

$$\xi = \bar{X}_n\left(1+\frac{1}{\tau^2 n}\right)^{-1} \quad \text{and} \quad D^2 = \frac{1}{n(1+\tau^2/n)}$$

we obtain that

$$R_0(\bar{X}_n) = \left(1 + \frac{1}{n\tau^2}\right)^{-1} \left(\frac{1}{n} + \bar{X}_n^2\right) \Phi\left(\sqrt{n}\,\bar{X}_n \left(1 + \frac{1}{n\tau^2}\right)^{-1/2}\right) + \frac{\bar{X}_n}{\sqrt{n}} \left(1 + \frac{1}{n\tau^2}\right)^{-3/2} \phi\left(\sqrt{n}\,\bar{X}_n \left(1 + \frac{1}{n\tau^2}\right)^{-1/2}\right).$$

In a similar fashion, we prove that the posterior risk associated with accepting H_1 is

$$R_1(\bar{X}_n) = \left(1 + \frac{1}{\tau^2 n}\right)^{-1} \left(\frac{1}{n} + \bar{X}_n^2\right) \Phi\left(-\sqrt{n}\,\bar{X}_n \left(1 + \frac{1}{n\tau^2}\right)^{-1/2}\right) - \frac{\bar{X}_n}{\sqrt{n}} \left(1 + \frac{1}{n\tau^2}\right)^{-3/2} \phi\left(\sqrt{n}\,\bar{X}_n \left(1 + \frac{1}{n\tau^2}\right)^{-1/2}\right).$$

The Bayes test procedure is to reject H_0 whenever $R_1(\bar{X}_n) \leq R_0(\bar{X}_n)$. Thus, H_0 should be rejected whenever

$$\left(\frac{1}{n} + \bar{X}_n^2\right) \left[2\Phi\left(\sqrt{n}\,\bar{X}_n \left(1 + \frac{1}{n\tau^2}\right)^{-1/2}\right) - 1\right] \geq -\frac{2\bar{X}_n}{\sqrt{n}} \left(1 + \frac{1}{n\tau^2}\right)^{-1/2} \phi\left(\sqrt{n}\,\bar{X}_n \left(1 + \frac{1}{n\tau^2}\right)^{-1/2}\right).$$

But this holds if, and only if, $\bar{X}_n \geq 0$. ■

Example 8.9. Let $X_1, X_2, \ldots$ be i.i.d. random variables having a normal distribution with mean μ, and variance $\sigma^2 = 1$. We wish to test $k = 3$ composite hypotheses $H_{-1} : -\infty < \mu < -1$; $H_0 : -1 \leq \mu \leq 1$; $H_1 : \mu > 1$. Let μ have a prior normal distribution, $\mu \sim N(0, \tau^2)$. Thus, let

$$\pi_{-1} = \Phi\left(-\frac{1}{\tau}\right), \quad \pi_0 = \Phi\left(\frac{1}{\tau}\right) - \Phi\left(-\frac{1}{\tau}\right)$$

and $\pi_1 = 1 - \Phi\left(\frac{1}{\tau}\right)$, be the prior probabilities of H_{-1}, H_0, and H_1, respectively. Furthermore, let

$$G_{-1}(\mu) = \begin{cases} \dfrac{\Phi(\mu/\tau)}{\Phi\left(-\frac{1}{\tau}\right)}, & \text{if } \mu < -1, \\ 1, & \text{if } \mu \geq -1, \end{cases}$$

$$G_0(\mu) = \begin{cases} 0, & \text{if } \mu < -1, \\ \dfrac{\Phi\left(\dfrac{\mu}{\tau}\right) - \Phi\left(-\dfrac{1}{\tau}\right)}{\Phi\left(\dfrac{1}{\tau}\right) - \Phi\left(-\dfrac{1}{\tau}\right)}, & \text{if } -1 \le \mu \le 1, \\ 1, & \text{if } \mu \ge 1, \end{cases}$$

and

$$G_1(\mu) = \begin{cases} 0, & \text{if } \mu < 1, \\ \dfrac{\Phi\left(\dfrac{\mu}{\tau}\right) - \Phi\left(\dfrac{1}{\tau}\right)}{1 - \Phi\left(\dfrac{1}{\tau}\right)}, & \text{if } 1 \le \mu. \end{cases}$$

The predictive likelihood functions of the three hypotheses are then

$$L_{-1}(\bar{X}_n) = \frac{1}{\pi_{-1}}\left(1 - \Phi\left(\frac{1 + n\tau^2 + n\tau^2\bar{X}_n}{\tau\sqrt{1 + n\tau^2}}\right)\right) \cdot \sqrt{\frac{n}{1 + n\tau^2}} \exp\left\{-\frac{n}{2(1 + n\tau^2)}\bar{X}_n^2\right\},$$

$$L_0(\bar{X}_n) = \frac{1}{\pi_0}\left(\Phi\left(\frac{1 + n\tau^2 - n\tau^2\bar{X}_n}{\tau\sqrt{1 + n\tau^2}}\right) + \Phi\left(\frac{1 + n\tau^2 + n\tau^2\bar{X}_n}{\tau\sqrt{1 + n\tau^2}}\right) - 1\right) \cdot$$
$$\cdot \sqrt{\frac{n}{1 + n\tau^2}} \exp\left\{-\frac{n}{2(1 + n\tau^2)}\bar{X}_n^2\right\},$$

and

$$L_1(\bar{X}_n) = \frac{1}{\pi_1}\left(1 - \Phi\left(\frac{1 + n\tau^2 - n\tau^2\bar{X}_n}{\tau\sqrt{1 + n\tau^2}}\right)\right) \cdot \sqrt{\frac{n}{1 + n\tau^2}} \exp\left\{-\frac{n}{2(1 + n\tau^2)}\bar{X}_n^2\right\}.$$

It follows that the posterior probabilities of H_j, $j = -1, 0, 1$ are as follows:

$$\pi_j(\bar{X}_n) = \begin{cases} 1 - \Phi\left(\dfrac{1 + n\tau^2(1 + \bar{X}_n)}{\tau\sqrt{1 + n\tau^2}}\right), & j = -1, \\ \Phi\left(\dfrac{1 + n\tau^2(1 - \bar{X}_n)}{\tau\sqrt{1 + n\tau^2}}\right) + \Phi\left(\dfrac{1 + n\tau^2(1 + \bar{X}_n)}{\tau\sqrt{1 + n\tau^2}}\right) - 1, & j = 0, \\ 1 - \Phi\left(\dfrac{1 + n\tau^2(1 - \bar{X}_n)}{\tau\sqrt{1 + n\tau^2}}\right), & j = 1. \end{cases}$$

Thus, $\pi_{-1}(\bar{X}_n) \geq 1-\epsilon$ if

$$\bar{X}_n \leq -\left(1+\frac{1}{n\tau^2}+Z_{1-\epsilon}\frac{\left(1+\frac{1}{n\tau^2}\right)^{1/2}}{\sqrt{n}\tau}\right).$$

Similarly, $\pi_1(\bar{X}_n) \geq 1-\epsilon$ if

$$\bar{X}_n \geq 1+\frac{1}{n\tau^2}+Z_{1-\epsilon}\frac{\sqrt{1+1/n\tau^2}}{\sqrt{n}\,\tau}.$$

Thus, if $b_n = 1+\frac{1}{n\tau^2}+Z_{1-\epsilon}\sqrt{1+\frac{1}{n\tau^2}}/\tau\sqrt{n}$ then b_n and $-b_n$ are outer stopping boundaries. In the region $(-b_n, b_n)$, we have two inner boundaries $(-c_n, c_n)$ such that if $|\bar{X}_n| < c_n$ then H_0 is accepted. The boundaries $\pm c_n$ can be obtained by solving the equation

$$\Phi\left(\frac{1+n\tau^2(1-c_n)}{\tau\sqrt{1+n\tau^2}}\right)+\Phi\left(\frac{1+n\tau^2(1+c_n)}{\tau\sqrt{1+n\tau^2}}\right)=2-\epsilon.$$

$c_n \geq 0$ and $c_n > 0$ only if $n > n_0$, where $\Phi\left(\frac{\sqrt{1+n_0\tau^2}}{\tau}\right) = 1-\frac{\epsilon}{2}$, or $n_0 = \left(Z^2_{1-\epsilon/2}-\frac{1}{\tau^2}\right)^+$. ■

Example 8.10. Consider the problem of estimating circular probabilities in the normal case. In Example 6.4, we derived the uniformly most accurate (UMA) lower confidence limit of the function

$$\psi(\sigma^2, \rho) = 1 - E_\rho\left\{P\left(J; \frac{1}{2\sigma^2}\right)\right\},$$

where J is a $NB\left(1-\frac{1}{\rho}, \frac{1}{2}\right)$ random variable for cases of **known** ρ. We derive here the Bayes lower credibility limit of $\psi(\sigma^2, \rho)$ for cases of known ρ. The minimal sufficient statistic is $T_{2n} = \sum_{i=1}^{n} X_i^2 + \frac{1}{\rho}\sum_{j=1}^{n} Y_j^2$. This statistic is distributed like $\sigma^2\chi^2[2n]$ or like $G\left(\frac{1}{2\sigma^2}, n\right)$. Let $\omega = \frac{1}{\sigma^2}$ and let $\omega \sim G(\tau, \nu)$. The posterior distribution of ω, given T_{2n}, is

$$\omega \mid T_{2n} \sim G(T_{2n}+\tau, n+\nu).$$

Accordingly, if $G^{-1}(p \mid T_{2n} + \tau, n + \nu)$ designates the pth quantile of this posterior distribution,

$$P\{\omega \geq G^{-1}(\alpha \mid T_{2n} + \tau, n + \nu) \mid T_{2n}\} = 1 - \alpha,$$

with probability one (with respect to the mixed prior distribution of T_{2n}). Thus, we obtain that a $1 - \alpha$ Bayes upper credibility limit for σ^2 is

$$\begin{aligned}\bar{\sigma}^2_\alpha &= \frac{1}{2G^{-1}(\alpha \mid T_{2n} + \tau, n + \nu)} \\ &= \frac{T_{2n} + \tau}{2G^{-1}(\alpha \mid 1, n + \nu)}.\end{aligned}$$

Note that if τ, and ν are close to zero then the Bayes credibility limit is very close to the non-Bayes UMA upper confidence limit derived in Example 7.4. Finally, the $(1 - \alpha)$ Bayes lower credibility limit for $\psi(\sigma^2, \rho)$ is $\psi(\bar{\sigma}^2_\alpha, \rho)$. ■

Example 8.11. We consider in the present example the problem of **inverse regression**. Suppose that the relationship between a controlled experimental variable x and an observed random variable $Y(x)$ is describable by a linear regression

$$Y(x) = \alpha + \beta x + \epsilon,$$

where ϵ is a random variable such that $E\{\epsilon\} = 0$ and $E\{\epsilon^2\} = \sigma^2$. The regression coefficients α and β are unknown. Given the results on n observations at $x_1, \ldots, x_n$, estimate the value of ξ at which $E\{Y(\xi)\} = \eta$, where η is a preassigned value. We derive here Bayes confidence limits for $\xi = (\eta - \alpha)/\beta$, under the assumption that m random variables are observed independently at x_1 and x_2, where $x_2 = x_1 + \Delta$. Both x_1 and Δ are determined by the design. Furthermore, we assume that the distribution of ϵ is $N(0, \sigma^2)$ and that (α, β) has a prior bivariate normal distribution with mean (α_0, β_0) and covariance matrix $V = (v_{ij}; i, j = 1, 2)$. For the sake of simplicity, we assume in the present example that σ^2 is known. The results can be easily extended to the case of unknown σ^2.

The minimal sufficient statistic is $(\bar{Y}_1, \bar{Y}_2)$ where $\bar{Y}_i$ is the mean of the m observations at x_i $(i = 1, 2)$. The posterior distribution of (α, β) given $(\bar{Y}_1, \bar{Y}_2)$ is the bivariate normal with mean vector

$$\begin{pmatrix}\alpha_1 \\ \beta_1\end{pmatrix} = \begin{pmatrix}\alpha_0 \\ \beta_0\end{pmatrix} + VX'\left(\frac{\sigma^2}{m}I + XVX'\right)^{-1} \cdot \begin{pmatrix}\bar{Y}_1 - (\alpha_0 + \beta_0 x_1) \\ \bar{Y}_2 - (\alpha_0 + \beta_0 x_2)\end{pmatrix},$$

where

$$X = \begin{pmatrix}1 & x_1 \\ 1 & x_2\end{pmatrix}$$

and I is the 2×2 identity matrix. Note that X is nonsingular. X is called the design matrix. The covariance matrix of the posterior distribution is

$$\Sigma = V - VX' \left(\frac{\sigma^2}{m} I + XVX' \right)^{-1} XV.$$

Let us denote the elements of Σ by Σ_{ij}, $i, j = 1, 2$. The problem is to determine the Bayes credibility interval to the parameter $\xi = (\eta - \alpha)/\beta$. Let $\underline{\xi}_\alpha$ and $\bar{\xi}_\alpha$ denote the limits of such a $(1 - \alpha)$ Bayes credibility interval. These limits should satisfy the posterior confidence level requirement

$$P \left\{ \underline{\xi}_\alpha \leq \frac{\eta - \alpha}{\beta} \leq \bar{\xi}_\alpha \mid \bar{Y}_1, \bar{Y}_2 \right\} \geq 1 - \alpha.$$

If we consider equal tail probabilities, these confidence limits are obtained by solving simultaneously the equations

$$\Phi \left(\frac{\eta - \alpha_1 - \beta_1 \underline{\xi}_\alpha}{\underline{D}^{1/2}} \right) = \alpha/2,$$
$$\Phi \left(\frac{\eta - \alpha_1 - \beta_1 \bar{\xi}_\alpha}{\bar{D}^{1/2}} \right) = 1 - \alpha/2,$$

where $\underline{D} = \Sigma_{11} + 2\underline{\xi}_\alpha \Sigma_{12} + \underline{\xi}_\alpha^2 \Sigma_{22}$ and similarly $\bar{D} = \Sigma_{11} + 2\bar{\xi}_\alpha \Sigma_{12} + \bar{\xi}_\alpha \Sigma_{22}$. By inverting, we can realize that the credibility limits $\underline{\xi}_\alpha$ and $\bar{\xi}_\alpha$ are the two roots of the quadratic equation

$$(\eta - \alpha_1 - \beta_1 \xi)^2 = \chi^2_{1-\alpha}[1](\Sigma_{11} + 2\xi \Sigma_{12} + \xi^2 \Sigma_{22}),$$

or

$$A\xi^2 - 2B\xi + C = 0,$$

where

$$A = \beta_1^2 - \chi^2_{1-\alpha}[1]\Sigma_{22},$$
$$B = \beta_1(\eta - \alpha_1)^2 + \chi^2_{1-\alpha}[1]\Sigma_{12},$$
$$C = (\eta - \alpha_1)^2 - \chi^2_{1-\alpha}[1]\Sigma_{11}.$$

The two roots (if they exist) are

$$\xi_{1,2} = \frac{(\eta - \alpha_1) + \chi^2_{1-\alpha}[1]\Sigma_{12}/\beta_1}{\beta_1 - \chi^2_{1-\alpha}[1]\Sigma_{22}/\beta_1} \pm \frac{(\chi^2_{1-\alpha}[1])^{1/2}\{(n-\alpha_1)^2\Sigma_{11} + 2(\eta-\alpha_1)\beta_1\Sigma_{12} + \beta_1^2\Sigma_{22} - \chi^2_{1-\alpha}[1]|\Sigma|\}^{1/2}}{\beta_1^2 - \chi^2_{1-\alpha}[1]\Sigma_{22}},$$

where $|\Sigma|$ denotes the determinant of the posterior covariance matrix. These credibility limits exist if the discriminant

$$\Delta^* = (\eta - \alpha_1, \beta_1)\Sigma\begin{pmatrix} \eta - \alpha_1 \\ \beta_1 \end{pmatrix} - \chi^2_{1-\alpha}[1]|\Sigma|$$

is nonnegative. After some algebraic manipulations, we obtain that

$$|\Sigma| = \frac{\sigma^2}{m}|V| \cdot \left|\left(\frac{\sigma^2}{m} + tr\{XVX'\}\right) I - XVX'\right|,$$

where tr$\{\cdot\}$ is the trace of the matrix in $\{\ \}$. Thus, if m is sufficiently large, $\Delta^* > 0$ and the two credibility limits exist with probability one. ■

Example 8.12. Suppose that $X_1, \ldots, X_{n+1}$ are i.i.d. random variables, having a Poisson distribution $P(\lambda)$, $0 < \lambda < \infty$. We ascribe λ a prior gamma distribution, i.e., $\lambda \sim G(\Lambda, \alpha)$.

After observing $X_1, \ldots, X_n$, the posterior distribution of λ, given $T_n = \sum_{i=1}^{n} X_i$ is $\lambda \mid T_n \sim G(\Lambda + n, T_n + \alpha)$. The predictive distribution of X_{n+1}, given T_n, is the negative-binomial, i.e.,

$$X_{n+1} \mid T_n \sim NB(\psi_n, T_n + \alpha),$$

where

$$\psi_n = \frac{\Lambda + n}{\Lambda + n + 1}.$$

Let $NB^{-1}(p; \psi, \alpha)$ denote the pth quantile of $NB(\psi, \alpha)$. The prediction interval for X_{n+1}, after observing $X_1, \ldots, X_n$, at level $1 - \alpha$, is

$$\left(NB^{-1}\left(\frac{\alpha}{2}; \psi_n, T_n + \alpha\right), NB^{-1}\left(1 - \frac{\alpha}{2}; \psi_n, T_n + \alpha\right)\right).$$

According to Equation (2.3.12), the pth quantile of $NB(\psi, \alpha)$ is $NB^{-1}(p \mid \psi, \alpha) =$ least integer k, $k \geq 1$, such that $I_{1-\psi}(\alpha, k+1) \geq p$. ■

Example 8.13. Suppose that an n-dimensional random vector $\mathbf{X}_n$ has the multinormal distribution $\mathbf{X}_n \mid \mu \sim N(\mu \mathbf{1}_n, \mathbf{V}_n)$, where $-\infty < \mu < \infty$ is unknown. The covariance matrix $\mathbf{V}_n$ is known. Assume that μ has a prior normal distribution, $\mu \sim N(\mu_0, \tau^2)$. The posterior distribution of μ, given $\mathbf{X}_n$, is $\mu \mid \mathbf{X}_n \sim N(\eta(\mathbf{X}_n), D_n)$, where

$$\eta(\mathbf{X}_n) = \mu_0 + \tau^2 1_n'(V_n + \tau^2 J_n)^{-1}(\mathbf{X}_n - \mu_0 1_n)$$

and

$$D_n = \tau^2(1 - \tau^2 1_n'(V_n + \tau^2 J_n)^{-1} 1_n).$$

Accordingly, the predictive distribution, of yet unobserved m-dimensional vector $\mathbf{Y}_m \mid \mu \sim N(\mu 1_m, V_m)$, is

$$\mathbf{Y}_n \mid \mathbf{X}_n \sim N(\eta(\mathbf{X}_n)1_m, V_m + D_n J_m).$$

Thus, a prediction region for $\mathbf{Y}_m$, at level $(1 - \alpha)$ is the ellipsoid of concentration

$$(\mathbf{Y}_m - \eta(\mathbf{X}_n)1_m)'(V_m + D_n J_m)^{-1}(\mathbf{Y}_m - \eta(\mathbf{X}_n)1_m) \leq \chi^2_{1-\alpha}[m].$$

■

Example 8.14. A new drug is introduced and the physician wishes to determine a lower prediction limit with confidence probability of $\gamma = 0.95$ for the number of patients in a group of $n = 10$ that will be cured. If X_n is the number of patients cured among n and if θ is the individual probability to be cured the model is binomial, i.e., $X_n \sim B(n, \theta)$. The lower prediction limit, for a given value of θ, is an integer $k_\gamma(\theta)$ such that $P_\theta\{X_n \geq k_\gamma(\theta)\} \geq \gamma$. If $B^{-1}(p; n, \theta)$ denotes the pth quantile of the binomial $B(n, \theta)$ then $k_\gamma(\theta) = \max(0, B^{-1}(1 - \gamma; n, \theta) - 1)$. Since the value of θ is unknown, we cannot determine $k_\gamma(\theta)$. Lower tolerance limits, which were discussed in Section 6.5, could provide estimates to the unknown $k_\gamma(\theta)$. A statistician may feel, however, that lower tolerance limits are too conservative, since he has good a priori information about θ. Suppose a statistician believes that θ is approximately equal to 0.8, and therefore, assigns θ a prior beta distribution $\beta(p, q)$ with mean 0.8 and variance 0.01. Setting the equations for the mean and variance of a $\beta(p, q)$ distribution (see Table 2.1 of Chapter 2), and solving for p and q, we obtain $p = 12$ and $q = 3$. We consider now the predictive distribution of X_n under $\beta(12, 3)$ prior distribution of θ. This predictive distribution has a probability function

$$p_H(j) = \binom{n}{j} \frac{B(12 + j, 3 + n - j)}{B(12, 3)}, \quad j = 0, \ldots, n.$$

For $n = 10$, we obtain the following predictive p.d.f. $p_H(j)$ and c.d.f. $f_H(j)$. According to this predictive distribution, the probability of at least 5 cures out of 10 patients is 0.972 and for at least 6 cures is 0.925.

j	$p_H(j)$	$F_H(j)$
0	0.000034	0.000034
1	0.000337	0.000378
2	0.001790	0.002160
3	0.006681	0.008841
4	0.019488	0.028329
5	0.046770	0.075099
6	0.094654	0.169752
7	0.162263	0.332016
8	0.231225	0.563241
9	0.256917	0.820158
10	0.179842	1.000000

■

Example 8.15. Suppose that in a given (rather simple) inventory system (see Example 8.2) the monthly demand, X of some commodity is a random variable having a Poisson distribution $P(\theta)$, $0 < \theta < \infty$. We wish to derive a Bayes estimator of the expected demand θ. In many of the studies on Bayes estimator of θ, a prior gamma distribution $G\left(\frac{1}{\tau}, \nu\right)$ is assumed for θ. The prior parameters τ and ν, $0 < \tau, \nu < \infty$, are specified. Note that the prior expectation of θ is $\nu\tau$ and its prior variance is $\nu\tau^2$. A large prior variance is generally chosen if the prior information on θ is vague. This yields a flat prior distribution. On the other hand, if the prior information on θ is strong in the sense that we have a high prior confidence that θ lies close to a value θ_0 say, pick $\nu\tau = \theta_0$ and $\nu\tau^2$ very small, by choosing τ to be small. In any case, the posterior distribution of θ, given a sample of n i.i.d. random variables $X_1, \ldots, X_n$, is determined in the following manner. $T_n = \sum_{i=1}^{n} X_i$ is a minimal sufficient statistic, where $T_n \sim P(n\theta)$. The derivation of the posterior density can be based on the p.d.f. of T_n. Thus, the product of the p.d.f. of T_n by the prior p.d.f. of θ is proportional to $\theta^{t+\nu-1}e^{-\theta(n+1/\tau)}$, where $T_n = t$. The factors that were omitted from the product of the p.d.fs are independent of θ and are, therefore, irrelevant. We recognize in the function $\theta^{t+\nu-1}e^{-\theta(n+1/\tau)}$ the kernel (the factor depending on θ) of a gamma p.d.f. Accordingly, the posterior distribution of θ, given T_n, is the gamma distribution $G\left(n + \frac{1}{\tau}, T_n + \nu\right)$. If we choose a squared-error loss function, then the posterior expectation is the Bayes estimator. We thus obtain the estimator $\hat{\theta} = (T_n + \nu)/\left(n + \frac{1}{\tau}\right)$. Note that the unbiased and the MLE of θ is T_n/n which is

not useful as long as $T_n = 0$, since we know that $\theta > 0$. If certain commodities have a very slow demand (a frequently encountered phenomenon among replacement parts) then T_n may be zero even when n is moderately large. On the other hand, the Bayes estimator θ is always positive. ■

Example 8.16. (a) Let $X_1, \ldots, X_N$ be i.i.d. random variables having a normal distribution $N(\theta, 1)$, $-\infty < \theta < \infty$. The minimal sufficient statistic is the sample mean $\bar{X}$. We assume that θ has a prior normal distribution $N(0, \tau^2)$. We derive the Bayes estimator for the zero-one loss function,

$$L(\hat{\theta}, \theta) = I\{\theta; |\hat{\theta} - \theta| \geq \delta\}.$$

The posterior distribution of θ given $\bar{X}$ is normal $N(\bar{X}(1 + 1/n\tau^2)^{-1}, (n + 1/\tau^2)^{-1})$. This can be verified by simple normal regression theory, recognizing that the joint distribution of $(\bar{X}, \theta)$ is the bivariate normal, with zero expectation and covariance matrix

$$V = \begin{pmatrix} \frac{1}{n} + \tau^2 & \tau^2 \\ \tau^2 & \tau^2 \end{pmatrix}.$$

Thus, the posterior risk is the posterior probability of the event $\{|\hat{\theta} - \theta| \geq \delta\}$. This is given by

$$R(\hat{\theta}, \tau^2) = 1 - \Phi\left(\frac{\hat{\theta} + \delta - \bar{X}(1 + 1/n\tau^2)^{-1}}{\left(n + \frac{1}{\tau^2}\right)^{-1/2}}\right) + \Phi\left(\frac{\hat{\theta} - \delta - \bar{X}(1 + \frac{1}{n\tau^2})^{-1}}{\left(n + \frac{1}{\tau^2}\right)^{-1/2}}\right).$$

We can show then (Zacks, 1971; p. 265) that the Bayes estimator of θ is the posterior expectation, i.e.,

$$\hat{\theta}(\bar{X}) = \bar{X}\left(1 + \frac{1}{n\tau^2}\right)^{-1}.$$

In this example, the minimization of the posterior variance and the maximization of the posterior probability of covering θ by the interval $(\hat{\theta} - \delta, \hat{\theta} + \delta)$ is the same. This is due to the normal prior and posterior distributions.

(b) Continuing with the same model, suppose that we wish to estimate the tail probability

$$\psi = P_\theta\{X \geq \xi_0\} = 1 - \Phi(\xi_0 - \theta) = \Phi(\theta - \xi_0).$$

Since the posterior distribution of $\theta - \xi_0$ given $\bar{X}$ is normal, the Bayes estimator for a squared-error loss is the posterior expectation

$$E_H\{\Phi(\theta - \xi_0) \mid \bar{X}\} = \Phi\left(\frac{\bar{X}\left(1 + \frac{1}{n\tau^2}\right)^{-1} - \xi_0}{\left(1 + \frac{\tau^2}{1 + n\tau^2}\right)^{1/2}}\right).$$

Note that this Bayes estimator is strongly consistent since, by the SLLN, $\bar{X} \to \theta$ almost surely (a.s.), and $\Phi(\cdot)$ is a continuous function. Hence, the Bayes estimator converges to $\Phi(\theta - \xi_0)$ a.s. as $n \to \infty$. It is interesting to compare this estimator to the minimum variance unbiased estimator (MVUE) and to the MLE of the tail probability. All these estimators are very close in large samples.

If the loss function is the absolute deviation, $|\hat{\psi} - \psi|$, rather than the squared-error, $(\hat{\psi} - \psi)^2$, then the Bayes estimator of ψ is the median of the posterior distribution of $\Phi(\theta - \xi_0)$. Since the Φ-function is strictly increasing this median is $\Phi(\theta_{.5} - \xi_0)$, where $\theta_{0.5}$ is the median of the posterior distribution of θ given $\bar{X}$. We thus obtain that the Bayes estimator for absolute deviation loss is

$$\hat{\psi} = \Phi\left(\bar{X}\left(1 + \frac{1}{n\tau^2}\right)^{-1} - \xi_0\right).$$

This is different from the posterior expectation. ■

Example 8.17. In this example, we derive Bayes estimators for the parameters μ and σ^2 in the normal model $N(\mu, \sigma^2)$ for squared-error loss. We assume that $X_1, \ldots, X_n$, given (μ, σ^2), are i.i.d. $N(\mu, \sigma^2)$. The minimal sufficient statistic is $(\bar{X}_n, Q)$, where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $Q = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. Let $\phi = 1/\sigma^2$ be the precision parameter, and consider the reparametrization $(\mu, \sigma^2) \to (\mu, \phi)$.

The likelihood function is

$$L(\mu, \phi) = \phi^{n/2} \exp\left\{-\frac{\phi}{2}[n(\mu - \bar{X}_n)^2 + Q]\right\}.$$

The following is a commonly assumed joint prior distribution for (μ, ϕ), namely,

$$\mu \mid \phi \sim N(\mu_0, \tau^2/\phi),$$

and

$$\phi \sim G\left(\frac{\psi}{2}, \frac{n_0}{2}\right), \quad n_0 > 2,$$

where n_0 is an integer, and $0 < \psi < \infty$. This joint prior distribution is called the Normal-Gamma prior, and denoted by $NG\left(\mu_0, \tau^2, \frac{n_0}{2}, \frac{\psi}{2}\right)$. Since $\bar{X}_n$ and Q are conditionally independent, given (μ, ϕ), and since the distribution of Q is independent of μ, the posterior distribution of $\mu \mid \bar{X}_n, \phi$ is normal with mean

$$\hat{\mu}_B = \frac{1}{1+\tau^2 n}\mu_0 + \frac{\tau^2 n}{1+\tau^2 n}\bar{X}_n$$

and variance

$$V\{\mu \mid \phi, T\} = \frac{\tau^2}{\phi(1+n\tau^2)}.$$

$\hat{\mu}_B$ is a Bayesian estimator of μ for the squared-error loss function. The posterior risk of $\hat{\mu}_B$ is

$$V\{\mu \mid \bar{X}_n, Q\} = E\{V\{\mu \mid \bar{X}_n, \phi\} \mid \bar{X}, Q\} + V\{E\{\mu \mid \bar{X}_n, \phi\} \mid \bar{X}, Q\}.$$

The second term on the RHS is zero. Thus, the posterior risk of $\hat{\mu}_B$ is

$$V\{\mu \mid \bar{X}, Q\} = \frac{\tau^2}{1+n\tau^2} E\left\{\frac{1}{\phi} \mid \bar{X}, Q\right\}.$$

The posterior distribution of ϕ depends only on Q. Indeed, if we denote generally by $p(\bar{X} \mid \mu, \phi)$ and $p(Q \mid \phi)$ the conditional p.d.f. of $\bar{X}$ and Q then

$$\begin{aligned} h(\mu, \phi \mid \bar{X}, Q) &\propto p(\bar{X} \mid \mu, \phi) p(Q \mid \phi) h(\mu \mid \phi) g(\phi) \\ &\propto h(\mu \mid \bar{X}, \phi) p(Q \mid \phi) g(\phi). \end{aligned}$$

Hence, the marginal posterior p.d.f. of ϕ is

$$\begin{aligned} h^*(\phi \mid \bar{X}, Q) &= \int_{-\infty}^{\infty} h(\mu, \phi \mid \bar{X}, Q) d\mu \\ &\propto p(Q \mid \phi) g(\phi). \end{aligned}$$

Thus, from our model, the posterior distribution of ϕ, given Q, is the gamma $G\left(\frac{Q+\psi}{2}, \frac{n+n_0-1}{2}\right)$. It follows that

$$E\left\{\frac{1}{\phi} \mid Q, \bar{X}\right\} = E\left\{\frac{1}{\phi} \mid Q\right\} = \frac{\psi + Q}{n_0 + n - 3}.$$

Thus, the posterior risk of $\hat{\mu}_B$ is

$$V\{\mu \mid \bar{X}, Q\} = \frac{\tau^2}{1 + n\tau^2} \cdot \frac{Q + \psi}{n + n_0 - 3}.$$

The Bayesian estimator of σ^2 is

$$\hat{\sigma}_B^2 = E\left\{\frac{1}{\phi} \mid Q\right\} = \frac{Q + \psi}{n + n_0 - 3}.$$

The posterior risk of $\hat{\sigma}_B^2$ is

$$\begin{aligned} V\left\{\frac{1}{\phi} \mid Q\right\} &= \frac{(Q+\psi)^2}{(n+n_0-3)(n+n_0-5)} - \frac{(Q+\psi)^2}{(n+n_0-3)^2} \\ &= \frac{2(Q+\psi)^2}{(n+n_0-3)^2(n+n_0-5)}, \end{aligned}$$

which is finite if $n + n_0 > 5$. ■

Example 8.18. Consider the model of the previous example, but with priorly independent parameters μ and ϕ, i.e., we assume that $h(\mu, \phi) = h(\mu)g(\phi)$, where

$$h(\mu) = \frac{1}{\sqrt{2\pi}\,\tau} \exp\left\{-\frac{1}{2\tau^2}(\mu - \mu_0)^2\right\}$$

and

$$g(\phi) = \frac{\left((\psi/2)^{\frac{n_0}{2}}\right)}{\Gamma\left(\frac{n_0}{2}\right)} \phi^{\frac{n_0}{2}-1} e^{-\phi\frac{\psi}{2}}.$$

If $p(\bar{X} \mid \mu, \phi)$ and $p(Q \mid \phi)$ are the p.d.f. of $\bar{X}$ and of Q, given μ, ϕ, respectively, then the joint posterior p.d.f. of (μ, ϕ), given $(\bar{X}, Q)$, is

$$h(\mu, \phi \mid \bar{X}, Q) = A(\bar{X}, Q)p(\bar{X} \mid \mu, \phi)p(Q \mid \phi)h(\mu)g(\phi),$$

where $A(\bar{X}, Q) > 0$ is a normalizing factor. The marginal posterior p.d.f. of μ, given $(\bar{X}, Q)$, is

$$h^*(\mu \mid \bar{X}, Q) = A(\bar{X}, Q)h(\mu) \cdot \int_0^\infty p(\bar{X} \mid \mu, \phi)p(Q \mid \phi)g(\phi)d\phi.$$

It is straightforward to show that the integral on the RHS of (8.4.18) is proportional to $\left[1 + \frac{n}{Q+\psi}(\mu - \bar{X})^2\right]^{-\frac{n_0+n}{2}}$. Thus,

$$h^*(\mu \mid \bar{X}, Q) = A^*(\bar{X}, Q) \exp\left\{-\frac{1}{2\tau^2}(\mu - \mu_0)^2\right\} \cdot \left[1 + \frac{n}{Q+\psi}(\mu - \bar{X})^2\right]^{-\frac{n_0+n}{2}}.$$

A simple analytic expression for the normalizing factor $A^*(\bar{X}, Q)$ is not available. One can resort to numerical integration to obtain the Bayesian estimator of μ, namely,

$$\hat{\mu}_B = A^*(\bar{X}, Q) \int_{-\infty}^\infty \mu e^{-\frac{1}{2\tau^2}(\mu-\mu_0)^2} \cdot \left[1 + \frac{n}{Q+\psi}(\mu - \bar{X})^2\right]^{-\frac{n_0+n}{2}} d\mu.$$

By the Lebesgue Dominated Convergence Theorem

$$\begin{aligned}\lim_{n\to\infty} \int_{-\infty}^\infty \mu \exp\left\{-\frac{1}{2\tau^2}(\mu - \mu_0)^2\right\} \cdot \left[1 + \frac{n}{Q+\psi}(\mu - \bar{X})^2\right]^{-\frac{n_0+n}{2}} d\mu \\ = \int_{-\infty}^\infty \mu \exp\left\{-\frac{1}{2\tau^2}(\mu - \mu_0)^2\right\} \lim_{n\to\infty} \cdot \\ \cdot \left[1 + \frac{n}{Q+\psi}(\mu - \bar{X})^2\right]^{-\frac{n_0+n}{2}} d\mu.\end{aligned}$$

Thus, for large values of n,

$$\hat{\mu}_B \approx \frac{\frac{\mu_0}{\tau^2} + \frac{\bar{X}}{\hat{\sigma}_n^2}}{\frac{1}{\tau^2} + \frac{1}{\hat{\sigma}^2}},$$

where $\hat{\sigma}_n^2 = \frac{Q+\psi}{n}$.

In a similar manner, we can show that the marginal posterior p.d.f. of ϕ is

$$g^*(\phi \mid \bar{X}, Q) = B^*(\bar{X}, Q) \frac{\phi^{\frac{n_0+n-1}{2}-1}}{\sqrt{1 + n\tau^2\phi}} \exp\left\{-\frac{\phi n}{2(1 + n\tau^2\phi)}(\bar{X} - \mu)^2 - \phi \cdot \frac{Q+\psi}{2}\right\},$$

where $B^*(\bar{X}, Q) > 0$ is a normalizing factor. Note that for large values of n, $g^*(\phi \mid \bar{X}, Q)$ is approximately the p.d.f. of $G\left(\frac{Q+\psi}{2}, \frac{n_0+n-1}{2}\right)$.

In Chapter 5, we discussed the least-squares and MVUEs of the parameters in linear models. Here, we consider Bayesian estimators for linear models. Comprehensive Bayesian analysis of various linear models is given in the books of Box and Tiao (1973) and of Zellner (1971). The analysis in Zellner's book (see Chapter III) follows a straightforward methodology of deriving the posterior distribution of the regression coefficients for informative and noninformative priors. Box and Tiao provide also geometrical representation of the posterior distributions (probability contours) and the HPD-regions of the parameters. Moreover, by analyzing the HPD-regions Box and Tiao establish the Bayesian justification to the analysis of variance and simultaneous confidence intervals of arbitrary contrasts (the Scheffé S-method). In Example 8.11, we derived the posterior distribution of the regression coefficients of the linear model $Y = \alpha + \beta x + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ and (α, β) have a prior normal distribution. In a similar fashion the posterior distribution of $\boldsymbol{\beta}$ in the multiple regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ can be obtained by assuming that $\boldsymbol{\epsilon} \sim N(0, V)$ and the prior distribution of $\boldsymbol{\beta}$ is $N(\boldsymbol{\beta}_0, \mathbf{B})$. By applying the multinormal theory, we readily obtain that the posterior distribution of $\boldsymbol{\beta}$, given $\mathbf{Y}$, is

$$\boldsymbol{\beta} \mid Y \sim N(\boldsymbol{\beta}_0 + \mathbf{B}\mathbf{X}'(\mathbf{V} + \mathbf{X}\mathbf{B}\mathbf{X}')^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0), \mathbf{B} - \mathbf{B}\mathbf{X}'(\mathbf{V} + \mathbf{X}\mathbf{B}\mathbf{X}')^{-1}\mathbf{X}\mathbf{B}).$$

This result is quite general and can be applied whenever the covariance matrix V is known. Often we encounter in the literature the $NG\left(\boldsymbol{\beta}_0, \tau^2, \frac{n_0}{2}, \frac{\psi}{2}\right)$ prior and the (observations) model

$$\mathbf{Y} \mid \boldsymbol{\beta}, \phi \sim N(\mathbf{X}\boldsymbol{\beta}, (1/\phi)I)$$

and

$$\boldsymbol{\beta} \mid \boldsymbol{\beta}_0, \tau^2 \sim N\left(\boldsymbol{\beta}_0, \frac{\tau^2}{\phi} I\right), \quad \text{and} \quad \phi \sim G\left(\frac{\psi}{2}, \frac{n_0}{2}\right).$$

This model is more general than the previous one, since presently the covariance matrices $\mathbf{V}$ and $\mathbf{B}$ are known only up to a factor of proportionality. Otherwise, the models are equivalent. If we replace $\mathbf{V}$ by $\frac{1}{\phi}\mathbf{V}^*$, where V^* is a **known** positive definite matrix, and ϕ, $0 < \phi < \infty$, is an unknown precision parameter then, by factoring $\mathbf{V}^* = \mathbf{C}^*\mathbf{C}^{*\prime}$, and letting $\mathbf{Y}^* = (\mathbf{C}^*)^{-1}\mathbf{Y}$, $\mathbf{X}^* = (\mathbf{C}^*)^{-1}\mathbf{X}$ we obtain

$$\mathbf{Y}^* \mid \boldsymbol{\beta}, \phi \sim N\left(\mathbf{X}^*\boldsymbol{\beta}, \frac{1}{\phi}\mathbf{I}\right).$$

Similarly, if $\mathbf{B} = \mathbf{D}\mathbf{D}'$ and $\boldsymbol{\beta}^* = \mathbf{D}^{-1}\boldsymbol{\beta}$ then

$$\boldsymbol{\beta}^* \mid \phi \sim N\left(\boldsymbol{\beta}_0^*, \frac{\tau^2}{\phi}\mathbf{I}\right).$$

If $\mathbf{X}^{**} = \mathbf{X}^*\mathbf{D}$ then the previous model, in terms of $\mathbf{Y}^*$ and $\mathbf{X}^{**}$, is reduced to

$$\mathbf{Y}^* = \mathbf{X}^{**}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}^*,$$

where $\mathbf{Y}^* = \mathbf{C}^{-1}\mathbf{Y}$, $\mathbf{X}^* = \mathbf{C}^{-1}\mathbf{X}\mathbf{D}$, $\boldsymbol{\beta}^* = \mathbf{D}^{-1}\boldsymbol{\beta}$, $\mathbf{V} = \mathbf{C}\mathbf{C}'$, and $\mathbf{B} = \mathbf{D}\mathbf{D}'$.

We obtained a linear model generalization of the results of Example 8.17. Indeed,

$$\boldsymbol{\beta} \mid \mathbf{Y}, \phi \sim N(\boldsymbol{\beta}_0 + \mathbf{X}'(\mathbf{I} + \tau^2\mathbf{X}\mathbf{X}')^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0), \frac{1}{\phi}[\mathbf{I} - \mathbf{X}'(\mathbf{I} + \tau^2\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}]).$$

Thus, the Bayesian estimator of $\boldsymbol{\beta}$, for the squared-error loss, $||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||^2$ is

$$\hat{\boldsymbol{\beta}}_B = (\mathbf{I} - \mathbf{X}'(\mathbf{I} + \tau^2\mathbf{X}\mathbf{X}')^{-1}\mathbf{X})\boldsymbol{\beta}_0 + \mathbf{X}'(\mathbf{I} + \tau^2\mathbf{X}\mathbf{X}')\mathbf{Y}.$$

As in Example 8.17, the conditional predictive distribution of $\mathbf{Y}$, given ϕ, is normal,

$$\mathbf{Y} \mid \phi \sim N\left(\mathbf{X}\boldsymbol{\beta}_0, \frac{1}{\phi}(\mathbf{I} + \tau^2\mathbf{X}\mathbf{X}')\right).$$

Hence, the marginal posterior distribution of ϕ, given $\mathbf{Y}$, is the gamma distribution, i.e.,

$$\phi \mid \mathbf{Y} \sim G\left(\frac{1}{2}(\psi + (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0)'(\mathbf{I} + \tau^2\mathbf{X}\mathbf{X}')^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0)), \frac{n + n_0}{2}\right),$$

where n is the dimension of $\mathbf{Y}$. Thus, the Bayesian estimator of ϕ is

$$\hat{\phi}_B = \frac{n + n_0}{\psi + (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0)'(\mathbf{I} + \tau^2\mathbf{X}\mathbf{X}')^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0)}.$$

Finally, if $\psi = n_0$ then the predictive distribution of $\mathbf{Y}$ is the multivariate $t[n_0; \mathbf{X}\boldsymbol{\beta}_0, I + \tau^2\mathbf{X}\mathbf{X}']$, defined in (2.13.12). ■

Example 8.19. The following is a random growth model. We follow the model assumptions of Section 8.4.3:

$$X_t = \theta_{0,t} + \theta_{1,t}t + \epsilon_t, \quad t = 1, 2, \ldots,$$

where $\theta_{0,t}$ and $\theta_{1,t}$ vary at random according to a random-walk model, i.e.,

$$\begin{pmatrix}\theta_{0,t}\\ \theta_{1,t}\end{pmatrix} = \begin{pmatrix}\theta_{0,t-1}\\ \theta_{1,t-1}\end{pmatrix} + \begin{pmatrix}\omega_{0,t}\\ \omega_{1,t}\end{pmatrix}.$$

Thus, let $\boldsymbol{\theta}_t = (\theta_{0,t}, \theta_{1,t})'$ and $\mathbf{a}_t' = (1, t)$. The dynamic linear model is thus

$$\begin{aligned} X_t &= \mathbf{a}_t'\boldsymbol{\theta}_t + \epsilon_t \\ \boldsymbol{\theta}_t &= \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad t = 1, 2, \ldots. \end{aligned}$$

Let $\boldsymbol{\eta}_t$ and C_t be the posterior mean and posterior covariance matrix of $\boldsymbol{\theta}_t$. We obtain the recursive equations

$$\begin{aligned} \boldsymbol{\eta}_t &= \boldsymbol{\eta}_{t-1} + \frac{1}{r_t}(X_t - \mathbf{a}_t'\boldsymbol{\eta}_{t-1})(C_{t-1} + \Omega)\mathbf{a}_t, \\ r_t &= \sigma^2 + \mathbf{a}_t'(C_{t-1} + \Omega)\mathbf{a}_t, \\ C_t &= C_{t-1} + \Omega - \frac{1}{r_t}(C_{t-1} + \Omega)\mathbf{a}_t\mathbf{a}_t'(C_{t-1} + \Omega), \end{aligned}$$

where $\sigma^2 = V\{\epsilon_t\}$ and Ω is the covariance matrix of $\boldsymbol{\omega}_t$. ■

Example 8.20. In order to illustrate the approximations of Section 8.5, we apply them here to a model in which the posterior p.d.f. can be computed exactly. Thus, let $X_1, \ldots, X_n$ be conditionally i.i.d. random variables, having a common Poisson distribution $P(\lambda)$, $0 < \lambda < \infty$. Let the prior distribution of λ be that of a gamma, $G(\Lambda, \alpha)$. Thus, the posterior distribution of λ, given $T_n = \sum_{i=1}^{n} X_i$ is like that of $G(n + \Lambda, \alpha + T_n)$, with p.d.f.

$$h_n(\lambda \mid T_n) = \frac{(n+\Lambda)^{\alpha+T_n}}{\Gamma(\alpha + T_n)}\lambda^{\alpha+T_n-1}e^{-\lambda(n+\Lambda)}, \quad 0 < \lambda < \infty.$$

The MLE of λ is $\hat{\lambda}_n = T_n/n$. In this model, $\mathbf{J}(\hat{\lambda}_n) = n/\bar{X}_n$. Thus, formula (8.5.11) yields the normal approximation

$$h_n^*(\lambda \mid T_n) = \frac{\sqrt{n}}{\sqrt{2\pi \cdot \bar{X}_n}} \exp\left\{-\frac{n}{2\bar{X}_n}(\lambda - \bar{X}_n)^2\right\}.$$

From large sample theory, we know that

$$\frac{G(n+\Lambda, \alpha+T_n) - \hat{\lambda}_n}{\hat{\lambda}_n^{1/2}} \xrightarrow[n\to\infty]{d} N(0, 1).$$

Thus, the approximation to $h(\lambda \mid T_n)$ given by

$$\hat{h}(\lambda \mid T_n) = \frac{\sqrt{n}\left(1+\frac{\Lambda}{n}\right)}{\sqrt{2\pi\left(\bar{X}_n+\frac{\alpha}{n}\right)}} \exp\left\{-\frac{1}{2}\frac{n\left(1+\frac{\Lambda}{n}\right)^2}{\left(\bar{X}_n+\frac{\alpha}{n}\right)}\left(\lambda-\frac{\bar{X}_n+\frac{\alpha}{n}}{1+\frac{\Lambda}{n}}\right)\right\}^2$$

should be better than $h_n^*(\lambda \mid T_n)$, if the sample size is not very large. ■

Example 8.21. We consider again the model of Example 8.20. In that model, for $0 < \lambda < \infty$,

$$\begin{aligned}\tilde{k}(\lambda) &= \lambda - \bar{X}_n \cdot \log\lambda + \lambda\frac{\Lambda}{n} - \frac{(\alpha-1)}{n}\log\lambda \\ &= \lambda\left(1+\frac{\Lambda}{n}\right) - \left(\bar{X}_n + \frac{\alpha-1}{n}\right)\log\lambda.\end{aligned}$$

Thus,

$$\tilde{k}'(\lambda) = \left(1+\frac{\Lambda}{n}\right) - \frac{\bar{X}_n+\frac{\alpha-1}{n}}{\lambda}$$

and the maximizer of $-n\tilde{k}(\lambda)$ is

$$\tilde{\lambda}_n = \frac{\bar{X}_n+\frac{\alpha-1}{n}}{1+\frac{\Lambda}{n}}.$$

Moreover,

$$\tilde{J}(\tilde{\lambda}_n) = \frac{\left(1+\frac{\Lambda}{n}\right)^2}{\bar{X}_n+\frac{\alpha-1}{n}}.$$

The normal approximation, based on $\tilde{\lambda}_n$ and $\tilde{J}(\tilde{\lambda}_n)$, is $N\left(\frac{\bar{X}_n+\frac{\alpha-1}{n}}{1+\frac{\Lambda}{n}}, \frac{\bar{X}_n+\frac{\alpha-1}{n}}{n\left(1+\frac{\Lambda}{n}\right)^2}\right)$. This is very close to the large sample approximation (8.5.15). The only difference is that α, in $\hat{h}(\lambda \mid T_n)$, is replaced by $\alpha' = \alpha - 1$. ■

Example 8.22. We consider again the model of Example 8.3. In that example, $Y_i \mid \lambda_i \sim P(\lambda_i)$, where $\lambda_i = e^{\mathbf{X}_i'\boldsymbol{\beta}}$, $i = 1, \ldots, n$. Let $(X) = (\mathbf{X}_1, \ldots, \mathbf{X}_n)'$ be the $n \times p$ matrix of covariates. The unknown parameter is $\boldsymbol{\beta} \in \mathbb{R}^{(p)}$. The prior distribution of $\boldsymbol{\beta}$ is normal, i.e., $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \tau^2 I)$. The likelihood function is, according to Equation (8.1.8),

$$L(\boldsymbol{\beta}; \mathbf{Y}_n(X)) = \exp\left\{\mathbf{W}_n'\boldsymbol{\beta} - \sum_{i=1}^{n} e^{\mathbf{X}_i'\boldsymbol{\beta}}\right\},$$

where $\mathbf{W}_n = \sum_{i=1}^{n} Y_i \mathbf{X}_i$. The prior p.d.f. is,

$$h(\boldsymbol{\beta}) = \frac{1}{(2\pi)^{p/2}\tau^{p/2}} \exp\left\{-\frac{1}{2\tau^2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\}.$$

Accordingly,

$$\tilde{k}(\boldsymbol{\beta}) = -\frac{1}{n}\left(\mathbf{W}_n'\boldsymbol{\beta} - \sum_{i=1}^{n} e^{\mathbf{x}_i'\boldsymbol{\beta}} - \frac{1}{2\tau^2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right).$$

Hence,

$$\nabla_{\boldsymbol{\beta}}\tilde{k}(\boldsymbol{\beta}) = -\frac{1}{n}\mathbf{W}_n + \frac{1}{n}\sum_{i=1}^{n} e^{\mathbf{x}_i'\boldsymbol{\beta}}\mathbf{x}_i + \frac{1}{n\tau^2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0),$$

and

$$\tilde{\mathbf{J}}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} e^{\mathbf{x}_i'\boldsymbol{\beta}}\mathbf{x}_i\mathbf{x}_i' + \frac{1}{n\tau^2} I.$$

The value $\tilde{\boldsymbol{\beta}}_n$ is the root of the equation

$$\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \tau^2\mathbf{W}_n - \tau^2 \sum_{i=1}^{n} e^{\mathbf{x}_i'\boldsymbol{\beta}}\mathbf{x}_i.$$

Note that

$$\tilde{\mathbf{J}}(\boldsymbol{\beta}) = \frac{1}{n\tau^2}(I + \tau^2(X)'\Delta(\boldsymbol{\beta})(X)),$$

where $\Delta(\boldsymbol{\beta})$ is an $n \times n$ diagonal matrix with ith diagonal element equal to $e^{\boldsymbol{\beta}'\mathbf{X}_i}$ ($i = 1, \ldots, n$). The matrix $\tilde{\mathbf{J}}(\boldsymbol{\beta})$ is positive definite for all $\boldsymbol{\beta} \in \mathbb{R}^{(p)}$. We can determine $\tilde{\boldsymbol{\beta}}_n$ by solving the equation iteratively, starting with the LSE, $\boldsymbol{\beta}_n^* = [(X)'(X)]^{-1}(X)'\mathbf{Y}_n^*$, where Y_n^* is a vector whose ith component is

$$Y_i^* = \begin{cases} \log(Y_i), & \text{if } Y_i > 0, \\ -\exp\{\mathbf{x}_i'\boldsymbol{\beta}_0\}, & \text{if } Y_i = 0, \end{cases} \qquad i = 1, \ldots, n.$$

The approximating p.d.f. for $h(\boldsymbol{\beta} \mid (X), \mathbf{Y}_n)$ is the p.d.f. of the p-variate normal $N\left(\tilde{\boldsymbol{\beta}}_n, \frac{1}{n}(\tilde{\mathbf{J}}(\tilde{\boldsymbol{\beta}}_n))^{-1}\right)$. This p.d.f. will be compared later numerically with a p.d.f. obtained by numerical integration and one obtained by simulation. ■

Example 8.23. Let (X_i, Y_i), $i = 1, \ldots, n$ be i.i.d. random vectors, having a standard bivariate normal distribution, i.e., $(X, Y)' \sim N\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$. The likelihood function of ρ is

$$L(\rho \mid T_n) = \frac{1}{(1-\rho^2)^{n/2}} \exp\left\{-\frac{1}{2(1-\rho^2)}[Q_X - 2\rho P_{XY} + Q_Y]\right\},$$

where $T_n = (Q_X, P_{XY}, Q_Y)$ and $Q_X = \sum_{i=1}^n X_i^2$, $Q_Y = \sum_{i=1}^n Y_i^2$, $P_{XY} = \sum_{i=1}^n X_i Y_i$. The Fisher information function for ρ is

$$I_n(\rho) = n\frac{1+\rho^2}{(1-\rho^2)^2}.$$

Using the Jeffreys prior

$$h(\rho) = \frac{(1+\rho^2)^{1/2}}{1-\rho^2}, \quad -1 < \rho < 1,$$

the Bayesian estimator of ρ for the squared error loss is

$$\hat{\rho}_B = \frac{\displaystyle\int_{-1}^{1} \rho \frac{(1+\rho^2)^{1/2}}{(1-\rho^2)^{\frac{n}{2}+1}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}[Q_X - 2\rho P_{XY} + Q_Y]\right\} d\rho}{\displaystyle\int_{-1}^{1} \frac{(1+\rho^2)^{1/2}}{(1-\rho^2)^{\frac{n}{2}+1}} \cdot \exp\left\{-\frac{1}{2(1-\rho^2)}[Q_X - 2\rho P_{XY} + Q_Y]\right\} d\rho}.$$

This estimator, for given values of Q_X, Q_Y and P_{XY}, can be evaluated accurately by 16-points Gaussian quadrature. For $n = 16$, we get from Table 2.54 of Abramowitz and Stegun (1968) the values

u_i	0.0950	0.2816	0.4580	0.6179	0.7554
ω_i	0.1895	0.1826	0.1692	0.1496	0.1246

u_i	0.8656	0.9446	0.9894
ω_i	0.0952	0.0623	0.0272

For negative values of u, we use $-u$; with the same weight, ω_i. For a sample of size $n = 10$, with $Q_X = 6.1448$, $Q_Y = 16.1983$, and $P_{XY} = 4.5496$, we obtain $\hat{\rho}_B = 0.3349$. ■

Example 8.24. In this example, we consider evaluating the integrals in $\hat{\rho}_B$ of Example 8.23 by simulations. We simulate 100 random variables $U_i \sim R(-1, 1)$ $i = 1, \ldots, 100$, and approximate the integrals in the numerator and denominator of $\hat{\rho}_\theta$ by averages. For $n = 100$, and the same values of Q_X, Q_Y, P_{XY}, as in Example 8.23, we obtain the approximation $\tilde{\rho}_B = 0.36615$. ■

Example 8.25. We return to Example 8.22 and compute the posterior expectation of $\boldsymbol{\beta}$ by simulation. Note that, for a large number M of simulation runs, $E\{\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{Y}\}$ is approximated by

$$\hat{E} = \frac{\sum_{j=1}^{M} \boldsymbol{\beta}_j \exp\left\{\mathbf{w}_n' \boldsymbol{\beta}_j - \sum_{i=1}^{n} e^{\mathbf{x}_i' \boldsymbol{\beta}_j}\right\}}{\sum_{j=1}^{M} \exp\left\{\mathbf{w}_n' \boldsymbol{\beta}_j - \sum_{i=1}^{n} e^{\mathbf{x}_i' \boldsymbol{\beta}_j}\right\}},$$

where $\boldsymbol{\beta}_j$ is a random vector, simulated from the $N(\boldsymbol{\beta}_0, \tau^2 I)$ distribution.

To illustrate the result numerically, we consider a case where the observed sample contains 40 independent observations; ten for each one of the for $\mathbf{x}$ vectors:

$$\mathbf{X}_1' = (1, -1, -1, 1),$$
$$\mathbf{X}_2' = (1, 1, -1, -1),$$
$$\mathbf{X}_3' = (1, -1, 1, -1),$$
$$\mathbf{X}_4' = (1, 1, 1, 1).$$

The observed values of $\mathbf{w}_{40}$ is $(6, 3, 11, 29)$. For $\boldsymbol{\beta}_0' = (0.1, 0.1, 0.5, 0.5)$, we obtain the following Bayesian estimators $\hat{E}$, with $M = 1000$,

Simulation			Normal Approximation		
$\hat{\boldsymbol{\beta}}$	$\tau = 0.01$	$\tau = 0.05$	$\tilde{\boldsymbol{\beta}}$	$\tau = 0.01$	$\tau = 0.05$
$\hat{\boldsymbol{\beta}}_1$	0.0946	0.0060	$\tilde{\boldsymbol{\beta}}_1$	0.0965	0.0386
$\hat{\boldsymbol{\beta}}_2$	0.0987	0.0842	$\tilde{\boldsymbol{\beta}}_2$	0.0995	0.0920
$\hat{\boldsymbol{\beta}}_3$	0.4983	0.4835	$\tilde{\boldsymbol{\beta}}_3$	0.4970	0.4510
$\hat{\boldsymbol{\beta}}_4$	0.5000	0.5021	$\tilde{\boldsymbol{\beta}}_4$	0.4951	0.4038

We see that when $\tau = 0.01$ the Bayesian estimators are very close to the prior mean $\boldsymbol{\beta}_0$. When $\tau = 0.05$ the Bayesian estimators might be quite different than $\boldsymbol{\beta}_0$. In Example 8.22, we approximated the posterior distribution of $\boldsymbol{\beta}$ by a normal distribution. For the values in this example the normal approximation yields similar results to those of the simulation. ■

Example 8.26. Consider the following repetitive problem. In a certain manufacturing process, a lot of N items is produced every day. Let M_j, $j = 1, 2, \ldots$, denote the number of defective items in the lot of the jth day. The parameters $M_1, M_2, \ldots$ are unknown. At the end of each day, a random sample of size n is selected without replacement from that day's lot for inspection. Let X_j denote the number of defectives observed in the sample of the jth day. The distribution of X_j is the hypergeometric $H(N, M_j, n)$, $j = 1, 2, \ldots$. Samples from different days are (conditionally) independent (given $M_1, M_2, \ldots$). In this problem, it is often reasonable to assume that the parameters $M_1, M_2, \ldots$ are independent random variables having the same binomial distribution $B(N, \theta)$. θ is the probability of defectives in the production process. It is assumed that θ does not change in time. The value of θ is, however, unknown. It is simple to verify that for a prior $B(N, \theta)$ distribution of M, and a squared-error loss function, the Bayes estimator of M_j is

$$\hat{M}_j = X_j + (N - n)\theta.$$

The corresponding Bayes risk is

$$\rho(\theta) = (N - n)\theta(1 - \theta).$$

A sequence of empirical Bayes estimators is obtained by substituting in $\hat{M}_j$ a consistent estimator of θ based on the results of the first $(j - 1)$ days. Under the above assumption on the prior distribution of $M_1, M_2, \ldots$, the predictive distribution of $X_1, X_2, \ldots$ is the binomial $B(n, \theta)$. A priori, for a given value of θ, $X_1, X_2, \ldots$ can be considered as i.i.d. random variables having the mixed distribution $B(n, \theta)$. Thus,

$\hat{p}_{j-1} = \dfrac{1}{n(j-1)} \sum_{i=1}^{j-1} X_i$, for $j \geq 2$, is a sequence of consistent estimators of θ. The corresponding sequence of empirical Bayes estimators is

$$\hat{\hat{M}}_j = X_j + (N-n)\hat{p}_{j-1}, \quad j \geq 2.$$

The posterior risk of $\hat{\hat{M}}_j$ given $(X_j, \hat{p}_{j-1})$ is

$$\begin{aligned}
\rho_j(\hat{\hat{M}}_j) &= E\{[M_j - \hat{\hat{M}}_j]^2 \mid \hat{p}_{j-1}, X_j\} \\
&= E\{[M_j - \hat{M}_j]^2 \mid \hat{p}_{j-1}, X_j\} + (\hat{\hat{M}}_j - \hat{M}_j)^2 \\
&= (N-n)\theta(1-\theta) + (\hat{\hat{M}}_j - \hat{M}_j)^2.
\end{aligned}$$

We consider now the conditional expectation of $\rho_j(\hat{\hat{M}}_j)$ given X_j. This is given by

$$\begin{aligned}
E\{\rho_j(\hat{\hat{M}}_j) \mid X_j\} &= (N-n)\theta(1-\theta) + (N-n)^2 E\{[\hat{p}_{j-1} - \theta]^2\} \\
&= (N-n)\theta(1-\theta)\left[1 + \frac{N-n}{n(j-1)}\right].
\end{aligned}$$

Notice that this converges as $j \to \infty$ to $\rho(\theta)$. ■

Example 8.27. This example shows the application of empirical Bayes techniques to the simultaneous estimation of many probability vectors. The problem was motivated by a problem of assessing the readiness probabilities of military units based on exercises of big units. For details, see Brier, Zacks, and Marlow (1986).

A large number, N, of units are tested independently on tasks that are classified into K categories. Each unit obtains on each task the value 1 if it is executed satisfactorily and the value 0 otherwise. Let $i, i = 1, \ldots, N$ be the index of the ith unit, and $j, j = 1, \ldots, k$ the index of a category. Unit i was tested on M_{ij} tasks of category j. Let X_{ij} denote the number of tasks in category j on which the ith unit received a satisfactory score. Let θ_{ij} denote the probability of the ith unit executing satisfactorily a task of category j. There are N parametric vectors $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iK})', i = 1, \ldots, N$, to be estimated.

The model is that, conditional on $\boldsymbol{\theta}_i$, $X_{i1}, \ldots, X_{iK}$ are independent random variables, having binomial distributions, i.e.,

$$X_{ij} \mid \theta_{ij} \sim \mathcal{B}(M_{ij}, \theta_{ij}).$$

In addition, the vectors $\boldsymbol{\theta}_i$ $(i = 1, \ldots, N)$ are i.i.d. random variables, having a common distribution. Since M_{ij} were generally large, but not the same, we have used first the variance stabilizing transformation

$$Y_{ij} = 2\sin^{-1}\left(\sqrt{\frac{X_{ij} + 3/8}{M_{ij} + 3/4}}\right), \quad i = 1, \ldots, N, \quad j = 1, \ldots, K.$$

For large values of M_{ij} the asymptotic distribution of Y_{ij} is $N(\eta_{ij}, 1/M_{ij})$, where $\eta_{ij} = 2\sin^{-1}(\sqrt{\theta_{ij}})$, as shown in Section 7.6.

Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iK})'$, $i = 1, \ldots, N$. The parametric empirical Bayes model is that $(\mathbf{Y}_i, \boldsymbol{\theta}_i)$ are i.i.d, $i = 1, \ldots, N$,

$$\mathbf{Y}_i \mid \boldsymbol{\theta}_i \sim N(\boldsymbol{\eta}_i, D_i), \quad i = 1, \ldots, N$$

and

$$\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_N \text{ are i.i.d. } N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iK})'$ and

$$D_i = \begin{bmatrix} \frac{1}{M_{i1}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{M_{iK}} \end{bmatrix}, \quad i = 1, \ldots, N.$$

The prior parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown. Note that if $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given then $\boldsymbol{\eta}, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_N$ are a posteriori independent, given $Y_1, \ldots, Y_N$. Furthermore, the posterior distribution of $\boldsymbol{\eta}_i$, given $\mathbf{Y}_i$, is

$$\boldsymbol{\eta}_i \mid \mathbf{Y}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N(\mathbf{Y}_i - \mathbf{B}_i(\mathbf{Y}_i - \boldsymbol{\mu}), (I - \mathbf{B}_i)\mathbf{D}_i),$$

where

$$\mathbf{B}_i = \mathbf{D}_i(\mathbf{D}_i + \boldsymbol{\Sigma})^{-1}, \quad i = 1, \ldots, N.$$

Thus, if $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given, the Bayesian estimator of $\boldsymbol{\eta}_i$, for a squared-error loss function $L(\boldsymbol{\eta}_i, \hat{\boldsymbol{\eta}}_i) = ||\hat{\boldsymbol{\eta}}_i - \boldsymbol{\eta}_i||^2$, is the posterior mean, i.e.,

$$\hat{\boldsymbol{\eta}}_i(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = B_i\boldsymbol{\mu} + (I - \mathbf{B}_i)\mathbf{Y}_i, \quad i = 1, \ldots, N.$$

The empirical Bayes method estimates $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from all the data. We derive now MLEs of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. These MLEs are then substituted in $\hat{\eta}_i(\mu, \boldsymbol{\Sigma})$ to yield empirical Bayes estimators of $\boldsymbol{\eta}_i$.

Note that $\mathbf{Y}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \mathbf{D}_i)$, $i = 1, \ldots, N$. Hence, the log-likelihood function of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, given the data $(\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$, is

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2}\sum_{i=1}^{N} \log|\boldsymbol{\Sigma} + \mathbf{D}_i| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{Y}_i - \boldsymbol{\mu})'(\boldsymbol{\Sigma} + \mathbf{D}_i)^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}).$$

The vector $\hat{\boldsymbol{\mu}}(\boldsymbol{\Sigma})$, which maximizes $l(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, for a given $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\mu}}(\boldsymbol{\Sigma}) = \left(\sum_{i=1}^{N}(\boldsymbol{\Sigma} \mathbf{D}_i)^{-1}\right)^{-1} \sum_{i=1}^{N}(\boldsymbol{\Sigma} + \mathbf{D}_i)^{-1}\mathbf{Y}_i.$$

Substituting $\hat{\mu}(\boldsymbol{\Sigma})$ in $l(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and finding $\boldsymbol{\Sigma}$ that maximizes the expression can yield the MLE $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$. Another approach, to find the MLE, is given by the *E–M algorithm*. The *E-M* algorithm considers the unknown parameters $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_N$ as missing data. The algorithm is an iterative process, having two phases in each iteration. The first phase is the *E*-phase, in which the conditional expectation of the likelihood function is determined, given the data and the current values of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In the next phase, the *M*-phase, the conditionally expected likelihood is maximized by determining the maximizing arguments $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$. More specifically, let

$$l^*(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_N, Y_1, \ldots, Y_N) = -\frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{\eta}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\eta}_i - \boldsymbol{\mu})$$

be the log-likelihood of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_N$ were known. Let $(\boldsymbol{\mu}^{(p)}, \boldsymbol{\Sigma}^{(p)})$ be the estimator of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ after p iterations, $p \geq 0$, where $\boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}$ are initial estimates.

In the $(p+1)$st iteration, we start (the *E*-phase) by determining

$$\begin{aligned} l^{**}(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_n, \boldsymbol{\mu}^{(p)}, \boldsymbol{\Sigma}^{(p)}) &= E\{l^*(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_N, \mathbf{Y}_1, \ldots, \mathbf{Y}_n) \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \boldsymbol{\mu}^{(p)}, \boldsymbol{\Sigma}^{(p)}\} \\ &= -\frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{N} E\{(\boldsymbol{\eta}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\eta}_i - \boldsymbol{\mu}) \mid \mathbf{Y}_i, \boldsymbol{\mu}^{(p)}, \boldsymbol{\Sigma}^{(p)}\}, \end{aligned}$$

where the conditional expectation is determined as though $\boldsymbol{\mu}^{(p)}$ and $\boldsymbol{\Sigma}^{(p)}$ are the true values. It is well known that if $E\{\mathbf{X}\} = \boldsymbol{\xi}$ and the covariance matrix of $\mathbf{X}$ is $\mathcal{C}(\mathbf{X})$ then $E\{\mathbf{X}'A\mathbf{X}\} = \boldsymbol{\mu}'A\boldsymbol{\mu} + \text{tr}\{\mathbf{AC}(\mathbf{X})\}$, where $\text{tr}\{\cdot\}$ is the trace of the matrix (see, Seber, 1977, p. 13). Thus,

$$\begin{aligned} E\{(\boldsymbol{\eta}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\eta}_i - \boldsymbol{\mu}) \mid \mathbf{Y}_i, \boldsymbol{\mu}^{(p)}, \boldsymbol{\Sigma}^{(p)}\} &= (\mathbf{W}_i^{(p)} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{W}_i^{(p)} - \boldsymbol{\mu}) \\ &\quad + \text{tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{V}_i^{(p)}\}, \end{aligned}$$

where, $\mathbf{W}_i^{(p)} = \hat{\boldsymbol{\eta}}(\boldsymbol{\mu}^{(p)}, \boldsymbol{\Psi}^{(p)})$, in which $\mathbf{B}_i^{(p)} = \mathbf{D}_i(\mathbf{D}_i + \boldsymbol{\Psi}^{(p)})^{-1}$, and $\mathbf{V}_i^{(p)} = \mathbf{B}_i^{(p)}\boldsymbol{\Psi}^{(p)}$, $i = 1, \ldots, N$. Thus,

$$l^{**}(\boldsymbol{\mu}, \boldsymbol{\Psi} \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_n, \boldsymbol{\mu}^{(p)}, \boldsymbol{\Psi}^{(p)}) = -\frac{N}{2}\log|\boldsymbol{\Psi}| - \frac{N}{2}\text{tr.}\{\boldsymbol{\Psi}^{-1}\bar{\mathbf{V}}^{(p)}\} - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{W}_i^{(p)} - \boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\mathbf{W}_i^{(p)} - \boldsymbol{\mu}),$$

where $\bar{\mathbf{V}}^{(p)} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{V}_i^{(p)}$. In phase-$M$, we determine $\boldsymbol{\mu}^{(p+1)}$ and $\boldsymbol{\Psi}^{(p+1)}$ by maximizing $l^{**}(\boldsymbol{\mu}, \boldsymbol{\Psi} \mid \cdots)$.

One can immediately verify that

$$\boldsymbol{\mu}^{(p+1)} = \bar{\mathbf{W}}^{(p)} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{W}_i^{(p)}.$$

Moreover,

$$l^{**}(\boldsymbol{\mu}^{(p+1)}, \boldsymbol{\Psi} \mid \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \boldsymbol{\mu}^{(p)}, \boldsymbol{\Psi}^{(p)}) = -\frac{N}{2}\log|\boldsymbol{\Psi}| - \frac{N}{2}\text{tr.}\{\boldsymbol{\Psi}^{-1}(\mathbf{C}^{(p)} + \bar{\mathbf{V}}^{(p)})\},$$

where $\mathbf{C}^{(p)} = (C_{jj'}^{(p)}; j, j' = 1, \ldots, K)$, and

$$C_{jj'}^{(p)} = \frac{1}{N}\sum_{i=1}^{N}(W_{ij}^{(p)} - \bar{W}_j^{(p)})(W_{ij'}^{(p)} - \bar{W}_{j'}^{(p)}).$$

Finally, the matrix maximizing l^{**} is

$$\boldsymbol{\Psi}^{(p+1)} = \mathbf{C}^{(p)} + \bar{\mathbf{V}}^{(p)}.$$

We can prove recursively, by induction on p, that

$$\mathbf{W}^{(0)} = \bar{\mathbf{Y}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{Y}_i,$$

$$\mathbf{W}^{(p+1)} = \bar{\mathbf{Y}} - \sum_{j=0}^{p}\bar{\mathbf{B}}^{(j)}\mathbf{K},$$

where $\bar{\mathbf{B}}^{(l)} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{B}_i^{(l)}$, $l = 0, 1, \ldots$, and

$$\mathbf{K} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{B}_i\mathbf{Y}_i - \bar{\mathbf{B}}\bar{\mathbf{Y}}).$$

Thus,

$$\begin{aligned}\lim_{p\to\infty}\bar{W}^{(p)} &= \bar{\mathbf{Y}} = (I - \bar{\mathbf{B}})^{-1}\mathbf{K} \\ &= \left(\sum_{i=1}^{N}(I - \mathbf{B}_i)\right)^{-1}\sum_{i=1}^{N}(I - \mathbf{B}_i)\mathbf{Y}_i.\end{aligned}$$

One continues iterating until $\boldsymbol{\mu}^{(p)}$ and $\boldsymbol{\Sigma}^{(p)}$ do not change significantly.

Brier, Zacks, and Marlow (1986) studied the efficiency of these empirical Bayes estimators, in comparison to the simple MLE, and to another type of estimator that will be discussed in Chapter 9. ■

PART III: PROBLEMS

Section 8.1

8.1.1 Let $\mathcal{F} = \{B(N, \theta); 0 < \theta < 1\}$ be the family of binomial distributions.

(i) Show that the family of beta prior distributions $\mathcal{H} = \{\beta(p, q); 0 < p, q < \infty\}$ is conjugate to $\mathcal{F}$.

(ii) What is the posterior distribution of θ given a sample of n i.i.d. random variables, having a distribution in $\mathcal{F}$?

(iii) What is the predicted distribution of X_{n+1}, given $(X_1, \ldots, X_n)$?

8.1.2 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a Pareto distribution, with p.d.f.

$$f(x; \nu) = \nu A^{\nu}/x^{\nu+1}, \quad 0 < A < x < \infty;$$

$0 < \nu < \infty$ (A is a specified positive constant).

(i) Show that the geometric mean, $G = \left(\prod_{i=1}^{n}X_i\right)^{1/n}$, is a minimal sufficient statistic.

(ii) Suppose that ν has a prior $G(\lambda, p)$ distribution. What is the posterior distribution of ν given $\mathbf{X}$?

(iii) What are the posterior expectation and posterior variance of ν given $\mathbf{X}$?

8.1.3 Let $\mathbf{X}$ be a p-dimensional vector having a multinormal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suppose that $\boldsymbol{\Sigma}$ is known and that $\boldsymbol{\mu}$ has a prior normal distribution $N(\boldsymbol{\mu}_0, \mathbf{V})$. What is the posterior distribution of $\boldsymbol{\mu}$ given $\mathbf{X}$?

8.1.4 Apply the results of Problem 3 to determine the posterior distribution of $\boldsymbol{\beta}$ in the normal multiple regression model of full rank, when σ^2 is known. More specifically, let

$$\mathbf{X} \sim N(\mathbf{A}\boldsymbol{\beta}, \sigma^2 I) \text{ and } \boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \mathbf{V}).$$

(i) Find the posterior distribution of $\boldsymbol{\beta}$ given $\mathbf{X}$.

(ii) What is the predictive distribution of $\mathbf{X}_2$ given $\mathbf{X}_1$, assuming that conditionally on $\boldsymbol{\beta}$, $\mathbf{X}_1$ and $\mathbf{X}_2$ are i.i.d.?

8.1.5 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a Poisson distribution, $P(\lambda)$, $0 < \lambda < \infty$. Compute the posterior probability $P\{\lambda \geq \bar{X}_n \mid \mathbf{X}_n\}$ corresponding to the Jeffreys prior.

8.1.6 Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables having a $N(0, \sigma^2)$ distribution, $0 < \sigma^2 < \infty$.

(i) Show that if $1/2\sigma^2 \sim G\left(\frac{1}{\tau}, \nu\right)$ then the posterior distribution of $1/2\sigma^2$ given $S = \Sigma X_i^2$ is also a gamma distribution.

(ii) What is $E\{\sigma^2 \mid S\}$ and $V\{\sigma^2 \mid S\}$ according to the above Bayesian assumptions.

8.1.7 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a $N(\mu, \sigma^2)$ distribution; $-\infty < \mu < \infty$, $0 < \sigma < \infty$. A normal-inverted gamma prior distribution for μ, σ^2 assumes that the conditional prior distribution of μ, given σ^2, is $N(\mu_0, \lambda\sigma^2)$ and that the prior distribution of $1/2\sigma^2$ is a $G\left(\frac{1}{\tau}, \nu\right)$. Derive the posterior joint distribution of (μ, σ^2) given $(\bar{X}, S^2)$, where $\bar{X}$ and S^2 are the sample mean and variance, respectively.

8.1.8 Consider again Problem 7 assuming that μ and σ^2 are priorly independent, with $\mu \sim N(\mu_0, D^2)$ and $1/2\sigma^2 \sim G\left(\frac{1}{\tau}, \nu\right)$. What are the posterior expectations and variances of μ and of σ^2, given $(\bar{X}, S^2)$?

8.1.9 Let $X \sim B(n, \theta)$, $0 < \theta < 1$. Suppose that θ has a prior beta distribution $\beta(\frac{1}{2}, \frac{1}{2})$. Suppose that the loss function associated with estimating θ by $\hat{\theta}$ is $L(\theta, \hat{\theta})$. Find the risk function and the posterior risk when $L(\hat{\theta}, \theta)$ is

(i) $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$;

(ii) $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2/\theta(1 - \theta)$;

(iii) $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$.

8.1.10 Consider a decision problem in which θ can assume the values in the set $\{\theta_0, \theta_1\}$. If i.i.d. random variables $X_1, \ldots, X_n$ are observed, their joint p.d.f. is $f(\mathbf{x}_n; \theta)$ where θ is either θ_0 or θ_1. The prior probability of $\{\theta = \theta_0\}$ is η. A statistician takes actions a_0 or a_1. The loss function associated with this decision problem is given by

$$L(\theta, a_0) = \begin{cases} 0, & \text{if } \theta = \theta_0, \\ K, & \text{if } \theta = \theta_1, \end{cases}$$

$$L(\theta, a_1) = \begin{cases} 1, & \text{if } \theta = \theta_0, \\ 0, & \text{if } \theta = \theta_1. \end{cases}$$

(i) What is the prior risk function, if the statistician takes action a_0 with probability ξ?

(ii) What is the posterior risk function?

(iii) What is the optimal action with no observations?

(iv) What is the optimal action after n observations?

8.1.11 The time till failure of an electronic equipment, T, has an exponential distribution, i.e., $T \sim G\left(\frac{1}{\tau}, 1\right)$; $0 < \tau < \infty$. The mean-time till failure, τ, has an inverted gamma prior distribution, $1/\tau \sim G(\Lambda, \nu)$. Given n observations on i.i.d. failure times $T_1, \ldots, T_n$, the action is an estimator $\hat{\tau}_n = t(T_1, \ldots, T_n)$. The loss function is $L(\hat{\tau}, \tau) = |\hat{\tau}_n - \tau|$. Find the posterior risk of $\hat{\tau}_n$. Which estimator will minimize the posterior risk?

Section 8.2

8.2.1 Let X be a random variable having a Poisson distribution, $P(\lambda)$. Consider the problem of testing the two simple hypotheses $H_0 : \lambda = \lambda_0$ against $H_1 : \lambda = \lambda_1$; $0 < \lambda_0 < \lambda_1 < \infty$.

(i) What is the form of the Bayes test $\phi_\pi(X)$, for a prior probability π of H_0 and costs c_1 and c_2 for errors of types I and II.

(ii) Show that $R_0(\pi) = c_1(1 - P([\xi(\pi)]; \lambda_0))$, where $P(j; \lambda)$ is the c.d.f. of $P(\lambda)$, $R_1(\pi) = c_2 P([\xi(\pi)]; \lambda_1)$,

$$\xi(\pi) = \left(\log \frac{\pi}{1-\pi} + \log \frac{c_1}{c_2} + \lambda_1 - \lambda_0 \right) / \log \left(\frac{\lambda_1}{\lambda_0} \right),$$

where $[x]$ is the largest integer not exceeding x.

(iii) Compute $(R_0(\pi), R_1(\pi))$ for the case of $c_1 = 1$, $c_2 = 3$; $\lambda_1/\lambda_0 = 2$, $\lambda_1 - \lambda_0 = 2$, and graph the lower boundary of the risk set R.

8.2.2 Let $X_1, \ldots, X_n$ be i.i.d. random variables having an exponential distribution $G(\lambda, 1)$, $0 < \lambda < \infty$. Consider the two composite hypotheses $H_0 : \lambda \leq \lambda_0$

against $H_1 : \lambda > \lambda_0$. The prior distribution of λ is $G\left(\frac{1}{\tau}, \nu\right)$. The loss functions associated with accepting H_i $(i = 0, 1)$ are

$$L_0(\lambda) = \begin{cases} 0, & \lambda \leq \lambda_0, \\ \lambda, & \lambda > \lambda_0, \end{cases}$$

and

$$L_1(\lambda) = \begin{cases} B\left(1 - \frac{\lambda}{\lambda_0}\right), & \lambda \leq \lambda_0, \\ 0, & \lambda < \lambda_0, \end{cases}$$

$0 < B < \infty$.

(i) Determine the form of the Bayes test of H_0 against H_1.

(ii) What is the Bayes risk?

8.2.3 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a binomial distribution $B(1, \theta)$. Consider the two composite hypotheses $H_0 : \theta \leq 1/2$ against $H_1 : \theta > 1/2$. The prior distribution of θ is $\beta(p, q)$. Compute the Bayes Factor in favor of H_1 for the cases of $n = 20$, $T = 15$ and

(i) $p = q = 1/2$ (Jeffreys prior);

(ii) $p = 1, q = 3$.

8.2.4 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a $N(\mu, \sigma^2)$ distribution. Let $Y_1, \ldots, Y_m$ be i.i.d. random variables having a $N(\eta, \rho\sigma^2)$ distribution, $-\infty < \mu, \eta < \infty$, $0 < \sigma^2, \rho < \infty$, The X-sample is independent of the Y-sample. Consider the problem of testing the hypothesis $H_0 : \rho \leq 1$, (μ, η, σ^2) arbitrary against $H_1 : \rho > 1$, (μ, η, σ^2) arbitrary. Determine the form of the Bayes test function for the formal prior p.d.f. $h(\mu, \eta, \sigma, \rho) \propto \frac{1}{\sigma^2} \cdot \frac{1}{\rho}$ and a loss function with $c_1 = c_2 = 1$.

8.2.5 Let $\mathbf{X}$ be a k-dimensional random vector having a multinomial distribution $M(n; \boldsymbol{\theta})$. We consider a Bayes test of $H_0 : \boldsymbol{\theta} = \frac{1}{k}\mathbf{1}$ against $H_1 : \boldsymbol{\theta} \neq \frac{1}{k}\mathbf{1}$. Let θ have a prior symmetric Dirichlet distribution (8.2.27) with $\nu = 1$ or 2 with equal hyper-prior probabilities.

(i) Compute the Bayes Factor in favor of H_1 when $k = 5, n = 50, X_1 = 7, X_2 = 12, X_3 = 9, X_4 = 15$, and $X_5 = 7$. [Hint: Approximate the values of $\Gamma(\nu k + n)$ by the Stirling approximation: $n! \approx e^{-n} n^n \sqrt{2\pi n}$, for large n.]

(ii) Would you reject H_0 if $c_1 = c_2 = 1$?

8.2.6 Let $X_1, X_2, \ldots$ be a sequence of i.i.d. normal random variables, $N(0, \sigma^2)$. Consider the problem of testing $H_0 : \sigma^2 = 1$ against $H_1 : \sigma^2 = 2$ sequentially. Suppose that $c_1 = 1$ and $c_2 = 5$, and the cost of observation is $c = 0.01$.

(i) Determine the functions $\rho^{(n)}(\pi)$ for $n = 1, 2, 3$.

(ii) What would be the Chernoff approximation to the SPRT boundaries (A, B)?

8.2.7 Let $X_1, X_2, \ldots$ be a sequence of i.i.d binomial random variables, $B(1, \theta)$, $0 < \theta < 1$. According to $H_1 : \theta = 0.3$. According to $H_1 : \theta = 0.7$. Let π, $0 < \pi < 1$, be the prior probability of H_0. Suppose that the cost for erroneous decision is $b = 10$[\$] (either type of error) and the cost of observation is $c = 0.1$[\$]. Derive the Bayes risk functions $\rho^{(i)}(\pi)$, $i = 1, 2, 3$ and the associated decision and stopping rules of the Bayes sequential procedure.

Section 8.3

8.3.1 Let $X \sim B(n, \theta)$ be a binomial random variable. Determine a $(1 - \alpha)$-level credibility interval for θ with respect to the Jeffreys prior $h(\theta) \propto \theta^{-1/2}(1 - \theta)^{-1/2}$, for the case of $n = 20$, $X = 17$, and $\alpha = 0.05$.

8.3.2 Consider the normal regression model (Problem 3, Section 2.9). Assume that σ^2 is known and that (α, β) has a prior bivariate normal distribution $N(\binom{\alpha_0}{\beta_0}, \Sigma)$.

(i) Derive the $(1 - \varepsilon)$ joint credibility region for (α, β).

(ii) Derive a $(1 - \varepsilon)$ credibility interval for $\alpha + \beta\xi_0$, when ξ_0 is specified.

(iii) What is the $(1 - \varepsilon)$ simultaneous credibility interval for $\alpha + \beta\xi$, for all ξ?

8.3.3 Consider Problem 4 of Section 8.2. Determine the $(1 - \alpha)$ upper credibility limit for the variance ratio ρ.

8.3.4 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a $N(\mu, \sigma^2)$ distribution and let $Y_1, \ldots, Y_n$ be i.i.d. random variables having a $N(\eta, \sigma^2)$ distribution. The Xs and the Ys are independent. Assume the formal prior for μ, η, and σ, i.e.,

$$h(\mu, \eta, \sigma) \propto \sigma^2, \quad -\infty < \mu, \ \eta < \infty, \ 0 < \sigma^2 < \infty.$$

(i) Determine a $(1 - \alpha)$ HPD-interval for $\delta = \mu - \eta$.

(ii) Determine a $(1 - \alpha)$ HPD-interval for σ^2.

8.3.5 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a $G(\lambda, 1)$ distribution and let $Y_1, \ldots, Y_m$ be i.i.d. random variables having a $G(\eta, 1)$ distribution. The

Xs and Ys are independent. Assume that λ and η are priorly independent having prior distributions $G\left(\frac{1}{\tau_1}, \nu_1\right)$ and $G\left(\frac{1}{\tau_2}, \nu_2\right)$, respectively. Determine a $(1-\alpha)$ HPD-interval for $\omega = \lambda/\eta$.

Section 8.4

8.4.1 Let $X \sim B(n, \theta)$, $0 < \theta < 1$. Suppose that the prior distribution of θ is $\beta(p, q)$, $0 < p, q < \infty$.

(i) Derive the Bayes estimator of θ for the squared-error loss.

(ii) What is the posterior risk and the prior risk of the Bayes estimator of (i)?

(iii) Derive the Bayes estimator of θ for the quadratic loss $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2/\theta(1-\theta)$, and the Jeffreys prior $\left(p = q = \frac{1}{2}\right)$.

(iv) What is the prior and the posterior risk of the estimator of (iii)?

8.4.2 Let $X \sim P(\lambda)$, $0 < \lambda < \infty$. Suppose that the prior distribution of λ is $G\left(\frac{1}{\tau}, \nu\right)$.

(i) Derive the Bayes estimator of λ for the loss function $L(\hat{\lambda}, \lambda) = a(\hat{\lambda} - \lambda)^+ + b(\hat{\lambda} - \lambda)^-$, where $(\cdot)^+ = \max(\cdot, \ )$ and $(\cdot)^- = -\min(\cdot, 0)$; $0 < a, b < \infty$.

(ii) Derive the Bayes estimator for the loss function $L(\hat{\lambda}, \lambda) = (\hat{\lambda} - \lambda)^2/\lambda$.

(iii) What is the limit of the Bayes estimators in (ii) when $\nu \to \frac{1}{2}$ and $\tau \to \infty$.

8.4.3 Let $X_1, \ldots, X_n, Y$ be i.i.d. random variables having a normal distribution $N(\mu, \sigma^2)$; $-\infty < \mu < \infty$, $0 < \sigma^2 < \infty$. Consider the Jeffreys prior with $h(\mu, \sigma^2) d\mu d\sigma^2 \propto d\mu d\sigma^2/\sigma^2$. Derive the γ-quantile of the predictive distribution of Y given $(X_1, \ldots, X_n)$.

8.4.4 Let $X \sim P(\lambda)$, $0 < \lambda < \infty$. Derive the Bayesian estimator of λ with respect to the loss function $L(\hat{\lambda}, \lambda) = (\hat{\lambda} - \lambda)^2/\lambda$, and a prior gamma distribution.

8.4.5 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a $B(1, \theta)$ distribution, $0 < \theta < 1$. Derive the Bayesian estimator of θ with respect to the loss function $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2/\theta(1-\theta)$, and a prior beta distribution.

8.4.6 In continuation of Problem 5, show that the posterior risk of $\hat{\theta} = \Sigma X/n$ with respect to $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2/\theta(1-\theta)$ is $1/n$ for all ΣX_i. This implies that the best sequential sampling procedure for this Bayes procedure is a fixed sample procedure. If the cost of observation is c, determine the optimal sample size.

8.4.7 Consider the normal-gamma linear model $NG\left(\boldsymbol{\beta}_0, \tau^2\frac{u_0}{2}, \frac{\psi}{2}\right)$ where $\mathbf{Y}$ is four-dimensional and

$$(\mathbf{X}) = \begin{bmatrix} 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

(i) What is the predictive distribution of $\mathbf{Y}$?

(ii) What is the Bayesian estimator $\hat{\phi}_B$?

8.4.8 As an alternative to the hierarchical model of Gelman et al. (1995) described in Section 8.4.2, assume that $n_1, \ldots, n_k$ are large. Make the variance stabilizing transformation

$$Y_i = 2\sin^{-1}\sqrt{\frac{J_i + 3/8}{n_i + 3/4}}, \quad i = 1, \ldots, k.$$

We consider now the normal model

$$\mathbf{Y} \mid \boldsymbol{\theta} \sim N(\boldsymbol{\eta}, D),$$

where $\mathbf{Y} = (Y_1, \ldots, Y_k)'$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)'$ with $\eta_i = 2\sin^{-1}(\sqrt{\theta_i})$, $i = 1, \ldots, k$. Moreover, D is a diagonal matrix, $D = \text{diag}\left\{\frac{1}{n_i}, i = 1, \ldots, k\right\}$. Assume a prior multinormal distribution for $\boldsymbol{\eta}$.

(i) Develop a credibility region for $\boldsymbol{\eta}$, and by inverse transformations (1:1) obtain credibility region for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.

8.4.9 Consider the normal random walk model, which is a special case of the dynamic linear model (8.4.6), given by

$$\begin{aligned} Y_n &= \theta_n + \epsilon_n, \\ \theta_n &= \theta_{n-1} + \omega_n, \quad n = 1, 2, \ldots, \end{aligned}$$

where $\theta_0 \sim N(\eta_0, c_0)$, $\{\epsilon_n\}$ are i.i.d. $N(0, \sigma^2)$ and $\{\omega_n\}$ are i.i.d. $N(0, \tau^2)$. Show that $\lim_{n\to\infty} c_n \to c^*$, where c_n is the posterior variance of θ_n, given $(Y_1, \ldots, Y_n)$. Find the formula of c^*.

Section 8.5

8.5.1 The integral

$$I = \int_0^1 e^{-\theta}\theta^X(1-\theta)^{n-X}d\theta$$

can be computed analytically or numerically. Analytically, it can be computed as

(i)

$$I = \sum_{j=0}^{\infty} \frac{(-1)^j}{j!} \int_0^1 \theta^{X+j}(1-\theta)^{n-X} d\theta$$

$$= \sum_{j=0}^{\infty} \frac{(-1)^j}{j!} B(X+j+1, n-X+1).$$

(ii) Make the transformation $\omega = \theta/(1-\theta)$, and write

$$I = \int_0^{\infty} \frac{e^{-\omega/(1+\omega)}}{(1+\omega)^2} \exp\{-n(\log(1+\omega) - \hat{p}_n \log \omega)\} d\omega,$$

where $\hat{p}_n = X/n$. Let $k(\omega) = \log(1+\omega) - \hat{p}_n \log(\omega)$ and $f(\omega) = \exp\left\{-\frac{\omega}{1+\omega}\right\}/(1+\omega)^2$. Find $\hat{\omega}$ which maximizes $-k(\omega)$. Use (8.5.3) and (8.5.4) to approximate I.

(iii) Approximate I numerically. Compare (i), (ii) and (iii) for $n = 20, 50$ and $X = 15, 37$. How good is the saddle-point approximation (8.5.9)?

8.5.2 Prove that if U_1, U_2 are two i.i.d. rectangular (0, 1) random variables then the Box Muller transformation (8.5.26) yields two i.i.d. $N(0, 1)$ random variables.

8.5.3 Consider the integral I of Problem [1]. How would you approximate I by simulation? How would you run the simulations so that, with probability ≥ 0.95, the absolute error is not greater than 1% of I.

Section 8.6

8.6.1 Let $(X_1, \theta_1), \ldots, (X_n, \theta_n), \ldots$ be a sequence of independent random vectors of which only the Xs are observable. Assume that the conditional distributions of X_i given θ_i are $B(1, \theta_i)$, $i = 1, 2, \ldots$, and that $\theta_1, \theta_2, \ldots$ are i.i.d. having some prior distribution $H(\theta)$ on (0, 1).

(i) Construct an empirical-Bayes estimator of θ for the squared-error loss.

(ii) Construct an empirical-Bayes estimator of θ for the squared-error loss, if it is assumed that $H(\theta)$ belongs to the family $\mathcal{H} = \{\beta(p, q) : 0 < p, q < \infty\}$.

8.6.2 Let $(X_1, \psi_1), \ldots, (X_n, \psi_n), \ldots$ be a sequence of independent random vectors of which only the Xs are observable. It is assumed that the conditional

distribution of X_i given ψ_i is $NB(\psi_i, \nu)$, ν known, $i = 1, 2, \ldots$. Moreover, it is assumed that $\psi_1, \psi_2, \ldots$ are i.i.d. having a prior distribution $H(\theta)$ belonging to the family $\mathcal{H}$ of beta distributions. Construct a sequence of empirical-Bayes estimators for the squared-error loss, and show that their posterior risks converges a.s. to the posterior risk of the true $\beta(p, q)$.

8.6.3 Let $(X_1, \lambda_1), \ldots, (X_n, \lambda_n), \ldots$ be a sequence of independent random vectors, where $X_i \mid \lambda_i \sim G(\lambda_i, 1), i = 1, 2, \ldots$, and $\lambda_1, \lambda_2, \ldots$ are i.i.d. having a prior $G\left(\frac{1}{\tau}, \nu\right)$ distribution; τ and ν unknown. Construct a sequence of empirical-Bayes estimators of λ_i, for the squared-error loss.

PART IV: SOLUTIONS OF SELECTED PROBLEMS

8.1.6 **(i)** Since $X \sim N(0, \sigma^2)$, $S = \sum_{i=1}^{n} X_i^2 \sim \sigma^2 \chi^2[n]$, or $S \sim 2\sigma^2 G\left(1, \frac{n}{2}\right)$. Thus, the density of S, given σ^2, is

$$f_S(x \mid \sigma^2) = \frac{1}{\Gamma(\frac{n}{2})(2\sigma^2)^{\frac{n}{2}}} x^{\frac{n}{2}-1} e^{-x/2\sigma^2}, \quad 0 < x < \infty.$$

Let $\phi = \frac{1}{2\sigma^2}$ and let the prior distribution of ϕ be like that of $G\left(\frac{1}{\tau}, \nu\right)$. Hence, the posterior distribution of ϕ, given S, is like that of $G\left(S + \frac{1}{\tau}, \frac{n}{2} + \nu\right)$.

(ii)
$$\begin{aligned} E\{\sigma^2 \mid S\} &= \frac{(S + \frac{1}{\tau})^{\frac{n}{2}+\nu}}{2\Gamma(\frac{n}{2} + \nu)} \int_0^\infty \phi^{\frac{n}{2}+\nu-2} e^{-\phi(S+E)} d\phi \\ &= \frac{(S + \frac{1}{\tau})^{\frac{n}{2}+\nu} \Gamma(\frac{n}{2} + \nu - 1)}{2\Gamma(\frac{n}{2} + \nu)(S + \frac{1}{\tau})^{\frac{n}{2}+\nu-1}} \\ &= \frac{S + \frac{1}{\tau}}{n + 2\nu - 2}. \end{aligned}$$

Similarly, we find that

$$V\{\sigma^2 \mid S\} = \frac{2(S + \frac{1}{\tau})^2}{(n + 2\nu - 2)^2(n + 2\nu - 4)}.$$

8.1.9

$$X \sim B(n, \theta), \quad 0 < \theta < 1,$$
$$\theta \sim \beta\left(\frac{1}{2}, \frac{1}{2}\right),$$
$$\theta \mid X \sim \beta\left(X + \frac{1}{2}, n - X + \frac{1}{2}\right).$$

Accordingly,

(i) $$E\{(\hat{\theta} - \theta)^2 \mid X\} = \hat{\theta}^2 - 2\hat{\theta}\frac{X + \frac{1}{2}}{n+1} + \frac{(X + \frac{1}{2})(X + \frac{3}{2})}{(n+1)(n+2)}.$$

(ii)

$$E\left\{\frac{\hat{\theta}^2}{\theta(1-\theta)} \mid X\right\} = \frac{n(n-1)\hat{\theta}^2}{(X - \frac{1}{2})(n - X - \frac{1}{2})} - \frac{2n\hat{\theta}}{n - X - \frac{1}{2}} + \frac{(X + \frac{1}{2})}{n(n - X - \frac{1}{2})}.$$

(iii)

$$E\{|\hat{\theta} - \theta| \mid X\} = 2\hat{\theta} I_{\hat{\theta}}\left(X + \frac{1}{2}, n - X + \frac{1}{2}\right) - 2\frac{X + \frac{1}{2}}{n+1} I_{\hat{\theta}}\left(X + \frac{3}{2}, n - X + \frac{1}{2}\right) - \frac{n - X + \frac{1}{2}}{n+1}.$$

8.2.1 **(i)** The prior risks are $R_0 = c_1\pi$ and $R_1 = c_2(1 - \pi)$, where π is the prior probability of H_0. These two risk lines intersect at $\pi^* = c_2/(c_1 + c_2)$. We have $R_0(\pi^*) = R_1(\pi^*)$. The posterior probability that H_0 is true is

$$\pi(X) = \frac{\pi e^{-\lambda_0}\lambda_0^x}{\pi e^{-\lambda_0}\lambda_0^x + (1 - \pi)e^{-\lambda_1}\lambda_1^x}.$$

The Bayes test function is

$$\phi_\pi(X) = \begin{cases} 1, & \text{if } \pi(x) < \pi^*, \\ 0, & \text{if } \pi(x) \geq \pi^*. \end{cases}$$

$\phi_\pi(X)$ is the probability of rejecting H_0. Note that $\phi_\pi(X) = 1$ if, and only if, $X > \xi(\pi)$, where

$$\xi(\pi) = \frac{\lambda_1 - \lambda_0 + \log\frac{\pi}{1-\pi} + \log\frac{c_1}{c_2}}{\log(\frac{\lambda_1}{\lambda_0})}.$$

(ii)
$$\begin{aligned} R_0(\pi) &= c_1 E\{\phi_\pi(X)\} \\ &= c_1 P\{X > \xi(\pi) \mid \lambda = \lambda_0\}. \end{aligned}$$

Let $P(j;\lambda)$ denote the c.d.f. of Poisson distribution with mean λ. Then

$$R_0(\pi) = c_1(1 - P([\xi(\pi)];\lambda_0)),$$

and

$$R_1(\pi) = c_2 P([\xi(\pi)];\lambda_1).$$

(iii)
$$\xi(\pi) = \frac{2 + \log\frac{\pi}{1-\pi} - \log(3)}{\log(2)}.$$

Then

$$R_0(\pi) = 1 - P\left(\left[1.300427 + \frac{\log\frac{\pi}{1-\pi}}{0.693147}\right];\lambda = 2\right),$$
$$R_1(\pi) = 3P\left(\left[1.300427 + \frac{\log\frac{\pi}{1-\pi}}{0.693147}\right];\lambda = 4\right).$$

8.3.2 We have a simple linear regression model $Y_i = \alpha + \beta x_i + \epsilon_i$, $i = 1, \ldots, n$; where ϵ_i are i.i.d. $N(0, \sigma^2)$. Assume that σ^2 is *known*. Let $(X) = (1_n, \mathbf{x}_n)$, where 1_n is an n-dimensional vector of 1s, $\mathbf{x}_n' = (x_1, \ldots, x_n)$. The model is

$$\mathbf{Y} = (X)\begin{pmatrix}\alpha\\ \beta\end{pmatrix} + \epsilon.$$

(i) Let $\boldsymbol{\theta} = \begin{pmatrix}\alpha\\ \beta\end{pmatrix}$, $\boldsymbol{\theta}_0\begin{pmatrix}\alpha_0\\ \beta_0\end{pmatrix}$. The prior distribution of $\boldsymbol{\theta}$ is $N(\boldsymbol{\theta}_0, \Sigma)$. Note that the covariance matrix of $\mathbf{Y}$ is $V[\mathbf{Y}] = \sigma^2 I + (X)\Sigma(X)'$. Thus, the posterior distribution of $\boldsymbol{\theta}$, given $\mathbf{Y}$, is $N(\eta(\mathbf{Y}), D)$, where

$$\boldsymbol{\eta}(\mathbf{Y}) = \boldsymbol{\theta}_0 + \Sigma(X)'(\sigma^2 I + (X)\Sigma(X)')^{-1}(\mathbf{Y} - (X)\boldsymbol{\theta}_0),$$

and

$$D = \Sigma - \Sigma(X)'(\sigma^2 I + (X)\Sigma(X)')^{-1}(X)\Sigma.$$

Accordingly

$$(\boldsymbol{\theta} - \boldsymbol{\eta}(\mathbf{Y}))' D^{-1} (\boldsymbol{\theta} - \boldsymbol{\eta}(\mathbf{Y}) \sim \chi^2[2].$$

Thus, the $(1-\alpha)$ credibility region for θ is

$$\{\boldsymbol{\theta} : (\boldsymbol{\theta} - \boldsymbol{\eta}(\mathbf{Y}))' D^{-1} (\boldsymbol{\theta} - \boldsymbol{\eta}(\mathbf{Y})) \leq \chi^2_{1-\varepsilon}[2]\}.$$

(ii) Let $\mathbf{w} = \begin{pmatrix} 1 \\ \xi \end{pmatrix}$ and $\mathbf{w}_0 = \begin{pmatrix} 1 \\ \xi_0 \end{pmatrix}$. $\alpha + \beta\xi_0 = \boldsymbol{\theta}'\mathbf{w}_0$. The posterior distribution of $\boldsymbol{\theta}'\mathbf{w}_0$, given $\mathbf{Y}$, is $N(\boldsymbol{\eta}(\mathbf{Y})'\mathbf{w}_0, \mathbf{w}_0' D \mathbf{w}_0)$. Hence, a $(1-\alpha)$ credibility interval for $\boldsymbol{\theta}'\mathbf{w}_0$, given $\mathbf{Y}$, has the limits $\boldsymbol{\eta}(\mathbf{Y})'\mathbf{w}_0 \pm z_{1-\varepsilon/2}(\mathbf{w}_0' D \mathbf{w}_0)^{1/2}$.

(iii) Simultaneous credibility interval for all ξ is by Sheffe's S-intervals

$$\boldsymbol{\eta}(\mathbf{Y})'\mathbf{w} \pm (2\chi^2_{1-\varepsilon}[2])^{1/2} (\mathbf{w}' D \mathbf{w})^{1/2}.$$

8.4.1

$$X \sim B(n, \theta), \quad 0 < \theta < 1.$$

The prior of θ is Beta(p, q), $0 < p, q < \infty$.

(i) The posterior distribution of θ, given X, is Beta$(p + X, q + n - X)$. Hence, the Bayes estimator of θ, for squared-error loss, is $\hat{\theta}_B = E(\theta \mid X) = \dfrac{X + p}{n + p + q}$.

(ii) The posterior risk is $V(\theta \mid X) = \dfrac{(X+p)(n-X+q)}{(n+p+q)^2(n+p+q+1)}$. The prior risk is $E_H\{\text{MSE}(\hat{\theta}_B)\} = \dfrac{pq}{(p+q)(p+q+1)(n+p+q)}$.

(iii) The Bayes estimator is

$$\hat{\theta}_B = \min\left(\frac{(X - \frac{1}{2})^+}{n-1}, 1\right).$$

8.4.3 The Jeffreys prior is $h(\mu, \sigma^2)d\mu d\sigma^2 = \frac{1}{\sigma^2} d\mu d\sigma^2$. A version of the likelihood function, given the minimal sufficient satistic $(\bar{X}, Q)$, where $Q = \sum_{i=1}^{n}(X_i - \bar{X})^2$, is

$$L(\mu, \sigma^2 \mid \bar{X}, Q) = \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{Q}{2\sigma^2} - \frac{n}{2\sigma^2}(\bar{X} - \mu)^2\right\}.$$

Thus, the posterior density under Jeffrey's prior is

$$h(\mu, \sigma^2) \mid \bar{X}, Q) = \frac{(\sigma^2)^{-(\frac{n}{2}+1)} \exp\{-\frac{Q}{2\sigma^2} - \frac{n}{2\sigma^2}(\bar{X} - \mu)^2\}}{\int_{-\infty}^{\infty}\int_{0}^{\infty} (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left\{-\frac{Q}{2\sigma^2} - \frac{n}{2\sigma^2}(\bar{X} - \mu)^2\right\} d\mu d\sigma^2}.$$

Now,

$$\int_{-\infty}^{\infty} e^{-\frac{n}{2\sigma^2}(\bar{X}-\mu)^2} d\mu = \sigma\sqrt{\frac{2\pi}{n}}.$$

Furthermore,

$$\int_{0}^{\infty} \frac{\sigma}{(\sigma^2)^{\frac{n}{2}+1}} e^{-\frac{Q}{2\sigma^2}} d\sigma^2 = \int_{0}^{\infty} \phi^{\frac{n-1}{2}-1} e^{-\phi\frac{Q}{2}} d\phi = \frac{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})}{Q^{\frac{n-1}{2}}}.$$

Accordingly,

$$h(\mu, \sigma^2 \mid \bar{X}, Q) = \frac{Q^{\frac{n-1}{2}}}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})} \cdot \frac{1}{(\sigma^2)^{\frac{n}{2}+1}} \cdot \exp\left\{-\frac{Q}{2\sigma^2} - \frac{n(\bar{X} - \mu)^2}{2\sigma^2}\right\}.$$

Thus, the predictive density of Y, given $[\bar{X}, Q]$, is

$$f_H(y \mid \bar{X}, Q) = \frac{Q^{\frac{n-1}{2}}}{2^{\frac{n}{2}}\sqrt{\pi}\,\Gamma(\frac{n-1}{2})} \int_{0}^{\infty} \frac{e^{-\frac{Q}{2\sigma^2}}}{(\sigma^2)^{\frac{n+1}{2}+1}} \cdot \cdot \left(\int_{-\infty}^{\infty} \exp\left\{-\frac{n(\bar{X} - \mu)^2}{2\sigma^2} - \frac{(y - \mu)^2}{2\sigma^2}\right\} d\mu\right) d\sigma^2,$$

or

$$f_H(y \mid \bar{X}, Q) = \frac{1}{\sqrt{Q}\, B(\frac{1}{2}, \frac{n-1}{2})} \left(1 + \frac{n}{n+1}\frac{(y - \bar{X})^2}{Q}\right)^{-\frac{n}{2}}.$$

Recall that the density of $t[n-1]$ is

$$f(t;n-1) = \frac{1}{\sqrt{n-1}\,B(\frac{1}{2},\frac{n-1}{2})}\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}}.$$

Thus, the γ-quantile of the predictive distribution $f_H(y \mid \bar{X}, Q)$ is $\bar{X} + \frac{t_\gamma[n-1]}{\sqrt{n-1}}\sqrt{Q\left(1+\frac{1}{n}\right)}$.

8.5.3

$$I = \int_0^1 e^{-\theta}\theta^X(1-\theta)^{n-X}d\theta = B(X+1,n-X+1)E\{e^{-U}\},$$

where $U \sim \beta(X+1,n-X+1)$. By simulation, we generate M i.i.d. values of $U_1,\ldots,U_M$, and estimate I by $\hat{I}_M = B(X+1,n-X+1)\cdot\frac{1}{M}\sum_{i=1}^{M}e^{-U_i}$. For large M, $\hat{I}_M \approx AN\left(I,\frac{1}{M}D\right)$, where $D = B^2(X+1,n-X+1)V\{e^{-U}\}$. We have to find M large enough so that $P\{|\hat{I}_M - I| \le 0.01I\} \ge 0.95$. According to the asymptotic normal distribution, we determine M so that

$$2\Phi\left(\frac{\sqrt{M}\,0.01E\{e^{-U}\}}{\sqrt{V\{e^{-U}\}}}\right) - 1 \ge 0.95,$$

or

$$M \ge \frac{\chi^2_{.95}[1]}{0.0001}\left(\frac{E\{e^{-2U}\}}{(E(e^{-U}))^2} - 1\right).$$

By the delta method, for large M,

$$E\{e^{-2U}\} \cong e^{-2\frac{X+1}{n+2}}\left(1+2\frac{(X+1)(n-X+1)}{(n+2)^2(n+3)}\right).$$

Similarly,

$$E\{e^{-U}\} \cong e^{-\frac{X+1}{n+2}}\left(1+\frac{1}{2}\frac{(X+1)(n-X+1)}{(n+2)^2(n+3)}\right).$$

Hence, M should be the smallest integer such that

$$M \ge \frac{\chi^2_{.95}[1]}{0.0001}\left(\frac{1+2\frac{(X+1)(n-X+1)}{(n+2)^2(n+3)}}{\left(1+\frac{1}{2}\frac{(X+1)(n-X+1)}{(n+2)^2(n+3)}\right)^2} - 1\right).$$

CHAPTER 9

Advanced Topics in Estimation Theory

PART I: THEORY

In the previous chapters, we discussed various classes of estimators, which attain certain optimality criteria, like minimum variance unbiased estimators (MVUE), asymptotic optimality of maximum likelihood estimators (MLEs), minimum mean-squared-error (MSE) equivariant estimators, Bayesian estimators, etc. In this chapter, we present additional criteria of optimality derived from the general statistical decision theory. We start with the game theoretic criterion of minimaxity and present some results on minimax estimators. We then proceed to discuss minimum risk equivariant and standard estimators. We discuss the notion of admissibility and present some results of Stein on the inadmissibility of some classical estimators. These examples lead to the so-called Stein-type and Shrinkage estimators.

9.1 MINIMAX ESTIMATORS

Given a class $\mathcal{D}$ of estimators, the risk function associated with each $d \in \mathcal{D}$ is $R(d, \theta)$, $\theta \in \Theta$. The maximal risk associated with d is $R^*(d) = \sup_{\theta \in \Theta} R(d, \theta)$. If in $\mathcal{D}$ there is an estimator d^* that minimizes $R^*(d)$ then d^* is called a **minimax** estimator. That is,

$$\sup_{\theta \in \Theta} R(d^*, \theta) = \inf_{d \in \mathcal{D}} \sup_{\theta \in \theta} R(d, \theta).$$

A minimax estimator may not exist in $\mathcal{D}$. We start with some simple results.

Lemma 9.1.1. *Let $\mathcal{F} = \{F(x; \theta), \theta \in \Theta\}$ be a family of distribution functions and $\mathcal{D}$ a class of estimators of θ. Suppose that $d^* \in \mathcal{D}$ and d^* is a Bayes estimator relative*

Examples and Problems in Mathematical Statistics, First Edition. Shelemyahu Zacks.

to some prior distribution $H^*(\theta)$ and that the risk function $R(d^*, \theta)$ does not depend on θ. Then d^* is a minimax estimator.

Proof. Since $R(d^*, \theta) = \rho^*$ for all θ in Θ, and d^* is Bayes against $H^*(\theta)$ we have

$$\begin{aligned} \rho^* &= \int R(d^*, \theta) h^*(\theta) d\theta = \inf_{d \in \mathcal{D}} \int R(d, \theta) h^*(\theta) d\theta \\ &\leq \sup_{\theta \in \Theta} \inf_{d \in \mathcal{D}} R(d, \theta) \leq \inf_{d \in \mathcal{D}} \sup_{\theta \in \Theta} R(d, \theta). \end{aligned} \tag{9.1.1}$$

On the other hand, since $\rho^* = R(d^*, \theta)$ for all θ

$$\rho^* = \sup_{\theta \in \Theta} R(d^*, \theta) \geq \inf_{d \in \mathcal{D}} \sup_{\theta \in \Theta} R(d, \theta). \tag{9.1.2}$$

From (9.1.1) and (9.1.2), we obtain that

$$\sup_{\theta \in \Theta} R(d^*, \theta) = \inf_{d \in \mathcal{D}} \sup_{\theta \in \Theta} R(d, \theta). \tag{9.1.3}$$

This means that d^* is minimax. QED

Lemma 9.1.1 can be generalized by proving that if there exists a sequence of Bayes estimators with prior risks converging to ρ^*, where ρ^* is a constant risk of d^*, then d^* is minimax. We obtain this result as a corollary of the following lemma.

Lemma 9.1.2. *Let $\{H_k; k \geq 1\}$ be a sequence of prior distributions on Θ and let $\{\hat{\theta}_k; k \geq 1\}$ be the corresponding sequence of Bayes estimators with prior risks $\rho(\hat{\theta}_k, H_k)$. If there exists an estimator d^* for which*

$$\sup_{\theta \in \Theta} R(\theta^*, \theta) \leq \limsup_{k \to \infty} \rho(\hat{\theta}_k, H_k), \tag{9.1.4}$$

then d^* is minimax.

Proof. If d^* is not a minimax estimator, there exists an estimator $\tilde{\theta}$ such that

$$\sup_{\theta \in \Theta} R(\tilde{\theta}, \theta) < \sup_{\theta \in \Theta} R(d^*, \theta). \tag{9.1.5}$$

Moreover, for each $k \geq 1$ since $\hat{\theta}_k$ is Bayes,

$$\begin{aligned} \rho(\hat{\theta}_k, H_k) &= \int R(\hat{\theta}_k, \theta) h_k(\theta) d\theta \\ &\leq \int R(\tilde{\theta}, \theta) h_k(\theta) d\theta \leq \sup_{\theta \in \Theta} R(\tilde{\theta}, \theta). \end{aligned} \tag{9.1.6}$$

But (9.1.5) in conjunction with (9.1.6) contradict (9.1.4). Hence, d^* is minimax. QED

9.2 MINIMUM RISK EQUIVARIANT, BAYES EQUIVARIANT, AND STRUCTURAL ESTIMATORS

In Section 5.7.1, we discussed the structure of models that admit equivariant estimators with respect to certain groups of transformations. In this section, we return to this subject and investigate minimum risk, Bayes and minimax equivariant estimators. The statistical model under consideration is specified by a sample space $\mathcal{X}$ and a family of distribution functions $\mathcal{F} = \{F(x; \theta); \theta \in \Theta\}$. Let $\mathcal{G}$ be a group of transformations that preserves the structure of the model, i.e., $g\mathcal{X} = \mathcal{X}$ for all $g \in \mathcal{G}$, and the induced group $\bar{\mathcal{G}}$ of transformations on Θ has the property that $\bar{g}\Theta = \Theta$ for all $\bar{g} \in \bar{\mathcal{G}}$. An equivariant estimator $\hat{\theta}(\mathbf{X})$ of θ was defined as one which satisfies the structural property that $\hat{\theta}(g\mathbf{X}) = \bar{g}\hat{\theta}(\mathbf{X})$ for all $g \in \mathcal{G}$.

In cases of various orbits of $\mathcal{G}$ in Θ, we may index the orbits by a parameter, say $\omega(\theta)$. The risk function of an equivariant estimator $\hat{\theta}(\mathbf{X})$ is then $R(\hat{\theta}, \omega(\theta))$. **Bayes equivariant** estimators can be considered. These are equivariant estimators that minimize the prior risk associated with θ, relative to a prior distribution $H(\theta)$. We assume that $\omega(\theta)$ is a function of θ for which the following prior risk exists, namely,

$$\begin{aligned} \rho(\hat{\theta}, H) &= \int_{\Theta} R(\hat{\theta}, \omega(\theta)) dH(\theta) \\ &= \int_{\Omega} R(\hat{\theta}, \omega) dK(\omega), \end{aligned} \tag{9.2.1}$$

where $K(\omega)$ is the prior distribution of $\omega(\theta)$, induced by $H(\theta)$. Let $U(\mathbf{X})$ be a maximal invariant statistic with respect to $\mathcal{G}$. Its distribution depends on θ only through $\omega(\theta)$. Suppose that $g(u; \omega)$ is the probability density function (p.d.f.) of $U(\mathbf{X})$ under ω. Let $k(\omega \mid U)$ be the posterior p.d.f. of ω given $U(\mathbf{X})$. The prior risk of θ can be written then as

$$\rho(\hat{\theta}, H) = E_{U|K}\{E_{\omega|U}\{R(\hat{\theta}, \omega)\}\}. \tag{9.2.2}$$

where $E_{\omega|U}\{R(\hat{\theta}, \omega)\}$ is the posterior risk of θ, given $U(X)$. **An equivariant estimator $\hat{\theta}_K$ is Bayes against $K(\omega)$ if it minimizes $E_{\omega|U}\{R(\hat{\theta}, \omega)\}$.**

As discussed earlier, the Bayes equivariant estimators are relevant only if there are different orbits of $\mathcal{G}$ in Θ. Another approach to the estimation problem, if there are no minimum risk equivariant estimators, is to derive formally the Bayes estimators with respect to invariant prior measures (like the Jeffreys improper priors). Such an approach to the above problem of estimating variance components was employed by Tiao and Tan (1965) and by Portnoy (1971). We discuss now formal Bayes estimators more carefully.

9.2.1 Formal Bayes Estimators for Invariant Priors

Formal Bayes estimators with respect to invariant priors are estimators that minimize the expected risk, when the prior distribution used is improper. In this section, we are concerned with invariant prior measures, such as the Jeffreys noninformative prior $h(\theta)d\theta \propto |I(\theta)|^{1/2}d\theta$. With such improper priors, the minimum risk estimators can often be formally derived as in Section 5.7. The resulting estimators are called **formal Bayes estimators**. For example, if $\mathcal{F} = \{F(x;\theta); -\infty < \theta < \infty\}$ is a family of location parameters distributions (of the translation type), i.e., the p.d.f.s are $f(x;\theta) = \phi(x-\theta)$ then, for the group $\mathcal{G}$ of real translations, the Jeffreys invariant prior is $h(\theta)d\theta \propto d\theta$. If the loss function is $L(\hat{\theta},\theta) = (\hat{\theta}-\theta)^2$, the formal Bayes estimator is

$$E\{\theta \mid \mathbf{X}\} = \frac{\int_{-\infty}^{\infty} \theta \prod_{i=1}^{n} \phi(X_i - \theta)d\theta}{\int_{-\infty}^{\infty} \prod_{i=1}^{n} \phi(X_i - \theta)d\theta}. \tag{9.2.3}$$

Making the transformation $Y = X_{(1)} - \theta$, where $X_{(1)} \leq \cdots \leq X_{(n)}$, we obtain

$$E\{\theta \mid \mathbf{X}\} = X_{(1)} - \frac{\int_{-\infty}^{\infty} y\phi(y) \prod_{i=2}^{n} \phi(X_{(i)} - X_{(1)} + y)dy}{\int_{-\infty}^{\infty} \phi(y) \prod_{i=2}^{n} \phi(X_{(i)} - X_{(1)} + y)dy}. \tag{9.2.4}$$

This is the Pitman estimator (5.7.10).

When $\mathcal{F}$ is a family of location and scale parameters, with p.d.f.s

$$f(x;\mu,\sigma) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right), \tag{9.2.5}$$

we consider the group $\mathcal{G}$ of real affine transformations $\mathcal{G} = \{[\alpha, \beta]; -\infty < \alpha < \infty, 0 < \beta < \infty\}$. The fisher information matrix of (μ, σ) for $\mathcal{F}$ is, if exists,

$$I(\mu, \sigma) = \frac{1}{\sigma^2}\begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}, \tag{9.2.6}$$

where

$$I_{11} = V\left\{\frac{\phi'(u)}{\phi(u)}\right\}, \tag{9.2.7}$$

$$I_{12} = I_{21} = \operatorname{cov}\left(\frac{\phi'(u)}{\phi(u)}, u\frac{\phi'(u)}{\phi(u)}\right), \tag{9.2.8}$$

and

$$I_{22} = V\left\{u^2\,\frac{\phi'(u)}{\phi(u)}\right\}. \tag{9.2.9}$$

Accordingly, $|I(\mu, \sigma)| \propto \dfrac{1}{\sigma^4}$ and the Jeffreys invariant prior is

$$h(\mu, \sigma)d\mu d\sigma \propto \frac{d\mu d\sigma}{\sigma^2}. \tag{9.2.10}$$

If the invariant loss function for estimating μ is $L(\hat{\mu}, \mu, \sigma) = \dfrac{(\hat{\mu} - \mu)^2}{\sigma^2}$ then the formal Bayes estimator of μ is

$$\begin{aligned} \hat{\mu} &= \frac{E\left\{\frac{\mu}{\sigma^2} \mid \mathbf{X}\right\}}{E\left\{\frac{1}{\sigma^2} \mid \mathbf{X}\right\}} \\ &= \frac{\int_{-\infty}^{\infty}\mu\int_0^{\infty}\frac{1}{\sigma^{n+4}}\prod_{i=1}^{n}\phi\left(\frac{x_i - \mu}{\sigma}\right)d\sigma d\mu}{\int_{-\infty}^{\infty}\int_0^{\infty}\frac{1}{\sigma^{n+4}}\prod_{i=1}^{n}\phi\left(\frac{x_i - \mu}{\sigma}\right)d\sigma d\mu}. \end{aligned} \tag{9.2.11}$$

Let $y_1 \leq \cdots \leq y_n$ represent realization of the order statistics $X_{(1)} \leq \cdots \leq X_{(n)}$. Consider the change of variables

$$\begin{aligned} u &= \frac{y_1 - \mu}{\sigma}, \\ v &= \frac{y_2 - y_1}{\sigma}, \\ z_i &= \frac{y_i - y_1}{y_2 - y_1}, \quad i \geq 3, \end{aligned} \tag{9.2.12}$$

then the formal Bayes estimator of μ is

$$\hat{\mu} = y_1 - (y_2 - y_1) \cdot \frac{\int_{-\infty}^{\infty} u\phi(u) \int_0^{\infty} v^n \phi(u+v) \prod_{i=3}^{n} \phi(u + vz_i) dv du}{\int_{-\infty}^{\infty} \phi(u) \int_0^{\infty} v^{n+1} \phi(u+v) \prod_{i=3}^{n} \phi(u + vz_i) dv du}. \tag{9.2.13}$$

For estimating σ, we consider the invariant loss function $L(\hat{\sigma}, \sigma) = (\hat{\sigma} - \sigma)^2/\sigma^2$. The formal Bayes estimator is then

$$\begin{aligned} \hat{\sigma} &= E\left\{\frac{1}{\sigma} \mid X\right\} / E\left\{\frac{1}{\sigma^2} \mid X\right\} \\ &= (y_2 - y_1) \frac{\int_{-\infty}^{\infty} \phi(u) \int_0^{\infty} v^n \phi(u+v) \prod_{i=3}^{n} \phi(u + vz_i) dv du}{\int_{-\infty}^{\infty} \phi(u) \int_0^{\infty} v^{n+1} \phi(u+v) \prod_{i=3}^{n} \phi(u + vz_i) dv du}. \end{aligned} \tag{9.2.14}$$

Note that the formal Bayes estimator (9.2.13) is equivalent to the Pitman estimator, for a location parameter family (with known scale parameter). The estimator (9.2.14) is the Pitman estimator for a scale parameter family.

Formal Bayes estimation can be used also when the model has parameters that are invariant with respect to the group of transformations $\mathcal{G}$. In the variance components model discussed in Example 9.3, the variance ratio $\rho = \tau^2/\sigma^2$ is such an invariant parameter. These parameters are called also **nuisance parameters** for the transformation model.

9.2.2 Equivariant Estimators Based on Structural Distributions

Fraser (1968) introduced structural distributions of parameters in cases of invariance structures, when all the parameters of the model can be transformed by the transformations in $\mathcal{G}$. Fraser's approach does not require the assignment of a prior

distribution to the unknown parameters. This approach is based on changing the variables of integration from those representing the observable random variables to those representing the parameters. We start the explanation by considering real parameter families. More specifically, let $\mathcal{F} = \{F(x;\theta); \theta \in \Theta\}$ be a family of distributions, where Θ is an interval on the real line. Let $\mathcal{G}$ be a group of one-to-one transformations, preserving the structure of the model. For the simplicity of the presentation, we assume that the distribution functions of $\mathcal{F}$ are absolutely continuous and the transformation in $\mathcal{G}$ can be represented as functions over $\mathcal{X} \times \Theta$. Choose in Θ a standard or reference point e and let U be a random variable, having the distribution $F(u;e)$, which is the standard distribution. Let $\phi(u)$ be the p.d.f. of the standard distribution. The structural model assumes that if a random variable X has a distribution function $F(x;\theta)$; when $\theta = \bar{g}e$, $g \in \mathcal{G}$, then $X = gU$. Thus, the structural model can be expressed in the formula

$$X = G(U, \theta), \quad \theta \in \Theta. \tag{9.2.15}$$

Assume that $G(u, \theta)$ is differentiable with respect to u and θ. Furthermore, let

$$u = G^{-1}(x;\theta); \quad x \in \mathcal{X}, \quad \theta \in \Theta. \tag{9.2.16}$$

The function $G(u, \theta)$ satisfies the equivariance condition that

$$gx = G(u, \bar{g}\theta), \quad \text{all } g \in \mathcal{G};$$

with an invariant inverse; i.e.,

$$u = G^{-1}(x, \theta) = G^{-1}(gx, \bar{g}\theta), \quad \text{all } g \in \mathcal{G}.$$

We consider now the variation of u as a function of θ for a fixed value of x. Writing the probability element of U at u in the form

$$\phi(u)du = \phi(G^{-1}(x, \theta)) \mid \frac{\partial}{\partial\theta} G^{-1}(x, \theta) \mid d\theta, \tag{9.2.17}$$

we obtain for every fixed x a distribution function for θ, over Θ, with p.d.f.

$$k(\theta; x) = \phi(G^{-1}(x, \theta))m(\theta, x), \tag{9.2.18}$$

where $m(\theta, x) = |\frac{\partial}{\partial\theta} G^{-1}(x, \theta)|$. The distribution function corresponding to $k(\theta, x)$ is called **the structural distribution** of θ given $X = x$. Let $L(\hat{\theta}(x), \theta)$ be an invariant loss function. The structural risk of $\hat{\theta}(x)$ is the expectation

$$R(\hat{\theta}(x)) = \int L(\hat{\theta}(x), \theta)k(\theta; x)d\theta. \tag{9.2.19}$$

An estimator $\theta_0(x)$ is called **minimum risk structural estimator** if it minimizes $R(\hat{\theta}(x))$. The p.d.f. (9.2.18) corresponds to one observation on X. Suppose that a sample of n independent identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ is represented by the point $\mathbf{x} = (x_1, \ldots, x_n)$. As before, θ is a real parameter. Let $V(\mathbf{X})$ be a maximal invariant statistic with respect to $\mathcal{G}$. The distribution of $V(\mathbf{X})$ is independent of θ. (We assume that Θ has one orbit of $\mathcal{G}$.) Let $k(\mathbf{v})$ be the joint p.d.f. of the maximal invariant $V(\mathbf{X})$. Let $u_1 = G^{-1}(x_1, \theta)$ and let $\phi(u \mid \mathbf{x})$ be the conditional p.d.f. of the standard variable $U = [\theta]^{-1}X$, given $V = \mathbf{v}$. This conditional p.d.f. of θ, for a given $\mathbf{x}$ is then, like in (9.2.18),

$$k(\theta; \mathbf{x}) = \phi(G^{-1}(x_1, \theta) \mid v(\mathbf{x}))m(\theta, x_1), \quad \theta \in \Theta. \tag{9.2.20}$$

If the model depends on a vector θ of parameters we make the appropriate generalizations as will be illustrated in Example 9.4.

We conclude the present section with some comment concerning minimum properties of formal Bayes and structural estimators. Girshick and Savage (1951) proved that if all equivariant estimators in the location parameter model have finite risk, then the Pitman estimator (9.2.24) is minimax. Generally, if a formal Bayes estimator with respect to an invariant prior measure (as the Jeffreys priors) and invariant loss function is an equivariant estimator, and if the parameter space Θ has only one orbit of $\mathcal{G}$, then the risk function of the formal Bayes estimator is constant over Θ. Moreover, if this formal Bayes estimator can be obtained as a limit of a sequence of proper Bayes estimators, or if there exists a sequence of proper Bayes estimators and the lower limit of their prior risks is not smaller than the risk of the formal Bayes estimator, then the formal Bayes is a minimax estimator. Several theorems are available concerning the minimax nature of the minimum risk equivariant estimators. The most famous is the **Hunt–Stein Theorem** (Zacks, 1971; p. 346).

9.3 THE ADMISSIBILITY OF ESTIMATORS

9.3.1 Some Basic Results

The class of all estimators can be classified according to the given risk function into two subclasses: **admissible and inadmissible** ones.

Definition. *An estimator $\hat{\theta}_1(x)$ is called inadmissible with respect to a risk function $R(\hat{\theta}, \theta)$ if there exists another estimator $\hat{\theta}_2(x)$ for which*

$$\begin{aligned} &(\boldsymbol{i}) \quad R(\hat{\theta}_2, \theta) \leq R(\hat{\theta}_1, \theta), \quad \textit{for all } \theta, \\ &(\boldsymbol{ii}) \quad R(\hat{\theta}_2, \theta') < R(\hat{\theta}_1, \theta'), \quad \textit{for some } \theta'. \end{aligned} \tag{9.3.1}$$

From the decision theoretic point of view inadmissible estimators are inferior. It is often not an easy matter to prove that a certain estimator is admissible. On the other hand, several examples exist of the inadmissibility of some commonly used

estimators. A few examples will be provided later in this section. We start, however, with a simple and important lemma.

Lemma 9.3.1 (Blyth, 1951). *If the risk function $R(\hat{\theta}, \theta)$ is continuous in θ for each $\hat{\theta}$, and if the prior distribution $H(\theta)$ has a positive p.d.f. at all θ then the Bayes estimator $\hat{\theta}_H(x)$ is admissible.*

Proof. By negation, if $\hat{\theta}_H(x)$ is inadmissible then there exists another estimator $\hat{\theta}^*(x)$ for which (9.3.1) holds. Let θ^* be a point at which the strong inequality (ii) of (9.3.1) holds. Since $R(\hat{\theta}, \theta)$ is continuous in θ for each $\hat{\theta}$, there exists a neighborhood $N(\theta^*)$ around θ^* over which the inequality (ii) holds for all $\theta \in N(\theta^*)$. Since $h(\theta) > 0$ for all θ, $P_H\{N(\theta^*)\} > 0$. Finally from inequality (i) we obtain that

$$\begin{aligned} \int_\Theta R(\theta^*, \theta)h(\theta)d\theta &= \int_{N(\theta^*)} R(\theta^*, \theta)h(\theta)d\theta + \int_{\bar{N}(\theta^*)} R(\theta^*, \theta)h(\theta)d\theta \\ &< \int_{N(\theta^*)} R(\hat{\theta}_H, \theta)h(\theta)d\theta + \int_{\bar{N}(\theta^*)} R(\hat{\theta}_H, \theta)h(\theta)d\theta. \end{aligned} \tag{9.3.2}$$

The left-hand side of (9.3.2) is the prior risk of θ^* and the right-hand side is the prior risk of $\hat{\theta}_H$. But this result contradicts the assumption that $\hat{\theta}_H$ is Bayes with respect to $H(\theta)$. QED

All the examples, given in Chapter 8, of proper Bayes estimators illustrate admissible estimators. Improper Bayes estimators are not necessarily admissible. For example, in the $N(\mu, \sigma^2)$ case, when both parameters are unknown, the formal Bayes estimator of σ^2 with respect to the Jeffreys improper prior $h(\sigma^2)d\sigma^2 \propto d\sigma^2/\sigma^2$ is $Q/(n-3)$, where $Q = \Sigma(X_i - \bar{X})^2$. This estimator is, however, inadmissible, since $Q/(n+1)$ has a smaller MSE for all σ^2. There are also admissible estimators that are not Bayes. For example, the sample mean $\bar{X}$ from a normal distribution $N(\theta, 1)$ is an admissible estimator with respect to a squared-error loss. However, $\bar{X}$ is not a proper Bayes estimator. It is a limit (as $k \to \infty$) of the Bayes estimators derived in Section 8.4, $\hat{\theta}_k = \bar{X}\left(1 + \frac{1}{nk}\right)^{-1}$. $\bar{X}$ is also an improper Bayes estimator with respect to the Jeffreys improper prior $h(\theta)d\theta \propto d\theta$. Indeed, for such an improper prior

$$\hat{\theta} = \frac{\int_{-\infty}^{\infty} \theta \exp\left\{-\frac{n}{2}(\bar{X} - \theta)^2\right\} d\theta}{\int_{-\infty}^{\infty} \exp\left\{-\frac{n}{2}(\bar{X} - \theta)^2\right\} d\theta} = \bar{X}. \tag{9.3.3}$$

The previous lemma cannot establish the admissibility of the sample mean $\bar{X}$. We provide here several lemmas that can be used.

Lemma 9.3.2. *Assume that the MSE of an estimator $\hat{\theta}_1$ attains the Cramér–Rao lower bound (under the proper regularity conditions) for all θ, $-\infty < \theta < \infty$, which is*

$$C_1(\theta) = B_1^2(\theta) + \frac{(1 + B_1'(\theta))^2}{I(\theta)}, \tag{9.3.4}$$

where $B_1(\theta)$ is the bias of $\hat{\theta}_1$. Moreover, if for any estimator $\hat{\theta}_2$ having a Cramér–Rao lower bound $C_2(\theta)$, the inequality $C_2(\theta) \leq C_1(\theta)$ for all θ implies that $B_2(\theta) = B_1(\theta)$ for all θ, then θ_1 is admissible.

Proof. If $\hat{\theta}_1$ is inadmissible, there exists an estimator $\hat{\theta}_2$ such that

$$R(\hat{\theta}_2, \theta) \leq R(\hat{\theta}_1, \theta), \quad \text{for all } \theta,$$

with a strict inequality at some θ'. Since $R(\hat{\theta}_1, \theta) = C_1(\theta)$ for all θ, we have

$$C_2(\theta) = R(\hat{\theta}_2, \theta) \leq R(\hat{\theta}_1, \theta) = C_1(\theta) \tag{9.3.5}$$

for all θ. But, according to the hypothesis, (9.3.5) implies that $B_1(\theta) = B_2(\theta)$ for all θ. Hence, $C_1(\theta) = C_2(\theta)$ for all θ. But this contradicts the assumption that $R(\hat{\theta}_2, \theta') < R(\hat{\theta}_1, \theta')$. Hence, $\hat{\theta}_1$ is admissible. QED

Lemma 9.3.2 can be applied to prove that, in the case of a sample from $N(0, \sigma^2)$, $S^2 = \frac{1}{n+2}\sum_{i=1}^{n} X_i^2$ is an admissible estimator of σ^2. (The MVUE and the MLE are inadmissible!) In such an application, we have to show that the hypotheses of Lemma 9.3.2 are satisfied. In the $N(0, \sigma^2)$ case, it requires lengthy and tedious computations (Zacks, 1971, p. 373). Lemma 9.3.2 is also useful to prove the following lemma (Girshick and Savage, 1951).

Lemma 9.3.3. *Let X be a one-parameter exponential type random variable, with p.d.f.*

$$f(x; \psi) = h(x) \exp\{\psi x - K(\psi)\},$$

$-\infty < \psi < \infty$. Then $\hat{\mu} = X$ is an admissible estimator of its expectation $\mu(\psi) = +K'(\psi)$, for the quadratic loss function $(\hat{\mu} - \mu)^2/\sigma^2(\psi)$; where $\sigma^2(\psi) = +K''(\psi)$ is the variance of X.

Proof. The proof of the present lemma is based on the following points. First X is an unbiased estimator of $\mu(\psi)$. Since the distribution of X is of the exponential type, its variance $\sigma^2(\psi)$ is equal to the Cramér–Rao lower bound, i.e.,

$$\begin{aligned}\sigma^2(\psi) &= (\mu'(\psi))^2/I(\psi) \\ &= (\sigma^2(\psi))^2/I(\psi).\end{aligned} \tag{9.3.6}$$

This implies that $I(\psi) = \sigma^2(\psi)$, which can be also derived directly. If $\tilde{\mu}(X)$ is any other estimator of $\mu(\psi)$ satisfying the Cramér–Rao regularity condition with variance $D^2(\psi)$, such that

$$D^2(\psi) \leq \sigma^2(\psi), \quad \text{all } -\infty < \psi < \infty, \tag{9.3.7}$$

then from the Cramér–Rao inequality

$$B^2(\psi) + \frac{(B'(\psi) + \mu'(\psi))^2}{\sigma^2(\psi)} \leq D^2(\psi) \leq \sigma^2(\psi), \quad \text{all } \psi, \tag{9.3.8}$$

where $B(\psi)$ is the bias function of $\tilde{\mu}(X)$. Thus, we arrived at the inequality

$$B^2(\psi)\sigma^2(\psi) + [B'(\psi) + \sigma^2(\psi)]^2 \leq \sigma^4(\psi), \tag{9.3.9}$$

all $-\infty < \psi < \infty$. This implies that

$$B'(\psi) \leq 0 \text{ and } B^2(\psi) + 2B'(\psi) \leq -\frac{(B'(\psi))^2}{\sigma^2(\psi)} \leq 0, \tag{9.3.10}$$

for all $-\infty < \psi < \infty$. From (9.3.10), we obtain that either $B(\psi) = 0$ for all ψ or

$$\frac{d}{d\psi}\left\{\frac{1}{B(\psi)}\right\} \geq \frac{1}{2}, \tag{9.3.11}$$

for all ψ such that $B(\psi) \neq 0$. Since $B(\psi)$ is a decreasing function, either $B(\psi) = 0$ for all $\psi \geq \psi_0$ or $B(\psi) \neq 0$ for all $\psi \geq \psi_0$. Let $G(\psi)$ be a function defined so that $G(\psi_0) = 1/B(\psi_0)$ and $G'(\psi) = 1/2$ for all $\psi \geq \psi_0$; i.e.,

$$G(\psi) = \frac{1}{B(\psi_0)} + \frac{1}{2}(\psi - \psi_0), \quad \text{all } \psi \geq \psi_0. \tag{9.3.12}$$

Since $1/B(\psi)$ is an increasing function and $\dfrac{d}{d\psi}(1/B(\psi)) \geq 1/2$, it is always above $G(\psi)$ on $\psi \geq \psi_0$. It follows that

$$\underline{\lim}_{\psi\to\infty}\,(1/B(\psi)) = \infty, \quad \text{or} \quad \overline{\lim}_{\psi\to\infty}\, B(\psi) = 0. \tag{9.3.13}$$

In a similar manner, we can show that $\overline{\lim_{\psi \to \infty}}(1/B(\psi)) = \infty$ or $\underline{\lim_{\psi \to -\infty}} B(\psi) = 0$. This implies that $B(\psi) = 0$ for all ψ. Finally, since the bias function of $\hat{\mu}(X) = X$ is also identically zero we obtain from the previous lemma that $\hat{\mu}(X)$ is admissible. QED

Karlin (1958) established sufficient condition for the admissibility of a linear estimator

$$\hat{\theta}_{\lambda,\gamma} = \frac{1}{1+\lambda} T + \frac{\gamma\lambda}{1+\lambda} \tag{9.3.14}$$

of the expected value of T in the one parameter exponential family, with p.d.f.

$$f(x;\theta) = \beta(\theta) e^{\theta T(x)}, \quad \underline{\theta} \leq \theta \leq \bar{\theta}.$$

Theorem 9.3.1 (Karlin). *Let X have a one-parameter exponential distribution, with $\underline{\theta} \leq \theta \leq \bar{\theta}$. Sufficient conditions for the admissibility of (9.3.14) as estimator of $E_\theta\{T(X)\}$, under squared-error loss, is*

$$\lim_{\theta^* \to \bar{\theta}} \int_{\theta_0}^{\theta^*} \frac{e^{-\gamma\lambda\theta}}{[\beta(\theta)]^\lambda} d\theta = \infty$$

or

$$\lim_{\theta^* \to \underline{\theta}} \int_{\theta^*}^{\theta_0} \frac{e^{-\gamma\lambda\theta}}{[\beta(\theta)]^\lambda} d\theta = \infty,$$

where $\underline{\theta} < \theta_0 < \bar{\theta}$.

For a proof of this theorem, see Lehmann and Casella (1998, p. 331).

Considerable amount of research was conducted on the question of the admissibility of formal or generalized Bayes estimators. Some of the important results will be discussed later. We address ourselves here to the question of the admissibility of equivariant estimators of the location parameter in the one-dimensional case. We have seen that the minimum risk equivariant estimator of a location parameter θ, when finite risk equivariant estimators exist, is the Pitman estimator

$$\hat{\theta}(\mathbf{X}) = X_{(1)} - E\{X_{(1)} \mid X_{(2)} - X_{(1)}, \ldots, X_{(n)} - X_{(1)}\}.$$

The question is whether this estimator is admissible. Let $\mathbf{Y} = (X_{(2)} - X_{(1)}, \ldots, X_{(n)} - X_{(1)})$ denote the maximal invariant statistic and let $f(x \mid \mathbf{y})$ the conditional distribution of $X_{(1)}$, when $\theta = 0$, given $\mathbf{Y} = \mathbf{y}$. Stein (1959) proved the following.

Theorem 9.3.2. *If $\hat{\theta}(\mathbf{X})$ is the Pitman estimator and*

$$E\{[E\{[X_{(1)} - E\{X_{(1)} \mid \mathbf{Y}\}]^2 \mid \mathbf{Y}\}]^{3/2}\} < \infty, \tag{9.3.15}$$

then $\hat{\theta}(\mathbf{X})$ is an admissible estimator of θ with respect to the squared-error loss.

We omit the proof of this theorem, which can be found in Stein's paper (1959) or in Zacks (1971, pp. 388–393). The admissibility of the Pitman estimator of a two-dimensional location parameter was proven later by James and Stein (1960). The Pitman estimator is not admissible, however, if the location parameter is a vector of order $p \geq 3$. This result, first established by Stein (1956) and by James and Stein (1960), will be discussed in the next section.

The Pitman estimator is a formal Bayes estimator. It is admissible in the real parameter case. The question is under what conditions formal Bayes estimators in general are admissible. Zidek (1970) established sufficient conditions for the admissibility of formal Bayes estimators having a bounded risk.

9.3.2 The Inadmissibility of Some Commonly Used Estimators

In this section, we discuss a few well-known examples of some MLE or best equivariant estimators that are inadmissible. The first example was developed by Stein (1956) and James and Stein (1960) established the inadmissibility of the MLE of the normal mean vector $\boldsymbol{\theta}$, in the $N(\theta, I)$ model, when the dimension of $\boldsymbol{\theta}$ is $p \geq 3$. The loss function considered is the squared-error loss, $L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = |\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}|^2$. This example opened a whole area of research and led to the development of a new type of estimator of a location vector, called the Stein estimators. Another example that will be presented establishes the inadmissibility of the best equivariant estimator of the variance of a normal distribution when the mean is unknown. This result is also due to Stein (1964). Other related results will be mentioned too.

I. The Inadmissibility of the MLE in the $N(\boldsymbol{\theta}, I)$ Case, With $p \geq 3$

Let $\mathbf{X}$ be a random vector of p components, with $p \geq 3$. Furthermore assume that $\mathbf{X} \sim N(\boldsymbol{\theta}, I)$. The assumption that the covariance matrix of X is I, is not a restrictive one, since if $\mathbf{X} \sim N(\boldsymbol{\theta}, V)$, with a known V, we can consider the case of $\mathbf{Y} = C^{-1}\mathbf{X}$, where $V = CC'$. Obviously, $\mathbf{Y} \sim N(\boldsymbol{\eta}, I)$ where $\boldsymbol{\eta} = C^{-1}\boldsymbol{\theta}$. Without loss of generality, we also assume that the sample size is $n = 1$. The MLE of $\boldsymbol{\theta}$ is $\mathbf{X}$ itself. Consider the squared-error loss function $L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = ||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2$. Since $\mathbf{X}$ is unbiased, the risk of the MLE is $R^* = p$ for all $\boldsymbol{\theta}$. We show now an estimator that has a risk function smaller than p for all $\boldsymbol{\theta}$, and when $\boldsymbol{\theta}$ is close to zero its risk is close to 2. The estimator suggested by Stein is

$$\hat{\boldsymbol{\theta}} = \left(1 - \frac{p-2}{\mathbf{X}'\mathbf{X}}\right) X. \tag{9.3.16}$$

This estimator is called the James–Stein estimator. The risk function of (9.3.16) is

$$\begin{aligned} R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= E_{\boldsymbol{\theta}} \left\{ \left|\left| \mathbf{X} - \boldsymbol{\theta} - \frac{p-2}{\mathbf{X}'\mathbf{X}} \mathbf{X} \right|\right|^2 \right\} \\ &= E_{\boldsymbol{\theta}}\{||\mathbf{X} - \boldsymbol{\theta}||^2\} - 2(p-2) E_{\boldsymbol{\theta}} \left(\frac{\mathbf{X}'(\mathbf{X} - \boldsymbol{\theta})}{\mathbf{X}'\mathbf{X}} \right) \\ &+ (p-2)^2 E_{\boldsymbol{\theta}} \left\{ \frac{1}{\mathbf{X}'\mathbf{X}} \right\}. \end{aligned} \tag{9.3.17}$$

The first term on the RHS of (9.3.17) is p. We notice that $\mathbf{X}'\mathbf{X} \sim \chi^2[p; \frac{1}{2}\boldsymbol{\theta}'\boldsymbol{\theta}]$. Accordingly,

$$\begin{aligned} E_{\boldsymbol{\theta}} \left\{ \frac{1}{\mathbf{X}'\mathbf{X}} \right\} &= E_{\boldsymbol{\theta}}\{E\{(\chi^2[p+2J])^{-1}\}\} \\ &= E_{\boldsymbol{\theta}} \left\{ \frac{1}{p-2+2J} \right\}, \end{aligned} \tag{9.3.18}$$

where $J \sim P(\frac{\boldsymbol{\theta}'\boldsymbol{\theta}}{2})$. We turn now to the second term on the RHS of (9.3.17). Let $U = \mathbf{X}'\boldsymbol{\theta}/||\boldsymbol{\theta}||$ and $V = \left|\left| \mathbf{X} - \frac{U}{||\boldsymbol{\theta}||}\boldsymbol{\theta} \right|\right|^2$. Note that $U \sim N(||\boldsymbol{\theta}||, 1)$ is independent of V and $V \sim \chi^2[p-1]$. Indeed, we can write

$$V = \mathbf{X}' \left(I - \frac{\boldsymbol{\theta}\boldsymbol{\theta}'}{||\boldsymbol{\theta}||^2} \right) \mathbf{X}, \tag{9.3.19}$$

where $A = (I - \boldsymbol{\theta}\boldsymbol{\theta}'/||\boldsymbol{\theta}||^2)$ is an idempotent matrix of rank $p-1$. Hence, $V \sim \chi^2[p-1]$. Moreover, $A\boldsymbol{\theta}/||\boldsymbol{\theta}|| = \mathbf{0}$. Hence, U and V are independent. Furthermore,

$$\begin{aligned} ||\mathbf{X}||^2 &= \left|\left| \mathbf{X} - U\frac{\boldsymbol{\theta}}{||\boldsymbol{\theta}||} + U\frac{\boldsymbol{\theta}}{||\boldsymbol{\theta}||} \right|\right|^2 \\ &= U^2 + V + 2\mathbf{X}'A\boldsymbol{\theta}/||\boldsymbol{\theta}|| = U^2 + V. \end{aligned} \tag{9.3.20}$$

We let $W = ||\mathbf{X}||^2$, and derive the p.d.f. of U/W. This is needed, since the second term on the RHS of (9.3.17) is $-2(p-2)[1 - ||\boldsymbol{\theta}||E_{\boldsymbol{\theta}}\{U/W\}]$. Since U and V are independent, their joint p.d.f. is

$$\begin{aligned} f(u, v \mid \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}(u - ||\theta||)^2 \right\} \cdot \\ \cdot \frac{1}{2^p - 1/2\Gamma(\frac{p-1}{2})} v^{\frac{p-3}{2}} \exp\left\{ -\frac{1}{2}v \right\}. \end{aligned} \tag{9.3.21}$$

Thus, the joint p.d.f. of U and W is

$$g(u, w; \boldsymbol{\theta}) = \frac{(w - u^2)^{\frac{p-3}{2}}}{\sqrt{2\pi}\, 2^{\frac{p-1}{2}} \Gamma(\frac{p-1}{2})} \exp\left\{-\frac{1}{2}||\boldsymbol{\theta}||^2 - \frac{1}{2}w + ||\boldsymbol{\theta}||u\right\}, \tag{9.3.22}$$

$0 \le u^2 \le w \le \infty$. The p.d.f. of $R = U/W$ is then

$$h(r \mid \boldsymbol{\theta}) = \int_0^\infty w g(rw, w; \boldsymbol{\theta}) dw. \tag{9.3.23}$$

Accordingly,

$$\begin{aligned} E_{\boldsymbol{\theta}}\left\{\frac{U}{W}\right\} = \frac{e^{-\frac{1}{2}||\boldsymbol{\theta}||^2}}{\sqrt{2\pi}\, 2^{\frac{p-1}{2}} \Gamma(\frac{p-1}{2})} \int_0^\infty dw \cdot \\ \cdot \int_{-\sqrt{w}}^{\sqrt{w}} \frac{u}{w}(w - u^2)^{\frac{p-3}{2}} \exp\left\{||\boldsymbol{\theta}||\, u - \frac{1}{2}w\right\} du. \end{aligned} \tag{9.3.24}$$

By making the change of variables to $t = u/\sqrt{w}$ and expanding $\exp\{||\boldsymbol{\theta}||t\sqrt{w}\}$ we obtain, after some manipulations,

$$\begin{aligned} E_\theta\left(\frac{(\mathbf{X} - \boldsymbol{\theta})'\mathbf{X}}{\mathbf{X}'\mathbf{X}}\right) &= 1 - ||\boldsymbol{\theta}|| E_\theta\left\{\frac{U}{W}\right\} \\ &= 1 - ||\boldsymbol{\theta}|| \frac{\exp\{-\frac{1}{2}||\boldsymbol{\theta}||^2\}}{\sqrt{\pi}} \sum_{j=0}^\infty \frac{\Gamma(\frac{1}{2})||\boldsymbol{\theta}||^{2j+1}}{2^{j+1}\Gamma(j+1)(\frac{p}{2} + j)} \\ &= \exp\left\{-\frac{1}{2}||\boldsymbol{\theta}||^2\right\} \sum_{j=0}^\infty \frac{(\frac{1}{2}||\boldsymbol{\theta}||^2)^j}{j!} \cdot \frac{p-2}{p-2+2j} \\ &= E_{\boldsymbol{\theta}}\left(\frac{p-2}{p-2+2J}\right), \end{aligned} \tag{9.3.25}$$

where $J \sim P(\frac{1}{2}\boldsymbol{\theta}'\boldsymbol{\theta})$. From (9.3.17), (9.3.18) and (9.3.25), we obtain

$$R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = p - E_{\boldsymbol{\theta}}\left(\frac{(p-2)^2}{p-2+2J}\right) < p, \quad \text{all } \theta. \tag{9.3.26}$$

Note that when $\boldsymbol{\theta} = \mathbf{0}$, $P_{\mathbf{0}}[J = 0] = 1$ and $R(\boldsymbol{\theta}, \mathbf{0}) = 2$. On the other hand, $\lim_{||\boldsymbol{\theta}|| \to \infty} R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = p$. The estimator $\hat{\boldsymbol{\theta}}$ given by (9.3.16) has smaller risk than the MLE for all $\boldsymbol{\theta}$ values. In the above development, there is nothing to tell us whether (9.3.16) is itself admissible. Note that (9.3.16) is not an equivariant estimator with respect to the group of real affine transformations, but it is equivariant with respect

to the group of orthogonal transformations (rotations). If the vector $\mathbf{X}$ has a known covariance matrix V, the estimator (9.3.16) should be modified to

$$\hat{\theta}(\mathbf{X}) = \left(1 - \frac{p-2}{\mathbf{X}'V^{-1}\mathbf{X}}\right)\mathbf{X}. \tag{9.3.27}$$

This estimator is equivariant with respect to the group $\mathcal{G}$ of nonsingular transformations $\mathbf{X} \to A\mathbf{X}$. Indeed, the covariance matrix of $\mathbf{Y} = A\mathbf{X}$ is $\Sigma = AVA'$. Therefore, $Y'\Sigma^{-1}Y = X'V^{-1}X$ for every $A \in \mathcal{G}$.

Baranchick (1973) showed, in a manner similar to the above, that in the usual multiple regression model with normal distributions the commonly used MLEs of the regression coefficients are inadmissible. More specifically, let $X_1, \ldots, X_n$ be a sample of n i.i.d. $(p+1)$ dimensional random vectors, having a multinormal distribution $N(\boldsymbol{\theta}, \Sigma)$. Consider the regression of $Y = X_1$ on $\mathbf{Z} = (X_2, \ldots, X_{p+1})'$. If we consider the partition $\boldsymbol{\theta}' = (\eta, \boldsymbol{\zeta}')$ and

$$\Sigma = \left(\begin{array}{c|c} \tau^2 & \mathbf{C}' \\ \hline \mathbf{C} & V \end{array}\right),$$

then the regression of Y on Z is given by

$$E\{Y \mid \mathbf{Z}\} = \alpha + \boldsymbol{\beta}'\mathbf{Z},$$

where $\alpha = \eta - \boldsymbol{\beta}'\boldsymbol{\zeta}$ and $\boldsymbol{\beta} = V^{-1}\mathbf{C}$. The problem is to estimate the vector of regression coefficients $\boldsymbol{\beta}$. The least-squares estimators (LSE) is $\hat{\boldsymbol{\beta}} = S^{-1}\left(\sum_{i=1}^{n} Y_i\mathbf{Z}_i - n\bar{Y}\bar{\mathbf{Z}}\right)$, where $S = \sum_{i=1}^{n}\mathbf{Z}_i\mathbf{Z}_i' - n\bar{\mathbf{Z}}\bar{\mathbf{Z}}'$. $Y_1, \ldots, Y_n$ and $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ are the sample statistics corresponding to $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

Consider the loss function

$$L(\hat{\alpha}, \hat{\boldsymbol{\beta}}; \alpha, \boldsymbol{\beta}, \Sigma) = [(\hat{\alpha} - \alpha) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{\zeta}]^2 + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'V(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \div (\tau^2 - \mathbf{C}'V^{-1}\mathbf{C}). \tag{9.3.28}$$

With respect to this loss function Baranchick proved that the estimators

$$\begin{aligned} \hat{\boldsymbol{\beta}}_c &= \left(1 - c\frac{1-R^2}{R^2}\right)\hat{\boldsymbol{\beta}} \\ \hat{\alpha}_c &= \bar{Y} - \hat{\boldsymbol{\beta}}_c'\mathbf{Z}, \end{aligned} \tag{9.3.29}$$

have risk functions smaller than that of the LSEs (MLEs) $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha} = \bar{Y} - \hat{\boldsymbol{\beta}}'\bar{Z}$, at all the parameter values, provided $c \in \left(0, \frac{2(p-2)}{n-p+2}\right)$ and $p \geq 3, n \geq p+2$. R^2 is the squared-multiple correlation coefficient given by

$$R^2 = \left(\sum_{i=1}^{n} Y_i \mathbf{Z}_i - n\bar{Y}\bar{\mathbf{Z}}\right)' S^{-1} \left(\sum_{i=1}^{n} Y_i \mathbf{Z}_i - n\bar{Y}\bar{\mathbf{Z}}\right) / \left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right).$$

The proof is very technical and is omitted. The above results of Stein and Baranchick on the inadmissibility of the MLEs can be obtained from the following theorem of Cohen (1966) that characterizes all the admissible linear estimate of the mean vector of multinormal distributions. The theorem provides only the conditions for the admissibility of the estimators, and contrary to the results of Stein and Baranchick, it does not construct alternative estimators.

Theorem 9.3.3 (Cohen, 1966). *Let $X \sim N(\boldsymbol{\theta}, I)$ where the dimension of $\mathbf{X}$ is p. Let $\hat{\boldsymbol{\theta}} = A\mathbf{X}$ be an estimator of $\boldsymbol{\theta}$, where A is a $p \times p$ matrix of known coefficients. Then $\hat{\boldsymbol{\theta}}$ is admissible with respect to the squared-error loss $||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2$ if and only if A is symmetric and its eigenvalues $\alpha_i (i = 1, \ldots, p)$ satisfy the inequality.*

$$0 \leq \alpha_i \leq 1, \quad \textit{for all } i = 1, \ldots, p, \tag{9.3.30}$$

with equality to 1 for at most two of the eigenvalues.

For a proof of the theorem, see Cohen (1966) or Zacks (1971, pp. 406–408). Note that for the MLE of $\boldsymbol{\theta}$ the matrix A is I, and all the eigenvalues are equal to 1. Thus, if $p \geq 3$, $\mathbf{X}$ is an inadmissible estimator. If we shrink the MLE towards the origin and consider the estimator $\boldsymbol{\theta}_\lambda = \lambda\mathbf{X}$ with $0 < \lambda < 1$ then the resulting estimator is admissible for any dimension p. Indeed, $\hat{\boldsymbol{\theta}}_\lambda$ is actually the Bayes estimator (8.4.31) with $A_1 = V = I$, $\Sigma = \tau^2 I$ and $A_2 = 0$. In this case, the Bayes estimator is $\hat{\boldsymbol{\beta}}_\tau = \frac{\tau^2}{1+\tau^2}\mathbf{X}$, where $0 < \tau < \infty$. We set $\lambda = \tau^2/(1+\tau^2)$. According to Lemma 9.3.1, this proper Bayes estimator is admissible. In Section 9.3.3, we will discuss more meaningful adjustment of the MLE to obtain admissible estimators of $\boldsymbol{\theta}$.

II. The Inadmissibility of the Best Equivariant Estimators of the Scale Parameter When the Location Parameter is Unknown

Consider first the problem of estimating the variance of a normal distribution $N(\mu, \sigma^2)$ when the mean μ is unknown. Let $X_1, \ldots, X_n$ be i.i.d. random variables having such distribution. Let $(\bar{X}, Q)$ be the minimal sufficient statistic, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $Q = \sum_{i=1}^{n}(X_i - \bar{X})^2$. We have seen that the minimum risk equivariant estimator, with

respect to the quadratic loss $L(\hat{\sigma}^2, \sigma^2) = \frac{1}{\sigma^4}(\hat{\sigma}^2 - \sigma^2)^2$ is $\hat{\sigma}_0^2 = Q/(n+1)$. Stein (1964) showed that this estimator is, however, inadmissible! The estimator

$$\hat{\sigma}_1^2 = \min\left(\frac{Q + n\bar{X}^2}{n+2}, \frac{Q}{n+1}\right) \tag{9.3.31}$$

has uniformly smaller risk function. We present here Stein's proof of this inadmissibility.

Let $S = Q + n\bar{X}^2$. Obviously, $S \sim \chi^2[n; n\mu^2/2\sigma^2] \sim \chi^2[n+2J]$ where $J \sim P(n\mu^2/2\sigma^2)$. Consider the scale equivariant estimators that are functions of (Q, S). Their structure is $f(Q, S) = S\phi(\frac{Q}{S})$. Moreover, the conditional distribution of Q/S given J is the beta distribution $\beta\left(\frac{n-1}{2}, \frac{1}{2} + J\right)$. Furthermore given J, Q/S and S are conditionally independent. Note that for $\hat{\sigma}_0^2$ we use the function $\phi_0\left(\frac{Q}{S}\right) = \frac{Q}{S(n+1)}$. Consider the estimator

$$\begin{aligned}\hat{\sigma}_1^2 &= \min\left\{\frac{Q}{n+1}, \frac{S}{n+2}\right\} \\ &= S\min\left\{\frac{1}{n+1}\cdot\frac{Q}{S}, \frac{1}{n+2}\right\}.\end{aligned} \tag{9.3.32}$$

Here, $\phi_1\left(\frac{Q}{S}\right) = \min\left\{\frac{1}{n+1}\frac{Q}{S}, \frac{1}{n+2}\right\}$. The risk function, for the quadratic loss $L(\hat{\sigma}, \sigma^2) = (\hat{\sigma}^2 - \sigma^2)^2/\sigma^4$ is, for any function $\phi\left(\frac{Q}{S}\right)$,

$$R(\phi) = E\left\{\left[\chi_2^2[n+2J]\phi\left(\frac{\chi_1^2[n-1]}{\chi_2^2[n+2J]}\right) - 1\right]^2\right\}, \tag{9.3.33}$$

where $Q \sim \sigma^2\chi_1^2[n-1]$ and $S \sim \sigma^2\chi_2^2[n+2J]$. Let $W = Q/S$. Then,

$$\begin{aligned}R(\phi) &= E\{E\{[\chi_2^2[n+2J]\phi(W) - 1]^2 \mid J, W\} \\ &= E\{\phi^2(W)(n+2J)(n+2J+2) - 2\phi(W)(n+2J) + 1\} \\ &= E\left\{(n+2J)(n+2J+2)\left[\phi(W) - \frac{1}{n+2J+2}\right]^2 + \frac{2}{n+2J+2}\right\}.\end{aligned} \tag{9.3.34}$$

We can also write,

$$\left(\phi_0(W) - \frac{1}{n+2J+2}\right)^2 = \left(\phi_0(W) - \phi_1(W)\right)^2 + \left(\phi_1(W) - \frac{1}{n+2J+2}\right)^2 + 2(\phi_0(W) - \phi_1(W))\left(\phi_1(W) - \frac{1}{n+2J+2}\right). \tag{9.3.35}$$

We notice that $\phi_1(W) \leq \phi_0(W)$. Moreover, if $\frac{W}{n+1} \leq \frac{1}{n+2}$ then $\phi_1(W) = \phi_0(W)$, and the first and third terms on the RHS of (9.3.35) are zero. Otherwise,

$$\phi_0(W) > \phi_1(W) = \frac{1}{n+2} \geq \frac{1}{n+2+2J}, \quad j = 0, 1, \ldots.$$

Hence,

$$\left(\phi_0(W) - \frac{1}{n+2J+2}\right)^2 \geq \left(\phi_1(W) - \frac{1}{n+2J+2}\right)^2 \tag{9.3.36}$$

for all J and W, with strict inequality on a (J, W) set having positive probability. From (9.3.34) and (9.3.36) we obtain that $R(\phi_1) < R(\phi_0)$. This proves that $\hat{\sigma}_0^2$ is inadmissible.

The above method of Stein can be used to prove the inadmissibility of equivariant estimators of the variance parameters also in other normal models. See for example, Klotz, Milton and Zacks (1969) for a proof of the inadmissibility of equivariant estimators of the variance components in Model II of ANOVA.

Brown (1968) studied the question of the admissibility of the minimum risk (best) equivariant estimators of the scale parameter σ in the general location and scale parameter model, with p.d.f.s $f(x; \mu, \sigma) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right); -\infty < \mu < \infty, 0 < \sigma < \infty$. The loss functions considered are invariant bowl-shaped functions, $L(\delta)$. These are functions that are nonincreasing for $\delta \leq \delta_0$ and nondecreasing for $\delta > \delta_0$ for some δ_0. Given the order statistic $X_{(1)} \leq \cdots \leq X_{(n)}$ of the sample, let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_{(i)}$, $S = \left\{\frac{1}{n}\sum_{i=1}^{n}(X_{(i)} - \bar{X})^2\right\}^{1/2}$ and $Z_i = (X_{(i)} - \bar{X})/S$, $i = 3, \ldots, n$. $\mathbf{Z} = (Z_3, \ldots, Z_n)'$ is a maximal invariant with respect to the group $\mathcal{G}$ of real affine transformations. The

best equivariant estimator of $\omega = \sigma^k$ is of the form $\hat{\omega}_0 = \phi_0(\mathbf{Z})S^k$, where $\phi_0(\mathbf{Z})$ is an optimally chosen function. Brown proved that the estimator

$$\hat{\omega}_1(S, \mathbf{Z}) = \begin{cases} \phi_1(\mathbf{Z})S^k, & \text{if } \left|\frac{\bar{X}}{S}\right| < K(\mathbf{Z}), \\ \phi_0(\mathbf{Z})S^k, & \text{if } \left|\frac{\bar{X}}{S}\right| \geq K(\mathbf{Z}), \end{cases} \tag{9.3.37}$$

where $K(\mathbf{Z})$ is appropriately chosen functions, and $\phi_1(\mathbf{Z}) < \phi_0(\mathbf{Z})$ has uniformly smaller risk than $\hat{\omega}_0$. This established the inadmissibility of the best equivariant estimator, when the location parameter is unknown, for general families of distributions and loss functions. Arnold (1970) provided a similar result in the special case of the family of shifted exponential distributions, i.e., $f(x; \mu, \sigma) = I\{x \geq \mu\}\frac{1}{\sigma}\exp\{-\frac{x-\mu}{\sigma}\}$, $-\infty < \mu < \infty, 0 < \sigma < \infty$.

Brewster and Zidek (1974) showed that in certain cases one can refine Brown's approach by constructing a sequence of improving estimators converging to a generalized Bayes estimator. The risk function of this estimator does not exceed that of the best equivariant estimators. In the normal case $N(\mu, \sigma^2)$, this estimator is of the form $\phi^*(Z)Q$, where

$$\phi^*(z) = E\{Q \mid Z \leq z\}/E\{Q^2 \mid Z \leq z\}, \tag{9.3.38}$$

with $Q = \sum_{i=1}^{n}(X_i - \bar{X})^2$, $Z = \sqrt{n}|\bar{X}|/\sqrt{Q}$. The conditional expectations in (9.3.38) are computed with $\mu = 0$ and $\sigma = 1$. Brewster and Zidek (1974) provided a general group theoretic framework for deriving such estimators in the general case.

9.3.3 Minimax and Admissible Estimators of the Location Parameter

In Section 9.3.1, we presented the James–Stein proof that the MLE of the location parameter vector in the $N(\theta, I)$ case with dimension $p \geq 3$ is inadmissible. It was shown that the estimator (9.3.16) is uniformly better than the MLE. The estimator (9.3.16) is, however, also inadmissible. Several studies have been published on the question of adjusting estimator (9.3.16) to obtain minimax estimators. In particular, see Berger and Bock (1976). Baranchick (1970) showed that a family of minimax estimators of θ is given by

$$\hat{\boldsymbol{\theta}}_\phi = \left(1 - \frac{(p-2)\phi(S)}{S}\right)\mathbf{X}, \tag{9.3.39}$$

where $S = \mathbf{X}'\mathbf{X}$ and $\phi(S)$ is a function satisfying the conditions:

$$\begin{aligned} &\textbf{(i)} \quad 0 \le \phi(S) \le 2; \\ &\textbf{(ii)} \quad \phi(S) \text{ is nondecreasing in } S. \end{aligned} \tag{9.3.40}$$

If the model is $N(\boldsymbol{\theta}, \sigma^2 I)$ with known σ^2 then the above result holds with $S = \mathbf{X}'\mathbf{X}/\sigma^2$. If σ^2 is unknown and $\hat{\sigma}^2$ is an estimator of σ^2 having a distribution like $\sigma^2 \cdot \chi^2[\nu]/(\nu + 2)$ then we substitute in (9.3.39) $S = \mathbf{X}'\mathbf{X}/\hat{\sigma}^2$. The minimaxity of (9.3.39) is established by proving that its risk function, for the squared-error loss, does not exceed the constant risk, $R^* = p$, of the MLE $\mathbf{X}$. Note that the MLE, $\mathbf{X}$, is also minimax. In addition, (9.3.39) can be improved by

$$\boldsymbol{\theta}_{\phi}^{+} = \left(1 - \frac{(p-2)\phi(\hat{S})}{\hat{S}}\right)^{+} \mathbf{X}, \tag{9.3.41}$$

where $a^+ = \max(a, 0)$. These estimators are not necessarily admissible. Admissible and minimax estimators of θ similar to (9.3.39) were derived by Strawderman (1972) for cases of known σ^2 and $p \ge 5$. These estimators are

$$\boldsymbol{\theta}_a(\mathbf{X}) = \left(1 - \frac{p - 2a + 2}{S} \cdot \phi(S)\right) \mathbf{X}, \tag{9.3.42}$$

where $\frac{1}{2} \le a \le 1$ for $p = 5$ and $0 \le a \le 1$ for $p \ge 6$, also

$$\phi(S) = \frac{G(\frac{S}{2}; \frac{p}{2} - a + 2)}{G(\frac{s}{s}; \frac{p}{2} - a + 1)},$$

in which $G(x; \nu) = P\{G(1, \nu) \le x\}$. In Example 9.9, we show that $\boldsymbol{\theta}_a(\mathbf{X})$ are generalized Bayes estimators for the squared-error loss and the hyper-prior model

$$\begin{aligned} &\textbf{(i)} \quad \boldsymbol{\theta} \mid \lambda \sim N\left(\mathbf{0}, \frac{1-\lambda}{\lambda} I\right), \quad 0 < \lambda \le 1; \\ &\textbf{(ii)} \quad \lambda \text{ has a generalized prior measure } G(\lambda) \text{ so that} \end{aligned} \tag{9.3.43}$$

$$dG(\lambda) \propto \lambda^{-a} d\lambda, \quad -\infty < a < \frac{p}{\lambda} + 1.$$

Note that if $a < 1$ then $G(\lambda)$ is a proper prior and $\hat{\boldsymbol{\theta}}(\mathbf{X})$ is a proper Bayes estimator. For a proof of the admissibility for the more general case of $a < \frac{p}{2} + 1$, see Lin (1974).

9.3.4 The Relationship of Empirical Bayes and Stein-Type Estimators of the Location Parameter in the Normal Case

Efron and Morris (1972a, 1972b, 1973) show the connection between the estimator (9.3.16) and empirical Bayes estimation of the mean vector of a multinormal distribution. Ghosh (1992) present a comprehensive comparison of Empirical Bayes and the Stein-type estimators, in the case where $\mathbf{X} \sim N(\boldsymbol{\theta}, I)$. See also the studies of Lindley and Smith (1972), Casella (1985), and Morris (1983).

Recall that in parametric empirical Bayes procedures, one estimates the unknown prior parameters, from a model of the predictive distribution of $\mathbf{X}$, and substitutes the estimators in the formulae of the Bayesian estimators. On the other hand, in hierarchical Bayesian procedures one assigns specific hyper prior distributions for the unknown parameters of the prior distributions. The two approaches may sometimes result with similar estimators.

Starting with the simple model of p-variate normal $\mathbf{X} \mid \boldsymbol{\theta} \sim N(\boldsymbol{\theta}, I)$, and $\boldsymbol{\theta} \sim N(\mathbf{0}, \tau^2 I)$, the Bayes estimator of $\boldsymbol{\theta}$, for the squared-error loss $L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = ||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2$, is $\hat{\boldsymbol{\theta}}_B = (1 - B)\mathbf{X}$, where $B = 1/(1 + \tau^2)$. The predictive distribution of $\mathbf{X}$ is $N(\mathbf{0}, B^{-1} I)$. Thus, the predictive distribution of $\mathbf{X}'\mathbf{X}$ is like that of $B^{-1}\chi^2[p]$. Thus, for $p > 3$, $(p-2)/\mathbf{X}'\mathbf{X}$ is predictive-unbiased estimator of B. Substituting this estimator for B in $\hat{\boldsymbol{\theta}}_B$ yields the parametric empirical Bayes estimator

$$\hat{\theta}_{EB}^{(1)} = \left(1 - \frac{p-2}{\mathbf{X}'\mathbf{X}}\right)\mathbf{X}. \tag{9.3.44}$$

$\hat{\theta}_{EB}$ derived here is identical with the James–Stein estimator (9.3.16). If we change the Bayesian model so that $\boldsymbol{\theta} \sim N(\mu \mathbf{1}, I)$, with μ known, then the Bayesian estimator is $\hat{\boldsymbol{\theta}}_B = (1 - B)\mathbf{X} + B\mu\mathbf{1}$, and the corresponding empirical Bayes estimator is

$$\begin{aligned} \hat{\theta}_{EB}^{(2)} &= \left(1 - \frac{p-2}{(X - \mu 1)'(\mathbf{X} - \mu 1)}\right)\mathbf{X} + \frac{(p-2)}{(\mathbf{X} - \mu\mathbf{1})'(\mathbf{X} - \mu\mathbf{1})}\mu\mathbf{1} \\ &= \mathbf{X} - \frac{p-2}{||\mathbf{X} - \mu\mathbf{1}||^2}(\mathbf{X} - \mu\mathbf{1}). \end{aligned} \tag{9.3.45}$$

If both μ and τ are unknown, the resulting empirical Bayes estimator is

$$\hat{\theta}_{EB}^{(3)} = \bar{X} - \frac{p-3}{\sum_{i=1}^{p}(X_i - \bar{X})^2}(\mathbf{X} - \bar{X}\mathbf{1}), \tag{9.3.46}$$

where $\bar{X} = \frac{1}{p}\sum_{i=1}^{p} X_i$.

Ghosh (1992) showed that the predictive risk function of $\hat{\theta}_{EB}^{(3)}$, namely, the trace of the MSE matrix $E\{(\hat{\boldsymbol{\theta}}_{EB}^{(3)} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_E - \boldsymbol{\theta})'\}$ is

$$R(\hat{\boldsymbol{\theta}}_{EB}^{(3)}, \boldsymbol{\theta}) = p - \frac{p-3}{1+\tau^2}. \tag{9.3.47}$$

Thus, $\hat{\boldsymbol{\theta}}_{EB}^{(3)}$ has smaller predictive risk than the MLE $\hat{\boldsymbol{\theta}}_{ML} = \mathbf{X}$ if $p \geq 4$.

The Stein-type estimator of the parametric vector $\boldsymbol{\beta}$ in the linear model $\mathbf{X} \sim A\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is

$$\hat{\boldsymbol{\beta}}^* = (1 - B)\hat{\boldsymbol{\beta}}, \tag{9.3.48}$$

where $\hat{\boldsymbol{\beta}}$ is the LSE and

$$B = \min(1, (p-2)\sigma^2 / \hat{\boldsymbol{\beta}}' A' A \hat{\boldsymbol{\beta}}). \tag{9.3.49}$$

It is interesting to compare this estimator of β with the ridge regression estimator (5.4.3), in which we substitute for the optimal k value the estimator $p\sigma^2/\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}$. The ridge regression estimators obtains the form

$$\hat{\boldsymbol{\beta}}^{**} = \left(I - \frac{\sigma^2 p}{\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}}(A'A)^{-1}\right)\hat{\boldsymbol{\beta}}. \tag{9.3.50}$$

There is some analogy but the estimators are obviously different. A comprehensive study of the property of the Stein-type estimators for various linear models is presented in the book of Judge and Bock (1978).

PART II: EXAMPLES

Example 9.1. Let X be a binomial $B(n, \theta)$ random variable. n is known, $0 < \theta < 1$. If we let θ have a prior beta distribution, i.e., $\theta \sim \beta(\nu_1, \nu_2)$ then the posterior distribution of θ given X is the beta distribution $\beta(\nu_1 + X, \nu_2 + n - X)$. Consider the linear estimator $\hat{\theta}_{\alpha,\beta} = \frac{\alpha}{n}X + \beta$. The MSE of $\hat{\theta}_{\alpha,\beta}$ is

$$\begin{aligned} R(\hat{\theta}_{\alpha,\beta}, \theta) = \beta^2 &+ \frac{\theta}{n}[1 - 2(1-\alpha) + (1-\alpha)^2 - 2n\beta(1-\alpha)] \\ &- \frac{\theta^2}{n}[1 - 2(1-\alpha) + (1-\alpha)^2(1-n)]. \end{aligned}$$

We can choose α^0 and β^0 so that $R(\hat{\theta}_{\alpha^0,\beta^0}, \theta) = (\beta^0)^2$. For this purpose, we set the equations

$$1 - 2(1-\alpha) + (1-\alpha)^2 - 2n\beta(1-\alpha) = 0,$$
$$1 - 2(1-\alpha) + (1-\alpha)^2(1-n) = 0.$$

The two roots are

$$\alpha^0 = \sqrt{n}/(1+\sqrt{n}),$$
$$\beta^0 = \frac{1}{2}/(1+\sqrt{n}).$$

With these constants, we obtain the estimator

$$\theta^* = \frac{1}{\sqrt{n}(1+\sqrt{n})}X + \frac{1}{2(1+\sqrt{n})},$$

with constant risk

$$R(\theta^*, \theta) = \frac{1}{4(1+\sqrt{n})^2}, \quad \text{for all } \theta.$$

We show now that θ^* is a minimax estimator of θ for a squared-error loss by specifying a prior beta distribution for which θ^* is Bayes.

The Bayes estimator for the prior $\beta(\nu_1, \nu_2)$ is

$$\hat{\theta}_{\nu_1,\nu_2} = \frac{\nu_1 + X}{\nu_1 + \nu_2 + n} = \frac{1}{\nu_1 + \nu_2 + n}X + \frac{\nu_1}{\nu_1 + \nu_2 + n}.$$

In particular, if $\nu_1 = \nu_2 = \frac{\sqrt{n}}{2}$ then $\hat{\theta}_{\nu_1,\nu_2} = \theta^*$. This proves that θ^* is minimax.

Finally, we compare the MSE of this minimax estimator with the variance of the MVUE, X/n, which is also an MLE. The variance of $\tilde{\theta} = X/n$ is $\theta(1-\theta)/n$. $V\{\hat{\theta}\}$ at $\theta = 1/2$ assumes its maximal value of $1/4n$. This value is larger than $R(\theta^*, \theta)$. Thus, we know that around $\theta = 1/2$ the minimax estimator has a smaller MSE than the MVUE. Actually, by solving the quadratic equation

$$\theta^2 - \theta + n/4(1+\sqrt{n})^2 = 0,$$

we obtain the two limits of the interval around $\theta = 1/2$ over which the minimax estimator is better. These limits are given by

$$\theta_{1,2} = \frac{1}{2}\left(1 \pm \frac{\sqrt{1+2\sqrt{n}}}{1+\sqrt{n}}\right).$$

■

Example 9.2.

A. Let $X_1, \ldots, X_n$ be i.i.d. random variables, distributed like $N(\theta, 1)$, $-\infty < \theta < \infty$. The MVUE (or MLE), $\bar{X}$, has a constant variance n^{-1}. Thus, if our loss function is the squared-error, $\bar{X}$ is minimax. Indeed, the estimators $\hat{\theta}_k = \bar{X}\left(1 + \frac{1}{nk}\right)^{-1}$, $k = 1, 2, \ldots$, are Bayes with respect to the prior $N(0, k)$ distributions. The risks of these Bayesian estimators are

$$\rho(\hat{\theta}_k, k) = \frac{1}{n}\left(1 + \frac{1}{nk}\right)^{-2} + \frac{k}{(1+nk)^2}, \quad k = 1, 2, \ldots.$$

But $\rho(\hat{\theta}_k, k) \to n^{-1}$ as $k \to \infty$. This proves that $\bar{X}$ is minimax.

B. Consider the problem of estimating the common mean, μ of two normal distributions, which were discussed in Example 5.24. We can show that $(X + Y)/2$ is a minimax estimator for the symmetric loss function $(\hat{\mu} - \mu)^2/\sigma^2 \max(1, \rho)$. If the loss function is $(\hat{\mu} - \mu)^2/\sigma^2$ then the minimax estimator is $\bar{X}$, regardless of $\bar{Y}$. This is due to the large risk when $\rho \to \infty$ (see details in Zacks (1971, p. 291)). ■

Example 9.3. Consider the problem of estimating the variance components in the Model II of analysis of variance. We have k blocks of n observations on the random variables, which are represented by the linear model

$$Y_{ij} = \mu + a_i + e_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n,$$

where e_{ij} are i.i.d. $N(0, \sigma^2)$; $a_1, \ldots, a_k$ are i.i.d. random variables distributed like $N(0, \tau^2)$, independently of $\{e_{ij}\}$. In Example 3.3, we have established that a minimal sufficient statistic is $T = \left(\sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij}^2, \sum_{i=1}^{k} \bar{Y}_i^2, \bar{\bar{Y}}\right)$. This minimal sufficient statistic can be represented by $T^* = (Q_e, Q_a, \bar{Y})$, where

$$Q_e = \sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_i)^2 \sim \sigma^2 \chi_1^2[k(n-1)],$$

$$Q_a = n\sum_{i=1}^{k}(\bar{Y}_i - \bar{\bar{Y}})^2 \sim \sigma^2(1 + n\rho)\chi_2^2[k-1],$$

and

$$kn\bar{\bar{Y}}^2 \sim \sigma^2(1 + n\rho)\chi_3^2\left[1; \frac{nk\mu^2}{2\sigma^2(1+n\rho)}\right];$$

$\rho = \tau^2/\sigma^2$ is the variance ratio, $\chi_i^2[\cdot]$ $i = 1, 2, 3$ are three independent chi-squared random variables. Consider the group $\mathcal{G}$ of real affine transformations, $\mathcal{G} = \{[\alpha, \beta]; -\infty < \alpha < \infty, 0 < \beta < \infty\}$, and the quadratic loss function $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2/\theta^2$. We notice that all the parameter points (μ, σ^2, τ^2) such that $\tau^2/\sigma^2 = \rho$ belong to the same orbit. The values of ρ, $0 < \rho < \infty$, index the various possible orbits in the parameter space. The maximal invariant reduction of T^* is

$$[\bar{\bar{Y}}, Q_e]^{-1}(\bar{\bar{Y}}, Q_e, Q_a) = \left(0, 1, \frac{Q_a}{Q_e}\right).$$

Thus, every equivariant estimator of σ^2 is the form

$$\hat{\sigma}_\psi^2 = Q_e \psi\left(\frac{Q_a}{Q_e}\right) = \frac{Q_e}{1+nk}(1 + Uf(U)),$$

where $U = Q_a/Q_e$, $\psi(U)$ and $f(U)$ are chosen functions. Note that the distribution of U depends only on ρ. Indeed, $U \sim (1+n\rho)\chi_2^2[k-1]/\chi_1^2[k(n-1)]$. The risk function of an equivariant estimator $\hat{\sigma}_\psi^2$ is (Zacks, 1970)

$$R(f, \rho) = \frac{2}{1+nk} + \frac{nk-1}{nk+1} E_\rho\left(U^2 \frac{(1+n\rho)^2}{(1+U+n\rho)^2} \cdot \left[f(U) - \frac{1}{1+n}\right]^2\right).$$

If $K(\rho)$ is any prior distribution of the variance ratio ρ, the prior risk $E_K\{R(f, \rho)\}$ is minimized by choosing $f(U)$ to minimize the posterior expectation given U, i.e.,

$$E_{\rho|U}\left\{U^2 \frac{(1+n\rho)^2}{(1+U+n\rho)^2}\left[f(U) - \frac{1}{1+n}\right]^2\right\}.$$

The function $f_K(U)$ that minimizes this posterior expectation is

$$f_K(U) = \frac{E_{\rho|U}\{(1+n\rho)(1+U+n\rho)^{-2}\}}{E_{\rho|U}\{(1+n\rho)^2(1+U+n\rho)^{-2}\}}.$$

The Bayes equivariant estimator of σ^2 is obtained by substituting $f_K(u)$ in $\hat{\sigma}_\psi^2$. For more specific results, see Zacks (1970b). ■

Example 9.4. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a location and scale parameter exponential distribution, i.e.,

$$X_i \sim \mu + \sigma G_i(1, 1), \quad i = 1, \ldots, n$$

$-\infty < \mu < \infty, 0 < \sigma < \infty$.

A minimal sufficient statistic is $(X_{(1)}, S_n)$, where $X_{(1)} \leq \cdots \leq X_{(n)}$ and $S_n = \frac{1}{n-1}\sum_{i=2}^{n}(X_{(i)} - X_{(1)})$. We can derive the structural distribution on the basis of the minimal sufficient statistic. Recall that $X_{(1)}$ and S_n are independent and

$$X_{(1)} \sim \mu + \sigma G_{(1)} \sim \mu + \sigma G(n, 1),$$

$$S_n \sim \frac{\sigma}{n-1}\sum_{i=2}^{n}(G_{(i)} - G_{(1)}) \sim \sigma G(n-1, n-1),$$

where $G_{(1)} \leq \cdots \leq G_{(n)}$ is an order statistic from a standard exponential distribution $G(1, 1)$, corresponding to $\mu = 0, \sigma = 1$.

The group of transformation under consideration is $\mathcal{G} = \{[a, b];\ -\infty < a, -\infty < a < \infty,\ 0 < b < \infty\}$. The standard point is the vector $(G_{(1)}, S_G)$, where $G_{(1)} = (X_{(1)} - \mu)/\sigma$ and $S_G = S_n/\sigma$. The Jacobian of this transformation is $J(X_{(1)}, S_n, \mu, \sigma) = \frac{S_n}{\sigma^3}$. Moreover, the p.d.f. of $(G_{(1)}, S_G)$ is

$$\phi(u, s) = \frac{n(n-1)^{n-1}}{(n-2)!}s^{n-2}\exp\{-nu - (n-1)s\};\ 0 \leq u \leq \infty,\ 0 \leq s \leq \infty.$$

Hence, the structural distribution of (μ, σ) given $(X_{(1)}, S_n)$ has the p.d.f.

$$k(\mu, \sigma; X_{(1)}, S_n) = I\{\mu \leq X_{(1)}\}\frac{n(n-1)^{n-1}}{(n-2)!} \cdot \frac{S_n^{n-1}}{\sigma^{n+1}}\exp\left\{-n\frac{X_{(1)} - \mu}{\sigma} - (n-1)\frac{S_n}{\sigma}\right\},$$

for $-\infty < \mu \leq X_{(1)}, 0 < \sigma < \infty$.

The minimum risk structural estimators in the present example are obtained in the following manner. Let $L(\hat{\mu}, \mu, \sigma) = (\hat{\mu} - \mu)^2/\sigma^2$ be the loss function for estimating μ. Then the minimum risk estimator is the μ-expectation. This is given by

$$\begin{aligned} E\{\mu \mid X_{(1)}, S_n\} &= \frac{n(n-1)^{n-1}}{(n-2)!}S_n^{n-1} \cdot \\ &\quad \cdot \int_0^{\infty}\frac{1}{\sigma^{n+1}}\exp\left\{-(n-1)\frac{S_n}{\sigma}\right\}d\sigma \int_{-\infty}^{X_{(1)}}\mu\exp\left\{-n\frac{X_{(1)} - \mu}{\sigma}\right\}d\mu \\ &= X_{(1)} - \frac{S_n^*}{n-2}, \end{aligned}$$

where $S_n^* = \frac{1}{n}\sum_{i=2}^{n}(X_{(i)} - X_{(1)}) = \frac{n-1}{n}S_n$.

It is interesting to notice that while the MLE of μ is $X_{(1)}$, the minimum risk structural estimator might be considerably smaller, but close to the Pitman estimator.

The minimum risk structural estimator of σ, for the loss function $L(\hat{\sigma}, \sigma) = (\hat{\sigma} - \sigma)^2/\sigma^2$, is given by

$$\hat{\sigma} = \frac{E\left\{\frac{1}{\sigma} \mid \mathbf{X}\right\}}{E\left\{\frac{1}{\sigma^2} \mid \mathbf{X}\right\}} = \frac{1/S_n}{\frac{n}{n-1} \cdot \frac{1}{S_n^2}} = \frac{S_n(n-1)}{n} = S_n^*.$$

One can show that $\hat{\sigma}$ is also the minimum risk equivariant estimator of σ. ■

Example 9.5. A minimax estimator might be inadmissible. We show such a case in the present example. Let $X_1, \ldots, X_n$ be i.i.d. random variables having a normal distribution, like $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. The objective is to estimate σ^2 with the quadratic loss $L(\hat{\sigma}^2, \sigma^2) = (\hat{\sigma}^2 - \sigma^2)^2/\sigma^4$. The best equivariant estimator with respect to the group $\mathcal{G}$ of real affine transformation is $\hat{\sigma}^2 = Q/(n+1)$, where $Q = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. This estimator has a constant risk $R(\hat{\sigma}^2, \sigma^2) = \frac{2}{n+1}$. Thus, $\hat{\sigma}_n^2$ is minimax. However, $\hat{\sigma}^2$ is dominated uniformly by the estimator (9.3.31) and is thus inadmissible. ■

Example 9.6. In continuation of Example 9.5, given the minimal sufficient statistics $(\bar{X}_n, Q)$, the Bayes estimator of σ^2, with respect to the squared error loss, and the prior assumption that μ and σ^2 are independent, $\mu \sim N(0, \tau^2)$ and $1/2\sigma^2 \sim G(\lambda, \nu)$, is

$$\hat{\sigma}^2_{\tau,\lambda,\nu}(\bar{X}, Q) = \frac{\int_0^\infty \theta^{n-\frac{5}{2}+\nu}(1+2\theta n\tau^2)^{-1/2}\exp\{-\frac{\theta n\bar{X}^2}{1+n\theta\tau^2} - \theta(Q+\lambda)\}}{\int_0^\infty \theta^{n-\frac{3}{2}+\nu}(1+2\theta n\tau^2)^{-1/2}\exp\{-\frac{\theta n\bar{X}^2}{1+n\theta\tau^2} - \theta(\theta+\lambda)\}d\theta}.$$

This estimator is admissible since $h(\mu, \sigma^2) > 0$ for all (μ, σ^2) and the risk function is continuous in (μ, σ^2). ■

Example 9.7. Let $X_1, \ldots, X_n$ be i.i.d. random variables, having the location parameter exponential density $f(x; \mu) = e^{-(x-\mu)}I\{x \geq \mu\}$. The minimal sufficient statistic is $X_{(1)} = \min_{1 \leq i \leq n}\{X_i\}$. Moreover, $X_{(1)} \sim \mu + G(n, 1)$. Thus, the UMVU estimator of μ is $\hat{\mu} = X_{(1)} - \frac{1}{n}$. According to (9.3.15), this estimator is admissible. Indeed, by Basu's

Theorem, the invariant statistic $\mathbf{Y} = (X_{(2)} - X_{(1)}, X_{(3)} - X_{(2)}, \ldots, X_{(n)} - X_{(n-1)})$ is independent of $X_{(1)}$. Thus

$$(X_{(1)} - E\{X_{(1)} \mid \mathbf{Y}\})^2 = \left(X_{(1)} - \mu - \frac{1}{n}\right)^2,$$

and

$$\begin{aligned} E\{(E\{(X_{(1)} - E\{X_{(1)} \mid \mathbf{Y}\})^2 \mid \mathbf{Y}\}^{3/2}\} &= E\left\{\left(X_{(1)} - \mu - \frac{1}{n}\right)^3\right\} \\ &= 6 - \frac{6}{n} + \frac{3}{n^2} - \frac{1}{n^3} < \infty. \end{aligned}$$

■

Example 9.8. Let $\mathbf{Y} \sim N(\boldsymbol{\theta}, I)$, where $\boldsymbol{\theta} = H\boldsymbol{\beta}$,

$$H = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

and $\boldsymbol{\beta}' = (\beta_1, \beta_2, \beta_3, \beta_4)$. Note that $\frac{1}{2}H$ is an orthogonal matrix. The LSE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (H'H)^{-1}H'\mathbf{Y} = \frac{1}{4}H'\mathbf{Y}$, and the LSE of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = H(H'H)^{-1}H' = \mathbf{Y}$. The eigenvalues of $H(H'H)^{-1}H'$ are $\alpha_i = 1$, for $i = 1, \ldots, 4$. Thus, according to Theorem 9.3.3, $\hat{\boldsymbol{\theta}}$ is admissible. ■

Example 9.9. Let $\mathbf{X} \sim N(\boldsymbol{\theta}, I)$. X is a p-dimensional vector. We derive the generalized Bayes estimators of $\boldsymbol{\theta}$, for squared-error loss, and the Bayesian hyper-prior (9.3.43). This hyper-prior is

$$\boldsymbol{\theta} \mid \lambda \sim N\left(\mathbf{0}, \frac{1-\lambda}{\lambda} I\right), \quad 0 < \lambda \leq 1,$$

and

$$dG(\lambda) \propto \lambda^{-a} d\lambda, \quad -\infty < a < \frac{p}{\lambda} + 1.$$

The joint distribution of $(\mathbf{X}', \boldsymbol{\theta}')'$, given λ, is

$$\begin{pmatrix} \mathbf{X} \\ \boldsymbol{\theta} \end{pmatrix} \mid \lambda \sim N\left(\mathbf{0}, \begin{bmatrix} \frac{1}{\lambda} I & \frac{1-\lambda}{\lambda} I \\ \frac{1-\lambda}{\lambda} I & \frac{1-\lambda}{\lambda} I \end{bmatrix}\right).$$

Hence, the conditional distribution of $\boldsymbol{\theta}$, given $(\mathbf{X}, \lambda)$ is

$$\boldsymbol{\theta} \mid \mathbf{X}, \lambda \sim N((1-\lambda)\mathbf{X}, (1-\lambda)I).$$

The marginal distribution of $\mathbf{X}$, given λ, is $N\left(\mathbf{0}, \frac{1}{\lambda}I\right)$. Thus, the density of the posterior distribution of λ, given $\mathbf{X}$, is

$$h(\lambda \mid \mathbf{X}) \propto \lambda^{\frac{p}{2}-a} \exp\left\{-\frac{\lambda}{2}S\right\}, \quad 0 < \lambda \leq 1.$$

where $S = \mathbf{X}'\mathbf{X}$. It follows that the generalized Bayes estimator of $\boldsymbol{\theta}$, given $\mathbf{X}$ is

$$\hat{\boldsymbol{\theta}}_a(\mathbf{X}) = (1 - E\{\lambda \mid \mathbf{X}\})\mathbf{X},$$

where

$$\begin{aligned} E\{\lambda \mid \mathbf{X}\} &= \frac{\int_0^1 \lambda^{\frac{p}{2}-a+1} e^{-\frac{\lambda}{2}S} d\lambda}{\int_0^1 \lambda^{\frac{p}{2}-a} e^{-\frac{\lambda}{2}S} d\lambda} \\ &= \frac{p-2a+2}{S} \cdot \frac{P\{G(1, \frac{p}{2}-a+2) \leq \frac{S}{2}\}}{P\{G(1, \frac{p}{2}-a+1) \leq \frac{S}{2}\}}. \end{aligned}$$

■

PART III: PROBLEMS

Section 9.1

9.1.1 Consider a finite population of N units. M units have the value $x = 1$ and the rest have the value $x = 0$. A random sample of size n is drawn **without** replacement. Let X be the number of sample units having the value $x = 1$. The conditional distribution of X, given M, n is $H(N, M, n)$. Consider the problem of estimating the parameter $P = M/N$, with a squared-error loss. Show that the linear estimator $\hat{P}_{\alpha,\beta} = \alpha\frac{X}{n} + \beta$, with $\alpha = \frac{1}{1+\sqrt{\frac{N-n}{n(N-1)}}}$ and $\beta = \frac{1}{2}(1-\alpha)$ has a constant risk.

9.1.2 Let $\mathcal{H}$ be a family of prior distributions on Θ. The Bayes risk of $H \in \mathcal{H}$ is $\rho(H) = \rho(\hat{\theta}_H, H)$, where $\hat{\theta}_H$ is a Bayesian estimator of θ, with respect to

H. H^* is called least-favorable in $\mathcal{H}$ if $\rho(H^*) = \sup_{H\in\mathcal{H}} \rho(H)$. Prove that if H is a prior distribution in $\mathcal{H}$ such that $\rho(H) = \sup_{\theta\in\Theta} R(\theta, \hat{\theta}_H)$ then

(i) $\hat{\theta}_H$ is minimax for $\mathcal{H}$;

(ii) H is least favorable in $\mathcal{H}$.

9.1.3 Let X_1 and X_2 be independent random variables $X_1 \mid \theta_1 \sim B(n, \theta_1)$ and $X_2 \mid \theta_2 \sim B(n, \theta_2)$. We wish to estimate $\delta = \theta_2 - \theta_1$.

(i) Show that for the squared error loss, the risk function $R(\hat{\delta}, \theta_1, \theta_2)$ of

$$\hat{\delta}(X_1, X_2) = \frac{\sqrt{2n}}{n(\sqrt{2n}+1)}(X_2 - X_1)$$

attains its supremum for all points (θ_1, θ_2) such that $\theta_1 + \theta_2 = 1$.

(ii) Apply the result of Problem 2 to show that $\hat{\delta}(X_1, X_2)$ is minimax.

9.1.4 Prove that if $\hat{\theta}(X)$ is a minimax estimator over Θ_1, where $\Theta_1 \subset \Theta$ and $\sup_{\theta\in\Theta_1} R(\hat{\theta}, \theta) = \sup_{\theta\in\Theta} R(\hat{\theta}, \theta)$, then $\hat{\theta}$ is minimax over Θ.

9.1.5 Let $X_1, \ldots, X_n$ be a random sample (i.i.d.) from a distribution with mean θ and variance $\sigma^2 = 1$. It is known that $|\theta| \le M$, $0 < M < \infty$. Consider the linear estimator $\hat{\theta}_{a,b} = a\bar{X}_n + b$, where $\bar{X}_n$ is the sample mean, with $0 \le a \le 1$.

(i) Derive the risk function of $\hat{\theta}_{a,b}$.

(ii) Show that

$$\sup_{\theta:|\theta|\le M} R(\hat{\theta}_{a,b}, \theta) = \max\{R(\hat{\theta}_{a,b}, -M), R(\hat{\theta}_{ab}, M)\}.$$

(iii) Show that $\hat{\theta}_{a^*} = a^*\bar{X}_n$, with $a^* = M^2/\left(M^2 + \frac{1}{n}\right)$ is minimax.

Section 9.2

9.2.1 Consider Problem 4, Section 8.3. Determine the Bayes estimator for μ, $\delta = \mu - \eta$ and σ^2 with respect to the improper prior $h(\mu, \eta, \sigma^2)$ specified there and the invariant loss functions

$$L(\hat{\mu}, \mu) = (\hat{\mu}, \mu)^2/\sigma^2, \quad L(\hat{\delta}, \delta) = (\hat{\delta} - \delta^2)/\sigma^2$$
$$\text{and} \quad L(\hat{\sigma}^2, \sigma^2) = (\hat{\sigma}^2 - \sigma^2)^2/\sigma^4,$$

respectively, and show that the Bayes estimators are equivariant with respect $\mathcal{G} = \{[\alpha, \beta]; -\infty < \alpha < \infty, 0 < \beta < \infty\}$.

9.2.2 Consider Problem 4, Section 8.2. Determine the Bayes equivariant estimator of the variance ratio ρ with respect to the improper prior distribution specified in the problem, the group $\mathcal{G} = \{[\alpha, \beta]; -\infty < \alpha < \infty, 0 < \beta < \infty\}$ and the squared-error loss function for ρ.

9.2.3 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a common rectangular distribution $R(\theta, \theta + 1)$, $-\infty < \theta < \infty$. Determine the minimum MSE equivariant estimator of θ with a squared-error loss $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$ and the group $\mathcal{G}$ of translations.

9.2.4 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a scale parameter distribution, i.e.,

$$\mathcal{F} = \left\{ \frac{1}{\sigma} \phi\left(\frac{x}{\sigma}\right), 0 < \sigma < \infty \right\}.$$

(i) Show that the Pitman estimator of σ is the same as the Formal Bayes estimator (9.2.14).

(ii) Derive the Pitman estimator of $\hat{\sigma}$ when $\phi(x) = e^{-x} I\{x \geq 0\}$.

Section 9.3

9.3.1 Minimax estimators are not always admissible. However, prove that the minimax estimator of θ in the $B(n, \theta)$ case, with squared-error loss function, is admissible.

9.3.2 Let $X \sim B(n, \theta)$. Show that $\hat{\theta} = X/n$ is an admissible estimator of θ

(i) for the squared-error loss;

(ii) for the quadratic loss $(\hat{\theta} - \theta)^2/\theta(1 - \theta)$.

9.3.3 Let X be a discrete random variable having a uniform distribution on $\{0, 1, \ldots, \theta\}$, where the parameter space is $\Theta = \{0, 1, 2, \ldots\}$. For estimating θ, consider the loss function $L(\hat{\theta}, \theta) = \theta(\hat{\theta} - \theta)^2$.

(i) Derive the Bayesian estimator $\hat{\theta}_H$, where H is a discrete prior distribution on Θ.

(ii) What is the Bayesian estimator of θ, when $h(\theta) = e^{-\tau} \frac{\tau^\theta}{\theta!}, \theta = 0, 1, \ldots,$ $0 < \tau < \infty$?

(iii) Show that $\hat{\theta} = X$ is a Bayesian estimator only for the prior distribution concentrated on $\theta = 0$, i.e., $P_H\{\theta = 0\} = 1$.

(iv) Compare the risk function of $\hat{\theta}$ with the risk function of $\hat{\theta}_1 = \max(1, X)$.

(v) Show that an estimator $\hat{\theta}_m = m$ is Bayesian against the prior H_m s.t. $P_{H_m}\{\theta = m\} = 1, m = 0, 2, \ldots$.

9.3.4 Let X be a random variable (r.v.) with mean θ and variance $\sigma^2, 0 < \sigma^2 < \infty$. Show that $\hat{\theta}_{a,b} = aX + b$ is an **inadmissible** estimator of θ, for the squared-error loss function, whenever

(i) $a > 1$, or

(ii) $a < 0$, or

(iii) $a = 1$ and $b \neq 0$.

9.3.5 Let $X_1, \ldots, X_n$ be i.i.d. random variables having a normal distribution with mean zero and variance $\sigma^2 = 1/\phi$.

(i) Derive the Bayes estimator of σ^2, for the squared-error loss, and gamma prior, i.e., $\phi \sim G\left(\frac{\psi}{2}, \frac{n_0}{2}\right)$.

(ii) Show that the risk of the Bayes estimator is finite if $n + n_0 > 4$.

(iii) Use Karlin's Theorem (Theorem 9.3.1) to establish the admissibility of this Bayes estimator, when $n + n_0 > 4$.

9.3.6 Let $X \sim B(n, \theta), 0 < \theta < 1$. For which values of λ and γ the linear estimator

$$\hat{\theta}_{\lambda,\gamma} = \frac{X}{n(1+\lambda)} + \frac{\gamma\lambda}{1+\lambda}$$

is admissible?

9.3.7 Prove that if an estimator has a constant risk, and is admissible then it is minimax.

9.3.8 Prove that if an estimator is unique minimax then it is admissible.

9.3.9 Suppose that $\mathcal{F} = \{F(x;\theta), \theta \in \Theta\}$ is invariant under a group of transformations $\mathcal{G}$. If Θ has only one orbit with respect to $\bar{\mathcal{G}}$ (transitive) then the minimum risk equivariant estimator is minimax and admissible.

9.3.10 Show that any unique Bayesian estimator is admissible.

9.3.11 Let $\mathbf{X} \sim N(\boldsymbol{\theta}, I)$ where the dimension of $\mathbf{X}$ is $p > 2$. Show that the risk function of $\hat{\boldsymbol{\theta}}_c = \left(1 - c\frac{p-2}{||\mathbf{X}||^2}\right)\mathbf{X}$, for the squared-error loss, is

$$R(\hat{\boldsymbol{\theta}}_c, \boldsymbol{\theta}) = 1 - \frac{p-2}{p} E_{\boldsymbol{\theta}} \left\{ \frac{c(2-c)}{|X|^2} \right\}.$$

For which values of c does $\hat{\boldsymbol{\theta}}_c$ dominate $\hat{\boldsymbol{\theta}} = \mathbf{X}$? (i.e., the risk of $\hat{\boldsymbol{\theta}}_c$ is uniformly smaller than that of $\mathbf{X}$).

9.3.12 Show that the James–Stein estimator (9.3.16) is dominated by

$$\hat{\boldsymbol{\theta}}^{+} = \left(1 - \frac{p-2}{||X||^2}\right)^{+} \mathbf{X},$$

where $a^{+} = \max(a, 0)$.

PART IV: SOLUTIONS OF SELECTED PROBLEMS

9.1.1

(i) The variance of $\frac{X}{n}$ is $\frac{1}{n}P(1-P)\left(1 - \frac{n-1}{N-1}\right)$. Thus, the MSE of $\hat{P}_{\alpha,\beta}$ is

$$\begin{aligned}\mathrm{MSE}\{\hat{P}_{\alpha,\beta}\} &= \frac{\alpha^2}{n}P(1-P)\left(1 - \frac{n-1}{N-1}\right) + (\beta - (1-\alpha)P)^2 \\ &= \left((1-\alpha)^2 - \frac{\alpha}{n}\left(1 - \frac{n-1}{N-1}\right)\right)P^2 \\ &\quad - 2\left((1-\alpha)\beta - \frac{\alpha^2}{2n}\left(1 - \frac{n-1}{N-1}\right)\right)P + \beta^2.\end{aligned}$$

Now, for $\hat{\alpha} = \left(1 + \sqrt{\frac{1}{n}\left(1 - \frac{n-1}{N-1}\right)}\right)^{-1}$ and $\hat{\beta} = \frac{1}{2}(1-\hat{\alpha})$, we get

$$\mathrm{MSE}\{\hat{P}_{\hat{\alpha},\hat{\beta}}\} = \hat{\beta}^2 = \frac{1}{4n}\left(1 - \frac{n-1}{N-1}\right)\left(1 + \sqrt{\left(1 - \frac{n-1}{N-1}\right)\frac{1}{n}}\right)^{-2}.$$

This is a constant risk (independent of P).

9.1.3

(i)

$$\begin{aligned}R(\hat{\delta}, \theta_1, \theta_2) &= \frac{2(\theta_1(1-\theta_1) + \theta_2(1-\theta_2)) + (\theta_1 - \theta_2)^2}{(1+\sqrt{2n})^2} \\ &= \frac{(\theta_1 + \theta_2)(2 - (\theta_1 + \theta_2))}{(1+\sqrt{2n})^2}.\end{aligned}$$

Let $w = \theta_1 + \theta_2$, $g(w) = 2w - w^2$ attains its maximum at $w = 1$. Thus, $R(\hat{\delta}, \theta_1, \theta_2)$ attains its supremum on $\{(\theta_1, \theta_2) : \theta_1 + \theta_2 = 1\}$, which is $R^{*}(\delta) = \dfrac{1}{(1+\sqrt{2n})^2}$.

(ii) Let θ_1, θ_2 be priorly independent, having the same prior distribution beta(a, a). Then, the Bayes estimator for this prior and squared-error loss is

$$\begin{aligned}\delta_B(X_1, X_2) &= E\{\theta_1 - \theta_2 \mid X_1, X_2\} = E\{\theta_1 \mid X_1\}E\{\theta_2 \mid X_2\} \\ &= \frac{X_1 + a - (X_2 + a)}{n + 2a} = \frac{X_1 - X_2}{n(1 + \frac{2a}{n})}.\end{aligned}$$

Let $a = \frac{n}{2}\left(1 + \frac{1}{\sqrt{2n}}\right)$ then the Bayes estimator $\delta_B(X_1, X_2)$ is equal to $\hat{\delta}(X_1, X_2)$. Hence,

$$\rho^*(\delta_B) = \sup_{0 \le \theta_1, \theta_2 \le 1} R(\hat{\delta}, \theta_1, \theta_2) = \rho^*(\delta) = \frac{1}{(1 + \sqrt{2n})^2}.$$

Thus, $\hat{\delta}(X_1, X_2)$ is minimax.

9.1.5

(i) $\hat{\theta}_{a,b} = a\bar{X}_n + b$. Hence,

$$\begin{aligned}R(\hat{\theta}_{a,b}, \theta) &= \text{MSE}\{\hat{\theta}_{a,b}\} \\ &= (1 - a)^2\left(\theta - \frac{b}{1 - a}\right)^2 + \frac{a^2}{n}.\end{aligned}$$

(ii) Since $|\theta| \le M < \infty$,

$$\begin{aligned}\sup_{\theta:|\theta| \le M} R(\hat{\theta}_{a,b}, \theta) &= \max\{R(\hat{\theta}_{a,b}, M), R(\hat{\theta}_{a,b}, -M)\} \\ &= \frac{a^2}{n} + (1 - a)^2 \max\left\{\left(M - \frac{b}{1 - a}\right)^2, \right. \\ &\qquad \left. \left(M + \frac{b}{1 - a}\right)^2\right\} \\ &= \frac{a^2}{n} + (1 - a)^2\left(M + \frac{|b|}{1 - a}\right)^2.\end{aligned}$$

(iii) For $b = 0$,

$$\sup_{|\theta| \le M} R(\hat{\theta}_{a,0}, \theta) = \frac{a^2}{n} + (1 - a)^2 M^2.$$

The value of a that minimizes this supremum is $a^* = M^2/(M^2 + 1/n)$. Thus, if $a = a^*$ and $b = 0$

$$\sup_{|\theta| \le M} R(\hat{\theta}_{a^*,0}, \theta) = \inf_{(a,b)} \sup_{|\theta| \le M} R(\hat{\theta}_{a,b}, \theta).$$

Thus, $\theta^* = a^* X$ is minimax.

9.3.3

$\Theta = \{0, 1, 2, \ldots\}$, $L(\hat{\theta}, \theta) = \theta(\hat{\theta} - \theta)^2$.

(i) $f(x;\theta) = \dfrac{1}{1+\theta} I\{x \in (0, 1, \ldots, \theta)\}$. Let $h(\theta)$ be a discrete p.d.f. on Θ. The posterior p.d.f. of θ, given $X = x$, is

$$h(\theta \mid x) = \frac{\frac{h(\theta)}{1+\theta} I\{\theta \ge x\}}{\sum_{j=x}^{\infty} \frac{h(j)}{1+j}}.$$

The posterior risk is

$$E_H\{\theta(\hat{\theta} - \theta)^2 \mid X\} = E\{\hat{\theta}^2(X)\theta - 2\theta^2\hat{\theta}(X) + \theta^3 \mid X\}.$$

Thus, the Bayes estimator is the integer part of

$$\begin{aligned}\hat{\theta}_H &= \frac{E\{\theta^2 \mid x\}}{E\{\theta \mid x\}} \\ &= \frac{\sum_{j=x}^{\infty} \frac{j^2}{1+j} h(j)}{\sum_{j=x}^{\infty} \frac{j}{1+j} h(j)} I(x \ge 1).\end{aligned}$$

(ii) If $h(j) = e^{-\tau} \dfrac{\tau^j}{j!}$, $j = 0, 1, \ldots$, the Bayesian estimator is the integer part of

$$B_H(x) = \frac{\sum_{j=x}^{\infty} \frac{j^2}{1+j} \cdot \frac{\tau^j}{j!}}{\sum_{j=x}^{\infty} \frac{j}{1+j} \cdot \frac{\tau^j}{j!}} I(x \ge 1).$$

Let $p(i;\tau)$ and $P(i,\tau)$ denote, respectively, the p.d.f. and c.d.f. of the Poisson distribution with mean τ. We have,

$$e^{-\tau}\sum_{j=x}^{\infty}\frac{j}{1+j}\frac{\tau^j}{j!} = 1 - P(x-1;\tau) - \frac{1}{\tau}(1-P(x;\tau)).$$

Similarly,

$$e^{-\tau}\sum_{j=x}^{\infty}\frac{j^2}{1+j}\frac{\tau^j}{j!} = \tau(1-P(x-2;\tau)) - (1-P(x-1;\tau)) + \frac{1}{\tau}(1-P(x;\tau)).$$

Thus,

$$B_H(x) = \frac{\tau(1-P(x-2;\tau)) - (1-P(x-1;\tau)) + \frac{1}{\tau}(1-P(x;\tau))}{1-P(x-1;\tau) - \frac{1}{\tau}(1-P(x;\tau))}.$$

Note that $B_H(x) > x$ for all $x \geq 1$. Indeed,

$$\sum_{j=x}^{\infty}\frac{j^2}{1+j}\cdot\frac{\tau^j}{j!} > x\sum_{j=x}^{\infty}\frac{j}{1+j}\cdot\frac{\tau^j}{j!},$$

for all $x \geq 1$.

(iii) If $P\{\theta = 0\} = 1$ then obviously $P\{X = 0\} = 1$ and $B_H(X) = 0 = X$. On the other hand, suppose that for some $j > 0$, $P\{\theta = j\} > 0$, then $P\{X = j\} = \frac{1}{1+j}$ and, with probability $\frac{1}{1+j}$, $B_H(X) > X$. Thus, $B_H(X) = X = 0$ if, and only, $P\{\theta = 0\} = 1$.

(iv) Let $\hat{\theta} = X$. The risk is then

$$R(\hat{\theta},\theta) = E_\theta\{(X-\theta)^2\} = \frac{1}{1+\theta}\sum_{j=0}^{\theta-1}(\theta-j)^2 = \frac{\theta(2\theta+1)}{6}.$$

Note that $\hat{R}(\theta,0) = 0$. Consider the estimator $\hat{\theta}_1 = \max(1, X)$. In this case, $R(\hat{\theta}_1, 0) = 1$. We can show that $R(\hat{\theta}_1,\theta) < R(\hat{\theta},\theta)$ for all $\theta \geq 1$. However, since $R(\hat{\theta},0) < R(\hat{\theta}_1,0)$, $\hat{\theta}_1$ is not better than $\hat{\theta}$.

References

Abramowitz, M., and Stegun, I. A. (1968). *Handbook of Mathematical Functions With Formulas, Graphs and Mathematical Tables*. New York: Dover Publications.

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons Ltd.

Aitchison, J., and Dunsmore, I. R. (1975). *Statistical Prediction Analysis*. Cambridge: Cambridge University Press.

Akahira, M. and Takeuchi, K. (1981). *Asymptotic Efficiency of Statistical Estimators: Concepts and Higher Order Asymptotic Efficiency*. Lecture Notes in Statistics, **7**. New York: Springer.

Anderson, T. W. (1958). *Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons Ltd.

Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. New York: John Wiley & Sons Ltd.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton, New Jersey: Princeton University Press.

Arnold, J. C. (1970). Inadmissibility of the usual scale estimate for a shifted exponential distribution. *J. Am. Stat. Assoc.*, **65**, 1260–1264.

Baranchick, A. J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *Ann. Math. Stat.*, **41**, 642–645.

Baranchick, A. J. (1973). Inadmissibility of the MLE in some multiple regression problems with three or more independent variables. *Ann. Stat.*, **1**, 312–321.

Barlow, R. E., and Proschan, F. (1966). Tolerance and confidence limits for classes of distributions based on failure rate. *Ann. Math. Stat.*, **37**, 1593–1601.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory*. New York: John Wiley & Sons Ltd.

Barndorff-Nielsen, O. E., and Cox, D. R. (1979). Edgeworth and saddlepoint approximations with statistical applications (with discussion). *J. R. Stat. Soc. B*, **41**, 279–312.

Barndorff-Nielsen, O. E., and Cox, D. R. (1994). *Inference and Asymptotics*. Monograph on Statistics and Applied Probability, No. 52. London: Chapman and Hall.

Examples and Problems in Mathematical Statistics, First Edition. Shelemyahu Zacks.

Barnett, V. (1973). *Comparative Statistical Inference*. New York: John Wiley & Sons Ltd.

Basu, D. (1975). Statistical information and likelihood. *Sankhya, A*, **37**, 1–71.

Berger, J. O., and Bock, M. E. (1976). Eliminating singularities of Stein-type estimators of location vectors. *J. Roy. Stat. Soc.*, B, **38**, 166–170.

Berk, R. H. (1967). A special group structure and equivariant estimation. *Ann. Math. Stat.*, **38**, 1436–1445.

Berk, R. H. (1973). Some asymptotic aspects of sequential analysis. *Ann. Stat.*, **1**, 1126–1138.

Berk, R. H. (1975a). Locally most powerful sequential tests. *Ann. Stat.*, **3**, 373–381.

Berk, R. H. (1975b). Comparing sequential and nonsequential tests. *Ann. Stat.*, **3**, 991–998.

Berk, R. H. (1976). Asymptotic efficiencies of sequential tests. *Ann. Stat.*, **4**, 891–900.

Bhapkar, V. P. (1972). On a measure of efficiency of an estimating equation. *Sankhya, A*, **34**, 467–472.

Bhattacharyya, A. (1946). On some analogues of the amount of information and their uses in statistical estimation. *Sankhya*, **8**, 1–14, 201–218, 315–328.

Bickel, P. J., and Doksum, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. San Francisco: Holden-Day.

Billah, M. B., and Saleh, A. K. M. E. (1998). Conflict between pretest estimators induced by three large sample tests under a regression model with student t-error. *The Statistician*, **47**, 1–14.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *Ann. Math. Stat.*, **18**, 105–110.

Blackwell, D., and Girshick, M. A. (1954). *Theory of Games and Statistical Decisions*. New York: John Wiley & Sons Ltd.

Blyth, C. R. (1951). On minimax statistical decision procedures and their admissibility. *Ann. Math. Stat.*, **22**, 22–42.

Blyth, C. R., and Roberts, D. M. (1972). On inequalities of Cramér-Rao type and admissibility proofs. *Proc. Sixth Berkeley Symp. Math. Stat. Prob.*, **I**, 17–30.

Borges, R., and Pfanzagl, J. (1965). One-parameter exponential families generated by transformation groups. *Ann. Math. Stat.*, **36**, 261–271.

Box, G. E. P., and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* , Massachusetts: Addison-Wesley.

Bratley, P., Fox, B. L., and Schrage, L. E. (1983). *A Guide to Simulation*. New York: Springer-Verlag.

Brewster, J. F., and Zidek, J. V. (1974). Improving on equivariant estimators. *Ann. Stat.*, **2**, 21–38.

Brier, S. S., Zacks, S., and Marlow, W. H. (1986). An application of empirical Bayes techniques to the simultaneous estimation of many probabilities. *Nav. Res. Logist. Q.*, **33**, 77–90.

Brown, L. D. (1964). Sufficient statistics in the case of independent random variables. *Ann. Math. Stat.*, **35**, 1456–1474.

Brown, L. D. (1968). Inadmissibility of the usual estimators of scale parameters in problems with unknown location and scale parameters. *Ann. Math. Stat.*, **39**, 29–48.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families With Applications in Statistical Decision Theory*. IMS Lecture Notes-Monograph Series, vol. 9. California: Haywood.

Brown, L. D., and Cohen, A. (1974). Point and confidence estimation of a common mean and recovery of interblock information. *Ann. Stat.*, **2**, 963–976.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *Am. Stat.*, **39**, 83–87.

Chapman, D. G., and Robbins, H. (1951). Minimum variance estimation without regularity assumptions. *Ann. Math. Stat.*, **22**, 581–586.

Chernoff, H. (1959). Sequential design of experiments. *Ann. Math. Stat.*, **30**, 755–770.

Chernoff, H. (1961). Sequential tests for the mean of a normal distribution. *Proc. Fourth Berkeley Symp. Math. Stat. Prob.*, **A**, 79–91.

Chernoff, H. (1965). Sequential tests for the mean of a normal distribution, III (small T). *Ann. Math. Stat.*, **36**, 28–54.

Chernoff, H. (1968). Optimal stochastic control. *Sankhya, A*, **30**, 221–252.

Chernoff, H., and Scheffé, H. (1952). A generalization of the Neyman-Pearson fundamental lemma. *Ann. Math. Stat.*, **23**, 213–225.

Chow, Y. S., and Robbins, H. (1965). On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *Ann. Math. Stat.*, **36**, 457–462.

Chow, Y. S., Robbins, H., and Siegmund, D. (1971). *Great Expectations: The Theory of Optimal Stopping*. Boston: Houghton Mifflin.

Cohen, A. (1966). All admissible linear estimates of the mean vector. *Ann. Math. Stat.*, **37**, 458–463.

Cohen, A., and Sackrowitz, H. B. (1974). On estimating the common mean of two normal distributions. *Ann. Stat.*, **2**, 1274–1282.

Cornfield, J. (1969). The Bayesian outlook and its applications. *Biometrics*, **25**, 617–657.

Cox, D. R., and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Cramér, H. (1946). A contribution to the theory of statistical estimation. *Skand Aktuar.*, **29**, 85–94.

Dantzig, G. B., and Wald, A. (1951). On the fundamental lemma of Neyman and Pearson. *Ann. Math. Stat.*, **22**, 87–93.

David, F. N., and Kendall, M. G. (1955). Tables of Symmetric Functions. *Biometrika*, **42**, 223.

David, H. A. (1970). *Order Statistics*. New York: John Wiley & Sons Ltd.

Davis, P. J., and Rabinowitz, P. (1984). *Methods of Numerical Integration*, 2nd edn. New York: Academic Press.

DeFinneti, B. (1974). *Theory of Probability*, vol. 1. New York: John Wiley & Sons Ltd.

DeGroot, M. H., and Raghavachari, M. (1970). Relations between Pitman efficiency and Fisher information. *Sankhya*, **32**, 319–324.

Denny, J. L. (1967). Sufficient conditions for a family of probabilities to be exponential. *Proc. Natl. Acad. Sci.*, **57**, 1184–1187.

Denny, J. L. (1969). Note on a theorem of Dynkin on the dimension of sufficient statistics. *Ann. Math. Statist.*, **40**, 1474–1476.

Draper, N., and Smith, H. (1966). *Applied Regression Analysis*. New York: John Wiley & Sons Ltd.

Dynkin, E. B. (1951). Necessary and sufficient statistics for a family of probability distributions. *Selected Translations in Math. Stat. Prob.*, **1**, 17–40.

Dynkin, E. B., and Yushkevich, A. A. (1969). *Markov Processes: Theorems and Problems*. New York: Plenum Press.

Eaton, M. L. (1989). *Group Invariance Applications in Statistics*. Regional Conferences Series in Probability and Statistics, vol. 1, IMS. California: Haywood.

Efron, B. (1975). Defining the curvature of a statistical problem (with application to second order efficiency). *Ann. Stat.*, **3**, 1189–1242 (with discussion).

Efron, B. (1978). The geometry of exponential families. *Ann. Stat.*, **6**, 362–376.

Efron, B., and Morris, C. (1971). Limiting the risk of Bayes and empirical estimators-Part I: the Bayes case. *J. Am. Stat. Assoc.*, **66**, 807–815.

Efron, B., and Morris C. (1973a). Combining possibly related estimation problems. *J. R. Stat. Soc. B*, **35**, 379–421.

Efron, B., and Morris, C. (1972a). Limiting the risk of Bayes and empirical Bayes estimators. *J. Am. Stat. Assoc.*, **67**, 103–109.

Efron, B., and Morris, C. (1972b). Empirical Bayes on vector observations: an extension of Stein's method. *Biometrika*, **59**, 335–347.

Efron, B., and Morris, C. (1973b). Stein's estimation rule and its competitors: an empirical Bayes approach. *J. Am. Stat. Assoc.*, **68**, 117–130.

Ellison, B. E. (1964). Two theorems of inference about the normal distribution with applications in acceptance sampling. *J. Am. Stat. Assoc.*, **59**, 89–95.

Evans, M., and Swartz, T. (2001). *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford: Oxford University Press.

Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*, vol. II. New York: John Wiley & Sons Ltd.

Fend, A. V. (1959). On the attainment of Cramér-Rao and Bhattacharya bounds for the variances of an estimate. *Ann. Math. Stat.*, **30**, 381–388.

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. New York: Academic Press.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. New York: Chapman and Hall.

Field, C., and Ronchetti, E. (1990). *Small Sample Asymptotics*. IMS Lecture Notes-Monograph Series, vol. 13. California: Haywood.

Fienberg, S. E. (1980). *The Analysis of Crossed-Classified Categorical Data*, 2nd edn. Boston, MA: MIT Press.

Finney, D. J. (1964). *Statistical Methods in Biological Assays*, 2nd edn. London: Griffin.

Fisher, R. A. (1922). On the mathematical foundation of theoretical statistics. *Phil. Trans. R. Soc. A*, **222**, 309–368.

Fisz, M. (1963). *Probability Theory and Mathematical Statistics*, 3rd edn. New York: John Wiley & Sons Ltd.

Fleiss, J. L. (1973). *Statistical Methods for Rates and Proportions*. New York: John Wiley & Sons Ltd.

Fraser, D. A. S. (1957). *Nonparametric Methods in Statistics*. New York: John Wiley & Sons Ltd.

Fraser, D. A. S. (1968). *The Structure of Inference*. New York: John Wiley & Sons Ltd.

Fréchet, M. (1943). Sur l'extension de certaines evaluations statistiques de petits enchantilons. *Rev. Int. Statisist.*, **11**, 182–205.

Galambos, J. (1978). *The Asymptotic Theory of Extreme Order Statistics*. New York: John Wiley & Sons Ltd.

Gamerman, D. (1997). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Text in Statistical Science. New York: Chapman and Hall.

Gastwirth, J. L. (1977). On robust procedures. *J. Am. Stat. Assoc.*, **61**, 929–948.

Geisser, S. (1993). *Predictive Inference: An Introduction*. Monographs on Statistics and Applied Probability, No. 55. New York: Chapman and Hall.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. New York: Chapman and Hall.

Ghosh, B. K. (1970). *Sequential Tests of Statistical Hypotheses*. Reading, MA: Addison-Wesley.

Ghosh, M. (1992). Hierarchical and Empirical Bayes Multivariate Estimation. *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, *IMS Lecture Notes—Monograph Series, vol. 17*.

Ghosh, M., Mukhopadhyay, N., and Sen, P. K. (1997). *Sequential Estimation*. New York: John Wiley & Sons Ltd.

Girshick, M. A., and Savage, L. J. (1951). Bayes and minimax estimates for quadratic loss functions. *Proc. Second Berkeley Symp. Math. Stat. Prob.*, **1**, 53–74.

Godambe, V. P. (1991). *Estimating Functions*. Oxford: Clarendon Press.

Gokhale, D. V., and Kullback, S. (1978). *The Information in Contingency Tables*. Textbooks and monographs, vol. 23. New York: Marcel Dekker.

Good, I. J. (1965). *The Estimation of Probabilities: An Assay on Modern Bayesian Methods*. Cambridge, MA: MIT Press.

Good, I. J. (1967). A Bayesian significance test for multinomial distributions. *J. R. Stat. Soc. B*, **28**, 399–431

Good, I. J. (l975). The Bayes factor against equiprobability of a multinomial population assuming a symmetric dirichlet prior. *Ann. Stat.*, **3**, 246–250.

Good, I. J., and Crook, J. F. (1974). The Bayes/non-Bayes compromise and multinomial distribution. *J. Am. Stat. Assoc.*, **69**, 711–720.

Graybill, F. (1961). *An Introduction to Linear Statistical Models*, Vol. I. New York: McGraw-Hill.

Graybill, F. A. (1976). *Theory and Application of the Linear Model*. Massachusetts: Duxbury Press.

Gross, A. J., and Clark, V. A. (1975). *Survival Distributions: Reliability Applications in the Biomedical Sciences*. New York: John Wiley & Sons Ltd.

Guenther, W. C. (1971). Tolerance intervals for univariate distributions. *Naval Res. Log. Quart.*, **19**, 309–333.

Gumbel, E. J. (1958). *Statistics of Extreme*. New York: Columbia University Press.

Guttman, I. (1970). Construction of beta content tolerance regions at confidence level gamma for large samples from the k-variate normal distribution. *Ann. Math. Stat.*, **41**, 376–400.

Haberman, S. J. (1974). *The Analysis of Frequency Data*. Chicago: The University of Chicago Press.

Hacking, I. (1965). *Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hald, A. (1952). *Statistical Theory With Engineering Applications*. New York: John Wiley & Sons Ltd.

Hall, W. J., Wijsman, R. A., and Ghosh, B. K. (1965). The relationship between sufficiency and invariance with applications in sequential analysis. *Ann. Math. Stat.*, **36**, 575–614.

Harrison, P. J., and Stevens, C. F. (1976). Bayesian forecasting. *J. R. Stat. Soc*. B, **38**, 205–247.

Hettmansperger, T. P. (1984). *Statistical Inference Based on Ranks*. New York: John Wiley & Sons Ltd.

Hoerl, A. E. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress*, **58**, 54–59.

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hoerl, A. E., Kennard, R. W., and Baldwin, K. F. (1975). Ridge regression: some simulations. *Comm. Stat.*, **4**, 105–123.

Holland, P. W. (1973). Covariance stabilizing transformations. *Ann. Stat.*, **1**, 84–92.

Huber, P. J. (1964). Robust estimation of the location parameter. *Ann. Math. Stat.*, **35**, 73–101.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Stat. Prob.*, **I**, 221–233.

James, W., and Stein, C. (1960). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Stat. Prob.*, **2**, 361–379 .

Jeffreys, H. (1961). *Theory of Probability*, 3rd edn. Oxford: Clarendon Press.

Jogdio, K., and Bohrer, R. (1973). Some simple examples and counter examples about the existence of optimal tests. *J. Am. Stat. Assoc.*, **68**, 679–682.

Johnson, N. L., and Kotz, S. (1969). *Distributions in Statistics*, Vol. I. *Discrete Distributions*, vol. II. *Continuous Univariate Distributions-1*, vol. III. *Continuous Univariate Distributions-2*. Boston: Houghton and Mifflin.

Johnson, N. L., and Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. New York: John Wiley & Sons Ltd.

Joshi, V. M. (1976). On the attainment of the Cramér-Rao lower bound. *Ann. Stat.*, **4**, 998–1002.

Judge, G. G., and Bock, M. E. (1978). *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. Amsterdam: North-Holland Publishing Co.

Karlin, S. (1956). Decision theory for polya type distributions. Case of two actions, I. *Third Berkeley Symp. Math. Stat. Prob.*, **1**, 115–128.

Karlin, S. (1958). Admissibility for estimation with quadratic loss. *Ann. Math. Stat.*, **29**, 406–436.

Karlin, S. and Rubin, H. (1956). The theory of decision procedures for distributions with monotone likelihood ratio. *Ann. Math. Stat.*, **27**, 272–300.

Khan, R. A. (1969). A general method for determining fixed-width confidence intervals. *Ann. Math. Stat.*, **40**, 704–709.

Kiefer, J. (1952). On minimum variance estimates. *Ann. Math. Stat.*, **23**, 627–629.

Kiefer, J., and Weiss, L. (1957). Some properties of generalized sequential probability ratio tests. *Ann. Math. Stat.*, **28**, 57–74.

Klotz, J. H., Milton, R. C., and Zacks, S. (1969). Mean square efficiency of estimators of variance components. *J. Am. Stat. Assoc.*, **64**, 1342–1349.

Kubokawa, T. (1987). *Estimation of The Common Means of Normal Distributions With Application to Regression and Design of Experiments*. Ph.D. Dissertation, University of Tsukuba.

Kullback, S. (1959). *Information Theory and Statistics*. New York: John Wiley & Sons Ltd.

Lai, T. L. (1973). Optimal stopping and sequential tests which minimize the maximum expected sample size. *Ann. Stat.*, **1**, 659–673.

Lancaster, H. O. (1969). *The chi-squared distributions*. New York: John Wiley & Sons Ltd.

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. Calif. Publ. Statist.*, **1**, 277–330.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. New York: Springer.

Lehmann, E. L. (1997). *Testing Statistical Hypotheses*, 2nd edn. Springer Texts in Statistics. New York: Springer.

Lehmann, E. L., and Casella, G. (1998). *Theory of Point Estimation*, 2nd edn. Springer Texts in Statistics. New York: Springer.

Lehmann, E. L., and Scheffé, H. (1950). Completeness, similar regions and unbiased estimation, I. *Sankhya*, **10**, 305–340.

Lehmann, E. L., and Scheffé, H. (1955). Completeness, similar regions and unbiased estimation, II. *Sankhya*, **15**, 219–236.

Lin, P. E. (1974). Admissible minimax estimators of the multivariate normal mean with squared error loss. *Commun. Stat.*, **3**, 95–100.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Stat.*, **27**, 986–1005.

Lindley, D. V., and Smith, A. F. M. (1972). Bayes estimates for the linear model. *J. R. Stat. Soc. B*, **34**, 1–42.

Lindsey, J. K. (1996). *Parametric Statistical Inference*. Oxford: Oxford Science Publications, Clarendon Press.

Lugannani, R., and Rice, S. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. *Adv. Appl. Probab.*, **12**, 475–490.

Maritz, J. (1970). *Empirical Bayes Methods*. London: Methuen.

Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics*, **12**, 55–67.

McLachlan, G. J., and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: John Wiley & Sons Ltd.

Miller, R. G. (1966). *Simultaneous Statistical Inference*. New York: McGraw-Hill.

Morris, C. (1983). Parametric Empirical Bayes Inference and Applications. *J. Am. Stat. Assoc.*, **78**, 47–65.

Mukhopadhyay, N., and de Silva, B. M. (2009). *Sequential Methods and Their Applicatio7ns*. Boca Raton, FL: CR Press.

Neyman, J. (1935). Sur un teorems concerente le cosidette statistiche sufficienti. *Inst. Ital. Atti. Giorn.*, **6**, 320–334.

Neyman, J., and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. R. Soc. A*, **231**, 289–337.

Neyman, J. and Pearson, E. S. (1936a). Contributions to the theory of testing statistical hypotheses, I. Unbiased critical regions of type A and type A(1). *Stat. Res. Mem.*, **1**, 1–37.

Neyman, J., and Pearson, E. S. (1936b). Sufficient statistics and uniformly most powerful tests of statistical hypotheses. *Stat. Res. Memo.*, **1**, 113–137.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrika*, **16**, 1–32.

Pfanzagl, J. (1972). Transformation groups and sufficient statistics. *Ann. Math. Stat.*, **43**, 553–568.

Pfanzagl, J. (1985). *Asymptotic Expansions for General Statistical Models*. Springer Lecture Notes in Statistics, Vol. 31. New York: Springer.

Pitman, E. J. G. (1948). *Notes on Nonparametric Statistical Inference*. Chapel Hill, NC: Institute of Statistics, University of North Carolina.

Pitman, E. J. G. (1979). *Some Basic Theory for Statistical Inference*. New York: Chapman and Hall.

Portnoy, S. (1971). Formal Bayes estimation with application to a random effect model. *Ann. Math. Stat.*, **42**, 1379–1402.

Raiffa, H., and Schlaifer, R. (1961). *Introduction to Statistical Decision Theory*. Cambridge: Harvard University Press.

Rao, C. R. (1945). Information and accuracy attainable in estimation of statistical parameters. *Bull. Cal. Math. Soc.*, **37**, 81–91.

Rao, C. R. (1947). Minimum variance and estimation of several parameters. *Proc. Camb. Phil. Soc.*, **43**, 280–283.

Rao, C. R. (1949). Sufficient statistics and minimum variance estimates. *Proc. Camb. Phil. Soc.*, **45**, 218–231.

Rao, C. R. (1963). Criteria of estimation in large samples. *Sankhya, A*, **25**, 189–206.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd edn. New York: John Wiley & Sons Ltd.

Reid, N. (1988). Saddlepoint Methods and Statistical Inference. *Statistical Science*, **3**, 213–238.

Reid, N. (1995). Likelihood and Bayesian Approximation Methods. In: Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors. *Bayesian Statistics*, Vol. 5. Oxford: Oxford University Press.

Robbins, H. (1956). The empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Stat. Prob.*, **1**, 157–164.

Robbins, H. (1964). The empirical approach to statistical decision problems. *Ann. Math. Stat.*, **35**, 1–20.

Rogatko, A., and Zacks, S. (1993). Ordering Genes: Controlling the Decision Error Probabilities. *Am. J. Hum. Genet.*, **52**, 947–957.

Rohatgi, V. K. (1976). *An Introduction to Probability Theory and Mathematical Statistics*. New York: John Wiley & Sons Ltd.

Sarhan, A. E., and Greenberg, B. G. (1962). *Contributions to Order Statistics*. New York: John Wiley & Sons Ltd.

Savage, L. J. (1962). *The Foundations of Statistical Inference*. London: Methuen.

Scheffé, H. (1970). Multiple testing versus multiple estimation. Improper confidence sets: estimation of directions and ratios. *Ann. Math. Stat.*, **41**, 1–29.

Scheffé, H. (1959). *The Analysis of Variance*. New York: John Wiley & Sons Ltd.

Schervish, M. J. (1995). *Theory of Statistics*. New York: Springer.

Schmetterer, L. (1974). *Introduction to Mathematical Statistics* (Revised English Edition). New York: Springer.

Searle, S. R. (1971). *Linear Models*. New York: John Wiley & Sons Ltd.

Seber, G. A. F. (1977). *Linear Regression Analysis*. New York: John Wiley & Sons Ltd.

Sen, A., and Srivastava, M. S. (1990). *Regression Analysis: Theory, Methods and Applications*. Springer Texts in Statistics. New York: Springer.

Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics: An Introduction With Applications*. New York: Chapman and Hall.

Shiryayev, A. N. (1973). *Statistical Sequential Analysis: Optimal Stopping Rules*. Translations of Math. Monographs, Vol. 38. American Math. Society, Providence, Rhode Island.

Shiryayev, A. N. (1984). *Probability*, Graduate Texts in Mathematics, No. 95. New York: Springer.

Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. New York: Springer.

Simons, G. (1968). On the cost of not knowing the variance when making a fixed-width confidence interval for the mean. *Ann. Math. Stat.*, **39**, 1946–1952.

Skovgaard, Ib M. (1990). *Analytic Statistical Models*. IMS Lecture Notes-Monograph Series, Vol. 15. California: Haywood.

Smith, A. F. M. (1973a). A general Bayesian linear model. *J. R. Stat. Soc. B*, **35**, 67–75.

Smith, A. F. M. (1973b). Bayes estimates in one-way and two-way models. *Biometrika*, **60**, 319–329.

Srivastava, M. S. (1971). On fixed-width confidence bounds for regression parameters. *Ann. Math. Stat.*, **42**, 1403–1411.

Starr, N. (1966). The performance of a sequential procedure for the fixed-width interval estimation of the mean. *Ann. Math. Stat.*, **37**, 36–50.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Stat.*, **16**, 243–258.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Stat. Prob.*, **1**, 197–206.

Stein, C. (1959). The admissibility of the Pitman's estimator for a single location parameter. *Ann. Math. Stat.*, **30**, 970–999.

Stein, C. (1962). A remark on the likelihood principle. *J. R. Stat. Soc. A*, 565–568.

Stein, C. (1964). Inadmissibility of the usual estimate of the variance of a normal distribution with unknown mean. *Ann. Inst. Stat. Math.*, **16**, 155–160.

Stein, C. (1986). *Approximate Computation of Expectations*. Lecture Notes-Monograph Series, Vol. 7. Hayward, CA: Institute of Mathematical Statistics.

Stone, J., and Conniffe, D. (1973). A critical view of ridge regression. *The Statistician*, **22**, 181–187.

Strawderman, W. E. (1972). On the existence of proper Bayes minimax estimators of the mean of a multivariate normal distribution. *Proc. Sixth Berkeley Symp. Math. Stat. Prob.*, **1**, 51–56.

Susarla, V. (1982). *Empirical Bayes Theory:Encyclopedia of Statistical Sciences*, vol. 2. New York: John Wiley & Sons Ltd, pp. 490–502.

Sverdrup, E. (1953). Similarity, minimaxity and admissibility of statistical test procedures. *Skand. Aktuar. Tidskrift*, **36**, 64–86.

Tan, P. (1969). A note on a theorem of Dynkin on necessary and sufficient statistics. *Canadian Math. Bulletin*, **12**, 347–351.

Tiao, G. C., and Tan, W. Y. (1965). Bayesian analysis of random-effect models in the analysis of variance. I. Posterior distribution of variance components. *Biometrika*, **51**, 219–230.

Tierney, L. J., and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities, *J. Am. Stat. Assoc.*, **81**, 82–86.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Ann. Math. Stat.*, **16**, 117–186.

Wald, A. (1947). *Sequential Analysis*. New York: John Wiley & Sons Ltd.

Watson, G. S. (1967). Linear least squares regression. *Ann. Math. Stat.*, **38**, 1679–1689.

West, M., and Harrison, P. J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd edn. New York: Springer.

West, M., Harrison, P. J., and Migon, H. S. (1985). Dynamic Generalized linear models and Bayesian forecasting (with discussion). *J. Am. Stat. Assoc.*, **80**, 73–83.

Wijsman, R. A. (1971). Exponentially bounded stopping time of SPRTs for composite hypotheses. *Ann. Math. Stat.*, **42**, 1859–1869.

Wijsman, R. A. (1970). Examples of exponentially bounded stopping time of invariant sequential probability ratio tests when the model may be false. *Proc. Sixth Berkeley Symp. Math. Stat. Prob.*, **1**, 109–128.

Wijsman, R. A. (1973). On the attainment of the Cramér-Rao lower bound. *Ann. Stat.*, **1**, 538–542.

Wijsman, R. A. (1990). *Invariant Measures on Groups and Their Use in Statistics*, Lecture Notes-Monograph Series, Vol. 14. Hayward, CA: Institute of Mathematical Statistics.

Wilks, S. S. (1962). *Mathematical Statistics*. New York: John Wiley & Sons Ltd.

Zacks, S. (1966). Unbiased estimation of the common mean of two normal distributions based on small samples. *J. Am. Stat. Assoc.*, **61**, 467–476.

Zacks, S. (1970a). Bayes and fiducial equivariant estimators of the common mean of two normal distributions. *Ann. Math. Stat.*, **41**, 59–69.

Zacks, S. (1970b). Bayes equivariant estimators of variance components. *Ann. Inst. Stat. Math.*, **22**, 27–40.

Zacks, S. (1971). *The Theory of Statistical Inference*. New York: John Wiley & Sons Ltd.

Zacks, S. (1992). *Introduction to Reliability Analysis: Probability Models and Statistical Methods*. New York: Springer.

Zacks, S. (1997). Adaptive Designs for Parametric Models. Chapter 5 in *Handbook of Statistics, Vol. 13: Design and Analysis of Experiments*. New York: North-Holland.

Zacks, S., and Solomon, H. (1976). On testing and estimating the interaction between treatments and environmental conditions in binomial experiments: The case of two stations. *Commun. Stat.*, **A5**, 197–223.

Zehna, P. W. (1966). Invariance of maximum likelihood estimation. *Ann. Math. Stat.*, **37**, 755.

Zelen, M. (1972). Exact significance tests for contingency tables embedded in a $2 * *N$ classification. *Sixth Berkeley Symp. Prob. Stat.*, **I**, 737–757.

Zelen, M., and Severo, N. C. (1968). Probability functions; Chapter 26 in *Abramowitz, M. and Stegun, I. A*. (1968).

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley & Sons Ltd.

Zyskind, G. (1967). On canonical forms, non-negative covariance matrices and best and simple least-squares linear estimators in linear models. *Ann. Math. Stat.*, **38**, 1092–1109.

Author Index

Abramowitz, M., 113, 382, 462, 470, 483, 510, 601, 611
Agresti, A., 275, 601
Aitchison, J., 410, 411, 601
Akahira, M., 444, 601
Anderson, T.W., 271, 332, 386, 601
Andrews, D.F., 352, 601
Arnold, J.C., 601

Baldwin, K.F., 337, 606
Baranchick, A.J., 582, 601
Barlow, R.E., 414, 601
Barndorff-Nielsen, O.E., 150, 206, 224, 226, 444, 601
Barnett, V., 486, 602
Basu, D., 206, 239, 602
Berger, J.O., 582, 602
Berk, R.H., 283, 342, 602
Bhapkar, V.P., 330, 602
Bhattacharyya, A., 325, 602
Bickel, P.J., 193, 305, 306, 601, 602
Billah, M.B., 349, 602
Bishop, Y.M.M., 275, 602
Blackwell, D., 322, 493, 602
Blyth, C.R., 571, 602
Bock, M.E., 349, 582, 585, 602, 606
Bohrer, R., 256, 606
Borges, R., 203, 602
Box, G.E.P., 501, 503, 537, 602
Bratley, P., 510, 602
Brewster, J.F., 582, 602
Brier, S.S., 545, 549, 602
Brown, L.D., 145, 202, 203, 373, 581, 602, 603

Carlin, J.B., 555, 605
Casella, G., 514, 574, 584, 603, 607
Chapman, D.G., 327, 603
Chernoff, H., 251, 497, 553, 603
Chow, Y.S., 283, 421, 603
Clark, V.A., 112, 341, 605
Cohen, A., 373, 575, 579, 603
Conniffe, D., 335, 609
Cornfield, J., 494, 603
Cox, D.R., 150, 224, 226, 248, 444, 601, 603
Cramér, H., 323, 603
Crook, J.F., 494, 605

Dantzig, G.B., 251, 603
David, F.N., 603
David, H.A., 120, 135, 178, 603
Davis, P.J., 603
de Silva, B.M., 421, 607
DeFinneti, B., 486, 603
DeGroot, M.H., 330, 603
Denny, J.L., 203, 603
Doksum, K.A., 193, 305, 306, 602
Draper, N., 332, 388, 603
Dunsmore, I.R., 410, 411, 601
Dynkin, E.B., 202, 203, 497, 604

Examples and Problems in Mathematical Statistics, First Edition. Shelemyahu Zacks.

Eaton, M.L., 342, 604
Efron, B., 146, 444, 514, 584, 604
Ellison, B.E., 604
Evans, M., 508, 604

Feller, W., 108, 604
Fend, A.V., 326, 604
Ferguson, T.S., 38, 441, 493, 604
Field, C., 150, 604
Fienberg, S.E., 275, 602, 604
Finney, D.J., 341, 604
Fisher, R.A., 193, 604
Fisz, M., 350, 604
Fleiss, J.L., 296, 468, 604
Fox, B.L., 510, 602
Fraser, D.A.S., 204, 342, 568, 604
Fréchet, M., 323, 605

Galambos, J., 605
Gamerman, D., 508, 605
Gastwirth, J.L., 351, 605
Geisser, S., 501, 605
Gelman, A., 503, 555, 605
Ghosh, B.K., 276, 279, 282, 342, 421, 497, 605, 606
Ghosh, M., 584, 585, 605
Girshick, M.A., 493, 570, 572, 602, 605
Godambe, V.P., 605
Gokhale, D.V., 275, 605
Good, I.J., 494, 605
Graybill, F.A., 133, 271, 332, 605
Greenberg, B.G., 335, 608
Gross, A.J., 112, 341, 605
Guenther, W.C., 410, 605
Gumbel, E.J., 113, 135, 605
Guttman, I., 605

Haberman, S.J., 275, 333, 605
Hacking, I., 486, 606
Hald, A., 606
Hall, W.J., 342, 606
Hampel, F.R., 601
Harrison, P.J., 506, 606, 610
Hettmansperger, T.P., 457, 606
Hinkley, D.V., 248, 603
Hoerl, A.E., 335–337, 606
Holland, P.W., 275, 445, 602, 606
Huber, P.J., 352, 442, 601, 606

James, W., 575, 606
Jeffreys, H., 488, 606
Jensen, 150
Jogdio, K., 256, 606
Johnson, N.L., 107, 115, 118, 137, 159, 606
Joshi, V.M., 325, 357, 606
Judge, G.G., 349, 585, 606

Kadane, J.B., 508, 610
Karlin, S., 251, 494, 574, 606
Kendall, M.G., 120, 603
Kennard, R.W., 335–337, 606
Khan, R.A., 421, 606
Kiefer, J., 283, 327, 606
Klotz, J.H., 581, 607
Kotz, S., 107, 115, 118, 137, 159, 606
Krishnan, T., 514, 607
Kubokawa, T., 373, 607
Kullback, S., 275, 605, 607

Lai, T.L., 283, 607
Lancaster, H.O., 273, 305, 306, 607
Le Cam, L., 442, 443, 607
Lehmann, E.L., 248, 256, 271, 322, 574, 607
Lin, P.E., 514, 583, 607
Lindley, D.V., 488, 502, 503, 584, 607
Lindsey, J.K., 340, 607
Lugannani, R., 607
Lugannani-Rice, R., 150

Maritz, J.L., 514, 607
Marlow, W.H., 545, 549, 602
Marquardt, D.W., 335, 337, 607
McLachlan, G.J., 514, 607
Migon, H.S., 506, 610
Miller, R.G., 415, 417, 607
Milton, R.C., 581, 607
Morris, C., 514, 584, 604, 607
Mukhopadhyay, N., 421, 605, 607

Neyman, J., 193, 248, 256, 369, 607, 608

Pearson, E.S., 248, 256, 607, 608
Pfanzagl, J., 203, 444, 602, 608
Pitman, E.J.G., 214, 330, 442, 608
Portnoy, S., 566, 608
Proschan, F., 414, 601

Rabinowitz, P., 603
Raghavachari, M., 330, 603
Raiffa, H., 487, 608
Rao, C.R., 322, 323, 332, 444, 608
Reid, N., 150, 447, 508, 608
Rice, S., 607
Robbins, H., 283, 327, 421, 513, 514, 603, 608
Roberts, D.M., 602
Rogatko, A., 500, 608
Rogers, W.H., 601
Rohatgi, V.K., 608
Ronchetti, E., 150, 604
Rubin, D. B., 605
Rubin, H., 494, 503, 555, 606

Sackrowitz, H.B., 373, 603
Saleh, A.K.M.E., 349, 602
Sarhan, A.E., 335, 608
Savage, L.J., 486, 570, 572, 605, 608
Scheffé, H., 251, 256, 304, 322, 417, 603, 607–609
Schervish, M.J., 203, 204, 442, 486, 609
Schlaifer, R., 487, 608
Schmetterer, L., 203, 248, 609
Schrage, L.E., 510, 602
Scott, E.L., 369, 608
Searle, S.R., 271, 332, 609
Seber, G.A.F., 332, 547, 609
Sen, A., 332, 609
Sen, P.K., 55, 421, 605, 609
Severo, N.C., 114, 118, 611
Shiryayev, A.N., 4, 40, 41, 46, 48, 53, 497, 609
Siegmund, D., 276
Siegmund, D., 283, 603, 609
Simons, G., 420, 609
Singer, J.M., 55, 609
Sirjaev, A.N., 609
Skovgaard, Ib M., 150, 609
Smith, A.F.M., 502, 503, 584, 607, 609
Smith, H., 332, 388, 603
Solomon, H., 293, 469, 610
Srivastava, M.S., 332, 421, 609
Starr, N., 421, 609
Stegun, I.A., 113, 382, 462, 470, 483, 510, 601, 611
Stein, C., 45, 47, 368, 418, 574, 575, 606, 609
Stern, H.S., 555, 605
Stevens, C.F., 506, 606
Stone, J., 335, 609
Strawderman, W.E., 575, 583, 609
Susarla, V., 514, 609
Sverdrup, E., 256, 610
Swartz, T., 508, 604

Takeuchi, K., 444, 601
Tan, P., 203, 610
Tan, W.Y., 566, 610
Tiao, G.C., 501, 503, 537, 566, 602, 610
Tierney, L.J., 508, 610
Tukey, J.W., 601

Wald, A., 251, 275, 283, 603, 610
Watson, G.S., 333, 610
Weiss, L., 283, 606
West, M., 506, 610
Wijsman, R.A., 277, 325, 342, 606, 610
Wilks, S.S., 261, 610
Wolfowitz, 283

Yushkevich, A.A., 497, 604

Zacks, S., 112, 203, 293, 341, 373, 381, 411, 421, 469, 494, 497, 500, 545, 549, 570, 572, 575, 579, 581, 587, 588, 602, 607, 608, 610
Zehna, P.W., 610
Zelen, M., 114, 118, 274, 610, 611
Zellner, A., 537, 611
Zidek, J.V., 575, 582, 602
Zyskind, G., 333, 611

Subject Index

$(1-\alpha)$ level p-content intervals, 411
(β, γ)-upper tolerance limit, 412
F-distributions, 179, 138
M-estimates, 351
P-value, 246
S-method, 417
α-trimmed means, 351
α-similar, 257
α-similar on the boundary, 257
σ-field, 2
σ-finite measure, 195
p-content prediction interval, 410
p-content tolerance intervals, 410
p-quantiles, 19
t-test, 285, 289, 290, 293, 294

absolutely continuous, 10, 196
acceptance probability, 281
admissibility of estimators, 570
algebra, 2
alternative hypothesis, 247
analysis of variance, 138, 139, 265, 266, 303, 415
ancillarity, 205
ancillary statistic, 205, 206, 210, 226, 233, 235, 355, 356, 375
association, 271, 272, 274, 296, 304, 305
asymptotic confidence intervals, 439, 479
asymptotic efficacy of T_n, 450
asymptotic efficiency of MLE, 443
asymptotic variance, 442
asymptotically efficient, 419, 442
asymptotically normal, 442
asymptotically size α test, 440
augmented symmetric functions, 120
autoregressive time-series, 172
average predictive error probability, 499

Bahadur's theorem, 204
Barndorff-Nielsen p^*-formula, 447, 508
Bartlett test, 302
Basu's theorem, 205, 206, 224, 591
Bayes, 337
Bayes decision function, 492
Bayes equivariant, 565, 566, 588, 594
Bayes factor, 494
Bayes procedures, 485
Bayes theorem, 8
Bayesian analysis, 138
Bayesian approach, 490
Bayesian estimation, 502
Bayesian inference, 485
Bayesian testing, 491
Bernoulli trials, 58, 106, 287
best linear combination of order statistics, 385
best linear unbiased estimators, 331
Beta distributions, 111
beta function, 107
Bhattacharyya lower bound, 326, 357, 383, 384
binomial distribution, 60, 67, 106, 273, 287

Examples and Problems in Mathematical Statistics, First Edition. Shelemyahu Zacks.

bivariate distribution function, 21
Bonferroni inequality, 271, 415
Borel σ-field, 3
Borel-Cantelli Lemma, 7
Box-Muller transformation, 511

C.R. regularity conditions, 323
canonical parameters, 145
Cantelli's theorem, 42
Cantor distribution, 11
Cauchy distribution, 63, 82, 184, 475
Central Limit Theorem, 44
central moments, 27
Chapman-Robbins inequality, 328, 384
characteristic function, 28, 32
Chebychev inequality, 27
Chebychev-Hermite polynomial, 147
chi-squared distribution, 112, 168
classical Bayesian model, 486
coefficient of correlation, 30
coefficient of determination, 172
coefficient of kurtosis, 178
complete class, 493
complete sufficient statistic, 203–205, 224, 239, 322
completeness, 203
conditional expectation, 23
conditional probability, 6
conditional test function, 259
confidence interval, 406
confidence level, 406
confidence limits, 406
confidence probability, 406
confidence region, 407
confluent-hypergeometric function, 292
conjugate families, 487
consistency of the MLE, 440, 442
consistent estimators, 439
contingency table, 274
continuity theorem, 41
contrasts, 416
converge in distribution, 36
convergence almost surely, 35
convergence in r-th mean, 35
convergence in distribution, 35
convergence in probability, 35
converges almost-surely, 36
converges in r-th mean, 36
convolution, 84, 118
covariance matrix, 31, 44
covariance stationary, 172
Cramér-Rao lower bound, 323, 325, 328, 329, 357, 383, 384, 398, 572, 573
Cramér-Rao regularity condition, 328
credibility intervals, 501
cross product ratio, 273, 425, 467
cumulants, 29
cumulants generating function, 29, 146
Cumulative Probability Integral Transformation, 110
curved exponential family, 146, 162, 180

decision function, 490
Definition of Sufficiency, 192
degrees of freedom, 112
delta method, 53
DeMorgan's laws, 74
di-gamma, 470
Dirichlet distribution, 119, 171, 552
discrete algebra, 2
dispersion, 19
distribution free (β, γ) upper tolerance limit, 412
distribution free confidence intervals, 412
distribution free test, 440
distribution of sums, 150
Dominated Convergence Theorem, 49
dynamic linear model, 504
Dynkin's regularity conditions, 202
Dynkin's theorem, 203

E-M algorithm, 547
Edgeworth approximation, 146, 148, 165, 166, 180, 181, 470
Edgeworth expansion, 147, 148, 165, 446, 447
efficiency function, 329
efficiency of multi-parameter estimator, 330
empirical Bayes, 584
empirical Bayes estimators, 513, 514, 544–546, 549, 556
empirical distribution function, 55
equivalent-likelihood partition, 201
equivariant, 343
equivariant estimator, 341, 343–346, 371–376, 380, 390, 403–405, 563, 565
error of Type I, 247
error of Type II, 247

estimating functions, 347
Euler constant, 117, 462
exchangeable, 502
exponential boundedness, 277
exponential conjugate, 149
exponential distribution, 73, 83, 112, 168, 170, 300, 301, 413, 428–430
exponential integral, 382
exponential type families, 144, 202
exponential type family, 144, 145, 162, 180, 251, 254, 256, 257, 288, 299, 300, 338
extreme value, 113

failure (hazard) rate function, 427
family of distributions, 191
Fatou's lemma, 49
Fatou's Theorem, 15
Fieller's method, 434
Fisher information function, 206–208, 225, 234, 235
Fisher information matrix, 212, 235, 236, 326, 330, 349, 358, 360
fixed width confidence intervals, 417
formal Bayes estimators, 566
fractiles, 19

Gamma distribution, 111, 220, 228
gamma function, 108
Gauss-Markov Theorem, 333
geometric random, 108
Glivenko-Cantelli's Theorem, 55
goodness of fit test, 305
guaranteed coverage tolerance intervals, 411

Hardy-Weinberg model, 232, 233, 235, 390, 403, 477
Hellinger distance, 214, 236
Hellinger's distance, 214
Helmert transformation, 177
hierarchical models, 502, 503
highest posterior density, 501
Holder's Inequality, 53
Hunt-Stein Theorem, 570
hypergeometric distribution, 107

idempotent matrix, 132
importance density, 511
importance sampling, 511
improper priors, 488, 503, 511
incomplete beta function, 107
incomplete beta function ratio, 69, 138
incomplete gamma function, 108
independence, 26
independent events, 6
information in an estimating function, 348
interaction, 267, 273
interquartile range, 19
Invariance Principle, 338
inverse regression, 527

James-Stein, 337
Jeffreys prior density, 489
Jensen's Inequality, 52, 210
joint density function, 22
joint distributions, 21

Kalman filter, 505
Karlin's admissibility theorem, 574
Karlin's Lemma, 252, 424
Khinchin WLLN, 89
Kullback-Leibler information, 206, 210, 227, 235, 241, 275, 299, 497
kurtosis, 29, 392

Lagrangian, 262, 331
Laplace distribution, 295, 376, 384, 385, 393, 400
Laplace method, 506
large sample confidence intervals, 445
large sample tests, 448
law of iterated logarithm, 48
law of total variance, 68
laws of large numbers, 42
least-squares, 332
least-squares estimators, 177, 262, 263, 301, 331
Lebesgue dominated convergence, 16
Lebesgue integral, 11, 13
likelihood functions, 200
likelihood ratio test, 260, 261, 274, 302, 303
likelihood statistic, 201, 202, 236, 237
Lindeberg-Feller Theorem, 46
linear models, 332
link functions, 340
local hypotheses, 473
location parameter, 110, 113, 300
log-convex, 427

log-normal distribution, 169, 181, 353, 366, 377, 411, 478, 515
log-normal distributions, 353
loss functions, 489
Lyapunov's Inequality, 53
Lyapunov's Theorem, 46, 72

main effects, 267
marginal distributions, 21
Markov WLLN, 89
maximal invariant statistic, 342
maximum likelihood estimator, 337
measurable space, 3
median, 19
minimal sufficient statistic, 200, 201
minimax estimators, 563
minimax test, 300
minimum chi-squared estimator, 390
minimum variance unbiased estimator, 322
Minkowsky's Inequality, 53
mixed effect model, 304
mixture, 11
mixture of the two types, 19
Model II of Analysis of Variance, 163, 224, 294, 303, 317, 430, 587
modes of convergence, 35
moment generating function, 28, 30
moment of order, 26
moment-equations estimators, 346
monotone convergence, 14
monotone likelihood ratio, 251
monotonic class, 3
multinomial distribution, 122, 137, 157, 172, 173, 216, 224, 271, 303, 304
multinormal distribution, 125, 216, 217
multivariate hypergeometric distributions, 124
multivariate negative binomial, 123, 124
multivariate normal distribution, 44
multivariate-t, 137

Negative-Binomial, 109, 191
Newton-Raphson method, 340
Neyman structure, 257
Neyman-Fisher Factorization Theorem, 193, 199
Neyman-Pearson Lemma, 248
non-central F, 140, 141, 269
non-central t, 136
non-central chi-squared, 130, 140
non-informative prior, 488
non-parametric test, 440
normal approximations, 114, 298, 307, 308
normal distribution, 113, 227, 229–231, 233, 265, 266, 283, 284, 286, 289, 290, 297, 302, 347, 360, 365, 366, 369, 373
normal regression, 261
nuisance parameter, 498, 568
nuisance parameters-unbiased tests, 256
null hypothesis, 247

observed significance level, 246
odds-ratio, 273
operating characteristic function, 282
orbit of $\mathcal{G}$, 343
order of magnitude in probability, 55
order statistics, 56, 133–135, 161, 177, 178
orthogonal subvectors, 214

parameter of non-centrality, 130
parameter space, 191
parametric empirical Bayes, 514
Pareto distribution, 549
partial correlation, 129, 174
partition of sample space, 2
Pascal distribution, 109, 167
Pitman ARE, 449, 450
Pitman estimator, 344, 346, 390, 391, 566, 568, 570, 574, 575, 589, 594
Pitman relative efficiency, 330
Poisson distributions, 108, 216, 236, 288, 291, 298, 355, 513, 515, 516, 529, 531, 539, 550, 551, 559
posterior distribution $H(\theta \mid X)$, 486
posterior risk, 491
power of a test, 247
pre-test estimators, 349
precision parameter, 517
predictive distributions, 487
predictive error probability, 499
prior distribution, 485
prior risk, 485
probability generating function, 29
probability measure, 9
probability model, 4
proportional-closeness, 432
psi function, 470

Radon-Nikodym, 196
Radon-Nikodym theorem, 10
random sample, 120, 195
random variable, 8, 9
random walk, 57, 77, 505
randomized test, 247
Rao's efficient score statistic, 449
Rao-Blackwell Lehmann-Scheffé Theorem, 322
rectangular distribution, 109, 215, 225, 230, 247, 284, 305, 350, 353, 356, 363, 364, 381, 384, 385, 390, 393
regression analysis, 133
regular, 193
regular family of distributions, 200
regular parametric family, 202
regularity conditions, 206
regularity conditions for estimating functions, 348
relative betting odds, 494
relative efficiency, 329
ridge regression, 335, 336, 363
ridge trace, 336
Riemann integrable, 11
risk function, 490
robust estimation, 349, 353

saddlepoint approximation, 146, 149, 439, 447
sample correlation, 142
sample median, 134
sample quantile, 56
sample range, 134
scale parameter, 110, 111, 113
Schwarz inequality, 30, 43, 52
score function, 206
second order deficiency, 444, 464, 478, 482
second order efficiency, 444
sensitivity, 214
sequential fixed-width interval estimation, 419
shape parameter, 111, 112
simulation, 510
simultaneous confidence intervals, 414
simultaneous coverage probability, 415
size of a test, 247
skewness, 29, 392
Slutsky's Theorem, 40
standard deviation, 27
standard normal distribution, 113
standard normal integral, 44
standard-errors, 135, 178
statistic, 192, 197
statistical model, 196
Stein type estimators, 584
Stein's two-state procedure, 418
Stieltjes-Riemann integral, 17
Stirling approximation, 92, 167
stopping variable, 276, 420
Strong Law of Large Numbers, 42
strongly consistent, 440
structural distributions, 568
structural estimators, 565
student's t-distribution, 135
subalgebra, 2
sufficient statistics, 192, 193, 199, 201
super efficient, 443
symmetric difference, 74

test function, 247
tetrachoric correlation, 174, 188
The Delta Method, 53
The Law of Total Variance, 30
tight family of distribution, 41
tolerance distributions, 339, 340, 369, 389
tolerance intervals, 407
total life, 428
transformations, 20, 118
trimean, 351
Type I censoring, 458

unbiased estimator, 322
unbiased test, 256, 257, 273, 301
uniform distribution, 67
uniform integrability, 48
uniformly most accurate confidence interval, 408
uniformly most accurate unbiased, 410
uniformly most powerful, 248
upper quartiles, 19
utility function, 489

variance, 27, 124
variance components, 163
variance stabilizing transformation, 442, 445, 466, 468, 478
variance-covariance matrix, 69

Wald Fundamental Identity, 280
Wald Sequential Probability Ratio Test, 275, 276, 280, 282, 283, 297, 306, 497, 498
Wald statistic, 448
Wald Theorem, 277
weak convergence, 38
Weak Law of Large Numbers, 41
Weibull distribution, 286
Weibull distributions, 112, 221, 233, 236, 331, 344, 367, 374, 378, 461
Wiener process, 497
Wilcoxon signed-rank test, 455
Wilks' likelihood ratio statistic, 449

WILEY SERIES IN PROBABILITY AND STATISTICS

ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

The ***Wiley Series in Probability and Statistics*** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

† ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data, *Second Edition*
AGRESTI · An Introduction to Categorical Data Analysis, *Second Edition*
AGRESTI · Categorical Data Analysis, *Third Edition*
ALTMAN, GILL, and McDONALD · Numerical Issues in Statistical Computing for the Social Scientist
AMARATUNGA and CABRERA · Exploration and Analysis of DNA Microarray and Protein Array Data
AMARATUNGA, CABRERA, and SHKEDY · Exploration and Analysis of DNA Microarray and Other High-Dimensional Data, *Second Edition*
ANDĚL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, *Third Edition*
* ANDERSON · The Statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG · Statistical Methods for Comparative Studies
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and DAVID (editors) · Advances in Biometry
ARNOLD, BALAKRISHNAN, and NAGARAJA · Records
* ARTHANARI and DODGE · Mathematical Programming in Statistics
* BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BAJORSKI · Statistics for Imaging, Optics, and Photonics
BALAKRISHNAN and KOUTRAS · Runs and Scans with Applications
BALAKRISHNAN and NG · Precedence-Type Tests and Applications
BARNETT · Comparative Statistical Inference, *Third Edition*
BARNETT · Environmental Statistics
BARNETT and LEWIS · Outliers in Statistical Data, *Third Edition*
BARTHOLOMEW, KNOTT, and MOUSTAKI · Latent Variable Models and Factor Analysis: A Unified Approach, *Third Edition*
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

BASILEVSKY · Statistical Factor Analysis and Related Methods: Theory and Applications

BATES and WATTS · Nonlinear Regression Analysis and Its Applications

BECHHOFER, SANTNER, and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons

BEIRLANT, GOEGEBEUR, SEGERS, TEUGELS, and DE WAAL · Statistics of Extremes: Theory and Applications

BELSLEY · Conditioning Diagnostics: Collinearity and Weak Data in Regression

† BELSLEY, KUH, and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity

BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, *Fourth Edition*

BERNARDO and SMITH · Bayesian Theory

BHAT and MILLER · Elements of Applied Stochastic Processes, *Third Edition*

BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications

BIEMER, GROVES, LYBERG, MATHIOWETZ, and SUDMAN · Measurement Errors in Surveys

BILLINGSLEY · Convergence of Probability Measures, *Second Edition*

BILLINGSLEY · Probability and Measure, *Anniversary Edition*

BIRKES and DODGE · Alternative Methods of Regression

BISGAARD and KULAHCI · Time Series Analysis and Forecasting by Example

BISWAS, DATTA, FINE, and SEGAL · Statistical Advances in the Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics

BLISCHKE and MURTHY (editors) · Case Studies in Reliability and Maintenance

BLISCHKE and MURTHY · Reliability: Modeling, Prediction, and Optimization

BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, *Second Edition*

BOLLEN · Structural Equations with Latent Variables

BOLLEN and CURRAN · Latent Curve Models: A Structural Equation Perspective

BOROVKOV · Ergodicity and Stability of Stochastic Processes

BOSQ and BLANKE · Inference and Prediction in Large Dimensions

BOULEAU · Numerical Methods for Stochastic Processes

* BOX and TIAO · Bayesian Inference in Statistical Analysis

BOX · Improving Almost Anything, *Revised Edition*

* BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement

BOX and DRAPER · Response Surfaces, Mixtures, and Ridge Analyses, *Second Edition*

BOX, HUNTER, and HUNTER · Statistics for Experimenters: Design, Innovation, and Discovery, *Second Editon*

BOX, JENKINS, and REINSEL · Time Series Analysis: Forcasting and Control, *Fourth Edition*

BOX, LUCEÑO, and PANIAGUA-QUIÑONES · Statistical Control by Monitoring and Adjustment, *Second Edition*

* BROWN and HOLLANDER · Statistics: A Biomedical Introduction

CAIROLI and DALANG · Sequential Stochastic Optimization

CASTILLO, HADI, BALAKRISHNAN, and SARABIA · Extreme Value and Related Models with Applications in Engineering and Science

CHAN · Time Series: Applications to Finance with R and S-Plus®, *Second Edition*

CHARALAMBIDES · Combinatorial Methods in Discrete Distributions

CHATTERJEE and HADI · Regression Analysis by Example, *Fourth Edition*

CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

CHERNICK · Bootstrap Methods: A Guide for Practitioners and Researchers, *Second Edition*
CHERNICK and FRIIS · Introductory Biostatistics for the Health Sciences
CHILÈS and DELFINER · Geostatistics: Modeling Spatial Uncertainty, *Second Edition*
CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies, *Second Edition*
CLARKE · Linear Models: The Theory and Application of Analysis of Variance
CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, *Second Edition*
* COCHRAN and COX · Experimental Designs, *Second Edition*
COLLINS and LANZA · Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences
CONGDON · Applied Bayesian Modelling
CONGDON · Bayesian Models for Categorical Data
CONGDON · Bayesian Statistical Modelling, *Second Edition*
CONOVER · Practical Nonparametric Statistics, *Third Edition*
COOK · Regression Graphics
COOK and WEISBERG · An Introduction to Regression Graphics
COOK and WEISBERG · Applied Regression Including Computing and Graphics
CORNELL · A Primer on Experiments with Mixtures
CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Third Edition*
COX · A Handbook of Introductory Statistical Methods
CRESSIE · Statistics for Spatial Data, *Revised Edition*
CRESSIE and WIKLE · Statistics for Spatio-Temporal Data
CSÖRGŐ and HORVÁTH · Limit Theorems in Change Point Analysis
DAGPUNAR · Simulation and Monte Carlo: With Applications in Finance and MCMC
DANIEL · Applications of Statistics to Industrial Experimentation
DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, *Eighth Edition*
* DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, *Second Edition*
DASU and JOHNSON · Exploratory Data Mining and Data Cleaning
DAVID and NAGARAJA · Order Statistics, *Third Edition*
* DEGROOT, FIENBERG, and KADANE · Statistics and the Law
DEL CASTILLO · Statistical Process Adjustment for Quality Control
DEMARIS · Regression with Social Data: Modeling Continuous and Limited Response Variables
DEMIDENKO · Mixed Models: Theory and Applications with R, *Second Edition*
DENISON, HOLMES, MALLICK and SMITH · Bayesian Methods for Nonlinear Classification and Regression
DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis
DEY and MUKERJEE · Fractional Factorial Plans
DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
* DODGE and ROMIG · Sampling Inspection Tables, *Second Edition*
* DOOB · Stochastic Processes
DOWDY, WEARDEN, and CHILKO · Statistics for Research, *Third Edition*
DRAPER and SMITH · Applied Regression Analysis, *Third Edition*
DRYDEN and MARDIA · Statistical Shape Analysis
DUDEWICZ and MISHRA · Modern Mathematical Statistics
DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, *Fourth Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations
EDLER and KITSOS · Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment
* ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
ENDERS · Applied Econometric Time Series, *Third Edition*
† ETHIER and KURTZ · Markov Processes: Characterization and Convergence
EVANS, HASTINGS, and PEACOCK · Statistical Distributions, *Third Edition*
EVERITT, LANDAU, LEESE, and STAHL · Cluster Analysis, *Fifth Edition*
FEDERER and KING · Variations on Split Plot and Split Block Experiment Designs
FELLER · An Introduction to Probability Theory and Its Applications, Volume I, *Third Edition*, Revised; Volume II, *Second Edition*
FITZMAURICE, LAIRD, and WARE · Applied Longitudinal Analysis, *Second Edition*
* FLEISS · The Design and Analysis of Clinical Experiments
FLEISS · Statistical Methods for Rates and Proportions, *Third Edition*
† FLEMING and HARRINGTON · Counting Processes and Survival Analysis
FUJIKOSHI, ULYANOV, and SHIMIZU · Multivariate Statistics: High-Dimensional and Large-Sample Approximations
FULLER · Introduction to Statistical Time Series, *Second Edition*
† FULLER · Measurement Error Models
GALLANT · Nonlinear Statistical Models
GEISSER · Modes of Parametric Statistical Inference
GELMAN and MENG · Applied Bayesian Modeling and Causal Inference from incomplete-Data Perspectives
GEWEKE · Contemporary Bayesian Econometrics and Statistics
GHOSH, MUKHOPADHYAY, and SEN · Sequential Estimation
GIESBRECHT and GUMPERTZ · Planning, Construction, and Statistical Analysis of Comparative Experiments
GIFI · Nonlinear Multivariate Analysis
GIVENS and HOETING · Computational Statistics
GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems
GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, *Second Edition*
GOLDSTEIN · Multilevel Statistical Models, *Fourth Edition*
GOLDSTEIN and LEWIS · Assessment: Problems, Development, and Statistical Issues
GOLDSTEIN and WOOFF · Bayes Linear Statistics
GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing
GROSS, SHORTLE, THOMPSON, and HARRIS · Fundamentals of Queueing Theory, *Fourth Edition*
GROSS, SHORTLE, THOMPSON, and HARRIS · Solutions Manual to Accompany Fundamentals of Queueing Theory, *Fourth Edition*
* HAHN and SHAPIRO · Statistical Models in Engineering
HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners
HALD · A History of Probability and Statistics and their Applications Before 1750
† HAMPEL · Robust Statistics: The Approach Based on Influence Functions
HARTUNG, KNAPP, and SINHA · Statistical Meta-Analysis with Applications
HEIBERGER · Computation for the Analysis of Designed Experiments
HEDAYAT and SINHA · Design and Inference in Finite Population Sampling
HEDEKER and GIBBONS · Longitudinal Data Analysis
HELLER · MACSYMA for Statisticians
HERITIER, CANTONI, COPT, and VICTORIA-FESER · Robust Methods in Biostatistics

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design, *Second Edition*
HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 2: Advanced Experimental Design
HINKELMANN (editor) · Design and Analysis of Experiments, Volume 3: Special Designs and Applications
HOAGLIN, MOSTELLER, and TUKEY · Fundamentals of Exploratory Analysis of Variance
* HOAGLIN, MOSTELLER, and TUKEY · Exploring Data Tables, Trends and Shapes
* HOAGLIN, MOSTELLER, and TUKEY · Understanding Robust and Exploratory Data Analysis
HOCHBERG and TAMHANE · Multiple Comparison Procedures
HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variance, *Third Edition*
HOEL · Introduction to Mathematical Statistics, *Fifth Edition*
HOGG and KLUGMAN · Loss Distributions
HOLLANDER, WOLFE, and CHICKEN · Nonparametric Statistical Methods, *Third Edition*
HOSMER and LEMESHOW · Applied Logistic Regression, *Second Edition*
HOSMER, LEMESHOW, and MAY · Applied Survival Analysis: Regression Modeling of Time-to-Event Data, *Second Edition*
HUBER · Data Analysis: What Can Be Learned From the Past 50 Years
HUBER · Robust Statistics
† HUBER and RONCHETTI · Robust Statistics, *Second Edition*
HUBERTY · Applied Discriminant Analysis, *Second Edition*
HUBERTY and OLEJNIK · Applied MANOVA and Discriminant Analysis, *Second Edition*
HUITEMA · The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, *Second Edition*
HUNT and KENNEDY · Financial Derivatives in Theory and Practice, *Revised Edition*
HURD and MIAMEE · Periodically Correlated Random Sequences: Spectral Theory and Practice
HUSKOVA, BERAN, and DUPAC · Collected Works of Jaroslav Hajek—with Commentary
HUZURBAZAR · Flowgraph Models for Multistate Time-to-Event Data
JACKMAN · Bayesian Analysis for the Social Sciences
† JACKSON · A User's Guide to Principle Components
JOHN · Statistical Methods in Engineering and Quality Assurance
JOHNSON · Multivariate Statistical Simulation
JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz
JOHNSON, KEMP, and KOTZ · Univariate Discrete Distributions, *Third Edition*
JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present
JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 1, *Second Edition*
JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 2, *Second Edition*
JOHNSON, KOTZ, and BALAKRISHNAN · Discrete Multivariate Distributions
JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE · The Theory and Practice of Econometrics, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

JUREK and MASON · Operator-Limit Distributions in Probability Theory
KADANE · Bayesian Methods and Ethics in a Clinical Trial Design
KADANE AND SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence
KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, *Second Edition*
KARIYA and KURATA · Generalized Least Squares
KASS and VOS · Geometrical Foundations of Asymptotic Inference
† KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis
KEDEM and FOKIANOS · Regression Models for Time Series Analysis
KENDALL, BARDEN, CARNE, and LE · Shape and Shape Theory
KHURI · Advanced Calculus with Applications in Statistics, *Second Edition*
KHURI, MATHEW, and SINHA · Statistical Tests for Mixed Linear Models
* KISH · Statistical Design for Research
KLEIBER and KOTZ · Statistical Size Distributions in Economics and Actuarial Sciences
KLEMELÄ · Smoothing of Multivariate Data: Density Estimation and Visualization
KLUGMAN, PANJER, and WILLMOT · Loss Models: From Data to Decisions, *Third Edition*
KLUGMAN, PANJER, and WILLMOT · Loss Models: Further Topics
KLUGMAN, PANJER, and WILLMOT · Solutions Manual to Accompany Loss Models: From Data to Decisions, *Third Edition*
KOSKI and NOBLE · Bayesian Networks: An Introduction
KOTZ, BALAKRISHNAN, and JOHNSON · Continuous Multivariate Distributions, Volume 1, *Second Edition*
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Supplement Volume
KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 1
KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 2
KOWALSKI and TU · Modern Applied U-Statistics
KRISHNAMOORTHY and MATHEW · Statistical Tolerance Regions: Theory, Applications, and Computation
KROESE, TAIMRE, and BOTEV · Handbook of Monte Carlo Methods
KROONENBERG · Applied Multiway Data Analysis
KULINSKAYA, MORGENTHALER, and STAUDTE · Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence
KULKARNI and HARMAN · An Elementary Introduction to Statistical Learning Theory
KUROWICKA and COOKE · Uncertainty Analysis with High Dimensional Dependence Modelling
KVAM and VIDAKOVIC · Nonparametric Statistics with Applications to Science and Engineering
LACHIN · Biostatistical Methods: The Assessment of Relative Risks, *Second Edition*
LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction
LAMPERTI · Probability: A Survey of the Mathematical Theory, *Second Edition*
LAWLESS · Statistical Models and Methods for Lifetime Data, *Second Edition*
LAWSON · Statistical Methods in Spatial Epidemiology, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

LE · Applied Categorical Data Analysis, *Second Edition*
LE · Applied Survival Analysis
LEE · Structural Equation Modeling: A Bayesian Approach
LEE and WANG · Statistical Methods for Survival Data Analysis, *Fourth Edition*
LEPAGE and BILLARD · Exploring the Limits of Bootstrap
LESSLER and KALSBEEK · Nonsampling Errors in Surveys
LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics
LIAO · Statistical Group Comparison
LIN · Introductory Stochastic Analysis for Finance and Insurance
LITTLE and RUBIN · Statistical Analysis with Missing Data, *Second Edition*
LLOYD · The Statistical Analysis of Categorical Data
LOWEN and TEICH · Fractal-Based Point Processes
MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, *Revised Edition*
MALLER and ZHOU · Survival Analysis with Long Term Survivors
MARCHETTE · Random Graphs for Statistical Pattern Recognition
MARDIA and JUPP · Directional Statistics
MARKOVICH · Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice
MARONNA, MARTIN and YOHAI · Robust Statistics: Theory and Methods
MASON, GUNST, and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science, *Second Edition*
McCULLOCH, SEARLE, and NEUHAUS · Generalized, Linear, and Mixed Models, *Second Edition*
McFADDEN · Management of Data in Clinical Trials, *Second Edition*
* McLACHLAN · Discriminant Analysis and Statistical Pattern Recognition
McLACHLAN, DO, and AMBROISE · Analyzing Microarray Gene Expression Data
McLACHLAN and KRISHNAN · The EM Algorithm and Extensions, *Second Edition*
McLACHLAN and PEEL · Finite Mixture Models
McNEIL · Epidemiological Research Methods
MEEKER and ESCOBAR · Statistical Methods for Reliability Data
MEERSCHAERT and SCHEFFLER · Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice
MENGERSEN, ROBERT, and TITTERINGTON · Mixtures: Estimation and Applications
MICKEY, DUNN, and CLARK · Applied Statistics: Analysis of Variance and Regression, *Third Edition*
* MILLER · Survival Analysis, *Second Edition*
MONTGOMERY, JENNINGS, and KULAHCI · Introduction to Time Series Analysis and Forecasting
MONTGOMERY, PECK, and VINING · Introduction to Linear Regression Analysis, *Fifth Edition*
MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness
MUIRHEAD · Aspects of Multivariate Statistical Theory
MULLER and STOYAN · Comparison Methods for Stochastic Models and Risks
MURTHY, XIE, and JIANG · Weibull Models
MYERS, MONTGOMERY, and ANDERSON-COOK · Response Surface Methodology: Process and Product Optimization Using Designed Experiments, *Third Edition*
MYERS, MONTGOMERY, VINING, and ROBINSON · Generalized Linear Models. With Applications in Engineering and the Sciences, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

NATVIG · Multistate Systems Reliability Theory With Applications
† NELSON · Accelerated Testing, Statistical Models, Test Plans, and Data Analyses
† NELSON · Applied Life Data Analysis
NEWMAN · Biostatistical Methods in Epidemiology
NG, TAIN, and TANG · Dirichlet Theory: Theory, Methods and Applications
OKABE, BOOTS, SUGIHARA, and CHIU · Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, *Second Edition*
OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis
PALTA · Quantitative Methods in Population Health: Extensions of Ordinary Regressions
PANJER · Operational Risk: Modeling and Analytics
PANKRATZ · Forecasting with Dynamic Regression Models
PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
PARDOUX · Markov Processes and Applications: Algorithms, Networks, Genome and Finance
PARMIGIANI and INOUE · Decision Theory: Principles and Approaches
* PARZEN · Modern Probability Theory and Its Applications
PEÑA, TIAO, and TSAY · A Course in Time Series Analysis
PESARIN and SALMASO · Permutation Tests for Complex Data: Applications and Software
PIANTADOSI · Clinical Trials: A Methodologic Perspective, *Second Edition*
POURAHMADI · Foundations of Time Series Analysis and Prediction Theory
POURAHMADI · High-Dimensional Covariance Estimation
POWELL · Approximate Dynamic Programming: Solving the Curses of Dimensionality, *Second Edition*
POWELL and RYZHOV · Optimal Learning
PRESS · Subjective and Objective Bayesian Statistics, *Second Edition*
PRESS and TANUR · The Subjectivity of Scientists and the Bayesian Approach
PURI, VILAPLANA, and WERTZ · New Perspectives in Theoretical and Applied Statistics
† PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming
QIU · Image Processing and Jump Regression Analysis
* RAO · Linear Statistical Inference and Its Applications, *Second Edition*
RAO · Statistical Inference for Fractional Diffusion Processes
RAUSAND and HØYLAND · System Reliability Theory: Models, Statistical Methods, and Applications, *Second Edition*
RAYNER, THAS, and BEST · Smooth Tests of Goodnes of Fit: Using R, *Second Edition*
RENCHER and SCHAALJE · Linear Models in Statistics, *Second Edition*
RENCHER and CHRISTENSEN · Methods of Multivariate Analysis, *Third Edition*
RENCHER · Multivariate Statistical Inference with Applications
RIGDON and BASU · Statistical Methods for the Reliability of Repairable Systems
* RIPLEY · Spatial Statistics
* RIPLEY · Stochastic Simulation
ROHATGI and SALEH · An Introduction to Probability and Statistics, *Second Edition*
ROLSKI, SCHMIDLI, SCHMIDT, and TEUGELS · Stochastic Processes for Insurance and Finance
ROSENBERGER and LACHIN · Randomization in Clinical Trials: Theory and Practice
ROSSI, ALLENBY, and McCULLOCH · Bayesian Statistics and Marketing
† ROUSSEEUW and LEROY · Robust Regression and Outlier Detection
ROYSTON and SAUERBREI · Multivariate Model Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modeling Continuous Variables

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

* RUBIN · Multiple Imputation for Nonresponse in Surveys
RUBINSTEIN and KROESE · Simulation and the Monte Carlo Method, *Second Edition*
RUBINSTEIN and MELAMED · Modern Simulation and Modeling
RUBINSTEIN, RIDDER, and VAISMAN · Fast Sequential Monte Carlo Methods for Counting and Optimization
RYAN · Modern Engineering Statistics
RYAN · Modern Experimental Design
RYAN · Modern Regression Methods, *Second Edition*
RYAN · Sample Size Determination and Power
RYAN · Statistical Methods for Quality Improvement, *Third Edition*
SALEH · Theory of Preliminary Test and Stein-Type Estimation with Applications
SALTELLI, CHAN, and SCOTT (editors) · Sensitivity Analysis
SCHERER · Batch Effects and Noise in Microarray Experiments: Sources and Solutions
* SCHEFFE · The Analysis of Variance
SCHIMEK · Smoothing and Regression: Approaches, Computation, and Application
SCHOTT · Matrix Analysis for Statistics, *Second Edition*
SCHOUTENS · Levy Processes in Finance: Pricing Financial Derivatives
SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization
* SEARLE · Linear Models
† SEARLE · Linear Models for Unbalanced Data
† SEARLE · Matrix Algebra Useful for Statistics
† SEARLE, CASELLA, and McCULLOCH · Variance Components
SEARLE and WILLETT · Matrix Algebra for Applied Economics
SEBER · A Matrix Handbook For Statisticians
† SEBER · Multivariate Observations
SEBER and LEE · Linear Regression Analysis, *Second Edition*
† SEBER and WILD · Nonlinear Regression
SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems
* SERFLING · Approximation Theorems of Mathematical Statistics
SHAFER and VOVK · Probability and Finance: It's Only a Game!
SHERMAN · Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties
SILVAPULLE and SEN · Constrained Statistical Inference: Inequality, Order, and Shape Restrictions
SINGPURWALLA · Reliability and Risk: A Bayesian Perspective
SMALL and McLEISH · Hilbert Space Methods in Probability and Statistical Inference
SRIVASTAVA · Methods of Multivariate Statistics
STAPLETON · Linear Statistical Models, *Second Edition*
STAPLETON · Models for Probability and Statistical Inference: Theory and Applications
STAUDTE and SHEATHER · Robust Estimation and Testing
STOYAN · Counterexamples in Probability, *Second Edition*
STOYAN, KENDALL, and MECKE · Stochastic Geometry and Its Applications, *Second Edition*
STOYAN and STOYAN · Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics
STREET and BURGESS · The Construction of Optimal Stated Choice Experiments: Theory and Methods
STYAN · The Collected Papers of T. W. Anderson: 1943–1985
SUTTON, ABRAMS, JONES, SHELDON, and SONG · Methods for Meta-Analysis in Medical Research

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

TAKEZAWA · Introduction to Nonparametric Regression
TAMHANE · Statistical Analysis of Designed Experiments: Theory and Applications
TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
THOMPSON · Empirical Model Building: Data, Models, and Reality, *Second Edition*
THOMPSON · Sampling, *Third Edition*
THOMPSON · Simulation: A Modeler's Approach
THOMPSON and SEBER · Adaptive Sampling
THOMPSON, WILLIAMS, and FINDLAY · Models for Investors in Real World Markets
TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
TSAY · Analysis of Financial Time Series, *Third Edition*
TSAY · An Introduction to Analysis of Financial Data with R
UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
† VAN BELLE · Statistical Rules of Thumb, *Second Edition*
VAN BELLE, FISHER, HEAGERTY, and LUMLEY · Biostatistics: A Methodology for the Health Sciences, *Second Edition*
VESTRUP · The Theory of Measures and Integration
VIDAKOVIC · Statistical Modeling by Wavelets
VIERTL · Statistical Methods for Fuzzy Data
VINOD and REAGLE · Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments
WALLER and GOTWAY · Applied Spatial Statistics for Public Health Data
WEISBERG · Applied Linear Regression, *Third Edition*
WEISBERG · Bias and Causation: Models and Judgment for Valid Comparisons
WELSH · Aspects of Statistical Inference
WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment
* WHITTAKER · Graphical Models in Applied Multivariate Statistics
WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting
WOODWORTH · Biostatistics: A Bayesian Introduction
WOOLSON and CLARKE · Statistical Methods for the Analysis of Biomedical Data, *Second Edition*
WU and HAMADA · Experiments: Planning, Analysis, and Parameter Design Optimization, *Second Edition*
WU and ZHANG · Nonparametric Regression Methods for Longitudinal Data Analysis
YIN · Clinical Trial Design: Bayesian and Frequentist Adaptive Methods
YOUNG, VALERO-MORA, and FRIENDLY · Visual Statistics: Seeing Data with Dynamic Interactive Graphics
ZACKS · Examples and Problems in Mathematical Statistics
ZACKS · Stage-Wise Adaptive Designs
* ZELLNER · An Introduction to Bayesian Inference in Econometrics
ZELTERMAN · Discrete Distributions—Applications in the Health Sciences
ZHOU, OBUCHOWSKI, and McCLISH · Statistical Methods in Diagnostic Medicine, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.