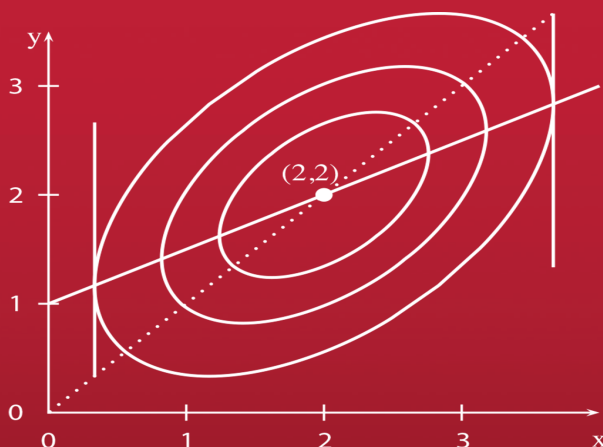


Texts in Statistical Science

Mathematical Statistics

Basic Ideas and
Selected Topics

Volume I
Second Edition



Peter J. Bickel
Kjell A. Doksum



CRC Press
Taylor & Francis Group

A CHAPMAN & HALL BOOK

Mathematical Statistics

Basic Ideas and
Selected Topics

Volume I

Second Edition

CHAPMAN & HALL/CRC

Texts in Statistical Science Series

Series Editors

Francesca Dominici, *Harvard School of Public Health, USA*

Julian J. Faraway, *University of Bath, UK*

Martin Tanner, *Northwestern University, USA*

Jim Zidek, *University of British Columbia, Canada*

Statistical Theory: A Concise Introduction

F. Abramovich and Y. Ritov

Practical Multivariate Analysis, Fifth Edition

A. Afifi, S. May, and V.A. Clark

Practical Statistics for Medical Research

D.G. Altman

Interpreting Data: A First Course in Statistics

A.J.B. Anderson

Introduction to Probability with R

K. Baclawski

Linear Algebra and Matrix Analysis for Statistics

S. Banerjee and A. Roy

Mathematical Statistics: Basic Ideas and Selected Topics, Volume I, Second Edition

P. J. Bickel and K. A. Doksum

Analysis of Categorical Data with R

C. R. Bilder and T. M. Loughin

Statistical Methods for SPC and TQM

D. Bissell

Introduction to Probability

J. K. Blitzstein and J. Hwang

Bayesian Methods for Data Analysis, Third Edition

B.P. Carlin and T.A. Louis

Second Edition

R. Caulcutt

The Analysis of Time Series: An Introduction, Sixth Edition

C. Chatfield

Introduction to Multivariate Analysis

C. Chatfield and A.J. Collins

Problem Solving: A Statistician's Guide, Second Edition

C. Chatfield

Statistics for Technology: A Course in Applied Statistics, Third Edition

C. Chatfield

Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians

R. Christensen, W. Johnson, A. Branscum, and T.E. Hanson

Modelling Binary Data, Second Edition

D. Collett

Modelling Survival Data in Medical Research, Third Edition

D. Collett

Introduction to Statistical Methods for Clinical Trials

T.D. Cook and D.L. DeMets

Applied Statistics: Principles and Examples

D.R. Cox and E.J. Snell

Multivariate Survival Analysis and Competing Risks

M. Crowder

Statistical Analysis of Reliability Data

M.J. Crowder, A.C. Kimber, T.J. Sweeting, and R.L. Smith

An Introduction to Generalized Linear Models, Third Edition

A.J. Dobson and A.G. Barnett

Nonlinear Time Series: Theory, Methods, and Applications with R Examples

R. Douc, E. Moulines, and D.S. Stoffer

Introduction to Optimization Methods and Their Applications in Statistics

B.S. Everitt

Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models

J.J. Faraway

Linear Models with R, Second Edition

J.J. Faraway

A Course in Large Sample Theory

T.S. Ferguson

Multivariate Statistics: A Practical Approach

B. Flury and H. Riedwyl

Readings in Decision Analysis

S. French

**Markov Chain Monte Carlo:
Stochastic Simulation for Bayesian Inference,
Second Edition**

D. Gamerman and H.F. Lopes

Bayesian Data Analysis, Third Edition

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson,
A. Vehtari, and D.B. Rubin

**Multivariate Analysis of Variance and
Repeated Measures: A Practical Approach for
Behavioural Scientists**

D.J. Hand and C.C. Taylor

Practical Longitudinal Data Analysis

D.J. Hand and M. Crowder

Logistic Regression Models

J.M. Hilbe

**Richly Parameterized Linear Models:
Additive, Time Series, and Spatial Models
Using Random Effects**

J.S. Hodges

Statistics for Epidemiology

N.P. Jewell

**Stochastic Processes: An Introduction,
Second Edition**

P.W. Jones and P. Smith

The Theory of Linear Models

B. Jørgensen

Principles of Uncertainty

J.B. Kadane

Graphics for Statistics and Data Analysis with R

K.J. Keen

Mathematical Statistics

K. Knight

Introduction to Multivariate Analysis:

Linear and Nonlinear Modeling

S. Konishi

**Nonparametric Methods in Statistics with SAS
Applications**

O. Korosteleva

**Modeling and Analysis of Stochastic Systems,
Second Edition**

V.G. Kulkarni

Exercises and Solutions in Biostatistical Theory

L.L. Kupper, B.H. Neelon, and S.M. O'Brien

Exercises and Solutions in Statistical Theory

L.L. Kupper, B.H. Neelon, and S.M. O'Brien

Design and Analysis of Experiments with R

J. Lawson

Design and Analysis of Experiments with SAS

J. Lawson

A Course in Categorical Data Analysis

T. Leonard

Statistics for Accountants

S. Letchford

**Introduction to the Theory of Statistical
Inference**

H. Liero and S. Zwanzig

Statistical Theory, Fourth Edition

B.W. Lindgren

**Stationary Stochastic Processes: Theory and
Applications**

G. Lindgren

**The BUGS Book: A Practical Introduction to
Bayesian Analysis**

D. Lunn, C. Jackson, N. Best, A. Thomas, and
D. Spiegelhalter

**Introduction to General and Generalized
Linear Models**

H. Madsen and P. Thyregod

Time Series Analysis

H. Madsen

Pólya Urn Models

H. Mahmoud

**Randomization, Bootstrap and Monte Carlo
Methods in Biology, Third Edition**

B.F.J. Manly

**Introduction to Randomized Controlled
Clinical Trials, Second Edition**

J.N.S. Matthews

**Statistical Methods in Agriculture and
Experimental Biology, Second Edition**

R. Mead, R.N. Curnow, and A.M. Hasted

Statistics in Engineering: A Practical Approach

A.V. Metcalfe

**Statistical Inference: An Integrated Approach,
Second Edition**

H. S. Migon, D. Gamerman, and
F. Louzada

Beyond ANOVA: Basics of Applied Statistics

R.G. Miller, Jr.

A Primer on Linear Models

J.F. Monahan

Applied Stochastic Modelling, Second Edition

B.J.T. Morgan

Elements of Simulation

B.J.T. Morgan

Probability: Methods and Measurement

A. O'Hagan

Introduction to Statistical Limit Theory

A.M. Polansky

Applied Bayesian Forecasting and Time Series Analysis

A. Pole, M. West, and J. Harrison

Statistics in Research and Development, Time Series: Modeling, Computation, and Inference

R. Prado and M. West

Introduction to Statistical Process Control

P. Qiu

Sampling Methodologies with Applications

P.S.R.S. Rao

A First Course in Linear Model Theory

N. Ravishanker and D.K. Dey

Essential Statistics, Fourth Edition

D.A.G. Rees

Stochastic Modeling and Mathematical Statistics: A Text for Statisticians and Quantitative Scientists

F.J. Samaniego

Statistical Methods for Spatial Data Analysis

O. Schabenberger and C.A. Gotway

Bayesian Networks: With Examples in R

M. Scutari and J.-B. Denis

Large Sample Methods in Statistics

P.K. Sen and J. da Motta Singer

Decision Analysis: A Bayesian Approach

J.Q. Smith

Analysis of Failure and Survival Data

P.J. Smith

Applied Statistics: Handbook of GENSTAT Analyses

E.J. Snell and H. Simpson

Applied Nonparametric Statistical Methods, Fourth Edition

P. Sprent and N.C. Smeeton

Data Driven Statistical Methods

P. Sprent

Generalized Linear Mixed Models: Modern Concepts, Methods and Applications

W. W. Stroup

Survival Analysis Using S: Analysis of Time-to-Event Data

M. Tableman and J.S. Kim

Applied Categorical and Count Data Analysis

W. Tang, H. He, and X.M. Tu

Elementary Applications of Probability Theory, Second Edition

H.C. Tuckwell

Introduction to Statistical Inference and Its Applications with R

M.W. Trosset

Understanding Advanced Statistical Methods

P.H. Westfall and K.S.S. Henning

Statistical Process Control: Theory and Practice, Third Edition

G.B. Wetherill and D.W. Brown

Generalized Additive Models: An Introduction with R

S. Wood

Epidemiology: Study Design and Data Analysis, Third Edition

M. Woodward

Practical Data Analysis for Designed Experiments

B.S. Yandell

Texts in Statistical Science

Mathematical Statistics

**Basic Ideas and
Selected Topics**

Volume I

Second Edition

Peter J. Bickel

University of California

Berkeley, California, USA

Kjell A. Doksum

University of Wisconsin

Madison, Wisconsin, USA



CRC Press

Taylor & Francis Group

Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2015 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Version Date: 20150115

International Standard Book Number-13: 978-1-4987-2381-7 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

and the CRC Press Web site at
<http://www.crcpress.com>

To Erich L. Lehmann

This page intentionally left blank

CONTENTS

PREFACE TO THE 2015 EDITION

xxi

1	STATISTICAL MODELS, GOALS, AND PERFORMANCE CRITERIA	1
1.1	Data, Models, Parameters, and Statistics	1
1.1.1	Data and Models	1
1.1.2	Parametrizations and Parameters	6
1.1.3	Statistics as Functions on the Sample Space	8
1.1.4	Examples, Regression Models	9
1.2	Bayesian Models	12
1.3	The Decision Theoretic Framework	16
1.3.1	Components of the Decision Theory Framework	17
1.3.2	Comparison of Decision Procedures	24
1.3.3	Bayes and Minimax Criteria	26
1.4	Prediction	32
1.5	Sufficiency	41
1.6	Exponential Families	49
1.6.1	The One-Parameter Case	49
1.6.2	The Multiparameter Case	53
1.6.3	Building Exponential Families	56
1.6.4	Properties of Exponential Families	58
1.6.5	Conjugate Families of Prior Distributions	62
1.7	Problems and Complements	66
1.8	Notes	95
1.9	References	96

2	METHODS OF ESTIMATION	99
2.1	Basic Heuristics of Estimation	99
2.1.1	Minimum Contrast Estimates; Estimating Equations	99
2.1.2	The Plug-In and Extension Principles	102
2.2	Minimum Contrast Estimates and Estimating Equations	107
2.2.1	Least Squares and Weighted Least Squares	107
2.2.2	Maximum Likelihood	114
2.3	Maximum Likelihood in Multiparameter Exponential Families	121
2.4	Algorithmic Issues	127
2.4.1	The Method of Bisection	127
2.4.2	Coordinate Ascent	129
2.4.3	The Newton–Raphson Algorithm	131
2.4.4	The EM (Expectation/Maximization) Algorithm	133
2.5	Problems and Complements	138
2.6	Notes	158
2.7	References	159
3	MEASURES OF PERFORMANCE	161
3.1	Introduction	161
3.2	Bayes Procedures	161
3.3	Minimax Procedures	170
3.4	Unbiased Estimation and Risk Inequalities	176
3.4.1	Unbiased Estimation, Survey Sampling	176
3.4.2	The Information Inequality	179
3.5	Nondecision Theoretic Criteria	188
3.5.1	Computation	188
3.5.2	Interpretability	189
3.5.3	Robustness	190
3.6	Problems and Complements	197
3.7	Notes	210
3.8	References	211
4	TESTING AND CONFIDENCE REGIONS	213
4.1	Introduction	213
4.2	Choosing a Test Statistic: The Neyman-Pearson Lemma	223
4.3	Uniformly Most Powerful Tests and Monotone Likelihood Ratio Models	227
4.4	Confidence Bounds, Intervals, and Regions	233

4.5	The Duality Between Confidence Regions and Tests	241
4.6	Uniformly Most Accurate Confidence Bounds	248
4.7	Frequentist and Bayesian Formulations	251
4.8	Prediction Intervals	252
4.9	Likelihood Ratio Procedures	255
4.9.1	Introduction	255
4.9.2	Tests for the Mean of a Normal Distribution-Matched Pair Experiments	257
4.9.3	Tests and Confidence Intervals for the Difference in Means of Two Normal Populations	261
4.9.4	The Two-Sample Problem with Unequal Variances	264
4.9.5	Likelihood Ratio Procedures for Bivariate Normal Distributions	266
4.10	Problems and Complements	269
4.11	Notes	295
4.12	References	295
5	ASYMPTOTIC APPROXIMATIONS	297
5.1	Introduction: The Meaning and Uses of Asymptotics	297
5.2	Consistency	301
5.2.1	Plug-In Estimates and MLEs in Exponential Family Models	301
5.2.2	Consistency of Minimum Contrast Estimates	304
5.3	First- and Higher-Order Asymptotics: The Delta Method with Applications	306
5.3.1	The Delta Method for Moments	306
5.3.2	The Delta Method for In Law Approximations	311
5.3.3	Asymptotic Normality of the Maximum Likelihood Estimate in Exponential Families	322
5.4	Asymptotic Theory in One Dimension	324
5.4.1	Estimation: The Multinomial Case	324
5.4.2	Asymptotic Normality of Minimum Contrast and M -Estimates	327
5.4.3	Asymptotic Normality and Efficiency of the MLE	331
5.4.4	Testing	332
5.4.5	Confidence Bounds	336
5.5	Asymptotic Behavior and Optimality of the Posterior Distribution	337
5.6	Problems and Complements	345
5.7	Notes	362
5.8	References	363

6	INFERENCE IN THE MULTIPARAMETER CASE	365
6.1	Inference for Gaussian Linear Models	365
6.1.1	The Classical Gaussian Linear Model	366
6.1.2	Estimation	369
6.1.3	Tests and Confidence Intervals	374
6.2	Asymptotic Estimation Theory in p Dimensions	383
6.2.1	Estimating Equations	384
6.2.2	Asymptotic Normality and Efficiency of the MLE	386
6.2.3	The Posterior Distribution in the Multiparameter Case	391
6.3	Large Sample Tests and Confidence Regions	392
6.3.1	Asymptotic Approximation to the Distribution of the Likelihood Ratio Statistic	392
6.3.2	Wald's and Rao's Large Sample Tests	398
6.4	Large Sample Methods for Discrete Data	400
6.4.1	Goodness-of-Fit in a Multinomial Model. Pearson's χ^2 Test	401
6.4.2	Goodness-of-Fit to Composite Multinomial Models. Contingency Tables	403
6.4.3	Logistic Regression for Binary Responses	408
6.5	Generalized Linear Models	411
6.6	Robustness Properties and Semiparametric Models	417
6.7	Problems and Complements	422
6.8	Notes	438
6.9	References	438
A	A REVIEW OF BASIC PROBABILITY THEORY	441
A.1	The Basic Model	441
A.2	Elementary Properties of Probability Models	443
A.3	Discrete Probability Models	443
A.4	Conditional Probability and Independence	444
A.5	Compound Experiments	446
A.6	Bernoulli and Multinomial Trials, Sampling With and Without Replacement	447
A.7	Probabilities on Euclidean Space	448
A.8	Random Variables and Vectors: Transformations	451
A.9	Independence of Random Variables and Vectors	453
A.10	The Expectation of a Random Variable	454
A.11	Moments	456
A.12	Moment and Cumulant Generating Functions	459

A.13 Some Classical Discrete and Continuous Distributions	460
A.14 Modes of Convergence of Random Variables and Limit Theorems	466
A.15 Further Limit Theorems and Inequalities	468
A.16 Poisson Process	472
A.17 Notes	474
A.18 References	475
B ADDITIONAL TOPICS IN PROBABILITY AND ANALYSIS	477
B.1 Conditioning by a Random Variable or Vector	477
B.1.1 The Discrete Case	477
B.1.2 Conditional Expectation for Discrete Variables	479
B.1.3 Properties of Conditional Expected Values	480
B.1.4 Continuous Variables	482
B.1.5 Comments on the General Case	484
B.2 Distribution Theory for Transformations of Random Vectors	485
B.2.1 The Basic Framework	485
B.2.2 The Gamma and Beta Distributions	488
B.3 Distribution Theory for Samples from a Normal Population	491
B.3.1 The χ^2 , F , and t Distributions	491
B.3.2 Orthogonal Transformations	494
B.4 The Bivariate Normal Distribution	497
B.5 Moments of Random Vectors and Matrices	502
B.5.1 Basic Properties of Expectations	502
B.5.2 Properties of Variance	503
B.6 The Multivariate Normal Distribution	506
B.6.1 Definition and Density	506
B.6.2 Basic Properties. Conditional Distributions	508
B.7 Convergence for Random Vectors: O_P and o_P Notation	511
B.8 Multivariate Calculus	516
B.9 Convexity and Inequalities	518
B.10 Topics in Matrix Theory and Elementary Hilbert Space Theory	519
B.10.1 Symmetric Matrices	519
B.10.2 Order on Symmetric Matrices	520
B.10.3 Elementary Hilbert Space Theory	521
B.11 Problems and Complements	524
B.12 Notes	538
B.13 References	539

C TABLES	541
Table I The Standard Normal Distribution	542
Table I' Auxilliary Table of the Standard Normal Distribution	543
Table II t Distribution Critical Values	544
Table III χ^2 Distribution Critical Values	545
Table IV F Distribution Critical Values	546
INDEX	547

VOLUME II CONTENTS

I	INTRODUCTION TO VOLUME II	1
I.1	Tests of Goodness of Fit and the Brownian Bridge	5
I.2	Testing Goodness of Fit to Parametric Hypotheses	5
I.3	Regular Parameters. Minimum Distance Estimates	6
I.4	Permutation Tests	8
I.5	Estimation of Irregular Parameters	8
I.6	Stein and Empirical Bayes Estimation	10
I.7	Model Selection	11
I.8	Problems and Complements	15
I.9	Notes	20
7	TOOLS FOR ASYMPTOTIC ANALYSIS	21
7.1	Weak Convergence in Function Spaces	21
7.1.1	Stochastic Processes and Weak Convergence	21
7.1.2	Maximal Inequalities	28
7.1.3	Empirical Processes on Function Spaces	31
7.2	The Delta Method in Infinite Dimensional Space	38
7.2.1	Influence Functions and the Gâteaux Derivative	38
7.2.2	The Quantile Process	47
7.3	Further Expansions	51
7.3.1	The von Mises Expansion	51
7.3.2	The Hoeffding/Analysis of Variance Expansion	54
7.4	Problems and Complements	62
7.5	Notes	71

8	DISTRIBUTION-FREE, UNBIASED AND EQUIVARIANT PROCEDURES	72
8.1	Introduction	72
8.2	Similarity and Completeness	73
8.2.1	Testing	73
8.2.2	Testing Optimality Theory	83
8.2.3	Estimation	86
8.3	Invariance, Equivariance and Minimax Procedures	91
8.3.1	Group Models	91
8.3.2	Group Models and Decision Theory	93
8.3.3	Characterizing Invariant Tests	95
8.3.4	Characterizing Equivariant Estimates	101
8.3.5	Minimaxity for Tests: Application to Group Models	102
8.3.6	Minimax Estimation, Admissibility, and Steinian Shrinkage	106
8.4	Problems and Complements	111
8.5	Notes	122
9	INFERENCE IN SEMIPARAMETRIC MODELS	123
9.1	ESTIMATION IN SEMIPARAMETRIC MODELS	123
9.1.1	Selected Examples	123
9.1.2	Regularization. Modified Maximum Likelihood	131
9.1.3	Other Modified and Approximate Likelihoods	140
9.1.4	Sieves and Regularization	143
9.2	Asymptotics. Consistency and Asymptotic Normality	149
9.2.1	A General Consistency Criterion	149
9.2.2	Asymptotics for Selected Models	151
9.3	Efficiency in Semiparametric Models	159
9.4	Tests and Empirical Process Theory	172
9.5	Asymptotic Properties of Likelihoods. Contiguity	177
9.6	Problems and Complements	189
9.7	Notes	205
10	MONTE CARLO METHODS	207
10.1	The Nature of Monte Carlo Methods	207
10.2	Three Basic Monte Carlo Methods	210
10.2.1	Simple Monte Carlo	211
10.2.2	Importance Sampling	212
10.2.3	Rejective Sampling	213

10.3	The Bootstrap	215
10.3.1	Bootstrap Samples and Bias Corrections	216
10.3.2	Bootstrap Variance and Confidence Bounds	220
10.3.3	The General i.i.d. Nonparametric Bootstrap	222
10.3.4	Asymptotic Theory for the Bootstrap	225
10.3.5	Examples where Efron's Bootstrap Fails. The m out of n Bootstraps	230
10.4	Markov Chain Monte Carlo	232
10.4.1	The Basic MCMC Framework	232
10.4.2	Metropolis Sampling Algorithms	233
10.4.3	The Gibbs Samplers	237
10.4.4	Speed of Convergence of MCMC	241
10.5	Applications of MCMC to Bayesian and Frequentist Inference	243
10.6	Problems and Complements	250
10.7	Notes	256
11	NONPARAMETRIC INFERENCE FOR FUNCTIONS OF ONE VARIABLE	257
11.1	Introduction	257
11.2	Convolution Kernel Estimates on R	258
11.2.1	Uniform Local Behavior of Kernel Density Estimates	261
11.2.2	Global Behavior of Convolution Kernel Estimates	263
11.2.3	Performance and Bandwidth Choice	264
11.2.4	Discussion of convolution kernel estimates	265
11.3	Minimum Contrast Estimates: Reducing Boundary Bias	266
11.4	Regularization and Nonlinear Density Estimates	272
11.4.1	Regularization and Roughness Penalties	272
11.4.2	Sieves. Machine Learning. Log Density Estimation	273
11.4.3	Nearest Neighbour Density Estimates	276
11.5	Confidence Regions	277
11.6	Nonparametric Regression for one Covariate	279
11.6.1	Estimation Principles	279
11.6.2	Asymptotic Bias and Variance Calculations	282
11.7	Problems and Complements	289
12	PREDICTION AND MACHINE LEARNING	299
12.1	Introduction	299
12.1.1	Statistical Approaches to Modeling and Analyzing Multidimensional data. Sieves	301

12.1.2	Machine Learning Approaches	305
12.1.3	Outline	307
12.2	Classification and Prediction	307
12.2.1	Multivariate Density and Regression Estimation	307
12.2.2	Bayes Rule and Nonparametric Classification	312
12.2.3	Sieve Methods	314
12.2.4	Machine Learning Approaches	316
12.3	Asymptotics	324
12.3.1	Optimal Prediction in Parametric Regression Models	326
12.3.2	Optimal Rates of Convergence for Estimation and Prediction in Nonparametric Models	329
12.3.3	The Gaussian White Noise (GWN) Model	338
12.3.4	Minmax Bounds on IMSE for Subsets of the GWN Model	340
12.3.5	Sparse Submodels	342
12.4	Oracle Inequalities	344
12.4.1	Stein's Unbiased Risk Estimate	346
12.4.2	Oracle Inequality for Shrinkage Estimators	347
12.4.3	Oracle Inequality and Adaptive Minimax Rate for Truncated Estimates	348
12.4.4	An Oracle Inequality for Classification	350
12.5	Performance and Tuning via Cross Validation	353
12.5.1	Cross Validation for Tuning Parameter Choice	354
12.5.2	Cross Validation for Measuring Performance	358
12.6	Model Selection and Dimension Reduction	359
12.6.1	A Bayesian Criterion for Model Selection	360
12.6.2	Inference after Model Selection	364
12.6.3	Dimension Reduction via Principal Component Analysis	366
12.7	Topics Untouched and Current Frontiers	367
12.8	Problems and Complements	371
D	APPENDIX D. SUPPLEMENTS TO TEXT	385
D.1	Probability Results	385
D.2	Supplements to Section 7.1	387
D.3	Supplement to Section 7.2	390
D.4	Supplement to Section 9.2.2	391
D.5	Supplement to Section 10.4	392
D.6	Supplement to Section 11.6	397

D.7 Supplement to Section 12.2.2	399
D.8 Problems and Complements	405
E SOLUTIONS FOR VOL. II	410
REFERENCES	423
SUBJECT INDEX	438

This page intentionally left blank

PREFACE TO THE 2015 EDITION

In this preface, we start with an overview of developments in statistics since the first (1977) edition, then give separate overviews of Volumes I and II of the second edition.

In the last 40 some years statistics has changed enormously under the impact of several forces:

- (1) The generation of what were once unusual types of data such as images, trees (phylogenetic and other), and other types of combinatorial objects.
- (2) The generation of enormous amounts of data—terabytes (the equivalent of 10^{12} characters) for an astronomical survey over three years.
- (3) The possibility of implementing computations of a magnitude that would have once been unthinkable.

The underlying sources of these changes have been the exponential change in computing speed (Moore's "law") and the development of devices (computer controlled) using novel instruments and scientific techniques (e.g., NMR tomography, gene sequencing). These techniques often have a strong intrinsic computational component. Tomographic data are the result of mathematically based processing. Sequencing is done by applying computational algorithms to raw gel electrophoresis data.

As a consequence the emphasis of statistical theory has shifted away from small sample optimality results in a number of directions:

- (1) Methods for inference based on larger numbers of observations and minimal assumptions—asymptotic methods in non- and semiparametric models, models with "infinite" number of parameters.
- (2) The construction of models for time series, temporal spatial series, and other complex data structures using sophisticated probability modeling but again relying for analytical results on asymptotic approximation. Multiparameter models are the rule.
- (3) The use of methods of inference involving simulation as a key element such as the bootstrap and Markov Chain Monte Carlo.

- (4) The development of techniques not describable in “closed mathematical form” but rather through elaborate algorithms for which problems of existence of solutions are important and far from obvious.
- (5) The study of the interplay between numerical and statistical considerations. Despite advances in computing speed, some methods run quickly in real time. Others do not and some though theoretically attractive cannot be implemented in a human lifetime.
- (6) The study of the interplay between the number of observations and the number of parameters of a model and the beginnings of appropriate asymptotic theories.

There have been other important consequences such as the extensive development of graphical and other exploratory methods for which theoretical development and connection with mathematics have been minimal. These will not be dealt with in our work.

In this edition we pursue our philosophy of describing the basic concepts of mathematical statistics relating theory to practice.

Volume I

This volume presents the basic classical statistical concepts at the Ph.D. level without requiring measure theory. It gives careful proofs of the major results and indicates how the theory sheds light on the properties of practical methods. The topics include estimation, prediction, testing, confidence sets, Bayesian analysis and the more general approach of decision theory.

We include from the start in Chapter 1 non- and semiparametric models, then go to parameters and parametric models stressing the role of identifiability. From the beginning we stress function-valued parameters, such as the density, and function-valued statistics, such as the empirical distribution function. We also, from the start, include examples that are important in applications, such as regression experiments. There is extensive material on Bayesian models and analysis and extended discussion of prediction and k -parameter exponential families. These objects that are the building blocks of most modern models require concepts involving moments of random vectors and convexity that are given in Appendix B.

Chapter 2 deals with estimation and includes a detailed treatment of maximum likelihood estimates (MLEs), including a complete study of MLEs in canonical k -parameter exponential families. Other novel features of this chapter include a detailed analysis, including proofs of convergence, of a standard but slow algorithm (coordinate descent) for convex optimization, applied, in particular to computing MLEs in multiparameter exponential families. We also give an introduction to the EM algorithm, one of the main ingredients of most modern algorithms for inference. Chapters 3 and 4 are on the theory of testing and confidence regions, including some optimality theory for estimation as well and elementary robustness considerations.

Chapter 5 is devoted to basic asymptotic approximations with one dimensional parameter models as examples. It includes proofs of consistency and asymptotic normality and optimality of maximum likelihood procedures in inference and a section relating Bayesian and frequentist inference via the Bernstein–von Mises theorem.

Finally, Chapter 6 is devoted to inference in multivariate (multiparameter) models. Included are asymptotic normality and optimality of maximum likelihood estimates, inference in the general linear model, Wilks theorem on the asymptotic distribution of the likelihood ratio test, the Wald and Rao statistics and associated confidence regions, and some parallels to the optimality theory and comparisons of Bayes and frequentist procedures given in the one dimensional parameter case in Chapter 5. Chapter 6 also develops the asymptotic joint normality of estimates that are solutions to estimating equations and presents Huber's Sandwich formula for the asymptotic covariance matrix of such estimates. Generalized linear models, including binary logistic regression, are introduced as examples. Robustness from an asymptotic theory point of view appears also. This chapter uses multivariate calculus in an intrinsic way and can be viewed as an essential prerequisite for the more advanced topics of Volume II.

Volume I includes Appendix A on basic probability and a larger Appendix B, which includes more advanced topics from probability theory such as the multivariate Gaussian distribution, weak convergence in Euclidean spaces, and probability inequalities as well as more advanced topics in matrix theory and analysis. The latter include the principal axis and spectral theorems for Euclidean space and the elementary theory of convex functions on R^d as well as an elementary introduction to Hilbert space theory. As in the first edition, we do not require measure theory but assume from the start that our models are what we call "regular." That is, we assume either a discrete probability whose support does not depend on the parameter set, or the absolutely continuous case with a density. Hilbert space theory is not needed, but for those who know this topic Appendix B points out interesting connections to prediction and linear regression analysis.

Appendix B is as self-contained as possible with proofs of most statements, problems, and references to the literature for proofs of the deepest results such as the spectral theorem. The reason for these additions are the changes in subject matter necessitated by the current areas of importance in the field.

For the first volume of the second edition we would like to add thanks to Jiangning Fan, Michael Jordan, Jianhua Huang, Ying Qing Chen, and Carl Spruill and the many students who were guinea pigs in the basic theory course at Berkeley. We also thank Faye Yeager for typing, Michael Ostland and Simon Cawley for producing the graphs, Yoram Gat for proofreading that found not only typos but serious errors, and Prentice Hall for generous production support.

Volume II

Volume II of the second edition will be forthcoming in 2015. It presents what we think are some of the most important statistical concepts, methods, and tools developed since the first edition. Topics to be included are: asymptotic efficiency in semiparametric models, semiparametric maximum likelihood estimation, survival analysis including Cox regression, classification, methods of inference based on sieve models, model selection, Monte Carlo methods such as the bootstrap and Markov Chain Monte Carlo, nonparametric curve estimation, and machine learning including support vector machines and classification and regression trees (CART).

The basic asymptotic tools that will be developed or presented, in part in the text and, in part in appendices, are weak convergence for random processes, elementary empirical process theory, and the functional delta method.

With the tools and concepts developed in this second volume students will be ready for advanced research in modern statistics.

We thank Akichika Ozeki and Sören Künzel for pointing out errors and John Kimmel and CRC Press for production support. We also thank Dee Frana and especially Anne Chong who typed 90% of Volume II for word processing.

Last and most important we would like to thank our wives, Nancy Kramer Bickel and Joan H. Fujimura, and our families for support, encouragement, and active participation in an enterprise that at times seemed endless, appeared gratifyingly ended in 1976 but has, with the field, taken on a new life.

Peter J. Bickel
bickel@stat.berkeley.edu
Kjell Doksum
doksum@stat.wisc.edu

STATISTICAL MODELS, GOALS, AND PERFORMANCE CRITERIA

1.1 DATA, MODELS, PARAMETERS AND STATISTICS

1.1.1 Data and Models

Most studies and experiments, scientific or industrial, large scale or small, produce data whose analysis is the ultimate object of the endeavor.

Data can consist of:

- (1) Vectors of scalars, measurements, and/or characters, for example, a single time series of measurements.
- (2) Matrices of scalars and/or characters, for example, digitized pictures or more routinely measurements of covariates and response on a set of n individuals—see Example 1.1.4 and Sections 2.2.1 and 6.1.
- (3) Arrays of scalars and/or characters as in contingency tables—see Chapter 6—or more generally multifactor multiresponse data on a number of individuals.
- (4) All of the above and more, in particular, functions as in signal processing, trees as in evolutionary phylogenies, and so on.

The goals of science and society, which statisticians share, are to draw useful information from data using everything that we know. The particular angle of mathematical statistics is to view data as the outcome of a random experiment that we model mathematically.

A detailed discussion of the appropriateness of the models we shall discuss in particular situations is beyond the scope of this book, but we will introduce general model diagnostic tools in Volume 2, Chapter 1. Moreover, we shall parenthetically discuss features of the sources of data that can make apparently suitable models grossly misleading. A generic source of trouble often called *gross errors* is discussed in greater detail in the section on robustness (Section 3.5.3). In any case all our models are generic and, as usual, “The Devil is in the details!” All the principles we discuss and calculations we perform should only be suggestive guides in successful applications of statistical analysis in science and policy. Subject matter specialists usually have to be principal guides in model formulation. A

priori, in the words of George Box (1979), “Models of course, are never true but fortunately it is only necessary that they be useful.”

In this book we will study how, starting with tentative models:

(1) We can conceptualize the data structure and our goals more precisely. We begin this in the simple examples that follow and continue in Sections 1.2–1.5 and throughout the book.

(2) We can derive methods of extracting useful information from data and, in particular, give methods that assess the generalizability of experimental results. For instance, if we observe an effect in our data, to what extent can we expect the same effect more generally? Estimation, testing, confidence regions, and more general procedures will be discussed in Chapters 2–4.

(3) We can assess the effectiveness of the methods we propose. We begin this discussion with decision theory in Section 1.3 and continue with optimality principles in Chapters 3 and 4.

(4) We can decide if the models we propose are approximations to the mechanism generating the data adequate for our purposes. Goodness of fit tests, robustness, and diagnostics are discussed in Volume 2, Chapter 1.

(5) We can be guided to alternative or more general descriptions that might fit better. Hierarchies of models are discussed throughout.

Here are some examples:

(a) We are faced with a population of N elements, for instance, a shipment of manufactured items. An unknown number $N\theta$ of these elements are defective. It is too expensive to examine all of the items. So to get information about θ , a sample of n is drawn without replacement and inspected. The data gathered are the number of defectives found in the sample.

(b) We want to study how a physical or economic feature, for example, height or income, is distributed in a large population. An exhaustive census is impossible so the study is based on measurements and a sample of n individuals drawn at random from the population. The population is so large that, for modeling purposes, we approximate the actual process of sampling without replacement by sampling with replacement.

(c) An experimenter makes n independent determinations of the value of a physical constant μ . His or her measurements are subject to random fluctuations (error) and the data can be thought of as μ plus some random errors.

(d) We want to compare the efficacy of two ways of doing something under similar conditions such as brewing coffee, reducing pollution, treating a disease, producing energy, learning a maze, and so on. This can be thought of as a problem of comparing the efficacy of two methods applied to the members of a certain population. We run $m + n$ independent experiments as follows: $m + n$ members of the population are picked at random and m of these are assigned to the first method and the remaining n are assigned to the second method. In this manner, we obtain one or more quantitative or qualitative measures of efficacy from each experiment. For instance, we can assign two drugs, A to m , and B to n , randomly selected patients and then measure temperature and blood pressure, have the patients rated qualitatively for improvement by physicians, and so on. Random variability

here would come primarily from differing responses among patients to the same drug but also from error in the measurements and variation in the purity of the drugs.

We shall use these examples to arrive at our formulation of statistical models and to indicate some of the difficulties of constructing such models. First consider situation (a), which we refer to as:

Example 1.1.1. Sampling Inspection. The mathematical model suggested by the description is well defined. A random experiment has been performed. The sample space consists of the numbers $0, 1, \dots, n$ corresponding to the number of defective items found. On this space we can define a random variable X given by $X(k) = k, k = 0, 1, \dots, n$. If $N\theta$ is the number of defective items in the population sampled, then by (A.13.6)

$$P[X = k] = \frac{\binom{N\theta}{k} \binom{N - N\theta}{n - k}}{\binom{N}{n}} \quad (1.1.1)$$

$$\text{if } \max(n - N(1 - \theta), 0) \leq k \leq \min(N\theta, n).$$

Thus, X has an hypergeometric, $\mathcal{H}(N\theta, N, n)$ distribution.

The main difference that our model exhibits from the usual probability model is that $N\theta$ is *unknown* and, in principle, can take on any value between 0 and N . So, although the sample space is well defined, we cannot specify the probability structure completely but rather only give a family $\{\mathcal{H}(N\theta, N, n)\}$ of probability distributions for X , any one of which could have generated the data actually observed. \square

Example 1.1.2. Sample from a Population. One-Sample Models. Situation (b) can be thought of as a generalization of (a) in that a quantitative measure is taken rather than simply recording “defective” or not. It can also be thought of as a limiting case in which $N = \infty$, so that sampling with replacement replaces sampling without. Formally, if the measurements are scalar, we observe x_1, \dots, x_n , which are modeled as realizations of X_1, \dots, X_n independent, identically distributed (i.i.d.) random variables with common unknown distribution function F . We often refer to such X_1, \dots, X_n as a *random sample* from F , and also write that X_1, \dots, X_n are i.i.d. as X with $X \sim F$, where “ \sim ” stands for “is distributed as.” The model is fully described by the set \mathcal{F} of distributions that we specify. The same model also arises naturally in situation (c). Here we can write the n determinations of μ as

$$X_i = \mu + \epsilon_i, \quad 1 \leq i \leq n \quad (1.1.2)$$

where $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$ is the vector of random errors. What should we assume about the distribution of ϵ , which together with μ completely specifies the joint distribution of X_1, \dots, X_n ? Of course, that depends on how the experiment is carried out. Given the description in (c), we postulate

(1) The value of the error committed on one determination does not affect the value of the error at other times. That is, $\epsilon_1, \dots, \epsilon_n$ are independent.

(2) The distribution of the error at one determination is the same as that at another. Thus, $\epsilon_1, \dots, \epsilon_n$ are identically distributed.

(3) The distribution of ϵ is independent of μ .

Equivalently X_1, \dots, X_n are a random sample and, if we let G be the distribution function of ϵ_1 and F that of X_1 , then

$$F(x) = G(x - \mu) \quad (1.1.3)$$

and the model is alternatively specified by \mathcal{F} , the set of F 's we postulate, or by $\{(\mu, G) : \mu \in R, G \in \mathcal{G}\}$ where \mathcal{G} is the set of all allowable error distributions that we postulate. Commonly considered \mathcal{G} 's are all distributions with center of symmetry 0, or alternatively all distributions with expectation 0. The classical default model is:

(4) The common distribution of the errors is $\mathcal{N}(0, \sigma^2)$, where σ^2 is unknown. That is, the X_i are a sample from a $\mathcal{N}(\mu, \sigma^2)$ population or equivalently $\mathcal{F} = \{\Phi(\frac{\cdot - \mu}{\sigma}) : \mu \in R, \sigma > 0\}$ where Φ is the standard normal distribution. \square

This default model is also frequently postulated for measurements taken on units obtained by random sampling from populations, for instance, heights of individuals or log incomes. It is important to remember that these are assumptions at best only approximately valid. All actual measurements are discrete rather than continuous. There are absolute bounds on most quantities—100 ft high men are impossible. Heights are always nonnegative. The Gaussian distribution, whatever be μ and σ , will have none of this.

Now consider situation (d).

Example 1.1.3. Two-Sample Models. Let $x_1, \dots, x_m; y_1, \dots, y_n$, respectively, be the responses of m subjects having a given disease given drug A and n other similarly diseased subjects given drug B . By convention, if drug A is a standard or placebo, we refer to the x 's as *control observations*. A placebo is a substance such as water that is expected to have no effect on the disease and is used to correct for the well-documented placebo effect, that is, patients improve even if they only think they are being treated. We let the y 's denote the responses of subjects given a new drug or treatment that is being evaluated by comparing its effect with that of the placebo. We call the y 's *treatment observations*.

Natural initial assumptions here are:

(1) The x 's and y 's are realizations of X_1, \dots, X_m a sample from F , and Y_1, \dots, Y_n a sample from G , so that the model is specified by the set of possible (F, G) pairs.

To specify this set more closely the critical *constant treatment* effect assumption is often made.

(2) Suppose that if treatment A had been administered to a subject response x would have been obtained. Then if treatment B had been administered to the same subject instead of treatment A , response $y = x + \Delta$ would be obtained where Δ does not depend on x . This implies that if F is the distribution of a control, then $G(\cdot) = F(\cdot - \Delta)$. We call this the *shift model* with parameter Δ .

Often the final simplification is made.

(3) The control responses are normally distributed. Then if F is the $\mathcal{N}(\mu, \sigma^2)$ distribution and G is the $\mathcal{N}(\mu + \Delta, \sigma^2)$ distribution, we have specified the Gaussian two sample model with equal variances. \square

How do we settle on a set of assumptions? Evidently by a mixture of experience and physical considerations. The advantage of piling on assumptions such as (1)–(4) of Example 1.1.2 is that, if they are true, we know how to combine our measurements to estimate μ in a highly efficient way and also assess the accuracy of our estimation procedure (Example 4.4.1). The danger is that, if they are false, our analyses, though correct for the model written down, may be quite irrelevant to the experiment that was actually performed. As our examples suggest, there is tremendous variation in the degree of knowledge and control we have concerning experiments.

In some applications we often have a tested theoretical model and the danger is small. The number of defectives in the first example clearly has a hypergeometric distribution; the number of α particles emitted by a radioactive substance in a small length of time is well known to be approximately Poisson distributed.

In others, we can be reasonably secure about some aspects, but not others. For instance, in Example 1.1.2, we can ensure independence and identical distribution of the observations by using different, equally trained observers with no knowledge of each other's findings. However, we have little control over what kind of distribution of errors we get and will need to investigate the properties of methods derived from specific error distribution assumptions when these assumptions are violated. This will be done in Sections 3.5.3 and 6.6.

Experiments in medicine and the social sciences often pose particular difficulties. For instance, in comparative experiments such as those of Example 1.1.3 the group of patients to whom drugs A and B are to be administered may be haphazard rather than a random sample from the population of sufferers from a disease. In this situation (and generally) it is important to *randomize*. That is, we use a random number table or other random mechanism so that the m patients administered drug A are a sample without replacement from the set of $m + n$ available patients. Without this device we could not know whether observed differences in drug performance might not (possibly) be due to unconscious bias on the part of the experimenter. All the severely ill patients might, for instance, have been assigned to B . The study of the model based on the minimal assumption of randomization is complicated and further conceptual issues arise. Fortunately, the methods needed for its analysis are much the same as those appropriate for the situation of Example 1.1.3 when F, G are assumed arbitrary. Statistical methods for models of this kind are given in Volume 2.

Using our first three examples for illustrative purposes, we now define the elements of a statistical model. A review of necessary concepts and notation from probability theory are given in the appendices.

We are given a random experiment with sample space Ω . On this sample space we have defined a random vector $\mathbf{X} = (X_1, \dots, X_n)$. When ω is the outcome of the experiment, $\mathbf{X}(\omega)$ is referred to as the *observations* or *data*. It is often convenient to identify the random vector \mathbf{X} with its realization, the data $\mathbf{X}(\omega)$. Since it is only \mathbf{X} that we observe, we need only consider its probability distribution. This distribution is assumed to be a member of a family \mathcal{P} of probability distributions on R^n . \mathcal{P} is referred to as the *model*. For instance, in Example 1.1.1, we observe X and the family \mathcal{P} is that of all hypergeometric distributions with sample size n and population size N . In Example 1.1.2, if (1)–(4) hold, \mathcal{P} is the

family of all distributions according to which X_1, \dots, X_n are independent and identically distributed with a common $\mathcal{N}(\mu, \sigma^2)$ distribution.

1.1.2 Parametrizations and Parameters

To describe \mathcal{P} we use a *parametrization*, that is, a map, $\theta \rightarrow P_\theta$ from a space of labels, the parameter space Θ , to \mathcal{P} ; or equivalently write $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Thus, in Example 1.1.1 we take θ to be the fraction of defectives in the shipment, $\Theta = \{0, \frac{1}{N}, \dots, 1\}$ and P_θ the $\mathcal{H}(N\theta, N, n)$ distribution. In Example 1.1.2 with assumptions (1)–(4) we have implicitly taken $\Theta = R \times R^+$ and, if $\theta = (\mu, \sigma^2)$, P_θ the distribution on R^n with density $\prod_{i=1}^n \frac{1}{\sigma} \varphi\left(\frac{x_i - \mu}{\sigma}\right)$ where φ is the standard normal density. If, still in this example, we know we are measuring a positive quantity in this model, we have $\Theta = R^+ \times R^+$. If, on the other hand, we only wish to make assumptions (1)–(3) with ϵ having expectation 0, we can take $\Theta = \{(\mu, G) : \mu \in R, G \text{ with density } g \text{ such that } \int xg(x)dx = 0\}$ and $P_{(\mu, G)}$ has density $\prod_{i=1}^n g(x_i - \mu)$.

When we can take Θ to be a nice subset of Euclidean space and the maps $\theta \rightarrow P_\theta$ are smooth, in senses to be made precise later, models \mathcal{P} are called *parametric*. Models such as that of Example 1.1.2 with assumptions (1)–(3) are called *semiparametric*. Finally, models such as that of Example 1.1.3 with only (1) holding and F, G taken to be arbitrary are called *nonparametric*. It's important to note that even nonparametric models make substantial assumptions—in Example 1.1.3 that X_1, \dots, X_m are independent of each other and Y_1, \dots, Y_n ; moreover, X_1, \dots, X_m are identically distributed as are Y_1, \dots, Y_n . The only truly nonparametric but useless model for $\mathbf{X} \in R^n$ is to assume that its (joint) distribution can be anything.

Note that there are many ways of choosing a parametrization in these and all other problems. We may take any one-to-one function of θ as a new parameter. For instance, in Example 1.1.1 we can use the number of defectives in the population, $N\theta$, as a parameter and in Example 1.1.2, under assumptions (1)–(4), we may parametrize the model by the first and second moments of the normal distribution of the observations (i.e., by $(\mu, \mu^2 + \sigma^2)$).

What parametrization we choose is usually suggested by the phenomenon we are modeling; θ is the fraction of defectives, μ is the unknown constant being measured. However, as we shall see later, the first parametrization we arrive at is not necessarily the one leading to the simplest analysis. Of even greater concern is the possibility that the parametrization is not one-to-one, that is, such that we can have $\theta_1 \neq \theta_2$ and yet $P_{\theta_1} = P_{\theta_2}$. Such parametrizations are called *unidentifiable*. For instance, in (1.1.2) suppose that we permit G to be arbitrary. Then the map sending $\theta = (\mu, G)$ into the distribution of (X_1, \dots, X_n) remains the same but $\Theta = \{(\mu, G) : \mu \in R, G \text{ has (arbitrary) density } g\}$. Now the parametrization is unidentifiable because, for example, $\mu = 0$ and $\mathcal{N}(0, 1)$ errors lead to the same distribution of the observations as $\mu = 1$ and $\mathcal{N}(-1, 1)$ errors. The critical problem with such parametrizations is that even with “infinite amounts of data,” that is, knowledge of the true P_θ , parts of θ remain unknowable. Thus, we will need to ensure that our parametrizations are *identifiable*, that is, $\theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} \neq P_{\theta_2}$.

Dual to the notion of a parametrization, a map from some Θ to \mathcal{P} , is that of a *parameter*, formally a map, ν , from \mathcal{P} to another space \mathcal{N} . A parameter is a feature $\nu(P)$ of the distribution of X . For instance, in Example 1.1.1, the fraction of defectives θ can be thought of as the mean of X/n . In Example 1.1.3 with assumptions (1)–(2) we are interested in Δ , which can be thought of as the difference in the means of the two populations of responses. In addition to the parameters of interest, there are also usually *nuisance parameters*, which correspond to other unknown features of the distribution of \mathbf{X} . For instance, in Example 1.1.2, if the errors are normally distributed with unknown variance σ^2 , then σ^2 is a nuisance parameter. We usually try to combine parameters of interest and nuisance parameters into a single grand parameter θ , which indexes the family \mathcal{P} , that is, make $\theta \rightarrow P_\theta$ into a parametrization of \mathcal{P} . Implicit in this description is the assumption that θ is a parameter in the sense we have just defined. But given a parametrization $\theta \rightarrow P_\theta$, θ is a parameter if and only if the parametrization is identifiable. Formally, we can define the (well defined) parameter $\theta : \mathcal{P} \rightarrow \Theta$ as the inverse of the map $\theta \rightarrow P_\theta$, from Θ to its range \mathcal{P} iff the latter map is 1-1, that is, if $P_{\theta_1} = P_{\theta_2}$ implies $\theta_1 = \theta_2$. Note that $\theta(P_\theta) = \theta$.

More generally, a function $q : \Theta \rightarrow \mathcal{N}$ can be identified with a parameter $\nu(P)$ iff $P_{\theta_1} = P_{\theta_2}$ implies $q(\theta_1) = q(\theta_2)$ and then $\nu(P_\theta) \equiv q(\theta)$.

Here are two points to note:

(1) A parameter can have many representations. For instance, in Example 1.1.2 with assumptions (1)–(4) the parameter of interest $\mu \equiv \mu(P)$ can be characterized as the mean of P , or the median of P , or the midpoint of the interquantile range of P , or more generally as the center of symmetry of P , as long as \mathcal{P} is the set of all Gaussian distributions.

(2) A vector parametrization that is unidentifiable may still have components that are parameters (identifiable). For instance, consider Example 1.1.2 again in which we assume the error ϵ to be Gaussian but with arbitrary mean Δ . Then P is parametrized by $\theta = (\mu, \Delta, \sigma^2)$, where σ^2 is the variance of ϵ . As we have seen this parametrization is unidentifiable and neither μ nor Δ are parameters in the sense we've defined. But $\sigma^2 = \text{Var}(X_1)$ evidently is and so is $\mu + \Delta$.

Sometimes the choice of \mathcal{P} starts by the consideration of a particular parameter. For instance, our interest in studying a population of incomes may precisely be in the mean income. When we sample, say with replacement, and observe X_1, \dots, X_n independent with common distribution, it is natural to write

$$X_i = \mu + \epsilon_i, \quad 1 \leq i \leq n$$

where μ denotes the mean income and, thus, $E(\epsilon_i) = 0$. The (μ, G) parametrization of Example 1.1.2 is now well defined and identifiable by (1.1.3) and $\mathcal{G} = \{G : \int x dG(x) = 0\}$.

Similarly, in Example 1.1.3, instead of postulating a constant treatment effect Δ , we can start by making the difference of the means, $\delta = \mu_Y - \mu_X$, the focus of the study. Then δ is identifiable whenever μ_X and μ_Y exist.

1.1.3 Statistics as Functions on the Sample Space

Models and parametrizations are creations of the statistician, but the true values of parameters are secrets of nature. Our aim is to use the data inductively, to narrow down in useful ways our ideas of what the “true” P is. The link for us are things we can compute, statistics. Formally, a *statistic* T is a map from the sample space \mathcal{X} to some space of values \mathcal{T} , usually a Euclidean space. Informally, $T(x)$ is what we can compute if we observe $X = x$. Thus, in Example 1.1.1, the fraction defective in the sample, $T(x) = x/n$. In Example 1.1.2 a common estimate of μ is the statistic $T(X_1, \dots, X_n) = \bar{X} \equiv \frac{1}{n} \sum_{i=1}^n X_i$, a common estimate of σ^2 is the statistic

$$s^2 \equiv \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

\bar{X} and s^2 are called the *sample mean* and *sample variance*. How we use statistics in estimation and other decision procedures is the subject of the next section.

For future reference we note that a statistic just as a parameter need not be real or Euclidean valued. For instance, a statistic we shall study extensively in Chapter 2 is the function valued statistic \hat{F} , called the *empirical distribution function*, which evaluated at $x \in R$ is

$$\hat{F}(X_1, \dots, X_n)(x) = \frac{1}{n} \sum_{i=1}^n 1(X_i \leq x)$$

where (X_1, \dots, X_n) are a sample from a probability P on R and $1(A)$ is the indicator of the event A . This statistic takes values in the set of all distribution functions on R . It estimates the function valued parameter F defined by its evaluation at $x \in R$,

$$F(P)(x) = P[X_1 \leq x].$$

Deciding which statistics are important is closely connected to deciding which parameters are important and, hence, can be related to model formulation as we saw earlier. For instance, consider situation (d) listed at the beginning of this section. If we suppose there is a single numerical measure of performance of the drugs and the difference in performance of the drugs for any given patient is a constant irrespective of the patient, then our attention naturally focuses on estimating this constant. If, however, this difference depends on the patient in a complex manner (the effect of each drug is complex), we have to formulate a relevant measure of the difference in performance of the drugs and decide how to estimate this measure.

Often the outcome of the experiment is used to decide on the model and the appropriate measure of difference. Next this model, which now depends on the data, is used to decide what estimate of the measure of difference should be employed (cf., for example, Mandel, 1964). Data-based model selection can make it difficult to ascertain or even assign a meaning to the accuracy of estimates or the probability of reaching correct conclusions. Nevertheless, we can draw guidelines from our numbers and cautiously proceed. These issues will be discussed further in Volume 2. In this volume we assume that the model has

been selected prior to the current experiment. This selection is based on experience with previous similar experiments (cf. Lehmann, 1990).

There are also situations in which selection of what data will be observed depends on the experimenter and on his or her methods of reaching a conclusion. For instance, in situation (d) again, patients may be considered one at a time, sequentially, and the decision of which drug to administer for a given patient may be made using the knowledge of what happened to the previous patients. The experimenter may, for example, assign the drugs alternatively to every other patient in the beginning and then, after a while, assign the drug that seems to be working better to a higher proportion of patients. Moreover, the statistical procedure can be designed so that the experimenter stops experimenting as soon as he or she has significant evidence to the effect that one drug is better than the other. Thus, the number of patients in the study (the sample size) is random. Problems such as these lie in the fields of *sequential analysis* and *experimental design*. They are not covered under our general model and will not be treated in this book. We refer the reader to Wetherill and Glazebrook (1986) and Kendall and Stuart (1966) for more information.

Notation. Regular models. When dependence on θ has to be observed, we shall denote the distribution corresponding to any particular parameter value θ by P_θ . Expectations calculated under the assumption that $\mathbf{X} \sim P_\theta$ will be written E_θ . Distribution functions will be denoted by $F(\cdot, \theta)$, density and frequency functions by $p(\cdot, \theta)$. However, these and other subscripts and arguments will be omitted where no confusion can arise.

It will be convenient to assume⁽¹⁾ from now on that in any parametric model we consider either:

- (1) All of the P_θ are continuous with densities $p(\mathbf{x}, \theta)$;
- (2) All of the P_θ are discrete with frequency functions $p(\mathbf{x}, \theta)$, and the set $\{\mathbf{x}_1, \mathbf{x}_2, \dots\} \equiv \{\mathbf{x}: p(\mathbf{x}, \theta) > 0\}$ is the same set for all $\theta \in \Theta$.

Such models will be called *regular parametric models*. In the discrete case we will use both the terms *frequency function* and *density* for $p(\mathbf{x}, \theta)$. See A.10.

1.1.4 Examples, Regression Models

We end this section with two further important examples indicating the wide scope of the notions we have introduced.

In most studies we are interested in studying relations between responses and several other variables not just treatment or control as in Example 1.1.3. This is the stage for the following.

Example 1.1.4. Regression Models. We observe $(\mathbf{z}_1, Y_1), \dots, (\mathbf{z}_n, Y_n)$ where Y_1, \dots, Y_n are independent. The distribution of the response Y_i for the i th subject or case in the study is postulated to depend on certain characteristics \mathbf{z}_i of the i th subject. Thus, \mathbf{z}_i is a d dimensional vector that gives characteristics such as sex, age, height, weight, and so on of the i th subject in a study. For instance, in Example 1.1.3 we could take \mathbf{z} to be the treatment label and write our observations as $(A, X_1), (A, X_m), (B, Y_1), \dots, (B, Y_n)$. This is obviously overkill but suppose that, in the study, drugs A and B are given at several

dose levels. Then, $d = 2$ and \mathbf{z}_i^T can denote the pair (Treatment Label, Treatment Dose Level) for patient i .

In general, \mathbf{z}_i is a nonrandom vector of values called a *covariate* vector or a vector of *explanatory variables* whereas Y_i is random and referred to as the *response variable* or *dependent variable* in the sense that its distribution depends on \mathbf{z}_i . If we let $f(y_i | \mathbf{z}_i)$ denote the density of Y_i for a subject with covariate vector \mathbf{z}_i , then the model is

$$(a) \quad p(y_1, \dots, y_n) = \prod_{i=1}^n f(y_i | \mathbf{z}_i).$$

If we let $\mu(\mathbf{z})$ denote the expected value of a response with given covariate vector \mathbf{z} , then we can write,

$$(b) \quad Y_i = \mu(\mathbf{z}_i) + \epsilon_i, \quad i = 1, \dots, n$$

where $\epsilon_i = Y_i - E(Y_i)$, $i = 1, \dots, n$. Here $\mu(\mathbf{z})$ is an unknown function from R^d to R that we are interested in. For instance, in Example 1.1.3 with the Gaussian two-sample model $\mu(A) = \mu$, $\mu(B) = \mu + \Delta$. We usually need to postulate more. A common (but often violated assumption) is

(1) The ϵ_i are identically distributed with distribution F . That is, the effect of \mathbf{z} on Y is through $\mu(\mathbf{z})$ only. In the two sample models this is implied by the constant treatment effect assumption. See Problem 1.1.8.

On the basis of subject matter knowledge and/or convenience it is usually postulated that

(2) $\mu(\mathbf{z}) = g(\beta, \mathbf{z})$ where g is known except for a vector $\beta = (\beta_1, \dots, \beta_d)^T$ of unknowns. The most common choice of g is the linear form,

(3) $g(\beta, \mathbf{z}) = \sum_{j=1}^d \beta_j z_j = \mathbf{z}^T \beta$ so that (b) becomes

$$(b') \quad Y_i = \mathbf{z}_i^T \beta + \epsilon_i, \quad 1 \leq i \leq n.$$

This is the *linear model*. Often the following final assumption is made:

(4) The distribution F of (1) is $\mathcal{N}(0, \sigma^2)$ with σ^2 unknown. Then we have the classical *Gaussian linear model*, which we can write in vector matrix form,

$$(c) \quad \mathbf{Y} \sim \mathcal{N}_n(\mathbf{Z}\beta, \sigma^2 J)$$

where $\mathbf{Z}_{n \times d} = (\mathbf{z}_1^T, \dots, \mathbf{z}_n^T)^T$ and J is the $n \times n$ identity.

Clearly, Example 1.1.3(3) is a special case of this model. So is Example 1.1.2 with assumptions (1)–(4). In fact by varying our assumptions this class of models includes any situation in which we have independent but not necessarily identically distributed observations. By varying the assumptions we obtain parametric models as with (1), (3) and (4) above, semiparametric as with (1) and (2) with F arbitrary, and nonparametric if we drop (1) and simply treat the \mathbf{z}_i as a label of the completely unknown distributions of Y_i . Identifiability of these parametrizations and the status of their components as parameters are discussed in the problems. \square

Finally, we give an example in which the responses are dependent.

Example 1.1.5. *Measurement Model with Autoregressive Errors.* Let X_1, \dots, X_n be the n determinations of a physical constant μ . Consider the model where

$$X_i = \mu + e_i, \quad i = 1, \dots, n$$

and assume

$$e_i = \beta e_{i-1} + \epsilon_i, \quad i = 1, \dots, n, \quad e_0 = 0$$

where ϵ_i are independent identically distributed with density f . Here the errors e_1, \dots, e_n are dependent as are the X 's. In fact we can write

$$X_i = \mu(1 - \beta) + \beta X_{i-1} + \epsilon_i, \quad i = 2, \dots, n, \quad X_1 = \mu + \epsilon_1. \quad (1.1.4)$$

An example would be, say, the elapsed times X_1, \dots, X_n spent above a fixed high level for a series of n consecutive wave records at a point on the seashore. Let $\mu = E(X_i)$ be the average time for an infinite series of records. It is plausible that e_i depends on e_{i-1} because long waves tend to be followed by long waves. A second example is consecutive measurements X_i of a constant μ made by the same observer who seeks to compensate for apparent errors. Of course, model (1.1.4) assumes much more but it may be a reasonable first approximation in these situations.

To find the density $p(x_1, \dots, x_n)$, we start by finding the density of e_1, \dots, e_n . Using conditional probability theory and $e_i = \beta e_{i-1} + \epsilon_i$, we have

$$\begin{aligned} p(e_1, \dots, e_n) &= p(e_1)p(e_2 | e_1)p(e_3 | e_1, e_2) \dots p(e_n | e_1, \dots, e_{n-1}) \\ &= p(e_1)p(e_2 | e_1)p(e_3 | e_2) \dots p(e_n | e_{n-1}) \\ &= f(e_1)f(e_2 - \beta e_1) \dots f(e_n - \beta e_{n-1}). \end{aligned}$$

Because $e_i = X_i - \mu$, the model for X_1, \dots, X_n is

$$p(x_1, \dots, x_n) = f(x_1 - \mu) \prod_{j=2}^n f(x_j - \beta x_{j-1} - (1 - \beta)\mu).$$

The default assumption, at best an approximation for the wave example, is that f is the $N(0, \sigma^2)$ density. Then we have what is called the AR(1) *Gaussian model*

$$p(x_1, \dots, x_n) = (2\pi)^{-\frac{1}{2}n} \sigma^{-n} \exp \left\{ -\frac{1}{2\sigma^2} \left[(x_1 - \mu)^2 + \sum_{i=2}^n (x_i - \beta x_{i-1} - (1 - \beta)\mu)^2 \right] \right\}.$$

We include this example to illustrate that we need not be limited by independence. However, save for a brief discussion in Volume 2, the conceptual issues of stationarity, ergodicity, and the associated probability theory models and inference for dependent data are beyond the scope of this book. \square

Summary. In this section we introduced the first basic notions and formalism of mathematical statistics, vector *observations* \mathbf{X} with unknown probability distributions P ranging over *models* \mathcal{P} . The notions of *parametrization* and *identifiability* are introduced. The general definition of *parameters* and *statistics* is given and the connection between parameters and parametrizations elucidated. This is done in the context of a number of classical examples, the most important of which is the workhorse of statistics, the *regression model*. We view statistical models as useful tools for learning from the outcomes of experiments and studies. They are useful in understanding how the outcomes can be used to draw inferences that go beyond the particular experiment. Models are approximations to the mechanisms generating the observations. How useful a particular model is is a complex mix of how good the approximation is and how much insight it gives into drawing inferences.

1.2 BAYESIAN MODELS

Throughout our discussion so far we have assumed that there is no information available about the true value of the parameter beyond that provided by the data. There are situations in which most statisticians would agree that more can be said. For instance, in the inspection Example 1.1.1, it is possible that, in the past, we have had many shipments of size N that have subsequently been distributed. If the customers have provided accurate records of the number of defective items that they have found, we can construct a frequency distribution $\{\pi_0, \dots, \pi_N\}$ for the proportion θ of defectives in past shipments. That is, π_i is the frequency of shipments with i defective items, $i = 0, \dots, N$. Now it is reasonable to suppose that the value of θ in the present shipment is the realization of a random variable θ with distribution given by

$$P[\theta = \frac{i}{N}] = \pi_i, \quad i = 0, \dots, N. \quad (1.2.1)$$

Our model is then specified by the joint distribution of the observed number X of defectives in the sample and the random variable θ . We know that, given $\theta = i/N$, X has the hypergeometric distribution $\mathcal{H}(i, N, n)$. Thus,

$$\begin{aligned} P[X = k, \theta = \frac{i}{N}] &= P[\theta = \frac{i}{N}]P[X = k \mid \theta = \frac{i}{N}] \\ &= \pi_i \frac{\binom{i}{k} \binom{N-i}{n-k}}{\binom{N}{n}}. \end{aligned} \quad (1.2.2)$$

This is an example of a *Bayesian* model.

There is a substantial number of statisticians who feel that it is always reasonable, and indeed necessary, to think of the true value of the parameter θ as being the realization of a random variable θ with a known distribution. This distribution does not always correspond to an experiment that is physically realizable but rather is thought of as a measure of the beliefs of the experimenter concerning the true value of θ before he or she takes any data.

Thus, the resulting statistical inference becomes subjective. The theory of this school is expounded by L. J. Savage (1954), Raiffa and Schlaiffer (1961), Lindley (1965), De Groot (1969), and Berger (1985). An interesting discussion of a variety of points of view on these questions may be found in Savage et al. (1962). There is an even greater range of viewpoints in the statistical community from people who consider all statistical statements as purely subjective to ones who restrict the use of such models to situations such as that of the inspection example in which the distribution of θ has an objective interpretation in terms of frequencies. Our own point of view is that subjective elements including the views of subject matter experts are an essential element in all model building. However, insofar as possible we prefer to take the frequentist point of view in validating statistical statements and avoid making final claims in terms of subjective posterior probabilities (see later). However, by giving θ a distribution purely as a theoretical tool to which no subjective significance is attached, we can obtain important and useful results and insights. We shall return to the Bayesian framework repeatedly in our discussion.

In this section we shall define and discuss the basic elements of Bayesian models. Suppose that we have a regular parametric model $\{P_\theta : \theta \in \Theta\}$. To get a Bayesian model we introduce a random vector θ , whose range is contained in Θ , with density or frequency function π . The function π represents our belief or information about the parameter θ before the experiment and is called the *prior density or frequency function*. We now think of P_θ as the conditional distribution of \mathbf{X} given $\theta = \theta$. The joint distribution of (θ, \mathbf{X}) is that of the outcome of a random experiment in which we first select $\theta = \theta$ according to π and then, given $\theta = \theta$, select \mathbf{X} according to P_θ . If both \mathbf{X} and θ are continuous or both are discrete, then by (B.1.3), (θ, \mathbf{X}) is appropriately continuous or discrete with density or frequency function,

$$f(\theta, \mathbf{x}) = \pi(\theta)p(\mathbf{x}, \theta). \quad (1.2.3)$$

Because we now think of $p(\mathbf{x}, \theta)$ as a conditional density or frequency function given $\theta = \theta$, we will denote it by $p(\mathbf{x} | \theta)$ for the remainder of this section.

Equation (1.2.2) is an example of (1.2.3). In the “mixed” cases such as θ continuous \mathbf{X} discrete, the joint distribution is neither continuous nor discrete.

The most important feature of a Bayesian model is the conditional distribution of θ given $\mathbf{X} = \mathbf{x}$, which is called the *posterior* distribution of θ . Before the experiment is performed, the information or belief about the true value of the parameter is described by the prior distribution. After the value \mathbf{x} has been obtained for \mathbf{X} , the information about θ is described by the posterior distribution.

For a concrete illustration, let us turn again to Example 1.1.1. For instance, suppose that $N = 100$ and that from past experience we believe that each item has probability .1 of being defective independently of the other members of the shipment. This would lead to the prior distribution

$$\pi_i = \binom{100}{i} (0.1)^i (0.9)^{100-i}, \quad (1.2.4)$$

for $i = 0, 1, \dots, 100$. Before sampling any items the chance that a given shipment contains

20 or more bad items is by the normal approximation with continuity correction, (A.15.10),

$$\begin{aligned} P[100\theta \geq 20] &= P \left[\frac{100\theta - 10}{\sqrt{100(0.1)(0.9)}} \geq \frac{10}{\sqrt{100(0.1)(0.9)}} \right] \\ &\approx 1 - \Phi \left(\frac{9.5}{3} \right) = 0.001. \end{aligned} \quad (1.2.5)$$

Now suppose that a sample of 19 has been drawn in which 10 defective items are found. This leads to

$$P[100\theta \geq 20 \mid X = 10] \approx 0.30. \quad (1.2.6)$$

To calculate the posterior probability given in (1.2.6) we argue loosely as follows: If before the drawing each item was defective with probability .1 and good with probability .9 independently of the other items, this will continue to be the case for the items left in the lot after the 19 sample items have been drawn. Therefore, $100\theta - X$, the number of defectives left after the drawing, is independent of X and has a $\mathcal{B}(81, 0.1)$ distribution. Thus,

$$\begin{aligned} P[100\theta \geq 20 \mid X = 10] &= P[100\theta - X \geq 10 \mid X = 10] \\ &= P \left[\frac{(100\theta - X) - 8.1}{\sqrt{81(0.9)(0.1)}} \geq \frac{1.9}{\sqrt{81(0.9)(0.1)}} \right] \\ &\approx 1 - \Phi(0.52) \\ &= 0.30. \end{aligned} \quad (1.2.7)$$

In general, to calculate the posterior, some variant of Bayes' Theorem (B.1.4) can be used. Specifically,

- (i) The posterior distribution is discrete or continuous according as the prior distribution is discrete or continuous.
- (ii) If we denote the corresponding (posterior) frequency function or density by $\pi(\theta \mid \mathbf{x})$, then

$$\begin{aligned} \pi(\theta \mid \mathbf{x}) &= \frac{\pi(\theta)p(\mathbf{x} \mid \theta)}{\sum_t \pi(t)p(\mathbf{x} \mid t)} && \text{if } \theta \text{ is discrete,} \\ &= \frac{\pi(\theta)p(\mathbf{x} \mid \theta)}{\int_{-\infty}^{\infty} \pi(t)p(\mathbf{x} \mid t)dt} && \text{if } \theta \text{ is continuous.} \end{aligned} \quad (1.2.8)$$

In the cases where θ and \mathbf{X} are both continuous or both discrete this is precisely Bayes' rule applied to the joint distribution of (θ, \mathbf{X}) given by (1.2.3). Here is an example.

Example 1.2.1. Bernoulli Trials. Suppose that X_1, \dots, X_n are indicators of n Bernoulli trials with probability of success θ where $0 < \theta < 1$. If we assume that θ has a priori distribution with density π , we obtain by (1.2.8) as posterior density of θ ,

$$\pi(\theta \mid x_1, \dots, x_n) = \frac{\pi(\theta)\theta^k(1-\theta)^{n-k}}{\int_0^1 \pi(t)t^k(1-t)^{n-k}dt} \quad (1.2.9)$$

for $0 < \theta < 1$, $x_i = 0$ or 1 , $i = 1, \dots, n$, $k = \sum_{i=1}^n x_i$.

Note that the posterior density depends on the data only through the total number of successes, $\sum_{i=1}^n X_i$. We also obtain the same posterior density if θ has prior density π and we only observe $\sum_{i=1}^n X_i$, which has a $\mathcal{B}(n, \theta)$ distribution given $\theta = \theta$ (Problem 1.2.9). We can thus write $\pi(\theta | k)$ for $\pi(\theta | x_1, \dots, x_n)$, where $k = \sum_{i=1}^n x_i$.

To choose a prior π , we need a class of distributions that concentrate on the interval $(0, 1)$. One such class is the two-parameter beta family. This class of distributions has the remarkable property that the resulting posterior distributions are again beta distributions. Specifically, upon substituting the $\beta(r, s)$ density (B.2.11) in (1.2.9) we obtain

$$\pi(\theta | k) = \frac{\theta^{r-1}(1-\theta)^{s-1}\theta^k(1-\theta)^{n-k}}{c} = \frac{\theta^{k+r-1}(1-\theta)^{n-k+s-1}}{c}. \quad (1.2.10)$$

The proportionality constant c , which depends on k , r , and s only, must (see (B.2.11)) be $B(k+r, n-k+s)$ where $B(\cdot, \cdot)$ is the beta function, and the posterior distribution of θ given $\sum X_i = k$ is $\beta(k+r, n-k+s)$.

As Figure B.2.2 indicates, the beta family provides a wide variety of shapes that can approximate many reasonable prior distributions though by no means all. For instance, non- U -shaped bimodal distributions are not permitted.

Suppose, for instance, we are interested in the proportion θ of “geniuses” ($IQ \geq 160$) in a particular city. To get information we take a sample of n individuals from the city. If n is small compared to the size of the city, (A.15.13) leads us to assume that the number X of geniuses observed has approximately a $\mathcal{B}(n, \theta)$ distribution. Now we may either have some information about the proportion of geniuses in similar cities of the country or we may merely have prejudices that we are willing to express in the form of a prior distribution on θ . We may want to assume that θ has a density with maximum value at 0 such as that drawn with a dotted line in Figure B.2.2. Or else we may think that $\pi(\theta)$ concentrates its mass near a small number, say 0.05. Then we can choose r and s in the $\beta(r, s)$ distribution, so that the mean is $r/(r+s) = 0.05$ and its variance is very small. The result might be a density such as the one marked with a solid line in Figure B.2.2.

If we were interested in some proportion about which we have no information or belief, we might take θ to be uniformly distributed on $(0, 1)$, which corresponds to using the beta distribution with $r = s = 1$. \square

A feature of Bayesian models exhibited by this example is that there are natural parametric families of priors such that the posterior distributions also belong to this family. Such families are called *conjugate*. Evidently the beta family is conjugate to the binomial. Another bigger conjugate family is that of finite mixtures of beta distributions—see Problem 1.2.16. We return to conjugate families in Section 1.6.

Summary. We present an elementary discussion of *Bayesian models*, introduce the notions of *prior* and *posterior* distributions and give *Bayes rule*. We also by example introduce the notion of a *conjugate family* of distributions.

1.3 THE DECISION THEORETIC FRAMEWORK

Given a statistical model, the information we want to draw from data can be put in various forms depending on the purposes of our analysis. We may wish to produce “best guesses” of the values of important parameters, for instance, the fraction defective θ in Example 1.1.1 or the physical constant μ in Example 1.1.2. These are *estimation* problems. In other situations certain P are “special” and we may primarily wish to know whether the data support “specialness” or not. For instance, in Example 1.1.3, P ’s that correspond to no treatment effect (i.e., placebo and treatment are equally effective) are special because the FDA (Food and Drug Administration) does not wish to permit the marketing of drugs that do no good. If μ_0 is the critical matter density in the universe so that $\mu < \mu_0$ means the universe is expanding forever and $\mu \geq \mu_0$ correspond to an eternal alternation of Big Bangs and expansions, then depending on one’s philosophy one could take either P ’s corresponding to $\mu < \mu_0$ or those corresponding to $\mu \geq \mu_0$ as special. Making determinations of “specialness” corresponds to *testing significance*. As the second example suggests, there are many problems of this type in which it’s unclear which of two disjoint sets of P ’s; \mathcal{P}_0 or \mathcal{P}_0^c is special and the general *testing problem* is really one of discriminating between \mathcal{P}_0 and \mathcal{P}_0^c . For instance, in Example 1.1.1 contractual agreement between shipper and receiver may penalize the return of “good” shipments, say, with $\theta < \theta_0$, whereas the receiver does not wish to keep “bad,” $\theta \geq \theta_0$, shipments. Thus, the receiver wants to discriminate and may be able to attach monetary costs to making a mistake of either type: “keeping the bad shipment” or “returning a good shipment.” In testing problems we, at a first cut, state which is supported by the data: “specialness” or, as it’s usually called, “hypothesis” or “nonspecialness” (or alternative).

We may have other goals as illustrated by the next two examples.

Example 1.3.1. Ranking. A consumer organization preparing (say) a report on air conditioners tests samples of several brands. On the basis of the sample outcomes the organization wants to give a ranking from best to worst of the brands (ties not permitted). Thus, if there are k different brands, there are $k!$ possible rankings or actions, one of which will be announced as more consistent with the data than others. \square

Example 1.3.2. Prediction. A very important class of situations arises when, as in Example 1.1.4, we have a vector \mathbf{z} , such as, say, (age, sex, drug dose)^T that can be used for prediction of a variable of interest Y , say a 50-year-old male patient’s response to the level of a drug. Intuitively, and as we shall see formally later, a reasonable prediction rule for an unseen Y (response of a new patient) is the function $\mu(\mathbf{z})$, the expected value of Y given \mathbf{z} . Unfortunately $\mu(\mathbf{z})$ is unknown. However, if we have observations (\mathbf{z}_i, Y_i) , $1 \leq i \leq n$, we can try to estimate the function $\mu(\cdot)$. For instance, if we believe $\mu(\mathbf{z}) = g(\beta, \mathbf{z})$ we can estimate β from our observations Y_i of $g(\beta, \mathbf{z}_i)$ and then plug our estimate of β into g . Note that we really want to estimate the function $\mu(\cdot)$; our results will guide the selection of doses of drug for future patients. \square

In all of the situations we have discussed it is clear that the analysis does not stop by specifying an estimate or a test or a ranking or a prediction function. There are many possible choices of estimates. In Example 1.1.1 do we use the observed fraction of defectives

X/n as our estimate or ignore the data and use historical information on past shipments, or combine them in some way? In Example 1.1.2 to estimate μ do we use the mean of the measurements, $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, or the median, defined as any value such that half the X_i are at least as large and half no bigger? The same type of question arises in all examples. The answer will depend on the model and, most significantly, on what criteria of performance we use. Intuitively, in estimation we care how far off we are, in testing whether we are right or wrong, in ranking what mistakes we've made, and so on. In any case, whatever our choice of procedure we need either a priori (before we have looked at the data) and/or a posteriori estimates of how well we're doing. In designing a study to compare treatments A and B we need to determine sample sizes that will be large enough to enable us to detect differences that matter. That is, we need a priori estimates of how well even the best procedure can do. For instance, in Example 1.1.3 even with the simplest Gaussian model it is intuitively clear and will be made precise later that, even if Δ is large, a large σ^2 will force a large m, n to give us a good chance of correctly deciding that the treatment effect is there. On the other hand, once a study is carried out we would probably want not only to estimate Δ but also know how reliable our estimate is. Thus, we would want a posteriori estimates of performance.

These examples motivate the decision theoretic framework: We need to

- (1) clarify the objectives of a study,
- (2) point to what the different possible actions are,
- (3) provide assessments of risk, accuracy, and reliability of statistical procedures,
- (4) provide guidance in the choice of procedures for analyzing outcomes of experiments.

1.3.1 Components of the Decision Theory Framework

As in Section 1.1, we begin with a statistical model with an observation vector \mathbf{X} whose distribution P ranges over a set \mathcal{P} . We usually take \mathcal{P} to be parametrized, $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$.

Action space. A new component is an *action space* \mathcal{A} of actions or decisions or claims that we can contemplate making. Here are action spaces for our examples.

Estimation. If we are estimating a real parameter such as the fraction θ of defectives, in Example 1.1.1, or μ in Example 1.1.2, it is natural to take $\mathcal{A} = \mathcal{R}$ though smaller spaces may serve equally well, for instance, $\mathcal{A} = \{0, \frac{1}{N}, \dots, 1\}$ in Example 1.1.1.

Testing. Here only two actions are contemplated: accepting or rejecting the “specialness” of P (or in more usual language the hypothesis $H : P \in \mathcal{P}_0$ in which we identify \mathcal{P}_0 with the set of “special” P 's). By convention, $\mathcal{A} = \{0, 1\}$ with 1 corresponding to rejection of H . Thus, in Example 1.1.3, taking action 1 would mean deciding that $\Delta \neq 0$.

Ranking. Here quite naturally $\mathcal{A} = \{\text{Permutations } (i_1, \dots, i_k) \text{ of } \{1, \dots, k\}\}$. Thus, if we have three air conditioners, there are $3! = 6$ possible rankings,

$$\mathcal{A} = \{(1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), (3, 2, 1)\}.$$

Prediction. Here \mathcal{A} is much larger. If Y is real, and $\mathbf{z} \in Z$, $\mathcal{A} = \{a : a \text{ is a function from } Z \text{ to } R\}$ with $a(\mathbf{z})$ representing the prediction we would make if the new unobserved Y had covariate value \mathbf{z} . Evidently Y could itself range over an arbitrary space \mathcal{Y} and then R would be replaced by \mathcal{Y} in the definition of $a(\cdot)$. For instance, if $Y = 0$ or 1 corresponds to, say, “does not respond” and “responds,” respectively, and $\mathbf{z} = (\text{Treatment, Sex})^T$, then $a(B, M)$ would be our prediction of response or no response for a male given treatment B .

Loss function. Far more important than the choice of action space is the choice of *loss function* defined as a function $l : \mathcal{P} \times \mathcal{A} \rightarrow R^+$. The interpretation of $l(P, a)$, or $l(\theta, a)$ if \mathcal{P} is parametrized, is the nonnegative loss incurred by the statistician if he or she takes action a and the true “state of Nature,” that is, the probability distribution producing the data, is P . As we shall see, although loss functions, as the name suggests, sometimes can genuinely be quantified in economic terms, they usually are chosen to qualitatively reflect what we are trying to do and to be mathematically convenient.

Estimation. In estimating a real valued parameter $\nu(P)$ or $q(\theta)$ if \mathcal{P} is parametrized the most commonly used loss function is,

Quadratic Loss: $l(P, a) = (\nu(P) - a)^2$ (or $l(\theta, a) = (q(\theta) - a)^2$).

Other choices that are, as we shall see (Section 5.1), less computationally convenient but perhaps more realistically penalize large errors less are *Absolute Value Loss*: $l(P; a) = |\nu(P) - a|$, and *truncated quadratic loss*: $l(P, a) = \min\{(\nu(P) - a)^2, d^2\}$. Closely related to the latter is what we shall call *confidence interval loss*, $l(P, a) = 0$, $|\nu(P) - a| \leq d$, $l(P, a) = 1$ otherwise. This loss expresses the notion that all errors within the limits $\pm d$ are tolerable and outside these limits equally intolerable. Although estimation loss functions are typically symmetric in ν and a , asymmetric loss functions can also be of importance. For instance, $l(P, a) = 1(\nu < a)$, which penalizes only overestimation and by the same amount arises naturally with lower confidence bounds as discussed in Example 1.3.3.

If $\nu = (\nu_1, \dots, \nu_d) = (q_1(\theta), \dots, q_d(\theta))$ and $\mathbf{a} = (a_1, \dots, a_d)$ are vectors, examples of loss functions are

$$l(\theta, \mathbf{a}) = \frac{1}{d} \sum (a_j - \nu_j)^2 = \text{squared Euclidean distance}/d$$

$$l(\theta, \mathbf{a}) = \frac{1}{d} \sum |a_j - \nu_j| = \text{absolute distance}/d$$

$$l(\theta, \mathbf{a}) = \max\{|a_j - \nu_j|, j = 1, \dots, d\} = \text{supremum distance.}$$

We can also consider function valued parameters. For instance, in the *prediction example* 1.3.2, $\mu(\cdot)$ is the parameter of interest. If we use $a(\cdot)$ as a predictor and the new \mathbf{z} has marginal distribution Q then it is natural to consider,

$$l(P, a) = \int (\mu(\mathbf{z}) - a(\mathbf{z}))^2 dQ(\mathbf{z}),$$

the expected squared error if a is used. If, say, Q is the empirical distribution of the \mathbf{z}_j in

the training set $(\mathbf{z}_1, Y), \dots, (\mathbf{z}_n, \mathbf{Y}_n)$, this leads to the commonly considered

$$l(P, a) = \frac{1}{n} \sum_{j=1}^n (\mu(\mathbf{z}_j) - a(\mathbf{z}_j))^2,$$

which is just n^{-1} times the squared Euclidean distance between the prediction vector $(a(\mathbf{z}_1), \dots, a(\mathbf{z}_n))^T$ and the vector parameter $(\mu(\mathbf{z}_1), \dots, \mu(\mathbf{z}_n))^T$.

Testing. We ask whether the parameter θ is in the subset Θ_0 or subset Θ_1 of Θ , where $\{\Theta_0, \Theta_1\}$, is a partition of Θ (or equivalently if $P \in \mathcal{P}_0$ or $P \in \mathcal{P}_1$). If we take action a when the parameter is in Θ_a , we have made the correct decision and the loss is zero. Otherwise, the decision is wrong and the loss is taken to equal one. This 0 – 1 loss function can be written as

$$\begin{aligned} 0 - 1 \text{ loss: } l(\theta, a) &= 0 \text{ if } \theta \in \Theta_a \text{ (The decision is correct)} \\ l(\theta, a) &= 1 \text{ otherwise (The decision is wrong).} \end{aligned}$$

Of course, other economic loss functions may be appropriate. For instance, in Example 1.1.1 suppose returning a shipment with $\theta < \theta_0$ defectives results in a penalty of s dollars whereas every defective item sold results in an r dollar replacement cost. Then the appropriate loss function is

$$\begin{aligned} l(\theta, 1) &= s \text{ if } \theta < \theta_0 \\ l(\theta, 1) &= 0 \text{ if } \theta \geq \theta_0 \\ l(\theta, 0) &= rN\theta. \end{aligned} \tag{1.3.1}$$

Decision procedures. We next give a representation of the process whereby the statistician uses the data to arrive at a decision. The data is a point $\mathbf{X} = \mathbf{x}$ in the outcome or sample space \mathcal{X} . We define a *decision rule* or *procedure* δ to be any function from the sample space taking its values in A . Using δ means that if $\mathbf{X} = \mathbf{x}$ is observed, the statistician takes action $\delta(\mathbf{x})$.

Estimation. For the problem of estimating the constant μ in the measurement model, we implicitly discussed two estimates or decision rules: $\delta_1(\mathbf{x}) =$ sample mean \bar{x} and $\delta_2(\mathbf{x}) = \hat{x} =$ sample median.

Testing. In Example 1.1.3 with X and Y distributed as $\mathcal{N}(\mu, \sigma^2)$ and $\mathcal{N}(\mu + \Delta, \sigma^2)$, respectively, if we are asking whether the treatment effect parameter Δ is 0 or not, then a reasonable rule is to decide $\Delta = 0$ if our estimate $\bar{x} - \bar{y}$ is close to zero, and to decide $\Delta \neq 0$ if our estimate is not close to zero. Here we mean close to zero relative to the variability in the experiment, that is, relative to the standard deviation σ . In Section 4.9.3 we will show how to obtain an estimate $\hat{\sigma}$ of σ from the data. The decision rule can now be written

$$\begin{aligned} \delta(\mathbf{x}, \mathbf{y}) &= 0 \text{ if } \frac{|\bar{x} - \bar{y}|}{\hat{\sigma}} < c \\ &1 \text{ if } \frac{|\bar{x} - \bar{y}|}{\hat{\sigma}} \geq c \end{aligned} \tag{1.3.2}$$

where c is a positive constant called the *critical value*. How do we choose c ? We need the next concept of the decision theoretic framework, the *risk* or *risk function*:

The risk function. If δ is the procedure used, l is the loss function, θ is the true value of the parameter, and $\mathbf{X} = \mathbf{x}$ is the outcome of the experiment, then the loss is $l(P, \delta(\mathbf{x}))$. We do not know the value of the loss because P is unknown. Moreover, we typically want procedures to have good properties not at just one particular \mathbf{x} , but for a range of plausible \mathbf{x} 's. Thus, we turn to the average or mean loss over the sample space. That is, we regard $l(P, \delta(\mathbf{X}))$ as a random variable and introduce the *risk function*

$$R(P, \delta) = E_P[l(P, \delta(\mathbf{X}))]$$

as the measure of the performance of the decision rule $\delta(\mathbf{X})$. Thus, for each δ , R maps \mathcal{P} or Θ to R^+ . $R(\cdot, \delta)$ is our a priori measure of the performance of δ . We illustrate computation of R and its a priori use in some examples.

Estimation. Suppose $\nu \equiv \nu(P)$ is the real parameter we wish to estimate and $\hat{\nu} \equiv \hat{\nu}(X)$ is our estimator (our decision rule). If we use quadratic loss, our risk function is called the *mean squared error* (MSE) of $\hat{\nu}$ and is given by

$$MSE(\hat{\nu}) = R(P, \hat{\nu}) = E_P(\hat{\nu}(X) - \nu(P))^2 \quad (1.3.3)$$

where for simplicity dependence on P is suppressed in MSE.

The MSE depends on the variance of $\hat{\nu}$ and on what is called the *bias* of $\hat{\nu}$ where

$$\text{Bias}(\hat{\nu}) = E(\hat{\nu}) - \nu$$

can be thought of as the “long-run average error” of $\hat{\nu}$. A useful result is

Proposition 1.3.1.

$$MSE(\hat{\nu}) = (\text{Bias } \hat{\nu})^2 + \text{Var}(\hat{\nu}).$$

Proof. Write the error as

$$(\hat{\nu} - \nu) = [\hat{\nu} - E(\hat{\nu})] + [E(\hat{\nu}) - \nu].$$

If we expand the square of the right-hand side keeping the brackets intact and take the expected value, the cross term will be zero because $E[\hat{\nu} - E(\hat{\nu})] = 0$. The other two terms are $(\text{Bias } \hat{\nu})^2$ and $\text{Var}(\hat{\nu})$. (If one side is infinite, so is the other and the result is trivially true.) \square

If $\text{Bias}(\hat{\nu}) = 0$, $\hat{\nu}$ is called *unbiased*. We next illustrate the computation and the a priori and a posteriori use of the risk function.

Example 1.3.3. *Estimation of μ (Continued).* Suppose X_1, \dots, X_n are i.i.d. measurements of μ with $\mathcal{N}(0, \sigma^2)$ errors. If we use the mean \bar{X} as our estimate of μ and assume quadratic loss, then

$$\begin{aligned} \text{Bias}(\bar{X}) &= E(\bar{X}) - \mu = 0 \\ \text{Var}(\bar{X}) &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{\sigma^2}{n} \end{aligned}$$

and, by Proposition 1.3.1

$$MSE(\bar{X}) = R(\mu, \sigma^2, \bar{X}) = \frac{\sigma^2}{n}, \quad (1.3.4)$$

which doesn't depend on μ .

Suppose that the precision of the measuring instrument σ^2 is known and equal to σ_0^2 or where realistically it is known to be $\leq \sigma_0^2$. Then (1.3.4) can be used for an a priori estimate of the risk of \bar{X} . If we want to be guaranteed $MSE(\bar{X}) \leq \varepsilon^2$ we can do it by taking at least $n_0 = (\sigma_0/\varepsilon)^2$ measurements.

If we have no idea of the value of σ^2 , planning is not possible but having taken n measurements we can then estimate σ^2 , for instance by $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$, or $n\hat{\sigma}^2/(n-1)$, an estimate we can justify later. The *a posteriori* estimate of risk $\hat{\sigma}^2/n$ is, of course, itself subject to random error.

Suppose that instead of quadratic loss we used the more natural⁽¹⁾ absolute value loss. Then

$$R(\mu, \sigma^2, \bar{X}) = E|\bar{X} - \mu| = E|\bar{\varepsilon}|$$

where $\varepsilon_i = X_i - \mu$. If, as we assumed, the ε_i are $\mathcal{N}(0, \sigma^2)$ then by (A.13.23), $(\sqrt{n}/\sigma)\bar{\varepsilon} \sim \mathcal{N}(0, 1)$ and

$$R(\mu, \sigma^2, \bar{X}) = \frac{\sigma}{\sqrt{n}} \int_{-\infty}^{\infty} |t| \varphi(t) dt = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{2}{\pi}}. \quad (1.3.5)$$

This harder calculation already suggests why quadratic loss is really favored. If we only assume, as we discussed in Example 1.1.2, that the ε_i are i.i.d. with mean 0 and variance $\sigma^2(P)$, then for quadratic loss, $R(P, \bar{X}) = \sigma^2(P)/n$ still, but for absolute value loss only approximate, analytic, or numerical and/or Monte Carlo computation, is possible. In fact, computational difficulties arise even with quadratic loss as soon as we think of estimates other than \bar{X} . For instance, define the *sample median* \hat{X} of the sample X_1, \dots, X_n as follows: When n is odd, $\hat{X} = X_{(k)}$ where $k = \frac{1}{2}(n+1)$ and $X_{(1)}, \dots, X_{(n)}$ denotes X_1, \dots, X_n ordered from smallest to largest. When n is even, $\hat{X} = \frac{1}{2}[X_{(r)} + X_{(r+1)}]$, where $r = \frac{1}{2}n$. Now $E(\hat{X} - \mu)^2 = E(\hat{\varepsilon}^2)$ can only be evaluated numerically (see Problem 1.3.6), or approximated asymptotically. \square

We next give an example in which quadratic loss and the breakup of MSE given in Proposition 1.3.1 is useful for evaluating the performance of competing estimators.

Example 1.3.4. Let μ_0 denote the mean of a certain measurement included in the U.S. census, say, age or income. Next suppose we are interested in the mean μ of the same measurement for a certain area of the United States. If we have no data for area A , a natural guess for μ would be μ_0 , whereas if we have a random sample of measurements X_1, X_2, \dots, X_n from area A , we may want to combine μ_0 and $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ into an estimator, for instance,

$$\hat{\mu} = (0.2)\mu_0 + (0.8)\bar{X}.$$

The choice of the weights 0.2 and 0.8 can only be made on the basis of additional knowledge about demography or the economy. We shall derive them in Section 1.6 through a

formal Bayesian analysis using a normal prior to illustrate a way of bringing in additional knowledge. Here we compare the performances of $\hat{\mu}$ and \bar{X} as estimators of μ using MSE. We easily find

$$\begin{aligned}\text{Bias}(\hat{\mu}) &= 0.2\mu_0 + 0.8\mu - \mu = 0.2(\mu_0 - \mu) \\ \text{Var}(\hat{\mu}) &= (0.8)^2 \text{Var}(\bar{X}) = (.64)\sigma^2/n \\ R(\mu, \hat{\mu}) &= \text{MSE}(\hat{\mu}) = .04(\mu_0 - \mu)^2 + (.64)\sigma^2/n.\end{aligned}$$

If μ is close to μ_0 , the risk $R(\mu, \hat{\mu})$ of $\hat{\mu}$ is smaller than the risk $R(\mu, \bar{X}) = \sigma^2/n$ of \bar{X} with the minimum relative risk $\inf\{MSE(\hat{\mu})/MSE(\bar{X}); \mu \in R\}$ being 0.64 when $\mu = \mu_0$. Figure 1.3.1 gives the graphs of $MSE(\hat{\mu})$ and $MSE(\bar{X})$ as functions of μ . Because we do not know the value of μ , using MSE, neither estimator can be proclaimed as being better than the other. However, if we use as our criteria the maximum (over μ) of the MSE (called the *minimax* criteria), then \bar{X} is optimal (Example 3.3.4).

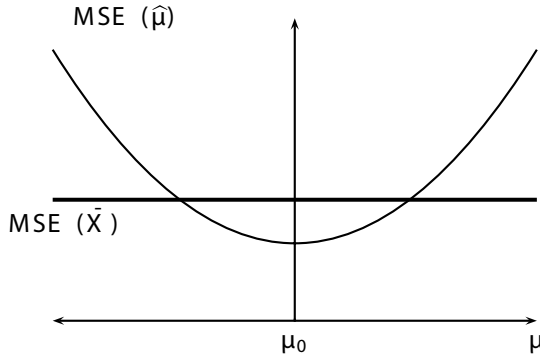


Figure 1.3.1. The mean squared errors of \bar{X} and $\hat{\mu}$. The two MSE curves cross at $\mu = \mu_0 \pm 3\sigma/\sqrt{n}$.

□

Testing. The test rule (1.3.2) for deciding between $\Delta = 0$ and $\Delta \neq 0$ can only take on the two values 0 and 1; thus, the risk is

$$R(\Delta, \delta) = l(\Delta, 0)P[\delta(\mathbf{X}, \mathbf{Y}) = 0] + l(\Delta, 1)P[\delta(\mathbf{X}, \mathbf{Y}) = 1],$$

which in the case of 0 – 1 loss is

$$\begin{aligned}R(\Delta, \delta) &= P[\delta(\mathbf{X}, \mathbf{Y}) = 1] \text{ if } \Delta = 0 \\ &= P[\delta(\mathbf{X}, \mathbf{Y}) = 0] \text{ if } \Delta \neq 0.\end{aligned}$$

In the general case \mathcal{X} and Θ denote the outcome and parameter space, respectively, and we are to decide whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$, where $\Theta = \Theta_0 \cup \Theta_1$, $\Theta_0 \cap \Theta_1 = \emptyset$. A *test*

function is a decision rule $\delta(\mathbf{X})$ that equals 1 on a set $C \subset \mathcal{X}$ called the *critical region* and equals 0 on the complement of C ; that is, $\delta(\mathbf{X}) = 1[\mathbf{X} \in C]$, where 1 denotes the indicator function. If $\delta(\mathbf{X}) = 1$ and we decide $\theta \in \Theta_1$ when in fact $\theta \in \Theta_0$, we call the error committed a *Type I error*, whereas if $\delta(\mathbf{X}) = 0$ and we decide $\theta \in \Theta_0$ when in fact $\theta \in \Theta_1$, we call the error a *Type II error*. Thus, the risk of $\delta(\mathbf{X})$ is

$$\begin{aligned} R(\theta, \delta) &= E(\delta(\mathbf{X})) = P(\delta(\mathbf{X}) = 1) \text{ if } \theta \in \Theta_0 \\ &= \text{Probability of Type I error} \\ R(\theta, \delta) &= P(\delta(\mathbf{X}) = 0) \text{ if } \theta \in \Theta_1 \\ &= \text{Probability of Type II error.} \end{aligned} \tag{1.3.6}$$

Finding good test functions corresponds to finding critical regions with small probabilities of error. In the *Neyman–Pearson* framework of statistical hypothesis testing, the focus is on first providing a small bound, say .05, on the probability of Type I error, and then trying to minimize the probability of a Type II error. The bound on the probability of a Type I error is called the *level of significance* and deciding Θ_1 is referred to as “Rejecting the hypothesis $H : \theta \in \Theta_0$ at level of significance α ”. For instance, in the treatments *A* and *B* example, we want to start by limiting the probability of falsely proclaiming one treatment superior to the other (deciding $\Delta \neq 0$ when $\Delta = 0$), and then next look for a procedure with low probability of proclaiming no difference if in fact one treatment is superior to the other (deciding $\Delta = 0$ when $\Delta \neq 0$).

This is not the only approach to testing. For instance, the loss function (1.3.1) and tests δ_k of the form, “Reject the shipment if and only if $X \geq k$,” in Example 1.1.1 lead to (Problem 1.3.18).

$$\begin{aligned} R(\theta, \delta) &= sP_\theta[X \geq k] + rN\theta P_\theta[X < k], \theta < \theta_0 \\ &= rN\theta P_\theta[X < k], \theta \geq \theta_0. \end{aligned} \tag{1.3.7}$$

Confidence Bounds and Intervals

Decision theory enables us to think clearly about an important hybrid of testing and estimation, confidence bounds and intervals (and more generally regions). Suppose our primary interest in an estimation type of problem is to give an upper bound for the parameter ν . For instance, an accounting firm examining accounts receivable for a firm on the basis of a random sample of accounts would be primarily interested in an upper bound on the total amount owed. If (say) X represents the amount owed in the sample and ν is the unknown total amount owed, it is natural to seek $\bar{\nu}(X)$ such that

$$P[\bar{\nu}(X) \geq \nu] \geq 1 - \alpha \tag{1.3.8}$$

for all possible distributions P of X . Such a $\bar{\nu}$ is called a $(1 - \alpha)$ upper confidence bound on ν . Here α is small, usually .05 or .01 or less. This corresponds to an a priori bound

on the risk of α on $\bar{\nu}(X)$ viewed as a decision procedure with action space R and loss function,

$$\begin{aligned} l(P, a) &= 0, \quad a \geq \nu(P) \\ &= 1, \quad a < \nu(P) \end{aligned}$$

an asymmetric estimation type loss function. The 0 – 1 nature makes it resemble a testing loss function and, as we shall see in Chapter 4, the connection is close. It is clear, though, that this formulation is inadequate because by taking $\bar{\nu} \equiv \infty$ we can achieve risk $\equiv 0$. What is missing is the fact that, though upper bounding is the primary goal, in fact it is important to get close to the truth—knowing that at most ∞ dollars are owed is of no use. The decision theoretic framework accommodates by adding a component reflecting this. For instance

$$\begin{aligned} l(P, a) &= a - \nu(P) \quad , a \geq \nu(P) \\ &= c \quad , a < \nu(P), \end{aligned}$$

for some constant $c > 0$. Typically, rather than this Lagrangian form, it is customary to first fix α in (1.3.8) and then see what one can do to control (say) $R(P, \bar{\nu}) = E(\bar{\nu}(X) - \nu(P))_+$, where $x_+ = x1(x \geq 0)$.

The same issue arises when we are interested in a *confidence interval* $[\underline{\nu}(X), \bar{\nu}(X)]$ for ν defined by the requirement that

$$P[\underline{\nu}(X) \leq \nu(P) \leq \bar{\nu}(X)] \geq 1 - \alpha$$

for all $P \in \mathcal{P}$. We shall go into this further in Chapter 4.

We next turn to the final topic of this section, general criteria for selecting “optimal” procedures.

1.3.2 Comparison of Decision Procedures

In this section we introduce a variety of concepts used in the comparison of decision procedures. We shall illustrate some of the relationships between these ideas using the following simple example in which Θ has two members, \mathcal{A} has three points, and the risk of all possible decision procedures can be computed and plotted. We conclude by indicating to what extent the relationships suggested by this picture carry over to the general decision theoretic model.

Example 1.3.5. Suppose we have two possible states of nature, which we represent by θ_1 and θ_2 . For instance, a component in a piece of equipment either works or does not work; a certain location either contains oil or does not; a patient either has a certain disease or does not, and so on. Suppose that three possible actions, a_1 , a_2 , and a_3 , are available. In the context of the foregoing examples, we could leave the component in, replace it, or repair it; we could drill for oil, sell the location, or sell partial rights; we could operate, administer drugs, or wait and see. Suppose the following loss function is decided on

TABLE 1.3.1. The loss function $l(\theta, a)$

		(Drill)	(Sell)	(Partial rights)
		a_1	a_2	a_3
(Oil)	θ_1	0	10	5
(No oil)	θ_2	12	1	6

Thus, if there is oil and we drill, the loss is zero, whereas if there is no oil and we drill, the loss is 12, and so on. Next, an experiment is conducted to obtain information about θ resulting in the random variable X with possible values coded as 0, 1, and frequency function $p(x, \theta)$ given by the following table

TABLE 1.3.2. The frequency function $p(x, \theta_i); i = 1, 2$

		Rock formation	
		x	
		0	1
(Oil)	θ_1	0.3	0.7
(No oil)	θ_2	0.6	0.4

Thus, X may represent a certain geological formation, and when there is oil, it is known that formation 0 occurs with frequency 0.3 and formation 1 with frequency 0.7, whereas if there is no oil, formations 0 and 1 occur with frequencies 0.6 and 0.4. We list all possible decision rules in the following table.

TABLE 1.3.3. Possible decision rules $\delta_i(x)$

		i								
		1	2	3	4	5	6	7	8	9
$x = 0$		a_1	a_1	a_1	a_2	a_2	a_2	a_3	a_3	a_3
$x = 1$		a_1	a_2	a_3	a_1	a_2	a_3	a_1	a_2	a_3

Here, δ_1 represents “Take action a_1 regardless of the value of X ,” δ_2 corresponds to “Take action a_1 , if $X = 0$; take action a_2 , if $X = 1$,” and so on.

The risk of δ at θ is

$$\begin{aligned} R(\theta, \delta) &= E[l(\theta, \delta(X))] = l(\theta, a_1)P[\delta(X) = a_1] \\ &\quad + l(\theta, a_2)P[\delta(X) = a_2] + l(\theta, a_3)P[\delta(X) = a_3]. \end{aligned}$$

For instance,

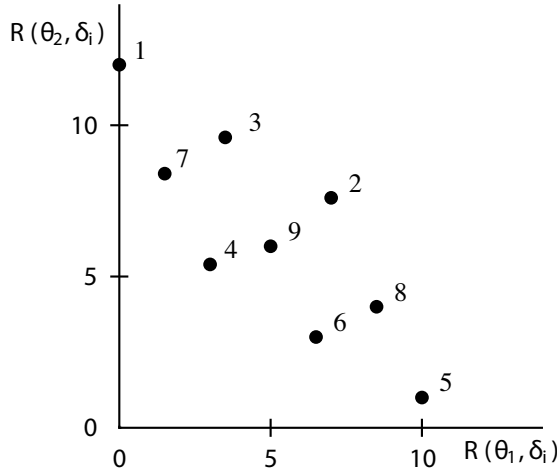
$$R(\theta_1, \delta_2) = 0(0.3) + 10(0.7) = 7$$

$$R(\theta_2, \delta_2) = 12(0.6) + 1(0.4) = 7.6.$$

If Θ is finite and has k members, we can represent the whole risk function of a procedure δ by a point in k -dimensional Euclidean space, $(R(\theta_1, \delta), \dots, R(\theta_k, \delta))$ and if $k = 2$ we can plot the set of all such points obtained by varying δ . The *risk points* $(R(\theta_1, \delta_i), R(\theta_2, \delta_i))$ are given in Table 1.3.4 and graphed in Figure 1.3.2 for $i = 1, \dots, 9$.

TABLE 1.3.4. Risk points $(R(\theta_1, \delta_i), R(\theta_2, \delta_i))$

i	1	2	3	4	5	6	7	8	9
$R(\theta_1, \delta_i)$	0	7	3.5	3	10	6.5	1.5	8.5	5
$R(\theta_2, \delta_i)$	12	7.6	9.6	5.4	1	3	8.4	4.0	6

**Figure 1.3.2.** The risk points $(R(\theta_1, \delta_i), R(\theta_2, \delta_i))$, $i = 1, \dots, 9$.

It remains to pick out the rules that are “good” or “best.” Criteria for doing this will be introduced in the next subsection. \square

1.3.3 Bayes and Minimax Criteria

The difficulties of comparing decision procedures have already been discussed in the special contexts of estimation and testing. We say that a procedure δ *improves* a procedure δ' if, and only if,

$$R(\theta, \delta) \leq R(\theta, \delta')$$

for all θ with strict inequality for some θ . It is easy to see that there is typically no rule δ that improves all others. For instance, in estimating $\theta \in R$ when $X \sim N(\theta, \sigma_0^2)$, if we ignore the data and use the estimate $\hat{\theta} = 0$, we obtain $MSE(\hat{\theta}) = \theta^2$. The absurd rule “ $\delta^*(X) = 0$ ” cannot be improved on at the value $\theta = 0$ because $E_0(\delta^2(X)) = 0$ if and only if $\delta(X) = 0$. Usually, if δ and δ' are two rules, neither improves the other. Consider, for instance, δ_4 and δ_6 in our example. Here $R(\theta_1, \delta_4) < R(\theta_1, \delta_6)$ but $R(\theta_2, \delta_4) > R(\theta_2, \delta_6)$.

The problem of selecting good decision procedures has been attacked in a variety of ways.

- (1) Narrow classes of procedures have been proposed using criteria such as considerations of symmetry, unbiasedness (for estimates and tests), or level of significance (for tests). Researchers have then sought procedures that improve all others within the class. We shall pursue this approach further in Chapter 3. Extensions of unbiasedness ideas may be found in Lehmann (1997, Section 1.5). Symmetry (or invariance) restrictions are discussed in Ferguson (1967).
- (2) A second major approach has been to compare risk functions by global criteria rather than on a pointwise basis. We shall discuss the Bayes and minimax criteria.

Bayes: The Bayesian point of view leads to a natural global criterion. Recall that in the Bayesian model θ is the realization of a random variable or vector $\boldsymbol{\theta}$ and that P_θ is the conditional distribution of \mathbf{X} given $\boldsymbol{\theta} = \theta$. In this framework $R(\theta, \delta)$ is just $E[l(\boldsymbol{\theta}, \delta(\mathbf{X})) \mid \boldsymbol{\theta} = \theta]$, the expected loss, if we use δ and $\boldsymbol{\theta} = \theta$. If we adopt the Bayesian point of view, we need not stop at this point, but can proceed to calculate what we expect to lose on the average as $\boldsymbol{\theta}$ varies. This quantity which we shall call the *Bayes risk of δ* and denote $r(\delta)$ is then, given by

$$r(\delta) = E[R(\boldsymbol{\theta}, \delta)] = E[l(\boldsymbol{\theta}, \delta(\mathbf{X}))]. \quad (1.3.9)$$

The second preceding identity is a consequence of the double expectation theorem (B.1.20) in Appendix B.

To illustrate, suppose that in the oil drilling example an expert thinks the chance of finding oil is .2. Then we treat the parameter as a random variable $\boldsymbol{\theta}$ with possible values θ_1, θ_2 and frequency function

$$\pi(\theta_1) = 0.2, \pi(\theta_2) = 0.8.$$

The Bayes risk of δ is, therefore,

$$r(\delta) = 0.2R(\theta_1, \delta) + 0.8R(\theta_2, \delta). \quad (1.3.10)$$

Table 1.3.5 gives $r(\delta_1), \dots, r(\delta_9)$ specified by (1.3.9).

TABLE 1.3.5. Bayes and maximum risks of the procedures of Table 1.3.3.

i	1	2	3	4	5	6	7	8	9
$r(\delta_i)$	9.6	7.48	8.38	4.92	2.8	3.7	7.02	4.9	5.8
$\max\{R(\theta_1, \delta_i), R(\theta_2, \delta_i)\}$	12	7.6	9.6	5.4	10	6.5	8.4	8.5	6

In the Bayesian framework δ is preferable to δ' if, and only if, it has smaller Bayes risk. If there is a rule δ^* , which attains the minimum Bayes risk, that is, such that

$$r(\delta^*) = \min_{\delta} r(\delta)$$

then it is called a *Bayes rule*. From Table 1.3.5 we see that rule δ_5 is the unique Bayes rule for our prior.

The method of computing Bayes procedures by listing all available δ and their Bayes risk is impracticable in general. We postpone the consideration of posterior analysis, the only reasonable computational method, to Section 3.2.

Note that the Bayes approach leads us to compare procedures on the basis of,

$$r(\delta) = \sum_{\theta} R(\theta, \delta) \pi(\theta),$$

if θ is discrete with frequency function $\pi(\theta)$, and

$$r(\delta) = \int R(\theta, \delta) \pi(\theta) d\theta,$$

if θ is continuous with density $\pi(\theta)$. Such comparisons make sense even if we do not interpret π as a prior density or frequency, but only as a weight function for averaging the values of the function $R(\theta, \delta)$. For instance, in Example 1.3.5 we might feel that both values of the risk were equally important. It is then natural to compare procedures using the simple average $\frac{1}{2}[R(\theta_1, \delta) + R(\theta_2, \delta)]$. But this is just Bayes comparison where π places equal probability on θ_1 and θ_2 .

Minimax: Instead of averaging the risk as the Bayesian does we can look at the worst possible risk. This is, we prefer δ to δ' , if and only if,

$$\sup_{\theta} R(\theta, \delta) < \sup_{\theta} R(\theta, \delta').$$

A procedure δ^* , which has

$$\sup_{\theta} R(\theta, \delta^*) = \inf_{\delta} \sup_{\theta} R(\theta, \delta),$$

is called *minimax* (minimizes the maximum risk).

The criterion comes from the general theory of two-person zero sum games of von Neumann.⁽²⁾ We briefly indicate “the game of decision theory.” Nature (Player I) picks a point $\theta \in \Theta$ independently of the statistician (Player II), who picks a decision procedure δ from \mathcal{D} , the set of all decision procedures. Player II then pays Player I, $R(\theta, \delta)$. The maximum risk of δ^* is the upper pure value of the game.

This criterion of optimality is very conservative. It aims to give maximum protection against the worst that can happen, Nature’s choosing a θ , which makes the risk as large as possible. The principle would be compelling, if the statistician believed that the parameter value is being chosen by a malevolent opponent who knows what decision procedure will be used. Of course, Nature’s intentions and degree of foreknowledge are not that clear and most statisticians find the minimax principle too conservative to employ as a general rule. Nevertheless, in many cases the principle can lead to very reasonable procedures.

To illustrate computation of the minimax rule we turn to Table 1.3.4. From the listing of $\max(R(\theta_1, \delta), R(\theta_2, \delta))$ we see that δ_4 is minimax with a maximum risk of 5.4.

Students of game theory will realize at this point that the statistician may be able to lower the maximum risk without requiring any further information by using a random

mechanism to determine which rule to employ. For instance, suppose that, in Example 1.3.5, we toss a fair coin and use δ_4 if the coin lands heads and δ_6 otherwise. Our expected risk would be,

$$\begin{aligned}\frac{1}{2}R(\theta, \delta_4) + \frac{1}{2}R(\theta, \delta_6) &= 4.75 \text{ if } \theta = \theta_1 \\ &= 4.20 \text{ if } \theta = \theta_2.\end{aligned}$$

The maximum risk 4.75 is strictly less than that of δ_4 .

Randomized decision rules: In general, if \mathcal{D} is the class of all decision procedures (nonrandomized), a *randomized decision procedure* can be thought of as a random experiment whose outcomes are members of \mathcal{D} . For simplicity we shall discuss only randomized procedures that select among a finite set $\delta_1, \dots, \delta_q$ of nonrandomized procedures. If the randomized procedure δ selects δ_i with probability λ_i , $i = 1, \dots, q$, $\sum_{i=1}^q \lambda_i = 1$, we then define

$$R(\theta, \delta) = \sum_{i=1}^q \lambda_i R(\theta, \delta_i). \quad (1.3.11)$$

Similarly we can define, given a prior π on Θ , the *Bayes risk* of δ

$$r(\delta) = \sum_{i=1}^q \lambda_i E[R(\theta, \delta_i)]. \quad (1.3.12)$$

A randomized Bayes procedure δ^* minimizes $r(\delta)$ among all randomized procedures. A randomized minimax procedure minimizes $\max_{\theta} R(\theta, \delta)$ among all randomized procedures.

We now want to study the relations between randomized and nonrandomized Bayes and minimax procedures in the context of Example 1.3.5. We will then indicate how much of what we learn carries over to the general case. As in Example 1.3.5, we represent the risk of any procedure δ by the vector $(R(\theta_1, \delta), R(\theta_2, \delta))$ and consider the *risk set*

$$S = \{(R(\theta_1, \delta), R(\theta_2, \delta)) : \delta \in \mathcal{D}^*\}$$

where \mathcal{D}^* is the set of all procedures, including randomized ones.

By (1.3.11),

$$S = \left\{ (r_1, r_2) : r_1 = \sum_{i=1}^9 \lambda_i R(\theta_1, \delta_i), r_2 = \sum_{i=1}^9 \lambda_i R(\theta_2, \delta_i), \lambda_i \geq 0, \sum_{i=1}^9 \lambda_i = 1 \right\}.$$

That is, S is the convex hull of the risk points $(R(\theta_1, \delta_i), R(\theta_2, \delta_i))$, $i = 1, \dots, 9$ (Figure 1.3.3).

If $\pi(\theta_1) = \gamma = 1 - \pi(\theta_2)$, $0 \leq \gamma \leq 1$, then all rules having Bayes risk c correspond to points in S that lie on the line

$$\gamma r_1 + (1 - \gamma) r_2 = c. \quad (1.3.13)$$

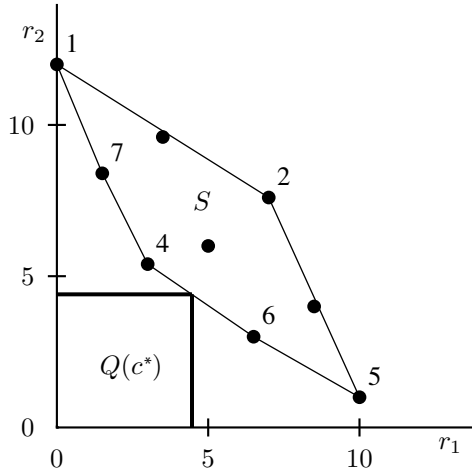


Figure 1.3.3. The convex hull S of the risk points $(R(\theta_1, \delta_i), R(\theta_2, \delta_i))$, $i = 1, \dots, 9$. The point where the square $Q(c^*)$ defined by (1.3.16) touches S is the risk point of the minimax rule.

As c varies, (1.3.13) defines a family of parallel lines with slope $-\gamma/(1-\gamma)$. Finding the Bayes rule corresponds to finding the smallest c for which the line (1.3.13) intersects S . This is that line with slope $-\gamma/(1-\gamma)$ that is tangent to S at the lower boundary of S . All points of S that are on the tangent are Bayes. Two cases arise:

- (1) The tangent has a unique point of contact with a risk point corresponding to a nonrandomized rule. For instance, when $\gamma = 0.2$, this point is $(10, 1)$, which is the risk point of the Bayes rule δ_5 (see Figure 1.3.3).
- (2) The tangent is the line connecting two “nonrandomized” risk points δ_i, δ_j . A point (r_1, r_2) on this line can be written

$$\begin{aligned} r_1 &= \lambda R(\theta_1, \delta_i) + (1-\lambda)R(\theta_1, \delta_j), \\ r_2 &= \lambda R(\theta_2, \delta_i) + (1-\lambda)R(\theta_2, \delta_j), \end{aligned} \quad (1.3.14)$$

where $0 \leq \lambda \leq 1$, and, thus, by (1.3.11) corresponds to the values

$$\begin{aligned} \delta &= \delta_i \text{ with probability } \lambda \\ &= \delta_j \text{ with probability } (1-\lambda), \quad 0 \leq \lambda \leq 1. \end{aligned} \quad (1.3.15)$$

Each one of these rules, as λ ranges from 0 to 1, is Bayes against π . We can choose two nonrandomized Bayes rules from this class, namely δ_i (take $\lambda = 1$) and δ_j (take $\lambda = 0$).

Because changing the prior π corresponds to changing the slope $-\gamma/(1-\gamma)$ of the line given by (1.3.13), the set B of all risk points corresponding to procedures Bayes with respect to some prior is just the lower left boundary of S (i.e., all points on the lower boundary of S that have as tangents the y axis or lines with nonpositive slopes).

To locate the risk point of the minimax rule consider the family of squares,

$$Q(c) = \{(r_1, r_2) : 0 \leq r_1 \leq c, 0 \leq r_2 \leq c\} \quad (1.3.16)$$

whose diagonal is the line $r_1 = r_2$. Let c^* be the smallest c for which $Q(c) \cap S \neq \emptyset$ (i.e., the first square that touches S). Then $Q(c^*) \cap S$ is either a point or a horizontal or vertical line segment. See Figure 1.3.3. It is the set of risk points of minimax rules because any point with smaller maximum risk would belong to $Q(c) \cap S$ with $c < c^*$ contradicting the choice of c^* . In our example, the first point of contact between the squares and S is the intersection between $r_1 = r_2$ and the line connecting the two points corresponding to δ_4 and δ_6 . Thus, the minimax rule is given by (1.3.14) with $i = 4, j = 6$ and λ the solution of

$$r_1 = \lambda R(\theta_1, \delta_4) + (1 - \lambda)R(\theta_1, \delta_6) = \lambda R(\theta_2, \delta_4) + (1 - \lambda)R(\theta_2, \delta_6) = r_2.$$

From Table 1.3.4, this equation becomes

$$3\lambda + 6.5(1 - \lambda) = 5.4\lambda + 3(1 - \lambda),$$

which yields $\lambda \cong 0.59$.

There is another important concept that we want to discuss in the context of the risk set. A decision rule δ is said to be *inadmissible* if there exists another rule δ' such that δ' improves δ . Naturally, all rules that are not inadmissible are called *admissible*. Using Table 1.3.4 we can see, for instance, that δ_2 is inadmissible because δ_4 improves it (i.e., $R(\theta_1, \delta_4) = 3 < 7 = R(\theta_1, \delta_2)$ and $R(\theta_2, \delta_4) = 5.4 < 7.6 = R(\theta_2, \delta_2)$).

To gain some insight into the class of all admissible procedures (randomized and non-randomized) we again use the risk set. A rule δ with risk point (r_1, r_2) is admissible, if and only if, there is no (x, y) in S such that $x \leq r_1$ and $y \leq r_2$, or equivalently, if and only if, $\{(x, y) : x \leq r_1, y \leq r_2\}$ has only (r_1, r_2) in common with S . From the figure it is clear that such points must be on the lower left boundary. In fact, the set of all lower left boundary points of S corresponds to the class of admissible rules and, thus, agrees with the set of risk points of Bayes procedures.

If Θ is finite, $\Theta = \{\theta_1, \dots, \theta_k\}$, we can define the risk set in general as

$$S = \{(R(\theta_1, \delta), \dots, R(\theta_k, \delta)) : \delta \in \mathcal{D}^*\}$$

where \mathcal{D}^* is the set of all randomized decision procedures. The following features exhibited by the risk set by Example 1.3.5 can be shown to hold generally (see Ferguson, 1967, for instance).

- (a) For any prior there is always a nonrandomized Bayes procedure, if there is a randomized one. Randomized Bayes procedures are mixtures of nonrandomized ones in the sense of (1.3.14).

- (b) The set B of risk points of Bayes procedures consists of risk points on the lower boundary of S whose tangent hyperplanes have normals pointing into the positive quadrant.
- (c) If Θ is finite and minimax procedures exist, they are Bayes procedures.
- (d) All admissible procedures are Bayes procedures.
- (e) If a Bayes prior has $\pi(\theta_i) > 0$ for all i , then any Bayes procedure corresponding to π is admissible.

If Θ is not finite there are typically admissible procedures that are not Bayes. However, under some conditions, all admissible procedures are either Bayes procedures or limits of Bayes procedures (in various senses). These remarkable results, at least in their original form, are due essentially to Wald. They are useful because the property of being Bayes is easier to analyze than admissibility.

Other theorems are available characterizing larger but more manageable classes of procedures, which include the admissible rules, at least when procedures with the same risk function are identified. An important example is the class of procedures that depend only on knowledge of a sufficient statistic (see Ferguson, 1967; Section 3.4). We stress that looking at randomized procedures is essential for these conclusions, although it usually turns out that all admissible procedures of interest are indeed nonrandomized. For more information on these topics, we refer to Blackwell and Girshick (1954) and Ferguson (1967).

Summary. We introduce the decision theoretic foundation of statistics including the notions of *action space*, *decision rule*, *loss function*, and *risk* through various examples including *estimation*, *testing*, *confidence bounds*, *ranking*, and *prediction*. The basic *bias-variance* decomposition of *mean square error* is presented. The basic global comparison criteria *Bayes* and *minimax* are presented as well as a discussion of optimality by restriction and notions of *admissibility*.

1.4 PREDICTION

The prediction Example 1.3.2 presented important situations in which a vector \mathbf{z} of covariates can be used to predict an unseen response Y . Here are some further examples of the kind of situation that prompts our study in this section. A college admissions officer has available the College Board scores at entrance and first-year grade point averages of freshman classes for a period of several years. Using this information, he wants to predict the first-year grade point averages of entering freshmen on the basis of their College Board scores. A stockholder wants to predict the value of his holdings at some time in the future on the basis of his past experience with the market and his portfolio. A meteorologist wants to estimate the amount of rainfall in the coming spring. A government expert wants to predict the amount of heating oil needed next winter. Similar problems abound in every field. The frame we shall fit them into is the following.

We assume that we know the joint probability distribution of a random vector (or variable) \mathbf{Z} and a random variable Y . We want to find a function g defined on the range of

\mathbf{Z} such that $g(\mathbf{Z})$ (the *predictor*) is “close” to Y . In terms of our preceding discussion, \mathbf{Z} is the information that we have and Y the quantity to be predicted. For example, in the college admissions situation, \mathbf{Z} would be the College Board score of an entering freshman and Y his or her first-year grade point average. The joint distribution of \mathbf{Z} and Y can be calculated (or rather well estimated) from the records of previous years that the admissions officer has at his disposal. Next we must specify what *close* means. One reasonable measure of “distance” is $(g(\mathbf{Z}) - Y)^2$, which is the *squared prediction error* when $g(\mathbf{Z})$ is used to predict Y . Since Y is not known, we turn to the *mean squared prediction error* (MSPE)

$$\Delta^2(Y, g(\mathbf{Z})) = E[g(\mathbf{Z}) - Y]^2$$

or its square root $\sqrt{E(g(\mathbf{Z}) - Y)^2}$. The MSPE is the measure traditionally used in the mathematical theory of prediction whose deeper results (see, for example, Grenander and Rosenblatt, 1957) presuppose it. The method that we employ to prove our elementary theorems *does* generalize to other measures of distance than $\Delta(Y, g(\mathbf{Z}))$ such as the mean absolute error $E(|g(\mathbf{Z}) - Y|)$ (Problems 1.4.7–11). Just how widely applicable the notions of this section are will become apparent in Remark 1.4.5 and Section 3.2 where the problem of MSPE prediction is identified with the optimal decision problem of Bayesian statistics with squared error loss.

The class \mathcal{G} of possible predictors g may be the nonparametric class \mathcal{G}_{NP} of all $g : R^d \rightarrow R$ or it may be to some subset of this class. See Remark 1.4.6. In this section we consider \mathcal{G}_{NP} and the class \mathcal{G}_L of linear predictors of the form $a + \sum_{j=1}^d b_j Z_j$.

We begin the search for the best predictor in the sense of minimizing MSPE by considering the case in which there is no covariate information, or equivalently, in which \mathbf{Z} is a constant; see Example 1.3.4. In this situation all predictors are constant and the best one is that number c_0 that minimizes $E(Y - c)^2$ as a function of c .

Lemma 1.4.1. $E(Y - c)^2$ is either ∞ for all c or is minimized uniquely by $c = \mu = E(Y)$. In fact, when $EY^2 < \infty$,

$$E(Y - c)^2 = \text{Var } Y + (c - \mu)^2. \quad (1.4.1)$$

Proof. $EY^2 < \infty$ if and only if $E(Y - c)^2 < \infty$ for all c ; see Problem 1.4.25. $EY^2 < \infty$ implies that μ exists, and by expanding

$$Y - c = (Y - \mu) + (\mu - c)$$

(1.4.1) follows because $E(Y - \mu) = 0$ makes the cross product term vanish. We see that $E(Y - c)^2$ has a unique minimum at $c = \mu$ and the lemma follows. \square

Now we can solve the problem of finding the best MSPE predictor of Y , given a vector \mathbf{Z} ; that is, we can find the g that minimizes $E(Y - g(\mathbf{Z}))^2$. By the substitution theorem for conditional expectations (B.1.16), we have

$$E[(Y - g(\mathbf{Z}))^2 \mid \mathbf{Z} = \mathbf{z}] = E[(Y - g(\mathbf{z}))^2 \mid \mathbf{Z} = \mathbf{z}]. \quad (1.4.2)$$

Let

$$\mu(\mathbf{z}) = E(Y \mid \mathbf{Z} = \mathbf{z}).$$

Because $g(\mathbf{z})$ is a constant, Lemma 1.4.1 assures us that

$$E[(Y - g(\mathbf{z}))^2 \mid \mathbf{Z} = \mathbf{z}] = E[(Y - \mu(\mathbf{z}))^2 \mid \mathbf{Z} = \mathbf{z}] + [g(\mathbf{z}) - \mu(\mathbf{z})]^2. \quad (1.4.3)$$

If we now take expectations of both sides and employ the double expectation theorem (B.1.20), we can conclude that

Theorem 1.4.1. *If \mathbf{Z} is any random vector and Y any random variable, then either $E(Y - g(\mathbf{Z}))^2 = \infty$ for every function g or*

$$E(Y - \mu(\mathbf{Z}))^2 \leq E(Y - g(\mathbf{Z}))^2 \quad (1.4.4)$$

for every g with strict inequality holding unless $g(\mathbf{Z}) = \mu(\mathbf{Z})$. That is, $\mu(\mathbf{Z})$ is the unique best MSPE predictor. In fact, when $E(Y^2) < \infty$,

$$E(Y - g(\mathbf{Z}))^2 = E(Y - \mu(\mathbf{Z}))^2 + E(g(\mathbf{Z}) - \mu(\mathbf{Z}))^2. \quad (1.4.5)$$

An important special case of (1.4.5) is obtained by taking $g(\mathbf{z}) = E(Y)$ for all \mathbf{z} . Write $\text{Var}(Y \mid \mathbf{z})$ for the variance of the condition distribution of Y given $\mathbf{Z} = \mathbf{z}$, that is, $\text{Var}(Y \mid \mathbf{z}) = E([Y - E(Y \mid \mathbf{z})]^2 \mid \mathbf{z})$, and recall (B.1.20), then (1.4.5) becomes,

$$\text{Var } Y = E(\text{Var}(Y \mid \mathbf{Z})) + \text{Var}(E(Y \mid \mathbf{Z})), \quad (1.4.6)$$

which is generally valid because if one side is infinite, so is the other.

Property (1.4.6) is linked to a notion that we now define: Two random variables U and V with $E|UV| < \infty$ are said to be *uncorrelated* if

$$E[V - E(V)][U - E(U)] = 0.$$

Equivalently U and V are uncorrelated if either $EV[U - E(U)] = 0$ or $EU[V - E(V)] = 0$. Let $\epsilon = Y - \mu(\mathbf{Z})$ denote the random prediction error, then we can write

$$Y = \mu(\mathbf{Z}) + \epsilon.$$

Proposition 1.4.1. *Suppose that $\text{Var } Y < \infty$, then*

- (a) ϵ is uncorrelated with every function of \mathbf{Z}
- (b) $\mu(\mathbf{Z})$ and ϵ are uncorrelated
- (c) $\text{Var}(Y) = \text{Var } \mu(\mathbf{Z}) + \text{Var } \epsilon$.

Proof. To show (a), let $h(\mathbf{Z})$ be any function of \mathbf{Z} , then by the iterated expectation theorem,

$$\begin{aligned} E\{h(\mathbf{Z})\epsilon\} &= E\{E[h(\mathbf{Z})\epsilon \mid \mathbf{Z}]\} \\ &= E\{h(\mathbf{Z})E[Y - \mu(\mathbf{Z}) \mid \mathbf{Z}]\} = 0 \end{aligned}$$

because $E[Y - \mu(\mathbf{Z}) \mid \mathbf{Z}] = \mu(\mathbf{Z}) - \mu(\mathbf{Z}) = 0$. Properties (b) and (c) follow from (a). \square

Note that Proposition 1.4.1(c) is equivalent to (1.4.6) and that (1.4.5) follows from (a) because (a) implies that the cross product term in the expansion of $E\{[Y - \mu(\mathbf{z})] + [\mu(\mathbf{z}) - g(\mathbf{z})]\}^2$ vanishes.

As a consequence of (1.4.6), we can derive the following theorem, which will prove of importance in estimation theory.

Theorem 1.4.2. *If $E(|Y|) < \infty$ but \mathbf{Z} and Y are otherwise arbitrary, then*

$$\text{Var}(E(Y \mid \mathbf{Z})) \leq \text{Var } Y. \quad (1.4.7)$$

If $\text{Var } Y < \infty$ strict inequality holds unless

$$Y = E(Y \mid \mathbf{Z}) \quad (1.4.8)$$

or equivalently unless Y is a function of \mathbf{Z} .

Proof. The assertion (1.4.7) follows immediately from (1.4.6). Equality in (1.4.7) can hold if, and only if,

$$E(\text{Var}(Y \mid \mathbf{Z})) = E(Y - E(Y \mid \mathbf{Z}))^2 = 0.$$

By (A.11.9) this can hold if, and only if, (1.4.8) is true. \square

Example 1.4.1. An assembly line operates either at full, half, or quarter capacity. Within any given month the capacity status does not change. Each day there can be 0, 1, 2, or 3 shutdowns due to mechanical failure. The following table gives the frequency function $p(z, y) = P(Z = z, Y = y)$ of the number of shutdowns Y and the capacity state Z of the line for a randomly chosen day. The row sums of the entries $p_Z(z)$ (given at the end of each row) represent the frequency with which the assembly line is in the appropriate capacity state, whereas the column sums $p_Y(y)$ yield the frequency of 0, 1, 2, or 3 failures among all days. We want to predict the number of failures for a given day knowing the state of the assembly line for the month. We find

$$E(Y \mid Z = 1) = \sum_{i=1}^3 iP[Y = i \mid Z = 1] = 2.45,$$

$$E(Y \mid Z = \frac{1}{2}) = 2.10, \quad E(Y \mid Z = \frac{1}{4}) = 1.20.$$

These fractional figures are not too meaningful as predictors of the natural number values of Y . But this predictor is also the right one, if we are trying to guess, as we reasonably might, the average number of failures per day in a given month. In this case if Y_i represents the number of failures on day i and Z the state of the assembly line, the best predictor is $E\left(30^{-1} \sum_{i=1}^{30} Y_i \mid Z\right) = E(Y \mid Z)$, also.

$z \backslash y$	$p(z, y)$				$p_Z(z)$
	0	1	2	3	
$\frac{1}{4}$	0.10	0.05	0.05	0.05	0.25
$\frac{1}{2}$	0.025	0.025	0.10	0.10	0.25
1	0.025	0.025	0.15	0.30	0.50
$p_Y(y)$	0.15	0.10	0.30	0.45	1

The MSPE of the best predictor can be calculated in two ways. The first is direct.

$$E(Y - E(Y | Z))^2 = \sum_z \sum_{y=0}^3 (y - E(Y | Z = z))^2 p(z, y) = 0.885.$$

The second way is to use (1.4.6) writing,

$$\begin{aligned}
 E(Y - E(Y | Z))^2 &= \text{Var } Y - \text{Var}(E(Y | Z)) \\
 &= E(Y^2) - E[(E(Y | Z))^2] \\
 &= \sum_y y^2 p_Y(y) - \sum_z [E(Y | Z = z)]^2 p_Z(z) \\
 &= 0.885
 \end{aligned}$$

as before. □

Example 1.4.2. *The Bivariate Normal Distribution. Regression toward the mean.* If (Z, Y) has a $\mathcal{N}(\mu_Z, \mu_Y, \sigma_Z^2, \sigma_Y^2, \rho)$ distribution, Theorem B.4.2 tells us that the conditional distribution of Y given $Z = z$ is $\mathcal{N}(\mu_Y + \rho(\sigma_Y/\sigma_Z)(z - \mu_Z), \sigma_Y^2(1 - \rho^2))$. Therefore, the best predictor of Y using Z is the linear function

$$\mu_0(Z) = \mu_Y + \rho(\sigma_Y/\sigma_Z)(Z - \mu_Z).$$

Because

$$E((Y - E(Y | Z = z))^2 | Z = z) = \sigma_Y^2(1 - \rho^2) \quad (1.4.9)$$

is independent of z , the MSPE of our predictor is given by,

$$E(Y - E(Y | Z))^2 = \sigma_Y^2(1 - \rho^2). \quad (1.4.10)$$

The qualitative behavior of this predictor and of its MSPE gives some insight into the structure of the bivariate normal distribution. If $\rho > 0$, the predictor is a monotone increasing function of Z indicating that large (small) values of Y tend to be associated with large (small) values of Z . Similarly, $\rho < 0$ indicates that large values of Z tend to go with small values of Y and we have negative dependence. If $\rho = 0$, the best predictor is just the constant μ_Y as we would expect in the case of independence. One minus the ratio of the MSPE of the best predictor of Y given Z to $\text{Var } Y$, which is the MSPE of the best constant predictor, can reasonably be thought of as a measure of dependence. The larger

this quantity the more dependent Z and Y are. In the bivariate normal case, this quantity is just ρ^2 . Thus, for this family of distributions the sign of the correlation coefficient gives the type of dependence between Z and Y , whereas its magnitude measures the degree of such dependence.

Because of (1.4.6) we can also write

$$\rho^2 = \frac{\text{Var } \mu_0(Z)}{\text{Var } Y}. \quad (1.4.11)$$

The line $y = \mu_Y + \rho(\sigma_Y/\sigma_Z)(z - \mu_Z)$, which corresponds to the best predictor of Y given Z in the bivariate normal model, is usually called the *regression* (line) of Y on Z . The term *regression* was coined by Francis Galton and is based on the following observation. Suppose Y and Z are bivariate normal random variables with the same mean μ , variance σ^2 , and positive correlation ρ . In Galton's case, these were the heights of a randomly selected father (Z) and his son (Y) from a large human population. Then the predicted height of the son, or the average height of sons whose fathers are the height Z , $(1 - \rho)\mu + \rho Z$, is closer to the population mean of heights μ than is the height of the father. Thus, tall fathers tend to have shorter sons; there is "regression toward the mean." This is compensated for by "progression" toward the mean among the sons of shorter fathers and there is no paradox. The variability of the predicted value about μ should, consequently, be less than that of the actual heights and indeed $\text{Var}((1 - \rho)\mu + \rho Z) = \rho^2\sigma^2$. Note that in practice, in particular in Galton's studies, the distribution of (Z, Y) is unavailable and the regression line is estimated on the basis of a sample $(Z_1, Y_1), \dots, (Z_n, Y_n)$ from the population. We shall see how to do this in Chapter 2. \square

Example 1.4.3. *The Multivariate Normal Distribution.* Let $\mathbf{Z} = (Z_1, \dots, Z_d)^T$ be a $d \times 1$ covariate vector with mean $\boldsymbol{\mu}_{\mathbf{Z}} = (\mu_1, \dots, \mu_d)^T$ and suppose that $(\mathbf{Z}^T, Y)^T$ has a $(d + 1)$ multivariate normal, $\mathcal{N}_{d+1}(\boldsymbol{\mu}, \Sigma)$, distribution (Section B.6) in which $\boldsymbol{\mu} = (\mu_{\mathbf{Z}}^T, \mu_Y)^T$, $\mu_Y = E(Y)$,

$$\Sigma = \begin{pmatrix} \Sigma_{\mathbf{Z}\mathbf{Z}} & \Sigma_{\mathbf{Z}Y} \\ \Sigma_{Y\mathbf{Z}} & \sigma_{YY} \end{pmatrix},$$

$\Sigma_{\mathbf{Z}\mathbf{Z}}$ is the $d \times d$ variance-covariance matrix $\text{Var}(\mathbf{Z})$,

$$\Sigma_{\mathbf{Z}Y} = (\text{Cov}(Z_1, Y), \dots, \text{Cov}(Z_d, Y))^T = \Sigma_{Y\mathbf{Z}}^T$$

and $\sigma_{YY} = \text{Var}(Y)$. Theorem B.6.5 states that the conditional distribution of Y given $\mathbf{Z} = \mathbf{z}$ is $\mathcal{N}(\mu_Y + (\mathbf{z} - \boldsymbol{\mu}_{\mathbf{Z}})^T \boldsymbol{\beta}, \sigma_{Y|Y|\mathbf{z}})$ where $\boldsymbol{\beta} = \Sigma_{\mathbf{Z}\mathbf{Z}}^{-1} \Sigma_{\mathbf{Z}Y}$ and $\sigma_{Y|Y|\mathbf{z}} = \sigma_{YY} - \Sigma_{Y\mathbf{Z}} \Sigma_{\mathbf{Z}\mathbf{Z}}^{-1} \Sigma_{\mathbf{Z}Y}$. Thus, the best predictor $E(Y | \mathbf{Z})$ of Y is the linear function

$$\mu_0(\mathbf{Z}) = \mu_Y + (\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}})^T \boldsymbol{\beta} \quad (1.4.12)$$

with MSPE

$$E[Y - \mu_0(\mathbf{Z})]^2 = E\{E[Y - \mu_0(\mathbf{Z})^2 | \mathbf{Z}]\} = E(\sigma_{Y|Y|\mathbf{Z}}) = \sigma_{YY} - \Sigma_{Y\mathbf{Z}} \Sigma_{\mathbf{Z}\mathbf{Z}}^{-1} \Sigma_{\mathbf{Z}Y}.$$

The quadratic form $\Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}$ is positive except when the joint normal distribution is degenerate, so the MSPE of $\mu_0(\mathbf{Z})$ is smaller than the MSPE of the constant predictor μ_Y . One minus the ratio of these MSPEs is a measure of how strongly the covariates are associated with Y . This quantity is called the *multiple correlation coefficient* (MCC), *coefficient of determination* or *population R-squared*. We write

$$MCC = \rho_{ZY}^2 = 1 - \frac{E[Y - \mu_0(\mathbf{Z})]^2}{\text{Var } Y} = \frac{\text{Var } \mu_0(\mathbf{Z})}{\text{Var } Y}$$

where the last identity follows from (1.4.6). By (1.4.11), the MCC equals the square of the usual correlation coefficient $\rho = \sigma_{ZY} / \sigma_Y^{\frac{1}{2}} \sigma_Z^{\frac{1}{2}}$ when $d = 1$.

For example, let Y and $\mathbf{Z} = (Z_1, Z_2)^T$ be the heights in inches of a 10-year-old girl and her parents (Z_1 = mother's height, Z_2 = father's height). Suppose⁽¹⁾ that $(\mathbf{Z}^T, Y)^T$ is trivariate normal with $\text{Var}(Y) = 6.39$

$$\Sigma_{\mathbf{ZZ}} = \begin{pmatrix} 7.74 & 2.92 \\ 2.92 & 6.67 \end{pmatrix}, \quad \Sigma_{\mathbf{ZY}} = (4.07, 2.98)^T.$$

Then the strength of association between a girl's height and those of her mother and father, respectively; and parents, are

$$\rho_{Z_1, Y}^2 = .335, \quad \rho_{Z_2, Y}^2 = .209, \quad \rho_{\mathbf{Z}, Y}^2 = .393.$$

In words, knowing the mother's height reduces the mean squared prediction error over the constant predictor by 33.5%. The percentage reductions knowing the father's and both parent's heights are 20.9% and 39.3%, respectively. In practice, when the distribution of $(\mathbf{Z}^T, Y)^T$ is unknown, the linear predictor $\mu_0(\mathbf{Z})$ and its MSPE will be estimated using a sample $(\mathbf{Z}_1^T, Y_1)^T, \dots, (\mathbf{Z}_n^T, Y_n)^T$. See Sections 2.1 and 2.2. \square

The best linear predictor. The problem of finding the best MSPE predictor is solved by Theorem 1.4.1. Two difficulties of the solution are that we need fairly precise knowledge of the joint distribution of \mathbf{Z} and Y in order to calculate $E(Y | \mathbf{Z})$ and that the best predictor may be a complicated function of \mathbf{Z} . If we are willing to sacrifice absolute excellence, we can avoid both objections by looking for a predictor that is best within a class of simple predictors. The natural class to begin with is that of linear combinations of components of \mathbf{Z} . We first do the one-dimensional case.

Let us call any random variable of the form $a + bZ$ a *linear predictor* and any such variable with $a = 0$ a *zero intercept linear predictor*. What is the best (zero intercept) linear predictor of Y in the sense of minimizing MSPE? The answer is given by:

Theorem 1.4.3. *Suppose that $E(Z^2)$ and $E(Y^2)$ are finite and Z and Y are not constant. Then the unique best zero intercept linear predictor is obtained by taking*

$$b = b_0 = \frac{E(ZY)}{E(Z^2)},$$

whereas the unique best linear predictor is $\mu_L(Z) = a_1 + b_1Z$ where

$$b_1 = \frac{\text{Cov}(Z, Y)}{\text{Var } Z}, \quad a_1 = E(Y) - b_1E(Z).$$

Proof. We expand $\{Y - bZ\}^2 = \{Y - [Z(b - b_0) + Zb_0]\}^2$ to get

$$E(Y - bZ)^2 = E(Y^2) + E(Z^2)(b - b_0)^2 - E(Z^2)b_0^2.$$

Therefore, $E(Y - bZ)^2$ is uniquely minimized by $b = b_0$, and

$$E(Y - b_0Z)^2 = E(Y^2) - \frac{[E(ZY)]^2}{E(Z^2)}. \quad (1.4.13)$$

To prove the second assertion of the theorem note that by (1.4.1),

$$E(Y - a - bZ)^2 = \text{Var}(Y - bZ) + (E(Y) - bE(Z) - a)^2.$$

Therefore, whatever be b , $E(Y - a - bZ)^2$ is uniquely minimized by taking

$$a = E(Y) - bE(Z).$$

Substituting this value of a in $E(Y - a - bZ)^2$ we see that the b we seek minimizes $E[(Y - E(Y)) - b(Z - E(Z))]^2$. We can now apply the result on zero intercept linear predictors to the variables $Z - E(Z)$ and $Y - E(Y)$ to conclude that b_1 is the unique minimizing value. \square

Remark 1.4.1. From (1.4.13) we obtain the proof of the Cauchy–Schwarz inequality (A.11.17) in the appendix. This is because $E(Y - b_0Z)^2 \geq 0$ is equivalent to the Cauchy–Schwarz inequality with equality holding if, and only if, $E(Y - b_0Z)^2 = 0$, which corresponds to $Y = b_0Z$. We could similarly obtain (A.11.16) directly by calculating $E(Y - a_1 - b_1Z)^2$. \square

Note that if $E(Y | Z)$ is of the form $a + bZ$, then $a = a_1$ and $b = b_1$, because, by (1.4.5), if the best predictor is linear, it must coincide with the best linear predictor. This is in accordance with our evaluation of $E(Y | Z)$ in Example 1.4.2. In that example nothing is lost by using linear prediction. On the other hand, in Example 1.4.1 the best linear predictor and best predictor differ (see Figure 1.4.1). A loss of about 5% is incurred by using the best linear predictor. That is,

$$\frac{E[Y - \mu_L(Z)]^2}{E[Y - \mu(Z)]^2} = 1.05.$$

Best Multivariate Linear Predictor. Our linear predictor is of the form

$$\mu_l(\mathbf{Z}) = a + \sum_{j=1}^d b_j Z_j = a + \mathbf{Z}^T \mathbf{b}$$

where $\mathbf{Z} = (Z_1, \dots, Z_d)^T$ and $\mathbf{b} = (b_1, \dots, b_d)^T$. Let

$$\beta = (E([\mathbf{Z} - E(\mathbf{Z})][\mathbf{Z} - E(\mathbf{Z})]^T))^{-1} E([\mathbf{Z} - E(\mathbf{Z})][Y - E(Y)]) = \Sigma_{\mathbf{Z}\mathbf{Z}}^{-1} \Sigma_{\mathbf{Z}Y}.$$

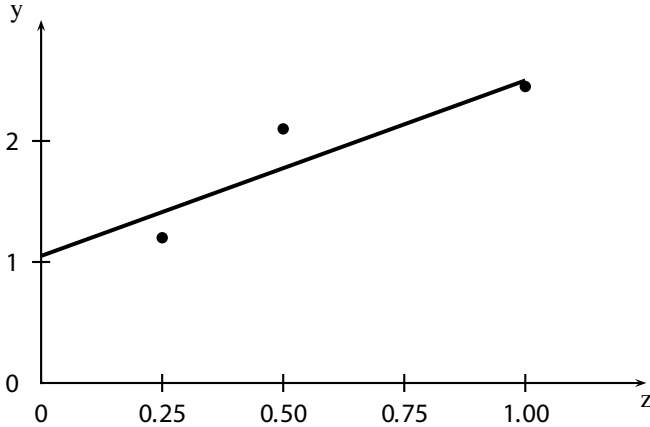


Figure 1.4.1. The three dots give the best predictor. The line represents the best linear predictor $y = 1.05 + 1.45z$.

Theorem 1.4.4. If EY^2 and $(E([\mathbf{Z} - E(\mathbf{Z})]^T[\mathbf{Z} - E(\mathbf{Z})]))^{-1}$ exist, then the unique best linear MSPE predictor is

$$\mu_L(\mathbf{Z}) = \mu_Y + (\mathbf{Z} - \mu_{\mathbf{Z}})^T \beta. \quad (1.4.14)$$

Proof. Note that $R(a, \mathbf{b}) \equiv E_p[Y - \mu_l(\mathbf{Z})]^2$ depends on the joint distribution P of $\mathbf{X} = (\mathbf{Z}^T, Y)^T$ only through the expectation μ and covariance Σ of \mathbf{X} . Let P_0 denote the multivariate normal, $\mathcal{N}(\mu, \Sigma)$, distribution and let $R_0(a, \mathbf{b}) = E_{P_0}[Y - \mu_l(\mathbf{Z})]^2$. By Example 1.4.3, $R_0(a, \mathbf{b})$ is minimized by (1.4.14). Because P and P_0 have the same μ and Σ , $R(a, \mathbf{b}) = R_0(a, \mathbf{b})$, and $R(a, \mathbf{b})$ is minimized by (1.4.14). \square

Remark 1.4.2. We could also have established Theorem 1.4.4 by extending the proof of Theorem 1.4.3 to $d > 1$. However, our new proof shows how second-moment results sometimes can be established by “connecting” them to the normal distribution. A third approach using calculus is given in Problem 1.4.19. \square

Remark 1.4.3. In the general, not necessarily normal, case the *multiple correlation coefficient* (MCC) or *coefficient of determination* is defined as the correlation between Y and the best linear predictor of Y ; that is,

$$\rho_{\mathbf{Z}Y}^2 = \text{Corr}^2(Y, \mu_L(\mathbf{Z})).$$

Thus, the MCC gives the strength of the *linear* relationship between \mathbf{Z} and Y . See Problem 1.4.17 for an overall measure of the strength of this relationship. \square

Remark 1.4.4. Suppose the model for $\mu(\mathbf{Z})$ is linear; that is,

$$\mu(\mathbf{Z}) = E(Y | \mathbf{Z}) = \alpha + \mathbf{Z}^T \beta$$

for unknown $\alpha \in R$ and $\beta \in R^d$. We want to express α and β in terms of moments of (\mathbf{Z}, Y) . Set $Z_0 = 1$. By Proposition 1.4.1(a), $\epsilon = Y - \mu(\mathbf{Z})$ and each of Z_0, \dots, Z_d are uncorrelated; thus,

$$E(Z_j[Y - (\alpha + \mathbf{Z}^T \beta)]) = 0, \quad j = 0, \dots, d. \quad (1.4.15)$$

Solving (1.4.15) for α and β gives (1.4.14) (Problem 1.4.23). Because the multivariate normal model is a linear model, this gives a new derivation of (1.4.12).

Remark 1.4.5. Consider the Bayesian model of Section 1.2 and the Bayes risk (1.3.8) defined by $r(\delta) = E[l(\theta, \delta(\mathbf{X}))]$. If we identify θ with Y and \mathbf{X} with \mathbf{Z} , we see that $r(\delta) = \text{MSPE}$ for squared error loss $l(\theta, \delta) = (\theta - \delta)^2$. Thus, the optimal MSPE predictor $E(\theta | \mathbf{X})$ is the Bayes procedure for squared error loss. We return to this in Section 3.2. \square

Remark 1.4.6. When the class \mathcal{G} of possible predictors g with $E|g(\mathbf{Z})| < \infty$ form a Hilbert space as defined in Section B.10 and there is a $g_0 \in \mathcal{G}$ such that

$$g_0 = \arg \inf \{ \Delta(Y, g(\mathbf{Z})) : g \in \mathcal{G} \},$$

then $g_0(\mathbf{Z})$ is called the *projection* of Y on the space \mathcal{G} of functions of \mathbf{Z} and we write $g_0(\mathbf{Z}) = \pi(Y | \mathcal{G})$. Moreover, $g(\mathbf{Z})$ and $h(\mathbf{Z})$ are said to be *orthogonal* if at least one has expected value zero and $E[g(\mathbf{Z})h(\mathbf{Z})] = 0$. With these concepts the results of this section are linked to the general Hilbert space results of Section B.10. Using the distance Δ and projection π notation, we can conclude that

$$\begin{aligned} \mu(\mathbf{Z}) &= \pi(Y | \mathcal{G}_{NP}), \quad \mu_L(\mathbf{Z}) = \pi(Y | \mathcal{G}_L) = \pi(\mu(\mathbf{Z}) | \mathcal{G}_L) \\ \Delta^2(Y, \mu_L(\mathbf{Z})) &= \Delta^2(\mu_L(\mathbf{Z}), \mu(\mathbf{Z})) + \Delta^2(Y, \mu(\mathbf{Z})) \end{aligned} \quad (1.4.16)$$

$$Y - \mu(\mathbf{Z}) \text{ is orthogonal to } \mu(\mathbf{Z}) \text{ and to } \mu_L(\mathbf{Z}). \quad (1.4.17)$$

Note that (1.4.16) is the Pythagorean identity. \square

Summary. We consider situations in which the goal is to predict the (perhaps in the future) value of a random variable Y . The notion of *mean squared prediction error* (MSPE) is introduced, and it is shown that if we want to predict Y on the basis of information contained in a random vector \mathbf{Z} , the optimal MSPE predictor is the conditional expected value of Y given \mathbf{Z} . The optimal MSPE predictor in the multivariate normal distribution is presented. It is shown to coincide with the optimal MSPE predictor when the model is left general but the class of possible predictors is restricted to be linear.

1.5 SUFFICIENCY

Once we have postulated a statistical model, we would clearly like to separate out any aspects of the data that are irrelevant in the context of the model and that may obscure our understanding of the situation.

We begin by formalizing what we mean by “a reduction of the data” $\mathbf{X} \in \mathcal{X}$. Recall that a *statistic* is any function of the observations generically denoted by $T(\mathbf{X})$ or T . The range of T is any space of objects \mathcal{T} , usually R or R^k , but as we have seen in Section 1.1.3, can also be a set of functions. If T assigns the same value to different sample points, then by recording or taking into account only the value of $T(\mathbf{X})$ we have a reduction of the data. Thus, $T(\mathbf{X}) = \bar{X}$ loses information about the X_i as soon as $n > 1$. Even $T(X_1, \dots, X_n) = (X_{(1)}, \dots, X_{(n)})$, loses information about the labels of the X_i . The idea of sufficiency is to reduce the data with statistics whose use involves no loss of information, in the context of a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$.

For instance, suppose that in Example 1.1.1 we had sampled the manufactured items in order, recording at each stage whether the examined item was defective or not. We could then represent the data by a vector $\mathbf{X} = (X_1, \dots, X_n)$ where $X_i = 1$ if the i th item sampled is defective and $X_i = 0$ otherwise. The total number of defective items observed, $T = \sum_{i=1}^n X_i$, is a statistic that maps many different values of (X_1, \dots, X_n) into the same number. However, it is intuitively clear that if we are interested in the proportion θ of defective items nothing is lost in this situation by recording and using only T .

One way of making the notion “a statistic whose use involves no loss of information” precise is the following. A statistic $T(\mathbf{X})$ is called *sufficient* for $P \in \mathcal{P}$ or the parameter θ if the conditional distribution of \mathbf{X} given $T(\mathbf{X}) = t$ does not involve θ . Thus, once the value of a sufficient statistic T is known, the sample $\mathbf{X} = (X_1, \dots, X_n)$ does not contain any further information about θ or equivalently P , *given that \mathcal{P} is valid*. We give a decision theory interpretation that follows. The most trivial example of a sufficient statistic is $T(\mathbf{X}) = \mathbf{X}$ because by any interpretation the conditional distribution of \mathbf{X} given $T(\mathbf{X}) = \mathbf{X}$ is point mass at \mathbf{X} .

Example 1.5.1. A machine produces n items in succession. Each item produced is good with probability θ and defective with probability $1 - \theta$, where θ is unknown. Suppose there is no dependence between the quality of the items produced and let $X_i = 1$ if the i th item is good and 0 otherwise. Then $\mathbf{X} = (X_1, \dots, X_n)$ is the record of n Bernoulli trials with probability θ . By (A.9.5),

$$P[X_1 = x_1, \dots, X_n = x_n] = \theta^t (1 - \theta)^{n-t} \quad (1.5.1)$$

where x_i is 0 or 1 and $t = \sum_{i=1}^n x_i$. By Example B.1.1, the conditional distribution of \mathbf{X} given $T = \sum_{i=1}^n X_i = t$ does not involve θ . Thus, T is a sufficient statistic for θ . \square

Example 1.5.2. Suppose that arrival of customers at a service counter follows a Poisson process with arrival rate (parameter) θ . Let X_1 be the time of arrival of the first customer, X_2 the time between the arrival of the first and second customers. By (A.16.4), X_1 and X_2 are independent and identically distributed exponential random variables with parameter θ . We prove that $T = X_1 + X_2$ is sufficient for θ . Begin by noting that according to Theorem B.2.3, whatever be θ , $X_1/(X_1 + X_2)$ and $X_1 + X_2$ are independent and the first of these statistics has a uniform distribution on $(0, 1)$. Therefore, the conditional distribution of $X_1/(X_1 + X_2)$ given $X_1 + X_2 = t$ is $\mathcal{U}(0, 1)$ whatever be t . Using our discussion in Section B.1.1 we see that given $X_1 + X_2 = t$, the conditional distribution of $X_1 = [X_1/(X_1 + X_2)](X_1 + X_2)$ and that of $X_1 t/(X_1 + X_2)$ are the same and we can conclude

that given $X_1 + X_2 = t$, X_1 has a $\mathcal{U}(0, t)$ distribution. It follows that, when $X_1 + X_2 = t$, whatever be θ , (X_1, X_2) is conditionally distributed as (X, Y) where X is uniform on $(0, t)$ and $Y = t - X$. Thus, $X_1 + X_2$ is sufficient. \square

In both of the foregoing examples considerable reduction has been achieved. Instead of keeping track of several numbers, we need only record one. Although the sufficient statistics we have obtained are “natural,” it is important to notice that there are many others that will do the same job. Being told that the numbers of successes in five trials is three is the same as knowing that the difference between the numbers of successes and the number of failures is one. More generally, if T_1 and T_2 are any two statistics such that $T_1(\mathbf{x}) = T_1(\mathbf{y})$ if and only if $T_2(\mathbf{x}) = T_2(\mathbf{y})$, then T_1 and T_2 provide the same information and achieve the same reduction of the data. Such statistics are called *equivalent*.

In general, checking sufficiency directly is difficult because we need to compute the conditional distribution. Fortunately, a simple necessary and sufficient criterion for a statistic to be sufficient is available. This result was proved in various forms by Fisher, Neyman, and Halmos and Savage. It is often referred to as the *factorization theorem* for sufficient statistics.

Theorem 1.5.1. *In a regular model, a statistic $T(\mathbf{X})$ with range \mathcal{T} is sufficient for θ if, and only if, there exists a function $g(t, \theta)$ defined for t in \mathcal{T} and θ in Θ and a function h defined on \mathcal{X} such that*

$$p(\mathbf{x}, \theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x}) \quad (1.5.2)$$

for all $\mathbf{x} \in \mathcal{X}$, $\theta \in \Theta$.

We shall give the proof in the discrete case. The complete result is established for instance by Lehmann (1997, Section 2.6).

Proof. Let $(\mathbf{x}_1, \mathbf{x}_2, \dots)$ be the set of possible realizations of \mathbf{X} and let $t_i = T(\mathbf{x}_i)$. Then T is discrete and $\sum_{i=1}^{\infty} P_{\theta}[T = t_i] = 1$ for every θ . To prove the sufficiency of (1.5.2), we need only show that $P_{\theta}[\mathbf{X} = \mathbf{x}_j | T = t_i]$ is independent of θ for every i and j . By our definition of conditional probability in the discrete case, it is enough to show that $P_{\theta}[\mathbf{X} = \mathbf{x}_j | T = t_i]$ is independent of θ on each of the sets $S_i = \{\theta : P_{\theta}[T = t_i] > 0\}$, $i = 1, 2, \dots$. Now, if (1.5.2) holds,

$$P_{\theta}[T = t_i] = \sum_{\{\mathbf{x}: T(\mathbf{x})=t_i\}} p(\mathbf{x}, \theta) = g(t_i, \theta) \sum_{\{\mathbf{x}: T(\mathbf{x})=t_i\}} h(\mathbf{x}). \quad (1.5.3)$$

By (B.1.1) and (1.5.2), for $\theta \in S_i$,

$$\begin{aligned} P_{\theta}[\mathbf{X} = \mathbf{x}_j | T = t_i] &= P_{\theta}[\mathbf{X} = \mathbf{x}_j, T = t_i] / P_{\theta}[T = t_i] \\ &= \frac{p(\mathbf{x}_j, \theta)}{P_{\theta}[T = t_i]} \\ &= \frac{g(t_i, \theta)h(\mathbf{x}_j)}{P_{\theta}[T = t_i]} \text{ if } T(\mathbf{x}_j) = t_i \\ &= 0 \text{ if } T(\mathbf{x}_j) \neq t_i. \end{aligned} \quad (1.5.4)$$

Applying (1.5.3) we arrive at,

$$\begin{aligned} P_\theta[\mathbf{X} = \mathbf{x}_j | T = t_i] &= 0 \text{ if } T(\mathbf{x}_j) \neq t_i \\ &= \frac{h(\mathbf{x}_j)}{\sum_{\{\mathbf{x}_k: T(\mathbf{x}_k) = t_i\}} h(\mathbf{x}_k)} \text{ if } T(\mathbf{x}_j) = t_i. \end{aligned} \quad (1.5.5)$$

Therefore, T is sufficient. Conversely, if T is sufficient, let

$$g(t_i, \theta) = P_\theta[T = t_i], h(\mathbf{x}) = P_\theta[\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t_i] \quad (1.5.6)$$

Then

$$p(\mathbf{x}, \theta) = P_\theta[\mathbf{X} = \mathbf{x}, T = T(\mathbf{x})] = g(T(\mathbf{x}), \theta)h(\mathbf{x}) \quad (1.5.7)$$

by (B.1.3). \square

Example 1.5.2 (continued). If X_1, \dots, X_n are the interarrival times for n customers, then the joint density of (X_1, \dots, X_n) is given by (see (A.16.4)),

$$p(x_1, \dots, x_n, \theta) = \theta^n \exp\left[-\theta \sum_{i=1}^n x_i\right] \quad (1.5.8)$$

if all the x_i are > 0 , and $p(x_1, \dots, x_n, \theta) = 0$ otherwise. We may apply Theorem 1.5.1 to conclude that $T(X_1, \dots, X_n) = \sum_{i=1}^n X_i$ is sufficient. Take $g(t, \theta) = \theta^n e^{-\theta t}$ if $t > 0$, $\theta > 0$, and $h(x_1, \dots, x_n) = 1$ if all the x_i are > 0 , and both functions $= 0$ otherwise. A whole class of distributions, which admits simple sufficient statistics and to which this example belongs, are introduced in the next section. \square

Example 1.5.3. Estimating the Size of a Population. Consider a population with θ members labeled consecutively from 1 to θ . The population is sampled with replacement and n members of the population are observed and their labels X_1, \dots, X_n are recorded. Common sense indicates that to get information about θ , we need only keep track of $X_{(n)} = \max(X_1, \dots, X_n)$. In fact, we can show that $X_{(n)}$ is sufficient. The probability distribution of \mathbf{X} is given by

$$p(x_1, \dots, x_n, \theta) = \theta^{-n} \quad (1.5.9)$$

if every x_i is an integer between 1 and θ and $p(x_1, \dots, x_n, \theta) = 0$ otherwise. Expression (1.5.9) can be rewritten as

$$p(x_1, \dots, x_n, \theta) = \theta^{-n} 1\{x_{(n)} \leq \theta\}, \quad (1.5.10)$$

where $x_{(n)} = \max(x_1, \dots, x_n)$. By Theorem 1.5.1, $X_{(n)}$ is a sufficient statistic for θ . \square

Example 1.5.4. Let X_1, \dots, X_n be independent and identically distributed random variables each having a normal distribution with mean μ and variance σ^2 , both of which are

unknown. Let $\theta = (\mu, \sigma^2)$. Then the density of (X_1, \dots, X_n) is given by

$$\begin{aligned} p(x_1, \dots, x_n, \theta) &= [2\pi\sigma^2]^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\} \\ &= [2\pi\sigma^2]^{-n/2} \left[\exp\left\{-\frac{n\mu^2}{2\sigma^2}\right\}\right] \left[\exp\left\{-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i\right)\right\}\right]. \end{aligned} \quad (1.5.11)$$

Evidently $p(x_1, \dots, x_n, \theta)$ is itself a function of $(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$ and θ only and upon applying Theorem 1.5.1 we can conclude that

$$T(X_1, \dots, X_n) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$$

is sufficient for θ . An equivalent sufficient statistic in this situation that is frequently used is

$$S(X_1, \dots, X_n) = \left[(1/n) \sum_{i=1}^n X_i, [1/(n-1)] \sum_{i=1}^n (X_i - \bar{X})^2\right],$$

where $\bar{X} = (1/n) \sum_{i=1}^n X_i$. The first and second components of this vector are called the *sample mean* and the *sample variance*, respectively. \square

Example 1.5.5. Suppose, as in Example 1.1.4 with $d = 2$, that Y_1, \dots, Y_n are independent, $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, with μ_i following the linear regression model

$$\mu_i = \beta_1 + \beta_2 z_i, \quad i = 1, \dots, n,$$

where we assume that the given constants $\{z_i\}$ are not all identical. Then $\theta = (\beta_1, \beta_2, \sigma^2)^T$ is identifiable (Problem 1.1.9) and

$$p(\mathbf{y}, \theta) = (2\pi\sigma^2)^{-n/2} \exp\left\{\frac{-\sum(\beta_1 + \beta_2 z_i)^2}{2\sigma^2}\right\} \exp\left\{\frac{-\sum Y_i^2 + 2\beta_1 \sum Y_i + 2\beta_2 \sum z_i Y_i}{2\sigma^2}\right\}.$$

Thus, $\mathbf{T} = (\sum Y_i, \sum Y_i^2, \sum z_i Y_i)$ is sufficient for θ . \square

Sufficiency and decision theory

Sufficiency can be given a clear operational interpretation in the decision theoretic setting. Specifically, if $T(\mathbf{X})$ is sufficient, we can, for any decision procedure $\delta(\mathbf{x})$, find a randomized decision rule $\delta^*(T(\mathbf{X}))$ depending only on $T(\mathbf{X})$ that does as well as $\delta(\mathbf{X})$ in the sense of having the same risk function; that is,

$$R(\theta, \delta) = R(\theta, \delta^*) \text{ for all } \theta. \quad (1.5.12)$$

By *randomized* we mean that $\delta^*(T(\mathbf{X}))$ can be generated from the value t of $T(\mathbf{X})$ and a random mechanism not depending on θ .

Here is an example.

Example 1.5.6. Suppose X_1, \dots, X_n are independent identically $\mathcal{N}(\theta, 1)$ distributed. Then

$$p(\mathbf{x}, \theta) = \exp\left\{n\theta\left(\bar{x} - \frac{1}{2}\theta\right)\right\} (2\pi)^{-\frac{1}{2}n} \exp\left\{-\frac{1}{2} \sum x_i^2\right\}$$

By the factorization theorem, \bar{X} is sufficient. Let $\delta(\mathbf{X}) = X_1$. Using only \bar{X} , we construct a rule $\delta^*(\mathbf{X})$ with the same risk = mean squared error as $\delta(\mathbf{X})$ as follows: Conditionally, given $\bar{X} = t$, choose $T^* = \delta^*(\mathbf{X})$ from the normal $\mathcal{N}(t, \frac{n-1}{n})$ distribution. Using Section B.1 and (1.4.6), we find

$$E(T^*) = E[E(T^*|\bar{X})] = E(\bar{X}) = \theta = E(X_1)$$

$$\text{Var}(T^*) = E[\text{Var}(T^*|\bar{X})] + \text{Var}[E(T^*|\bar{X})] = \frac{n-1}{n} + \frac{1}{n} = 1 = \text{Var}(X_1).$$

Thus, $\delta^*(\mathbf{X})$ and $\delta(\mathbf{X})$ have the same mean squared error. \square

The proof of (1.5.12) follows along the lines of the preceding example: Given $T(\mathbf{X}) = t$, the distribution of $\delta(\mathbf{X})$ does not depend on θ . Now draw δ^* randomly from this conditional distribution. This $\delta^*(T(\mathbf{X}))$ will have the same risk as $\delta(\mathbf{X})$ because, by the double expectation theorem,

$$R(\theta, \delta^*) = E\{E[\ell(\theta, \delta^*(T))|T]\} = E\{E[\ell(\theta, \delta(\mathbf{X}))|T]\} = R(\theta, \delta).$$

\square

Sufficiency and Bayes models

There is a natural notion of sufficiency of a statistic T in the Bayesian context where in addition to the model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ we postulate a prior distribution Π for Θ .

In Example 1.2.1 (Bernoulli trials) we saw that the posterior distribution given $\mathbf{X} = \mathbf{x}$ is the same as the posterior distribution given $T(\mathbf{X}) = \sum_{i=1}^n X_i = k$, where $k = \sum_{i=1}^n X_i$. In this situation we call T Bayes sufficient.

Definition. $T(\mathbf{X})$ is *Bayes sufficient* for Π if the posterior distribution of θ given $\mathbf{X} = \mathbf{x}$ is the same as the posterior (conditional) distribution of θ given $T(\mathbf{X}) = T(\mathbf{x})$ for all \mathbf{x} .

Equivalently, θ and \mathbf{X} are independent given $T(\mathbf{X})$.

Theorem 1.5.2. (Kolmogorov). *If $T(\mathbf{X})$ is sufficient for θ , it is Bayes sufficient for every Π .*

This result and a partial converse is the subject of Problem 1.5.14.

Minimal sufficiency

For any model there are many sufficient statistics: Thus, if X_1, \dots, X_n is a $\mathcal{N}(\mu, \sigma^2)$ sample $n \geq 2$, then $T(\mathbf{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ and $S(\mathbf{X}) = (X_1, \dots, X_n)$ are both sufficient. But $T(\mathbf{X})$ provides a greater reduction of the data. We define the statistic $T(\mathbf{X})$ to be *minimally sufficient* if it is sufficient and provides a greater reduction of the data

than any other sufficient statistic $S(\mathbf{X})$, in that, we can find a transformation r such that $T(\mathbf{X}) = r(S(\mathbf{X}))$.

Example 1.5.1 (continued). In this Bernoulli trials case, $T = \sum_{i=1}^n X_i$ was shown to be sufficient. Let $S(\mathbf{X})$ be any other sufficient statistic. Then by the factorization theorem we can write $p(\mathbf{x}, \theta)$ as

$$p(\mathbf{x}, \theta) = g(S(\mathbf{x}), \theta)h(\mathbf{x})$$

Combining this with (1.5.1), we find

$$\theta^T (1 - \theta)^{n-T} = g(S(\mathbf{x}), \theta)h(\mathbf{x}) \text{ for all } \theta.$$

For any two fixed θ_1 and θ_2 , the ratio of both sides of the foregoing gives

$$(\theta_1/\theta_2)^T [(1 - \theta_1)/(1 - \theta_2)]^{n-T} = g(S(\mathbf{x}), \theta_1)/g(S(\mathbf{x}), \theta_2).$$

In particular, if we set $\theta_1 = 2/3$ and $\theta_2 = 1/3$, take the log of both sides of this equation and solve for T , we find

$$T = r(S(\mathbf{x})) = \{\log[2^n g(S(\mathbf{x}), 2/3)/g(S(\mathbf{x}), 1/3)]\}/2 \log 2.$$

Thus, T is minimally sufficient. □

The likelihood function

The preceding example shows how we can use $p(\mathbf{x}, \theta)$ for different values of θ and the factorization theorem to establish that a sufficient statistic is minimally sufficient. We define the likelihood function L for a given observed data vector \mathbf{x} as

$$L_{\mathbf{x}}(\theta) = p(\mathbf{x}, \theta), \theta \in \Theta.$$

Thus, $L_{\mathbf{x}}$ is a map from the sample space \mathcal{X} to the class \mathcal{T} of functions $\{\theta \rightarrow p(\mathbf{x}, \theta) : \mathbf{x} \in \mathcal{X}\}$. It is a statistic whose values are functions; if $\mathbf{X} = \mathbf{x}$, the statistic L takes on the value $L_{\mathbf{x}}$. In the discrete case, for a given θ , $L_{\mathbf{x}}(\theta)$ gives the probability of observing the point \mathbf{x} . In the continuous case it is approximately proportional to the probability of observing a point in a small rectangle around \mathbf{x} . However, when we think of $L_{\mathbf{x}}(\theta)$ as a function of θ , it gives, for a given observed \mathbf{x} , the “likelihood” or “plausibility” of various θ . The formula (1.2.8) for the posterior distribution can then be remembered as Posterior \propto (Prior) \times (Likelihood) where the sign \propto denotes proportionality as functions of θ .

Example 1.5.4 (continued). In this $\mathcal{N}(\mu, \sigma^2)$ example, the likelihood function (1.5.11) is determined by the two-dimensional sufficient statistic

$$T = (T_1, T_2) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right).$$

Set $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$, then

$$L_{\mathbf{x}}(\theta) = (2\pi\theta_2)^{-n/2} \exp\left\{-\frac{n\theta_1^2}{2\theta_2}\right\} \exp\left\{-\frac{1}{2\theta_2}(t_2 - 2\theta_1 t_1)\right\}.$$

Now, as a function of θ , $L_{\mathbf{x}}(\cdot)$ determines (t_1, t_2) because, for example,

$$t_2 = -2 \log L_{\mathbf{x}}(0, 1) - n \log 2\pi$$

with a similar expression for t_1 in terms of $L_{\mathbf{x}}(0, 1)$ and $L_{\mathbf{x}}(1, 1)$ (Problem 1.5.17). Thus, L is a statistic that is equivalent to (t_1, t_2) and, hence, itself sufficient. By arguing as in Example 1.5.1 (continued) we can show that T and, hence, L is minimal sufficient. \square

In fact, a statistic closely related to L solves the minimal sufficiency problem in general. Suppose there exists θ_0 such that

$$\{\mathbf{x} : p(\mathbf{x}, \theta) > 0\} \subset \{\mathbf{x} : p(\mathbf{x}, \theta_0) > 0\}$$

for all θ . Let $\Lambda_{\mathbf{x}} = \frac{L_{\mathbf{x}}}{L_{\mathbf{x}}(\theta_0)}$. Thus, $\Lambda_{\mathbf{x}}$ is the function valued statistic that at θ takes on the value $\frac{p(\mathbf{x}, \theta)}{p(\mathbf{x}, \theta_0)}$, the *likelihood ratio* of θ to θ_0 . Then $\Lambda_{\mathbf{x}}$ is minimal sufficient. See Problem 1.5.12 for a proof of this theorem of Dynkin, Lehmann, and Scheffé.

The “irrelevant” part of the data

We can always rewrite the original \mathbf{X} as $(T(\mathbf{X}), S(\mathbf{X}))$ where $S(\mathbf{X})$ is a statistic needed to uniquely determine \mathbf{x} once we know the sufficient statistic $T(\mathbf{x})$. For instance, if $T(\mathbf{X}) = \bar{X}$ we can take $S(\mathbf{X}) = (X_1 - \bar{X}, \dots, X_n - \bar{X})$, the *residuals*; or if $T(\mathbf{X}) = (X_{(1)}, \dots, X_{(n)})$, the order statistics, $S(\mathbf{X}) = (R_1, \dots, R_n)$, the *ranks*, where $R_i = \sum_{j=1}^n 1(X_j \leq X_i)$. $S(\mathbf{X})$ becomes irrelevant (*ancillary*) for inference if $T(\mathbf{X})$ is known *but only if \mathcal{P} is valid*. Thus, in Example 1.5.5, if $\sigma^2 = 1$ is postulated, \bar{X} is sufficient, but if in fact $\sigma^2 \neq 1$ all information about σ^2 is contained in the residuals. If, as in the Example 1.5.4, σ^2 is assumed unknown, $(\bar{X}, \sum_{i=1}^n (X_i - \bar{X})^2)$ is sufficient, but if in fact the common distribution of the observations is not Gaussian all the information needed to estimate this distribution is contained in the corresponding $S(\mathbf{X})$ —see Problem 1.5.13. If \mathcal{P} specifies that X_1, \dots, X_n are a random sample, $(X_{(1)}, \dots, X_{(n)})$ is sufficient. But the ranks are needed if we want to look for possible dependencies in the observations as in Example 1.1.5.

Summary. Consider an experiment with observation vector $\mathbf{X} = (X_1, \dots, X_n)$. Suppose that \mathbf{X} has distribution in the class $\mathcal{P} = \{P_{\theta} : \theta \in \Theta\}$. We say that a statistic $T(\mathbf{X})$ is *sufficient* for $P \in \mathcal{P}$, or for the parameter θ , if the conditional distribution of \mathbf{X} given $T(\mathbf{X}) = t$ does not involve θ . Let $p(\mathbf{X}, \theta)$ denote the frequency function or density of \mathbf{X} . The *factorization theorem* states that $T(\mathbf{X})$ is sufficient for θ if and only if there exist functions $g(t, \theta)$ and $h(\mathbf{X})$ such that

$$p(\mathbf{X}, \theta) = g(T(\mathbf{X}), \theta)h(\mathbf{X}).$$

We show the following result: If $T(\mathbf{X})$ is sufficient for θ , then for any decision procedure $\delta(\mathbf{X})$, we can find a randomized decision rule $\delta^*(T(\mathbf{X}))$ depending only on the value of $t = T(\mathbf{X})$ and not on θ such that δ and δ^* have identical risk functions. We define a statistic $T(\mathbf{X})$ to be *Bayes sufficient* for a prior π if the posterior distribution of θ given $\mathbf{X} = \mathbf{x}$ is the same as the posterior distribution of θ given $T(\mathbf{X}) = T(\mathbf{x})$ for all \mathbf{X} . If

$T(\mathbf{X})$ is sufficient for θ , it is Bayes sufficient for θ . A sufficient statistic $T(\mathbf{X})$ is *minimally sufficient* for θ if for any other sufficient statistic $S(\mathbf{X})$ we can find a transformation r such that $T(\mathbf{X}) = r(S(\mathbf{X}))$. The likelihood function is defined for a given data vector of observations \mathbf{X} to be the function of θ defined by $L_{\mathbf{X}}(\theta) = p(\mathbf{X}, \theta)$, $\theta \in \Theta$. If $T(\mathbf{X})$ is sufficient for θ , and if there is a value $\theta_0 \in \Theta$ such that

$$\{\mathbf{x} : p(\mathbf{x}, \theta) > 0\} \subset \{\mathbf{x} : p(\mathbf{x}, \theta_0) > 0\}, \quad \theta \in \Theta,$$

then, by the factorization theorem, the *likelihood ratio*

$$\Lambda_{\mathbf{X}}(\theta) = \frac{L_{\mathbf{X}}(\theta)}{L_{\mathbf{X}}(\theta_0)}$$

depends on \mathbf{X} through $T(\mathbf{X})$ only. $\Lambda_{\mathbf{X}}(\theta)$ is a minimally sufficient statistic.

1.6 EXPONENTIAL FAMILIES

The binomial and normal models considered in the last section exhibit the interesting feature that there is a natural sufficient statistic whose dimension as a random vector is independent of the sample size. The class of families of distributions that we introduce in this section was first discovered in statistics independently by Koopman, Pitman, and Darmois through investigations of this property⁽¹⁾. Subsequently, many other common features of these families were discovered and they have become important in much of the modern theory of statistics.

Probability models with these common features include normal, binomial, Poisson, gamma, beta, and multinomial regression models used to relate a response variable Y to a set of predictor variables. More generally, these families form the basis for an important class of models called generalized linear models. We return to these models in Chapter 2. They will reappear in several connections in this book.

1.6.1 The One-Parameter Case

The family of distributions of a model $\{P_{\theta} : \theta \in \Theta\}$, is said to be a *one-parameter exponential family*, if there exist real-valued functions $\eta(\theta)$, $B(\theta)$ on Θ , real-valued functions T and h on R^q , such that the density (frequency) functions $p(x, \theta)$ of the P_{θ} may be written

$$p(x, \theta) = h(x) \exp\{\eta(\theta)T(x) - B(\theta)\} \quad (1.6.1)$$

where $x \in \mathcal{X} \subset R^q$. Note that the functions η , B , and T are not unique.

In a one-parameter exponential family the random variable $T(X)$ is sufficient for θ . This is clear because we need only identify $\exp\{\eta(\theta)T(x) - B(\theta)\}$ with $g(T(x), \theta)$ and $h(x)$ with itself in the factorization theorem. We shall refer to T as a *natural sufficient statistic* of the family.

Here are some examples.

Example 1.6.1. *The Poisson Distribution.* Let P_θ be the Poisson distribution with unknown mean θ . Then, for $x \in \{0, 1, 2, \dots\}$,

$$p(x, \theta) = \frac{\theta^x e^{-\theta}}{x!} = \frac{1}{x!} \exp\{x \log \theta - \theta\}, \quad \theta > 0. \quad (1.6.2)$$

Therefore, the P_θ form a one-parameter exponential family with

$$q = 1, \eta(\theta) = \log \theta, B(\theta) = \theta, T(x) = x, h(x) = \frac{1}{x!}. \quad (1.6.3)$$

Example 1.6.2. *The Binomial Family.* Suppose X has a $\mathcal{B}(n, \theta)$ distribution, $0 < \theta < 1$. Then, for $x \in \{0, 1, \dots, n\}$

$$\begin{aligned} p(x, \theta) &= \binom{n}{x} \theta^x (1 - \theta)^{n-x} \\ &= \binom{n}{x} \exp\left[x \log\left(\frac{\theta}{1 - \theta}\right) + n \log(1 - \theta)\right]. \end{aligned} \quad (1.6.4)$$

Therefore, the family of distributions of X is a one-parameter exponential family with

$$q = 1, \eta(\theta) = \log\left(\frac{\theta}{1 - \theta}\right), B(\theta) = -n \log(1 - \theta), T(x) = x, h(x) = \binom{n}{x}. \quad (1.6.5)$$

□

Here is an example where $q = 2$.

Example 1.6.3. Suppose $X = (Z, Y)^T$ where $Y = Z + \theta W$, $\theta > 0$, Z and W are independent $\mathcal{N}(0, 1)$. Then

$$\begin{aligned} f(x, \theta) &= f(z, y, \theta) = f(z) f_\theta(y | z) = \varphi(z) \theta^{-1} \varphi((y - z)\theta^{-1}) \\ &= (2\pi\theta)^{-1} \exp\left\{-\frac{1}{2}[z^2 + (y - z)^2\theta^{-2}]\right\} \\ &= (2\pi)^{-1} \exp\left\{-\frac{1}{2}z^2\right\} \exp\left\{-\frac{1}{2}\theta^{-2}(y - z)^2 - \log \theta\right\}. \end{aligned}$$

This is a one-parameter exponential family distribution with

$$q = 2, \eta(\theta) = -\frac{1}{2}\theta^{-2}, B(\theta) = \log \theta, T(x) = (y - z)^2, h(x) = (2\pi)^{-1} \exp\left\{-\frac{1}{2}z^2\right\}.$$

□

The families of distributions obtained by sampling from one-parameter exponential families are themselves one-parameter exponential families. Specifically, suppose X_1, \dots, X_m are independent and identically distributed with common distribution P_θ ,

where the P_θ form a one-parameter exponential family as in (1.6.1). If $\{P_\theta^{(m)}\}$, $\theta \in \Theta$, is the family of distributions of $\mathbf{X} = (X_1, \dots, X_m)$ considered as a random vector in R^{mq} and $p(\mathbf{x}, \theta)$ are the corresponding density (frequency) functions, we have

$$\begin{aligned} p(\mathbf{x}, \theta) &= \prod_{i=1}^m h(x_i) \exp[\eta(\theta)T(x_i) - B(\theta)] \\ &= \left[\prod_{i=1}^m h(x_i) \right] \exp \left[\eta(\theta) \sum_{i=1}^m T(x_i) - mB(\theta) \right] \end{aligned} \quad (1.6.6)$$

where $\mathbf{x} = (x_1, \dots, x_m)$. Therefore, the $P_\theta^{(m)}$ form a one-parameter exponential family. If we use the superscript m to denote the corresponding T , η , B , and h , then $q^{(m)} = mq$, and

$$\begin{aligned} \eta^{(m)}(\theta) &= \eta(\theta), \\ T^{(m)}(\mathbf{x}) &= \sum_{i=1}^m T(x_i), B^{(m)}(\theta) = mB(\theta), h^{(m)}(\mathbf{x}) = \prod_{i=1}^m h(x_i). \end{aligned} \quad (1.6.7)$$

Note that the natural sufficient statistic $T^{(m)}$ is one-dimensional whatever be m . For example, if $\mathbf{X} = (X_1, \dots, X_m)$ is a vector of independent and identically distributed $\mathcal{P}(\theta)$ random variables and $P_\theta^{(m)}$ is the family of distributions of \mathbf{x} , then the $P_\theta^{(m)}$ form a one-parameter exponential family with natural sufficient statistic $T^{(m)}(\mathbf{X}) = \sum_{i=1}^m T(X_i)$.

Some other important examples are summarized in the following table. We leave the proof of these assertions to the reader.

TABLE 1.6.1

Family of distributions		$\eta(\theta)$	$T(x)$
$\mathcal{N}(\mu, \sigma^2)$	σ^2 fixed	μ/σ^2	x
	μ fixed	$-1/2\sigma^2$	$(x - \mu)^2$
$\Gamma(p, \lambda)$	p fixed	$-\lambda$	x
	λ fixed	$(p - 1)$	$\log x$
$\beta(r, s)$	r fixed	$(s - 1)$	$\log(1 - x)$
	s fixed	$(r - 1)$	$\log x$

The statistic $T^{(m)}(X_1, \dots, X_m)$ corresponding to the one-parameter exponential family of distributions of a sample from any of the foregoing is just $\sum_{i=1}^m T(X_i)$.

In our first Example 1.6.1 the sufficient statistic $T^{(m)}(X_1, \dots, X_m) = \sum_{i=1}^m X_i$ is distributed as $\mathcal{P}(m\theta)$. This family of Poisson distributions is one-parameter exponential whatever be m . In the discrete case we can establish the following general result.

Theorem 1.6.1. Let $\{P_\theta\}$ be a one-parameter exponential family of discrete distributions with corresponding functions T , η , B , and h , then the family of distributions of the statistic $T(X)$ is a one-parameter exponential family of discrete distributions whose frequency

functions may be written

$$h^*(t) \exp\{\eta(\theta)t - B(\theta)\}$$

for suitable h^* .

Proof. By definition,

$$\begin{aligned} P_\theta[T(x) = t] &= \sum_{\{x: T(x)=t\}} p(x, \theta) \\ &= \sum_{\{x: T(x)=t\}} h(x) \exp[\eta(\theta)T(x) - B(\theta)] \\ &= \exp[\eta(\theta)t - B(\theta)] \left\{ \sum_{\{x: T(x)=t\}} h(x) \right\}. \end{aligned} \quad (1.6.8)$$

If we let $h^*(t) = \sum_{\{x: T(x)=t\}} h(x)$, the result follows. \square

A similar theorem holds in the continuous case if the distributions of $T(X)$ are themselves continuous.

Canonical exponential families. We obtain an important and useful reparametrization of the exponential family (1.6.1) by letting the model be indexed by η rather than θ . The exponential family then has the form

$$q(x, \eta) = h(x) \exp[\eta T(x) - A(\eta)], \quad x \in \mathcal{X} \subset R^q \quad (1.6.9)$$

where $A(\eta) = \log \int \cdots \int h(x) \exp[\eta T(x)] dx$ in the continuous case and the integral is replaced by a sum in the discrete case. If $\theta \in \Theta$, then $A(\eta)$ must be finite, if q is definable. Let \mathcal{E} be the collection of all η such that $A(\eta)$ is finite. Then as we show in Theorem 1.6.3, \mathcal{E} is either an interval or all of R and the class of models (1.6.9) with $\eta \in \mathcal{E}$ contains the class of models with $\theta \in \Theta$. The model given by (1.6.9) with η ranging over \mathcal{E} is called the *canonical one-parameter exponential family generated by T and h* . \mathcal{E} is called the *natural parameter space* and T is called the *natural sufficient statistic*.

Example 1.6.1. (continued). The Poisson family in canonical form is

$$q(x, \eta) = (1/x!) \exp\{\eta x - \exp[\eta]\}, \quad x \in \{0, 1, 2, \dots\},$$

where $\eta = \log \theta$,

$$\exp\{A(\eta)\} = \sum_{x=0}^{\infty} (e^{\eta x}/x!) = \sum_{x=0}^{\infty} (e^\eta)^x / x! = \exp(e^\eta),$$

and $\mathcal{E} = R$. \square

Here is a useful result.

Theorem 1.6.2. If X is distributed according to (1.6.9) and η is an interior point of \mathcal{E} , the moment-generating function of $T(X)$ exists and is given by

$$M(s) = \exp[A(s + \eta) - A(\eta)]$$

for s in some neighborhood of 0.

Moreover,

$$E(T(X)) = A'(\eta), \quad \text{Var}(T(X)) = A''(\eta).$$

Proof. We give the proof in the continuous case. We compute

$$\begin{aligned} M(s) &= E(\exp(sT(X))) = \int \cdots \int h(x) \exp[(s + \eta)T(x) - A(\eta)] dx \\ &= \{\exp[A(s + \eta) - A(\eta)]\} \int \cdots \int h(x) \exp[(s + \eta)T(x) - A(s + \eta)] dx \\ &= \exp[A(s + \eta) - A(\eta)] \end{aligned}$$

because the last factor, being the integral of a density, is one. The rest of the theorem follows from the moment generating property of $M(s)$ (see Section A.12). \square

Here is a typical application of this result.

Example 1.6.4 Suppose X_1, \dots, X_n is a sample from a population with density

$$p(x, \theta) = (x/\theta^2) \exp(-x^2/2\theta^2), \quad x > 0, \theta > 0.$$

This is known as the *Rayleigh* distribution. It is used to model the density of “time until failure” for certain types of equipment. Now

$$\begin{aligned} p(\mathbf{x}, \theta) &= \left(\prod_{i=1}^n (x_i/\theta^2) \right) \exp\left(-\sum_{i=1}^n x_i^2/2\theta^2\right) \\ &= \left(\prod_{i=1}^n x_i \right) \exp\left[\frac{-1}{2\theta^2} \sum_{i=1}^n x_i^2 - n \log \theta^2\right]. \end{aligned}$$

Here $\eta = -1/2\theta^2$, $\theta^2 = -1/2\eta$, $B(\theta) = n \log \theta^2$ and $A(\eta) = -n \log(-2\eta)$. Therefore, the natural sufficient statistic $\sum_{i=1}^n X_i^2$ has mean $-n/\eta = 2n\theta^2$ and variance $n/\eta^2 = 4n\theta^4$. Direct computation of these moments is more complicated. \square

1.6.2 The Multiparameter Case

Our discussion of the “natural form” suggests that one-parameter exponential families are naturally indexed by a one-dimensional real parameter η and admit a one-dimensional sufficient statistic $T(x)$. More generally, Koopman, Pitman, and Darmois were led in their investigations to the following family of distributions, which is naturally indexed by a k -dimensional parameter and admit a k -dimensional sufficient statistic.

A family of distributions $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, $\Theta \subset R^k$, is said to be a *k-parameter exponential family*, if there exist real-valued functions η_1, \dots, η_k and B of $\boldsymbol{\theta}$, and real-valued functions T_1, \dots, T_k , h on R^q such that the density (frequency) functions of the $P_{\boldsymbol{\theta}}$ may be written as,

$$p(x, \boldsymbol{\theta}) = h(x) \exp\left[\sum_{j=1}^k \eta_j(\boldsymbol{\theta}) T_j(x) - B(\boldsymbol{\theta})\right], \quad x \in \mathcal{X} \subset R^q. \quad (1.6.10)$$

By Theorem 1.5.1, the vector $\mathbf{T}(X) = (T_1(X), \dots, T_k(X))^T$ is sufficient. It will be referred to as a *natural sufficient statistic* of the family.

Again, suppose $\mathbf{X} = (X_1, \dots, X_m)^T$ where the X_i are independent and identically distributed and their common distribution ranges over a k -parameter exponential family given by (1.6.10). Then the distributions of \mathbf{X} form a k -parameter exponential family with natural sufficient statistic

$$\mathbf{T}^{(m)}(\mathbf{X}) = \left(\sum_{i=1}^m T_1(X_i), \dots, \sum_{i=1}^m T_k(X_i) \right)^T.$$

Example 1.6.5. The Normal Family. Suppose that $P_\theta = \mathcal{N}(\mu, \sigma^2)$, $\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$. The density of P_θ may be written as

$$p(x, \theta) = \exp\left[\frac{\mu}{\sigma^2}x - \frac{x^2}{2\sigma^2} - \frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right], \quad (1.6.11)$$

which corresponds to a two-parameter exponential family with $q = 1$, $\theta_1 = \mu$, $\theta_2 = \sigma^2$, and

$$\begin{aligned} \eta_1(\theta) &= \frac{\mu}{\sigma^2}, \quad T_1(x) = x, \quad \eta_2(\theta) = -\frac{1}{2\sigma^2}, \quad T_2(x) = x^2, \\ B(\theta) &= \frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2)\right), \quad h(x) = 1. \end{aligned}$$

If we observe a sample $\mathbf{X} = (X_1, \dots, X_m)$ from a $\mathcal{N}(\mu, \sigma^2)$ population, then the preceding discussion leads us to the natural sufficient statistic

$$\left(\sum_{i=1}^m X_i, \sum_{i=1}^m X_i^2 \right)^T,$$

which we obtained in the previous section (Example 1.5.4). □

Again it will be convenient to consider the “biggest” families, letting the model be indexed by $\boldsymbol{\eta} = (\eta_1, \dots, \eta_k)^T$ rather than θ . Thus, *the canonical k -parameter exponential family generated by \mathbf{T} and h is*

$$q(x, \boldsymbol{\eta}) = h(x) \exp\{\mathbf{T}^T(x)\boldsymbol{\eta} - A(\boldsymbol{\eta})\}, \quad x \in \mathcal{X} \subset R^q$$

where $\mathbf{T}(x) = (T_1(x), \dots, T_k(x))^T$ and, in the continuous case,

$$A(\boldsymbol{\eta}) = \log \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x) \exp\{\mathbf{T}^T(x)\boldsymbol{\eta}\} dx.$$

In the discrete case, $A(\boldsymbol{\eta})$ is defined in the same way except integrals over R^q are replaced by sums. In either case, we define the natural parameter space as

$$\mathcal{E} = \{\boldsymbol{\eta} \in R^k : -\infty < A(\boldsymbol{\eta}) < \infty\}.$$

Example 1.6.5. ($\mathcal{N}(\mu, \sigma^2)$ continued). In this example, $k = 2$, $\mathbf{T}^T(x) = (x, x^2) = (T_1(x), T_2(x))$, $\eta_1 = \mu/\sigma^2$, $\eta_2 = -1/2\sigma^2$, $A(\boldsymbol{\eta}) = \frac{1}{2}[(-\eta_1^2/2\eta_2) + \log(\pi/(-\eta_2))]$, $h(x) = 1$ and $\mathcal{E} = R \times R^- = \{(\eta_1, \eta_2) : \eta_1 \in R, \eta_2 < 0\}$.

Example 1.6.6. *Linear Regression.* Suppose as in Examples 1.1.4 and 1.5.5 that Y_1, \dots, Y_n are independent, $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, with $\mu_i = \beta_1 + \beta_2 z_i$, $i = 1, \dots, n$. From Example 1.5.5, the density of $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ can be put in canonical form with $k = 3$, $\mathbf{T}(\mathbf{Y}) = (\Sigma Y_i, \Sigma z_i Y_i, \Sigma Y_i^2)^T$, $\eta_1 = \beta_1/\sigma^2$, $\eta_2 = \beta_2/\sigma^2$, $\eta_3 = -1/2\sigma^2$,

$$A(\boldsymbol{\eta}) = \frac{-n}{4\eta_3}[\eta_1^2 + \hat{m}_2\eta_2^2 + \bar{z}\eta_1\eta_2 + 2\log(\pi/-\eta_3)],$$

and $\mathcal{E} = \{(\eta_1, \eta_2, \eta_3) : \eta_1 \in R, \eta_2 \in R, \eta_3 < 0\}$, where $\hat{m}_2 = n^{-1}\Sigma z_i^2$.

Example 1.6.7. *Multinomial Trials.* We observe the outcomes of n independent trials where each trial can end up in one of k possible categories. We write the outcome vector as $\mathbf{X} = (X_1, \dots, X_n)^T$ where the X_i are i.i.d. as X and the sample space of each X_i is the k categories $\{1, 2, \dots, k\}$. Let $T_j(\mathbf{x}) = \sum_{i=1}^n 1[X_i = j]$, and $\lambda_j = P(X_i = j)$. Then $p(\mathbf{x}, \boldsymbol{\lambda}) = \prod_{j=1}^k \lambda_j^{T_j(\mathbf{x})}$, $\boldsymbol{\lambda} \in \Lambda$, where Λ is the simplex $\{\boldsymbol{\lambda} \in R^k : 0 < \lambda_j < 1, j = 1, \dots, k, \sum_{j=1}^k \lambda_j = 1\}$. It will often be more convenient to work with unrestricted parameters. In this example, we can achieve this by the reparametrization

$$\lambda_j = e^{\alpha_j} / \sum_{j=1}^k e^{\alpha_j}, j = 1, \dots, k, \boldsymbol{\alpha} \in R^k.$$

Now we can write the likelihood as

$$q_0(\mathbf{x}, \boldsymbol{\alpha}) = \exp\left\{\sum_{j=1}^k \alpha_j T_j(\mathbf{x}) - n \log \sum_{j=1}^k \exp(\alpha_j)\right\}.$$

This is a k -parameter canonical exponential family generated by T_1, \dots, T_k and $h(\mathbf{x}) = \prod_{i=1}^n 1[x_i \in \{1, \dots, k\}]$ with canonical parameter $\boldsymbol{\alpha}$ and $\mathcal{E} = R^k$. However $\boldsymbol{\alpha}$ is not identifiable because $q_0(\mathbf{x}, \boldsymbol{\alpha} + c\mathbf{1}) = q_0(\mathbf{x}, \boldsymbol{\alpha})$ for $\mathbf{1} = (1, \dots, 1)^T$ and all c . This can be remedied by considering

$$\mathbf{T}_{(k-1)}(\mathbf{x}) \equiv (T_1(\mathbf{x}), \dots, T_{k-1}(\mathbf{x}))^T,$$

$\eta_j = \log(\lambda_j/\lambda_k) = \alpha_j - \alpha_k$, $1 \leq j \leq k-1$, and rewriting

$$q(\mathbf{x}, \boldsymbol{\eta}) = \exp\{\mathbf{T}_{(k-1)}^T(\mathbf{x})\boldsymbol{\eta} - n \log(1 + \sum_{j=1}^{k-1} e^{\eta_j})\}$$

where

$$\lambda_j = \frac{e^{\eta_j}}{1 + \sum_{j=1}^{k-1} e^{\eta_j}} = \frac{e^{\alpha_j}}{\sum_{j=1}^k e^{\alpha_j}}, \quad j = 1, \dots, k-1.$$

Note that $q(\mathbf{x}, \boldsymbol{\eta})$ is a $k - 1$ parameter canonical exponential family generated by $\mathbf{T}_{(k-1)}$ and $h(\mathbf{x}) = \prod_{i=1}^n 1[x_i \in \{1, \dots, k\}]$ with canonical parameter $\boldsymbol{\eta}$ and $\mathcal{E} = R^{k-1}$. Moreover, the parameters $\eta_j = \log(P_{\boldsymbol{\eta}}[X = j]/P_{\boldsymbol{\eta}}[X = k])$, $1 \leq j \leq k - 1$, are identifiable. Note that the model for \mathbf{X} is unchanged. \square

1.6.3 Building Exponential Families

Submodels

A *submodel* of a k -parameter canonical exponential family $\{q(\mathbf{x}, \boldsymbol{\eta}); \boldsymbol{\eta} \in \mathcal{E} \subset R^k\}$ is an exponential family defined by

$$p(x, \boldsymbol{\theta}) = q(x, \boldsymbol{\eta}(\boldsymbol{\theta})) \quad (1.6.12)$$

where $\boldsymbol{\theta} \in \Theta \subset R^l$, $l \leq k$, and $\boldsymbol{\eta}$ is a map from Θ to a subset of R^k . Thus, if X is discrete taking on k values as in Example 1.6.7 and $\mathbf{X} = (X_1, \dots, X_n)^T$ where the X_i are i.i.d. as X , then *all* models for \mathbf{X} are exponential families because they are submodels of the multinomial trials model.

Affine transformations

If \mathcal{P} is the canonical family generated by $\mathbf{T}_{k \times 1}$ and h and \mathbf{M} is the affine transformation from R^k to R^l defined by

$$\mathbf{M}(\mathbf{T}) = M_{\ell \times k} \mathbf{T} + \mathbf{b}_{\ell \times 1},$$

it is easy to see that the family generated by $\mathbf{M}(\mathbf{T}(X))$ and h is the subfamily of \mathcal{P} corresponding to

$$\Theta = [\boldsymbol{\eta}^{-1}](\mathcal{E}) \subset R^\ell$$

and

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = M^T \boldsymbol{\theta}.$$

Similarly, if $\Theta \subset R^\ell$ and $\boldsymbol{\eta}(\boldsymbol{\theta}) = B_{k \times \ell} \boldsymbol{\theta} \in R^k$, then the resulting submodel of \mathcal{P} above is a submodel of the exponential family generated by $B^T \mathbf{T}(X)$ and h . See Problem 1.6.17 for details. Here is an example of affine transformations of $\boldsymbol{\theta}$ and \mathbf{T} .

Example 1.6.8. Logistic Regression. Let Y_i be independent binomial, $\mathcal{B}(n_i, \lambda_i)$, $1 \leq i \leq n$. If the λ_i are unrestricted, $0 < \lambda_i < 1$, $1 \leq i \leq n$, this, from Example 1.6.2, is an n -parameter canonical exponential family with $\mathcal{Y}_i \equiv$ integers from 0 to n_i generated by $\mathbf{T}(Y_1, \dots, Y_n) = \mathbf{Y}$, $h(\mathbf{y}) = \prod_{i=1}^n \binom{n_i}{y_i} 1(0 \leq y_i \leq n_i)$. Here $\eta_i = \log \frac{\lambda_i}{1-\lambda_i}$, $A(\boldsymbol{\eta}) = \sum_{i=1}^n n_i \log(1 + e^{\eta_i})$. However, let $x_1 < \dots < x_n$ be specified levels and

$$\eta_i(\boldsymbol{\theta}) = \theta_1 + \theta_2 x_i, \quad 1 \leq i \leq n, \quad \boldsymbol{\theta} = (\theta_1, \theta_2)^T \in R^2. \quad (1.6.13)$$

This is a linear transformation $\boldsymbol{\eta}(\boldsymbol{\theta}) = B_{n \times 2} \boldsymbol{\theta}$ corresponding to $B_{n \times 2} = (\mathbf{1}, \mathbf{x})$, where $\mathbf{1}$ is $(1, \dots, 1)^T$, $\mathbf{x} = (x_1, \dots, x_n)^T$. Set $M = B^T$, then this is the two-parameter canonical exponential family generated by $M\mathbf{Y} = (\sum_{i=1}^n Y_i, \sum_{i=1}^n x_i Y_i)^T$ and h with

$$A(\theta_1, \theta_2) = \sum_{i=1}^n n_i \log(1 + \exp(\theta_1 + \theta_2 x_i)).$$

This model is sometimes applied in experiments to determine the toxicity of a substance. The Y_i represent the number of animals dying out of n_i when exposed to level x_i of the substance. It is assumed that each animal has a random toxicity threshold X such that death results if and only if a substance level on or above X is applied. Assume also:

- (a) No interaction between animals (independence) in relation to drug effects
- (b) The distribution of X in the animal population is *logistic*; that is,

$$P[X \leq x] = [1 + \exp\{-(\theta_1 + \theta_2 x)\}]^{-1}, \quad (1.6.14)$$

$\theta_1 \in R, \theta_2 > 0$. Then (and only then),

$$\log(P[X \leq x]/(1 - P[X \leq x])) = \theta_1 + \theta_2 x$$

and (1.6.13) holds. □

Curved exponential families

Exponential families (1.6.12) with the range of $\boldsymbol{\eta}(\boldsymbol{\theta})$ restricted to a subset of dimension l with $l \leq k - 1$, are called *curved exponential families* provided they do not form a canonical exponential family in the $\boldsymbol{\theta}$ parametrization.

Example 1.6.9. *Gaussian with Fixed Signal-to-Noise Ratio.* In the normal case with X_1, \dots, X_n i.i.d. $\mathcal{N}(\mu, \sigma^2)$, suppose the ratio $|\mu|/\sigma$, which is called the *coefficient of variation* or *signal-to-noise ratio*, is a known constant $\lambda_0 > 0$. Then, with $\theta = \mu$, we can write

$$p(\mathbf{x}, \theta) = \exp \left\{ \lambda_0^2 \theta^{-1} T_1 - \frac{1}{2} \lambda_0^2 \theta^{-2} T_2 - \frac{1}{2} n [\lambda_0^2 + \log(2\pi \lambda_0^{-2} \theta^2)] \right\}$$

where $T_1 = \sum_{i=1}^n x_i$, $T_2 = \sum_{i=1}^n x_i^2$, $\eta_1(\theta) = \lambda_0^2 \theta^{-1}$ and $\eta_2(\theta) = -\frac{1}{2} \lambda_0^2 \theta^{-2}$. This is a curved exponential family with $l = 1$. □

In Example 1.6.8, the $\boldsymbol{\theta}$ parametrization has dimension 2, which is less than $k = n$ when $n > 3$. However, $p(\mathbf{x}, \boldsymbol{\theta})$ in the $\boldsymbol{\theta}$ parametrization is a canonical exponential family, so it is not a curved family.

Example 1.6.10. *Location-Scale Regression.* Suppose that Y_1, \dots, Y_n are independent, $Y_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. If each μ_i ranges over R and each σ_i^2 ranges over $(0, \infty)$, this is by Example 1.6.5 a $2n$ -parameter canonical exponential family model with $\eta_i = \mu_i/\sigma_i^2$, and $\eta_{n+i} = -1/2\sigma_i^2$, $i = 1, \dots, n$, generated by

$$T(\mathbf{Y}) = (Y_1, \dots, Y_n, Y_1^2, \dots, Y_n^2)^T$$

and $h(\mathbf{Y}) = 1$. Next suppose that (μ_i, σ_i^2) depend on the value z_i of some covariate, say,

$$\mu_i = \theta_1 + \theta_2 z_i, \sigma_i^2 = \theta_3 (\theta_1 + \theta_2 z_i)^2, z_1 < \cdots < z_n$$

for unknown parameters $\theta_1 \in R$, $\theta_2 \in R$, $\theta_3 > 0$ (e.g., Bickel, 1978; Carroll and Ruppert, 1988, Sections 2.1–2.5; and Snedecor and Cochran, 1989, Section 15.10). For $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$, the map $\boldsymbol{\eta}(\boldsymbol{\theta})$ is

$$\eta_i(\boldsymbol{\theta}) = \theta_3^{-1} (\theta_1 + \theta_2 z_i)^{-1}, \eta_{n+i}(\boldsymbol{\theta}) = \frac{1}{2} \theta_3^{-1} (\theta_1 + \theta_2 z_i)^{-2}, i = 1, \dots, n.$$

Because $\sum_{i=1}^n \eta_i(\boldsymbol{\theta}) Y_i + \sum_{i=1}^n \eta_{n+i}(\boldsymbol{\theta}) Y_i^2$ cannot be written in the form $\sum_{j=1}^3 \eta_j^*(\boldsymbol{\theta}) T_j^*(\mathbf{Y})$ for some $\eta_j^*(\boldsymbol{\theta})$, $T_j^*(\mathbf{Y})$, then $p(\mathbf{y}, \boldsymbol{\theta}) = q(\mathbf{y}, \boldsymbol{\eta}(\boldsymbol{\theta}))$ as defined in (6.1.12) is not an exponential family model, but a curved exponential family model with $l = 3$. \square

Models in which the variance $\text{Var}(Y_i)$ depends on i are called *heteroscedastic* whereas models in which $\text{Var}(Y_i)$ does not depend on i are called *homoscedastic*. Thus, Examples 1.6.10 and 1.6.6 are heteroscedastic and homoscedastic models, respectively.

We return to curved exponential family models in Section 2.3.

Supermodels

We have already noted that the exponential family structure is preserved under i.i.d. sampling. Even more is true. Let Y_j , $1 \leq j \leq n$, be independent, $Y_j \in \mathcal{Y}_j \subset R^q$, with an exponential family density

$$q_j(y_j, \boldsymbol{\theta}) = \exp\{\mathbf{T}_j^T(y_j) \boldsymbol{\eta}(\boldsymbol{\theta}) - B_j(\boldsymbol{\theta})\} h_j(y_j), \boldsymbol{\theta} \in \Theta \subset R^k.$$

Then $\mathbf{Y} \equiv (Y_1, \dots, Y_n)^T$ is modeled by the exponential family generated by $\mathbf{T}(\mathbf{Y}) = \sum_{j=1}^n \mathbf{T}_j(Y_j)$ and $\prod_{j=1}^n h_j(y_j)$, with parameter $\boldsymbol{\eta}(\boldsymbol{\theta})$, and $B(\boldsymbol{\theta}) = \sum_{j=1}^n B_j(\boldsymbol{\theta})$.

In Example 1.6.8 note that (1.6.13) exhibits Y_j as being distributed according to a two-parameter family generated by $T_j(Y_j) = (Y_j, x_j Y_j)$ and we can apply the supermodel approach to reach the same conclusion as before.

1.6.4 Properties of Exponential Families

Theorem 1.6.1 generalizes directly to k -parameter families as does its continuous analogue. We extend the statement of Theorem 1.6.2.

Recall from Section B.5 that for any random vector $\mathbf{T}_{k \times 1}$, we define

$$M(\mathbf{s}) \equiv E e^{\mathbf{s}^T \mathbf{T}}$$

as the moment-generating function, and

$$E(\mathbf{T}) \equiv (E(T_1), \dots, E(T_k))^T$$

$$\text{Var}(\mathbf{T}) = \|\text{Cov}(T_a, T_b)\|_{k \times k}.$$

Theorem 1.6.3. *Let \mathcal{P} be a canonical k -parameter exponential family generated by (\mathbf{T}, h) with corresponding natural parameter space \mathcal{E} and function $A(\boldsymbol{\eta})$. Then*

- (a) \mathcal{E} is convex
- (b) $A : \mathcal{E} \rightarrow R$ is convex
- (c) *If \mathcal{E} has nonempty interior \mathcal{E}^0 in R^k and $\boldsymbol{\eta}_0 \in \mathcal{E}^0$, then $\mathbf{T}(X)$ has under $\boldsymbol{\eta}_0$ a moment-generating function M given by*

$$M(\mathbf{s}) = \exp\{A(\boldsymbol{\eta}_0 + \mathbf{s}) - A(\boldsymbol{\eta}_0)\}$$

valid for all \mathbf{s} such that $\boldsymbol{\eta}_0 + \mathbf{s} \in \mathcal{E}$. Since $\boldsymbol{\eta}_0$ is an interior point this set of \mathbf{s} includes a ball about $\mathbf{0}$.

Corollary 1.6.1. *Under the conditions of Theorem 1.6.3*

$$E_{\boldsymbol{\eta}_0} \mathbf{T}(X) = \dot{A}(\boldsymbol{\eta}_0)$$

$$\text{Var}_{\boldsymbol{\eta}_0} \mathbf{T}(X) = \ddot{A}(\boldsymbol{\eta}_0)$$

where $\dot{A}(\boldsymbol{\eta}_0) = (\frac{\partial A}{\partial \eta_1}(\boldsymbol{\eta}_0), \dots, \frac{\partial A}{\partial \eta_k}(\boldsymbol{\eta}_0))^T$, $\ddot{A}(\boldsymbol{\eta}_0) = \|\frac{\partial^2 A}{\partial \eta_a \partial \eta_b}(\boldsymbol{\eta}_0)\|$.

The corollary follows immediately from Theorem B.5.1 and Theorem 1.6.3(c).

Proof of Theorem 1.6.3. We prove (b) first. Suppose $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \mathcal{E}$ and $0 \leq \alpha \leq 1$. By the Hölder inequality (B.9.4), for any $u(x), v(x), h(x) \geq 0$, $r, s > 0$ with $\frac{1}{r} + \frac{1}{s} = 1$,

$$\int u(x)v(x)h(x)dx \leq \left(\int u^r(x)h(x)dx\right)^{\frac{1}{r}} \left(\int v^s(x)h(x)dx\right)^{\frac{1}{s}}.$$

Substitute $\frac{1}{r} = \alpha$, $\frac{1}{s} = 1 - \alpha$, $u(x) = \exp(\alpha \boldsymbol{\eta}_1^T \mathbf{T}(x))$, $v(x) = \exp((1 - \alpha) \boldsymbol{\eta}_2^T \mathbf{T}(x))$ and take logs of both sides to obtain, (with ∞ permitted on either side),

$$A(\alpha \boldsymbol{\eta}_1 + (1 - \alpha) \boldsymbol{\eta}_2) \leq \alpha A(\boldsymbol{\eta}_1) + (1 - \alpha) A(\boldsymbol{\eta}_2) \quad (1.6.15)$$

which is (b). If $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \mathcal{E}$ the right-hand side of (1.6.15) is finite. Because

$$\int \exp(\boldsymbol{\eta}^T \mathbf{T}(x)) h(x) dx > 0$$

for all $\boldsymbol{\eta}$ we conclude from (1.6.15) that $\alpha \boldsymbol{\eta}_1 + (1 - \alpha) \boldsymbol{\eta}_2 \in \mathcal{E}$ and (a) follows. Finally (c) is proved in exactly the same way as Theorem 1.6.2. \square

The formulae of Corollary 1.6.1 give a classical result in Example 1.6.6.

Example 1.6.7. (*continued*). Here, using the α parametrization,

$$A(\alpha) = n \log \left(\sum_{j=1}^k e^{\alpha_j} \right)$$

and

$$E_{\lambda}(T_j(\mathbf{X})) = nP_{\lambda}[X = j] \equiv n\lambda_j = ne^{\alpha_j} / \sum_{\ell=1}^k e^{\alpha_{\ell}}$$

$$\begin{aligned} \text{Cov}_{\lambda}(T_i, T_j) &= \frac{\partial^2 A}{\partial \alpha_i \partial \alpha_j}(\alpha) = -n \frac{e^{\alpha_j} e^{\alpha_i}}{(\sum_{\ell=1}^k e^{\alpha_{\ell}})^2} = -n\lambda_i \lambda_j, \quad i \neq j \\ \text{Var}_{\lambda}(T_i) &= \frac{\partial^2 A}{\partial \alpha_i^2}(\alpha) = n\lambda_i(1 - \lambda_i). \end{aligned}$$

□

The rank of an exponential family

Evidently every k -parameter exponential family is also k' -dimensional with $k' > k$. However, there is a minimal dimension.

An exponential family is of *rank* k iff the generating statistic \mathbf{T} is k -dimensional and 1, $T_1(X), \dots, T_k(X)$ are linearly independent with positive probability. Formally, $P_{\eta}[\sum_{j=1}^k a_j T_j(X) = a_{k+1}] < 1$ unless all a_j are 0.

Note that $P_{\theta}(A) = 0$ or $P_{\theta}(A) < 1$ for some θ iff the corresponding statement holds for all θ because $0 < \frac{p(x, \theta_1)}{p(x, \theta_2)} < \infty$ for all x, θ_1, θ_2 such that $h(x) > 0$.

Going back to Example 1.6.7 we can see that the multinomial family is of rank at most $k - 1$. It is intuitively clear that $k - 1$ is in fact its rank and this is seen in Theorem 1.6.4 that follows. Similarly, in Example 1.6.8, if $n = 1$, and $\eta_1(\theta) = \theta_1 + \theta_2 x_1$ we are writing the one-parameter binomial family corresponding to Y_1 as a two-parameter family with generating statistic $(Y_1, x_1 Y_1)$. But the rank of the family is 1 and θ_1 and θ_2 are not identifiable. However, if we consider \mathbf{Y} with $n \geq 2$ and $x_1 < x_n$ the family as we have seen remains of rank ≤ 2 and is in fact of rank 2. Our discussion suggests a link between rank and identifiability of the η parameterization. We establish the connection and other fundamental relationships in Theorem 1.6.4.

Theorem 1.6.4. Suppose $\mathcal{P} = \{q(x, \eta); \eta \in \mathcal{E}\}$ is a canonical exponential family generated by $(\mathbf{T}_{k \times 1}, h)$ with natural parameter space \mathcal{E} such that \mathcal{E} is open. Then the following are equivalent.

- (i) \mathcal{P} is of rank k .
- (ii) η is a parameter (identifiable).
- (iii) $\text{Var}_{\eta}(\mathbf{T})$ is positive definite.

(iv) $\eta \rightarrow \dot{A}(\eta)$ is 1-1 on \mathcal{E} .

(v) A is strictly convex on \mathcal{E} .

Note that, by Theorem 1.6.3, because \mathcal{E} is open, \dot{A} is defined on all of \mathcal{E} .

Proof. We give a detailed proof for $k = 1$. The proof for $k > 1$ is then sketched with details left to a problem. Let $\sim (\cdot)$ denote “ (\cdot) is false.” Then

$\sim(\text{i}) \Leftrightarrow P_\eta[a_1 T = a_2] = 1$ for $a_1 \neq 0$. This is equivalent to $\text{Var}_\eta(T) = 0 \Leftrightarrow \sim(\text{iii})$

$\sim(\text{ii}) \Leftrightarrow$ There exist $\eta_1 \neq \eta_2$ such that $P_{\eta_1} = P_{\eta_2}$.

Equivalently

$$\exp\{\eta_1 T(x) - A(\eta_1)\}h(x) = \exp\{\eta_2 T(x) - A(\eta_2)\}h(x).$$

Taking logs we obtain $(\eta_1 - \eta_2)T(X) = A(\eta_2) - A(\eta_1)$ with probability 1 $\equiv \sim(\text{i})$. We, thus, have (i) \equiv (ii) \equiv (iii). Now (iii) $\Rightarrow A''(\eta) > 0$ by Theorem 1.6.2 and, hence, $A'(\eta)$ is strictly monotone increasing and 1-1. Conversely, $A''(\eta_0) = 0$ for some η_0 implies that $T \equiv c$, with probability 1, for all η , by our remarks in the discussion of rank, which implies that $A''(\eta) = 0$ for all η and, hence, A' is constant. Thus, (iii) \equiv (iv) and the same discussion shows that (iii) \equiv (v).

Proof of the general case sketched

I. $\sim(\text{i}) \equiv \sim(\text{iii})$

$\sim(\text{i}) \equiv P_\eta[\mathbf{a}^T \mathbf{T} = c] = 1$ for some $\mathbf{a} \neq 0$, all η

$\sim(\text{iii}) \equiv \mathbf{a}^T \text{Var}_\eta(\mathbf{T})\mathbf{a} = \text{Var}_\eta(\mathbf{a}^T \mathbf{T}) = 0$ for some $\mathbf{a} \neq 0$, all $\eta \equiv (\sim \text{i})$

II. $\sim(\text{ii}) \equiv \sim(\text{i})$

$\sim(\text{ii}) \equiv P_{\eta_1} = P_{\eta_0}$ some $\eta_1 \neq \eta_0$. Let

$$\mathcal{Q} = \{P_{\eta_0 + c(\eta_1 - \eta_0)} : \eta_0 + c(\eta_1 - \eta_0) \in \mathcal{E}\}.$$

\mathcal{Q} is the exponential family (one-parameter) generated by $(\eta_1 - \eta_0)^T \mathbf{T}$. Apply the case $k = 1$ to \mathcal{Q} to get $\sim(\text{ii}) \equiv \sim(\text{i})$.

III. (iv) \equiv (v) \equiv (iii)

Properties (iv) and (v) are equivalent to the statements holding for every \mathcal{Q} defined as previously for arbitrary η_0, η_1 . \square

Corollary 1.6.2. Suppose that the conditions of Theorem 1.6.4 hold and \mathcal{P} is of rank k . Then

(a) \mathcal{P} may be uniquely parametrized by $\mu(\eta) \equiv E_\eta \mathbf{T}(X)$ where μ ranges over $\dot{A}(\mathcal{E})$,

(b) $\log q(x, \eta)$ is a strictly concave function of η on \mathcal{E} .

Proof. This is just a restatement of (iv) and (v) of the theorem. \square

The relation in (a) is sometimes evident and the μ parametrization is close to the initial parametrization of classical \mathcal{P} . Thus, the $\mathcal{B}(n, \theta)$ family is parametrized by $E(X)$, where X is the Bernoulli trial, the $\mathcal{N}(\mu, \sigma_0^2)$ family by $E(X)$. For $\{\mathcal{N}(\mu, \sigma^2)\}$, $E(X, X^2) = (\mu, \sigma^2 + \mu^2)$, which is obviously a 1-1 function of (μ, σ^2) . However, the relation in (a) may be far from obvious (see Problem 1.6.21). The corollary will prove very important in estimation theory. See Section 2.3. We close the present discussion of exponential families with the following example.

Example 1.6.11. The p Variate Gaussian Family. An important exponential family is based on the multivariate Gaussian distributions of Section B.6. Recall that $\mathbf{Y}_{p \times 1}$ has a p variate Gaussian distribution, $\mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$, with mean $\boldsymbol{\mu}_{p \times 1}$ and positive definite variance covariance matrix $\Sigma_{p \times p}$, iff its density is

$$f(\mathbf{Y}, \boldsymbol{\mu}, \Sigma) = |\det(\Sigma)|^{-p/2} (2\pi)^{-p/2} \exp\left\{-\frac{1}{2}(\mathbf{Y} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{Y} - \boldsymbol{\mu})\right\}. \quad (1.6.16)$$

Rewriting the exponent we obtain

$$\begin{aligned} \log f(\mathbf{Y}, \boldsymbol{\mu}, \Sigma) &= -\frac{1}{2} \mathbf{Y}^T \Sigma^{-1} \mathbf{Y} + (\Sigma^{-1} \boldsymbol{\mu})^T \mathbf{Y} \\ &\quad - \frac{1}{2} (\log |\det(\Sigma)| + \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu}) - \frac{p}{2} \log \pi. \end{aligned} \quad (1.6.17)$$

The first two terms on the right in (1.6.17) can be rewritten

$$-(\sum_{1 \leq i < j \leq p} \sigma^{ij} Y_i Y_j + \frac{1}{2} \sum_{i=1}^p \sigma^{ii} Y_i^2) + \sum_{i=1}^p (\sum_{j=1}^p \sigma^{ij} \mu_j) Y_i$$

where $\Sigma^{-1} \equiv \|\sigma^{ij}\|$, revealing that this is a $k = p(p+3)/2$ parameter exponential family with statistics $(Y_1, \dots, Y_p, \{Y_i Y_j\}_{1 \leq i < j \leq p})$, $h(\mathbf{Y}) \equiv 1$, $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma)$, $B(\boldsymbol{\theta}) = \frac{1}{2} (\log |\det(\Sigma)| + \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu})$. By our supermodel discussion, if $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ are iid $\mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$, then $\mathbf{X} \equiv (\mathbf{Y}_1, \dots, \mathbf{Y}_n)^T$ follows the $k = p(p+3)/2$ parameter exponential family with $\mathbf{T} = (\Sigma_i \mathbf{Y}_i, \Sigma_i \mathbf{Y}_i \mathbf{Y}_i^T)$, where we identify the second element of \mathbf{T} , which is a $p \times p$ symmetric matrix, with its distinct $p(p+1)/2$ entries. It may be shown (Problem 1.6.29) that \mathbf{T} (and $h \equiv 1$) generate this family and that the rank of the family is indeed $p(p+3)/2$, generalizing Example 1.6.5, and that \mathcal{E} is open, so that Theorem 1.6.4 applies. \square

1.6.5 Conjugate Families of Prior Distributions

In Section 1.2 we considered beta prior distributions for the probability of success in n Bernoulli trials. This is a special case of *conjugate families* of priors, families to which the posterior after sampling also belongs.

Suppose X_1, \dots, X_n is a sample from the k -parameter exponential family (1.6.10), and, as we always do in the Bayesian context, write $p(\mathbf{x} | \boldsymbol{\theta})$ for $p(\mathbf{x}, \boldsymbol{\theta})$. Then

$$p(\mathbf{x} | \boldsymbol{\theta}) = \left[\prod_{i=1}^n h(x_i) \right] \exp\left\{ \sum_{j=1}^k \eta_j(\boldsymbol{\theta}) \sum_{i=1}^n T_j(x_i) - nB(\boldsymbol{\theta}) \right\}. \quad (1.6.18)$$

where $\theta \in \Theta$, which is k -dimensional. A conjugate exponential family is obtained from (1.6.18) by letting n and $t_j = \sum_{i=1}^n T_j(x_i)$, $j = 1, \dots, k$, be “parameters” and treating θ as the variable of interest. That is, let $\mathbf{t} = (t_1, \dots, t_{k+1})^T$ and

$$\begin{aligned}\omega(\mathbf{t}) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left\{\sum_{j=1}^k t_j \eta_j(\theta) - t_{k+1} B(\theta)\right\} d\theta_1 \cdots d\theta_k \\ \Omega &= \{(t_1, \dots, t_{k+1}) : 0 < \omega(t_1, \dots, t_{k+1}) < \infty\}\end{aligned}\quad (1.6.19)$$

with integrals replaced by sums in the discrete case. We assume that Ω is nonempty (see Problem 1.6.36), then

Proposition 1.6.1. *The $(k+1)$ -parameter exponential family given by*

$$\pi_{\mathbf{t}}(\theta) = \exp\left\{\sum_{j=1}^k \eta_j(\theta) t_j - t_{k+1} B(\theta) - \log \omega(\mathbf{t})\right\} \quad (1.6.20)$$

where $\mathbf{t} = (t_1, \dots, t_{k+1}) \in \Omega$, is a conjugate prior to $p(\mathbf{x}|\theta)$ given by (1.6.18).

Proof. If $p(\mathbf{x}|\theta)$ is given by (1.6.18) and π by (1.6.20), then

$$\begin{aligned}\pi(\theta|\mathbf{x}) &\propto p(\mathbf{x}|\theta) \pi_{\mathbf{t}}(\theta) \propto \exp\left\{\sum_{j=1}^k \eta_j(\theta) \left(\sum_{i=1}^n T_j(x_i) + t_j\right) - (t_{k+1} + n) B(\theta)\right\} \\ &\propto \pi_{\mathbf{s}}(\theta),\end{aligned}\quad (1.6.21)$$

where

$$\mathbf{s} = (s_1, \dots, s_{k+1})^T = \left(t_1 + \sum_{i=1}^n T_1(x_i), \dots, t_k + \sum_{i=1}^n T_k(x_i), t_{k+1} + n\right)^T$$

and \propto indicates that the two sides are proportional functions of θ . Because two probability densities that are proportional must be equal, $\pi(\theta|\mathbf{x})$ is the member of the exponential family (1.6.20) given by the last expression in (1.6.21) and our assertion follows. \square

Remark 1.6.1. Note that (1.6.21) is an updating formula in the sense that as data x_1, \dots, x_n become available, the parameter \mathbf{t} of the prior distribution is updated to $\mathbf{s} = (\mathbf{t} + \mathbf{a})$, where $\mathbf{a} = (\sum_{i=1}^n T_1(x_i), \dots, \sum_{i=1}^n T_k(x_i), n)^T$. \square

It is easy to check that the beta distributions are obtained as conjugate to the binomial in this way.

Example 1.6.12. Suppose X_1, \dots, X_n is a $\mathcal{N}(\theta, \sigma_0^2)$ sample, where σ_0^2 is known and θ is unknown. To choose a prior distribution for θ , we consider the conjugate family of the model defined by (1.6.20). For $n = 1$

$$p(x|\theta) \propto \exp\left\{\frac{\theta x}{\sigma_0^2} - \frac{\theta^2}{2\sigma_0^2}\right\}. \quad (1.6.22)$$

This is a one-parameter exponential family with

$$T_1(x) = x, \eta_1(\theta) = \frac{\theta}{\sigma_0^2}, B(\theta) = \frac{\theta^2}{2\sigma_0^2}.$$

The conjugate two-parameter exponential family given by (1.6.20) has density

$$\pi_{\mathbf{t}}(\theta) = \exp\left\{\frac{\theta}{\sigma_0^2}t_1 - \frac{\theta^2}{2\sigma_0^2}t_2 - \log \omega(t_1, t_2)\right\}. \quad (1.6.23)$$

Upon completing the square, we obtain

$$\pi_{\mathbf{t}}(\theta) \propto \exp\left\{-\frac{t_2}{2\sigma_0^2}\left(\theta - \frac{t_1}{t_2}\right)^2\right\}. \quad (1.6.24)$$

Thus, $\pi_{\mathbf{t}}(\theta)$ is defined only for $t_2 > 0$ and all t_1 and is the $\mathcal{N}(t_1/t_2, \sigma_0^2/t_2)$ density. Our conjugate family, therefore, consists of all $\mathcal{N}(\eta_0, \tau_0^2)$ distributions where η_0 varies freely and τ_0^2 is positive.

If we start with a $\mathcal{N}(\eta_0, \tau_0^2)$ prior density, we must have in the (t_1, t_2) parametrization

$$t_2 = \frac{\sigma_0^2}{\tau_0^2}, \quad t_1 = \frac{\eta_0 \sigma_0^2}{\tau_0^2}. \quad (1.6.25)$$

By (1.6.21), if we observe $\Sigma X_i = s$, the posterior has a density (1.6.23) with

$$t_2(n) = \frac{\sigma_0^2}{\tau_0^2} + n, \quad t_1(s) = \frac{\eta_0 \sigma_0^2}{\tau_0^2} + s.$$

Using (1.6.24), we find that $\pi(\theta|\mathbf{x})$ is a normal density with mean

$$\mu(s, n) = \frac{t_1(s)}{t_2(n)} = \left(\frac{\sigma_0^2}{\tau_0^2} + n\right)^{-1} \left[s + \frac{\eta_0 \sigma_0^2}{\tau_0^2}\right] \quad (1.6.26)$$

and variance

$$\tau_0^2(n) = \frac{\sigma_0^2}{t_2(n)} = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma_0^2}\right)^{-1}. \quad (1.6.27)$$

Note that we can rewrite (1.6.26) intuitively as

$$\mu(s, n) = w_1 \bar{x} + w_2 \eta_0 \quad (1.6.28)$$

where $w_1 = n\tau_0^2(n)/\sigma_0^2$, $w_2 = \tau_0^2(n)/\tau_0^2$ so that $w_2 = 1 - w_1$. \square

These formulae can be generalized to the case \mathbf{X}_i i.i.d. $\mathcal{N}_p(\boldsymbol{\theta}, \Sigma_0)$, $1 \leq i \leq n$, Σ_0 known, $\boldsymbol{\theta} \sim \mathcal{N}_p(\boldsymbol{\eta}_0, \tau_0^2 \mathbf{I})$ where $\boldsymbol{\eta}_0$ varies over R^p , τ_0^2 is scalar with $\tau_0 > 0$ and \mathbf{I} is the $p \times p$ identity matrix (Problem 1.6.37). Moreover, it can be shown (Problem 1.6.30) that the $\mathcal{N}_p(\boldsymbol{\lambda}, \Gamma)$ family with $\boldsymbol{\lambda} \in R^p$ and Γ symmetric positive definite is a conjugate family

to $\mathcal{N}_p(\boldsymbol{\theta}, \Sigma_0)$, but a richer one than we've defined in (1.6.20) except for $p = 1$ because $\mathcal{N}_p(\boldsymbol{\lambda}, \Gamma)$ is a $p(p+3)/2$ rather than a $p+1$ parameter family. In fact, the conditions of Proposition 1.6.1 are often too restrictive. In the one-dimensional Gaussian case the members of the Gaussian conjugate family are unimodal and symmetric and have the same shape. It is easy to see that one can construct conjugate priors for which one gets reasonable formulae for the parameters indexing the model and yet have as great a richness of the shape variable as one wishes by considering finite mixtures of members of the family defined in (1.6.20). See Problems 1.6.31 and 1.6.32.

Discussion

Note that the uniform $\mathcal{U}(\{1, 2, \dots, \theta\})$ model of Example 1.5.3 is *not* covered by this theory. The natural sufficient statistic $\max(X_1, \dots, X_n)$, which is one-dimensional whatever be the sample size, is not of the form $\sum_{i=1}^n T(X_i)$. In fact, the family of distributions in this example and the family $\mathcal{U}(0, \theta)$ are not exponential. Despite the existence of classes of examples such as these, starting with Koopman, Pitman, and Darmois, a theory has been built up that indicates that under suitable regularity conditions families of distributions, which admit k -dimensional sufficient statistics for all sample sizes, must be k -parameter exponential families. Some interesting results and a survey of the literature may be found in Brown (1986). Problem 1.6.10 is a special result of this type.

Summary. $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, $\Theta \subset R^k$, is a k -parameter exponential family of distributions if there are real-valued functions η_1, \dots, η_k and B on Θ , and real-valued functions T_1, \dots, T_k, h on R^q such that the density (frequency) function of $P_{\boldsymbol{\theta}}$ can be written as

$$p(x, \boldsymbol{\theta}) = h(x) \exp\left[\sum_{j=1}^k \eta_j(\boldsymbol{\theta}) T_j(x) - B(\boldsymbol{\theta})\right], x \in \mathcal{X} \subset R^q. \quad (1.6.29)$$

$\mathbf{T}(X) = (T_1(X), \dots, T_k(X))^T$ is called the *natural sufficient statistic* of the family. The *canonical k -parameter exponential family generated by \mathbf{T} and h* is

$$q(x, \boldsymbol{\eta}) = h(x) \exp\{\mathbf{T}^T(x) \boldsymbol{\eta} - A(\boldsymbol{\eta})\}$$

where

$$A(\boldsymbol{\eta}) = \log \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} h(x) \exp\{\mathbf{T}^T(x) \boldsymbol{\eta}\} dx$$

in the continuous case, with integrals replaced by sums in the discrete case. The set

$$\mathcal{E} = \{\boldsymbol{\eta} \in R^k : -\infty < A(\boldsymbol{\eta}) < \infty\}$$

is called the *natural parameter space*. The set \mathcal{E} is convex, the map $A : \mathcal{E} \rightarrow R$ is convex. If \mathcal{E} has a nonempty interior in R^k and $\boldsymbol{\eta}_0 \in \mathcal{E}$, then $\mathbf{T}(X)$ has for $X \sim P_{\boldsymbol{\eta}_0}$ the moment-generating function

$$\psi(\mathbf{s}) = \exp\{A(\boldsymbol{\eta}_0 + \mathbf{s}) - A(\boldsymbol{\eta}_0)\}$$

for all \mathbf{s} such that $\boldsymbol{\eta}_0 + \mathbf{s}$ is in \mathcal{E} . Moreover $E_{\boldsymbol{\eta}_0}[\mathbf{T}(X)] = \dot{A}(\boldsymbol{\eta}_0)$ and $Var_{\boldsymbol{\eta}_0}[T(X)] = \ddot{A}(\boldsymbol{\eta}_0)$ where \dot{A} and \ddot{A} denote the gradient and Hessian of A .

An exponential family is said to be of *rank* k if \mathbf{T} is k -dimensional and $1, T_1, \dots, T_k$ are linearly independent with positive P_{θ} probability for some $\theta \in \Theta$. If \mathcal{P} is a canonical exponential family with \mathcal{E} open, then the following are equivalent:

- (i) \mathcal{P} is of rank k ,
- (ii) η is identifiable,
- (iii) $\text{Var}_{\eta}(\mathbf{T})$ is positive definite,
- (iv) the map $\eta \rightarrow \dot{A}(\eta)$ is 1 - 1 on \mathcal{E} ,
- (v) A is strictly convex on \mathcal{E} .

A family \mathcal{F} of prior distributions for a parameter vector θ is called a *conjugate family* of priors to $p(x | \theta)$ if the posterior distribution of θ given \mathbf{x} is a member of \mathcal{F} . The $(k + 1)$ -parameter exponential family

$$\pi_{\mathbf{t}}(\theta) = \exp\left\{\sum_{j=1}^k \eta_j(\theta)t_j - B(\theta)t_{k+1} - \log \omega\right\}$$

where

$$\omega = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\{\Sigma \eta_j(\theta)t_j - B(\theta)t_{k+1}\} d\theta,$$

and

$$\mathbf{t} = (t_1, \dots, t_{k+1}) \in \Omega = \{(t_1, \dots, t_{k+1}) \in R^{k+1} : 0 < \omega < \infty\},$$

is conjugate to the exponential family $p(x|\theta)$ defined in (1.6.29).

1.7 PROBLEMS AND COMPLEMENTS

Problems for Section 1.1

1. Give a formal statement of the following models identifying the probability laws of the data and the parameter space. State whether the model in question is parametric or nonparametric.

(a) A geologist measures the diameters of a large number n of pebbles in an old stream bed. Theoretical considerations lead him to believe that the logarithm of pebble diameter is normally distributed with mean μ and variance σ^2 . He wishes to use his observations to obtain some information about μ and σ^2 but has in advance no knowledge of the magnitudes of the two parameters.

(b) A measuring instrument is being used to obtain n independent determinations of a physical constant μ . Suppose that the measuring instrument is known to be biased to the positive side by 0.1 units. Assume that the errors are otherwise identically distributed normal random variables with known variance.

(c) In part (b) suppose that the amount of bias is positive but unknown. Can you perceive any difficulties in making statements about μ for this model?

(d) The number of eggs laid by an insect follows a Poisson distribution with unknown mean λ . Once laid, each egg has an unknown chance p of hatching and the hatching of one egg is independent of the hatching of the others. An entomologist studies a set of n such insects observing both the number of eggs laid and the number of eggs hatching for each nest.

2. Are the following parametrizations identifiable? (Prove or disprove.)

(a) The parametrization of Problem 1.1.1(c).

(b) The parametrization of Problem 1.1.1(d).

(c) The parametrization of Problem 1.1.1(d) if the entomologist observes *only* the number of eggs hatching but not the number of eggs laid in each case.

3. Which of the following parametrizations are identifiable? (Prove or disprove.)

(a) X_1, \dots, X_p are independent with $X_i \sim \mathcal{N}(\alpha_i + \nu, \sigma^2)$.

$$\theta = (\alpha_1, \alpha_2, \dots, \alpha_p, \nu, \sigma^2)$$

and P_θ is the distribution of $\mathbf{X} = (X_1, \dots, X_p)$.

(b) Same as (a) with $\alpha = (\alpha_1, \dots, \alpha_p)$ restricted to

$$\{(a_1, \dots, a_p) : \sum_{i=1}^p a_i = 0\}.$$

(c) X and Y are independent $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$, $\theta = (\mu_1, \mu_2)$ and we observe $Y - X$.

(d) X_{ij} , $i = 1, \dots, p$; $j = 1, \dots, b$ are independent with $X_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma^2)$ where $\mu_{ij} = \nu + \alpha_i + \lambda_j$, $\theta = (\alpha_1, \dots, \alpha_p, \lambda_1, \dots, \lambda_b, \nu, \sigma^2)$ and P_θ is the distribution of X_{11}, \dots, X_{pb} .

(e) Same as (d) with $(\alpha_1, \dots, \alpha_p)$ and $(\lambda_1, \dots, \lambda_b)$ restricted to the sets where $\sum_{i=1}^p \alpha_i = 0$ and $\sum_{j=1}^b \lambda_j = 0$.

4. (a) Let U be any random variable and V be any other nonnegative random variable. Show that

$$F_{U+V}(t) \leq F_U(t) \text{ for every } t.$$

(If F_X and F_Y are distribution functions such that $F_X(t) \leq F_Y(t)$ for every t , then X is said to be *stochastically larger* than Y .)

(b) As in Problem 1.1.1 describe formally the following model. Two groups of n_1 and n_2 individuals, respectively, are sampled at random from a very large population. Each

member of the second (treatment) group is administered the same dose of a certain drug believed to lower blood pressure and the blood pressure is measured after 1 hour. Each member of the first (control) group is administered an equal dose of a placebo and then has the blood pressure measured after 1 hour. It is known that the drug either has no effect or lowers blood pressure, but the distribution of blood pressure in the population sampled before and after administration of the drug is quite unknown.

5. The number n of graduate students entering a certain department is recorded. In each of k subsequent years the number of students graduating and of students dropping out is recorded. Let N_i be the number dropping out and M_i the number graduating during year i , $i = 1, \dots, k$. The following model is proposed.

$$P_\theta[N_1 = n_1, M_1 = m_1, \dots, N_k = n_k, M_k = m_k] \\ = \frac{n!}{n_1! \dots n_k! m_1! \dots m_k! r!} \mu_1^{n_1} \dots \mu_k^{n_k} \nu_1^{m_1} \dots \nu_k^{m_k} \rho^r$$

where

$$\mu_1 + \dots + \mu_k + \nu_1 + \dots + \nu_k + \rho = 1, \quad 0 < \mu_i < 1, \quad 0 < \nu_i < 1, \quad 1 \leq i \leq k \\ n_1 + \dots + n_k + m_1 + \dots + m_k + r = n$$

and $\theta = (\mu_1, \dots, \mu_k, \nu_1, \dots, \nu_k)$ is unknown.

(a) What are the assumptions underlying this model?

(b) θ is very difficult to estimate here if k is large. The simplification $\mu_i = \pi(1 - \mu)^{i-1}$, $\nu_i = (1 - \pi)(1 - \nu)^{i-1}$ for $i = 1, \dots, k$ is proposed where $0 < \pi < 1$, $0 < \mu < 1$, $0 < \nu < 1$ are unknown. What assumptions underlie the simplification?

6. Which of the following models are regular? (Prove or disprove.)

(a) P_θ is the distribution of X when X is uniform on $(0, \theta)$, $\Theta = (0, \infty)$.

(b) P_θ is the distribution of X when X is uniform on $\{0, 1, 2, \dots, \theta\}$, $\Theta = \{1, 2, \dots\}$.

(c) Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$. Let $Y = 1$ if $X \leq 1$ and $Y = X$ if $X > 1$. $\theta = (\mu, \sigma^2)$ and P_θ is the distribution of Y .

(d) Suppose the possible control responses in an experiment are $0.1, 0.2, \dots, 0.9$ and they occur with frequencies $p(0.1), p(0.2), \dots, p(0.9)$. Suppose the effect of a treatment is to increase the control response by a fixed amount θ . Let P_θ be the distribution of a treatment response.

7. Show that $Y - c$ has the same distribution as $-Y + c$, if and only if, the density or frequency function p of Y satisfies $p(c + t) = p(c - t)$ for all t . Both Y and p are said to be *symmetric* about c .

Hint: If $Y - c$ has the same distribution as $-Y + c$, then $P(Y \leq t + c) = P(-Y \leq t - c) = P(Y \geq c - t) = 1 - P(Y < c - t)$.

8. Consider the two sample models of Examples 1.1.3(2) and 1.1.4(1).

(a) Show that if $Y \sim X + \delta(X)$, $\delta(x) = 2\mu + \Delta - 2x$ and $X \sim \mathcal{N}(\mu, \sigma^2)$, then $G(\cdot) = F(\cdot - \Delta)$. That is, the two cases $\delta(x) \equiv \Delta$ and $\delta(x) = 2\mu + \Delta - 2x$ yield the same distribution for the data (X_1, \dots, X_n) , (Y_1, \dots, Y_n) . Therefore, $G(\cdot) = F(\cdot - \Delta)$ does not imply the constant treatment effect assumption.

(b) In part (a), suppose X has a distribution F that is not necessarily normal. For what type of F is it possible to have $G(\cdot) = F(\cdot - \Delta)$ for both $\delta(x) \equiv \Delta$ and $\delta(x) = 2\mu + \Delta - 2x$?

(c) Suppose that $Y \sim X + \delta(X)$ where $X \sim \mathcal{N}(\mu, \sigma^2)$ and $\delta(x)$ is continuous. Show that if we assume that $\delta(x) + x$ is strictly increasing, then $G(\cdot) = F(\cdot - \Delta)$ implies that $\delta(x) \equiv \Delta$.

9. Collinearity: Suppose $Y_i = \sum_{j=1}^p z_{ij}\beta_j + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ independent, $1 \leq i \leq n$. Let $\mathbf{z}_j \equiv (z_{1j}, \dots, z_{nj})^T$.

(a) Show that $(\beta_1, \dots, \beta_p)$ are identifiable iff $\mathbf{z}_1, \dots, \mathbf{z}_p$ are not collinear (linearly independent).

(b) Deduce that $(\beta_1, \dots, \beta_p)$ are not identifiable if $n < p$, that is, if the number of parameters is larger than the number of observations.

10. Let $X = (\min(T, C), I(T \leq C))$ where T, C are independent,

$$P[T = j] = p(j), \quad j = 0, \dots, N,$$

$$P[C = j] = r(j), \quad j = 0, \dots, N$$

and (p, r) vary freely over $\mathcal{F} = \{(p, r) : p(j) > 0, r(j) > 0, 0 \leq j \leq N, \sum_{j=0}^N p(j) = 1, \sum_{j=0}^N r(j) = 1\}$ and N is known. Suppose X_1, \dots, X_n are observed i.i.d. according to the distribution of X .

Show that $\{p(j) : j = 0, \dots, N\}, \{r(j) : j = 0, \dots, N\}$ are identifiable.

Hint: Consider “hazard rates” for $Y \equiv \min(T, C)$,

$$P[Y = j, Y = T \mid Y \geq j].$$

11. The Scale Model. Positive random variables X and Y satisfy a *scale model* with parameter $\delta > 0$ if $P(Y \leq t) = P(\delta X \leq t)$ for all $t > 0$, or equivalently, $G(t) = F(t/\delta)$, $\delta > 0, t > 0$.

(a) Show that in this case, $\log X$ and $\log Y$ satisfy a shift model with parameter $\log \delta$.

(b) Show that if X and Y satisfy a shift model with parameter Δ , then e^X and e^Y satisfy a scale model with parameter e^Δ .

(c) Suppose a scale model holds for X, Y . Let $c > 0$ be a constant. Does $X' = X^c$, $Y' = Y^c$ satisfy a scale model? Does $\log X', \log Y'$ satisfy a shift model?

12. The Lehmann Two-Sample Model. In Example 1.1.3 let X_1, \dots, X_m and Y_1, \dots, Y_n denote the survival times of two groups of patients receiving treatments A and B . $S_X(t) =$

$P(X > t) = 1 - F(t)$ and $S_Y(t) = P(Y > t) = 1 - G(t)$, $t > 0$, are called the *survival functions*. For the A group, survival beyond time t is modeled to occur if the events $T_1 > t, \dots, T_a > t$ all occur, where T_1, \dots, T_a are unobservable and i.i.d. as T with survival function S_0 . Similarly, for the B group, $Y > t$ occurs iff $T'_1 > t, \dots, T'_b > t$ where T'_1, \dots, T'_b are i.i.d. as T .

(a) Show that $S_Y(t) = S_X^{b/a}(t)$.

(b) By extending (b/a) from the rationals to $\delta \in (0, \infty)$, we have the *Lehmann model*

$$S_Y(t) = S_X^\delta(t), \quad t > 0. \quad (1.7.1)$$

Equivalently, $S_Y(t) = S_0^\Delta(t)$ with $\Delta = a\delta$, $t > 0$. Show that if S_0 is continuous, then $X' = -\log S_0(X)$ and $Y' = -\log S_0(Y)$ follow an exponential scale model (see Problem 1.1.11) with scale parameter δ^{-1} .

Hint: By Problem B.2.12, $S_0(T)$ has a $\mathcal{U}(0, 1)$ distribution; thus, $-\log S_0(T)$ has an exponential distribution. Also note that $P(X > t) = S_0^a(t)$.

(c) Suppose that T and Y have densities $f_0(t)$ and $g(t)$. Then $h_0(t) = f_0(t)/S_0(t)$ and $h_Y(t) = g(t)/S_Y(t)$ are called the *hazard rates* of T and Y . Moreover, $h_Y(t) = \Delta h_0(t)$ is called the *Cox proportional hazard model*. Show that $h_Y(t) = \Delta h_0(t)$ if and only if $S_Y(t) = S_0^\Delta(t)$.

13. A proportional hazard model. Let $f(t | \mathbf{z}_i)$ denote the density of the survival time Y_i of a patient with covariate vector \mathbf{z}_i and define the *regression survival* and *hazard* functions of Y_i as

$$S_Y(t | \mathbf{z}_i) = \int_t^\infty f(y | \mathbf{z}_i) dy, \quad h(t | \mathbf{z}_i) = f(t | \mathbf{z}_i) / S_Y(t | \mathbf{z}_i).$$

Let T denote a survival time with density $f_0(t)$ and hazard rate $h_0(t) = f_0(t)/P(T > t)$. The *Cox proportional hazard model* is defined as

$$h(t | \mathbf{z}) = h_0(t) \exp\{g(\boldsymbol{\beta}, \mathbf{z})\} \quad (1.7.2)$$

where $h_0(t)$ is called the *baseline hazard function* and g is known except for a vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ of unknowns. The most common choice of g is the linear form $g(\boldsymbol{\beta}, \mathbf{z}) = \mathbf{z}^T \boldsymbol{\beta}$. Set $\Delta = \exp\{g(\boldsymbol{\beta}, \mathbf{z})\}$.

(a) Show that (1.7.2) is equivalent to $S_Y(t | \mathbf{z}) = S_T^\Delta(t)$.

(b) Assume (1.7.2) and that $F_0(t) = P(T \leq t)$ is known and strictly increasing. Find an increasing function $Q(t)$ such that the regression survival function of $Y' = Q(Y)$ does not depend on $h_0(t)$.

Hint: See Problem 1.1.12.

(c) Under the assumptions of (b) above, show that there is an increasing function $Q^*(t)$ such that if $Y_i^* = Q^*(Y_i)$, then

$$Y_i^* = -g(\boldsymbol{\beta}, \mathbf{z}_i) + \epsilon_i$$

for some appropriate ϵ_i . Specify the distribution of ϵ_i .

Hint: See Problems 1.1.11 and 1.1.12.

14. In Example 1.1.2 with assumptions (1)–(4), the parameter of interest can be characterized as the median $\nu = F^{-1}(0.5)$ or mean $\mu = \int_{-\infty}^{\infty} x dF(x) = \int_0^1 F^{-1}(u) du$. Generally, μ and ν are regarded as *centers* of the distribution F . When F is not symmetric, μ may be very much pulled in the direction of the longer tail of the density, and for this reason, the median is preferred in this case. Examples are the distribution of income and the distribution of wealth. Here is an example in which the mean is extreme and the median is not. Suppose the monthly salaries of state workers in a certain state are modeled by the *Pareto* distribution with distribution function

$$\begin{aligned} F(x, \theta) &= 1 - (x/c)^{-\theta}, & x \geq c \\ &= 0, & x < c \end{aligned}$$

where $\theta > 0$ and $c = 2,000$ is the minimum monthly salary for state workers. Find the median ν and the mean μ for the values of θ where the mean exists. Show how to choose θ to make $\mu - \nu$ arbitrarily large.

15. Let X_1, \dots, X_m be i.i.d. F, Y_1, \dots, Y_n be i.i.d. G , where the model $\{(F, G)\}$ is described by

$$\psi(X_1) = Z_1, \quad \psi(Y_1) = Z'_1 + \Delta,$$

where ψ is an unknown strictly increasing differentiable map from R to R , $\psi' > 0$, $\psi(\pm\infty) = \pm\infty$, and Z_1 and Z'_1 are independent random variables.

(a) Suppose Z_1, Z'_1 have a $\mathcal{N}(0, 1)$ distribution. Show that both ψ and Δ are identifiable.

(b) Suppose Z_1 and Z'_1 have a $\mathcal{N}(0, \sigma^2)$ distribution with σ^2 unknown. Are ψ and Δ still identifiable? If not, what parameters are?

Hint: (a) $P[X_1 \leq t] = \Phi(\psi(t))$.

Problems for Section 1.2

1. Merging Opinions. Consider a parameter space consisting of two points θ_1 and θ_2 , and suppose that for given θ , an experiment leads to a random variable X whose frequency function $p(x | \theta)$ is given by

$\theta \backslash x$	0	1
θ_1	0.8	0.2
θ_2	0.4	0.6

Let π be the prior frequency function of θ defined by $\pi(\theta_1) = \frac{1}{2}$, $\pi(\theta_2) = \frac{1}{2}$.

(a) Find the posterior frequency function $\pi(\theta | x)$.

(b) Suppose X_1, \dots, X_n are independent with frequency function $p(x | \theta)$. Find the posterior $\pi(\theta | x_1, \dots, x_n)$. Observe that it depends only on $\sum_{i=1}^n x_i$.

(c) Same as (b) except use the prior $\pi_1(\theta_1) = .25$, $\pi_1(\theta_2) = .75$.

(d) Give the values of $P(\theta = \theta_1 \mid \sum_{i=1}^n X_i = .5n)$ for the two priors π and π_1 when $n = 2$ and 100 .

(e) Give the most probable values $\hat{\theta} = \arg \max_{\theta} \pi(\theta \mid \sum_{i=1}^n X_i = k)$ for the two priors π and π_1 . Compare these $\hat{\theta}$'s for $n = 2$ and 100 .

(f) Give the set on which the two $\hat{\theta}$'s disagree. Show that the probability of this set tends to zero as $n \rightarrow \infty$. Assume $X \sim p(x) = \sum_{i=1}^2 \pi(\theta_i) p(x \mid \theta_i)$. For this convergence, does it matter which prior, π or π_1 , is used in the formula for $p(x)$?

2. Consider an experiment in which, for given $\theta = \theta$, the outcome X has density $p(x \mid \theta) = (2x/\theta^2)$, $0 < x < \theta$. Let π denote a prior density for θ .

(a) Find the posterior density of θ when $\pi(\theta) = 1$, $0 \leq \theta \leq 1$.

(b) Find the posterior density of θ when $\pi(\theta) = 3\theta^2$, $0 \leq \theta \leq 1$.

(c) Find $E(\theta \mid x)$ for the two priors in (a) and (b).

(d) Suppose X_1, \dots, X_n are independent with the same distribution as X . Find the posterior density of θ given $X_1 = x_1, \dots, X_n = x_n$ when $\pi(\theta) = 1$, $0 \leq \theta \leq 1$.

3. Let X be the number of failures before the first success in a sequence of Bernoulli trials with probability of success θ . Then $P_{\theta}[X = k] = (1 - \theta)^k \theta$, $k = 0, 1, 2, \dots$. This is called the *geometric distribution* ($\mathcal{G}(\theta)$). Suppose that for given $\theta = \theta$, X has the geometric distribution

(a) Find the posterior distribution of θ given $X = 2$ when the prior distribution of θ is uniform on $\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$.

(b) Relative to (a), what is the most probable value of θ given $X = 2$? Given $X = k$?

(c) Find the posterior distribution of θ given $X = k$ when the prior distribution is beta, $\beta(r, s)$.

4. Let X_1, \dots, X_n be distributed as

$$p(x_1, \dots, x_n \mid \theta) = \frac{1}{\theta^n}$$

where x_1, \dots, x_n are natural numbers between 1 and θ and $\Theta = \{1, 2, 3, \dots\}$.

(a) Suppose θ has prior frequency,

$$\pi(j) = \frac{c(a)}{j^a}, \quad j = 1, 2, \dots,$$

where $a > 1$ and $c(a) = [\sum_{j=1}^{\infty} j^{-a}]^{-1}$. Show that

$$\pi(j \mid x_1, \dots, x_n) = \frac{c(n+a, m)}{j^{n+a}}, \quad j = m, m+1, \dots,$$

where $m = \max(x_1, \dots, x_n)$, $c(b, t) = [\sum_{j=t}^{\infty} j^{-b}]^{-1}$, $b > 1$.

(b) Suppose that $\max(x_1, \dots, x_n) = x_1 = m$ for all n . Show that $\pi(m | x_1, \dots, x_n) \rightarrow 1$ as $n \rightarrow \infty$ whatever be a . Interpret this result.

5. In Example 1.2.1 suppose n is large and $(1/n) \sum_{i=1}^n x_i = \bar{x}$ is not close to 0 or 1 and the prior distribution is beta, $\beta(r, s)$. Justify the following approximation to the posterior distribution

$$P[\theta \leq t | X_1 = x_1, \dots, X_n = x_n] \approx \Phi\left(\frac{t - \tilde{\mu}}{\tilde{\sigma}}\right)$$

where Φ is the standard normal distribution function and

$$\tilde{\mu} = \frac{n}{n+r+s} \bar{x} + \frac{r}{n+r+s}, \quad \tilde{\sigma}^2 = \frac{\tilde{\mu}(1-\tilde{\mu})}{n+r+s}.$$

Hint: Let $\beta(a, b)$ denote the posterior distribution. If a and b are integers, then $\beta(a, b)$ is the distribution of $(a\bar{V}/b\bar{W})[1 + (a\bar{V}/b\bar{W})]^{-1}$, where $V_1, \dots, V_a, W_1, \dots, W_b$ are independent standard exponential. Next use the central limit theorem and Slutsky's theorem.

6. Show that a conjugate family of distributions for the Poisson family is the gamma family.

7. Show rigorously using (1.2.8) that if in Example 1.1.1, $D = N\theta$ has a $\mathcal{B}(N, \pi_0)$ distribution, then the posterior distribution of D given $X = k$ is that of $k + \mathbf{Z}$ where \mathbf{Z} has a $\mathcal{B}(N - n, \pi_0)$ distribution.

8. Let (X_1, \dots, X_{n+k}) be a sample from a population with density $f(x | \theta)$, $\theta \in \Theta$. Let θ have prior density π . Show that the conditional distribution of $(\theta, X_{n+1}, \dots, X_{n+k})$ given $X_1 = x_1, \dots, X_n = x_n$ is that of $(Y, \mathbf{Z}_1, \dots, \mathbf{Z}_k)$ where the marginal distribution of Y equals the posterior distribution of θ given $X_1 = x_1, \dots, X_n = x_n$, and the conditional distribution of the \mathbf{Z}_i 's given $Y = t$ is that of sample from the population with density $f(x | t)$.

9. Show in Example 1.2.1 that the conditional distribution of θ given $\sum_{i=1}^n X_i = k$ agrees with the posterior distribution of θ given $X_1 = x_1, \dots, X_n = x_n$, where $\sum_{i=1}^n x_i = k$.

10. Suppose X_1, \dots, X_n is a sample with $X_i \sim p(x | \theta)$, a regular model and integrable as a function of θ . Assume that $A = \{x : p(x | \theta) > 0\}$ does not involve θ .

(a) Show that the family of priors

$$\pi(\theta) = \prod_{i=1}^N p(\xi_i | \theta) \bigg/ \int_{\Theta} \prod_{i=1}^N p(\xi_i | \theta) d\theta$$

where $\xi_i \in A$ and $N \in \{1, 2, \dots\}$ is a conjugate family of prior distributions for $p(\mathbf{x} | \theta)$ and that the posterior distribution of θ given $\mathbf{X} = \mathbf{x}$ is

$$\pi(\theta | \mathbf{x}) = \prod_{i=1}^{N'} p(\xi'_i | \theta) \bigg/ \int_{\Theta} \prod_{i=1}^{N'} p(\xi'_i | \theta) d\theta$$

where $N' = N + n$ and $(\xi'_1, \dots, \xi'_{N'}) = (\xi_1, \dots, \xi_N, x_1, \dots, x_n)$.

(b) Use the result (a) to give $\pi(\theta)$ and $\pi(\theta | \mathbf{x})$ when

$$\begin{aligned} p(x | \theta) &= \theta \exp\{-\theta x\}, x > 0, \theta > 0 \\ &= 0 \text{ otherwise.} \end{aligned}$$

11. Let $p(x | \theta) = \exp\{-(x - \theta)\}$, $0 < \theta < x$ and let $\pi(\theta) = 2 \exp\{-2\theta\}$, $\theta > 0$. Find the posterior density $\pi(\theta | x)$.

12. Suppose $p(\mathbf{x} | \theta)$ is the density of i.i.d. X_1, \dots, X_n , where $X_i \sim \mathcal{N}(\mu_0, \frac{1}{\theta})$, μ_0 is known, and $\theta = \sigma^{-2}$ is (called) the *precision* of the distribution of X_i .

(a) Show that $p(\mathbf{x} | \theta) \propto \theta^{\frac{1}{2}n} \exp(-\frac{1}{2}t\theta)$ where $t = \sum_{i=1}^n (X_i - \mu_0)^2$ and \propto denotes “proportional to” as a function of θ .

(b) Let $\pi(\theta) \propto \theta^{\frac{1}{2}(\lambda-2)} \exp\{-\frac{1}{2}\nu\theta\}$, $\nu > 0$, $\lambda > 0$; $\theta > 0$. Find the posterior distribution $\pi(\theta | \mathbf{x})$ and show that if λ is an integer, given \mathbf{x} , $\theta(t + \nu)$ has a $\chi^2_{\lambda+n}$ distribution. Note that, unconditionally, $\nu\theta$ has a χ^2_λ distribution.

(c) Find the posterior distribution of σ .

13. Show that if X_1, \dots, X_n are i.i.d. $\mathcal{N}(\mu, \sigma^2)$ and we formally put $\pi(\mu, \sigma) = \frac{1}{\sigma}$, then the posterior density $\pi(\mu | \bar{x}, s^2)$ of μ given (\bar{x}, s^2) is such that $\sqrt{n} \frac{(\mu - \bar{X})}{s} \sim t_{n-1}$. Here $s^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$.

Hint: Given μ and σ , \bar{X} and s^2 are independent with $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ and $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$. This leads to $p(\bar{x}, s^2 | \mu, \sigma^2)$. Next use Bayes rule.

14. In a Bayesian model where X_1, \dots, X_n, X_{n+1} are i.i.d. $f(x | \theta)$, $\theta \sim \pi$, the *predictive distribution* is the marginal distribution of X_{n+1} . The *posterior predictive distribution* is the conditional distribution of X_{n+1} given X_1, \dots, X_n .

(a) If f and π are the $\mathcal{N}(\theta, \sigma_0^2)$ and $\mathcal{N}(\theta_0, \tau_0^2)$ densities, compute the predictive and posterior predictive distribution.

(b) Discuss the behavior of the two predictive distributions as $n \rightarrow \infty$.

15. The *Dirichlet distribution* is a conjugate prior for the multinomial. The Dirichlet distribution, $\mathcal{D}(\boldsymbol{\alpha})$, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_r)^T$, $\alpha_j > 0$, $1 \leq j \leq r$, has density

$$f_{\boldsymbol{\alpha}}(\mathbf{u}) = \frac{\Gamma(\sum_{j=1}^r \alpha_j)}{\prod_{j=1}^r \Gamma(\alpha_j)} \prod_{j=1}^r u_j^{\alpha_j-1}, 0 < u_j < 1, \sum_{j=1}^r u_j = 1.$$

Let $\mathbf{N} = (N_1, \dots, N_r)$ be multinomial

$$\mathcal{M}(n, \boldsymbol{\theta}), \boldsymbol{\theta} = (\theta_1, \dots, \theta_r)^T, 0 < \theta_j < 1, \sum_{j=1}^r \theta_j = 1.$$

Show that if the prior $\pi(\theta)$ for θ is $\mathcal{D}(\alpha)$, then the posterior $\pi(\theta \mid \mathbf{N} = \mathbf{n})$ is $\mathcal{D}(\alpha + \mathbf{n})$, where $\mathbf{n} = (n_1, \dots, n_r)$.

Problems for Section 1.3

1. Suppose the possible states of nature are θ_1, θ_2 , the possible actions are a_1, a_2, a_3 , and the loss function $l(\theta, a)$ is given by

$\theta \backslash a$	a_1	a_2	a_3
θ_1	0	1	2
θ_2	2	0	1

Let X be a random variable with frequency function $p(x, \theta)$ given by

$\theta \backslash x$	0	1
θ_1	p	$(1 - p)$
θ_2	q	$(1 - q)$

and let $\delta_1, \dots, \delta_9$ be the decision rules of Table 1.3.3. Compute and plot the risk points when

(a) $p = q = .1$,

(b) $p = 1 - q = .1$.

(c) Find the minimax rule among $\delta_1, \dots, \delta_9$ for the preceding case (a).

(d) Suppose that θ has prior $\pi(\theta_1) = 0.5$, $\pi(\theta_2) = 0.5$. Find the Bayes rule for case (a).

2. Suppose that in Example 1.3.5, a new buyer makes a bid and the loss function is changed to

$\theta \backslash a$	a_1	a_2	a_3
θ_1	0	7	4
θ_2	12	1	6

(a) Compute and plot the risk points in this case for each rule $\delta_1, \dots, \delta_9$ of Table 1.3.3.

(b) Find the minimax rule among $\{\delta_1, \dots, \delta_9\}$.

(c) Find the minimax rule among the randomized rules.

(d) Suppose θ has prior $\pi(\theta_1) = \gamma$, $\pi(\theta_2) = 1 - \gamma$. Find the Bayes rule when (i) $\gamma = 0.5$ and (ii) $\gamma = 0.1$.

3. The problem of selecting the better of two treatments or of deciding whether the effect of one treatment is beneficial or not often reduces to the problem of deciding whether $\theta < 0$, $\theta = 0$ or $\theta > 0$ for some parameter θ . See Example 1.1.3. Let the actions corresponding to deciding whether $\theta < 0$, $\theta = 0$ or $\theta > 0$ be denoted by $-1, 0, 1$, respectively and suppose the loss function is given by (from Lehmann, 1957)

$\theta \backslash a$	-1	0	1
< 0	0	c	$b + c$
$= 0$	b	0	b
> 0	$b + c$	c	0

where b and c are positive. Suppose \mathbf{X} is a $\mathcal{N}(\theta, 1)$ sample and consider the decision rule

$$\delta_{r,s}(\mathbf{X}) = \begin{cases} -1 & \text{if } \bar{X} < r \\ 0 & \text{if } r \leq \bar{X} \leq s \\ 1 & \text{if } \bar{X} > s. \end{cases}$$

(a) Show that the risk function is given by

$$\begin{aligned} R(\theta, \delta_{r,s}) &= c\bar{\Phi}(\sqrt{n}(r - \theta)) + b\bar{\Phi}(\sqrt{n}(s - \theta)), & \theta < 0 \\ &= b\bar{\Phi}(\sqrt{n}s) + b\Phi(\sqrt{n}r), & \theta = 0 \\ &= c\Phi(\sqrt{n}(s - \theta)) + b\Phi(\sqrt{n}(r - \theta)), & \theta > 0 \end{aligned}$$

where $\bar{\Phi} = 1 - \Phi$, and Φ is the $\mathcal{N}(0, 1)$ distribution function.

(b) Plot the risk function when $b = c = 1$, $n = 1$ and

$$(i) r = -s = -1, (ii) r = -\frac{1}{2}s = -1.$$

For what values of θ does the procedure with $r = -s = -1$ have smaller risk than the procedure with $r = -\frac{1}{2}s = -1$?

4. Stratified sampling. We want to estimate the mean $\mu = E(X)$ of a population that has been divided (stratified) into s mutually exclusive parts (strata) (e.g., geographic locations or age groups). Within the j th stratum we have a sample of i.i.d. random variables X_{1j}, \dots, X_{n_jj} ; $j = 1, \dots, s$, and a stratum sample mean \bar{X}_j ; $j = 1, \dots, s$. We assume that the s samples from different strata are independent. Suppose that the j th stratum has $100p_j\%$ of the population and that the j th stratum population mean and variances are μ_j and σ_j^2 . Let $N = \sum_{j=1}^s n_j$ and consider the two estimators

$$\hat{\mu}_1 = N^{-1} \sum_{j=1}^s \sum_{i=1}^{n_j} X_{ij}, \quad \hat{\mu}_2 = \sum_{j=1}^s p_j \bar{X}_j$$

where we assume that p_j , $1 \leq j \leq s$, are known.

(a) Compute the biases, variances, and MSEs of $\hat{\mu}_1$ and $\hat{\mu}_2$. How should n_j , $0 \leq j \leq s$, be chosen to make $\hat{\mu}_1$ unbiased?

(b) *Neyman allocation.* Assume that $0 < \sigma_j^2 < \infty$, $1 \leq j \leq s$, are known (estimates will be used in a later chapter). Show that the strata sample sizes that minimize $MSE(\hat{\mu}_2)$ are given by

$$n_k = N \frac{p_k \sigma_k}{\sum_{j=1}^s p_j \sigma_j}, \quad k = 1, \dots, s. \quad (1.7.3)$$

Hint: You may use a Lagrange multiplier.

(c) Show that $MSE(\hat{\mu}_1)$ with $n_k = p_k N$ minus $MSE(\hat{\mu}_2)$ with n_k given by (1.7.3) is $N^{-1} \sum_{j=1}^s p_j (\sigma_j - \bar{\sigma})^2$, where $\bar{\sigma} = \sum_{j=1}^s p_j \sigma_j$.

5. Let \bar{X}_b and \hat{X}_b denote the sample mean and the sample median of the sample $X_1 - b, \dots, X_n - b$. If the parameters of interest are the population mean and median of $X_i - b$, respectively, show that $MSE(\bar{X}_b)$ and $MSE(\hat{X}_b)$ are the same for all values of b (the MSEs of the sample mean and sample median are *invariant* with respect to shift).

6. Suppose that X_1, \dots, X_n are i.i.d. as $X \sim F$, that \hat{X} is the median of the sample, and that n is odd. We want to estimate “the” median ν of F , where ν is defined as a value satisfying $P(X \leq \nu) \geq \frac{1}{2}$ and $P(X \geq \nu) \geq \frac{1}{2}$.

(a) Find the MSE of \hat{X} when

- (i) F is discrete with $P(X = a) = P(X = c) = p$, $P(X = b) = 1 - 2p$, $0 < p < 1$, $a < b < c$.

Hint: Use Problem 1.3.5. The answer is $MSE(\hat{X}) = [(a-b)^2 + (c-b)^2]P(S \geq k)$ where $k = .5(n+1)$ and $S \sim \mathcal{B}(n, p)$.

- (ii) F is uniform, $\mathcal{U}(0, 1)$.

Hint: See Problem B.2.9.

- (iii) F is normal, $\mathcal{N}(0, 1)$, $n = 1, 5, 25, 75$.

Hint: See Problem B.2.13. Use a numerical integration package.

(b) Compute the relative risk $RR = MSE(\hat{X})/MSE(\bar{X})$ in question (i) when $b = 0$, $a = -\Delta$, $b = \Delta$, $p = .20, .40$, and $n = 1, 5, 15$.

(c) Same as (b) except when $n = 15$, plot RR for $p = .1, .2, .3, .4, .45$.

(d) Find $E|\hat{X} - b|$ for the situation in (i). Also find $E|\bar{X} - b|$ when $n = 1$, and 2 and compare it to $E|\hat{X} - b|$.

(e) Compute the relative risks $MSE(\hat{X})/MSE(\bar{X})$ in questions (ii) and (iii).

7. Let X_1, \dots, X_n be a sample from a population with values

$$\theta - 2\Delta, \theta - \Delta, \theta, \theta + \Delta, \theta + 2\Delta; \Delta > 0.$$

Each value has probability .2. Let \bar{X} and \hat{X} denote the sample mean and median. Suppose that n is odd.

(a) Find $MSE(\hat{X})$ and the relative risk $RR = MSE(\hat{X})/MSE(\bar{X})$.

(b) Evaluate RR when $n = 1, 3, 5$.

Hint: By Problem 1.3.5, set $\theta = 0$ without loss of generality. Next note that the distribution of \hat{X} involves Bernoulli and multinomial trials.

8. Let X_1, \dots, X_n be a sample from a population with variance σ^2 , $0 < \sigma^2 < \infty$.

(a) Show that $s^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ is an unbiased estimator of σ^2 .

Hint: Write $(X_i - \bar{X})^2 = ([X_i - \mu] - [\bar{X} - \mu])^2$, then expand $(X_i - \bar{X})^2$ keeping the square brackets intact.

(b) Suppose $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

(i) Show that $MSE(s^2) = 2(n-1)^{-1} \sigma^4$.

(ii) Let $\hat{\sigma}_c^2 = c \sum_{i=1}^n (X_i - \bar{X})^2$. Show that the value of c that minimizes $MSE(\hat{\sigma}_c^2)$ is $c = (n+1)^{-1}$.

Hint for question (b): Recall (Theorem B.3.3) that $\sigma^{-2} \sum_{i=1}^n (X_i - \bar{X})^2$ has a χ_{n-1}^2 distribution. You may use the fact that $E(X_i - \mu)^4 = 3\sigma^4$.

9. Let θ denote the proportion of people working in a company who have a certain characteristic (e.g., being left-handed). It is known that in the state where the company is located, 10% have the characteristic. A person in charge of ordering equipment needs to estimate θ and uses

$$\hat{\theta} = (.2)(.10) + (.8)\hat{p}$$

where $\hat{p} = X/n$ is the proportion with the characteristic in a sample of size n from the company. Find $MSE(\hat{\theta})$ and $MSE(\hat{p})$. If the true θ is θ_0 , for what θ_0 is

$$MSE(\hat{\theta})/MSE(\hat{p}) < 1?$$

Give the answer for $n = 25$ and $n = 100$.

10. In Problem 1.3.3(a) with $b = c = 1$ and $n = 1$, suppose θ is discrete with frequency function $\pi(0) = \pi(-\frac{1}{2}) = \pi(\frac{1}{2}) = \frac{1}{3}$. Compute the Bayes risk of $\delta_{r,s}$ when

(a) $r = -s = -1$

(b) $r = -\frac{1}{2}s = -1$.

Which one of the rules is the better one from the Bayes point of view?

11. A decision rule δ is said to be *unbiased* if

$$E_{\theta}(l(\theta, \delta(\mathbf{X}))) \leq E_{\theta}(l(\theta', \delta(\mathbf{X})))$$

for all $\theta, \theta' \in \Theta$.

(a) Show that if θ is real and $l(\theta, a) = (\theta - a)^2$, then this definition coincides with the definition of an unbiased estimate of θ .

(b) Show that if we use the 0-1 loss function in testing, then a test function is unbiased in this sense if, and only if, the *power function*, defined by $\beta(\theta, \delta) = E_{\theta}(\delta(\mathbf{X}))$, satisfies

$$\beta(\theta', \delta) \geq \sup\{\beta(\theta, \delta) : \theta \in \Theta_0\},$$

for all $\theta' \in \Theta_1$.

12. In Problem 1.3.3, show that if $c \leq b$, $z > 0$, and

$$r = -s = -z \left(\frac{b}{b+c} \right) / \sqrt{n},$$

then $\delta_{r,s}$ is unbiased

13. A (behavioral) *randomized test* of a hypothesis H is defined as any statistic $\varphi(\mathbf{X})$ such that $0 \leq \varphi(\mathbf{X}) \leq 1$. The interpretation of φ is the following. If $\mathbf{X} = \mathbf{x}$ and $\varphi(\mathbf{x}) = 0$ we decide Θ_0 , if $\varphi(\mathbf{x}) = 1$, we decide Θ_1 ; but if $0 < \varphi(\mathbf{x}) < 1$, we perform a Bernoulli trial with probability $\varphi(\mathbf{x})$ of success and decide Θ_1 if we obtain a success and decide Θ_0 otherwise.

Define the nonrandomized test δ_u , $0 < u < 1$, by

$$\begin{aligned} \delta_u(\mathbf{X}) &= 1 && \text{if } \varphi(\mathbf{X}) \geq u \\ &= 0 && \text{if } \varphi(\mathbf{X}) < u. \end{aligned}$$

Suppose that $U \sim \mathcal{U}(0, 1)$ and is independent of \mathbf{X} . Consider the following randomized test δ : Observe U . If $U = u$, use the test δ_u . Show that δ agrees with φ in the sense that,

$$P_\theta[\delta(\mathbf{X}) = 1] = 1 - P_\theta[\delta(\mathbf{X}) = 0] = E_\theta(\varphi(\mathbf{X})).$$

14. *Convexity of the risk set.* Suppose that the set of decision procedures is finite. Show that if δ_1 and δ_2 are two randomized procedures, then, given $0 < \alpha < 1$, there is a randomized procedure δ_3 such that $R(\theta, \delta_3) = \alpha R(\theta, \delta_1) + (1 - \alpha)R(\theta, \delta_2)$ for all θ .

15. Suppose that $P_{\theta_0}(B) = 0$ for some event B implies that $P_\theta(B) = 0$ for all $\theta \in \Theta$. Further suppose that $l(\theta_0, a_0) = 0$. Show that the procedure $\delta(X) \equiv a_0$ is admissible.

16. In Example 1.3.4, find the set of μ where $MSE(\hat{\mu}) \leq MSE(\bar{X})$. Your answer should depend on n, σ^2 and $\delta = |\mu - \mu_0|$.

17. In Example 1.3.4, consider the estimator

$$\hat{\mu}_w = w\mu_0 + (1 - w)\bar{X}.$$

If n, σ^2 and $\delta = |\mu - \mu_0|$ are known,

- (a) find the value of w_0 that minimizes $MSE(\hat{\mu}_w)$,
- (b) find the minimum relative risk of $\hat{\mu}_{w_0}$ to \bar{X} .

18. For Example 1.1.1, consider the loss function (1.3.1) and let δ_k be the decision rule “reject the shipment iff $X \geq k$.”

- (a) Show that the risk is given by (1.3.7).
- (b) If $N = 10$, $s = r = 1$, $\theta_0 = .1$, and $k = 3$, plot $R(\theta, \delta_k)$ as a function of θ .
- (c) Same as (b) except $k = 2$. Compare δ_2 and δ_3 .

19. Consider a decision problem with the possible states of nature θ_1 and θ_2 , and possible actions a_1 and a_2 . Suppose the loss function $\ell(\theta, a)$ is

$\theta \backslash a$	a_1	a_2
θ_1	0	2
θ_2	3	1

Let X be a random variable with probability function $p(x | \theta)$

$\theta \backslash x$	0	1
θ_1	0.2	0.8
θ_2	0.4	0.6

(a) Compute and plot the risk points of the nonrandomized decision rules. Give the minimax rule among the nonrandomized decision rules.

(b) Give and plot the risk set S . Give the minimax rule among the randomized decision rules.

(c) Suppose θ has the prior distribution defined by $\pi(\theta_1) = 0.1$, $\pi(\theta_2) = 0.9$. What is the Bayes decision rule?

Problems for Section 1.4

1. An urn contains four red and four black balls. Four balls are drawn at random without replacement. Let Z be the number of red balls obtained in the first two draws and Y the total number of red balls drawn.

(a) Find the best predictor of Y given Z , the best linear predictor, and the best zero intercept linear predictor.

(b) Compute the MSPEs of the predictors in (a).

2. In Example 1.4.1 calculate explicitly the best zero intercept linear predictor, its MSPE, and the ratio of its MSPE to that of the best and best linear predictors.

3. In Problem B.1.7 find the best predictors of Y given X and of X given Y and calculate their MSPEs.

4. Let U_1, U_2 be independent standard normal random variables and set $Z = U_1^2 + U_2^2$, $Y = U_1$. Is Z of any value in predicting Y ?

5. Give an example in which the best linear predictor of Y given Z is a constant (has no predictive value) whereas the best predictor of Y given Z predicts Y perfectly.

6. Give an example in which Z can be used to predict Y perfectly, but Y is of no value in predicting Z in the sense that $\text{Var}(Z | Y) = \text{Var}(Z)$.

7. Let Y be any random variable and let $R(c) = E(|Y - c|)$ be the *mean absolute prediction error*. Show that either $R(c) = \infty$ for all c or $R(c)$ is minimized by taking c to be any number such that $P[Y \geq c] \geq \frac{1}{2}$, $P[Y \leq c] \geq \frac{1}{2}$. A number satisfying these restrictions is called a *median* of (the distribution of) Y . The midpoint of the interval of such c is called the conventionally defined median or simply just *the median*.

Hint: If $c < c_0$,

$$E|Y - c_0| = E|Y - c| + (c - c_0)\{P[Y \geq c_0] - P[Y < c_0]\} + 2E[(c - Y)1[c < Y < c_0]].$$

8. Let Y have a $\mathcal{N}(\mu, \sigma^2)$ distribution. **(a)** Show that $E(|Y - c|) = \sigma \mathcal{Q}[|c - \mu|/\sigma]$ where $\mathcal{Q}(t) = 2[\varphi(t) + t\Phi(t)] - t$.

(b) Show directly that μ minimizes $E(|Y - c|)$ as a function of c .

9. If Y and Z are any two random variables, exhibit a best predictor of Y given Z for mean absolute prediction error.

10. Suppose that Z has a density p , which is symmetric about c , $p(c + z) = p(c - z)$ for all z . Show that c is a median of Z .

11. Show that if (Z, Y) has a bivariate normal distribution the best predictor of Y given Z in the sense of MSPE coincides with the best predictor for mean absolute error.

12. Many observed biological variables such as height and weight can be thought of as the sum of unobservable genetic and environmental variables. Suppose that Z, Y are measurements on such a variable for a randomly selected father and son. Let Z', Z'', Y', Y'' be the corresponding genetic and environmental components $Z = Z' + Z'', Y = Y' + Y''$, where (Z', Y') have a $\mathcal{N}(\mu, \mu, \sigma^2, \sigma^2, \rho)$ distribution and Z'', Y'' are $\mathcal{N}(\nu, \tau^2)$ variables independent of each other and of (Z', Y') .

(a) Show that the relation between Z and Y is weaker than that between Z' and Y' ; that is, $|\text{Corr}(Z, Y)| < |\rho|$.

(b) Show that the error of prediction (for the best predictor) incurred in using Z to predict Y is greater than that incurred in using Z' to predict Y' .

13. Suppose that Z has a density p , which is symmetric about c and which is *unimodal*; that is, $p(z)$ is nonincreasing for $z \geq c$.

(a) Show that $P[|Z - t| \leq s]$ is maximized as a function of t for each $s > 0$ by $t = c$.

(b) Suppose (Z, Y) has a bivariate normal distribution. Suppose that if we observe $Z = z$ and predict $\mu(z)$ for Y our loss is 1 unit if $|\mu(z) - Y| > s$, and 0 otherwise. Show that the predictor that minimizes our expected loss is again the best MSPE predictor.

14. Let Z_1 and Z_2 be independent and have exponential distributions with density $\lambda e^{-\lambda z}$, $z > 0$. Define $Z = Z_2$ and $Y = Z_1 + Z_1 Z_2$. Find

(a) The best MSPE predictor $E(Y | Z = z)$ of Y given $Z = z$

(b) $E(E(Y | Z))$

(c) $\text{Var}(E(Y | Z))$

(d) $\text{Var}(Y | Z = z)$

(e) $E(\text{Var}(Y | Z))$

(f) The best linear MSPE predictor of Y based on $Z = z$.

Hint: Recall that $E(Z_1) = E(Z_2) = 1/\lambda$ and $\text{Var}(Z_1) = \text{Var}(Z_2) = 1/\lambda^2$.

15. Let $\mu(\mathbf{z}) = E(Y | \mathbf{Z} = \mathbf{z})$. Show that

$$\text{Var}(\mu(\mathbf{Z}))/\text{Var}(Y) = \text{Corr}^2(Y, \mu(\mathbf{Z})) = \max_g \text{Corr}^2(Y, g(\mathbf{Z}))$$

where $g(\mathbf{Z})$ stands for any predictor.

16. Show that $\rho_{\mathbf{Z}Y}^2 = \text{Corr}^2(Y, \mu_L(\mathbf{Z})) = \max_{g \in \mathcal{L}} \text{Corr}^2(Y, g(\mathbf{Z}))$ where \mathcal{L} is the set of linear predictors.

17. One minus the ratio of the smallest possible MSPE to the MSPE of the constant predictor is called *Pearson's correlation ratio* $\eta_{\mathbf{Z}Y}^2$; that is,

$$\eta_{\mathbf{Z}Y}^2 = 1 - E[Y - \mu(\mathbf{Z})]^2 / \text{Var}(Y) = \text{Var}(\mu(\mathbf{Z})) / \text{Var}(Y).$$

(See Pearson, 1905, and Doksum and Samarov, 1995, on estimation of $\eta_{\mathbf{Z}Y}^2$.)

(a) Show that $\eta_{\mathbf{Z}Y}^2 \geq \rho_{\mathbf{Z}Y}^2$, where $\rho_{\mathbf{Z}Y}^2$ is the population multiple correlation coefficient of Remark 1.4.3.

Hint: See Problem 1.4.15.

(b) Show that if Z is one-dimensional and h is a 1-1 increasing transformation of Z , then $\eta_{h(Z)Y}^2 = \eta_{ZY}^2$. That is, η^2 is invariant under such h .

(c) Let $\epsilon_L = Y - \mu_L(\mathbf{Z})$ be the linear prediction error. Show that, in the linear model of Remark 1.4.4, ϵ_L is uncorrelated with $\mu_L(\mathbf{Z})$ and $\eta_{\mathbf{Z}Y}^2 = \rho_{\mathbf{Z}Y}^2$.

18. *Predicting the past from the present.* Consider a subject who walks into a clinic today, at time t , and is diagnosed with a certain disease. At the same time t a diagnostic indicator Z_0 of the severity of the disease (e.g., a blood cell or viral load measurement) is obtained. Let S be the unknown date in the past when the subject was infected. We are interested in the time $Y_0 = t - S$ from infection until detection. Assume that the conditional density of Z_0 (the present) given $Y_0 = y_0$ (the past) is

$$\mathcal{N}(\mu + \beta y_0, \sigma^2),$$

where μ and σ^2 are the mean and variance of the severity indicator Z_0 in the population of people without the disease. Here βy_0 gives the mean increase of Z_0 for infected subjects over the time period y_0 ; $\beta > 0$, $y_0 > 0$. It will be convenient to rescale the problem by introducing $Z = (Z_0 - \mu)/\sigma$ and $Y = \beta Y_0/\sigma$.

(a) Show that the conditional density $f(z | y)$ of Z given $Y = y$ is $\mathcal{N}(y, 1)$.

(b) Suppose that Y has the exponential density

$$\pi(y) = \lambda \exp\{-\lambda y\}, \quad \lambda > 0, \quad y > 0.$$

Show that the conditional distribution of Y (the past) given $Z = z$ (the present) has density

$$\pi(y | z) = (2\pi)^{-\frac{1}{2}} c^{-1} \exp \left\{ -\frac{1}{2} [y - (z - \lambda)]^2 \right\}, \quad y > 0$$

where $c = \Phi(z - \lambda)$. This density is called the *truncated (at zero) normal*, $\mathcal{N}(z - \lambda, 1)$, density.

Hint: Use Bayes' Theorem.

(c) Find the conditional density $\pi_0(y_0 | z_0)$ of Y_0 given $Z_0 = z_0$.

(d) Find the best predictor of Y_0 given $Z_0 = z_0$ using mean absolute prediction error $E|Y_0 - g(Z_0)|$.

Hint: See Problems 1.4.7 and 1.4.9.

(e) Show that the best MSPE predictor of Y given $Z = z$ is

$$E(Y | Z = z) = c^{-1} \varphi(\lambda - z) - (\lambda - z).$$

(In practice, all the unknowns, including the “prior” π , need to be estimated from cohort studies; see Berman, 1990, and Normand and Doksum, 2001).

19. Establish 1.4.14 by setting the derivatives of $R(a, b)$ equal to zero, solving for (a, b) , and checking convexity.

20. Let Y be the number of heads showing when X fair coins are tossed, where X is the number of spots showing when a fair die is rolled. Find

(a) The mean and variance of Y .

(b) The MSPE of the optimal predictor of Y based on X .

(c) The optimal predictor of Y given $X = x$, $x = 1, \dots, 6$.

21. Let \mathbf{Y} be a vector and let $r(\mathbf{Y})$ and $s(\mathbf{Y})$ be real valued. Write $\text{Cov}[r(\mathbf{Y}), s(\mathbf{Y}) | \mathbf{z}]$ for the covariance between $r(\mathbf{Y})$ and $s(\mathbf{Y})$ in the conditional distribution of $(r(\mathbf{Y}), s(\mathbf{Y}))$ given $\mathbf{Z} = \mathbf{z}$.

(a) Show that if $\text{Cov}[r(\mathbf{Y}), s(\mathbf{Y})] < \infty$, then

$$\text{Cov}[r(\mathbf{Y}), s(\mathbf{Y})] = E\{\text{Cov}[r(\mathbf{Y}), s(\mathbf{Y}) | \mathbf{Z}]\} + \text{Cov}\{E[r(\mathbf{Y}) | \mathbf{Z}], E[s(\mathbf{Y}) | \mathbf{Z}]\}.$$

(b) Show that (a) is equivalent to (1.4.6) when $r = s$.

(c) Show that if Z is real, $\text{Cov}[r(\mathbf{Y}), Z] = \text{Cov}\{E[r(\mathbf{Y}) | Z], Z\}$.

(d) Suppose $Y_1 = a_1 + b_1 Z_1 + W$ and $Y_2 = a_2 + b_2 Z_2 + W$, where Y_1 and Y_2 are responses of subjects 1 and 2 with common influence W and separate influences Z_1 and Z_2 , where Z_1, Z_2 and W are independent with finite variances. Find $\text{Corr}(Y_1, Y_2)$ using (a).

(e) In the preceding model (d), if $b_1 = b_2$ and Z_1, Z_2 and W have the same variance σ^2 , we say that there is a 50% overlap between Y_1 and Y_2 . In this case what is $\text{Corr}(Y_1, Y_2)$?

(f) In model (d), suppose that Z_1 and Z_2 are $\mathcal{N}(\mu, \sigma^2)$ and $W \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Find the optimal predictor of Y_2 given (Y_1, Z_1, Z_2) .

22. In Example 1.4.3, show that the MSPE of the optimal predictor is $\sigma_Y^2(1 - \rho_{ZY}^2)$.

23. Verify that solving (1.4.15) yields (1.4.14).

24. (a) Let $w(y, \mathbf{z})$ be a positive real-valued function. Then $[y - g(\mathbf{z})]^2/w(y, \mathbf{z}) = \delta_w(y, g(\mathbf{z}))$ is called *weighted squared prediction error*. Show that the mean weighted squared prediction error is minimized by $\mu_0(\mathbf{Z}) = E_0(Y | \mathbf{Z})$, where

$$p_0(y, \mathbf{z}) = cp(y, \mathbf{z})/w(y, \mathbf{z})$$

and c is the constant that makes p_0 a density. Assume that

$$E\delta_w(Y, g(\mathbf{Z})) < \infty$$

for some g and that p_0 is a density.

(b) Suppose that given $Y = y$, $Z \sim \mathcal{B}(n, y)$, $n \geq 2$, and suppose that Y has the beta, $\beta(r, s)$, density. Find $\mu_0(Z)$ when (i) $w(y, z) = 1$, and (ii) $w(y, z) = y(1 - y)$, $0 < y < 1$.
Hint: See Example 1.2.9.

25. Show that $EY^2 < \infty$ if and only if $E(Y - c)^2 < \infty$ for all c .

Hint: Whatever be Y and c ,

$$\frac{1}{2}Y^2 - c^2 \leq (Y - c)^2 = Y^2 - 2cY + c^2 \leq 2(Y^2 + c^2).$$

Problems for Section 1.5

1. Let X_1, \dots, X_n be a sample from a Poisson, $\mathcal{P}(\theta)$, population where $\theta > 0$.

(a) Show directly that $\sum_{i=1}^n X_i$ is sufficient for θ .

(b) Establish the same result using the factorization theorem.

2. Let n items be drawn in order without replacement from a shipment of N items of which $N\theta$ are bad. Let $X_i = 1$ if the i th item drawn is bad, and $= 0$ otherwise. Show that $\sum_{i=1}^n X_i$ is sufficient for θ directly and by the factorization theorem.

3. Suppose X_1, \dots, X_n is a sample from a population with one of the following densities.

(a) $p(x, \theta) = \theta x^{\theta-1}$, $0 < x < 1$, $\theta > 0$. This is the beta, $\beta(\theta, 1)$, density.

(b) $p(x, \theta) = \theta a x^{a-1} \exp(-\theta x^a)$, $x > 0$, $\theta > 0$, $a > 0$.

This is known as the *Weibull* density.

(c) $p(x, \theta) = \theta a^\theta / x^{(\theta+1)}$, $x > a$, $\theta > 0$, $a > 0$.

This is known as the *Pareto* density.

In each case, find a real-valued sufficient statistic for θ , a fixed.

4. (a) Show that T_1 and T_2 are equivalent statistics if, and only if, we can write $T_2 = H(T_1)$ for some 1-1 transformation H of the range of T_1 into the range of T_2 . Which of the following statistics are equivalent? (Prove or disprove.)

(b) $\prod_{i=1}^n x_i$ and $\sum_{i=1}^n \log x_i$, $x_i > 0$

(c) $\sum_{i=1}^n x_i$ and $\sum_{i=1}^n \log x_i$, $x_i > 0$

(d) $(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2)$ and $(\sum_{i=1}^n x_i, \sum_{i=1}^n (x_i - \bar{x})^2)$

(e) $(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^3)$ and $(\sum_{i=1}^n x_i, \sum_{i=1}^n (x_i - \bar{x})^3)$.

5. Let $\theta = (\theta_1, \theta_2)$ be a bivariate parameter. Suppose that $T_1(\mathbf{X})$ is sufficient for θ_1 whenever θ_2 is fixed and known, whereas $T_2(\mathbf{X})$ is sufficient for θ_2 whenever θ_1 is fixed and known. Assume that θ_1, θ_2 vary independently, $\theta_1 \in \Theta_1$, $\theta_2 \in \Theta_2$ and that the set $S = \{\mathbf{x} : p(\mathbf{x}, \theta) > 0\}$ does not depend on θ .

(a) Show that if T_1 and T_2 do not depend on θ_2 and θ_1 respectively, then $(T_1(\mathbf{X}), T_2(\mathbf{X}))$ is sufficient for θ .

(b) Exhibit an example in which $(T_1(\mathbf{X}), T_2(\mathbf{X}))$ is sufficient for θ , $T_1(\mathbf{X})$ is sufficient for θ_1 whenever θ_2 is fixed and known, but $T_2(\mathbf{X})$ is not sufficient for θ_2 , when θ_1 is fixed and known.

6. Let X take on the specified values v_1, \dots, v_k with probabilities $\theta_1, \dots, \theta_k$, respectively. Suppose that X_1, \dots, X_n are independently and identically distributed as X . Suppose that $\theta = (\theta_1, \dots, \theta_k)$ is unknown and may range over the set $\Theta = \{(\theta_1, \dots, \theta_k) : \theta_i \geq 0, 1 \leq i \leq k, \sum_{i=1}^k \theta_i = 1\}$. Let N_j be the number of X_i which equal v_j .

(a) What is the distribution of (N_1, \dots, N_k) ?

(b) Show that $\mathbf{N} = (N_1, \dots, N_{k-1})$ is sufficient for θ .

7. Let X_1, \dots, X_n be a sample from a population with density $p(x, \theta)$ given by

$$\begin{aligned} p(x, \theta) &= \frac{1}{\sigma} \exp \left\{ - \left(\frac{x - \mu}{\sigma} \right) \right\} \text{ if } x \geq \mu \\ &= 0 \text{ otherwise.} \end{aligned}$$

Here $\theta = (\mu, \sigma)$ with $-\infty < \mu < \infty$, $\sigma > 0$.

(a) Show that $\min(X_1, \dots, X_n)$ is sufficient for μ when σ is fixed.

(b) Find a one-dimensional sufficient statistic for σ when μ is fixed.

(c) Exhibit a two-dimensional sufficient statistic for θ .

8. Let X_1, \dots, X_n be a sample from some continuous distribution F with density f , which is unknown. Treating f as a parameter, show that the order statistics $X_{(1)}, \dots, X_{(n)}$ (cf. Problem B.2.8) are sufficient for f .

9. Let X_1, \dots, X_n be a sample from a population with density

$$\begin{aligned} f_\theta(x) &= a(\theta)h(x) \text{ if } \theta_1 \leq x \leq \theta_2 \\ &= 0 \text{ otherwise} \end{aligned}$$

where $h(x) \geq 0$, $\theta = (\theta_1, \theta_2)$ with $-\infty < \theta_1 \leq \theta_2 < \infty$, and $a(\theta) = \left[\int_{\theta_1}^{\theta_2} h(x) dx \right]^{-1}$ is assumed to exist. Find a two-dimensional sufficient statistic for this problem and apply your result to the $\mathcal{U}[\theta_1, \theta_2]$ family of distributions.

10. Suppose X_1, \dots, X_n are i.i.d. with density $f(x, \theta) = \frac{1}{2}e^{-|x-\theta|}$. Show that $(X_{(1)}, \dots, X_{(n)})$, the order statistics, are minimal sufficient.

Hint: $\frac{\partial}{\partial \theta} \log L_{\mathbf{X}}(\theta) = -\sum_{i=1}^n \text{sgn}(X_i - \theta)$, $\theta \notin \{X_1, \dots, X_n\}$, which determines $X_{(1)}, \dots, X_{(n)}$.

11. Let X_1, X_2, \dots, X_n be a sample from the uniform, $\mathcal{U}(0, \theta)$, distribution. Show that $X_{(n)} = \max\{X_i; 1 \leq i \leq n\}$ is minimal sufficient for θ .

12. *Dynkin, Lehmann, Scheffé's Theorem.* Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ where P_θ is discrete concentrated on $\mathcal{X} = \{x_1, x_2, \dots\}$. Let $p(x, \theta) \equiv P_\theta[X = x] \equiv L_x(\theta) > 0$ on \mathcal{X} . Show that $\frac{L_{\mathbf{X}}(\cdot)}{L_{\mathbf{X}}(\theta_0)}$ is minimal sufficient.

Hint: Apply the factorization theorem.

13. Suppose that $\mathbf{X} = (X_1, \dots, X_n)$ is a sample from a population with continuous distribution function $F(x)$. If $F(x)$ is $N(\mu, \sigma^2)$, $T(\mathbf{X}) = (\bar{X}, \hat{\sigma}^2)$, where $\hat{\sigma}^2 = n^{-1} \sum (X_i - \bar{X})^2$, is sufficient, and $S(\mathbf{X}) = (X'_{(1)}, \dots, X'_{(n)})$, where $X'_{(i)} = (X_{(i)} - \bar{X})/\hat{\sigma}$, is “irrelevant” (ancillary) for (μ, σ^2) . However, $S(\mathbf{X})$ is exactly what is needed to estimate the “shape” of $F(x)$ when $F(x)$ is unknown. The shape of F is represented by the equivalence class $\mathcal{F} = \{F((\cdot - a)/b) : b > 0, a \in R\}$. Thus a distribution G has the same shape as F iff $G \in \mathcal{F}$. For instance, one “estimator” of this shape is the scaled empirical distribution function

$$\begin{aligned} \hat{F}_s(x) &= j/n, \quad x'_{(j)} \leq x < x'_{(j+1)}, \quad j = 1, \dots, n-1 \\ &= 0, \quad x < x'_{(1)} \\ &= 1, \quad x \geq x'_{(n)}. \end{aligned}$$

Show that for fixed x , $\hat{F}_s((x - \bar{x})/\hat{\sigma})$ converges in probability to $F(x)$. Here we are using F to represent \mathcal{F} because every member of \mathcal{F} can be obtained from F .

14. *Kolmogorov's Theorem.* We are given a regular model with Θ finite.

(a) Suppose that a statistic $T(\mathbf{X})$ has the property that for any prior distribution on θ , the posterior distribution of θ depends on \mathbf{x} only through $T(\mathbf{x})$. Show that $T(\mathbf{X})$ is sufficient.

(b) Conversely show that if $T(\mathbf{X})$ is sufficient, then, for any prior distribution, the posterior distribution depends on \mathbf{x} only through $T(\mathbf{x})$.

Hint: Apply the factorization theorem.

15. Let X_1, \dots, X_n be a sample from $f(x - \theta)$, $\theta \in R$. Show that the order statistics are minimal sufficient when f is the density *Cauchy* $f(t) = 1/\pi(1 + t^2)$.

16. Let $X_1, \dots, X_m; Y_1, \dots, Y_n$ be independently distributed according to $\mathcal{N}(\mu, \sigma^2)$ and $\mathcal{N}(\eta, \tau^2)$, respectively. Find minimal sufficient statistics for the following three cases:

- (i) μ, η, σ, τ are arbitrary: $-\infty < \mu, \eta < \infty, 0 < \sigma, \tau$.
- (ii) $\sigma = \tau$ and μ, η, σ are arbitrary.
- (iii) $\mu = \eta$ and μ, σ, τ are arbitrary.

17. In Example 1.5.4, express t_1 as a function of $L_x(0, 1)$ and $L_x(1, 1)$.

Problems to Section 1.6

1. Prove the assertions of Table 1.6.1.

2. Suppose X_1, \dots, X_n is as in Problem 1.5.3. In each of the cases (a), (b) and (c), show that the distribution of \mathbf{X} forms a one-parameter exponential family. Identify η, B, T , and h .

3. Let X be the number of failures before the first success in a sequence of Bernoulli trials with probability of success θ . Then $P_\theta[X = k] = (1 - \theta)^k \theta$, $k = 0, 1, 2, \dots$. This is called the *geometric distribution* ($\mathcal{G}(\theta)$).

(a) Show that the family of geometric distributions is a one-parameter exponential family with $T(x) = x$.

(b) Deduce from Theorem 1.6.1 that if X_1, \dots, X_n is a sample from $\mathcal{G}(\theta)$, then the distributions of $\sum_{i=1}^n X_i$ form a one-parameter exponential family.

(c) Show that $\sum_{i=1}^n X_i$ in part (b) has a *negative binomial* distribution with parameters (n, θ) defined by $P_\theta[\sum_{i=1}^n X_i = k] = \binom{n+k-1}{k} (1 - \theta)^k \theta^n$, $k = 0, 1, 2, \dots$ (The negative binomial distribution is that of the number of failures before the n th success in a sequence of Bernoulli trials with probability of success θ .)

Hint: By Theorem 1.6.1, $P_\theta[\sum_{i=1}^n X_i = k] = c_k (1 - \theta)^k \theta^n$, $0 < \theta < 1$. If

$$\sum_{k=0}^{\infty} c_k \omega^k = \frac{1}{(1 - \omega)^n}, \quad 0 < \omega < 1, \quad \text{then} \quad c_k = \frac{1}{k!} \frac{d^k}{d\omega^k} (1 - \omega)^{-n} \Big|_{\omega=0}.$$

4. Which of the following families of distributions are exponential families? (Prove or disprove.)

- (a) The $\mathcal{U}(0, \theta)$ family

(b) $p(x, \theta) = \{\exp[-2 \log \theta + \log(2x)]\} 1[x \in (0, \theta)]$

(c) $p(x, \theta) = \frac{1}{9}, x \in \{0.1 + \theta, \dots, 0.9 + \theta\}$

(d) The $\mathcal{N}(\theta, \theta^2)$ family, $\theta > 0$

(e) $p(x, \theta) = 2(x + \theta)/(1 + 2\theta), 0 < x < 1, \theta > 0$

(f) $p(x, \theta)$ is the conditional frequency function of a binomial, $\mathcal{B}(n, \theta)$, variable X , given that $X > 0$.

5. Show that the following families of distributions are two-parameter exponential families and identify the functions η, B, T , and h .

(a) The beta family.

(b) The gamma family.

6. Let X have the Dirichlet distribution, $\mathcal{D}(\alpha)$, of Problem 1.2.15.

Show the distribution of X form an r -parameter exponential family and identify η, B, T , and h .

7. Let $\mathbf{X} = ((X_1, Y_1), \dots, (X_n, Y_n))$ be a sample from a bivariate normal population. Show that the distributions of \mathbf{X} form a five-parameter exponential family and identify η, B, T , and h .

8. Show that the family of distributions of Example 1.5.3 is not a one parameter exponential family.

Hint: If it were, there would be a set A such that $p(x, \theta) > 0$ on A for all θ .

9. Prove the analogue of Theorem 1.6.1 for discrete k -parameter exponential families.

10. Suppose that $f(x, \theta)$ is a positive density on the real line, which is continuous in x for each θ and such that if (X_1, X_2) is a sample of size 2 from $f(\cdot, \theta)$, then $X_1 + X_2$ is sufficient for θ . Show that $f(\cdot, \theta)$ corresponds to a one-parameter exponential family of distributions with $T(x) = x$.

Hint: There exist functions $g(t, \theta), h(x_1, x_2)$ such that $\log f(x_1, \theta) + \log f(x_2, \theta) = g(x_1 + x_2, \theta) + h(x_1, x_2)$. Fix θ_0 and let $r(x, \theta) = \log f(x, \theta) - \log f(x, \theta_0)$, $q(x, \theta) = g(x, \theta) - g(x, \theta_0)$. Then, $q(x_1 + x_2, \theta) = r(x_1, \theta) + r(x_2, \theta)$, and hence, $[r(x_1, \theta) - r(0, \theta)] + [r(x_2, \theta) - r(0, \theta)] = r(x_1 + x_2, \theta) - r(0, \theta)$.

11. Use Theorems 1.6.2 and 1.6.3 to obtain moment-generating functions for the sufficient statistics when sampling from the following distributions.

(a) normal, $\theta = (\mu, \sigma^2)$

(b) gamma, $\Gamma(p, \lambda), \theta = \lambda, p$ fixed

(c) binomial

(d) Poisson

(e) negative binomial (see Problem 1.6.3)

(f) gamma, $\Gamma(p, \lambda), \theta = (p, \lambda)$.

12. Show directly using the definition of the rank of an exponential family that the multinomial distribution, $\mathcal{M}(n; \theta_1, \dots, \theta_k)$, $0 < \theta_j < 1$, $1 \leq j \leq k$, $\sum_{j=1}^k \theta_j = 1$, is of rank $k - 1$.

13. Show that in Theorem 1.6.3, the condition that \mathcal{E} has nonempty interior is equivalent to the condition that \mathcal{E} is not contained in any $(k - 1)$ -dimensional hyperplane.

14. Construct an exponential family of rank k for which \mathcal{E} is not open and \dot{A} is not defined on all of \mathcal{E} . Show that if $k = 1$ and $\mathcal{E}^0 \neq \emptyset$ and \dot{A}, \ddot{A} are defined on all of \mathcal{E} , then Theorem 1.6.3 continues to hold.

15. Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ where P_θ is discrete and concentrated on $\mathcal{X} = \{x_1, x_2, \dots\}$, and let $p(x, \theta) = P_\theta[X = x]$. Show that if \mathcal{P} is a (discrete) canonical exponential family generated by (\mathbf{T}, h) and $\mathcal{E}^0 \neq \emptyset$, then \mathbf{T} is minimal sufficient.

Hint: $\frac{\partial \log L_X(\boldsymbol{\eta})}{\partial \eta_j} = T_j(X) - E_{\boldsymbol{\eta}} T_j(X)$. Use Problem 1.5.12.

16. Life testing. Let X_1, \dots, X_n be independently distributed with exponential density $(2\theta)^{-1}e^{-x/2\theta}$ for $x \geq 0$, and let the ordered X 's be denoted by $Y_1 \leq Y_2 \leq \dots \leq Y_n$. It is assumed that Y_1 becomes available first, then Y_2 , and so on, and that observation is continued until Y_r has been observed. This might arise, for example, in life testing where each X measures the length of life of, say, an electron tube, and n tubes are being tested simultaneously. Another application is to the disintegration of radioactive material, where n is the number of atoms, and observation is continued until r α -particles have been emitted. Show that

(i) The joint distribution of Y_1, \dots, Y_r is an exponential family with density

$$\frac{1}{(2\theta)^r} \frac{n!}{(n-r)!} \exp \left[-\frac{\sum_{i=1}^r y_i + (n-r)y_r}{2\theta} \right], \quad 0 \leq y_1 \leq \dots \leq y_r.$$

(ii) The distribution of $[\sum_{i=1}^r Y_i + (n-r)Y_r]/\theta$ is χ^2 with $2r$ degrees of freedom.

(iii) Let Y_1, Y_2, \dots denote the time required until the first, second, ... event occurs in a Poisson process with parameter $1/2\theta'$ (see A.16). Then $Z_1 = Y_1/\theta'$, $Z_2 = (Y_2 - Y_1)/\theta'$, $Z_3 = (Y_3 - Y_2)/\theta'$, ... are independently distributed as χ^2 with 2 degrees of freedom, and the joint density of Y_1, \dots, Y_r is an exponential family with density

$$\frac{1}{(2\theta')^r} \exp \left(-\frac{y_r}{2\theta'} \right), \quad 0 \leq y_1 \leq \dots \leq y_r.$$

The distribution of Y_r/θ' is again χ^2 with $2r$ degrees of freedom.

(iv) The same model arises in the application to life testing if the number n of tubes is held constant by replacing each burned-out tube with a new one, and if Y_1 denotes the time at which the first tube burns out, Y_2 the time at which the second tube burns out, and so on, measured from some fixed time. The lifetimes are assumed to be exponentially distributed.

Hint (ii): The random variables $Z_i = (n - i + 1)(Y_i - Y_{i-1})/\theta$ ($i = 1, \dots, r$) are independently distributed as χ^2_2 , $Y_0 = 0$, and $[\sum_{i=1}^r Y_i + (n - r)Y_r]/\theta = \sum_{i=1}^r Z_i$.

17. Suppose that $(\mathbf{T}_{k \times 1}, h)$ generate a canonical exponential family \mathcal{P} with parameter $\boldsymbol{\eta}_{k \times 1}$ and $\mathcal{E} = R^k$. Let

$$\mathcal{Q} = \{\mathcal{Q}_\theta : \mathcal{Q}_\theta = P\boldsymbol{\eta} \text{ with } \boldsymbol{\eta} = B_{k \times l}\boldsymbol{\theta}_{l \times 1} + \mathbf{c}_{k \times 1}\}, \quad l \leq k.$$

(a) Show that \mathcal{Q} is the exponential family generated by $\Pi_L \mathbf{T}$ and $h \exp\{c^T \mathbf{T}\}$, where Π_L is the projection matrix onto $\mathcal{L} = \{\boldsymbol{\eta} : \boldsymbol{\eta} = B\boldsymbol{\theta}, \boldsymbol{\theta} \in R^l\}$.

(b) Show that if \mathcal{P} has full rank k and B is of rank l , then \mathcal{Q} has full rank l .

Hint: If B is of rank l , you may assume

$$\Pi_L = B[B^T B]^{-1} B^T.$$

18. Suppose Y_1, \dots, Y_n are independent with $Y_i \sim \mathcal{N}(\beta_1 + \beta_2 z_i, \sigma^2)$, where z_1, \dots, z_n are covariate values not all equal. (See Example 1.6.6.) Show that the family has rank 3. Give the mean vector and the variance matrix of \mathbf{T} .

19. Logistic Regression. We observe $(\mathbf{z}_1, Y_1), \dots, (\mathbf{z}_n, Y_n)$ where the Y_1, \dots, Y_n are independent, $Y_i \sim \mathcal{B}(n_i, \lambda_i)$. The success probability λ_i depends on the characteristics \mathbf{z}_i of the i th subject, for example, on the covariate vector $\mathbf{z}_i = (\text{age, height, blood pressure})^T$. The function $l(u) = \log[u/(1 - u)]$ is called the *logit* function. In the logistic linear regression model it is assumed that $l(\lambda_i) = \mathbf{z}_i^T \boldsymbol{\beta}$ where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^T$ and \mathbf{z}_i is $d \times 1$. Show that $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ follow an exponential model with rank d iff $\mathbf{z}_1, \dots, \mathbf{z}_d$ are not collinear (linearly independent) (cf. Examples 1.1.4, 1.6.8 and Problem 1.1.9).

20. (a) In part II of the proof of Theorem 1.6.4, fill in the details of the arguments that \mathcal{Q} is generated by $(\boldsymbol{\eta}_1 - \boldsymbol{\eta}_0)^T \mathbf{T}$ and that $\sim(\text{ii}) \equiv \sim(\text{i})$.

(b) Fill in the details of part III of the proof of Theorem 1.6.4.

21. Find $\mu(\boldsymbol{\eta}) = E\boldsymbol{\eta}^T \mathbf{T}(X)$ for the gamma, $\Gamma(\alpha, \lambda)$, distribution, where $\boldsymbol{\theta} = (\alpha, \lambda)$.

22. Let X_1, \dots, X_n be a sample from the k -parameter exponential family distribution (1.6.10). Let $\mathbf{T} = (\sum_{i=1}^n T_1(X_i), \dots, \sum_{i=1}^n T_k(X_i))$ and let

$$\mathcal{S} = \{(\eta_1(\boldsymbol{\theta}), \dots, \eta_k(\boldsymbol{\theta})) : \boldsymbol{\theta} \in \Theta\}.$$

Show that if \mathcal{S} contains a subset of $k + 1$ vectors $\mathbf{v}_0, \dots, \mathbf{v}_{k+1}$ so that $\mathbf{v}_i - \mathbf{v}_0$, $1 \leq i \leq k$, are not collinear (linearly independent), then \mathbf{T} is minimally sufficient for $\boldsymbol{\theta}$.

23. Using (1.6.20), find a conjugate family of distributions for the gamma and beta families.

(a) With one parameter fixed.

(b) With both parameters free.

24. Using (1.6.20), find a conjugate family of distributions for the normal family using as parameter $\theta = (\theta_1, \theta_2)$ where $\theta_1 = E_\theta(X)$, $\theta_2 = 1/(\text{Var}_\theta X)$ (cf. Problem 1.2.12).

25. Consider the linear Gaussian regression model of Examples 1.5.5 and 1.6.6 except with σ^2 known. Find a conjugate family of prior distributions for $(\beta_1, \beta_2)^T$.

26. Using (1.6.20), find a conjugate family of distributions for the multinomial distribution. See Problem 1.2.15.

27. Let \mathcal{P} denote the canonical exponential family generated by \mathbf{T} and h . For any $\boldsymbol{\eta}_0 \in \mathcal{E}$, set $h_0(x) = q(x, \boldsymbol{\eta}_0)$ where q is given by (1.6.9). Show that \mathcal{P} is also the canonical exponential family generated by \mathbf{T} and h_0 .

28. *Exponential families are maximum entropy distributions.* The entropy $h(f)$ of a random variable X with density f is defined by

$$h(f) = E(-\log f(X)) = - \int_{-\infty}^{\infty} [\log f(x)] f(x) dx.$$

This quantity arises naturally in information theory; see Section 2.2.2 and Cover and Thomas (1991). Let $S = \{x : f(x) > 0\}$.

(a) Show that the canonical k -parameter exponential family density

$$f(x, \boldsymbol{\eta}) = \exp \left\{ \eta_0 + \sum_{j=1}^k \eta_j r_j(x) - A(\boldsymbol{\eta}) \right\}, \quad x \in S$$

maximizes $h(f)$ subject to the constraints

$$f(x) \geq 0, \quad \int_S f(x) dx = 1, \quad \int_S f(x) r_j(x) dx = \alpha_j, \quad 1 \leq j \leq k,$$

where η_0, \dots, η_k are chosen so that f satisfies the constraints.

Hint: You may use Lagrange multipliers. Maximize the integrand.

(b) Find the maximum entropy densities when $r_j(x) = x^j$ and (i) $S = (0, \infty)$, $k = 1$, $\alpha_1 > 0$; (ii) $S = R$, $k = 2$, $\alpha_1 \in R$, $\alpha_2 > 0$; (iii) $S = R$, $k = 3$, $\alpha_1 \in R$, $\alpha_2 > 0$, $\alpha_3 \in R$.

29. As in Example 1.6.11, suppose that $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ are i.i.d. $\mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu}$ varies freely in R^p and Σ ranges freely over the class of all $p \times p$ symmetric positive definite matrices. Show that the distribution of $\mathbf{Y} = (Y_1, \dots, Y_n)$ is the $p(p+3)/2$ canonical exponential family generated by $h = 1$ and the $p(p+3)/2$ statistics

$$T_j = \sum_{i=1}^n Y_{ij}, \quad 1 \leq j \leq p; \quad T_{jl} = \sum_{i=1}^n Y_{ij} Y_{il}, \quad 1 \leq j \leq l \leq p$$

where $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{ip})$. Show that \mathcal{E} is open and that this family is of rank $p(p+3)/2$.

Hint: Without loss of generality, take $n = 1$. We want to show that $h = 1$ and the $m = p(p+3)/2$ statistics $T_j(\mathbf{Y}) = Y_j$, $1 \leq j \leq p$, and $T_{jl}(\mathbf{Y}) = Y_j Y_l$, $1 \leq j \leq l \leq p$,

generate $\mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$. As Σ ranges over all $p \times p$ symmetric positive definite matrices, so does Σ^{-1} . Next establish that for symmetric matrices M ,

$$\int \exp\{-\mathbf{u}^T M \mathbf{u}\} d\mathbf{u} < \infty \text{ iff } M \text{ is positive definite}$$

by using the spectral decomposition (see B.10.1.2)

$$M = \sum_{j=1}^p \lambda_j \mathbf{e}_j \mathbf{e}_j^T \text{ for } \mathbf{e}_1, \dots, \mathbf{e}_p \text{ orthogonal, } \lambda_j \in R.$$

To show that the family has full rank m , use induction on p to show that if Z_1, \dots, Z_p are i.i.d. $\mathcal{N}(0, 1)$ and if $B_{p \times p} = (b_{jl})$ is symmetric, then

$$P\left(\sum_{j=1}^p a_j Z_j + \sum_{j,l} b_{jl} Z_j Z_l = c\right) = P(\mathbf{a}^T \mathbf{Z} + \mathbf{Z}^T B \mathbf{Z} = c) = 0$$

unless $\mathbf{a} = \mathbf{0}$, $B = 0$, $c = 0$. Next recall (Appendix B.6) that since $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$, then $\mathbf{Y} = S\mathbf{Z}$ for some nonsingular $p \times p$ matrix S .

30. Show that if $\mathbf{X}_1, \dots, \mathbf{X}_n$ are i.i.d. $\mathcal{N}_p(\boldsymbol{\theta}, \Sigma_0)$ given $\boldsymbol{\theta}$ where Σ_0 is known, then the $\mathcal{N}_p(\boldsymbol{\lambda}, \Gamma)$ family is conjugate to $\mathcal{N}_p(\boldsymbol{\theta}, \Sigma_0)$, where $\boldsymbol{\lambda}$ varies freely in R^p and Γ ranges over all $p \times p$ symmetric positive definite matrices.

31. Conjugate Normal Mixture Distributions. A Hierarchical Bayesian Normal Model. Let $\{(\mu_j, \tau_j) : 1 \leq j \leq k\}$ be a given collection of pairs with $\mu_j \in R$, $\tau_j > 0$. Let $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ be a random pair with $\lambda_j = P((\boldsymbol{\mu}, \boldsymbol{\sigma}) = (\mu_j, \tau_j))$, $0 < \lambda_j < 1$, $\sum_{j=1}^k \lambda_j = 1$. Let $\boldsymbol{\theta}$ be a random variable whose conditional distribution given $(\boldsymbol{\mu}, \boldsymbol{\sigma}) = (\mu_j, \tau_j)$ is normal, $\mathcal{N}(\mu_j, \tau_j^2)$. Consider the model $X = \boldsymbol{\theta} + \epsilon$, where $\boldsymbol{\theta}$ and ϵ are independent and $\epsilon \sim \mathcal{N}(0, \sigma_0^2)$, σ_0^2 known. Note that $\boldsymbol{\theta}$ has the prior density

$$\pi(\boldsymbol{\theta}) = \sum_{j=1}^k \lambda_j \varphi_{\tau_j}(\boldsymbol{\theta} - \mu_j) \quad (1.7.4)$$

where φ_τ denotes the $\mathcal{N}(0, \tau^2)$ density. Also note that $(X | \boldsymbol{\theta})$ has the $\mathcal{N}(\boldsymbol{\theta}, \sigma_0^2)$ distribution.

(a) Find the posterior

$$\pi(\boldsymbol{\theta} | x) = \sum_{j=1}^k P((\boldsymbol{\mu}, \boldsymbol{\sigma}) = (\mu_j, \tau_j) | x) \pi(\boldsymbol{\theta} | (\mu_j, \tau_j), x)$$

and write it in the form

$$\sum_{j=1}^k \lambda_j(x) \varphi_{\tau_j(x)}(\boldsymbol{\theta} - \mu_j(x))$$

for appropriate $\lambda_j(x)$, $\tau_j(x)$ and $\mu_j(x)$. This shows that (1.7.4) defines a conjugate prior for the $\mathcal{N}(\theta, \sigma_0^2)$ distribution.

(b) Let $X_i = \theta + \epsilon_i$, $1 \leq i \leq n$, where θ is as previously and $\epsilon_1, \dots, \epsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma_0^2)$. Find the posterior $\pi(\theta | x_1, \dots, x_n)$, and show that it belongs to class (1.7.4).

Hint: Consider the sufficient statistic for $p(\mathbf{x} | \theta)$.

32. A Hierarchical Binomial–Beta Model. Let $\{(r_j, s_j) : 1 \leq j \leq k\}$ be a given collection of pairs with $r_j > 0$, $s_j > 0$, let (R, S) be a random pair with $P(R = r_j, S = s_j) = \lambda_j$, $0 < \lambda_j < 1$, $\sum_{j=1}^k \lambda_j = 1$, and let θ be a random variable whose conditional density $\pi(\theta, r, s)$ given $R = r$, $S = s$ is beta, $\beta(r, s)$. Consider the model in which $(X | \theta)$ has the binomial, $\mathcal{B}(n, \theta)$, distribution. Note that θ has the prior density

$$\pi(\theta) = \sum_{j=1}^k \lambda_j \pi(\theta, r_j, s_j). \quad (1.7.5)$$

Find the posterior

$$\pi(\theta | x) = \sum_{j=1}^k P(R = r_j, S = s_j | x) \pi(\theta | (r_j, s_j), x)$$

and show that it can be written in the form $\sum \lambda_j(x) \pi(\theta, r_j(\mathbf{x}), s_j(\mathbf{x}))$ for appropriate $\lambda_j(x)$, $r_j(x)$ and $s_j(x)$. This shows that (1.7.5) defines a class of conjugate priors for the $\mathcal{B}(n, \theta)$ distribution.

33. Let $p(x, \eta)$ be a one parameter canonical exponential family generated by $T(x) = x$ and $h(x)$, $x \in \mathcal{X} \subset R$, and let $\psi(x)$ be a nonconstant, nondecreasing function. Show that $E_\eta \psi(X)$ is strictly increasing in η .

Hint:

$$\begin{aligned} \frac{\partial}{\partial \eta} E_\eta \psi(X) &= \text{Cov}_\eta(\psi(X), X) \\ &= \frac{1}{2} E\{(X - X')[\psi(X) - \psi(X')]\} \end{aligned}$$

where X and X' are independent identically distributed as X (see A.11.12).

34. Let (X_1, \dots, X_n) be a stationary Markov chain with two states 0 and 1. That is,

$$P[X_i = \epsilon_i | X_1 = \epsilon_1, \dots, X_{i-1} = \epsilon_{i-1}] = P[X_i = \epsilon_i | X_{i-1} = \epsilon_{i-1}] = p_{\epsilon_{i-1} \epsilon_i}$$

where $\begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$ is the matrix of transition probabilities. Suppose further that

(i) $p_{00} = p_{11} = p$, so that, $p_{10} = p_{01} = 1 - p$.

(ii) $P[X_1 = 0] = P[X_1 = 1] = \frac{1}{2}$.

(a) Show that if $0 < p < 1$ is unknown this is a full rank, one-parameter exponential family with $T = N_{00} + N_{11}$ where $N_{ij} \equiv$ the number of transitions from i to j . For example, 01011 has $N_{01} = 2$, $N_{11} = 1$, $N_{00} = 0$, $N_{10} = 1$.

(b) Show that $E(T) = (n - 1)p$ (by the method of indicators or otherwise).

35. A Conjugate Prior for the Two-Sample Problem. Suppose that X_1, \dots, X_n and Y_1, \dots, Y_n are independent $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$ samples, respectively. Consider the prior π for which for some $r > 0$, $k > 0$, $r\sigma^{-2}$ has a χ_k^2 distribution and given σ^2 , μ_1 and μ_2 are independent with $\mathcal{N}(\xi_1, \sigma^2/k_1)$ and $\mathcal{N}(\xi_2, \sigma^2/k_2)$ distributions, respectively, where $\xi_j \in R$, $k_j > 0$, $j = 1, 2$. Show that π is a conjugate prior.

36. The inverse Gaussian density, $IG(\mu, \lambda)$, is

$$f(x, \mu, \lambda) = [\lambda/2\pi]^{1/2} x^{-3/2} \exp\{-\lambda(x - \mu)^2/2\mu^2 x\}, \quad x > 0, \quad \mu > 0, \quad \lambda > 0.$$

(a) Show that this is an exponential family generated by $\mathbf{T}(X) = -\frac{1}{2}(X, X^{-1})^T$ and $h(x) = (2\pi)^{-1/2} x^{-3/2}$.

(b) Show that the canonical parameters η_1, η_2 are given by $\eta_1 = \mu^{-2}\lambda$, $\eta_2 = \lambda$, and that $A(\eta_1, \eta_2) = -[\frac{1}{2}\log(\eta_2) + \sqrt{\eta_1\eta_2}]$, $\mathcal{E} = [0, \infty) \times (0, \infty)$.

(c) Find the moment-generating function of \mathbf{T} and show that $E(X) = \mu$, $\text{Var}(X) = \mu^{-3}\lambda$, $E(X^{-1}) = \mu^{-1} + \lambda^{-1}$, $\text{Var}(X^{-1}) = (\lambda\mu)^{-1} + 2\lambda^{-2}$.

(d) Suppose $\mu = \mu_0$ is known. Show that the gamma family, $\Gamma(\alpha, \beta)$, is a conjugate prior.

(e) Suppose that $\lambda = \lambda_0$ is known. Show that the conjugate prior formula (1.6.20) produces a function that is not integrable with respect to μ . That is, Ω defined in (1.6.19) is empty.

(f) Suppose that μ and λ are both unknown. Show that (1.6.20) produces a function that is not integrable; that is, Ω defined in (1.6.19) is empty.

37. Let X_1, \dots, X_n be i.i.d. as $X \sim \mathcal{N}_p(\boldsymbol{\theta}, \Sigma_0)$ where Σ_0 is known. Show that the conjugate prior generated by (1.6.20) is the $\mathcal{N}_p(\boldsymbol{\eta}_0, \tau_0^2 \mathbf{I})$ family, where $\boldsymbol{\eta}_0$ varies freely in R^p , $\tau_0^2 > 0$ and \mathbf{I} is the $p \times p$ identity matrix.

38. Let $X_i = (Z_i, Y_i)^T$ be i.i.d. as $X = (Z, Y)^T$, $1 \leq i \leq n$, where X has the density of Example 1.6.3. Write the density of X_1, \dots, X_n as a canonical exponential family and identify T , h , A , and \mathcal{E} . Find the expected value and variance of the sufficient statistic.

39. Suppose that Y_1, \dots, Y_n are independent, $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, $n \geq 4$.

(a) Write the distribution of Y_1, \dots, Y_n in canonical exponential family form. Identify \mathbf{T} , h , $\boldsymbol{\eta}$, A , and \mathcal{E} .

(b) Next suppose that μ_i depends on the value z_i of some covariate and consider the submodel defined by the map $\boldsymbol{\eta} : (\theta_1, \theta_2, \theta_3)^T \rightarrow (\boldsymbol{\mu}^T, \sigma^2)^T$ where $\boldsymbol{\eta}$ is determined by

$$\mu_i = \exp\{\theta_1 + \theta_2 z_i\}, \quad z_1 < z_2 < \dots < z_n; \quad \sigma^2 = \theta_3$$

where $\theta_1 \in R$, $\theta_2 \in R$, $\theta_3 > 0$. This model is sometimes used when μ_i is restricted to be positive. Show that $p(\mathbf{y}, \boldsymbol{\theta})$ as given by (1.6.12) is a curved exponential family model with $l = 3$.

40. Suppose Y_1, \dots, Y_n are independent exponentially, $\mathcal{E}(\lambda_i)$, distributed survival times, $n \geq 3$.

(a) Write the distribution of Y_1, \dots, Y_n in canonical exponential family form. Identify \mathbf{T} , h , $\boldsymbol{\eta}$, A , and \mathcal{E} .

(b) Recall that $\mu_i = E(Y_i) = \lambda_i^{-1}$. Suppose μ_i depends on the value z_i of a covariate. Because $\mu_i > 0$, μ_i is sometimes modeled as

$$\mu_i = \exp\{\theta_1 + \theta_2 z_i\}, \quad i = 1, \dots, n$$

where not all the z 's are equal. Show that $p(\mathbf{y}, \boldsymbol{\theta})$ as given by (1.6.12) is a curved exponential family model with $l = 2$.

1.8 NOTES

Note for Section 1.1

(1) For the measure theoretically minded we can assume more generally that the P_θ are all dominated by a σ finite measure μ and that $p(x, \theta)$ denotes $\frac{dP_\theta}{d\mu}$, the Radon Nikodym derivative.

Notes for Section 1.3

(1) More natural in the sense of measuring the Euclidean distance between the estimate $\hat{\theta}$ and the “truth” θ . Squared error gives much more weight to those $\hat{\theta}$ that are far away from θ than those close to θ .

(2) We define the lower boundary of a convex set simply to be the set of all boundary points r such that the set lies completely on or above any tangent to the set at r .

Note for Section 1.4

(1) Source: Hodges, Jr., J. L., D. Kretch, and R. S. Crutchfield. (1975)

Notes for Section 1.6

(1) Exponential families arose much earlier in the work of Boltzmann in statistical mechanics as laws for the distribution of the states of systems of particles—see Feynman (1963), for instance. The connection is through the concept of entropy, which also plays a key role in information theory—see Cover and Thomas (1991).

(2) The restriction that's $x \in R^q$ and that these families be discrete or continuous is artificial. In general if μ is a σ finite measure on the sample space \mathcal{X} , $p(x, \theta)$ as given by (1.6.1) can be taken to be the density of X with respect to μ —see Lehmann (1997), for instance.

This permits consideration of data such as images, positions, and spheres (e.g., the Earth), and so on.

Note for Section 1.7

(1) $\mathbf{u}^T M \mathbf{u} > 0$ for all $p \times 1$ vectors $\mathbf{u} \neq 0$.

1.9 REFERENCES

- BERGER, J. O., *Statistical Decision Theory and Bayesian Analysis* New York: Springer, 1985.
- BERMAN, S. M., "A Stochastic Model for the Distribution of HIV Latency Time Based on T4 Counts," *Biometrika*, 77, 733–741 (1990).
- BICKEL, P. J., "Using Residuals Robustly I: Tests for Heteroscedasticity, Nonlinearity," *Ann. Statist.* 6, 266–291 (1978).
- BLACKWELL, D. AND M. A. GIRSHICK, *Theory of Games and Statistical Decisions* New York: Wiley, 1954.
- BOX, G. E. P., "Sampling and Bayes Inference in Scientific Modelling and Robustness (with Discussion)," *J. Royal Statist. Soc. A* 143, 383–430 (1979).
- BROWN, L., *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, IMS Lecture Notes—Monograph Series, Hayward, 1986.
- CARROLL, R. J. AND D. RUPPERT, *Transformation and Weighting in Regression* New York: Chapman and Hall, 1988.
- COVER, T. M. AND J. A. THOMAS, *Elements of Information Theory* New York: Wiley, 1991.
- DE GROOT, M. H., *Optimal Statistical Decisions* New York: McGraw–Hill, 1969.
- DOKSUM, K. A. AND A. SAMAROV, "Nonparametric Estimation of Global Functionals and a Measure of the Explanatory Power of Covariates in Regression," *Ann. Statist.* 23, 1443–1473 (1995).
- FERGUSON, T. S., *Mathematical Statistics* New York: Academic Press, 1967.
- FEYNMAN, R. P., *The Feynman Lectures on Physics*, v. 1, R. P. Feynman, R. B. Leighton, and M. Sands, Eds., Ch. 40 *Statistical Mechanics of Physics* Reading, MA: Addison-Wesley, 1963.
- GRENNANDER, U. AND M. ROSENBLATT, *Statistical Analysis of Stationary Time Series* New York: Wiley, 1957.
- HODGES, JR., J. L., D. KRETCH AND R. S. CRUTCHFIELD, *Statlab: An Empirical Introduction to Statistics* New York: McGraw–Hill, 1975.
- KENDALL, M. G. AND A. STUART, *The Advanced Theory of Statistics*, Vols. II, III New York: Hafner Publishing Co., 1961, 1966.
- LEHMANN, E. L., "A Theory of Some Multiple Decision Problems, I and II," *Ann. Math. Statist.* 22, 1–25, 547–572 (1957).
- LEHMANN, E. L., "Model Specification: The Views of Fisher and Neyman, and Later Developments," *Statist. Science* 5, 160–168 (1990).
- LEHMANN, E. L., *Testing Statistical Hypotheses*, 2nd ed. New York: Springer, 1997.

- LINDLEY, D. V., *Introduction to Probability and Statistics from a Bayesian Point of View*, Part I: *Probability*; Part II: *Inference* London: Cambridge University Press, 1965.
- MANDEL, J., *The Statistical Analysis of Experimental Data* New York: J. Wiley & Sons, 1964.
- NORMAND, S-L. AND K. A. DOKSUM, "Empirical Bayes Procedures for a Change Point Problem with Application to HIV/AIDS Data," *Empirical Bayes and Likelihood Inference*, 67–79, Editors: S. E. Ahmed and N. Reid. New York: Springer, Lecture Notes in Statistics, 2001.
- PEARSON, K., "On the General Theory of Skew Correlation and Nonlinear Regression," *Proc. Roy. Soc. London* 71, 303 (1905). (Draper's Research Memoirs, Dulan & Co, Biometrics Series II.)
- RAIFFA, H. AND R. SCHLAIFFER, *Applied Statistical Decision Theory*, Division of Research, Graduate School of Business Administration, Harvard University, Boston, 1961.
- SAVAGE, L. J., *The Foundations of Statistics*, J. Wiley & Sons, New York, 1954.
- SAVAGE, L. J. ET AL., *The Foundation of Statistical Inference* London: Methuen & Co., 1962.
- SNEDECOR, G. W. AND W. G. COCHRAN, *Statistical Methods*, 8th Ed. Ames, IA: Iowa State University Press, 1989.
- WETHERILL, G. B. AND K. D. GLAZEBROOK, *Sequential Methods in Statistics* New York: Chapman and Hall, 1986.

This page intentionally left blank

References

1 1. STATISTICAL MODELS, GOALS, AND PERFORMANCE CRITERIA

✎ ✎ ✎ ✎

✎ ✎ ✎ ✎ LINDLEY, D. V., Introduction to Probability and Statistics from a Bayesian Point of View, Part I: Probability; Part II: Inference London: Cambridge University Press, 1965. MANDEL, J., The Statistical Analysis of Experimental Data New York: J. Wiley & Sons, 1964. NORMAND, S-L. AND K. A. DOKSUM, "Empirical Bayes Procedures for a Change Point Problem with Application to HIV/AIDS Data," Empirical Bayes and Likelihood Inference, 67-79, Editors: S. E. Ahmed and N. Reid. New York: Springer, Lecture Notes in Statistics, 2001. PEARSON, K., "On the General Theory of Skew Correlation and Nonlinear Regression," Proc. Roy. Soc. London 71, 303 (1905). (Draper's Research Memoirs, Dulan & Co, Biometrics Series II.) RAIFFA, H. AND R. SCHLAIFFER, Applied Statistical Decision Theory, Division of Research, Graduate School of Business Administration, Harvard University, Boston, 1961. SAVAGE, L. J., The Foundations of Statistics, J. Wiley & Sons, New York, 1954. SAVAGE, L. J. ET AL., The Foundation of Statistical Inference London: Methuen & Co., 1962. SNEDECOR, G. W. AND W.G. COCHRAN, Statistical Methods, 8th Ed. Ames, IA: Iowa State University Press, 1989. WETHERILL, G. B. AND K. D. GLAZEBROOK, Sequential Methods in Statistics New York: Chapman and Hall, 1986. This page intentionally left blank

2 2. METHODS OF ESTIMATION

⌘ ⌘ ⌘ ⌘

⌘ ⌘ ⌘ ⌘ RUPPERT, D., AND M. P. WAND, "Multivariate Locally Weighted Least Squares Regression," Ann. Statist., 22, 1346-1370 (1994). SEBER, G. A. F., AND C.J. WILD, Nonlinear Regression New York: Wiley, 1989. SHANNON, C. E., "A Mathematical Theory of Communication," Bell System Tech. Journal, 27, 379- 243, 623-656 (1948). SNEDECOR, G. W., AND W. COCHRAN, Statistical Methods, 6th ed. Ames, IA: Iowa State University Press, 1967. STIGLER, S., The History of Statistics Cambridge, MA: Harvard University Press, 1986. WEISBERG, S., Applied Linear Regression, 2nd ed. New York: Wiley, 1985. WU, C. F. J., "On the Convergence Properties of the EMAlgorithm," Ann. Statist., 11, 95-103 (1983).


3 3. MEASURES OF PERFORMANCES

⌘ ⌘ ⌘ ⌘

⌘ ⌘ ⌘ ⌘ HAMPEL, F., "The Influence Curve and Its Role in Robust Estimation," J. Amer. Statist. Assoc., 69, 383-393 (1974). HAMPEL, F., E. RONCHETTI, P. ROUSSEUW, AND W. STAHEL, Robust Statistics: The Approach Based on Influence Functions New York: J. Wiley & Sons, 1986. HANSEN, M. H., AND B. YU, "Model Selection and the Principle of Minimum Description Length," J. Amer. Statist. Assoc., 96, 746-774 (2001). HOGG, R., "Adaptive Robust Procedures," J. Amer. Statist. Assoc., 69, 909-927 (1974). HUBER, P., Robust Statistics New York: Wiley, 1981. HUBER, P., "Robust Statistics: A Review," Ann. Math. Statist., 43, 1041-1067 (1972). JAECKEL, L. A., "Robust Estimates of Location," Ann. Math. Statist., 42, 1020-1034 (1971). JEFFREYS, H., Theory of Probability, 2nd ed. London: Oxford University Press, 1948. KARLIN, S., Mathematical Methods and Theory in Games, Programming, and Economics Reading, MA: Addison-Wesley, 1959. LEHMANN, E. L., Testing Statistical Hypotheses New York: Springer, 1986. LEHMANN, E. L., AND G. CASELLA, Theory of Point Estimation, 2nd ed. New York: Springer, 1998. LINDLEY, D. V., Introduction to Probability and Statistics from a Bayesian Point of View, Part I: Probability; Part II: Inference, Cambridge University Press, London, 1965. LINDLEY, D.V., "Decision Analysis and Bioequivalence Trials," Statistical Science, 13, 136-141 (1998). NORBERG, R., "Hierarchical Credibility: Analysis of a Random Effect Linear Model with Nested Classification," Scand. Actuarial J., 204-222 (1986). RISSANEN, J., "Stochastic Complexity (With Discussions)," J. Royal Statist. Soc. B, 49, 223-239 (1987). SAVAGE, L. J., The Foundations of Statistics New York: J. Wiley & Sons, 1954. SHIBATA, R., "Bootstrap Estimate of Kullback-Leibler Information for Model Selection," Statistica Sinica, 7, 375-394 (1997). STEIN, C., "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Distribution," Proc. Third Berkeley Symposium on Math. Statist. and Probability, 1, University of California Press, 197-206 (1956). TUKEY, J. W., Exploratory Data Analysis Reading, MA: Addison-Wesley, 1972. WALLACE, C. S., AND P. R. FREEMAN, "Estimation and Inference by Compact Coding (With Discussions)," J. Royal Statist. Soc. B, 49, 240-251 (1987). WIJSMAN, R. A., "On the Attainment of the Cramér-Rao Lower Bound," Ann. Math. Statist., 1, 538-542 (1973).

4 4. TESTING AND CONFIDENCE REGIONS: BASIC THEORY



 BICKEL, P., E. HAMMEL, AND J. W. O'CONNELL, "Is there a sex bias in graduate admissions?" *Science*, 187, 398-404 (1975). BOX, G. E. P., *Apology for Ecumenism in Statistics and Scientific Inference, Data Analysis and Robustness*, G. E. P. Box, T. Leonard, and C. F. Wu, Editors New York: Academic Press, 1983. BROWN, L. D., T. CAI, AND A. DASGUPTA, "Interval estimation for a binomial proportion," *Statistical Science*, 101-128 (2001). DOKSUM, K. A. AND G. SIEVERS, "Plotting with confidence: Graphical comparisons of two populations," *Biometrika*, 63, 421-434 (1976). DOKSUM, K. A., G. FENSTAD, AND R. AABERGE, "Plots and tests for symmetry," *Biometrika*, 64, 473-487 (1977). DURBIN, J., "Distribution theory for tests based on the sample distribution function," *Regional Conference Series in Applied Math.*, 9, SIAM, Philadelphia, Pennsylvania (1973). FERGUSON, T., *Mathematical Statistics. A Decision Theoretic Approach* New York: Academic Press, 1967. FISHER, R. A., *Statistical Methods for Research Workers*, 13th ed. New York: Hafner Publishing Company, 1958. HALD, A., *Statistical Theory with Engineering Applications* New York: J. Wiley & Sons, 1952. HEDGES, L. V. AND I. OLKIN, *Statistical Methods for Meta-Analysis* Orlando, FL: Academic Press, 1985. JEFFREYS, H., *The Theory of Probability* Oxford: Oxford University Press, 1961. LEHMANN, E. L., *Testing Statistical Hypotheses*, 2nd ed. New York: Springer, 1997. POPPER, K. R., *Conjectures and Refutations; the Growth of Scientific Knowledge*, 3rd ed. New York: Harper and Row, 1968. PRATT, J., "Length of confidence intervals," *J. Amer. Statist. Assoc.*, 56, 549-567 (1961). SACKROWITZ, H. AND E. SAMUEL-CAHN, "P values as random variables—Expected P values," *The American Statistician*, 53, 326-331 (1999). STEPHENS, M., "EDF statistics for goodness of fit," *J. Amer. Statist.*, 69, 730-737 (1974). STEIN, C., "A two-sample test for a linear hypothesis whose power is independent of the variance," *Ann. Math. Statist.*, 16, 243-258 (1945). TATE, R. F. AND G. W. KLETT, "Optimal confidence intervals for the variance of a normal distribution," *J. Amer. Statist. Assoc.*, 54, 674-682 (1959). VAN ZWET, W. R. AND J. OSTERHOFF, "On the combination of independent test statistics," *Ann. Math. Statist.*, 38, 659-680 (1967). WALD, A., *Sequential Analysis* New York: Wiley, 1947. WALD, A., *Statistical Decision Functions* New York: Wiley, 1950. WANG, Y., "Probabilities of the type I errors of the Welch tests," *J. Amer. Statist. Assoc.*, 66, 605-608 (1971). WELCH, B., "Further notes on

Mrs. Aspin's tables," *Biometrika*, 36, 243-246 (1949).
WETHERILL, G. B. AND K. D. GLAZEBROOK, *Sequential Methods*
in Statistics New York: Chapman and Hall, 1986.

5 5. ASYMPTOTIC APPROXIMATIONS

BERGER, J., Statistical Decision Theory and Bayesian Analysis New York: Springer-Verlag, 1985.

BHATTACHARYA, R. H. AND R. RANGA RAO, Normal Approximation and Asymptotic Expansions New York: Wiley, 1976.

BILLINGSLEY, P., Probability and Measure New York: Wiley, 1979.

BOX, G. E. P., "Non-normality and tests on variances," Biometrika, 40, 318-324 (1953).

CRAMER, H., Mathematical Methods of Statistics Princeton, NJ: Princeton University Press, 1946.

DAVID, F. N., Tables of the Correlation Coefficient, Cambridge University Press, reprinted in Biometrika Tables for Statisticians (1966), Vol. I, 3rd ed., H. O. Hartley and E. S. Pearson, Editors Cambridge: Cambridge University Press, 1938.

FERGUSON, T. S., A Course in Large Sample Theory New York: Chapman and Hall, 1996.

FISHER, R. A., "Theory of statistical estimation," Proc. Camb. Phil. Soc., 22, 700-725 (1925).

FISHER, R. A., Statistical Inference and Scientific Method, Vth Berkeley Symposium, 1958.

HAMMERSLEY, J. M. AND D. C. HANSCOMB, Monte Carlo Methods London: Methuen & Co., 1964.

HOEFFDING, W., "Probability inequalities for sums of bounded random variables," J. Amer. Statist. Assoc., 58, 13-80 (1963).

HUBER, P. J., The Behavior of the Maximum Likelihood Estimator Under Non-Standard Conditions, Proc. Vth Berk. Symp. Math. Statist. Prob., Vol. 1 Berkeley, CA: University of California Press, 1967.

LE CAM, L. AND G. L. YANG, Asymptotics in Statistics, Some Basic Concepts New York: Springer, 1990.

LEHMANN, E. L., Elements of Large-Sample Theory New York: Springer-Verlag, 1999.

LEHMANN, E. L. AND G. CASELLA, Theory of Point Estimation
New York: Springer-Verlag, 1998.

NEYMAN, J. AND E. L. SCOTT, "Consistent estimates based on
partially consistent observations," *Econometrica*, 16, 1-32
(1948).

RAO, C. R., Linear Statistical Inference and Its
Applications, 2nd ed. New York: J. Wiley & Sons, 1973.

RUDIN, W., Mathematical Analysis, 3rd ed. New York: McGraw
Hill, 1987.

SCHERVISCH, M., Theory of Statistics New York: Springer,
1995.

SERFLING, R. J., Approximation Theorems of Mathematical
Statistics New York: J. Wiley & Sons, 1980.

STIGLER, S., The History of Statistics: The Measurement of
Uncertainty Before 1900 Cambridge, MA: Harvard University
Press, 1986.

WILSON, E. B. AND M. M. HILFERTY, "The distribution of chi
square," *Proc. Nat. Acad. Sci., U.S.A.*, 17, p. 684 (1931).
This page intentionally left blank

6 6. INFERENCE IN THE MULTIPARAMETER CASE

◊ ◊ ◊ ◊

◊ ◊ ◊ ◊ DIXON, W. AND F. MASSEY, Introduction to Statistical Analysis, 3rd ed. New York: McGraw-Hill, 1969. FISHER, R. A., Statistical Methods for Research Workers, 13th ed. New York: Hafner, 1958. GRAYBILL, F., An Introduction to Linear Statistical Models, Vol. I New York: McGraw-Hill, 1961. HABERMAN, S., The Analysis of Frequency Data Chicago: University of Chicago Press, 1974. HALD, A., Statistical Theory with Engineering Applications New York: Wiley, 1952. HUBER, P. J., "The behavior of the maximum likelihood estimator under nonstandard conditions," Proc. Fifth Berkeley Symp. Math. Statist. Prob. 1, Univ. of California Press, 221-233 (1967). KASS, R., J. KADANE AND L. TIERNEY, "Approximate marginal densities of nonlinear functions," Biometrika, 76, 425-433 (1989). KOENKER, R. AND V. D'OREY, "Computing regression quantiles," J. Roy. Statist. Soc. Ser. C, 36, 383-393 (1987). LAPLACE, P.-S., "Sur quelques points du systéme du monde," Memoires de l'Académie des Sciences de Paris (Reprinted in Oeuvres Complètes, 11, 475-558. Gauthier-Villars, Paris) (1789). MALLOWS, C., "Some comments on C p," Technometrics, 15, 661-675 (1973). MCCULLAGH, P. AND J. A. NELDER, Generalized Linear Models London: Chapman and Hall, New York, 1983; second edition, 1989. PORTNOY, S. AND R. KOENKER, "The Gaussian Hare and the Laplacian Tortoise: Computability of squared-error versus absolute-error estimators," Statistical Science, 12, 279-300 (1997). RAO, C. R., Linear Statistical Inference and Its Applications, 2nd ed. New York: J. Wiley & Sons, 1973. ROBERTSON, T., F. T. WRIGHT, AND R. L. DYKSTRA, Order Restricted Statistical Inference New York: Wiley, 1988. SCHEFFÉ, H., The Analysis of Variance New York: Wiley, 1959. SCHERVISCH, M., Theory of Statistics New York: Springer, 1995. STIGLER, S., The History of Statistics: The Measurement of Uncertainty Before 1900 Cambridge, MA: Harvard University Press, 1986. WEISBERG, S., Applied Linear Regression, 2nd ed. New York: Wiley, 1985. This page intentionally left blank

A. A REVIEW OF BASIC PROBABILITY THEORY

⌘ ⌘ ⌘ ⌘

⌘ ⌘ ⌘ ⌘

⌘ ⌘ ⌘ ⌘

⌘ ⌘ ⌘ ⌘

⌘ ⌘ ⌘ ⌘ A.8 RANDOM VARIABLES AND VECTORS: TRANSFORMATIONS

Although sample spaces can be very diverse, the statistician is usually interested primarily in one or more numerical characteristics of the sample point that has occurred. For example, we measure the weight of pigs drawn at random from a population, the time to breakdown and length of repair time for a randomly chosen machine, the yield per acre of a field of wheat in a given year, the concentration of a certain pollutant in the atmosphere, and so on. In the probability model, these quantities will correspond to random variables and vectors. A.8.1A random variable X is a function from Ω to R such that the set $\{\omega : X(\omega) \in B\} = X^{-1}(B)$ is in A for every $B \in B$. (1) A.8.2A random vector $X = (X_1, \dots, X_k)^T$ is k -tuple of random variables, or equivalently a function from Ω to R^k such that the set $\{\omega : X(\omega) \in B\} = X^{-1}(B)$ is in A for every $B \in B^k$. (1) For $k = 1$ random vectors are just random variables. The event $X^{-1}(B)$ will usually be written $[X \in B]$ and $P([X \in B])$ will be written $P[X \in B]$. The probability distribution of a random vector X is, by definition, the probability measure P_X in the model (R^k, B^k, P_X) given by $P_X(B) = P[X \in B]$. (A.8.3) A.8.4A random vector is said to have a continuous or discrete distribution (or to be continuous or discrete) according to whether its probability distribution is continuous or discrete. Similarly, we will refer to the frequency function, density, d.f., and so on of a random vector when we are, in fact, referring to those features of its probability distribution. The subscript X or X will be used for densities, d.f.'s, and so on to indicate which vector or variable they correspond to unless the reference is clear from the context in which case they will be omitted. The probability of any event that is expressible purely in terms of X can be calculated if we know only the probability distribution of X . In the discrete case this means we need only know the frequency function and in the continuous case the density. Thus, from (A.7.5) and (A.7.8) $P[X \in A] = \sum_{x \in A} p(x)$, if X is discrete $= \int_A p(x)dx$, if X is continuous. (A.8.5) When we are interested in particular random variables or vectors, we will describe them purely in terms

of their probability distributions without any further specification of the underlying sample space on which they are defined. The study of realor vector-valued functions of a random vector X is central in the theory of probability and of statistics. Here is the formal definition of such transformations. Let g be any function from \mathbb{R}^k to \mathbb{R}^m , $k, m \geq 1$, such that (2) $g^{-1}(B) = \{y \in \mathbb{R}^k : g(y) \in B\} \in \mathcal{B}^k$ for every $B \in \mathcal{B}^m$. Then the random transformation $g(X)$ is defined by $g(X)(\omega) = g(X(\omega))$.

(A.8.6) An example of a transformation often used in statistics is $g = (g_1, g_2)'$ with $g_1(X) = k^{-1} \sum_{i=1}^k X_i = \bar{X}$ and $g_2(X) = k^{-1} \sum_{i=1}^k (X_i - \bar{X})^2$. Another common example is $g(X) = (\min\{X_i\}, \max\{X_i\})'$. The probability distribution of $g(X)$ is completely determined by that of X through $P[g(X) \in B] = P[X \in g^{-1}(B)]$.

(A.8.7) If X is discrete with frequency function p_X , then $g(X)$ is discrete and has frequency function $p_{g(X)}(t) = \sum \{x: g(x)=t\} p_X(x)$.

(A.8.8) Suppose that X is continuous with density p_X and g is real-valued and one-to-one (3) on an open set S such that $P[X \in S] = 1$. Furthermore, assume that the derivative g' of g exists and does not vanish on S . Then $g(X)$ is continuous with density given by $p_{g(X)}(t) = p_X(g^{-1}(t)) |g'(g^{-1}(t))|$ (A.8.9) for $t \in g(S)$, and 0 otherwise. This is called the change of variable formula. If $g(X) = \sigma X + \mu$, $\sigma \neq 0$, and X is continuous, then $p_{g(X)}(t) = \frac{1}{|\sigma|} p_X\left(\frac{t-\mu}{\sigma}\right)$.

(A.8.10) From (A.8.8) it follows that if $(X, Y)^T$ is a discrete random vector with frequency function $p_{(X,Y)}$, then the frequency function of X , known as the marginal frequency function, is given by (4) $p_X(x) = \sum_y p_{(X,Y)}(x, y)$.

(A.8.11) Similarly, if $(X, Y)^T$ is continuous with density $p_{(X,Y)}$, it may be shown (as a consequence of (A.8.7) and (A.7.8)) that X has a marginal density function given by $p_X(x) = \int_{-\infty}^{\infty} p_{(X,Y)}(x, y) dy$. (5) (A.8.12) These notions generalize to the case $Z = (X, Y)$, a random vector obtained by putting two random vectors together. The (marginal) frequency or density of X is found as in (A.8.11) and (A.8.12) by summing or integrating out over y in $p_{(X,Y)}(x, y)$. Discrete random variables may be used to approximate continuous ones arbitrarily closely and vice versa.

◊ ◊ ◊ ◊

◊ ◊ ◊ ◊

◊ ◊ ◊ ◊

◊ ◊ ◊ ◊ A.11.3 The distribution of a random variable is typically uniquely specified by its moments. This is the

case, for example, if the random variable possesses a moment generating function (cf. (A.12.1)).

A.11.4 The k th central moment of X is by definition $E[(X - E(X))^k]$, the k th moment of $(X - E(X))$, and is denoted by μ_k .

A.11.5 The second central moment is called the variance of X and will be written $\text{Var } X$. The nonnegative square root of $\text{Var } X$ is called the standard deviation of X . The standard deviation measures the spread of the distribution of X about its expectation. It is also called a measure of scale. Another measure of the same type is $E(|X - E(X)|)$, which is often referred to as the mean deviation. The variance of X is finite if and only if the second moment of X is finite (cf. (A.11.15)). If a and b are constants, then by (A.10.7) $\text{Var}(aX + b) = a^2 \text{Var } X$.

(A.11.6) (One side of the equation exists if and only if the other does.)

A.11.7 If X is any random variable with well-defined (finite) mean and variance, the standardized version or Z-score of X is the random variable $Z = (X - E(X)) / \sqrt{\text{Var } X}$. By (A.10.7) and (A.11.6) it follows then that $E(Z) = 0$ and $\text{Var } Z = 1$.

(A.11.8) A.11.9 If $E(X^2) = 0$, then $X = 0$. If $\text{Var } X = 0$, $X = E(X)$ (a constant). These results follow, for instance, from (A.15.2).

A.11.10 The third and fourth central moments are used in the coefficient of skewness γ_1 and the kurtosis γ_2 , which are defined by $\gamma_1 = \mu_3 / \sigma^3$, $\gamma_2 = (\mu_4 / \sigma^4) - 3$ where $\sigma^2 = \text{Var } X$. See also Section A.12 where γ_1 and γ_2 are expressed in terms of cumulants. These descriptive measures are useful in comparing the shapes of various frequently used densities.

A.11.11 If $Y = a + bX$ with $b > 0$, then the coefficient of skewness and the kurtosis of Y are the same as those of X . If $X \sim N(\mu, \sigma^2)$, then $\gamma_1 = \gamma_2 = 0$.

A.11.12 It is possible to generalize the notion of moments to random vectors. For simplicity we consider the case $k = 2$. If X_1 and X_2 are random variables and i, j are natural numbers, then the product moment of order (i, j) of X_1 and X_2 is, by definition, $E(X_1^i X_2^j)$. The central product moment of order (i, j) of X_1 and X_2 is again by definition $E[(X_1 - E(X_1))^i (X_2 - E(X_2))^j]$. The central product moment of order $(1, 1)$ is $E[(X_1 - E(X_1))(X_2 - E(X_2))]$. Review called the covariance of X_1 and X_2 and is written $\text{Cov}(X_1, X_2)$. By expanding the product $(X_1 - E(X_1))(X_2 - E(X_2))$ and using (A.10.3) and (A.10.7), we obtain the relations, $\text{Cov}(aX_1 + bX_2, cX_3 + dX_4) = ac \text{Cov}(X_1, X_3) + bc \text{Cov}(X_2, X_3) + ad \text{Cov}(X_1, X_4) + bd \text{Cov}(X_2, X_4)$ (A.11.13) and $\text{Cov}(X_1, X_2) = E(X_1 X_2) - E(X_1)E(X_2)$.

(A.11.14) If X'_1 and X'_2 are distributed as X_1 and X_2 and are independent of X_1 and X_2 , then $\text{Cov}(X_1, X_2) = \frac{1}{2} E(X_1 - X'_1)(X_2 - X'_2)$. If we put $X_1 = X_2 = X$ in (A.11.14), we get the formula $\text{Var } X = E(X^2) - [E(X)]^2$.

(A.11.15) The covariance is defined whenever X_1 and X_2 have finite variances and in

that case $|\text{Cov}(X_1, X_2)| \leq \sqrt{(\text{Var} X_1)(\text{Var} X_2)}$ (A.11.16) with equality holding if and only if (1) X_1 or X_2 is a constant or (2) $(X_1 - E(X_1)) = \text{Cov}(X_1, X_2) / \text{Var} X_2 (X_2 - E(X_2))$. This is the correlation inequality. It may be obtained from the Cauchy-Schwarz inequality, $|E(Z_1 Z_2)| \leq \sqrt{E(Z_1^2)E(Z_2^2)}$ (A.11.17) for any two random variables Z_1, Z_2 such that $E(Z_1^2) < \infty, E(Z_2^2) < \infty$. Equality holds if and only if one of Z_1, Z_2 equals 0 or $Z_1 = aZ_2$ for some constant a . The correlation inequality corresponds to the special case $Z_1 = X_1 - E(X_1), Z_2 = X_2 - E(X_2)$. A proof of the Cauchy-Schwarz inequality is given in Remark 1.4.1. The correlation of X_1 and X_2 , denoted by $\text{Corr}(X_1, X_2)$, is defined whenever X_1 and X_2 are not constant and the variances of X_1 and X_2 are finite by $\text{Corr}(X_1, X_2) = \text{Cov}(X_1, X_2) / \sqrt{(\text{Var} X_1)(\text{Var} X_2)}$. (A.11.18) The correlation of X_1 and X_2 is the covariance of the standardized versions of X_1 and X_2 . The correlation inequality is equivalent to the statement $|\text{Corr}(X_1, X_2)| \leq 1$. (A.11.19) Equality holds if and only if X_2 is linear function ($X_2 = a + bX_1, b \neq 0$) of X_1 .

✎ ✎ ✎ ✎

✎ ✎ ✎ ✎


✎ ✎ ✎ ✎ distributions, which arise frequently in probability and statistics, and list some of their properties. Following the name of each distribution we give a shorthand notation that will sometimes be used as will obvious abbreviations such as “binomial (n, θ)” for “the binomial distribution with parameter (n, θ)”. The symbol p as usual stands for a frequency or density function. If anywhere below p is not specified explicitly for some value of x it shall be assumed that p vanishes at that point. Similarly, if the value of the distribution function F is not specified outside some set, it is assumed to be zero to the “left” of the set and one to the “right” of the set. I. Discrete Distributions The binomial distribution with parameters n and θ : $B(n, \theta)$. $p(k) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$, $k = 0, 1, \dots, n$. (A.13.1) The parameter n can be any integer ≥ 0 whereas θ may be any number in $[0, 1]$. A.13.2 If X is the total number of successes obtained in n Bernoulli trials with probability of success θ , then X has a $B(n, \theta)$ distribution (see (A.6.3)). If X has a $B(n, \theta)$ distribution, then $E(X) = n\theta$, $\text{Var} X = n\theta(1 - \theta)$. (A.13.3) Higher-order moments may be computed from the moment generating function $M_X(t) = [\theta e^t + (1-\theta)]^n$. (A.13.4) A.13.5 If X_1, X_2, \dots, X_k are independent random variables distributed as $B(n_1, \theta), B(n_2, \theta), \dots, B(n_k, \theta)$, respectively, then $X_1 + X_2 + \dots + X_k$ has a $B(n_1$

$\dots + n k, \theta$ distribution. This result may be derived by using (A.12.5) and (A.12.6) in conjunction with (A.13.4). The hypergeometric distribution with parameters D, N , and $n : H(D, N, n)$. $p(k) = \binom{D}{k} \binom{N-D}{n-k} / \binom{N}{n}$ (A.13.6) for k a natural number with $\max(0, n - (N - D)) \leq k \leq \min(n, D)$. The parameters D and n may be any natural numbers that are less than or equal to the natural number N . A.13.7 If X is the number of defectives (special objects) in a sample of size n taken without replacement from a population with D defectives and $N - D$ nondefectives, then X has an $H(D, N, n)$ distribution (see (A.6.10)). If the sample is taken with replacement, X has a $B(n, D/N)$ distribution. Review If X has an $H(D, N, n)$ distribution, then $E(X) = n D / N$, $\text{Var} X = n D / N (1 - D / N) N - n N - 1$. (A.13.8) Formulae (A.13.8) may be obtained directly from the definition (A.13.6). An easier way is to use the interpretation (A.13.7) by writing $X = \sum_{j=1}^n I_j$ where $I_j = 1$ if the j th object sampled is defective and 0 otherwise, and then applying formulae (A.10.4), (A.10.7), and (A.11.20). The Poisson distribution with parameter $\lambda : P(\lambda)$. $p(k) = e^{-\lambda} \lambda^k / k!$ (A.13.9) for $k = 0, 1, 2, \dots$. The parameter λ can be any positive number. If X has a $P(\lambda)$ distribution, then $E(X) = \text{Var} X = \lambda$. (A.13.10) The moment generating function of X is given by $M_X(t) = e^{\lambda(e^t - 1)}$. (A.13.11) A.13.12 If X_1, X_2, \dots, X_n are independent random variables with $P(\lambda_1), P(\lambda_2), \dots, P(\lambda_n)$ distributions, respectively, then $X_1 + X_2 + \dots + X_n$ has the $P(\lambda_1 + \lambda_2 + \dots + \lambda_n)$ distribution. This result may be derived in the same manner as the corresponding fact for the binomial distribution. The multinomial distribution with parameters $n, \theta_1, \dots, \theta_q : M(n, \theta_1, \dots, \theta_q)$. $p(k_1, \dots, k_q) = n! / (k_1! \dots k_q!) \theta_1^{k_1} \dots \theta_q^{k_q}$ (A.13.13) whenever k_i are nonnegative integers such that $\sum_{i=1}^q k_i = n$. The parameter n is any natural number while $(\theta_1, \dots, \theta_q)$ is any vector in $\theta = \{(\theta_1, \dots, \theta_q) : \theta_i \geq 0, 1 \leq i \leq q, \sum_{i=1}^q \theta_i = 1\}$. A.13.14 If $X = (X_1, \dots, X_q)'$, where X_i is the number of times outcome ω_i occurs in n multinomial trials with probabilities $(\theta_1, \dots, \theta_q)$, then X has a $M(n, \theta_1, \dots, \theta_q)$ distribution (see (A.6.6)). If X has a $M(n, \theta_1, \dots, \theta_q)$ distribution, $E(X_i) = n\theta_i$, $\text{Var} X_i = n\theta_i(1 - \theta_i)$, $\text{Cov}(X_i, X_j) = -n\theta_i\theta_j$, $i \neq j$, $i, j = 1, \dots, q$. (A.13.15)

Review

These results may either be derived directly or by a representation such as that discussed in (A.13.8) and an application of formulas (A.10.4), (A.10.7), (A.13.13), and

(A.11.20). A.13.16 If X has a $M(n, \theta_1, \dots, \theta_q)$ distribution, then $(X_{i_1}, \dots, X_{i_s}, n - \sum_{j=1}^s X_{i_j})'$ has a $M(n, \theta_{i_1}, \dots, \theta_{i_s}, 1 - \sum_{j=1}^s \theta_{i_j})$ distribution for any set $\{i_1, \dots, i_s\} \subset \{1, \dots, q\}$. Therefore, X_j has $B(n, \theta_j)$ distributions for each j and more generally $\sum_{j=1}^s X_{i_j}$ has a $B(n, \sum_{j=1}^s \theta_{i_j})$ distribution if $s < q$. These remarks follow from the interpretation (A.13.14). II. Continuous Distributions

Before beginning our listing we introduce some convenient notations: $X \sim F$ will mean that X is a random variable with d.f. F , and $X \sim p$ will similarly mean that X has density or frequency function p . Let Y be a random variable with d.f. F . Let F_μ be the d.f. of $Y + \mu$. The family $F_L = \{F_\mu : -\infty < \mu < \infty\}$ is called a location parameter family, μ is called a location parameter, and we say that Y generates F_L . By definition, for any μ , $X \sim F_\mu \Leftrightarrow X - \mu \sim F$. Therefore, for any μ, γ , $F_\mu(x) = F(x - \mu) = F_\theta(x - \mu) = F_\gamma(x + (\gamma - \mu))$ and all calculations involving F_μ can be referred back to F or any other member of the family. Similarly, if Y generates F_L so does $Y + \gamma$ for any fixed γ . If Y has a first moment, it follows that we may without loss of generality (as far as generating F_L goes) assume that $E(Y) = 0$. Then if $X \sim F_\mu$, $E(X) = \mu$. Similarly let $F * \sigma$ be the d.f. of σY , $\sigma > 0$. The family $F_S = \{F * \sigma : \sigma > 0\}$ is called a scale parameter family, σ is a scale parameter, and Y is said to generate F_S . By definition, for any $\sigma > 0$, $X \sim F * \sigma \Leftrightarrow X/\sigma \sim F$. Again all calculations involving one member of the family can be referred back to any other because for any $\sigma, \tau > 0$, $F * \sigma(x) = F * \tau(\tau x / \sigma)$. If Y generates F_S and Y has a first moment different from 0, we may without loss of generality take $E(Y) = 1$ and, hence, if $X \sim F * \sigma$, then $E(X) = \sigma$. Alternatively, if Y has a second moment, we may select F as being the unique member of the family F_S having $\text{Var } Y = 1$ and then $X \sim F * \sigma \Leftrightarrow \text{Var } X = \sigma^2$. Finally, define $F_{\mu, \sigma}$ as the d.f. of $\sigma Y + \mu$. The family $F_{L,S} = \{F_{\mu, \sigma} : -\infty < \mu < \infty, \sigma > 0\}$ is called a location-scale parameter family, μ is called a location parameter, and σ a scale parameter, and Y is said to generate $F_{L,S}$. From $F_{\mu, \sigma}(x) = F((x - \mu)/\sigma) = F_{\gamma, \tau}(\tau(x - \mu)/\sigma + \gamma)$, we see as before how to refer calculations involving one member of the family back to any other. Without loss of generality, if Y has a second moment, we may take $E(Y) = 0$, $\text{Var } Y = 1$. 

Review Then if $X \sim F_{\mu, \sigma}$, we obtain $E(X) = \mu$, $\text{Var } X = \sigma^2$. Clearly $F_\mu = F_{\mu, 1}$, $F * \sigma = F_{0, \sigma}$. The relation between the density of $F_{\mu, \sigma}$ and that of F is given by (A.8.10). All the families of densities we now define are location-scale or scale families. The normal (Gaussian) distribution with parameters μ and $\sigma^2 : N(\mu, \sigma^2)$. $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x - \mu)^2\}$. (A.13.17) The









parameter μ can be any real number while σ is positive. The normal distribution with $\mu = 0$ and $\sigma = 1$ is known as the standard normal distribution. Its density will be denoted by $\phi(z)$ and its d.f. by $\Phi(z)$. A.13.18 The family of $N(\mu, \sigma^2)$ distributions is a location-scale family. If Z has a $N(0, 1)$ distribution, then $\sigma Z + \mu$ has a $N(\mu, \sigma^2)$ distribution, and conversely if X has a $N(\mu, \sigma^2)$ distribution, then $(X - \mu)/\sigma$ has a standard normal distribution. If X has a $N(\mu, \sigma^2)$ distribution, then $E(X) = \mu$, $\text{Var} X = \sigma^2$. (A.13.19) More generally, all moments may be obtained from $M_X(t) = \exp\{\mu t + \frac{1}{2}\sigma^2 t^2\}$ (A.13.20) for $-\infty < t < \infty$. In particular if $\mu = 0$, $\sigma^2 = 1$, then $M_X(t) = \sum_{k=0}^{\infty} \frac{(2k)!}{2^k k!} \frac{t^{2k}}{(2k)!}$ (A.13.21) and, hence, in this case we can conclude from (A.12.4) that $E(X^k) = 0$ if $k \geq 0$ is odd $E(X^k) = k! 2^{k/2} (k/2)!$ if $k \geq 0$ is even. (A.13.22) A.13.23 If X_1, \dots, X_n are independent normal random variables such that $E(X_i) = \mu_i$, $\text{Var} X_i = \sigma_i^2$, and c_1, \dots, c_n are any constants that are not all 0, then $\sum_{i=1}^n c_i X_i$ has a $N(c_1 \mu_1 + \dots + c_n \mu_n, c_1^2 \sigma_1^2 + \dots + c_n^2 \sigma_n^2)$ distribution. This follows from (A.13.20), (A.12.5), and (A.12.6). Further information about the normal distribution may be found in Section A.15 and Appendix B. The exponential distribution with parameter λ : $E(\lambda)$.

• • • •

• • • •

• • • •

• • • • be interpreted as the relative frequency of occurrence of A in n independent repetitions of the experiment in which A is an event and the Bernoulli law is now evidently a statement of the type we wanted. Bernoulli's proof of this result was rather complicated and it remained for the Russian mathematician Chebychev to give a two-line argument. His generalization of Bernoulli's result is based on an inequality that has proved to be of the greatest importance in probability and statistics. Chebychev's Inequality If X is any random variable, then $P[|X| \geq a] \leq E(X^2)/a^2$. (A.15.2) The Bernoulli law follows readily from (A.15.2) and (A.13.3) via the calculation $p[|S_n/n - p| \geq q] \leq E(S_n/n - p)^2/q^2 = \text{Var } S_n/n^2/q^2 = p(1-p)/nq^2 \rightarrow 0$ as $n \rightarrow \infty$. (A.15.3) A generalization of (A.15.2), which contains various important and useful inequalities, is the following. Let g be a nonnegative function on R such that g is nondecreasing on the range of a random variable Z . Then $P[Z \geq a] \leq E(g(Z))/g(a)$. (A.15.4) If we put $Z = |X|$,

$g(t) = t^2$ if $t \geq 0$ and 0 otherwise, we get (A.15.2). Other important cases are obtained by taking $Z = |X|$ and $g(t) = t$ if $t \geq 0$ and 0 otherwise (Markov's inequality), and $Z = X$ and $g(t) = e^{st}$ for $s > 0$ and all real t (Bernstein's inequality, see B.9.5 for the binomial case Bernstein's inequality). Proof of (A.15.4). Note that by the properties of g , $g(a)1[Z \geq a] \leq g(Z)1[Z \geq a] \leq g(Z)$. (A.15.5) Therefore, by (A.10.8) $g(a)P[Z \geq a] = E(g(a)1[Z \geq a]) \leq E(g(Z))$, (A.15.6) which is equivalent to (A.15.4). * The following result, which follows from Chebychev's inequality, is a useful generalization of Bernoulli's law. Khintchin's (Weak) Law of Large Numbers Let $\{X_i\}$, $i \geq 1$, be a sequence of independent identically distributed random variables with finite mean μ and define $S_n = \sum_{i=1}^n X_i$. Then $S_n/n \xrightarrow{P} \mu$. (A.15.7)         Review Upon taking the X_i to be indicators of binomial trials, we obtain (A.15.1). De Moivre-Laplace Theorem Suppose that $\{S_n\}$ is a sequence of random variables such that for each n , S_n has a $B(n, p)$ distribution where $0 < p < 1$. Then $S_n - np \sqrt{np(1-p)} \xrightarrow{L} Z$, (A.15.8) where Z has a standard normal distribution. That is, the standardized versions of S_n converge in law to a standard normal random variable. If we write $S_n - np \sqrt{np(1-p)} = \sqrt{n} \sqrt{p(1-p)} (S_n/n - p)$ and use (A.14.9), it is easy to see that (A.15.8) implies (A.15.1). The De Moivre-Laplace theorem is generalized by the following. Central Limit Theorem Let $\{X_i\}$ be a sequence of independent identically distributed random variables with (common) expectation μ and variance σ^2 such that $0 < \sigma^2 < \infty$. Then, if $S_n = \sum_{i=1}^n X_i$ $S_n - n\mu \sigma \sqrt{n} \xrightarrow{L} Z$, (A.15.9) where Z has the standard normal distribution. The last two results are most commonly used in statistics as approximation theorems. Let k and l be nonnegative integers. The De Moivre-Laplace theorem is used as $P[k \leq S_n \leq l] = P[k - 1/2 \leq S_n \leq l + 1/2] = P[k - np - 1/2 \sqrt{npq} \leq S_n - np \sqrt{npq} \leq l - np + 1/2 \sqrt{npq}] \approx \Phi(l - np + 1/2 \sqrt{npq}) - \Phi(k - np - 1/2 \sqrt{npq})$ (A.15.10) where $q = 1-p$. The $1/2$ appearing in $k - 1/2$ and $l + 1/2$ is called the continuity correction. We have an excellent idea of how good this approximation is. An illustrative discussion is given in Feller (1968, pp. 187-188). A rule of thumb is that for most purposes the approximation can be used when np and $n(1-p)$ are both larger than 5. Only when the X_i are integer-valued is the first step of (A.15.10) followed. Otherwise (A.15.9) is applied in the form $P[a \leq S_n \leq b] \approx \Phi(b - n\mu \sqrt{n\sigma}) - \Phi(a - n\mu \sqrt{n\sigma})$. (A.15.11)



REFERENCES

- AÏT-SAHALIA, Y., BICKEL, P.J. and STOKER, T.M., Goodness-of-fit tests for kernel regression with an application to option implied volatilities, *Journal of Econometrics* 105, 363–412, 2001.
- AALEN, O., Nonparametric inference for a family of counting processes, *Ann. Statist.* 6, 701–726, 1978.
- ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D.L. and JOHNSTONE, I.M., Adapting to unknown sparsity by controlling the false discovery rate, *Ann. Statist.* 34, 584–653, 2006.
- AKAIKE, H., Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* 21, 243–247, 1969.
- AMINI A.A. and WAINWRIGHT M.J., High-dimensional analysis of semidefinite relaxations for sparse principal components, *Ann. Statist.* 37, 2877–2921, 2009.
- AMIT, Y. and GEMAN, D., Shape quantization and recognition with randomized trees. *Neural computation* 9, 1545–1588, 1997.
- ANDERSEN, P.K., BORGAN, O., GILL, R.D. and KEIDING, N., Censoring, truncation, and filtering in statistical models based on counting processes, *Contemporary Mathematics* 80, 19–60, Providence, American Mathematical Society, 1988.
- ANDERSEN, P.K., BORGAN, O., GILL, R.D. and KEIDING, N., *Statistical Models Based on Counting Processes*. New York, Springer, 1993.
- ANDERSEN, P.K. and GILL, R.D., Cox’s regression model for counting processes: A large sample study. *Ann. Statist.* 10, 1100–1120, 1982.
- ANDERSON, T.W., *Introduction to Multivariate Statistical Analysis*, 3rd ed. J. Wiley, 2003.
- ANDREWS, D.F., BICKEL, P.J., HAMPEL, F.R., HUBER, P.J., ROGERS, W.H. and TUKEY, J. W., *Robust estimates of Location: Survey and Advances*. Princeton, NJ, Princeton University Press, 1972.
- APOSTOL, T.M., *Mathematical Analysis*, 1st ed., Addison Wesley, 1957.
- ARLOT, S. and CELISSE, A., A survey of cross validation procedures for model selection, *Statistics Surveys* 4, 40–79, 2010.

- ASSOUAD, P., Deux remarques sur l'estimation, *L.R. Acad. Sci. Paris, ser I*, 296, 1021–1024, 1983.
- BABU, G.J. and FEIGELSON, E.D., *Astrostatistics*, London, Chapman and Hall, 1996.
- BARANIUK, R.G., CEVHER, V., and WAKIN, M.B., Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective, *Proceedings of the IEEE* 98, 959–971, 2010.
- BARLOW, R.E. and PROSCHAN, F., Inequalities for linear combinations of order statistics from restricted families, *Ann. Math. Statist.* 37, 1574–1591, 1966.
- BARRON, A., BIRGÉ, L. and MASSART, P., Risk bounds for model selection via penalization, *Probab. Theory Related Fields* 113(3), 301–413, 1999.
- BELL, C.B. and DOKSUM, K.A., “Optimal” one-sample distribution-free tests and their two-sample extensions, *Ann. Math. Statist.* 37, 120–132, 1966.
- BELL, C.B., BLACKWELL, D. and BREIMAN, L., On the completeness of order statistics, *Ann. Math. Statist.* 31, 794–796, 1960.
- BENJAMINI, Y. and HOCHBERG, Y., Controlling the false discovery rate: A practical and powerful approach to multiple testing, *J. Royal Statist. Soc. Ser. B* 57, 289–300, 1995.
- BERGER, J.O., *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., New York, Springer, 1985.
- BERNARDO, J.M. and SMITH, A.F.M., *Bayesian Theory*, New York, Wiley, 1994.
- BICKEL, P.J., Some contributions to the theory of order statistics, *Proceedings of Vth Berkeley Symposium on Probability and Statistics*, 575–592, 1967.
- BICKEL, P.J., Tests for monotone failure rate II, *Ann. Math. Statist.* 40, 1250–1260, 1970.
- BICKEL, P. J., Minimax estimation of the mean of a normal distribution when the parameter space is restricted”, *Ann. Statist.* 9, 1301–1309, 1981.
- BICKEL, P.J. and DOKSUM, K.A., Tests for monotone failure rate based on normalized spacings, *Ann. Math. Statist.* 40, 1216–1235, 1969.
- BICKEL, P. J. and DOKSUM, K. A., *Mathematical Statistics. Basic Ideas and Selected Topics*, 1st ed., Oakland, CA, Holden–Day, 1977.
- BICKEL, P.J. and FREEDMAN, D.A. Some asymptotic theory for the bootstrap, *Ann. Statist.* 9, 1196–1217, 1981.
- BICKEL, P.J., GÖTZE, F. and VAN ZWET, W.R., Resampling fewer than n observations: Gains, losses, and remedies for losses, *Statistica Sinica* 7, 1–31, 1997.
- BICKEL, P.J., KLAASSEN, C.A.J., RITOV, Y. and WELLNER, J.A., *Efficient and Adaptive Estimation for Semiparametric Models*, John Hopkins Univ. Press, Baltimore, 1993. Reissued by Springer, New York, 1998.
- BKRW, 1993, 1998). Short for previous entry.
- BICKEL, P.J. and KRIEGER, Confidence bands for a distribution function using the bootstrap, *J. Amer. Statist. Assoc.* 84, no.405, 95–100, 1989.

- BICKEL, P. J. and LEVINA, E., Regularized estimation of large covariance matrices, *Ann. Statist.* 36, 199–227, 2008a.
- BICKEL, P. J. and LEVINA, E., Covariance regularization by thresholding, *Ann. Statist.* 36(6), 2577–2604, 2008b.
- BICKEL, P. J. and LI, B., Regularization in statistics, Discussants: A.B. Tsybakov, S.A. van de Geer, B. Yu, T. Valdés, C. Rivero, J. Fan, and A. van der Vaart, *Test*, 271–344, 2006.
- BICKEL, P. J. and RITOV, Y., Estimating integrated squared density derivatives: Sharp best order of convergence estimates, *Sankhya*, A381–393, 1988.
- BICKEL, P.J. and RITOV, Y., Local asymptotic normality of ranks and covariates in transformation models, *Festschrift for Lucien Le Cam*, D. Pollard and G.L. Yang, eds., New York, Springer, 1997.
- BICKEL, P.J., RITOV, Y. and STOKER, T., Tailor-made tests for goodness of fit to semiparametric hypotheses, *Ann. Statist.* 34, 721–741, 2006.
- BICKEL, P. J., RITOV, Y. and ZAKAI, A., Some theory for generalized boosting algorithms, *J. Mach. Learn. Res.* 7, 705–732, 2006.
- BICKEL, P. J. and ROSENBLATT, M., On some global measures of the deviation of density function estimates, *Ann. Statist.* 1, 1071–1095, 1973.
- BICKEL, P. J. and SAKOV, A., On the choice of m in the m out of n bootstrap and confidence bounds for extrema, *Statistica Sinica* 18, 967–985, 2008.
- BILLINGSLEY, P., *Convergence of Probability Measures*, New York, Wiley, 1968.
- BIRKHOFF, G. and MACLANE, S., *A Survey of Modern Algebra*, AKP Classics, 1998.
- BIRNBAUM, A., JOHNSTONE, I.M., NADLER, B. and PAUL, D., Minimax bounds for sparse PCA with noisy high-dimensional data, *Ann. Statist.* 41, 1055–1084, 2013.
- BISHOP, C., *Pattern Recognition and Machine Learning*, New York, Springer, 2006.
- BJERVE, S., DOKSUM, K.A. and YANDELL, B.S., Uniform confidence bounds for regression based on a simple moving average, *Scand. J. Statist.* 12, 159–169, 1995.
- BLACKWELL, D. and GIRSHICK, M.A. *Theory of Games and Statistical Decisions*. Wiley, New York, 1954. Reissued by Dover, New York, 1979.
- BLYTH, C.R., On minimax statistical procedures and their admissibility, *Ann. Math. Statist.* 22, 22–42, 1951.
- BOX, G.E.P and COX, D.R., An analysis of transformations, *J. Royal Statist. Soc. Ser. B* 26, 211–252, 1964.
- BOX G.E.P., HUNTER, W.G. and HUNTER, J.S. *Statistics for Experimenters*, New York, Wiley, 1978.
- BOYD, S. and VANDENBERGHE, I., *Convex Optimization*, Cambridge University Press, Cambridge, 2004.
- BREIMAN, L., *Probability*, Reading, MA, Addison-Wesley 1968. Reprinted in SIAM Classics in Applied Mathematics, 2000.

- BREIMAN, L., Random forests, *Machine learning* 45, 5–32, 2001.
- BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. and STONE, J., *Classification and Regression Trees*, Belmont, Wadsworth, 1984.
- BRILLINGER, D., *Time Series. Data Analysis and Theory*. Reprint of the 1981 edition. *Classics in Applied Mathematics* 36, Philadelphia, PA, *Society for Industrial and Applied Mathematics (SIAM)*, 2001.
- BROOKS, S., GELMAN, A., JONES, G. and MENG, X.L., eds., *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.
- BROWN, L., CARTER, A., LOW, M., and ZHANG, C., Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift, *Ann. Statist.* 32, 2399–2430, 2004.
- BROWN, L. and LOW, M., Asymptotic equivalence of nonparametric regression and white noise, *Ann. Statist.* 24, 2384–2398, 1996.
- BUHLMANN, P. and KÜNSCH, H.R., The blockwise bootstrap for general parameters of a stationary time series, *Scand. J. Statist.* 22, 35–54, 1995.
- BUHLMANN, P. and VAN DE GEER, S., *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Heidelberg, Springer, 2011.
- CASELLA, G. and STRAWDERMAN, W. E., Estimating a bounded normal mean, *Ann. Statist.* 9, 870–878, 1981.
- CHANDRASEKARAN, V. and JORDAN, M.I., Computational and statistical tradeoffs via convex relaxation. *PNAS* 110, 1181–1190, 2013.
- CHEN, H., Convergence rates for parametric components in a partly linear model, *Ann. Statist.* 16, 136–146, 1988.
- CHEN, A. and BICKEL, P.J., *Efficient Independent Component Analysis*, Technical Report, UC Berkeley, 2003.
- CHERNOFF, H., GASTWIRTH, J.L. and JOHNS, M.V. Asymptotic distribution of linear combinations of functions of order statistics with applications to estimation. *Ann. Math. Statist.* 38, 52–72, 1967.
- CHUNG, K.L., The strong law of large numbers, *Proc. 2nd Berkeley Symp. Math. Statist. Prob.*, 341–352, Berkeley, Univ. California Press, 1951.
- CLEVELAND, W., Robust locally weighted regression and smoothing scatterplots, *J. Amer. Statist. Assoc.* 74, 829–836, 1979.
- CLEVELAND, W. and DEVLIN, S., Locally weighted regression: An approach to regression analysis by local fitting, *J. Amer. Statist. Assoc.* 83, 596–610, 1988.
- COMON, P., Independent component analysis, a new concept?, *Signal Processing* 36, 287–314, 1994.
- COVER, T.M. and HART, P.E., Nearest neighbor pattern classification, *IEEE Transactions on Information Theory* 13, 21–27, 1967.
- COVER, T.M. and THOMAS, J.A., *Elements of Information Theory*, New York, Wiley, 1991.

- COX, D.R., Regression models and life-tables, *J. Royal Statist. Soc. Ser. B* 34, No. 2, 187–220, 1972.
- COX, D.R., Partial likelihood, *Biometrika* 62, 269–276, 1975.
- COX, D.R. and OAKES, D., *Analysis of Survival Data*, New York, Chapman and Hall, 1984.
- DARLING, D.A. The Cramér–Smirnov test in the parametric case, *Ann. Inst. Statist. Math.* 26, 1–20, 1955.
- DAUBECHIES, T., *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.
- DE BOOR, C., *A Practical Guide to Splines*. New York, Springer, 1978.
- DE HAAN, L. and FERREIRA, A., *Extreme Value Theory. An Introduction*, New York, Springer, 2006.
- DE MONTRICHER, G.F., TAPIA, R.A. and THOMPSON, J.R., Nonparametric maximum likelihood estimation of probability densities by penalty function methods, *Ann. Statist.* 3, 1329–1348, 1975.
- DEVROYE, L., GYÖRFI, L. and LUGOSI, G., *A Probabilistic Theory of Pattern Recognition, Applications of Mathematics 31*, New York, Springer-Verlag, 1996.
- DIACONIS, P., The Markov chain Monte Carlo revolution. *Bull. American Math. Soc.* 46, 179–205, 2009.
- DIACONIS, P. and FREEDMAN, D. On the consistency of Bayes estimates (with a discussion and a rejoinder by the authors), *Ann. Statist.* 14, 1–67, 1986.
- DIACONIS, P. and STROOCK, D., Geometric bounds for eigenvalues of Markov chains, *The Annals of Applied Probability* 1(1), 36–61, 1991.
- DIACONIS, P. and STRUMFELS, S., Algebraic algorithms for sampling from conditional distributions, *The Annals of Statist.* 26, 363–397, 1998.
- DOKSUM, K.A., An extension of partial likelihood methods for proportional hazard models to general transformation models, *Ann. Statist.* 15, 325–345, 1987.
- DOKSUM, K.A. and OZEKI, A., Semiparametric models and likelihood - the power of ranks, *Optimality. The Third Erich L. Lehmann Symposium*, Javier Rojo, ed. IMS Lecture Notes-Monograph Series, 67–92, 2009.
- DOKSUM, K.A. and NABEYA, S., Estimation in proportional hazard and log-linear models, *J. Statistical Planning and Inference*, 297–303, 1984.
- DOKSUM, K.A. and SAMAROV, A., Nonparametric estimation of global functionals and a measure of explanatory power of covariates in regression, *Ann. Statist.* 23, 1443–1473, 1995.
- DONOHOO, D. and JOHNSTONE, I.M., Minimax risk over ℓ_p -balls for ℓ_q -error, *Probab. Theory Related Fields* 99, 277–303, 1994.
- DONOHOO, D. and JOHNSTONE, I., Adapting to unknown smoothness via wavelet shrinkage, *J. Amer. Statist. Assoc.* 90, 1200–1224, 1995.
- DONOHOO, D. and LU, R.C., Geometrizing Rates of Convergence II, III, *Ann. Statist.* 19, 633–667, 1991.

- DONSKER, M.D., Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.* 23, 277–281, 1952.
- DOOB, J.L., Heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.* 20, 393–403, 1949.
- DOOB, J.L. *Stochastic Processes*, Wiley, New York, 1953.
- DOUKHAN, R. *Mixing: Properties and Examples. Springer Lecture Notes in Statistics*, New York, Springer-Verlag, 1995.
- DUDOIT, S., SHAFFER, J.P., and BOLDRICK, J.C., Multiple hypothesis testing in microarray experiments, *Statist. Sci.* 18(1), 71–103, 2003.
- DURBIN, J. Distribution theory for tests based on the sample distribution function. *Regional Conference Series in Applied Mathematics* 9, SIAM Philadelphia, PA, 1973.
- EFRON, B., Bootstrap methods: Another look at the jackknife, *Ann. Statist.* 7(1), 1–26, 1979.
- EFRON, B., Estimation the error rate of a prediction rule: Improvement on cross-validation, *J. Amer. Statist. Assoc.* 78(382), 316–331, 1983.
- EFRON, B., Microarrays, empirical Bayes and the two-groups model, *Statistical Science* 2, 197–223, 2008.
- EFRON, B., *Large Scale Inference. Empirical Bayes Methods for Estimation, Testing and Prediction*, Cambridge University Press, Cambridge, 2010.
- EFRON, B. and MORRIS, C.N., Stein's estimation rule and its competitors – an empirical Bayes approach, *J. Amer. Statist. Assoc.* 68, 117–130, 1973.
- EFRON, B. and TIBSHIRANI, R.J., *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability 57, Chapman and Hall, New York, 1993.
- EL KAROUI, N., Recent results about the largest eigenvalue of random covariance matrices and statistical application, *Acta Physica Polonica B* 36, 2005.
- EL KAROUI, N., Operator norm consistent estimation of large dimensional sparse covariance matrices, *Ann. Stat.* 36, 2717–2756, 2008.
- ENGLE, R.F., GRANGER, C.W.J., RICE, J. and WEISS, A., Semiparametric estimates of the relation between weather and electricity sales, *J. Amer. Statist. Assoc.* 81, 310–320, 1986.
- FAN, J. and GIJBELS, I., *Local Polynomial Modelling and Its Applications*. London, Chapman and Hall, 1996.
- FAN, J., HÄRDLE, W. and MAMMEN, E., Direct estimation of low dimensional components in additive models, *Ann. Statist.* 26, 943–971, 1998.
- FAN, J. and LI, R., New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis, *J. Amer. Statist. Assoc.* 99, 710–723, 2004.
- FAN, J., HAN, X. and GU, W., Estimating false discovery proportion under arbitrary covariance dependence, *J. Amer. Statist. Assoc.* 107, 1019–1035, 2012.

- FANO, R. M., *Class Notes for Transmission of Information*, Course 6.574, MIT, Cambridge MA, 1952.
- FELLER, W., *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd ed., New York, Wiley, 1968.
- FITHIAN W., SUN, D. and TAYLOR, J., Optimal inference after model selection, arXiv preprint arXiv:1410.2597s, 2014.
- FIX, E. and HODGES, J., Discriminatory analysis—nonparametric discrimination: Consistency properties. Technical Report 21-49-004 4, Randolph Field, Texas, U.S. Air Force, School of Aviation Medicine, 1951.
- FREUND, Y. and SCHAPIRE, R., A decision-theoretic generalization of online learning and application to boosting, *Journal of Computer and System Sciences* 55, 119–139, 1997.
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R., Additive logistic regression: A statistical view of boosting (with discussion), *Ann. Statist.* 28, 337–407, 2000.
- GELMAN, A., CARLIN, J.B., STERN, H.S. and RUBIN, D.B., *Bayesian Data Analysis. 2nd Ed.*, Boca Raton, FL, Chapman & Hall/CRC, 2004.
- GELMAN, A., CARLIN, J.B., STERN, H.S., DUNSON, D.B., VEHTARI, A. and RUBIN, D.B., *Bayesian Data Analysis*, Vol. 2, Boca Raton, FL, CRC Press, 2014.
- GEMAN, S. and GEMAN, D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 721–741, 1984.
- GEYER, C.J., *Optimization of Functions, Markov Chain Monte Carlo in Practice.*, p.241, CRC Press, 1995.
- GHOSH, J.K. and RAMAMOORTHY, R.V., *Bayesian nonparametrics*. New York, Springer 2003.
- GILKS, W.R. (ed.), *Markov Chain Monte Carlo in Practice*. CRC Press, 1995.
- GILL, R.D., VARDI, Y. and WELLNER, J.A., Large sample theory of empirical distributions in biased sampling models, *Ann. Statist.* 16, 1069–1112, 1988.
- GINÉ, E. and ZINN, J., Bootstrapping general empirical measures. *Ann. Probab.* 18, 851–869, 1990.
- GOOD, I. J. and R. A. GASKINS, Nonparametric roughness penalties for probability densities, *Biometrika* 58, 255–277, 1971.
- GÖTZE, F., Abstract. *Bulletin of Institute of Mathematical Statistics*, 1993.
- GRENNANDER, U., *Abstract Inference*, Wiley, New York, 1981.
- GRIMMETT, G. and STIRZAKER, D., *Probability and Random Processes*, Oxford Univ. Press, Oxford, 2001.
- GYÖRFI L., KOHLER, M., KRZYZAK, A. and WALK, H., *A Distribution-Free Theory of Nonparametric Regression*, New York, Springer, 2002.
- HÁJEK, J., A characterization of limiting distributions of regular estimates, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 14, 323–330, 1970.

- HÁJEK, J., Local asymptotic minimax and admissibility in estimation. *Proceedings of the Sixth Berkeley Symposium on Math. Statist. and Prob.*, 1, 175–194, 1972.
- HÁJEK, J. and SÍDÁK, Z., *Theory of Rank Tests*, Academic Press, New York, 1967.
- HALL, P. *The Bootstrap and Edgeworth Expansion*. Springer, New York, 1997.
- HALL, P. and HYDE, C.C., *Martingale Theory and its Application*, Academic Press, 1980.
- HALL, P., On the number of bootstrap simulations required to construct a confidence interval, *Ann. Statist.* 14, 1453–1462, 1986.
- HAMMERSLEY, J.M. and HANDSCOMB, D.C., *Monte Carlo Methods*. Chapman and Hall, London, 1965.
- HÄRDLE, W., LIANG, H. and GAO, J., *Partially Linear Models*, Springer, New York, 2000.
- HASTIE, T. and TIBSHIRANI, R., *Generalised Additive Models*, Chapman and Hall, London, 1990.
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J., *The Elements of Statistical Learning*, 2nd ed., New York, Springer, 2001, 2009.
- HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M., *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC Press, 2015.
- HODGES, J.L., Jr. and LEHMANN, E.L., Some applications of the Cramèr Rao inequality, *Proc. Second Berkeley Symp. on Math. Statist. and Prob.*, 1, Univ. of California Press, 1951, 13–22.
- HOEFFDING, W., A class of statistics with asymptotically normal distributions. *Ann. Math. Statist.* 19, 293–325, 1948.
- HOEFFDING, W., Optimum nonparametric tests, *Proc. of 2nd Berkeley Symposium on Mathematical Statistics and Probability*, Univ. California Press, 1951, 83–92.
- HOEFFDING, W., Probability inequalities for sums of bounded random variables.” *J. Amer. Statist. Assoc.* 58, 13–30, 1963.
- HOERL, A.E. and KENNARD, R.W., Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, 12, 55–67, 1970.
- HOLM, S., A simple sequentially rejective multiple test procedure, *Scand. J. Statist.* 6, 65–70, 1979.
- HUBER, P., The behaviour of maximum likelihood estimates under non standard conditions. *Proc. of 5th Berkeley Symposium on Math. Statist. and Prob.*, 221–234, 1967.
- HUBER, P., *Robust Statistics*, New York, Wiley, 1981.
- HYVARINEN, A., KARHUNEN, J. and OJA, E., *Independent Component Analysis*, New York, John Wiley and Sons, 2001.
- IBRAGIMOV, I. and HASMINSKII, R.Z., *Statistical Estimation: Asymptotic Theory*, New York, Springer, 1981.
- IBRAGIMOV, I. and LINNIK, Y., *Independent and Stationary Sequences of Random Variables*, WoltersNordhoff Publishing, Groningen, 1971.

- JAMES, W. and STEIN, C., Estimation with quadratic loss, *Proc. Fourth Berkeley Symposium on Math. Statist. and Prob., 1*, Univ. of California Press, 1961, 311–319.
- JIANG, J. and DOKSUM, K.A., Empirical plug-in curve and surface estimates”, *Mathematical and Statistical Methods in Reliability*, B.H. Lindquist and K.A. Doksum, eds., New Jersey, World Scientific, 2003.
- JOHNSON, R.A. and WICHERN, D. W., *Applied Multivariate Statistical Analysis*, Upper Saddle River, New Jersey, Prentice-Hall, 2003.
- JOHNSTONE, I.M. and LU, A.Y., On consistency and sparsity for principal components analysis in high dimensions, *J. Amer. Statist. Assoc.*, 104, 682–693, 2009.
- JOHNSTONE, I.M. and SILVERMAN, B.W., Empirical Bayes selection of wavelet thresholds, *Ann. Statist.* 33, 1700–1752, 2005.
- JUNG, S. and MARRON, J.S., PCA consistency in high dimension, low sample size context, *Ann. Statist.* 37, 4104–4130, 2009.
- KAGAN, A.M., LINNIK, Y.V. and RAO, C.R., *Characterization Problems of Mathematical Statistics*, New York, Wiley, 1973.
- KALBFLEISCH, J.D. and Prentice, R.L., Marginal likelihoods based on Cox’s regression and life model, *Biometrika* 60, 267–278, 1973.
- KALBFLEISCH, J.D. and Prentice, R.L., *The Statistical Analysis of Failure Time Data*, 2nd ed., Wiley, New York, 2002.
- KALBFLEISCH, J. and SPROTT, D.A., Application of likelihood methods involving a large number of parameters (with discussion), *J. Royal Statist. Soc. Ser. B* 32, 175–208, 1970.
- KERKYACHARIAN, G. and PICKARD, D., Wavelet shrinkage: Asymptopia? (with discussion), *J. Royal Statist. Soc. Ser. B* 57, 201–337, 1995.
- KIEFER, J. and WOLFOWITZ, J. Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters, *Ann. Math. Statist.* 27, 887–906, 1956.
- KLAASSEN, C.A.J. and WELLNER, J. A., Efficient estimation in the bivariate normal copula model: Normal margins are least favourable, *Bernoulli* 3, 55–77, 1997.
- KOSOROK, M.R., *Introduction to Empirical Processes and Semiparametric Inference*, New York, Springer, 2008.
- KÜNSCH, H.R., The jackknife and the bootstrap for general stationary observations, *Ann. Statist.* 17, 1217–1241, 1989.
- LAWLESS, J.E., *Statistical Models and Methods for Lifetime Data*, New York, Wiley, 1982.
- LE CAM, L., On the asymptotic theory of estimation and testing hypotheses, *Proc. Third Berkeley Symp. Math. Statist. Prob. 1*, Univ. California Press, Berkeley, 1956, 129–156.
- LE CAM, L., Locally asymptotically normal families of distributions, *Univ. California Publ. Statist.* 3, 37–98,, 1960.
- LE CAM, L., Sufficiency and approximate sufficiency, *Ann. Math. Statist.* 35, 1419–1455, 1964.

- LE CAM, L., Limits of experiments, *Proc. Sixth Berkeley Symp. Math. Statist. Prob. 1*, Univ. California Press, Berkeley, 1972, 245–261.
- LE CAM, L., *Asymptotic Methods in Statistical Decision Theory*, New York, Springer-Verlag, 1986.
- LE CAM, L. and YANG, G.L., *Asymptotics in Statistics; Some Basic Concepts*, New York, Springer, 1990.
- LEHMANN, E.L., Consistency and unbiasedness of certain nonparametric tests, *Ann. Math. Statist.* 22, No. 2, 165–179, 1951.
- LEHMANN, E.L., The power of rank tests, *Ann. Math. Statist.* 24, 23–43, 1953.
- LEHMANN, E.L., Ordered families of distributions, *Ann. Math. Statist.* 26, 399–419, 1955.
- LEHMANN, E.L., *Nonparametrics. Statistical Methods Based on Ranks*, New York, Springer, 2006.
- LEHMANN, E.L. and CASELLA, G., *Theory of Point Estimation*, 2nd ed., New York, Springer, 1998.
- LEHMANN, E.L. and ROMANO, J., *Testing Statistical Hypotheses*, 3rd ed., New York, Springer, 2005.
- LEHMANN, E.L. and SCHEFFÉ, H., Completeness, similar regions, and unbiased estimation, *Part 1, Sankhyā*, 10, 305–340, *Part 2, Sankhyā* 15, 219–236, 1950, 1955.
- LEHMANN, E.L., ROMANO, J.P. and SHAFFER, J.P., On optimality of stepdown and stepup multiple test procedures, *Ann. Statist.* 33(3), 1084–1108, 2005.
- LEVIT, B.Y., Infinite dimensional informational lower bounds, *Theory Prob. Applic.* 23, 388–394, 1978.
- LIN, D.Y. and ZENG, D., Correcting for population stratification in genomewide association studies, *J. Amer. Statist. Assoc.* 106, 997–1008, 2011.
- LITTLE, R.J.A. and RUBIN, D.B., *Statistical Analysis with Missing Data*, New York, Wiley, 1986. Republished by New York, Wiley, 2014.
- LIU, J.S., *Monte Carlo Strategies in Scientific Computing*. New York, Springer, 2001.
- LOADER, C., *Local Regression and Likelihood*, Springer, New York, 1999.
- MALLOWS, C.L., A note on asymptotic joint normality, *Ann. Math. Statist.*, 43, 508–515, 1972.
- MALLOWS, C.L., Some comments on C_p , *Technometrics* 15, 661–675, 1973.
- MAMMEN, E. and TSYBAKOV, A.B., Asymptotic minimax recovery of sets with smooth boundaries, *Ann. Statist.* 23, 502–524, 1995.
- MAMMEN, E. and van de GEER, S.A., Penalized quasi-likelihood estimation in partial linear models, *Ann. Statist.* 25, 1014–1035, 1997.
- MARCINKIEWICZ, J. and ZYGMUND, A., Quelques inégalités pour les opérations lineaires, *Fundamenta Mathematica* 32, 113–121, 1939.
- MARDIA, K.V., KENT, J.T. and BIBBY, J.M., *Multivariate Analysis*, Academic Press, 1979.

- MARRON, J.S., Optimal rates of convergence to Bayes risk in nonparametric discrimination, *Ann. Statist.*, 11, 1142–1155, 1983.
- MECKLIN, C.J. and MUNDFROM, D.J., An appraisal and bibliography of tests for multivariate normality, *International Stat. Review* 72, 123–138, 2004.
- MEIR, R and RÄTSCH, G., An introduction to boosting and leveraging, In *Advanced Lectures on Machine Learning*, S. Mendelson and A. Smola, eds., Lecture Notes in Computer Science, Springer, 2003, 119–184.
- METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E., Equations of state calculations by fast computing machines, *J. Chemical Physics* 21, 1087–1091, 1953.
- MEYN, S.P. and TWEEDIE, R.L., *Markov Chains and Stochastic Stability*. New York, Springer, 1993.
- MORGAN, J.N. and SONQUIST, J.A., Problems in the analysis of survey data, and a proposal, *J. Amer. Statist. Assoc.* 58, 415, 1963.
- MURPHY, S.A., Asymptotic theory for the frailty model, *Ann. Statist.* 23, No. 1, 182–198, 1995.
- MURPHY, S.A., Consistency in a proportional hazards model incorporating a random effect, *Ann. Statist.* 22, 712–731, 1994.
- MURPHY, S.A. and van der VAART, A.W., On profile likelihood (with discussion), *J. Amer. Statist. Assoc.* 95, 449–485, 2000.
- MURPHY, S.A., ROSSINI, T.J. and van der VAART, A.W., MLE in the proportional odds model, *J. Amer. Statist. Assoc.* 92, 968–976, 1997.
- NACHBIN, L., *The Haar Integral*, Von Nostrand, New York, 1965.
- NADARAYA, E.A., *Nonparametric Estimation of Probability Densities and Regression Curves*, Boston, Kluwer Academic Publishers, 1989.
- NUSSBAUM, M., Asymptotic equivalence of density estimation and Gaussian white noise, *Annals of Statistics*, 24, 2399–2430, 1996.
- OSTLAND, M. *A Monte Carlo algorithm applied to travel time estimation and vehicle matching*, UC Berkeley Thesis, 1999.
- OWEN, A., Empirical likelihood ratio confidence intervals for a single functional, *Biometrika* 75, 237–249, 1988.
- OWEN, A., *Empirical Likelihood*, Chapman and Hall, 2001.
- PARNER, E., Asymptotic theory for the correlated gamma-frailty model, *Ann. Statist.* 26, 183–214, 1998.
- PASULA, H., RUSSELL, S., OSTLAND, M. and RITOV, Y., Tracking many objects with many sensors, Proc. IJCAI-99, 1999.
- PATTERSON, N., PRICE, A.L. and REICH, D., Population structure and eigenanalysis, *PLOS Genetics* 12, 2074–2093, 2006.

- PAUL, D., Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, *Statistica Sinica* 17, 1617–1642, 2007.
- PEARL, J., *Causality, Models, Reasoning and Inference*, 2nd ed., Cambridge University Press, Cambridge, 2009.
- PETTY, K., BICKEL, P., JIANG, J., OSTLAND, M., RICE, J., RITOV, Y. and SCHOENBERG, F., Accurate estimation of travel times from single-loop detectors. *Transportation Research Part A: Policy and Practice*, 32(1), 1–17, 1998.
- PETTY, K., OSTLAND, M., KWON, J., RICE, J. and BICKEL, P., A new methodology for evaluating incident detection algorithms, *Transportation Research Part C: Emerging Technologies* 10, 189–204, 2002.
- PINSKER, M.S., Optimal filtering of square integrable signals in Gaussian white noise, *Problems of Information Transmission*, 16, 120–133, 1980.
- POLITIS, D.N. and ROMANO, J.P., Large sample confidence regions based on subsamples under minimal assumptions (in resampling), *Ann. Statist.* 22, No. 4, 2031–2050, 1994.
- POLITIS, D.N., ROMANO, J.P. and WOLF, M., *Subsampling*. New York, Springer, 1999.
- POLLARD, D. *Convergence of Stochastic Processes*, New York, Springer-Verlag, 1984.
- POLLARD, D., Empirical Processes: Theory and Applications, NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 2, Hayward, CA., IMS, 1990.
- PRICE, A., PATTERSON, N., PLENGE, R., WEINBLATT, M., SHADICK, N. and REICH, D., Principal components analysis corrects for stratification in genome-wide association studies, *Nature Genetics* 38, 904–909, 2006.
- PROSCHAN, F. and PYKE, R., Tests for monotone failure rate, *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* 3, Univ. of California Press, 1967.
- QUENOUILLE, M.H., Approximate Tests of Correlation in Time-Series, *J. Royal Statist. Soc. Ser. B* 11, 68–84, 1949.
- QUINLAN, J.R., Simplifying decision trees, *International Journal of Man-Machine Studies* 27(3), 221–234, 1987.
- QUINLAN, J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann Series in Machine Learning, San Francisco, Morgan Kaufmann Publishers, 1993.
- RAO, C.R., *Linear Statistical Inference and Its Applications*, 2nd ed., Wiley, New York, 1973.
- RAO, C.R. and SHINOZAKI, N., Precision of individual estimators in simultaneous estimation of parameters, *Biometrika* 65, 23–30, 1978.
- RAO, P., *Nonparametric Functional Estimation*, Orlando, Academic Press, 1983.
- REID, N., A conversation with Sir David Cox, *Statistical Science* 9, 439–455, 1994.
- RIPLEY, B.D., *Stochastic Simulation*, New York, Wiley, 1987.
- RIPLEY, B.D., *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, 1996.

- RISSANEN, J., Universal Prior for Integers and Estimation by Minimum Description Length, *Ann. Statist.* 11(2), 416–431, 1983.
- ROBBINS, H. An empirical Bayes approach to statistics, *Proc. Third Berkeley Symp. Math. Statist. and Prob* 1, Univ. of Calif. Press, Berkeley CA, 1956, 157–164.
- ROBBINS, H., The empirical Bayes approach to statistical decision problems, *Ann. Math. Statist.* 35, 1–20, 1964.
- ROCKAFELLAR, R.T., *Convex Analysis*, Princeton, NJ, Princeton University Press, 1969.
- ROSENTHAL, J., Asymptotic variance and convergence rates of nearly periodic Markov chain Monte Carlo algorithms, *J. Amer. Statist. Assoc.* 98, 169–177, 2003.
- RUBIN, D.B., Comments on J. Neyman and causal inference in experiments and observational studies. On the application of probability theory to agriculture experiments. *Statist. Sci.* 4, 472–480, 1990.
- RUDEMO, M., Empirical choice of histograms and kernel density estimators *Scand. J. Statist.* 9, 65–78, 1982.
- RUDIN, C., DAUBECHIES, I. and SCHAPIRE, R.E., The dynamics of AdaBoost: Cyclic behavior and convergence of margins, *J. Mach. Learn. Res.* 5, 1557–1595, 2003.
- RUPPERT, D. and WAND, M.P., Multivariate weighted least squares regression, *Ann. Statist.* 22, 1346–1370, 1994.
- RUPPERT, D., WAND, M.P., HOLST, U. and HÖSSJER, O., Local polynomial variance-function estimation. *Technometrics* 39, 262–273, 1997.
- SACKS, J., WELCH, W.J., MITCHELL, T.J. and WYNN, H.P., Design and analysis of computer experiments, *Statist. Sci.* 4, 1989.
- SAVAGE, I.R., Contributions to the theory of rank order statistics, the two-sample case, *Ann. Math Statist.* 27, 590–615, 1956.
- SAVAGE, I.R., Contributions to the theory of rank order statistics – the “trend” case, *Ann. Math Statist.* 28, 968–977, 1957.
- SAVAGE, I.R., Lehmann alternatives, Proceedings of Conference on Nonparametric Statistical Inference, Budapest, Hungary, 1980, 795–821.
- SCHAPIRE, R.E., The strength of weak learnability, *Machine Learning* 5, 197–227, 1990.
- SCHERVISH, M., *Theory of Statistics*. New York, Springer, 1995.
- SCHWARZ, G., Estimating the dimension of a model, *Ann. Statist.* 6(2), 461–464, 1978.
- SCORNET, E., BIAU, G., and VERT, J., Consistency of random forests, *Ann. Statist.* 43, 1716–1741, 2015.
- SCOTT, D. W., On optimal and data-based histograms, *Biometrika* 66, 605–610, 1979.
- SCOTT, D.W., *Multivariate Density Estimation. Theory, Practice, and Visualization*. Wiley, New York, 1992.
- SEBER, G.A.F, *Multivariate Observations*, New York, Wiley, 1984.

- SERFLING, R.J., *Approximation Theorems of Mathematical Statistics*, New York, Wiley, 1980.
- SHAO, J. and TU, D., *The Jackknife and Bootstrap*, New York, Springer, 1995.
- SHEN, D., SHEN, H., and MARRON, J. S., Consistency of sparse PCA in high dimension, low sample size contexts, Technical report, 2011. Available at <http://arxiv.org/pdf/1104.4289v1.pdf>
- SHIBATA, R., An optimal selection of regression variables. *Biometrika* 68, 45–54, 1981.
- SHORACK, G.R., Weak convergence of the general quantile process in $\|\cdot/q\|$ -metrics. *Bull. Inst. Math. Statist.* 11, 60–71, 1982.
- SHORACK, G.R. and WELLNER, J.A., *Empirical Processes with Applications to Statistics*, New York, Wiley, 1986.
- SIBUYA, M., Generating doubly exponential random numbers, *Ann. Inst. Statist. Math. Tokyo Suppl.* VI–7, 1968.
- SILVERMAN, B., *Density Estimation for Statistics and Data Analysis*, London, Chapman and Hall, 1986.
- SKLAR, A., Fonctions de repartition a n dimensions et leurs marges, *L'Institut de Statistique de L'Universite de Paris*, 8, 1959, 229–231.
- SPIÓTVOLL, E., On the optimality of some multiple comparison procedures”, *Ann. Math. Statist.* 43, 398–411, 1972.
- STANLEY, R.P., *Enumerative Combinations*, Wadsworth, Monterey, 1986.
- STEIN, C.M., Efficient nonparametric testing and estimation, *Proc. Third Berkeley Symp. Math. Statist. Prob. 1*, Univ. California Press, Berkeley, 1956a, 187–195.
- STEIN, C.M., Inadmissibility of the usual estimator for the mean of a multivariate distribution *Proc. Third Berkeley Symp. Math. Statist. and Prob 1*, Univ. of Calif. Press, Berkeley CA, 1956b, 197–206.
- STEIN, C.M., Estimation of the mean of a multivariate normal distribution, *Ann. Statist.*, 9, 1135–1151, 1981.
- STEIN, C.M., Approximate computation of expectations, *Lecture Notes and Monograph Series V.7*, Hayward, CA, Institute of Mathematical Statistics, 1986.
- STIGLER, S., Linear functions of order statistics.” *Ann. Math. Statist.* 40, 770–788, 1969.
- STIGLER, S.M., Completeness and unbiased estimation, *The American Statistician* 26, 28–29, 1972.
- STONE, C.J., Optimal global rates of convergence for nonparametric regression, *Ann. Statist.* 10, 1040–1053, 1982.
- STONE, C.J., HANSEN, M., KOOPERBERG, C. and TRUONG, Y., Polynomial splines and their tensor products (with discussion and rejoinder by authors and Jianhua Z. Huang), *Ann. Statist.* 25, 1341–1470, 1997.
- STONE, C. J. and KOO, J. Y., Logspline density estimation, *Contemporary Mathematics* 59, 1–15, 1986.

- STOREY, J.D., The optimal discovery procedure: A new approach to simultaneous significance testing, *J. Roy. Statist. Soc. Ser. B* 69, 347–368, 2007.
- SUN, W. and CAI, T.T., Oracle and adaptive compound decision rules for false discovery rate control, *J. Amer. Statist. Assoc.* 102, 901–912, 2007.
- TALAGRAND, M., Sharper bounds for Gaussian and empirical processes. *Ann. Prob.* 22, 28–76, 1994.
- TAPIA, R. and THOMPSON, J., *Nonparametric Probability Density Estimation*, Baltimore, John Hopkins University Press, 1978.
- TIBSHIRANI, R., Regression shrinkage and selection via the lasso, *J. Roy. Statist. Soc. Ser. B* 58, 267–288, 1996.
- TIERNEY, L., Introduction to general state space Markov chain Theory, *Markov Chain Monte Carlo in Practice.*, CRC Press, 1995, p. 59.
- TIKHONOV, A., Solution of incorrectly formulated problems and the regularization method, *Soviet Math. Dokl.* 4, 1035–1038, 1963.
- TSIATIS, A.A., A large sample study of Cox’s regression model, *Ann. Statist.* 9, 93–108, 1981.
- TSYBAKOV, A., *Introduction to Nonparametric Estimation*, New York, Springer, 2008.
- TUKEY, J.W., Bias and confidence in not-quite large samples, *Ann. Math. Statist.* 29, 614, 1958.
- VAPNIK, V.N., *The Nature of Statistical Learning Theory*, 2nd Ed., 1998, New York, Springer, 1996.
- VAPNIK, V.N., *Statistical Learning Theory*. New York, Wiley, 1998.
- VAN DE GEER, S.A., *Applications of Empirical Processes Theory*, Vol. 6 of Cambridge Ser. in Statist. and Probabilistic Mathematics, 2000(a).
- VAN DE GEER, S.A., *Empirical Processes in M-Estimation*, Cambridge Univ. Press, Cambridge, 2000(b).
- VAN DER LAAN, M.J. and ROBINS, J.M. *Unified Methods for Censored Longitudinal Data and Causality*, New York, Springer, 2003.
- VAN DER VAART, A.W., *Asymptotic Statistics*, Cambridge, Cambridge Univ. Press, 1998.
- VAN DER VAART, A.W. and WELLNER, J., *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York, Springer, 1996.
- VAN ZWET, W.R., *Convex Transformation of Random Variables*, Math. Centrum, Amsterdam, 1964.
- VAN ZWET, W.R., A Berry-Esseen bound for symmetric statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 66, 425–440, 1984.
- VAUPEL, J.W., MANTON, K.G. and STALLARD, E., The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography* 16, 439–454, 1979.

- VENTER, J. and DE WET, T., Asymptotic distributions of certain test statistics, *South African Statist. J.* 6, 135–149, 1972.
- VIOLLAZ, A.J., Nonparametric estimation of probability density functions based on orthogonal expansions, *Rev. Mat. Uni. Complut. Madrid*, 41–84, 1989.
- VON MISES, R., On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.* 18, 309–348, 1947.
- VON NEUMANN, J., Various techniques used in connection with random digits, *National Bureau of Standards Applied Mathematics Series 12*, 36–38, 1951.
- WACHTER, K.W., The strong limits of random matrix spectra for sample matrices of independent elements”, *The Annals of Probability* 6, 1–18, 1978.
- WAHBA, G., On the numerical solution of Fredholm integral equations of the first kind, Tech. Report, Math. Research Center, Univ. of Wisconsin, Madison, 1969.
- WAHBA, G., *Spline Models for Observational Data*, SIAM, Philadelphia, 1990.
- WAINWRIGHT, M.J. and JORDAN, M.I., Graphical models, exponential families, and variational inference, *Foundations and Trends® in Machine Learning* 1, 2008, 1–305.
- WALD, A., Test of statistical hypothesis concerning several parameters when the number of observations is large, *Transactions of the American Math. Soc.* 54, 426–482, 1943.
- WALD, A., *Statistical Decision Functions*, Oxford, Wiley, 1950.
- WAND, M. P. and JONES, M. C., *Kernel Smoothing*, London, Chapman and Hall, 1995.
- WANG, Y., A likelihood ratio test against stochastic ordering in several populations, *J. Amer. Statist. Assoc.* 91, 1676–1683, 1996.
- WIDDER, D.V., *The Laplace Transform*, Princeton NJ, Princeton University Press, 1941.
- WIJSMAN, R., Invariant measures on groups and their use in statistics, Hayward, CA, IMS Lecture Notes, 1990.
- WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T., A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, *Biostatistics* 10, 515–534, 2009.
- YANG, F., On high dimensional data analysis and biomedical genomics, Ph.D. Thesis, Dept. of Statistics, Univ. of Wisconsin, 2013.
- YANG, F., DOKSUM, K. and TSUI, K.W., Principal component analysis (PCA) for high dimensional data. PCA is dead. Long live PCA. In: *Proceedings for Workshop on Perspectives on High Dimensional Data Analysis II, Montreal*, S.E. Ahmed, ed., Contemporary Mathematics, American Mathematical Society, Providence, RI, 2014.
- YUAN, M. and LIN, Y., Model selection and estimation in regression with grouped variables. *J. Royal Statist. Soc. Ser. B* 68(1), 49–67, 2007.
- ZENG, D. and LIN, D.Y., Efficient estimation for the accelerated failure time model. *J. Amer. Statist. Assoc.* 102, 1387–1396, 2007.
- ZENG, D. and LIN, D.Y., Maximum likelihood estimation in semiparametric regression models with censored data, *J.R. Statist. Soc. B* 69, 507–564, 2007a.

- ZHAO, P., ROCHA, G. and YU, B., The composite absolute penalties family for grouped and hierarchical variable selection, *Ann. Statist.* 37, 3468-3497, 2009.
- ZOU, H., The adaptive Lasso and its oracle properties, *J. Amer. Statist. Assoc.* 101, 1418-1429, 2006.

