

STATISTICS

Concepts and Methods

Third Edition

Ditlev Monrad
E. James Harner
William F. Stout
with Xuming He and
Louis A. Roussos



Möbius Publishing, Ltd.
Champaign, Illinois



Statistics: Concepts and Methods, Third Edition

This book is published by Möbius Publishing, Ltd.
1802 South Duncan Road, Champaign, IL 61822
Telephone: (217) 398-9086
www.8-mobius.com

Design and composition: Publication Services, Inc.
Cover design: David W. Eynon

Copyright © 2008 by Möbius Publishing, Ltd.
All rights reserved. No part of this publication may
be reproduced, stored, or transmitted by any means
without the prior written permission of the publisher.

1 2 3 4 5 6 7 8 9 10 01 03

Produced in the United States of America.

ISBN 13:1978-1-928583-11-0

CONTENTS

Preface	vii
About the Authors	xi
Acknowledgments	xiii
Part I Describing Data	1
1 Exploring Data by Graphical Methods	3
1.1 The Science of Statistics	5
1.2 Displaying Small Sets of Numbers: Dotplots and Stem-and-Leaf Displays	14
1.3 Graphing Categorical Data	25
1.4 Frequency Histograms	30
1.5 Density Histograms	38
1.6 Misusing Statistics	46
Chapter Review Exercises	61
2 Summarizing Data by Numerical Measures: Center and Spread	67
2.1 The Center of a Data Set	69
2.2 Mean versus Median versus Mode as a Measure of Center	78
2.3 Measuring the Spread of a Data Set: The Standard Deviation	90
2.4 The Normal Approximation for Data	108
2.5 Boxplot: The Five-Number Summary	116
Chapter Review Exercises	132
3 Linear Relationships: Regression and Correlation	139
3.1 Scatterplots	142
3.2 The Correlation Coefficient	147
3.3 Regression	163
3.4 The Question of Causation	194
Chapter Review Exercises	197
Part II Probability Modeling and Obtaining Data	203
4 Probabilities and Simulation	205
4.1 Experimental Probability	207

4.2	Probability Models	215
4.3	Simulation: A Powerful Tool for Learning and Doing Statistics	241
4.4	Simulating Random Sampling via a Box Model	262
4.5	Random Sampling with or without Replacement	269
	Chapter Review Exercises	277
5	Expected Value and Simulation	285
5.1	The Expected Value (Theoretical Mean) of a Random Variable: Their Simulation	287
5.2	Using Five-Step Simulation to Estimate Mean Values	299
5.3	The Standard Deviation of a Random Variable	308
	Chapter Review Exercises	316
6	Probability Distributions: The Essentials	319
6.1	The Binomial Distribution	321
6.2	The Geometric Distribution	331
6.3	The Poisson Distribution	335
6.4	The Normal Probability Distribution	347
	Chapter Review Exercises	367
7	Obtaining Data: Random Sampling and Randomized Experiments	373
7.1	Introduction	374
7.2	Survey Sampling from a Real Population: Probability Sampling versus Non-Probability Sampling	380
7.3	Experimental Design: Observational Studies versus Randomized Experiments	396
7.4	Randomized Block Designs, Including Matched-Pairs Designs	410
	Chapter Review Exercises	415
Part III	Statistical Inference: Estimation and Hypothesis Testing	421
8	Confidence Interval Estimation	423
8.1	Bias and Chance Error	424
8.2	The Large Sample Distributions of \bar{X} and \hat{p}	438
8.3	The Standard Error	458
8.4	Large-Sample Confidence Interval for the Population Proportion p	465
	Chapter Review Exercises	474

9	An Introduction to Hypothesis Testing	481
9.1	The Null Hypothesis and the Alternative Hypothesis	486
9.2	Tests for a Population Proportion	489
9.3	Tests for Randomized Controlled Experiments Producing Sample Proportions	500
	Chapter Review Exercises	510
10	Estimating with Confidence	516
10.1	Confidence Intervals for a Population Mean μ when δ is known	519
10.2	Confidence Intervals for a Population Mean μ when δ is unknown	526
10.3	Large-Sample Confidence Intervals for a Population Proportion ρ	532
10.4	Using Bootstrapping to Obtain Large-Sample Confidence Intervals	535
10.5	Confidence Interval for the Difference Between Two Population Means $\mu_x - \mu_y$	537
10.6	Large-Sample Confidence Interval for the Difference $p_1 - p_2$ between Two Population Proportions	544
10.7	Confidence Interval for the Differences of Two Population Means in the matched-Pairs Design Case: $\mu_D = \mu_X - \mu_Y$	546
10.8	Point Estimate for the Population Variance δ^2 and SD δ ; Unbiasedness and Biasedness of the Various Estimators	548
	Chapter Review Exercises	552
11	More on Hypothesis Testing	557
11.1	Tests for a Population Mean	558
11.2	Tests for Equality of Two Population proportions and Equality of Two Population Means	576
11.3	One-Sided and Two-Sided Hypothesis Testing	591
11.4	Significance Testing Versus Acceptance/Rejection Testing: Concepts And Methods	596
	Chapter Review Exercises	609
12	Chi-Square Testing	615
12.1	Is the Die Fair?	617
12.2	How Big a Difference in the D Statistic Makes a Difference?	620
12.3	The Chi-Square Statistic	625
12.4	Real-Life Chi-Square Examples	630
12.5	The Chi-Square Density	639

12.6	The Chi-Square Distribution and Its Use for Chi-Square Testing	648
12.7	Unequal Expected Frequencies	656
12.8	Chi-Square Tests of Independence and Homogeneity for a Two-Way Contingency Table	667
	Chapter Review Exercises	679
13	Inference About Regression	686
13.1	Inference about the Regression Line Slope	688
13.2	Confidence Interval for Regression-Based Prediction of Y Given x and for Estimation of the Line $E(Y x)$	713
13.3	Applying Regression to Nonlinear Relationships by Transforming the Variables	717
	Chapter Review Exercises	727
	Glossary	731
	Appendices:	
A	Computationally Generated Random Digits	754
B	Random Number Tables	756
C	Chi-Square Probabilities	760
D	Linear Interpolation	761
E	Normal Probabilities	763
F	Students t Probabilities	765
G	Binomial Probabilities	766
H	F-Distribution Probabilities	774
I	Bonferroni Confidence Intervals	777
J	Cumulative Poisson Probabilities	778
	Index	781

PREFACE

The authors strongly believe that the most effective way for beginning statistics students to learn statistics well is for them to directly experience its central concepts and standard procedures by closely observing and actively interacting with simulated and real data sets. The particular advantage of having a student simulate data sets is that the student then actually experiences the underlying mechanism—that is, the probability model—producing his or her data. Then students experience firsthand the process of statistically analyzing data, and in this way their statistical expertise develops. Actively learning statistics by direct contact and interaction with data is facilitated in this textbook by the demonstration of statistical concepts using the five-step simulation method throughout. Carrying out these five steps, students design simulation studies, and often also solve their statistical problems, facilitated by the textbook's powerful supporting software. This software is fast, Web-based, point-and-click, and instructional. It produces data-driven demonstrations of fundamental statistical concepts and illustrates highly accurate simulationbased statistical analyses that are heavily used by professional statisticians in the modern era of statistics.

A vital complement to the experiential, simulation-based, data-driven presentation of statistical concepts and methods is a clear and understandable combined verbal and formula-based presentation of the formal, logical, and deductive aspects of the subject, using only the moderate amount of mathematical formalism required to make the concepts and procedures understandable. In particular, the normal population sampling and the central limit theorem and other large-sample approaches to statistical inference are given prominence just as they are in a traditional statistics textbook. However, the modern simulation-based approach to inference is also clearly and completely presented. Inference thus has three distinct approaches: normal population, central limit theorem/large-sample, and simulation-based.

In the textbook, the theoretical formal aspects of each major topic are intermingled with rich real-data-based and, when appropriate, simulation-based examples of concepts and methods. The basic aspects of probability that are so important to a statistics course (and often very difficult for beginning students to master) are introduced as empirical phenomena and only then are abstracted and expanded into a body of useful probability concepts and accompanying formulas in Chapters 4, 5, and 6. The capstone of this empirical emphasis is the five-step computer-based simulation that is used throughout. Simulation allows students to transcend (and often avoid) abstract mathematical formalism and yet gain a deep understanding of the empirical role that probability plays in statistical thinking. In Chapter 6 the core probability models (normal, binomial, geometric, and Poisson) are thoroughly discussed with an emphasis on their role in providing descriptions or models of realworld random phenomena.

Pedagogically, the textbook stands on three legs: (1) carefully crafted and thorough verbal explanations of concepts and procedures; (2) clear, formula-based descriptions and explanations of statistical concepts and procedures, including both traditional procedures and the more recently emphasized simulation-based procedures; and (3) immersion in the world of numerical chance phenomena and their statistical analyses. This immersion is achieved through (i) textbook-supplied and student-created simulation studies, made possible by the textbook's fast, sophisticated, and easy-to-use software, and (ii) interesting and compelling examples rich in real-world data.

To be most effective, an introductory statistics textbook must strike a balance: On the one hand, it must help the student to learn basic statistical concepts in depth and, ultimately, to think statistically; on the other hand, it must also expose the student to the evolving body of

statistical procedures that are widely used in practice and are judged to constitute statistical literacy. This balance, a central goal of this book, equips students with both real statistical discernment and the needed familiarity with the body of statistical procedures so widely used in science, business, and government, and so often reported in the media. Without this balance, the course becomes only a “liberal arts” introduction to statistics, failing to equip students with the capability to analyze data sets. The textbook is divided into three parts: (1) using numerical and graphical techniques to explore and discover possibly meaningful and important patterns in data; (2)(i) obtaining quality data for a planned statistical study via experimental designs that use randomization and via surveys that use random sampling; (ii) actively experiencing the fundamental role of probability models in statistics, with the mathematical formalism of probability nicely interwoven with the rich and active experiencing of randomness via simulations; and (3) using statistical inference as a tool to draw valid and useful conclusions about important real-world settings from quality data via estimation and hypothesis testing.

In today’s technology-rich learning environments, an introductory statistics textbook must be designed to make profound and meaningful use of available computer-based technology, both to be pedagogically sound and to help students carry out statistical inferences with real and sometimes sizable data sets. Instructional software must be platform independent, extremely easy to use for all (point-and-click is vital, because there is no time available in a typical jam-packed statistical course syllabus for a week or more of computer software training!), and thoroughly integrated with the textbook. In a statistics course, the software should make the concepts and methods come alive, be a rich source of data, and make complex statistical computations and large-scale, simulation-based demonstrations of concepts and simulation-based inferences fast and easy to carry out. The software must be an “active learning” tool, to use an overworked but apt phrase. The accompanying simulation-based, Web-delivered, customized instructional software does all of these things, using the five-step method to allow students to construct simulations that facilitate learning.

The textbook perhaps breaks new ground with its in-depth treatment of bootstrap-based and other resampling inferential procedures that require simulation. The bootstrap approach is one of the most important advances of the past 25 years in statistical methodology and has become an essential component of the practicing statistician’s toolbox. Hence, it needs to be taught at the elementary level. Consistent with the instructional approach of the book, the bootstrap method is simulation-based and, as such, meshes seamlessly with the five-step method used throughout. The bootstrap approach of this textbook is natural and intuitive, which makes it easy for the beginning statistics student to understand. The bootstrap provides the student with a powerful, cutting-edge, and widely used method of statistical inference. The bootstrap together with the traditional small-sample, normal population inferential approaches (such as the t -test) and large-sample inferential approaches (based on the central limit theorem and on the associated large-sample result for the chi-square statistic) provide the student with enormous inferential power.

The book is written to be accessible to all students having at least a modest exposure to algebra; intermediate algebra suffices. Although mathematical concepts and formulas appear frequently, and students are encouraged to really think about what they are learning from such formulas, the amount of formal mathematical background needed is minimal. The book should work very well in any noncalculus-based introductory statistics course. The instructor has the luxury, if he or she desires, of downplaying large-sample and normal population-based inference and the formal structure of probability theory. In fact, an instructor can even design a course that puts most of its focus on the simulation-based approach to statistical inference. The textbook provides a choice between the major stress being placed on simulation, and on the other hand, a balanced approach to sampling from a simulation, normal population, and

large sample theory.

Several influential, nationally circulated reports have stressed the need for new emphases in the teaching of statistics. Most notably, the widely influential American Statistical Association/Mathematical Association of America Cobb report recommends that the teaching of statistics be heavily based on data and proposes that more emphasis be placed on statistical concepts than on abstract theory. Further, it stresses active learning and simulation-based learning. This textbook is tailor-made to address these valid recommendations and requirements, not only because of its simulation-based approach but also because of numerous data-driven examples and exercises (provided after each section), many encouraging students to use the instructional software. Thus, students have ample opportunity to practice using the concepts and methods learned in the text.

Part I of the book begins with three chapters describing how one explores and summarizes important patterns occurring in data—data being the focus of statistics—with tables and graphs and by means of numerical (statistical) indices. This includes a descriptive introduction to linear regression in Chapter 3.

Part II presents probability modeling empirically and formally. It also describes how statisticians (and students) can plan a study whose goal is to collect quality data from which valid statistical inferences can be made. Chapters 4 and 5 provide a heavily empirical and simulation-based introduction to probability through emphasis on probability and probability distributions, and on the mean and standard deviation of a random variable or of its probability distribution. This is appropriate because probability is the logical underpinning of inferential statistics. In particular, probability, expected value, and standard deviation are introduced as empirical concepts, with large-scale simulated samples (routinely 10,000 simulations conducted, and even 100,000 occasionally) often used to accurately estimate unknown theoretical probabilities, expected values, and standard deviations. This is facilitated by student usage of the supporting instructional software.

Chapter 6 supplements the basic introduction to probability in Chapters 4 and 5 with four core probability distributions—normal, binomial, geometric, and Poisson—and the standard large-sample relationships, involving them, such as the normal approximation to the binomial. Some instructors may wish to de-emphasize or omit portions of Chapter 6. Chapter 7 tells the student how one obtains good data for a planned statistical study through probability sampling of real populations and by well-designed randomized experiments. One section of the chapter provides information on how students can validly collect data for their own planned statistical studies.

Part III then covers statistical inference, focusing on confidence intervals and hypothesis testing. Chapter 8 presents an introduction to estimation, beginning with a discussion of the two sources of estimation error, namely bias and chance variation. Then the large-sample estimation of a population mean μ and of a population proportion p are developed using the central limit theorem (CLT) for \bar{X} .

Chapter 9 introduces hypothesis testing. After an explanation of fundamental concepts, testing of $H_0 : p = p_0$ is discussed via both the simulation approach and the large sample CLT for \hat{p} approach. The chapter closes with a novel simulation approach to randomized controlled experiment hypothesis testing when population members each display one of two characteristics (e.g., recover from illness or die).

Chapters 10 and 11 then cover the important confidence interval and hypothesis testing procedures in the standard one- and two-population settings involving means and proportions, stressing normal populations inference, large-sample inference, and simulation-based inference.

Chapter 12 then presents chi-square testing, beginning with the more widely applicable (not requiring a large sample) simulation-based approach. Then the traditional large-sample,

chi-square distribution approach (using the chi-square distribution) is presented. Coverage includes both the usual multinomial distribution chi-square test (null hypothesis of a specified many-sided, possibly unfair die) and contingency table-based chi-square tests of independence and multiple-population homogeneity.

Chapter 13 complements the descriptive treatment of linear regression in Chapter 3 with an inference-based treatment of linear regression.

The entire book can easily be covered in a two-semester course. However, a rich one-semester course can be nicely carried out, with careful planning and judicious choices of topics covered. If desired, the instructor can omit Chapter 12 without loss of continuity. In this regard, starred (*) sections or chapters can also be omitted without loss of continuity.

This textbook has been heavily influenced by instructor and student input based on classroom use of earlier editions and other textbooks written by some of the authors. In particular, Chapter 7 on data collection was developed in response to such input.

One unique aspect of the textbook is the integration of the printed material with easy-to-use simulation software. They combine to enrich the learning experience as well as allow easy access to a large set of simulation-based inferential procedures often used in modern statistical practice. When use of the software is especially appropriate, a computer symbol sometimes appears in the margin. A hand-held calculator will often prove useful for ordinary data compilation, and, a symbol of a calculator may appear in such contexts in the margin. A key graphic indicates material related to the chapter's Key Problem.

ABOUT THE AUTHORS

Ditlev Monrad is an Associate Professor in the Department of Statistics and the Department of Mathematics at the University of Illinois at Urbana-Champaign (UIUC). His research is in the area of theoretical probability. A dedicated teacher, Prof. Monrad has served for many years as an undergraduate and graduate advisor for statistics majors at UIUC. He has extensive experience in teaching introductory statistics and probability at the 4-year college level.

William F. Stout is a Professor of Statistics at UIUC, where he has been on the faculty since 1967. He is an internationally acclaimed researcher in the application of statistics to the fields of educational and psychological measurement. He has been the thesis research advisor of 17 Ph.D. students working on the interface of statistical and educational measurement. As the founder of the University of Illinois Statistical Laboratory for Educational and Psychological Measurement, Prof. Stout has led the development of new and widely applicable theories and methods for improving standardized testing. He is a past president of the International Psychometric Society and was the director of an Education Testing Service research group that developed statistical tools to carry out skills-level formative assessments of students by using standardized tests. Prof. Stout is the author of over 60 books and published research papers. He has held National Foundation posts continuously for almost 40 years.

E. James Harner is a Professor and Chair of the Department of Statistics at West Virginia University. His principal research efforts are in environmental statistics, statistical computing environments, and dynamic graphics, though he is increasingly focusing on learning-based probability models and bioinformatics. He has over 40 published research papers and has co-developed several statistical software and Web-based learning environments. Currently, he is directing the development of IDEAL (Intelligent Distributed Environment for Adaptive Learning), a Webbased learning environment that will incorporate advanced tutoring and assessment subsystems into college instruction.

Contributing Authors

Xuming He is a Professor in the Department of Statistics at UIUC. Prof. He is the director of the department's statistical consulting service (Illinois Statistics Office). In addition, Prof. He serves as a consultant to the Argonne National Laboratory and is a member of the editorial board for the Journal of Multivariate Analysis, Statistica Sinica, and Statistics and Probability Letters. A instructor highly ranked by students, Prof. He has taught statistics courses at both undergraduate and graduate levels since 1989.

Prof. He's principal research areas are regression models of larger dimensions, regression splines and constrained models, robust methods in linear models, and asymptotics of psychometric model estimation. Much of his research has been conducted under grants from the National Security Agency and the National Science Foundation.

Louis Roussos is a senior psychometric scientist at the Measured Progress testing company. Once an aerospace engineer who developed mathematical models for acoustics and vibrations research, his major research interest now centers on analyzing educational and psychological tests. Among his many accomplishments, Prof. Roussos received the 1997 National Council of Measurement in Education (NCME) Outstanding Dissertation Award and the 1999 American Psychological Association's Division 5 Dissertation Award. He also received the NCME Outstanding Application of Measurement Theory Award in 2006.

With extensive statistics and mathematics teaching experience, Prof. Roussos has both pedagogical expertise and test analysis competence. He has developed new statistical tech-

niques and software for dimensionality analysis of standardized tests and test bias, and has designed computerized adaptive standardized tests. Prof. Roussos has developed new statistical procedures that constitute a major advance in test-based formative skills-level assessment.

ACKNOWLEDGMENTS

The authors wish to acknowledge their appreciation for the superb, in-depth intellectual and technical support provided by Möbius Communications in the development and preparation of this book. Möbius Communications Product Manager Louise Toft developed the Professional Profiles that open each chapter and was responsible for many of the practical details involved in taking a manuscript and turning it into a book. Brandon M. Warga, Ben Coblentz, Foti Kutil, Pamela J. Broderick-Rhoades, Peter A. Nelson, Nannette Monet, Jessica Matthews, Jyothirmai Gubili, and Alysia Cooley worked as a team to help turn the materials from the different authors into a unified textbook.

The authors wish to thank the publishers of the following materials for granting permission for their use:

Chapter 1: opening photos courtesy of Dr. Andrea Donnellan.

Chapter 2: opening photos courtesy of Dr. Andrew Harris.

Chapter 3: opening photos courtesy of the National Safety Council.

Chapter 4: opening photos courtesy of Dr. Paul W. Chodas and the Solar System Dynamic Group at the Jet Propulsion Laboratory.

Chapter 5: opening illustrations courtesy of Dr. J. Steven Landefeld and the Bureau of Economic Analysis.

Chapter 6: opening photos courtesy of the Jet Propulsion Laboratory.

Chapter 7: opening photos courtesy of Bruce W. Hoynoski and Nielsen Media Research.

Chapter 8: opening photos courtesy of Office of Migratory Bird Management and the Duck Stamp Office, U. S. Fish and Wildlife Service.

Chapter 9: opening photos courtesy of Dr. Sean Todd and the Allied Whale Marine Mammal Research Facility at the College of the Atlantic.

Chapter 10: opening photos courtesy of Dr. Harold Brooks. The map is from the National Severe Storms Laboratory Web site, www.nssl.noaa.gov/~brooks.concannon and was used with the permission of the National Severe Storms Laboratory.

Chapter 11: opening photos courtesy of Dr. Jian Zhang.

—

—

|