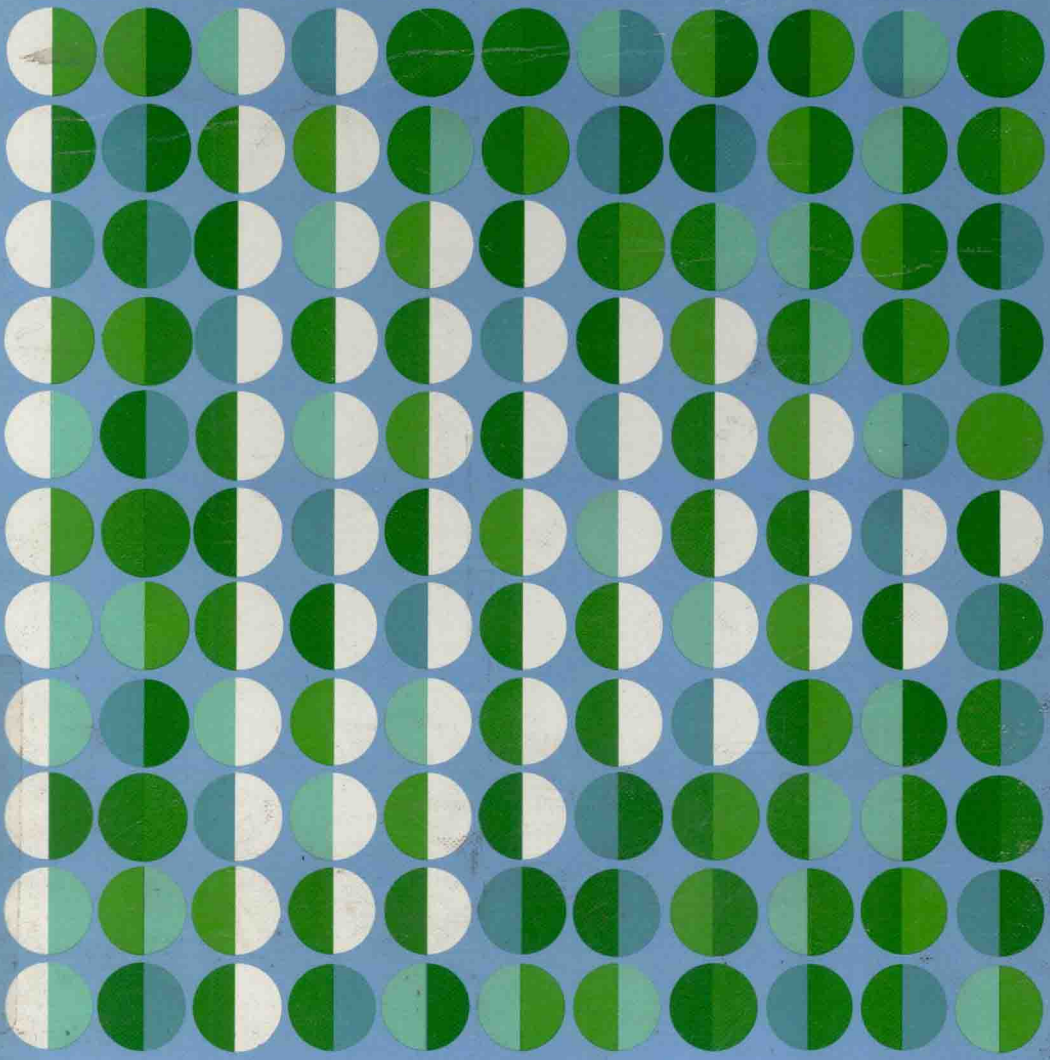
AN INTRODUCTION TO CONTEMPORARY STATISTICS

Lambert H. Koopmans



LAMBERT H. KOOPMANS
University of New Mexico

AN INTRODUCTION TO CONTEMPORARY STATISTICS

DUXBURY PRESS
Boston, Massachusetts

An Introduction to Contemporary Statistics was prepared for publication by the following people:

Production Editor: Barbara Gracia
Copy Editor: Carol Beal
Interior Designer: Catherine Dorin
Cover Designer: Catherine Dorin

Duxbury Press
A Division of Wadsworth, Inc.

Library of Congress Cataloging in Publication Data

Koopmans, Lambert Herman, 1930-
An introduction to contemporary statistics.

Includes bibliographical references and index.
1. Mathematical statistics. I. Title.
QA276.A1K66 519.5 80-19725
ISBN 0-87872-292-0

Printed in the United States of America
1 2 3 4 5 6 7 8 9—85 84 83 82 81

Preface

The course that motivated the writing of this text was designed in the early 1970s to serve as a beginning course for majors in several departments at the University of New Mexico—psychology, sociology, economics, political science, and business. Subsequently the course also attracted majors in such subjects as nursing, biology, anthropology, and geology. The faculties in these departments wanted their students to be able to carry out and interpret the basic methods of inference on data they would be exposed to in later courses and their subsequent professional lives, as well as to be able to read and understand the applications of statistics in the literature of their fields. Thus from the beginning the main thrust of the course was statistical methodology. Since the classical methods remain predominant in most of these fields, the main emphasis was, and still is, on these methods. They form the core of this book.

I began teaching the course in 1973. In each succeeding semester I tried to add some feeling for the currency and importance of statistical application by bringing in data from my own research and consulting and from that of friends and colleagues. (Often the data found their way into the classroom before the consulting job was completed.) Much of that data and the associated problems are given in this book in examples and exercises. This fact, rather than any connection with the local chamber of commerce,

will account for the frequent references to people and places in and around Albuquerque, New Mexico.

It is (becoming) common knowledge among practicing statisticians that the classical statistical methods can fail, sometimes rather badly, when indiscriminately applied to real data. Since real data are what students will actually face when they embark on their own research, and because curative measures are now available, I feel that there is no legitimate defense for not discussing these difficulties and providing strategies for overcoming them. This objective is achieved in this text by providing robust and nonparametric alternatives to all the classical procedures.

However, the available corrective measures are of little value unless we can identify when they are needed. Moreover, inference is just one phase of an adequate examination of data. Much of what we are looking for in a data set—and often new things not being specifically looked for—are found in a preliminary exploration of the data. New tools, fashioned by J. W. Tukey, are introduced in the early chapters of the book, and they make the exploration of data both easy and informative. These tools have an added diagnostic function that makes it possible to see in advance when a classical inference method will be in trouble and will need modification.

I became acquainted with exploratory methods in 1975 during a visiting professorship at Princeton University, which allowed me to study Tukey's ideas in some depth. Shortly thereafter I began to use these methods extensively in my research and consulting. Their usefulness became immediately apparent. The methods were gradually introduced into my elementary statistics course beginning in 1976 and have, by a process of evolution, almost completely replaced the more traditional methods of descriptive statistics. Students are first introduced to the investigation of statistical problems as an exploratory endeavor in which the data can provide new lines of insight and understanding rather than the more traditional description and summary. This treatment is then followed in later chapters by the appropriate classical inference method. Thus although the usual goals of a beginning course in statistics are retained, they are approached through the use of exploratory tools. Except for this difference, the text follows the lines of a traditional course in elementary statistics and can be readily adapted for use in such a course by an appropriate selection of topics.

Computing difficulties arose with the introduction of real data sets in class. At first I attempted to have students use the university computer. This ploy was soon abandoned because of the administrative and pedagogical difficulties associated with large class sizes and the necessity of teaching computing techniques along with statistics in a course already badly stretched for time with required material. I also must confess to a (difficult to articulate and justify) bias toward hand analysis of data. I feel that everyone learning statistics should grub around with his or her data firsthand in at least one course. A current partial justification for this feeling is that many

of the exploratory methods presented in this book are still relatively rare in standard computer packages.* However, for myself, I suspect that the bias will persist even when this is no longer true.

One reason for this suspicion is that remarkably powerful statistical computers now exist that are inexpensive and portable enough to fit into a pocket—hand calculators. Consequently, we need not be tied to the availability of a computer terminal. Statistical analyses can be done where and when one likes. I began to recommend the purchase of calculators for my classes in the early 1970s. At first, cost was a drawback. However, while the ethics of requiring (or even recommending) that students buy an expensive calculator was being debated, the problem evaporated. It is now possible to purchase a calculator with much more than enough computing power to do everything we will need for around $20. The calculator makes it unnecessary to learn to use square root tables, and it equalizes everyone's arithmetic ability, making it possible to concentrate on learning statistics.

I currently *require* the students in my classes to have hand calculators. In fact, I go so far as to recommend particular features. A description of the recommended calculators is given in succeeding paragraphs. Keystrokes for efficiently carrying out the necessary computations on these calculators are given in the text.

Clearly there are limits to the usefulness of hand analysis in statistics, even with the most elaborate hand calculators currently available. Beyond a certain point the volume of data and/or calculations become so great that nothing but the use of a modern computer is feasible. In fact, the computer has made possible some statistical analyses that were only wishfully thought of earlier in the history of the subject. Moreover, many applications have been and are currently being invented specifically for them. These ideas are discussed at various points in the text. However, as this text will demonstrate, a large core of the most important methods in statistics can be conveniently applied by hand to data sets of sizes commonly seen in practice. Consequently, both computers and hand calculators should be familiar tools available in those situations for which each is most useful.

The usefulness of calculators will be amply demonstrated in the exercises. Solutions to the exercises given in the body of the text, as well as to the odd-numbered exercises to be found at the end of each chapter, are provided at the end of the text. Solutions to the even-numbered exercises are given in a separate instructor's manual.

*A step toward the solution of this problem has been taken with the publication of *Applications, Basics, and Computing of Exploratory Data Analysis* by Paul F. Velleman and David C. Hoaglin, Duxbury Press, Boston, Mass., 1981. The computer routines from this book are used in the current releases (80.1 and 80.2) of *Minitab,* developed by Thomas A. Ryan Jr., Pennsylvania State University.

Features of the Recommended Hand Calculators

The following keys and features are currently available on the Sharp EL 506 and the Texas Instrument TI-35, both of which presently cost less than $25. Both are algebraic notation scientific calculators with liquid crystal displays (LCD). As much as I admire reverse Polish notation (one of my first calculators was a Hewlitt-Packard HP-55), the less expensive calculators with the required features use algebraic notation. The long battery life of an LCD calculator makes it truly portable and prevents embarrassing battery fatality during exams.

The standard function keys that constitute the bare essentials for statistical calculations are $+$, $-$, $\times$, $\div$, x^2, $\sqrt{\ }$ (square root), 1/x (or x^{-1}, the reciprocal), and $+/-$ (change sign). Other features used in the text are parentheses ( ), the exponent function EXP, a memory with keys $x \rightarrow$ M (transfer display to memory), RM (recall memory), M+ (add display to memory). A statistical mode makes possible the convenient computation of the statistical functions $\bar{x}$ (sample mean) and s (sample standard deviation). The El 506 also has the useful feature of displaying n (sample size), Σx (sum of entries), and Σx^2 (sum of squares of entries) in the statistical mode.

These calculators also possess keys to compute the following functions and their inverses: ln (log to base e), log (log to base 10), sin, cos, tan, and y^x. We will use a couple of these features in the text. Other available features we will not use are $\sqrt[3]{\ }$ (cube root), π (display 3.14159 . . . in the register), and $n!$ (factorial).

Acknowledgments

The roots of indebtedness for a book such as this are deep and, for me, extend back to my teachers at the University of California, Berkeley. Their influence is ever present in my writing. In particular, I had the privilege of learning about statistical inference from one of its creators, Jerzy Neyman. His ideas are basic to the second part of the text.

I am grateful for a long-time acquaintanceship with John Tukey in which he has provided me with access to his work and with important guidance and inspiration. His early publications on time series analysis were instrumental in my writing a book on that subject some years ago. Now his ideas on exploratory data analysis and robust statistical inference form the core of part I and much of part III of this book. I am indebted to Geoffrey Watson for the invitation to visit Princeton, where I was able to learn about these things first hand.

I was fortunate to attend a series of lectures by Oscar Kempthorne at New Mexico State University during the time I was writing part II of the text. His views on statistical inference, expressed through discussion and in the later reading of papers he kindly made available to me, came at just the right time to help me clarify my own thoughts on this important topic. His comments, and those of Scott Urquhart of New Mexico State, provided much important food for thought when it was needed.

During the writing of the entire manuscript I was fortunate to have Jon Receconi, a member of one of my recent elementary statistics classes, provide me with the students' viewpoint of the material as it was being produced. His kind but firm admonition of "it isn't as clear as it was in class" was the source of extensive rewritings of many sections and in some cases, whole chapters of the text. Jon has since decided to enter the ministry—a decision he assures me is unrelated to his work on this book.

Exercise solutions were provided by my students Kathy Hsi and Donna Jacobi. Donna also checked, and frequently improved, the numerical accuracy of the examples given in the text.

I have been aided by many local friends and colleagues who have supplied me with data sets, research material, comments, and encouragement. Ron Iman of Sandia Corporation kindly provided me with reprints and preprints of his work, which play an important role in part III of the text. I have particularly benefited from discussions with Francis Wall, Dick Prairie, and Ron Schrader.

In preparing final manuscript drafts, comments and suggestions from reviewers were particularly helpful. I would like to acknowledge the following reviewers: Thomas A. Aiuppa, University of Wisconsin, La Crosse; Paul Alper, College of St. Thomas; Jeffrey B. Birch, Virginia Polytechnic Institute and State University; Thomas A. Louis, Harvard School of Public Health; David R. Lund, University of Wisconsin, Eau Claire; John M. Rogers, California Poly, San Luis Obispo; Stephen B. Vardeman, Purdue University; and Paul F. Velleman, Cornell University.

Several people connected with Duxbury Press deserve commendation for their work on this book. Carol Beal did a magnificent job of copy editing. Barbara Gracia efficiently managed the many details of production, and editor, Pat Fitzgerald, supplied several key ideas. The impressive physical design of the book was accomplished by Catherine Dorin. Special thanks are due to managing editor Jerry Lyons and local representative Nancy Tandberg who have been associated with this project for more than two years now. Their advice and support have been much appreciated.

Finally, I must confess that I am just one member of a two-person writing team—and not always the most hard working one at that. The other member, my wife Sharon, has typed and retyped the manuscript from barely legible handwritten originals and has carried out many of the other tasks required to produce a book, including providing the author with encouragement when the writing went poorly, without ever losing her good humor and confidence in the project. To her, this book is lovingly dedicated.

LHK

Contents