6th edition

STATISTICS

A FIRST COURSE

John E. Freund & Gary A. Simon

SIXTH
EDITION

Statistics

A First Course

John E. Freund
Arizona State University

Gary A. Simon
New York University

Prentice Hall
Upper Saddle River, NJ 07458

Library of Congress Cataloging-in-Publication Data

Freund, John E.
Statistics : a first course. — 6th ed. / John E. Freund
Gary A. Simon.
p. cm.
Includes bibliographical references and index.
ISBN 0-13-083024-0
1. Statistics. I. Simon, Gary A. II. Title.
QA276.12.F74 1995
519.5—dc20 94-27691
CIP

To Doug and John and to Judy

Tim Bozik, editor-in-chief
Jerome Grant, acquisition editor
Jeanne Hoeting, managing editor
Nancy Bauer/JoAnne Wendelken, editorial assistants
Judith A. Matz-Coniglio, interior design
Paula Maylahn, cover design
Nicholas Romanelli, production editor
Trudy Pisciotti, manufacturing buyer

Simon & Schuster / A Viacom Company
Upper Saddle River, New Jersey 07458

Printed in the United States of America
10 9 8 7 6 5 4

ISBN 0-13-083024-0

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*

Statistics

A First Course

SIXTH EDITION

Preface

The teaching of statistics continues to improve. Courses have moved beyond descriptive statistics into issues of inference. Computers have eliminated the drudgery of traditional statistics courses, allowing the address of topics that could not have been considered previously.

Formerly directed at college juniors and seniors, statistics is now taught to college freshman and sophomores and also in enlightened high school programs. Our purpose, as in the previous five editions, is to reach the student earlier in his or her educational adventure. This objective will be apparent from the organization of the subject matter, the simplified language, the format and notation and, above all, the examples and exercises. The high level of mathematical precision of the previous editions has been enhanced without complicating the presentation.

Exercises are placed at the ends of sections, or closely related sections, to make it easier for both instructors and students to locate exercises related to given topics. There are five review sections, intended to reinforce concepts of a particular section or chapter.

Important formulas, definitions, and rules are highlighted within colored boxes. The normal and *t* distributions are repeated on the inside back cover; the inside front cover contains an index to the boxed formulas.

This edition introduces the small sample test of the difference between two means when the standard deviations are not assumed equal. Approximately 125 exercises are new (or revised) in this edition.

Computers are mentioned frequently. The printouts shown in the text are intended to impress on the reader the statistical software that will facilitate most of the techniques discussed. These programs will sometimes provide information that is otherwise not be readily accesible and may be preferred, for instance, when given statistical tables are inadequate. Most of these printouts were obtained with the widely used MINITAB program. The programs STATISTIX and ECONOMETRICS TOOLKIT were also used for several of the printouts. However, it is not necessary that the reader have access to these or any particular statistical package.

The use of a computer in conjunction with the text is entirely optional. However, to facilitate the reader's use of a computer, special exercises are preceded by a computer icon (💻). These exercises deal primarily with problems that are not solved trivially without a computer. The instructor may, or may not, choose to assign them, without loss of continuity.

As in the previous editions, controversy has not been avoided. The reader is exposed to the weaknesses of statistical techniques, as well as their strengths. It is hoped that this honest approach will provide a stimulus as well as a challenge.

The authors are indebted to Professor E. S. Pearson and the Biometrika trustees for permission to reproduce parts of Tables 8 and 18 from their *Biometrika Tables for Statisticians;* to Prentice-Hall, Inc., to reproduce part of Table 2 of R. A. Johnson and W. W. Wichern's *Applied Multivaritate Statistical Analysis;* to the Addison-Wesley Publishing Company to base Table VII on Table 11.4 of D. B. Owen's *Handbook of Statistical* Tables; and to the editor of the *Annals of Mathematical Statistics* to reproduce the material in Table VIII. The changes in this edition are my responsibility.

We express our appreciation to Richard Manning Smith for his work on the fourth edition. We also thank Phyllis Barnidge for her care in checking the manuscript. Our gratitude is also extended to the many students and colleagues whose suggestions have contributed greatly to this and to previous editions; in particular to:

Dale O. Everson, *University of Idaho*
Frank Gunnip, *Oakland Community College*
Tom Brieske, *Georgia State University*
John Van Iwaarden, *Hope College*
Arthur Larry Wright, *University of Arizona*
E. Ray Bobo, *Georgetown University*
John Heinen, *Regis College*
Lloyd B. Smith, Jr., *Lenoir-Rhyne College*
Raymond Mueller, *DeVry Institute of Technology*
Ian Walton, *Mission College*
Jake Uhrich, *South Puget Sound Community College*
Tom Hogenkamp, *Erie Community College*
Ronald Friesen, *Bluffton College*
Steven L. Harvey, *Seminole Junior College*
James Siefert, *ICM School of Business*
Mark Sherrin, *Flagler College*

Gary A. Simon

Contents

Preface *xiii*

1 Introduction *1*

1.1 Statistics, Past and Present *2*
1.2 The Study of Statistics *3*
1.3 Statistics, What Lies Ahead *6*
Solutions of Practice Exercises *7*

2 Date Display and Organization *8*

2.1 Listing Numeric Vaules *9*
2.2 Frequency Distributions *13*
2.3 Graphic Presentations *23*
Solutions of Practice Exercises *31*

3 Statistical Descriptions *33*

3.1 Samples and Populations *34*
3.2 Measures of Location: The Mean *35*
3.3 Measures of Location: The Weighted Mean *38*
3.4 Measures of Location: The Median and Other Fractiles *43*
3.5 Measures of Location: The Mode *52*
3.6 Measures of Variation: The Range *55*
3.7 Measures of Variation: The Standard Deviation *57*
3.8 Some Applications of the Standard Deviation *63*
3.9 The Description of Grouped Data ⋆ *68*
3.10 Some Further Descriptions ⋆ *75*
3.11 Technical Note (Summations) *80*
Solutions of Practice Exercises *83*

Review: Chapters 1, 2, & 3 *89*

⋆ All sections marked with a star may be considered optional.

4 Possibilities and Probabilities *96*

4.1 Counting *97*
4.2 Permutations *103*
4.3 Combinations *108*
4.4 Probability *112*
4.5 Mathematical Expectation *120*
4.6 A Decision Problem * *124*
Solutions of Practice Exercises *131*

5 Some Rules of Probability *134*

5.1 The Sample Space *135*
5.2 Events *137*
5.3 Some Basic Rules of Probability *146*
5.4 Probabilities and Odds *149*
5.5 Addition Rules *155*
5.6 Conditional Probability *163*
5.7 Independent Events *165*
5.8 Multiplication Rules *166*
5.9 Bayes' Theorem * *175*
Solutions of Practice Exercises *180*

Review: Chapters 4 & 5 *185*

6 Probability Distributions *193*

6.1 Probability Distributions *194*
6.2 The Binomial Distribution *196*
6.3 The Hypergeometric Distribution *206*
6.4 The Poisson Distribution *213*
6.5 The Multinomial Distribution *220*
6.6 The Mean of a Probability Distribution *222*
6.7 The Standard Deviation of a Probability Distribution *222*
6.8 Chebyshev's Theorem *229*
Solutions of Practice Exercises *232*

7 The Normal Distribution *236*

7.1 Continuous Distributions *237*
7.2 The Normal Distribution *240*

7.3 Some Applications 253
7.4 The Normal Approximation to the Binomial Distribution 258
Solutions of Practice Exercises 264

8 Sampling and Sampling Distributions 266

8.1 Random Sampling 267
8.2 Sampling Distributions 272
8.3 The Standard Error of the Mean 277
8.4 The Central Limit Theorem 282
Solutions of Practice Exercises 289

Review: Chapters 6, 7, & 8 292

9 The Analysis of Measurements 299

9.1 The Estimation of Means 300
9.2 Confidence Intervals for Means 307
9.3 Confidence Intervals for Means (Small Samples) 309
9.4 Confidence Intervals for Standard Deviations * 314
9.5 Test of Hypotheses 315
9.6 Significance Tests 321
9.7 Tests Concerning Means 327
9.8 Tests Concerning Means (Small Samples) 334
9.9 Differences Between Means 340
9.10 Differences Between Means (Small Samples) 343
9.11 Other Issues in Comparing Two Means * 345
9.12 Differences Between Means (Paired Data) 352
9.13 Differences Among *k* Means * 358
9.14 Analysis of Variance * 363
Solutions of Practice Exercises 373

10 The Analysis of Count Data 381

10.1 The Estimation of Proportions 382
10.2 Tests Concerning Proportions 387
10.3 Tests Concerning Proportions (Large Samples) 391
10.4 Differences Between Proportions 394
10.5 Differences Among Proportions 397
10.6 Contingency Tables 403

10.7 Tables with Ordered Categories *413*
10.8 Goodness of Fit *417*
Solutions of Practice Exercises 422

Review: Chapters 9 & 10 *430*

11 The Analysis of Paired Data *438*

11.1 Curve Fitting *439*
11.2 The Method of Least Squares *441*
11.3 Regression Analysis * *453*
11.4 The Coefficient of Correlation *461*
11.5 The Interpretation of r *465*
11.6 Correlations for Count Data *467*
11.7 A Significance Test for r *469*
Solutions of Practice Exercises 475

12 Nonparametric Tests *479*

12.1 The One-Sample Sign Test *480*
12.2 The Paired-Sample Sign Test *482*
12.3 The Sign Test (Large Samples) *483*
12.4 Rank Sums: The U Test *486*
12.5 Rank Sums: The U Test (Large Samples) * *490*
12.6 Rank Sums: The H Test * *495*
12.7 Tests of Randomness: Runs *497*
12.8 Tests of Randomness: Runs (Large Samples) * *499*
12.9 Tests of Randomness: Runs Above and Below the Median *501*
12.10 Rank Correlation *504*
Solutions of Practice Exercises 510

Review: Chapters 11 & 12 *518*

Bibliography *525*

Statistical Tables *529*

Answers to Odd-Numbered Exercises *554*

Index *I-1*

1 Introduction

1.1 Statistics, Past and Present *2*
1.2 The Study of Statistics *3*
1.3 Statistics, What Lies Ahead *6*

To see what statistics is and what statisticians do, let us refer to a recent edition of a popular dictionary.

> **Statistics** A branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data.
>
> **Statistician** One versed in or engaged in compiling collections of data.

The first definition makes it clear that statistics covers a lot of ground—it concerns how data are obtained, how they are manipulated, and how they are put to use. A shortcoming of this definition is that statistics is not really limited to masses of data. Indeed, one of the major achievements of modern statistics is that it enables us to squeeze useful information out of relatively meager sets of data; for instance, limited data on the severity of a rare disease, difficult-to-obtain data on the environmental impact of nuclear waste disposal methods, or very scarce data on the chemistry of distant stars.

The second definition leaves a great deal to be desired, for a statistician can be anyone from a clerk filing records on births and deaths or a person keeping tract of baseball batting averages and football pass completions, to a consultant who applies sophisticated decision-making techniques on the managerial level or a scholar who develops the mathematical theory on which the ever-growing body of statistical methods is based. **Thus, statistics provides opportunities for persons with very little formal training and also for those with advanced college degrees.**

In Sections 1.1 and 1.3 we discuss the recent growth of statistics and its expected development in the future. Section 1.2 stresses the need for the study of statistics.

1.1 STATISTICS, PAST AND PRESENT

The origin of the material we shall study in this book may be traced to two areas of interest which, on the surface, have very little in common: government (political science) and games of chance.

Governments have long used censuses to count persons and property. A famous example is the census reported in the *Domesday Book* of William of Normandy, completed in the year 1086, which covered most of England, listing its economic resources, including property owners and the land which they owned. In the first U.S. census, in 1790, government agents merely counted the population, but more recent U.S. censuses have become much wider in scope, providing a wealth of information about the population and the economy, and they are conducted every ten years. The most recent one was conducted in 1990.

The problem of describing, summarizing, and analyzing census data led to the development of methods which, until recently, constituted almost all there was to the subject of statistics. These methods, which originally consisted mainly of presenting data in the form of tables and charts, constitute what we now call **descriptive statistics.** This includes anything done to data which is designed to summarize, or describe, them without going any further, that is, without trying to infer anything that goes beyond the data themselves. For instance, if a newspaper reports net paid circulations of 172,316 in 1980 and 207,185 in 1990 and we perform the necessary calculations to show that there was an increase of 20.2%, our work belongs to the field of descriptive statistics. This would not be the case, however, if we used the given data to predict the newspaper's circulation in the year 2000.

Although descriptive statistics is an important branch of statistics and it continues to be widely used, statistical information usually arises from samples (from observations made on only part of a large set of items), and this means that its analysis will require generalizations which go beyond the data. As a result, an important feature of the growth of statistics in this century has been the shift in emphasis from methods which merely describe to methods which serve to make generalizations, that is, a shift in emphasis from descriptive statistics to the methods of **statistical inference.**

Such methods are required, for instance, to predict the operating life span of a sewing machine (on the basis of the performance of several such machines); to estimate the 1999 assessed value of all privately owned property in Orange Country, California (on the basis of business trends, population projections, and so forth); to compare the effectiveness of two reducing diets (on the basis of the weight losses of persons who have been on the diets); to determine the optimum dose of a medication (on the basis of tests performed with volunteer patients from selected hospitals); or to predict the flow of traffic on a freeway which has not yet been built (on the basis of past traffic counts on alternate routes).

In each of the situations described in the preceding paragraph, there are uncertainties because there is only partial, incomplete, or indirect information, and it is with the use of the methods of statistical inference that we judge the merits of the results and, perhaps, suggest a "most profitable" choice, a "most promising" prediction, or a "most reasonable" course of action.

In view of the uncertainties, we handle problems like these with statistical methods which find their origin in games of chance. Although the mathematical study of games of chance dates back to the seventeenth century, it was not until the early part of the nineteenth century that the theory developed for "heads or tails," for example, or "red or black" or "even or odd," was applied also to real-life situations where the outcomes were "boy or girl," "life or death," "pass or fail," and so forth. Thus, **probability theory** was applied to many problems in the behavioral, natural, and social sciences, and nowadays it provides an important tool for the analysis of any situation (in science, in business, or in everyday life) which in some way involves an element of uncertainty or chance. In particular, it provides the basis for the methods which we use when we generalize from observed data, namely, when we use the methods of statistical inference.

1.2 THE STUDY OF STATISTICS

There are two reasons why the scope of statistics and the need to study statistics have grown enormously in the last few decades. One reason is the increasingly quantitative approach employed in all the sciences, as well as in business and in many other activities which directly affect our lives. This includes the use of mathematical techniques in the evaluation of antipollution controls, in inventory planning, in the analysis of cloud formations, in the study of diet and longevity, in the evaluation of teaching techniques, and so forth.

The other reason is that the amount of statistical information that is collected, processed, and disseminated to the public for one reason or another has increased almost beyond comprehension, and what part of it is "good" statistics and what part is "bad" statistics is anybody's guess. To act as watchdogs, more and more persons with some knowledge of statistics are needed to take an active part in the collection of the data, in the analysis of the data, and, what is equally important, in all the preliminary planning. Without the latter, it is frightening to think of all the things that can go wrong. The results of costly studies can be completely useless if questions are ambiguous or asked in the wrong way, for example, or if instruments are poorly adjusted or if all relevant factors are not taken into account.

In contrast to this text, which presents a general introduction to the subject of statistics, numerous books have been written on business statistics, educational statistics, medical statistics, psychological statistics, . . . , and even on statistics for historians. Although problems arising in these various disciplines will sometimes require special statistical techniques, none of the basic methods discussed in this text is restricted to any particular field of application. In the same way in which 3 + 3 = 6 regardless of whether we are adding dollar amounts, horses, or trees, the methods we shall present provide appropriate **statistical models** regardless of whether the data are insurance premiums, tax payments, reaction times, test scores, or humidity readings. To emphasize this point, the examples and the exercises in this text were chosen to cover a wide spectrum of applications.

EXERCISES

(Exercises 1.1 and 1.6 are practice exercises; their complete solutions are given on page 7)

1.1 In four successive history tests a student received grades of 45, 73, 77, and 86. Which conclusions can be obtained from these figures by purely descriptive methods and which require generalizations? Explain your answers.

(a) Only one of the grades exceeds 85.
(b) The student's grades increased from each test to the next.
(c) The student must have studied harder for each successive test.
(d) The difference between the highest and lowest grades is 41.

1.2 Mary and Jean are real estate salespersons. In the first three months of 1994 Mary sold 3, 6, and 2 one-family homes and Jean sold 4, 0, and 5 one-family homes. Which of the following conclusions can be obtained from these figures by purely descriptive methods and which require generalizations? Explain your answers.

(a) During the three months Mary sold more one-family homes than Jean.
(b) Mary is a better real estate salesperson than Jean.
(c) Mary sold at least two one-family homes during each of the three months.
(d) Jean probably took her annual vacation during the second month.

1.3 The paid attendance of a minor league baseball team's first four home games was 5,308, 4,030, 6,386, and 5,770 in the year 1992 and 6,274, 5,883, 7,615 and 1,312 in the year 1993. Which of the following conclusions can be obtained from these figures by purely descriptive methods and which require generalizations? Explain your answers.

(a) The fourth 1993 figure was probably recorded incorrectly and should have been 7,312 instead of 1,312.
(b) Among the eight games, the paid attendance for any one game was highest in 1993.

(c) Among the eight games, the paid attendance in 1993 exceeded 6,000 more often than in 1992.

(d) Since the paid attendance at each of the first three home games was higher in 1993 than in 1992, the weather must have been better on those days.

1.4 Driving the same model car, five persons averaged 22.5, 21.7, 23.0, 22.5, and 21.8 miles per gallon. Which of the following conclusions can be obtained by purely descriptive methods and which require generalizations? Explain your answers.

(a) More often than any of the other figures, the drivers averaged 22.5 miles per gallon.

(b) The second and fifth persons must have done more city driving than the others.

(c) None of the averages differs from 22.0 by more than 1.0.

(d) If the whole experiment were repeated, none of the drivers would average less than 21.0 or more than 24.0 miles per gallon.

1.5 The three oranges which a person bought at a supermarket weighed 9, 8, and 13 ounces. Which of the following conclusions can be obtained from these data by purely descriptive methods and which require generalizations? Explain your answers.

(a) The average weight of the three oranges in 10 ounces.

(b) The average weight of oranges sold at that supermarket in 10 ounces.

1.6 "Bad" statistics may well result from asking questions in the wrong way or of the wrong persons. Explain why the following may lead to useless data:

(a) To determine public sentiment about a certain foreign trade restriction, an interviewer asks voters: "Do you feel that this unfair practice should be stopped?"

(b) In order to predict a municipal election, a public opinion poll telephones persons selected haphazardly from the city's telephone directory.

(c) In a study of art appreciation, persons are asked whether they like Indian art.

1.7 "Bad" statistics may also result from asking questions in the wrong place or at the wrong time. Explain why the following may lead to useless data:

(a) A house-to-house survey is made during weekday mornings to study consumer reaction to certain convenience foods.

(b) To predict an election, a poll taker interviews persons coming out of a building which houses the national headquarters of a political party.

(c) To determine what the average person spends on a vacation, a researcher interviews the passengers on a luxury cruise.

1.8 Explain why the following may lead to useless data:

(a) To determine the proportion of improperly sealed cans of coffee, a quality-control inspector examines every 50th can coming off a production line.

(b) To determine the average annual income of its graduates 10 years after graduation, a college's alumni office sent questionnaires in 1990 to all members of the class of 1980, and the estimate was based on the questionnaires returned.

(c) To study executives' reaction to its copying machines, the Xerox corporation hires a research organization to ask executives the question: How do you like using Xerox copies?

1.9 A statistically minded lawyer has his office on the third floor of a very tall office building, and whenever he leaves his office he records whether the first elevator which stops at his floor is going up or coming down. Having done this for some time, he discovers that the vast majority of the time the first elevator which stops is going down. Comment on his conclusion that fewer elevators are going up than are coming down.

1.3 STATISTICS, WHAT LIES AHEAD

In Section 1.1 we indicated how the emphasis in statistics has shifted from summarizing data by means of charts and tables to making inferences (that is, generalizations) on the basis of partial, incomplete, or indirect information. This is not meant to imply, however, that the subject of statistics has now become stable and inflexible, and that it has ceased to grow. Aside from the fact that new statistical techniques are constantly being developed to meet particular needs, the whole philosophy of statistics continues to be in a state of change. For example, attempts have been made to treat all problems of statistical inference within the framework of a unified theory called **decision theory,** which, so to speak, covers everything "from cradle to grave." One of the main features of this theory is that we must account for all the consequences which can arise when we base decisions on statistical data. This poses serious problems, as it is generally difficult, if not impossible, to put cash values on the consequences of one's actions. For instance, how can we put a cash value on the consequences of the decision whether or not to market a new medication, especially if the wrong decision may well involve the loss of human lives?

There are also statisticians who suggest that the emphasis has swung too far from descriptive statistics to statistical inference; rightly so, they feel that the solution of many problems requires only descriptive methods. To accommodate their needs, some new descriptive techniques have recently been developed under the general heading of **exploratory data analysis.** These will be introduced in Sections 2.1 and 3.4.

We have mentioned all this mainly to impress upon the reader that statistics, like most other fields of learning, is not static. Indeed, it is difficult to picture what a beginning course in statistics will be like twenty years hence. Certain aspects will probably still be the same, and that includes the role of probability theory in the foundations of statistics as well as certain data-summarizing techniques which have been very useful in the past and will undoubtedly continue to be widely used in the future.

SOLUTIONS OF PRACTICE EXERCISES

1.1 (a) The conclusion merely describes the data, as it can be seen that 86 exceeds 85, but 45, 73, and 77 do not.

(b) The conclusion merely describes the data, since 73 exceeds 45, 77 exceeds 73, and 86 exceeds 77.

(c) This is a generalization, as there can be many other reasons for the increases in the grades. For instance, the student may have felt better physically when he got the higher grades, or he may have been just lucky in studying the exact material asked for in the tests.

(d) This conclusion merely describes the data; the highest grade is 86, the lowest grade is 45, and their difference is $86 - 45 = 41$.

1.6 (a) This is called ''begging the question,'' because the interviewer suggests to the voters that the practice is, in fact, unfair.

(b) Persons selected from a telephone directory will generally not provide a satisfactory cross section of all persons eligible to vote. One reason is that some persons choose to have unlisted numbers, and the tendency to choose unlisted numbers may be related to political judgments.

(c) The term ''Indian art'' is ambiguous; some persons may respond with reference to the work of American Indians, while others may be thinking about art produced in India.