

0212
H114
V.1

9860111

Shelby J. Haberman

Advanced Statistics

Volume I: Description of Populations



E9860111



Springer

Shelby J. Haberman
Northwestern University
Department of Statistics
Evanston, IL 60208
USA

With 9 illustrations.

Library of Congress Cataloging-in-Publication Data

Haberman, Shelby J.

Advanced statistics / Shelby J. Haberman.

p. cm. — (Springer series in statistics)

Includes bibliographical references and index.

Contents: v. 1. Description of populations

ISBN 0-387-94717-5 (v. 1 : hardcover : alk. paper)

1. Mathematical statistics. I. Title. II. Series.

QA276.H18 1996

519.5—dc20

96-10601

Printed on acid-free paper.

© 1996 Springer-Verlag New York, Inc.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

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

Production managed by Robert Wexler; manufacturing supervised by Jacqui Ashri.

Photocomposed pages prepared from the author's LaTeX file.

Printed and bound by Edwards Brothers, Inc., Ann Arbor, MI.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1

ISBN 0-387-94717-5 Springer-Verlag New York Berlin Heidelberg SPIN 10490451

Springer Series in Statistics

Advisors:

P. Bickel, P. Diggle, S. Fienberg, Krickeberg,
I. Olkin, N. Wermuth, S. Zeger

Springer

New York

Berlin

Heidelberg

Barcelona

Budapest

Hong Kong

London

Milan

Paris

Santa Clara

Singapore

Tokyo

Springer Series in Statistics

- Andersen/Borgan/Gill/Keiding*: Statistical Models Based on Counting Processes.
Andrews/Herzberg: Data: A Collection of Problems from Many Fields for the Student and Research Worker.
Anscombe: Computing in Statistical Science through APL.
Berger: Statistical Decision Theory and Bayesian Analysis, 2nd edition.
Bollmar/Zacks: Prediction Theory for Finite Populations.
Bremaud: Point Processes and Queues: Martingale Dynamics.
Brockwell/Davis: Time Series: Theory and Methods, 2nd edition.
Daley/Vere-Jones: An Introduction to the Theory of Point Processes.
Dzhaparidze: Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series.
Fahrmeir/Tutz: Multivariate Statistical Modelling Based on Generalized Linear Models.
Farrell: Multivariate Calculation.
Federer: Statistical Design and Analysis for Intercropping Experiments.
Fienberg/Hoaglin/Kruskal/Tanur (Eds.): A Statistical Model: Frederick Mosteller's Contributions to Statistics, Science and Public Policy.
Fisher/Sen: The Collected Works of Wassily Hoeffding.
Good: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses.
Goodman/Kruskal: Measures of Association for Cross Classifications.
Grandell: Aspects of Risk Theory.
Haberman: Advanced Statistics, Volume I: Description of Populations.
Hall: The Bootstrap and Edgeworth Expansion.
Härdle: Smoothing Techniques: With Implementation in S.
Hartigan: Bayes Theory.
Heyer: Theory of Statistical Experiments.
Huet/Bouvier/Gruet/Jolivet: Statistical Tools for Nonlinear Regression: A Practical Guide with S-PLUS Examples.
Jolliffe: Principal Component Analysis.
Kolen/Brennan: Test Equating: Methods and Practices.
Kotz/Johnson (Eds.): Breakthroughs in Statistics Volume I.
Kotz/Johnson (Eds.): Breakthroughs in Statistics Volume II.
Kres: Statistical Tables for Multivariate Analysis.
Le Cam: Asymptotic Methods in Statistical Decision Theory.
Le Cam/Yang: Asymptotics in Statistics: Some Basic Concepts.
Longford: Models for Uncertainty in Educational Testing.
Manoukian: Modern Concepts and Theorems of Mathematical Statistics.
Miller, Jr.: Simultaneous Statistical Inference, 2nd edition.
Mosteller/Wallace: Applied Bayesian and Classical Inference: The Case of *The Federalist Papers*.

(continued after index)

Preface

Statistics is a discipline devoted to the description of numerical measurements on members of populations. Given an accurate population census, statistics is concerned with the development of population parameters designed to summarize information on population members provided by numerical measurements. In practice, much less is often available than an accurate census. Commonly, measurement errors exist and only a sample of population members is available. In such cases, statistics considers approximation of population parameters by means of information gathered from the sample or information gathered by approximate rather than exact measurements. Thus, it is reasonable to take the position that use of sampling and approximate measurements is useless unless a statistician knows what to do if exact measurements on the whole population are available. Given this position, it follows that a book on statistics should begin by considering methods of population description. This approach is taken in this two-volume work. The first volume concerns definition and study of parameters useful for the description of measurements on population members. The second volume considers the use of samples to approximate population parameters.

Chapter 1 provides a basic introduction to measures of size, location, and dispersion. Desirable properties of such measures are described, and some basic examples are considered. Some readers may wish to omit proofs concerning order extensions and unions of measures of size.

Chapter 2 considers the Daniell integral, a very important special case of a measure of size, and the expectation, a Daniell integral which is also a measure of location. The use of Daniell integrals is natural given the

emphasis on measurements of location and size. The customary approach based on measure theory that is commonly encountered in graduate texts in statistics is somewhat more indirect. Although it is important to be familiar with such basic theorems such as the monotone and dominated convergence theorems, it is much less important to be familiar with their proofs. Consequently, it is possible for the reader to omit proofs of these results. Similarly, proofs related to the Daniell extension can be omitted if necessary.

Chapter 3 considers the problem of defining population variables which are readily studied by statistical methods. Measurable functions and random variables are defined and studied, and some basic approaches for their description are considered. Of particular note are descriptions of random variables by use of histograms and cumulative distribution functions and by use of expectations of bounded continuous transformations. Some readers, already familiar with measure theory, may find the derivations of Daniell integrals from measures to be particularly informative. Other readers may wish to avoid such derivations. Chapter 3 includes a substantial amount of material from classical mathematical analysis. It is important for the reader to understand that numerous commonly encountered functions in mathematics are Baire functions, and it is important to understand that Baire functions of measurable functions are themselves measurable functions under quite general conditions. The proofs of these results can be omitted if necessary.

Chapter 4 develops Lebesgue integrals by use of weak convergence. These Daniell integrals are shown to generalize the Riemann integrals of calculus. It is also shown that Lebesgue integrals have a central role in the development of a large class of important Daniell integrals and expectations. The concept of weak convergence is very important in this chapter and in Volume 2; however, the reader can reasonably consider omitting proofs concerning weak convergence, especially those which exploit local compactness and properties of upper and lower semicontinuous functions. Product integrals are important in statistical work; however, the reader may consider omitting the proofs of properties of these integrals.

Chapter 5 explores the classical problem of least squares. Basic dispersion measures, such as variances and standard deviations, are defined, and their properties are studied. Simple linear regression and multiple linear regression are described, and a very general description of the least squares method is provided. The treatment here emphasizes population description rather than statistical models. The reader may wish to omit proofs concerning the infinite-dimensional case.

Results are applied in Chapter 6 to the study of such basic statistical concepts as independence and conditional expectations. This treatment emphasizes the study of independence and conditional expectations in terms of predicting variables.

Chapter 7 discusses quantiles and measures of location and scale based on quantiles. The chapter considers basic problems of stability of parameters in the face of measurement errors. The material on least absolute error is not extensively used in the remainder of the book, so that it is relatively easy to omit.

Chapter 8 considers uses of moments and related functions to describe distributions of measurable functions and random variables. Distance measures based on moments are developed, and moment-generating functions, cumulant-generating functions, cumulants, and characteristic functions are applied to population description. General formulas relating moments and cumulants are not often used, although results up to fourth moments and fourth cumulants are quite important. Consequently some of the most general formulas can be omitted. In addition, characterization of a distribution by using moment generating functions or by using moments is not used very often in the remainder of the book, so this topic can be omitted if necessary.

In the second volume, Chapter 9 provides a general discussion of approximation of distributions by using sampling. Convergence in distribution is developed, and classical limiting results such as the central limit theorem are presented. Chapter 10 describes simple random sampling with replacement, the basic sampling method used in statistical inference. Chapter 11 describes such alternative sampling methods as simple random sampling without replacement and stratified random sampling. Chapter 12 considers confidence weights. These weights are closely related to conventional confidence intervals. Chapter 13 explores assessment of models. This chapter differs somewhat from conventional treatments of hypothesis tests to the extent that models are assessed both in terms of validity and in terms of their value as approximations. Chapter 14 examines inferences concerning least squares. Chapter 15 considers inferences for quantiles. Chapter 16 examines prediction of nominal and ordinal variables.

This book is intended for use by graduate students in statistics and by professional statisticians. The reader is assumed to have a good knowledge of analysis and linear algebra, so that open sets, continuous functions, differentials, Riemann integrals, matrices, and vectors are familiar terms. A prior background in statistics is not a formal requirement, although previous training in statistics will obviously be helpful. A moderate familiarity with statistical packages or computer languages is extremely helpful to a student who wishes to work through the exercises. The book makes significant mathematical demands on the reader to the extent that strong efforts are made to provide rigorous statements and proofs of results. To assist the reader with notation, the beginning of the index provides contains a list of mathematical symbols used, together with the page reference for the first use of the notation. For many readers, an initial reading of the book may be accomplished by skipping proofs of results.

This book differs considerably from conventional general books on statistics. The emphasis on description of populations separates this book from

such advanced works in statistics as Rao (1973), Cox and Hinkley (1974), and Bickel and Doksum (1977). This emphasis on population description is shared throughout the many editions of Kendall and Stuart (1977, 1979) and Kendall, Stuart, and Ord (1983); however, that three-volume treatment of statistics is too large to be readily used as a text and is not especially rigorous in presenting mathematical results.

This book also differs from conventional texts because it emphasizes the development and measurement of population parameters without using limited probability models. In this framework, the book carefully considers in Volume 2 use of a sample mean to estimate a population mean, for such an estimator can be used with at least some success whenever the population mean is defined. On the other hand, use of the sample mean to estimate the population median under the assumption that the population distribution is normal is not advocated here because the procedure can be quite unsatisfactory if the population distribution is not normal. The emphasis on statistical procedures which can be used under quite general conditions reflects the emphasis on robustness in Tukey (1962) and Huber (1981); however, this book provides somewhat more emphasis on population parameters which are readily interpreted.

The coverage of statistical topics is sufficiently broad that a student completing a course based on this book should be able to apply a number of standard statistical procedures and should have some reasonable knowledge of conditions under which they are appropriate. Numerical examples are provided to ensure that the student has practice in analyzing data via the methods presented in this book.

Preparation of this book has greatly benefited from the use of a preliminary draft in a graduate class at Northwestern University. The students helped greatly in clarifying which material needed more work and in finding errors in typing or substance. Tom Severini reviewed the manuscript and provided many helpful suggestions. The remaining errors are all the author's responsibility. Research for this book was partially supported by National Science Foundation grants DMS-867373, DMS-8900018, and DMS-9303713.

Springer Series in Statistics

(continued from p. ii)

- Pollard*: Convergence of Stochastic Processes.
- Pratt/Gibbons*: Concepts of Nonparametric Theory.
- Read/Cressie*: Goodness-of-Fit Statistics for Discrete Multivariate Data.
- Reinsel*: Elements of Multivariate Time Series Analysis.
- Reiss*: A Course on Point Processes.
- Reiss*: Approximate Distributions of Order Statistics: With Applications to Non-parametric Statistics.
- Rieder*: Robust Asymptotic Statistics.
- Rosenbaum*: Observational Studies.
- Ross*: Nonlinear Estimation.
- Sachs*: Applied Statistics: A Handbook of Techniques, 2nd edition.
- Särndal/Swensson/Wretman*: Model Assisted Survey Sampling.
- Schervish*: Theory of Statistics.
- Seneta*: Non-Negative Matrices and Markov Chains, 2nd edition.
- Shao/Tu*: The Jackknife and Bootstrap.
- Siegmund*: Sequential Analysis: Tests and Confidence Intervals.
- Simonoff*: Smoothing Methods in Statistics.
- Tanner*: Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, 3rd edition.
- Tong*: The Multivariate Normal Distribution.
- van der Vaart/Wellner*: Weak Convergence and Empirical Processes: With Applications to Statistics.
- Vapnik*: Estimation of Dependences Based on Empirical Data.
- Weerahandi*: Exact Statistical Methods for Data Analysis.
- West/Harrison*: Bayesian Forecasting and Dynamic Models.
- Wolter*: Introduction to Variance Estimation.
- Yaglom*: Correlation Theory of Stationary and Related Random Functions I: Basic Results.
- Yaglom*: Correlation Theory of Stationary and Related Random Functions II: Supplementary Notes and References.

9704/4

Contents

Preface	v
1 Populations, Measurements, and Parameters	1
1.1 Basic parameters	5
1.1.1 Infima, suprema, and ranges	5
1.1.2 Relative suprema, infima, and ranges	7
1.1.3 Sums	10
1.1.4 Partial sums	12
1.1.5 Counts and fractions	14
1.1.6 Conditional counts and fractions	14
1.1.7 Weighted sums	15
1.1.8 Characterization of weighted sums	21
1.1.9 Distributions	22
1.1.10 Inverse distributions	27
1.2 Measurement of size and dispersion	29
1.2.1 Order extensions	30
1.2.2 Measures of location	41
1.2.3 Homogeneity	43
1.2.4 Homogeneous measures of size and location	46
1.2.5 Positive and absolute homogeneity	46
1.2.6 Additivity	50
1.2.7 Finite additivity	54
1.2.8 Linear subspaces and linear functionals	56
1.2.9 Linear combinations	58

1.2.10	Vector-valued variables and positive linear functionals	61
1.2.11	Subadditive and superadditive functions	62
1.2.12	Linear lattices	66
1.2.13	The maximum norm	69
1.2.14	Seminorms derived from positive linear functionals	70
1.2.15	Mean deviations about the mean	72
1.2.16	Additive measures of size on linear lattices	73
1.3	Uncertainty and positive linear functionals	74
1.4	Exercises	82
2	Expectations and Daniell Integrals	87
2.1	Examples of Daniell integrals	89
2.2	Limits and Daniell integrals	96
2.2.1	Limits and sums	96
2.2.2	The monotone convergence theorem	100
2.2.3	Suprema and infima	103
2.2.4	Dominated convergence	107
2.3	From Daniell preintegrals to Daniell integrals	109
2.3.1	Dominance	111
2.3.2	Countable bounding	112
2.3.3	The Daniell extension	117
2.3.4	Closed Daniell integrals	120
2.4	Exercises	123
3	Random Variables and Measurable Functions	127
3.1	Measurable functions	127
3.1.1	Examples of real measurable functions	128
3.1.2	Limits of measurable vector variables	132
3.1.3	Sequentially-closed sets	133
3.1.4	Linear lattices	134
3.1.5	Finite measures and probabilities	138
3.1.6	Measure and probability distributions	141
3.1.7	Daniell integrals generated from finite measures	143
3.1.8	Real measurable functions with Daniell integrals	150
3.2	Regular Daniell integrals and continuous functions	152
3.2.1	Regular Daniell integrals and Baire functions	158
3.2.2	Baire sets	164
3.2.3	Open and closed sets	164
3.2.4	Intervals	169
3.2.5	Compact sets	176
3.2.6	Continuous real functions with compact support	177
3.2.7	Monotone functions	183
3.3	Intervals and distributions	183
3.3.1	Histograms	185
3.3.2	Cumulative distribution functions	186

3.4 Exercises	190
4 Construction of Daniell Integrals	199
4.1 Integral-generating linear lattices	199
4.1.1 Weak convergence	201
4.1.2 Continuous functions with compact support	204
4.1.3 Lebesgue integrals for the real line	210
4.1.4 Lebesgue integrals for subpopulations of the line	211
4.1.5 Lebesgue integrals and uniform expectations	217
4.1.6 Continuous integrals	220
4.2 Transformations and Lebesgue integrals	220
4.2.1 Monotone transformations	221
4.2.2 Generation of real random variables	225
4.3 Product integrals and Lebesgue integrals	227
4.3.1 Repeated integration	233
4.3.2 Tensor products	235
4.3.3 Differentiable transformations	244
4.4 Exercises	250
5 Least Squares	265
5.1 Square-integrable functions	265
5.1.1 Second moments	266
5.1.2 Matrix computations and second moments	267
5.1.3 Square seminorms	269
5.1.4 The variance and standard deviation	270
5.2 Mean-squared error and least-squares predictors	276
5.2.1 Minimum mean-squared error	277
5.2.2 Comparison of minimum mean-squared error	280
5.2.3 Least-squares predictors	281
5.2.4 General properties of least-squares predictors	284
5.2.5 Normal equations	287
5.2.6 Orthogonal linear subspaces	289
5.3 Simple linear regression	293
5.3.1 Correlation	295
5.3.2 Rank correlation	298
5.4 Multiple linear regression	304
5.4.1 Multiple correlation	308
5.4.2 Partial correlation	309
5.5 Least-squares prediction for infinite-dimensional linear subspaces	310
5.6 Exercises	317
6 Independence and Dependence	325
6.1 Independence and dependence	325
6.1.1 Examples of independence	327

6.1.2	Mutual independence	330
6.1.3	Product independence	333
6.2	Conditional expectations	335
6.2.1	Conditional expectations and least squares	335
6.2.2	Prediction by discrete random variables	341
6.2.3	Product integrals	347
6.2.4	Linear regression models and conditional expectations	355
6.3	Exercises	357
7	Quantiles	367
7.1	Definition of quantiles	367
7.1.1	Quantile functions	371
7.1.2	Characterization of distributions by quantiles	373
7.2	Measures of location	373
7.2.1	Monotone-increasing transformations	375
7.2.2	Errors in measurement	376
7.2.3	Trimmed means	378
7.3	Measures of dispersion based on quantile functions	383
7.3.1	Half-ranges	384
7.3.2	Mean deviations about the median	385
7.3.3	Trimmed variances and standard deviations	389
7.3.4	Mean differences	392
7.4	Medians and prediction by mean absolute error	396
7.5	Exercises	400
8	Moments	405
8.1	Moments for real measurable functions	405
8.1.1	Seminorms	407
8.1.2	Distance measurement	410
8.1.3	Central moments	411
8.1.4	Skewness	414
8.1.5	Polynomial regression	416
8.1.6	Power series	419
8.1.7	The normal cumulative distribution function	421
8.1.8	Normal quantiles	423
8.2	Moment-generating functions	427
8.2.1	Moments and moment-generating functions	429
8.2.2	Complex moment-generating functions	433
8.2.3	Sums of independent variables	441
8.2.4	Cumulants	441
8.2.5	Characteristic functions	446
8.2.6	Characterization of distributions	447
8.3	Multivariate moments	451
8.3.1	Power series	452

8.3.2 Moment-generating functions	454
8.4 Exercises	458
Bibliography	467
Index	475

List of Tables

1.1	State Populations in 1980	2
1.2	The Population of the United States by Region	4
1.3	The Number of Inhabitants per State, Classified by Region	9
1.4	Live Births in 1985 in the United States by Age of Mother	23
3.1	States Classified by Number of 1980 Inhabitants	185
5.1	State Populations in Thousands from 1940 to 1980	299
5.2	Means and Standard Deviations of the Natural Logarithms of State Populations in Thousands	300
5.3	Covariances of the Natural Logarithms of State Populations in Thousands	301
5.4	Correlations of the Natural Logarithms of State Populations in Thousands	301
5.5	Partial Correlations of the Natural Logarithms of State Populations in Thousands Given the Natural Logarithm of State Populations in 1940	311
5.6	Annual Precipitation in Inches in Central Park from 1868 through 1970	319
6.1	State Populations in 1980 by Region	345
6.2	Conditional Expectations and Standard Deviations of State Populations by Region	346

6.3	Inhabitants of the United States in 1970 Classified by Type of Residence and Sex	359
6.4	Inhabitants of the United States in 1990 Classified by Region of Residence and Race	362
7.1	Conditional Medians and Mean Deviations for State Populations by Region	399
8.1	The Number of Daily Newspapers in the United States from 1920 to 1970	418
8.2	Computation of $\Phi(\cdot)$	423