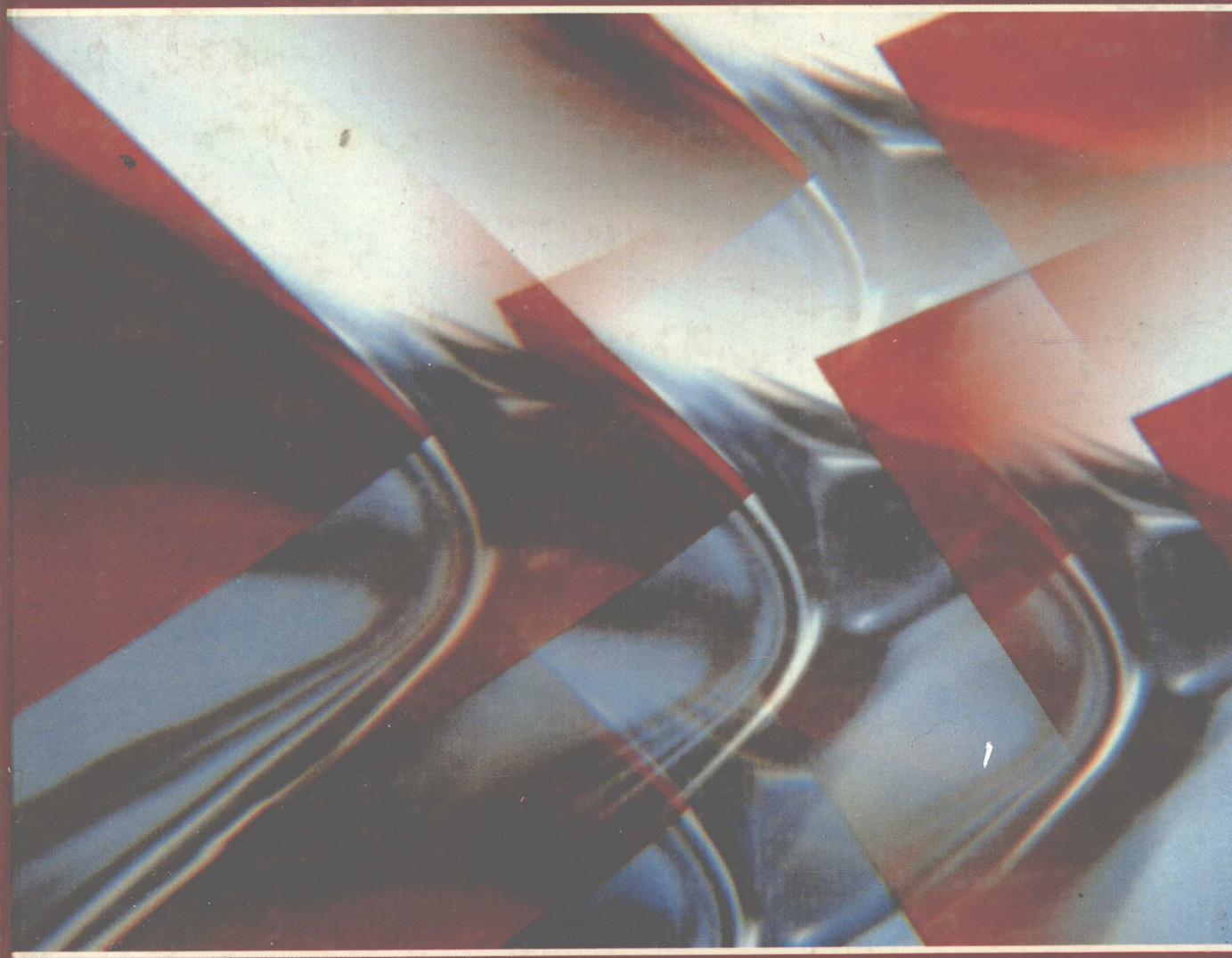
APPLIED STATISTICS

A First Course



Mark L. Berenson / David M. Levine
David Rindskopf

APPLIED STATISTICS
A First Course

Mark L. Berenson
Department of Statistics and Computer Information Systems
Baruch College

David M. Levine
Department of Statistics and Computer Information Systems
Baruch College

David Rindskopf
Departments of Educational Psychology and Psychology
City University of New York, Graduate Center

PRENTICE HALL, *Englewood Cliffs, New Jersey 07632*

Library of Congress Cataloging-in-Publication Data

Berenson, Mark L.
Applied statistics.

Bibliography: p.
Includes index.
1. Statistics. I. Levine, David M.
II. Rindskopf, David. III. Title.
QA276.12.B46 1988 519.5 87-7233
ISBN 0-13-041476-X

Editorial/production supervision: Eleanor Perz
Interior design: Levavi & Levavi
Cover design: Maureen Eide
Manufacturing buyer: Barbara Kittle

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

ISBN 0-13-041476-X 01

PRENTICE-HALL INTERNATIONAL (UK) LIMITED, *London*
PRENTICE-HALL OF AUSTRALIA PTY. LIMITED, *Sydney*
PRENTICE-HALL CANADA INC., *Toronto*
PRENTICE-HALL HISPANOAMERICANA, S.A., *Mexico*
PRENTICE-HALL OF INDIA PRIVATE LIMITED, *New Delhi*
PRENTICE-HALL OF JAPAN, INC., *Tokyo*
SIMON & SCHUSTER ASIA PTE. LTD., *Singapore*
EDITORA PRENTICE-HALL DO BRASIL, LTDA., *Rio de Janeiro*

To my parents, the memory of my father, and to my wife, Rhoda, and daughters, Kathy and Lori.

M.L.B.

To my parents, to my wife, Marilyn, and my daughter, Sharyn.

D.M.L.

To my parents, the memory of my mother, and to my wife, Toni.

D.R.

Preface

When planning an introductory text in statistics that is intended to transcend a wide range of academic disciplines, the authors must decide how the text will differ from those already available and what contribution it will make to the field of study. In developing *Applied Statistics*: *A First Course*, our goal was to present a practical, data-analytic approach to statistical applications so that fundamental methods and concepts could be taught to students in all disciplines in a one-semester, noncalculus-based introductory course. Hence, the distinguishing characteristics of this text are its innovative approaches combined with its internal pedagogical devices.

Innovations

- *Using Case Studies as Chapter Motivators*: Detailed case studies involving extensive analysis based on actual or realistic data are used to motivate the discussions throughout the text.
- *Using a Survey/Database for Integrating Course Topics*: Perhaps the most difficult aspect for a student taking a statistics course is to perceive its continuity (i.e., how one set of topics interrelates with another). The text attempts to overcome this problem by utilizing the results from an alumni association membership survey (Case Study A) in 139 end-of-chapter Database Problems to show the interrelationships between descriptive statistics, cross-tabulations, statistical inference, ANOVA, nonparametrics, and regression and correlation.
- *Using Computer Packages in a Statistics Course*: The rapid expansion in the use of computer packages to assist in data analysis has posed a dilemma for statistics educators in terms of how to best employ software for a statistics course. This text goes beyond merely illustrating computer output. Rather, it devotes an entire chapter (Chapter 15—Using Computers for Statistical Analysis) to demonstrate how computer software is actually used to assist in data analysis. Utilizing questions from the alumni association membership survey (Case Study A) that cover the entire range of topics presented in Chapters 3 through 14 of the text, three widely used commercially available packages (SPSS^X, Minitab, and STAT-

GRAPHICS) are described and the resulting (annotated) computer outputs are interpreted. Hence, this chapter is not intended to make the reader proficient at any particular package; it is, however, intended to familiarize the reader with the use of statistical software and to illustrate that learning the syntax of a command-driven package (SPSS^X or Minitab) or learning to manipulate the displays from a menu-driven package (STATGRAPHICS) are not difficult. Moreover, since the chapter utilizes statistical procedures that had been covered throughout the text, it serves to demonstrate how the various topics are interwoven when one is performing a statistical analysis of a large data set.

- *End-of-Section Problems and End-of-Chapter Problems*: 725 additional problems are presented both at the end of sections and at the end of chapters for purposes of pedagogy. Most problems contain, on average, four or five parts. However, emphasis is not on "number crunching" but on understanding and interpretation. It is essential that students be able to express what they have learned. *Literacy* is enhanced by numerous thought-provoking questions involving discussion, letter-writing, memos, and reports. Problems that are either especially thought-provoking or have no "exact solution" are flagged with a "✎" symbol.

Internal Pedagogical Devices

- *Writing Style*: The material is written in a conversational, narrative type of style in order to emphasize both the concepts and methods of applied statistics. The basic philosophy is to "write for the student, not for the professor." Each particular concept is thoroughly developed with detailed examples. Moreover, to reduce student anxiety in dealing with end-of-section Problems or Supplementary Problems, humorous names are interjected as the situation merits.
- *Realistic Applications*: To motivate student interest, the material in the text is developed by referring to realistic type applications, to applications with actual data, and to applications selected from the alumni association membership survey (Case Study A).
- *Pedagogical Tools*: Numerous important pedagogical tools (boxed formulas, boxed examples and solutions, drawings, tabular interpretations, annotated computer output) have been utilized throughout the text to enhance the student's ability to comprehend the concepts and methods of applied statistics. Use of color is given for emphasis.
- *Reading and Interpreting Statistical Tables*: Each of the statistical tables shown in Appendix C is examined in depth when it is initially presented. Detailed explanations and illustrations are provided to enable the student to read and properly interpret the particular tables.
- *Highlighted Examples with Solutions*: Throughout the text several examples are provided with solutions "boxed off" in color for emphasis. This pedagogical feature permits the student to quickly review a particular topic.

- *End-of-Text Material*: To assist the student, the following end-of-text material is included:

a. Appendix A lists the rules for arithmetic and algebraic operations and discusses summation notation in depth.
b. Appendix B provides a list of statistical symbols and notation.
c. Appendix C provides an extensive set of statistical tables.
d. Answers to Selected Problems—To provide students with immediate feedback, answers to specific problems (designated with a ● symbol) selected from the end of sections and the end of chapters throughout the text are given.
e. Index—a detailed index consisting of more than 450 entries is provided along with extensive cross-referencing.

It is our hope and anticipation that the unique approaches taken in this text will make the study of introductory statistics more meaningful, rewarding, and comprehensible for all readers.

We are extremely grateful to the many organizations and companies that generously allowed us to use their actual data for developing problems and examples throughout our text. In particular, we would like to cite Time Inc. (publisher of *Fortune*), The American Association of University Professors (publisher of *Academe*), and CBS Inc. (publisher of *Road & Track*). Moreover, we would like to thank the Biometrika Trustees, American Cyanamid Company, The Rand Corporation, and the Institute of Mathematical Statistics for their kind permission to publish various tables in Appendix C. Furthermore, we would like to acknowledge SPSS Inc., Minitab Inc., and STSC Inc. for their permission to describe syntax and present computer output in this text.

We wish to express our thanks to Dennis Hogan, Carol Sobel, Kate Moore, and Eleanor Perz of the editorial staff at Prentice Hall for their continued encouragement. We also wish to thank Professors Robert F. Brown, UCLA; Larry J. Ringer, Texas A & M University; Thomas A. O'Connor, University of Louisville; and Phillip McGill, Illinois Central College for their constructive comments during the development of this textbook. Finally, we wish to thank our wives and children for their patience, understanding, love, and assistance in making this project a reality. It is to them that we dedicate this book.

Mark L. Berenson
David M. Levine
David Rindskopf

Contents

Preface *xiii*

1 Introduction *1*

1.1 What Is Modern Statistics? 1
1.2 The Growth and Development of Modern Statistics 2
1.3 Why Study Modern Statistics? 3
1.4 Examples of Applications of Statistics 4

2 Data Collection *6*

2.1 Introduction: The Need for Research 6
2.2 Sources of Data for Research 7
2.3 Obtaining Data through Survey Research 8
2.4 Designing the Questionnaire Instrument 13
Case A—The Alumni Association President's Study 13
2.5 Choosing the Sample Size for the Survey 18
2.6 Types of Samples 19
2.7 Drawing the Simple Random Sample 19
2.8 Obtaining the Responses 24
2.9 Data Preparation: Editing, Coding, and Transcribing 24
2.10 Data Collection: A Review and a Preview 27

3 Describing and Summarizing Data *35*

3.1 Introduction: What's Ahead 35
Case B—The Heldman Township Comptroller's Problem 36
3.2 Exploring the Data 36
3.3 Properties of Quantitative Data 37
3.4 Measures of Central Tendency 38
3.5 Measures of Dispersion 49
3.6 Shape 58

3.7 Calculating Descriptive Summary Measures from a Population 59
Case B—The Heldman Township Comptroller's Problem Revisited 67

4 Data Presentation: Tables and Charts *78*

4.1 Introduction: What's Ahead 78
4.2 Tabulating Quantitative Data: The Frequency Distribution 79
4.3 Tabulating Quantitative Data: The Relative Frequency Distribution and Percentage Distribution 87
4.4 Graphing Quantitative Data: The Histogram and Polygon 91
4.5 Qualitative Data Presentation: Summary Tables 99
4.6 Qualitative Data Presentation: Bar Charts and Pie Diagrams 100
4.7 Qualitative Data Presentation: Cross-Classification Tables 104

5 An Introduction to Exploratory Data Analysis Techniques *116*

5.1 Introduction: What's Ahead 116
5.2 The Stem-and-Leaf Display 116
5.3 Obtaining and Studying the Quartiles 120
5.4 The Midhinge—A Measure of Central Tendency 123
5.5 The Interquartile Range—A Measure of Dispersion 123
5.6 Five-Number Summaries 124
5.7 The Box-and-Whisker Plot 125
5.8 Descriptive Statistics: An Overview 127

6 Basic Probability *132*

6.1 Introduction: What's Ahead 132
Case C—The Medical Researcher's Study of Common Cold Remedies 132
6.2 Objective and Subjective Probability 133
6.3 Basic Probability Concepts 134
6.4 Simple or Marginal Probability 135
6.5 Joint Probability 138
6.6 Addition Rule 140
6.7 Conditional Probability 144
6.8 Multiplication Rule 146
6.9 Counting Rules 149

7 Basic Probability Distributions *159*

7.1 Introduction: What's Ahead 159
Case D—The Governor's Football Lottery Program 160
7.2 Mathematical Expectation for Discrete Random Variables 161

7.3 Binomial Distribution 170
7.4 Mathematical Models of Continuous Random Variables: The Probability Density Function 184
7.5 The Normal Distribution 185
Case D—The Governor's Football Lottery Program Revisited 193
7.6 The Normal Distribution as an Approximation to Various Discrete Probability Distributions 207
7.7 Summary 211

8 Sampling Distributions *216*
8.1 Introduction: What's Ahead 216
Case E—The Quality Control Director's Problem 217
8.2 The Sampling Distribution of the Mean 217
8.3 The Sampling Distribution of the Proportion 231
8.4 Sampling from Finite Populations 235
8.5 Summary and Overview 238

9 Estimation *240*
9.1 Introduction: What's Ahead 240
Case F—The Personnel Director's Employee Benefits Study 241
9.2 Confidence-Interval Estimate of the Mean (σ_X Known) 241
9.3 Confidence-Interval Estimate of the Mean (σ_X Unknown) 247
9.4 Confidence-Interval Estimate of the Mean Difference in Matched or Paired Samples 254
9.5 Confidence-Interval Estimate of the Proportion 257
9.6 Sample-Size Determination for the Mean 259
9.7 Sample-Size Determination for the Proportion 262
9.8 Estimation and Sample-Size Determination for Finite Populations 265

10 Hypothesis Testing for Quantitative Variables *273*
10.1 Introduction: What's Ahead 273
Case G—The University President's Evaluation of Student Grades and Employee Information 273
10.2 The Hypothesis-Testing Procedure 274
10.3 Type I and Type II Errors 276
10.4 Test of Hypothesis for the Mean (σ_X Known) 279
10.5 Summarizing the Steps of Hypothesis Testing 281
10.6 Test of Hypothesis for the Mean (σ_X Unknown) 282
10.7 One-Tailed Tests 284
10.8 Testing for the Difference Between the Means of Two Independent Populations 286

10.9 Testing for the Mean Difference in Matched or Paired Samples 291
10.10 A Connection Between Confidence Intervals and Hypothesis Testing 299
10.11 The *p*-Value Approach to Hypothesis Testing 303

11 Hypothesis Testing for Qualitative Variables *308*

11.1 Introduction: What's Ahead 308
Case H—The Marketing Director's Microcomputer Survey 308
11.2 Test of Hypothesis for a Proportion (One Sample) 309
11.3 Testing for a Difference Between Proportions from Two Independent Populations: Using the Normal Approximation 312
11.4 Testing for a Difference Between Proportions from Two Independent Populations: Using the Chi-Square Test 315
11.5 Testing the Difference Among the Proportions from *C* Independent Populations 321
11.6 Chi-Square Test of Independence in the $R \times C$ Table 324
11.7 Testing for a Difference Between Two Proportions in Matched or Paired Populations: The McNemar Test 328

12 The Analysis of Variance *337*

12.1 Introduction: What's Ahead 337
Case I—The Educational Researcher's Problem 338
12.2 The Logic Behind the Analysis of Variance 338
12.3 The *F* Distribution 342
12.4 The Analysis-of-Variance Table 346
12.5 Computational Methods 347
12.6 Assumptions of the Analysis of Variance 351
12.7 Contrasts Among Group Means (Optional Section) 352
12.8 Overview of Experimental Design 356

13 Simple Linear Regression and Correlation *362*

13.1 Introduction: What's Ahead 362
13.2 Simple Linear Regression 363
Case J—The Placement Officer's Problem 363
13.3 The Scatter Diagram 365
13.4 Developing the Simple Linear Regression Equation 365
13.5 The Standard Error of Estimate 377
13.6 Partitioning the Total Variation 379
13.7 Correlation Analysis: Measuring the Strength of the Association 385
13.8 Inferential Methods in Regression and Correlation: An Overview 389

13.9 Testing the Significance of the Linear Relationship 389
13.10 Confidence-Interval Estimation 396
13.11 Assumptions of Simple Linear Regression and Correlation 403
13.12 Residual Analysis (Optional Section) 406
13.13 Summary 412

14 Nonparametric Methods *419*

14.1 Introduction: What's Ahead 419
14.2 Classical versus Nonparametric Procedures 420
14.3 Advantages and Disadvantages of Using Nonparametric Methods 422
Case K—The Advertising Account Executive's Problem 423
14.4 Wald-Wolfowitz One-Sample Runs Test for Randomness 425
14.5 Wilcoxon Signed-Ranks Test 432
14.6 Wilcoxon Rank-Sum Test 442
14.7 Kruskal-Wallis Test for c Independent Samples 448
14.8 Spearman's Rank Correlation Procedure 453
14.9 Summary 458

15 Using Computers for Statistical Analysis *466*

15.1 Introduction: What's Ahead 466
15.2 The Role of the Computer 466
15.3 Writing a Computer Program (Optional Section) 467
15.4 Introduction to Statistical Packages 470
15.5 Using SPSS^X 472
15.6 Using Minitab 479
15.7 Using STATGRAPHICS 485
15.8 Summary 497

Appendices:

A. Review of Arithmetic, Algebra, and Summation Notation 498
B. Statistical Symbols and Greek Alphabet 504
C. Tables 505

Answers to Selected Problems (•) *534*

Index *551*

Chapter 1

Introduction

1.1 WHAT IS MODERN STATISTICS?

A century ago H. G. Wells commented that "statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." Each day of our lives we are exposed to a wide assortment of numerical information pertaining to such phenomena as stock market activity, unemployment rates, medical research findings, opinion poll results, weather forecasts, and sports data. Frequently, such information has a profound effect on our lives.

The subject of modern statistics encompasses the collection, presentation, and characterization of information to assist in both data analysis and the decision-making process.

1.2 THE GROWTH AND DEVELOPMENT OF MODERN STATISTICS

The growth and development of modern statistics can be traced to two separate phenomena: the needs of government to collect information on its citizenry (see References 1 and 2) and the development of the mathematics of probability theory.

The collection of data began at least as early as recorded history. During the Egyptian, Greek, and Roman civilizations information was obtained primarily for the purposes of taxation and military conscription. In the Middle Ages, church institutions often kept records concerning births, deaths, and marriages. In America, although various records were kept back to colonial times (see Reference 3), the Federal Constitution required the taking of a census every ten years beginning in 1790. Today, these data are used for many purposes including congressional apportionment and the allocation of federal funds.

1.2.1 Descriptive Statistics

These and other needs for data on a nationwide basis were closely intertwined with the development of the subject of descriptive statistics.

> **Descriptive statistics can be defined as those methods involving the collection, presentation, and characterization of a set of data in order to properly describe the various features of that set of data.**

Although descriptive statistical methods are important for presenting and characterizing information (see Chapters 2–5), it has been the development of inferential statistical methods as an outgrowth of probability theory that has led to the great expansion in the application of statistics in all fields of research today.

1.2.2 Inferential Statistics

The initial impetus for the formulation of the mathematics of probability theory came from the investigation of games of chance during the Renaissance. The foundations of the subject of probability can be traced back to the middle of the seventeenth century in the correspondence between the mathematician Pascal and the gambler Chevalier de Mere (see Reference 1). These and other developments by such mathematicians as Bernoulli, DeMoivre, and Gauss were the forerunners of the subject of inferential statistics. However, it has only been since the turn of this century that statisticians such as Pearson, Fisher, Gosset, Neyman, Wald, and Tukey have pioneered in the development of the methods of inferential statistics that are so widely applied in so many fields today.

Inferential statistics can be defined as those methods that make possible the estimation of a characteristic of a population or the making of a decision concerning a population based only on sample results.

To clarify this, a few definitions are necessary.

A population (or universe) is the totality of items or things under consideration.

A sample is the portion of the population that is selected for analysis.

A parameter is a summary measure that is computed to describe a characteristic of an entire population.

A statistic is a summary measure that is computed to describe a characteristic from only a sample of the population.

Thus, one major aspect of inferential statistics is the process of using sample statistics to draw conclusions about the true population parameters.

The need for inferential statistical methods derives from the need for sampling. As a population becomes large, it is usually too costly, too time consuming, and too cumbersome to obtain our information from the entire population. Decisions pertaining to the characteristics of the population have to be based on the information contained in but a sample of that population. Probability theory provides the link by ascertaining the likelihood that the results from the sample reflect the results in the population.

These ideas can be clearly seen in the example of a political poll. If the pollster wishes to estimate the percentage of the votes a candidate will receive in a particular election, he or she will not interview each of the thousands (or even millions) of voters. Instead, a sample of voters will be selected. Based on the outcome from the sample, conclusions will be drawn concerning the entire population of voters. Appended to these conclusions will be a probability statement specifying the likelihood or confidence that the results from the sample reflect the true voting behavior in the population.

1.3 WHY STUDY MODERN STATISTICS?

With the expansion in the use of both large-scale mainframe computers and minicomputer systems as well as the advent of the personal computer, the use of statistical methods as an aid to data analysis and decision making has grown dramatically over the past decade and will continue to grow in the future.

By studying the subject of modern statistics, we will obtain an appreciation for and understanding of those techniques which are used on the numerical information we encounter in both our professional and nonprofessional lives. The concepts and

methods described in this text are intended to provide a fundamental background in the subject of modern statistics and its applications in a wide variety of disciplines.

1.4 EXAMPLES OF APPLICATIONS OF STATISTICS

- *Agriculture*: A university's Agriculture College develops a new type of corn which it hopes will increase yield. But it knows that the number of bushels per acre it will get from the seed will vary, depending on factors such as the weather, the type of soil, and the equipment used by the farmer. How can it get a good estimate of what the average yield will be for the new corn, and whether the new corn is better than types already in use?
- *Business*: An airline needs to decide how many reservations to take for its flights. If it takes too many, then it will alienate customers who show up and are told there is no room for them on the flight. If it takes only as many reservations as there are seats, it will lose money because some people who reserve seats will not show up. How can the airline estimate the number of "no-shows" in order to plan the optimal number of reservations to allow—that is, the number which will minimize the number of people with reservations who get denied a seat, while maximizing the number of seats filled on the flights?
- *Health*: A toothpaste manufacturer wants to show that the use of its product reduces the incidence of tooth decay. The number of cavities which people have depends on many factors, such as what they eat, their genetic characteristics, how well they take care of their teeth, and the amount of fluorine in their water. How might studies be designed to prove that one toothpaste is better than another, given that there would be variability among people in the number of cavities they have, even if they all used the same toothpaste, let alone different ones?
- *Education*: Developers of a new computer language, PORKY, claim that students who learn to program in PORKY will also show improvement in academic subjects, reasoning skills, and IQ. Students, of course, vary tremendously in all of these areas, even when they are in the same schools and classes. Given this variability among students, how might we investigate whether students learning PORKY also improve in other areas?
- *Physics*: Scientists measuring the speed of light set up an apparatus which measures how long it takes a beam of light to return after being reflected in a mirror. Although they use instruments as precise as can be made, they find that the length of time varies slightly from trial to trial. Given that they know, from theory, that the measurement should be the same each time if the experiment is done perfectly, how should they estimate the speed of light from the many (different) measurements they have? How can they express the level of accuracy of their estimate?

What the above applications have in common is that they ask for a conclusion about a situation in which there is uncertainty because not all of the relevant factors can be measured or controlled, and because we cannot study the whole population. This causes the outcome to be variable, and thus not completely predictable. For example, the yield of corn, the number of passengers actually showing up, the number of cavities, the number of items correct on a test, and the measurement of

the speed of light all involve measurements of quantities which are uncertain. If we knew how many cavities each person should have under "normal" conditions, then testing the effect of using a new toothpaste would be easy.

Problems

For Problems 1.1 to 1.6, specify the general problem to be solved, the specific inference to be made, what the population is, and (if you are describing the results of an actual published study) what the weaknesses of the study might be. Where appropriate, tell what parameters are of primary interest and what statistics are used to arrive at a conclusion.

1.1 Describe three applications of statistics to business.

1.2 Describe three applications of statistics to sports.

1.3 Describe three applications of statistics to political science or public administration.

1.4 Describe three applications of statistics to psychology.

1.5 Describe three applications of statistics to education.

1.6 Describe three applications of statistics to medicine or medical research.

1.7 Political polls taken before elections have some aspects which are descriptive, and others which are inferential. What is descriptive, and what is inferential, about the results reported in such polls? What assumptions are usually made when inferences are made from these polls?

1.8 We often hear or read statements such as "One-third of the children in America go hungry each night." Presuming that these statements have some factual basis (that is, they are not just invented), tell how you might count or estimate the number of children going hungry each night.

1.9 When a doctor gives you a blood test, or takes your temperature or blood pressure, he or she makes a decision about whether the results are abnormal. How do you think the criteria for normality are arrived at for such tests? What are some possible problems which might arise in establishing what is normal?

References

1. Pearson, E. S., ed., *The History of Statistics in the Seventeenth and Eighteenth Centuries* (New York: Macmillan, 1978).
2. Pearson, E. S., and M. G. Kendall, eds., *Studies in the History of Statistics and Probability* (Darien, Conn.: Hafner, 1970).
3. Wattenberg, B. E., ed., *Statistical History of the United States: From Colonial Times to the Present* (New York: Basic Books, 1976).