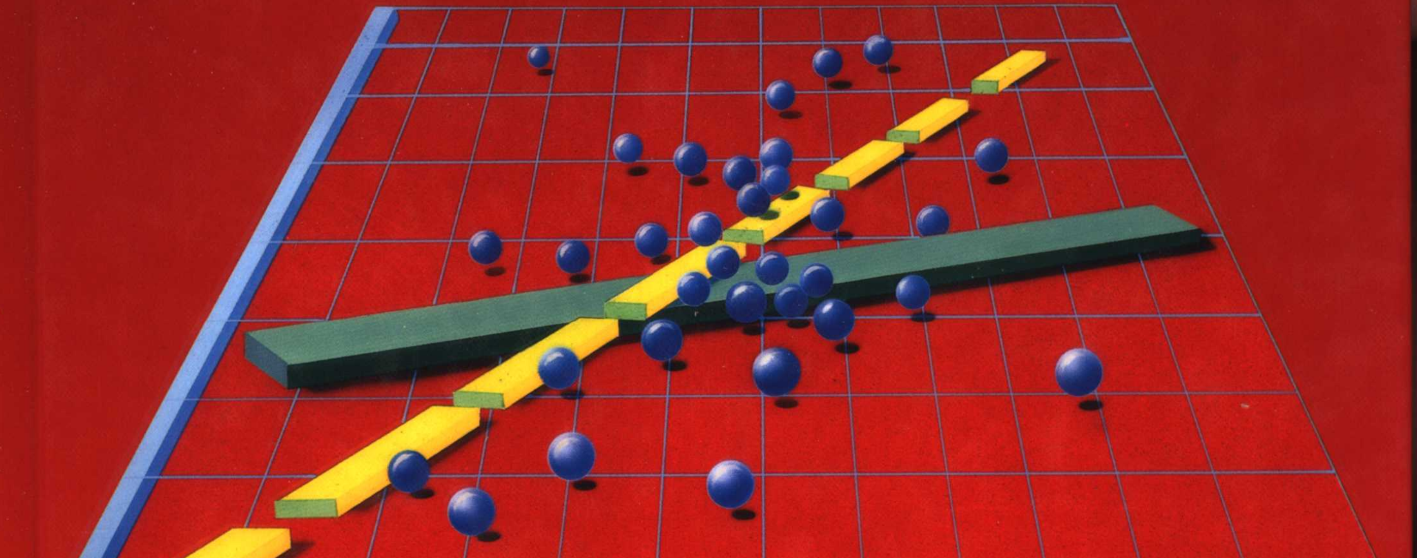
INTRODUCTION to the PRACTICE of STATISTICS

SECOND EDITION



DAVID S. MOORE GEORGE P. McCABE

Introduction to the Practice of Statistics

• • •

SECOND EDITION

David S. Moore
George P. McCabe
PURDUE UNIVERSITY

W. H. FREEMAN AND COMPANY
NEW YORK

Cover illustration by Salem Krieger

Library of Congress Cataloging in Publication Data

Moore, David S.
Introduction to the practice of statistics / David S. Moore, George P. McCabe—2nd ed.
p. cm.
Includes bibliographical references.
ISBN 0-7167-2250-X
1. Mathematical statistics. I. McCabe, George P. II. Title.
QA276. 12.M65 1993
519.5—dc20 92-22880
CIP

Printed in the United States of America

Ninth printing 1996, MB

Preface

We are pleased that so many students and teachers have been receptive to a text that attempts to focus on data and on statistical reasoning. Our second edition has nonetheless been revised more thoroughly than many new editions, though without altering the substance and style of the text. In this preface we will first describe our overall philosophy and then discuss the changes that appear in the new edition.

Introduction to the Practice of Statistics is an elementary but serious introduction to modern statistics for general college audiences. This book is elementary in the level of mathematics required and in the statistical procedures presented. It is serious because our aim is to help readers think about data and use statistical methods with understanding. Students need only a working knowledge of algebra; that is, they must be able to read and use formulas without a detailed explanation of each step.

Statistics is interesting and useful because it is a means of using data to gain insight into real problems. As the continuing revolution in computing relieves the burden of calculating and graphing, an emphasis on statistical concepts and on insight from data becomes both more important and more practical. We have seen many statistical mistakes, but few that involved simply getting a calculation wrong. We therefore ask students to think about the background of the data, the design of the study that produced the data, the possible effect of outlying observations on their conclusions, and the reasoning that lies behind standard methods of inference. Users of statistics who form these habits from the beginning are well prepared to learn and use more advanced methods.

The title of the book expresses our intent to introduce readers to statistics as it is used in practice. Statistics in practice is concerned with gaining understanding from data; it is focused on problem solving rather than on methods that may be useful in specific settings. A text cannot fully imitate practice, because it must teach specific methods in a logical order and must use data that are not the reader's own. Nonetheless, our interest and experience in applying statistics have influenced the nature of this book in several ways.

Focus on Statistical Reasoning and Data We share the emerging consensus among statisticians that statistical education should focus on data and on statistical reasoning rather than on either the presentation of as many methods as possible or the mathematical theory of inference. Understanding statistical reasoning should be the most important objective of any reader. We attempt to present statistics as a coherent discipline with important modes of thought that recur in many specific settings. The first two chapters, for example, concern the art of organizing and exploring data. We hope that readers will draw from these chapters some strategies for understanding data and not just a kit of useful tools. Later chapters similarly attempt to make clear the fundamental modes of thought in designs for data production and in the probabilistic reasoning of formal inference.

An emphasis on data accompanies the emphasis on conceptual understanding. Not all numbers are data. The number 10.3 alone is meaningless; it acquires meaning when we are told that it is the birth weight of a child in pounds or the percent of teenagers who are unemployed. Because context makes numbers meaningful, our examples and exercises are presented in the context of real-world problems. We often comment on issues of statistical practice raised by particular examples. The data presented in examples and in exercises are mostly real and, even when not, are based on real problems. Many of these problems are drawn from those brought to the statistical consulting service at Purdue University by students and faculty from many disciplines. We hope that the presence of background information, even in exercises intended for routine drill, will encourage readers to consider the meaning of their calculations as well as the calculations themselves.

Computers and Statistical Calculations Statistical calculations are in practice performed by software packages on a computer. Many instructors make use of statistical software in the first course. We use either Minitab or SAS in our own teaching, the choice depending on the needs of the students; many other packages are equally satisfactory. We have included some topics that reflect the dominance of software in practice, such as the interpretation of normal quantile plots and an explanation of the two-sample *t* statistic with approximate degrees of freedom. Some exercises require graphs and calculations that are tedious without a computer, and others present computer output as a basis for further work. But we have been careful to make the book easily usable by students without access to computing facilities. A scientific calculator with statistical functions is helpful, but any calculator that will give the mean and standard deviation from keyed-in data is adequate.

Judgment in Statistics Statistics in practice requires judgment. It is easy to list the mathematical assumptions that justify use of a particular procedure, but not always easy to decide when the procedure can be used in practice. Because judgment is developed by experience, an introductory course should present firm guidelines and not make unreasonable demands on the judgment of students. We have given guidelines—for example on using the t procedures for means and avoiding the F procedures for variances—that we follow ourselves. Although we have cited the literature on which our recommendations rest, we recognize that other statisticians may follow different guidelines. We prefer to give imperfect rules rather than avoid the issue of when procedures are in fact useful in analyzing real problems. Similarly, some exercises require the use of judgment in addition to right-or-wrong calculations and conclusions. Both students and teachers should recognize that not every part of every exercise has a single correct answer. We enjoy and encourage classroom discussion of questions of interpretation. But, as is appropriate in a first course, most exercises are straightforward even at the cost of some oversimplification.

Teaching Experiences In writing this book, we drew on our experience with two groups of students. In teaching general undergraduates from a variety of disciplines, we cover the first nine chapters omitting all but a very few of the starred sections. This results in a modern introduction to basic statistics that is quite standard in content, though with some reordering of material and with a strong emphasis on data and reasoning. The Annenberg/Corporation for Public Broadcasting telecourse *Against All Odds: Inside Statistics* uses the text in this way. Our second group of students are advanced undergraduates and beginning graduate students in such fields as education, social sciences, and health sciences. These students are more scientifically sophisticated than general undergraduates, though not more mathematically advanced. In teaching this second group, we cover nearly the entire text except for Sections 4.4 and 10.2. The starred sections therefore help distinguish between two somewhat different courses that can be taught from this book. Of course, any starred section can be omitted without impeding the reading of later chapters.

The Second Edition

The focus of the changes in this Second Edition is to make the book more accessible to students and more teachable for instructors. In undertaking the revision, we have been guided by our own teaching experience and by comments from many students and teachers. There are numerous small improvements in the writing, simpler notation where possible, and several interesting new real data sets. More important are several major changes:

- Readers will find a substantial reorganization early in the book. Chapter 2 of the first edition, which dealt with data on a variable changing over time, has vanished. Because most teachers were treating these topics only lightly, they are now integrated with other material, resulting in a clearer path for students. The most important improvement is a single unified treatment of fitting lines to data early in the new Chapter 2. Plots against time now appear in Chapter 1 as part of data analysis for a single variable; exponential growth is an optional topic in Chapter 2, as an example of transforming data to obtain linearity; and statistical control is discussed in Chapter 5 along with $\bar{x}$ control charts. As a result, Chapters 1 and 2 present a more straightforward introduction to the tools and tactics of data analysis.
- We have modified the presentation of probability in Chapter 4 at the two points at which students encountered the most difficulty: The introduction of random variables is now delayed until the second section to allow better digestion of probability basics, and the exposition of the rules for means and variances of random variables has been slowed for easier comprehension. In addition, Bayes's rule appears in an optional section for those who need it. In Chapter 5, we make a clearer distinction between sample counts and sample proportions in discussing sampling distributions.
- We have completely rewritten the presentations of the chi-square test for two-way tables in Section 8.3 and of inference in the simple linear regression setting in Section 9.1. Students will now find a more direct path to the main results, with refinements placed in later optional subsections. Instructors and students will enjoy new data sets such as the lean of the Tower of Pisa over time and Florida manatees killed by power boats against boat registrations.
- We tried to help teachers who (as we recommend) use software. A data disk available in several formats contains all large and many moderate-size data sets. Exercises that are feasible only if students use software are separated out as "computer exercises" at the end of each chapter. Many of these exercises are new. A completely new Minitab Guide is available.

The practice of even elementary statistics is more intellectually challenging than is suggested by the limited mathematics that is required.

We present real or realistic data and at least hint at the complexities of statistics in practice; we regularly ask students to draw a substantial conclusion rather than simply report the result of a calculation, and we try to explain the reasoning (but not the mathematics) behind recipes. Most students and teachers report that the added effort is rewarded by added understanding. In this new edition we have done our best to remove unnecessary challenges. There is, to paraphrase Euclid's advice to King Ptolemy of Egypt long ago, no royal road to learning. We hope that this road to learning is now a bit smoother.

Supplements

Several supplements are available free to adopters of the Second Edition of *Introduction to the Practice of Statistics:*

- The *Instructor's Guide* contains the following: overviews and teaching suggestions for each chapter; suggested excerpts from the telecourse *Against All Odds* and tips for using these videos in a nontelecourse environment; sample examinations with solutions; and worked-out solutions to all exercises.
- A *data disk,* available in several formats, enables instructors to easily enter all our larger data sets and many moderate-size data sets into their local computer system.
- A *test bank,* available both in computerized and printed versions, for generating quizzes and exams.
- *Transparency masters* of many of the text's figures and tables.

In addition, students may purchase a *Minitab Guide,* written by Betsy Greenberg and Mark Serva of the University of Texas, Austin, that accompanies the text and gives detailed instructions in using the Minitab statistical software.

We have not developed software specifically to accompany this text because we strongly encourage the use of professionally written software, which is both more capable and more transportable to other settings. W. H. Freeman and Company offers a special discount on the student version of *Data Desk, Learning Data Analysis with Data Desk,* by Paul F. Velleman, to users of this text.

The Telecourse *Against All Odds*

Against All Odds: Inside Statistics is a 26-program telecourse on statistics and its applications sponsored by the Annenberg/Corporation for Public Broadcasting Project. This telecourse, developed by David S.

Moore, offers a visual introduction to modern statistics that closely parallels *Introduction to the Practice of Statistics.* In addition to broadcast of the series by public television stations, the programs are inexpensively available on videotape to individuals and institutions. Instructors who do not follow the telecourse format may find excerpts from this unique video series valuable as supplements to classroom instruction. More information and demonstration videotapes can be obtained by telephoning 1-800-LEARNER.

Two supplements are available that link the television programs to this text.

- The *Telecourse Faculty Guide* contains an overview of the material in each video program and provides detailed advice on implementing a telecourse.

- The *Telecourse Study Guide,* available for sale to students, guides students in their study of the course material. Each unit in the study guide corresponds to a video program, and provides an overview of the content, learning objectives, assigned reading and exercises from this text, and self-test questions with fully worked solutions. The Study Guide also contains sample examinations with worked solutions. Instructors who are not using the video material may find the large number of additional exercises in the Study Guide useful.

Acknowledgments

We are grateful to colleagues who commented on the manuscript and to students who studied from it. In particular, we would like to thank the following colleagues who offered specific comments on the second edition:

Richard Berk, UCLA
Trudy Ann Cameron, UCLA
Philip B. Ender, UCLA
Eugene A. Enneking, Portland State University
James Finch, University of San Francisco
Evan Fisher, Lafayette College
Chris Franklin, University of Georgia
Chris Freiling, UCLA
Gavin G. Gregory, University of Texas at El Paso
Pete Herron, Suffolk County Community College
Piet de Jong, University of British Columbia
Ita G. G. Kreft, UCLA

N. T. Longford, Educational Testing Service
J. David Mason, University of Utah
David Meyers, Bakersfield College
Cheryl McKeeman, Vancouver Community College
Mary R. Parker, Austin Community College
Clifford O. Pope, Atlantic Union College
Thomas P. Ryan, Texas A & M University
W. Robert Stephenson, Iowa State University
Frances P. Stewart, University of South Carolina
George Sturm, Grand Valley State University
Nola D. Tracy, McNeese State University
Roy H. Williams, Memphis State University
Bob Woodle, Jamestown College
Mary Rising, University of Louisville

We would also like to acknowledge the ongoing influence of reviewers of the first edition:

Rudolf J. Freund, Texas A & M University
Bernard Harris, University of Wisconsin, Madison
David Hildebrand, University of Pennsylvania
Lambert Koopmans, University of New Mexico
David R. Lund, University of Wisconsin, Eau Claire
Donald E. Ramirez, University of Virginia
Charles W. Sinclair, Portland State University
Jessica Utts, University of California, Davis
Robert Wardrop, University of Wisconsin, Madison
Anne Watkins, Los Angeles Pierce College
Jeffrey Witmer, Oberlin College
Mary Sue Younger, University of Tennessee, Knoxville
Douglas Zahn, Florida State University

Most of all, we are grateful to the many people in varied disciplines and occupations with whom we have worked to gain understanding from data. They provided both material for this book and the experience that enabled us to write it. Perhaps even more important, working with people from many fields has constantly reminded us of the importance of statistical fundamentals in an age when computer routines and professional advice quickly handle detailed questions. If the publisher would allow it, we would call this book "What you should know before you talk to a statistician." We hope that users and potential users of statistics will find it helpful.

David S. Moore

George P. McCabe

Introduction: What Is Statistics?

Statistics is the science of collecting, organizing, and interpreting numerical facts, which we call *data*. We are bombarded by data in our everyday life. Most of us associate "statistics" with the bits of data that appear in news reports: baseball batting averages, imported car sales, the latest poll of the president's popularity, and the average high temperature for today's date. Advertisements often claim that data show the superiority of the advertiser's product. All sides in public debates about economics, education, and social policy argue from data. Yet the usefulness of statistics goes far beyond these everyday examples.

The study and collection of data are important in the work of many professions, so that training in the science of statistics is valuable preparation for a variety of careers. Each month, for example, government statistical offices release the latest numerical information on unemployment and inflation. Economists and financial advisors as well as policy makers in government and business study these data in order to make informed decisions. Doctors must understand the origin and trustworthiness of the data that appear in medical journals if they are to offer their patients the most effective treatment. Politicians rely on data from polls of public opinion. Business decisions are based on market research data that reveal consumer tastes. Farmers study data from field trials of new crop varieties. Engineers gather data on the quality and reliability of manufactured products. Most areas of academic study make use of numbers, and therefore also make use of the methods of statistics.

We can no more escape data than we can avoid the use of words. Just as words on a page are meaningless to the illiterate or confusing to the partially educated, so data do not interpret themselves but must be read with understanding. Just as a writer can arrange words into convincing arguments or incoherent nonsense, so data can be compelling, misleading, or simply irrelevant. Numerical literacy, the ability to follow and understand numerical arguments, is important for everyone. The ability to express yourself numerically, to be an author rather than just a reader, is a vital skill in many professions and areas of study. The study of statistics is therefore essential to a sound education. We must learn how to read data, critically and with comprehension; we must learn how to produce data that provide clear answers to important questions; and we must learn sound methods for drawing trustworthy conclusions based on data.

Historically, the ideas and methods of statistics developed gradually as society grew interested in collecting and using data for a variety of applications. The earliest origins of statistics lie in the desire of rulers to count the number of inhabitants or measure the value of taxable land in their domains. As the physical sciences developed in the seventeenth and eighteenth centuries, the importance of careful measurements of weights, distances, and other physical quantities grew. Astronomers and surveyors striving for exactness had to deal with variation in their measurements. Many measurements should be better than a single measurement, even though they vary among themselves. How can we best combine many varying observations? Statistical methods that are still important were invented in order to analyze scientific measurements.

By the nineteenth century, the agricultural, life, and behavioral sciences also began to rely on data to answer fundamental questions. How are the heights of parents and children related? Does a new variety of wheat produce higher yields than the old, and under what conditions of rainfall and fertilizer? Can a person's mental ability and behavior be measured just as we measure height and reaction time? Effective methods for dealing with such questions developed slowly and with much debate.[1]

As methods for producing and understanding data grew in number and sophistication, the new discipline of statistics took shape in the twentieth century. Ideas and techniques that originated in the collection of government data, in the study of astronomical or biological measurements, and in the attempt to understand heredity or intelligence came together to form a unified "science of data." That science of data—statistics—is the topic of this text.

The first two chapters deal with statistical methods for organizing and describing data. These chapters progress from simpler to more complex data. Chapter 1 examines data on a single variable, Chapter 2 is devoted to relationships among two or more variables. You will learn both how to examine data produced by others and how to organize and summarize your own data. These summaries will be first graphical, then numerical, then when appropriate in the form of a mathematical model that gives a compact description of the overall pattern of the data. Chapter 3 outlines arrangements (called designs) for producing data that answer specific questions. The principles presented in this chapter will help you to design proper samples and experiments, and to evaluate such investigations in your field of study.

The remaining seven chapters discuss statistical inference—formal methods for drawing conclusions from properly produced data. Statistical inference uses the language of probability to describe how reliable its conclusions are, so some basic facts about probability are needed to understand inference. Probability is the subject of Chapters 4 and 5. Chapter 6, perhaps the most important chapter in the text, introduces the reasoning of statistical inference. We emphasize that effective inference is based on good procedures for producing data (Chapter 3), careful examination of the data (Chapters 1 and 2), and an understanding of the nature of statistical inference as discussed in Chapter 6. Chapters 7 through 10 describe some of the most common specific methods of inference: for drawing conclusions about means and proportions from one and two samples, about relations in categorical data, regression and correlation, and analysis of variance.

The practice of statistics involves the use of many recipes for numerical calculation, some quite simple and some very complex. As you learn how to use these recipes, remember that the goal of statistics is not calculation for its own sake, but gaining understanding from numbers. Many of the calculations can be automated by a calculator or computer, but you must supply the understanding. Chapters 7 to 10 present only a few of the many specific procedures for inference. The more complex procedures are always carried out by computers using specialized software. A thorough grasp of the principles of statistics will enable you to quickly learn more advanced methods as needed. On the other hand, a fancy computer analysis carried out without attention to basic principles will often produce elaborate nonsense. As you read, seek to understand the principles as well as the necessary details of methods and recipes.

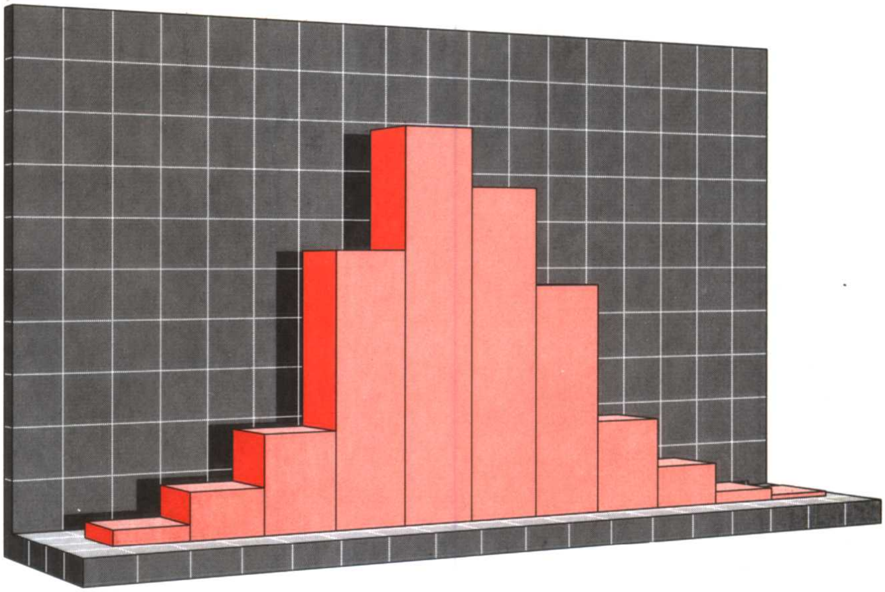
NOTE

[1] The rise of statistics from the physical, life, and behavioral sciences is described in detail by S. M. Stigler, *The History of Statistics: The Measurement of Uncertainty Before 1900,* Harvard-Belknap, Cambridge, Mass. 1986. Much of the information in the brief historical notes appearing throughout the text is drawn from this book.

Introduction to the Practice of Statistics

CHAPTER

1



Looking at Data—Distributions

1.1 DISPLAYING DISTRIBUTIONS
1.2 DESCRIBING DISTRIBUTIONS
1.3 THE NORMAL DISTRIBUTIONS

Contents

Preface xv

Introduction: What Is Statistics? xxiii

CHAPTER 1
Looking at Data: Distributions 1

1.1 Displaying Distributions 3
- Measurement 3
- Variation 5
- Stemplots 6
- Histograms 10
- Looking At Data 14
- Time Plots 16
- Section 1.1 Exercises 20

1.2 Describing Distributions 36
- Measuring Center 36
- Resistant Measures of Spread 40
- The Standard Deviation 46
- Changing the Unit of Measurement 49
- Section 1.2 Exercises 52

1.3 The Normal Distributions 58
- Density Curves 59
- Normal Distributions 63
- Normal Distribution Calculations 67
- Assessing Normality 73
- Section 1.3 Exercises 79

Chapter 1 Exercises 86

Chapter 1 Computer Exercises 90

Starred sections are optional.

CHAPTER 2
Looking at Data: Relationships 94

2.1 Scatterplots 98
Interpreting Scatterplots 98
Smoothing Scatterplots* 104
Categorical Explanatory Variables 107
Section 2.1 Exercises 109

2.2 Least Squares Regression 117
Fitting a Line to Data 118
Least-Squares Regression 121
Residuals 125
Outliers and Influential Observations 131
Section 2.2 Exercises 136

2.3 An Application: Exponential Growth* 147
The Nature of Exponential Growth 147
The Logarithm Transformation 150
Residuals Again 152
Section 2.3 Exercises 156

2.4 Correlation 161
Computing the Correlation 162
Correlation in the Regression Setting 168
Interpreting Correlation and Regression 171
Section 2.4 Exercises 177

2.5 Relations in Categorical Data 183
Analyzing Two-Way Tables 184
Simpson's Paradox 188
Section 2.5 Exercises 191

2.6 The Question of Causation 197
Smoking and Lung Cancer 198
Establishing Causation 202
Section 2.6 Exercises 204

Chapter 2 Exercises 206

Chapter 2 Computer Exercises 212

CHAPTER 3
Producing Data 220

3.1 First Steps 223
The Need for Design 223
Sampling 225
Experiments 225
Section 3.1 Exercises 227