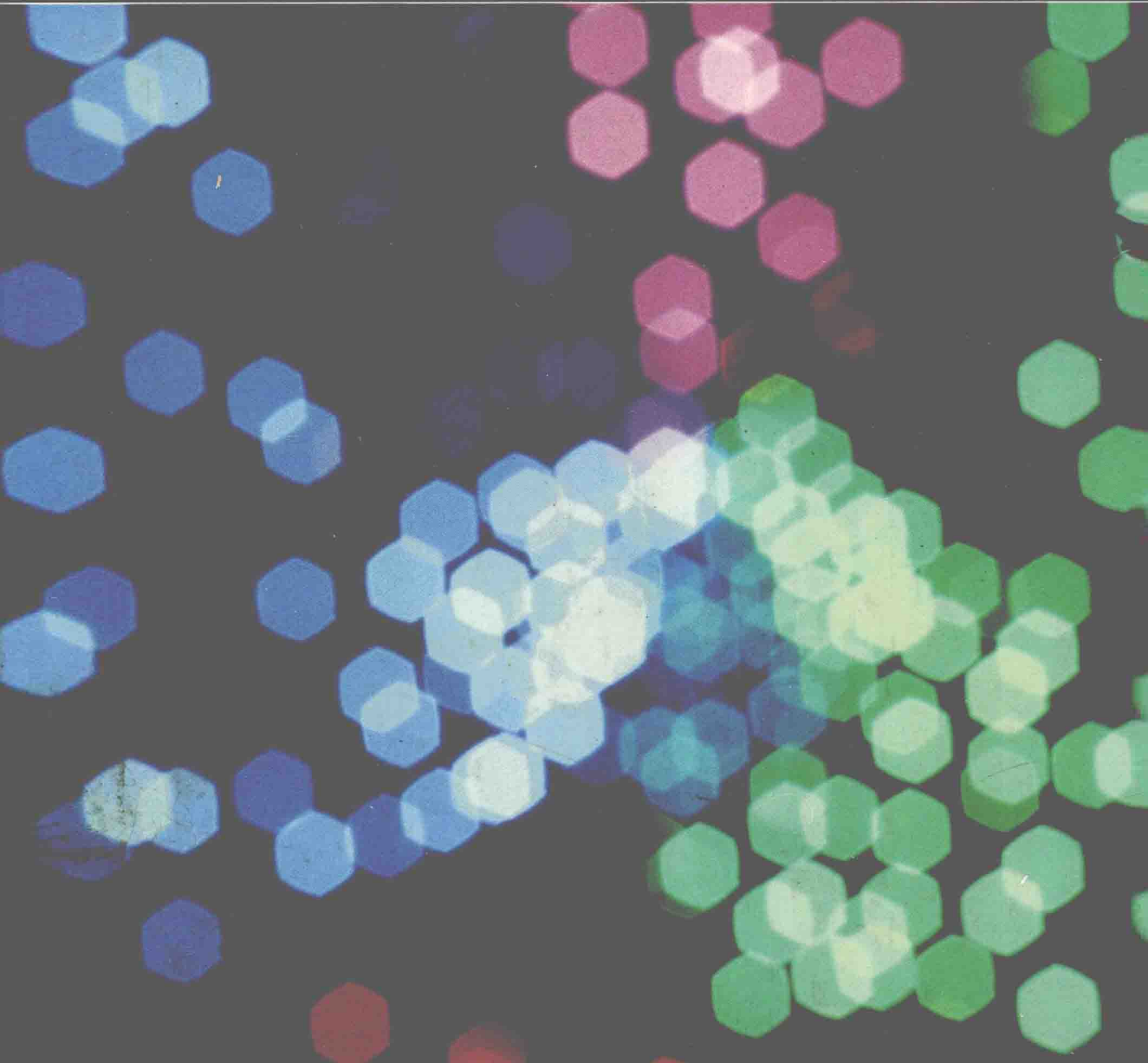
AN INTRODUCTION TO

STATISTICS

WITH DATA ANALYSIS

SHELLEY RASMUSSEN



FREE COPY FREE COPY FREE COPY

AN INTRODUCTION TO STATISTICS WITH DATA ANALYSIS

Shelley Rasmussen

Brooks/Cole Publishing Company
Pacific Grove, California

Brooks/Cole Publishing Company
A Division of Wadsworth, Inc.

© 1992 by Wadsworth, Inc., Belmont, California 94002. All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transcribed, in any form or by any means—electronic, mechanical, photocopying, recording, or otherwise—without the prior written permission of the publisher, Brooks/Cole Publishing Company, Pacific Grove, California 93950, a division of Wadsworth, Inc.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Rasmussen, Shelley, [date]
An introduction to statistics with data analysis / by Shelley Rasmussen.
p. cm.
Includes bibliographical references and index.
ISBN 0-534-13578-1 :
1. Mathematical statistics. I. Title.
QA276.R375 1991
519.5—dc20 91-9035
CIP

International Student Edition ISBN: 0-534-98585-8

Sponsoring Editor: *Michael Sugarman*
Marketing Representative: *John Moroney*
Editorial Assistant: *Lainie Giuliano*
Production Editor: *Ben Greensfelder*
Manuscript Editor: *Susan Reiland*
Permissions Editor: *Carline Haga*
Interior Design: *Vernon T. Boes*
Cover Design: *Michael A. Rogondino*
Cover Photo: *Lee Hocker*
Art Coordinator: *Cloyce Wall*
Typesetting: *G&S Typesetters, Inc.*
Cover Printing: *Lehigh Press Lithographers/Autoscreen*
Printing and Binding: *R. R. Donnelley & Sons Company*

MINITAB is a registered trademark of Minitab.

STATA is a registered trademark of Computing Resource Center.

About the Author

Shelley Rasmussen received a Ph.D. in Statistics from the University of Michigan. She has taught statistics at the Massachusetts Institute of Technology and universities within the state systems of Texas, New Hampshire, and Massachusetts. As a practicing statistician, she has worked in the pharmaceutical industry and for a cancer research center. She currently teaches and consults in statistics, quality control, and experimental design for industries involved in engineering, high technology, and new product development.

PREFACE

This book is intended for a one- or two-semester introduction to statistics. The discussion is not calculus-based; the only prerequisite is high school algebra.

The emphasis is on the art of statistical thinking. I believe that a course emphasizing statistical thinking about applied problems ought to be anyone's introduction to statistics, no matter what major or year in college. Everyone should understand the usefulness of statistics in addressing real-world problems. Such an understanding would enrich the lives of all students and motivate some to further study in the theory and application of statistics.

Almost all of the examples and exercises in this book are based on real data sets. In a few cases I felt forced to invent a data set to illustrate an idea, because I did not have a real example at hand. Even then I based the example on a realistic application. As a student and a teacher, I have always appreciated real examples in references. I believe that students will be more motivated to study statistics if its usefulness is immediately apparent. This will be most obvious in a book if examples and exercises illustrate the use of statistics in real investigations.

Data analysis is introduced at the beginning of the book, in Part I, and used throughout. Data analysis involves the use of simple graphical and tabular techniques to gain an understanding of the information in a data set. Regrettably, techniques of data analysis are not familiar to many college graduates, not to mention high school graduates. At a recent multidisciplinary workshop for a select group of exceptional high school teachers (funded by the National Science Foundation and run by the Tsongas Industrial History Center in Lowell, Massachusetts), one social studies teacher did not understand why we would ever want to graph data and several others said they always skipped graphs in textbooks. My response was that they were missing the opportunity to help their students to understand the many graphical presentations of data, some good and some bad, that appear daily in the media.

In data analysis and formal statistical analysis, the more carefully a data set is collected, the more useful information can be derived from it. When we use the ideas of experimental design, we plan a study in order to address the

此为试读,需要完整PDF请访问: www.ertongbook.com

questions of interest as efficiently as possible. A well-designed study often needs very little formal statistical analysis. A poorly designed study may yield little useful information no matter how much we massage the data. The importance of data collection and experimental design is emphasized throughout the book.

In formal statistical analysis, we use a sample of data to make inferences about a larger population. These inferences take the form of probability statements about the population, based on what we see in the sample. (We have to make certain assumptions about the sample in order for these probability statements to make sense; a good experimental design helps to assure the validity of some of these assumptions.) Since probability statements form the basis of formal statistical inference, we have to discuss some probability. I have kept this discussion to a minimum. Part II contains the essential concepts in probability that we need for statistical inference. Two optional sections, Sections 6-4 and 6-6, contain interesting applications of probability that are not used again later. The reader who does not want to cover the median test (Section 11-5) or Fisher's exact test (Section 16-5) can skip the discussion of the hypergeometric probability distributions in Section 7-3.

A number of topics and techniques of formal statistical inference are presented in Part III. Classical analysis that depends on the assumption of Gaussian (normal) data is discussed for each appropriate application. In addition, for many applications I have included one and sometimes two alternatives to the classical analysis. Section 10-4, for instance, discusses nonparametric inferences about a population mean or median, based on ranks; Section 10-5 covers inferences about a population median based on signs; Section 14-4 discusses robust inferences about two or more variances. I think it is important for students to realize that not all data sets follow a Gaussian distribution and that there are straightforward alternatives to the classical analysis for many applications. Readers who want to consider only classical analyses, however, may skip the sections on alternative approaches without loss of continuity.

Many students and friends helped me by providing data sets, reviews, suggestions, and encouragement during the writing of this book. Among them are Paul Catalano, Dennie Clarke-Hundley, Paul Gavelis, Janet LaBonte, Nicole LaVallee, Mary Lundquist, Alex Olsen, Michele Walsh, and Penny Angus Yepez. Miin-Show Chao helped with a number of computer runs. Lee Panas contributed data sets, reviewed chapters, provided useful advice, and solved all the exercises for the solutions manual.

I appreciate the contributions of the many reviewers who patiently read the various versions of the manuscript, each version better than the previous one in large part because of their comments and advice. These reviewers include: Dr. Richard Alo, University of Houston; Professor David Banks, Carnegie-Mellon University; Dr. Lynne Billard, University of Georgia; Dr. Bill Korin, The American University; Professor Robert Lacher, South Dakota State University; Professor Ed Landauer, Clackamas Community College; Ms. Mary Parker, Austin Community College; Professor Robert Schaefer, Miami University; Professor Paul Speckman, University of Missouri; Professor Jeff Spielman, Roanoke College; Professor George Terrell, Virginia Polytechnic Institute; and Dr. Cindy van Es, Cornell University. I am grateful to the developmental editor, John Bergez, whose detailed criticisms of my organization and writing style, though punishing, helped me with revisions that made the book much easier to read. I am also grateful to copy editor Susan Reiland for her careful reading of the manuscript and many helpful suggestions.

My thanks to all the investigators cited in the references, from whose work I benefited. My apologies to anyone who should have been cited and was not. My thanks also to all of the teachers, colleagues, and students from whom I have learned probability and statistics, especially Michael Woodroofe of the University of Michigan.

I am grateful most of all to my family for all of their love and support. My mother, Jackie Guernsey; my sister, Pam Grant; my brothers, John Rasmussen and Bill Guernsey; and my mother-in-law, Mary Olsen, provided much encouragement. My children, James, Emily, and Vin were extremely patient and supportive. My husband, Dick Olsen, deserves the most gratitude, for encouraging me to start this project, for providing advice and inspiration, and for giving me the time (the children were all preschoolers when I started) to do it. Thank you very much!

Shelley Rasmussen

CONTENTS

PART ONE
DATA ANALYSIS

CHAPTER 1

Introduction 1

CHAPTER 2

Studying One Variable at a Time: Lists, Tables, and Plots 27

PART TWO PROBABILITY

PART THREE
STATISTICAL INFERENCE

此为试读,需要完整PDF请访问: www.ertongbook.com

CHAPTER 1

Introduction

IN THIS CHAPTER

- Statistics
- Data analysis
- Case, variable, data value
- Quantitative and qualitative variables
- Unit of measurement
- Missing value

Statistics are numbers. Statisticians use numbers (or statistics) to expand our knowledge of the universe, if only a very small part of the universe. We are all statisticians when we use numbers in this way. This book is about such use of numbers. It is not intended as a comprehensive manual of statistical techniques, but rather as an introduction to the art of statistical thinking.

By a **statistic** we mean either a number—a numerical piece of information or datum—or a number calculated from a set of data values.

When practicing the *art of statistics,* we use numerical information to increase our knowledge in some way. Used in this sense, statistics refers to the branch of mathematics dealing with theory and techniques of collecting, organizing, and interpreting numerical information.

By **statistics** we mean either a collection of numerical information, or the branch of mathematics dealing with theory and techniques of collecting, organizing, and interpreting numerical information.

We may use information from a market analysis to select cities for introducing a new product. Or, we might study racing forms to decide how to place a bet in the next horse race. Perhaps we want to examine individual or team performance in major-league baseball. In each of these cases, we study a collection of information, called a *data set.*

A **data set** is a collection of information.

When we try to make sense of a data set, we are engaging in *data analysis.*

By **data analysis** we mean making sense of a data set.

Baseball is extremely conducive to data analysis, since baseball statistics are readily available by player and by team. A baseball fan might study individual variables such as batting average: What is a typical batting average for a player in the major leagues? What is an exceptionally good (or poor) batting average? The fan might also examine relationships between variables: What is the relationship between team batting average and winning percentage? Is this relationship different for the American League than for the National League?

Data analysis involves studying variables and relationships between variables in a collection of information. Often we want to do more. We may want to use a sample of information to learn about a larger population. For instance, we might want to use a sample of the thousands of parts produced in a day to decide whether too much gold is being electroplated onto components used in personal computer hardware. Or, we may want to conduct a taste test of two products in a sample of consumers to make decisions regarding product preference in a larger group of consumers. We might want to compare a new treatment with a standard treatment in patients with a particular form of cancer. In each of these cases, it is impractical to study the entire population (parts electroplated in a day, consumers in a product market, or cancer patients). Instead, we look at a sample or subset of the population. We use the information from the sample to learn about the population. This is *statistical inference.*

By **statistical inference** we mean drawing conclusions about a population based on a sample from that population.

The **population** is the group or collection of interest to us.

A **sample** is a subset of the population. We use the observations in the sample to learn about the population.

Data analysis can aid in statistical inference. Medical researchers routinely study characteristics of patients with a particular form of cancer. They look for relationships among such variables as age, sex, stage of illness, response to treatment, and survival.

Estimation is a part of statistical inference. Investigators might use average survival time for patients in a sample to estimate average survival time for all patients in the population. They might then calculate a range of reasonable values for this average survival time. Interpreting such a range of reasonable values, called a confidence interval, depends on ideas in probability.

Statistical inference also involves hypothesis testing. In testing hypotheses, we compare two statements about the state of nature, such as:

Average survival with the new treatment is the same as for the standard treatment.

Average survival with the new treatment is longer than for the standard treatment.

Which of these two statements does the sample support? To decide, we use ideas in probability.

Both estimation and hypothesis testing use probability. We make probability statements about the population based on what we see in the sample. For these statements to make sense, the sample must be similar to the population, a *representative sample.* Suppose the cancer patients in a sample all have very advanced disease. Then researchers cannot make inferences about a larger population that includes patients with less advanced disease. This leads to the idea of experimental design. We want to collect a sample, or carry out an experiment, so that statistical inferences make sense.

1-1 An Overview of the Book

Our study of statistics begins with data analysis. Though the techniques are fairly simple, they can provide a lot of insight into a collection of information.

Some data analysis tools are tabular. A table can summarize certain types of information in a data set. For instance, we might use a table, called a frequency table, to display the number of baseball players with 1991 salaries in each of several intervals (say, less than \$500,000, \$500,000 to \$1,000,000, and so on). Other tools of data analysis are graphical. A histogram, sometimes called a bar graph, is a graphical tool for displaying the information in a frequency table. We could use such a graph to display the information on numbers of players per salary range, instead of listing these numbers in a table.

此为试读，需要完整PDF请访问：www.ertongbook.com