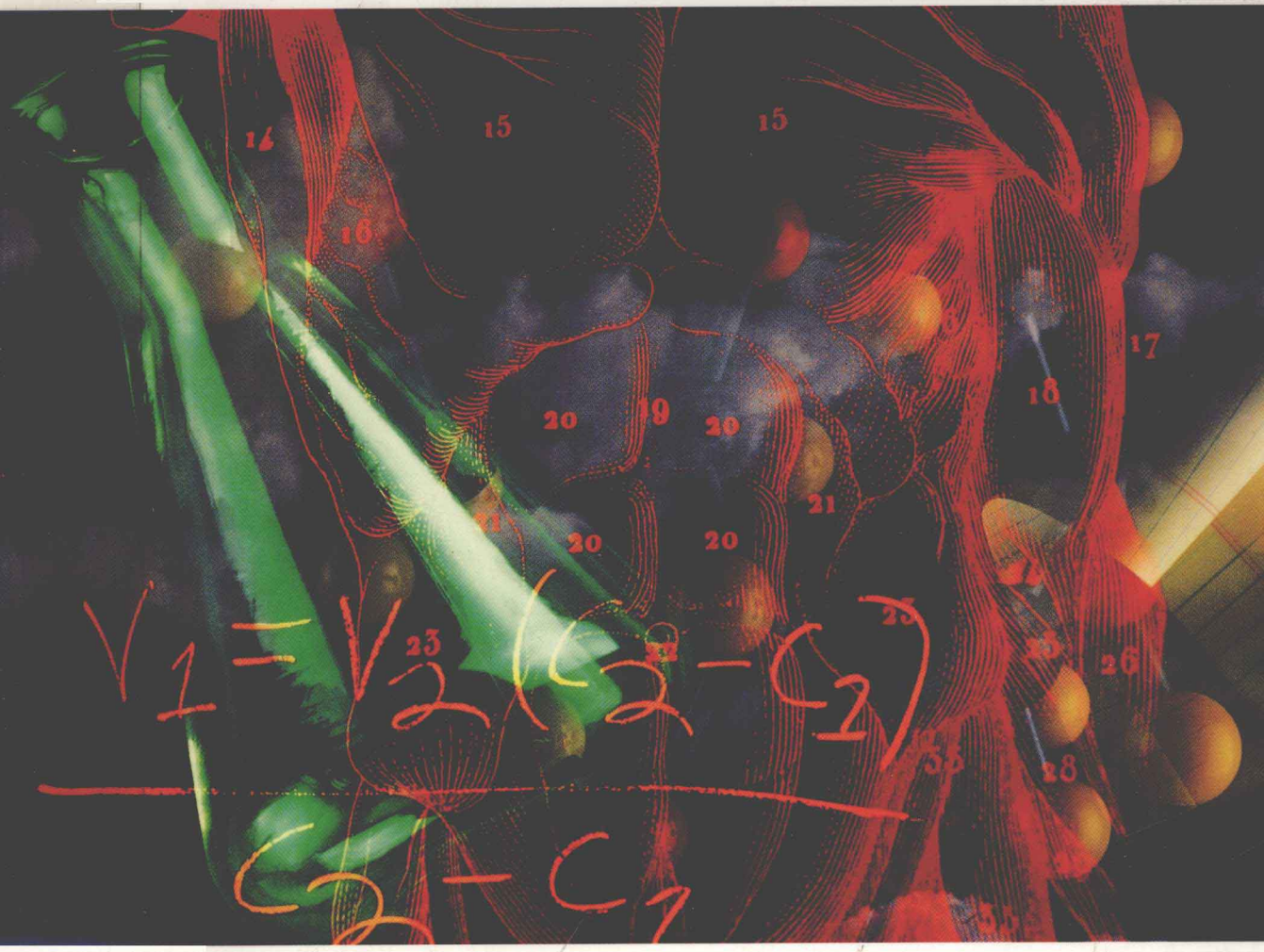
FUNDAMENTALS of Biostatistics



Bernard Rosner

FOURTH EDITION

FUNDAMENTALS OF BIOSTATISTICS

BERNARD ROSNER
HARVARD UNIVERSITY

Duxbury Press
An Imprint of Wadsworth Publishing Company
An International Thomson Publishing Company

ITP™

Belmont • Albany • Bonn • Boston • Cincinnati • Detroit • London • Madrid • Melbourne
Mexico City • New York • Paris • San Francisco • Singapore • Tokyo • Toronto • Washington

This book is dedicated to my wife Cynthia and my children Sarah, David, and Laura.

Statistics Editor: Alexander Kugushev
Editorial Assistant: Jennifer Burger
Production Editor: The Book Company
Print Buyer: Karen Hunt
Copy Editor: Linda Purrington
Technical Illustrator: TUCKERdesign
Cover: William Reuter
Compositor: Jonathan Peck Typographers
Printer: Banta Co., Harrisonburg, VA

This book is printed on acid-free recycled paper.

Printed in the United States of America

For more information, contact:

Wadsworth Publishing Company
10 Davis Drive
Belmont, California 94002

International Thomson Publishing
Berkshire House 168-173
High Holborn
London, WC1V7AA
England

Thomas Nelson Australia
102 Dodds Street
South Melbourne 3205
Victoria, Australia

Nelson Canada
1120 Birchmount Road
Scarborough, Ontario
Canada M1K 5G4

International Thomson Publishing GmbH
Königwinterer Strasse 418
53227 Bonn
Germany

International Thomson Publishing Asia
221 Henderson Road
#05-10
Singapore 0315

International Thomson Publishing-Japan
Hirakawacho-cho Kyowa Building, 3F
2-2-1 Hirakawacho-cho
Chiyoda-ku, 102 Tokyo
Japan

4 5 6 7 8 9 10—01 00 99 98 97

Library of Congress Cataloging-in-Publication Data
Rosner, Bernard (Bernard A.)
Fundamentals of biostatistics / Bernard Rosner,—4th ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-531-20840-8
1. Biometry. 2. Medical statistics. I. Title
QH323.5.R674 1994
574'.01'5185—dc20 94-13154
CIP

PREFACE

I have written this introductory-level biostatistics text for upper-level undergraduate or graduate students interested in medicine or other health-related areas. This book requires no previous background in statistics, and its mathematical level assumes only a knowledge of algebra.

Fundamentals of Biostatistics evolved from a set of notes that I used in a course in biostatistics taught to Harvard University undergraduates and Harvard Medical School students over the past fifteen years. I wrote this book to help motivate students to master the statistical methods that are most often used in the medical literature. From the student's viewpoint, it is important that the example material used to develop these methods is representative of what actually exists in the literature. Therefore, most examples and exercises used in this book are either based on actual articles from the medical literature or on actual medical research problems I have encountered during my consulting experience at the Harvard Medical School.

The Approach

Most other introductory statistics texts either use a completely nonmathematical, cookbook approach or develop the material in a rigorous, sophisticated mathematical framework. In this book I have attempted to follow an intermediate course, minimizing the amount of mathematical formulation and yet giving complete explanations of all the important concepts. *Every* new concept is developed systematically through completely worked out examples from current medical research problems. In addition, computer output is introduced where appropriate to illustrate these concepts.

The material in this book is suitable for either a one- or two-semester course in biostatistics. The material in Chapters 1 through 8 and Chapter 10 is suitable for a one-semester course. The instructor may select appropriate material from the other chapters as time permits.

Changes in the Fourth Edition

There are a total of 21 new sections and 9 additional sections with substantial revisions in the Fourth Edition. The new features include:

- An expanded set of computer exercises based on real data sets has been developed. The data sets are on a diskette that is provided with the book.
- A case study on lead exposure in children used in several chapters throughout the book.
- An extended discussion of randomized clinical trials including
  (a) design features (Section 6.4.1)
  (b) sample size issues (Section 10.7.3)
- One-Sample Inference for the Poisson distribution (Section 6.8 and 7.11)
- Sample size estimation based on confidence interval width (Section 7.7.3)
- Outlier detection techniques (Section 8.9)
- The one-way ANOVA random effects model (Section 9.6)

- The cross-over design (Section 9.7)
- A discussion of the most popular study designs in biomedical research (Section 10.3)
- Measures of effect for categorical data (Section 10.4)
- The hypergeometric distribution (Section 10.5.1)
- Issues in epidemiologic research include confounding, standardization, and effect modification (Sections 10.9 and 10.10)
- The Mantel extension test (Section 10.10.4)
- Power and sample size estimation for stratified categorical data (Section 10.11)
- Greatly expanded section on assessing goodness of fit of regression models including residual analysis for both simple linear regression (Section 11.6) and multiple linear regression (Section 11.7.3)
- Partial regression coefficients (Section 11.7.1)
- Partial residual plots (Section 11.7.3)
- Relationship between t test methods, analysis of variance, analysis of covariance, and regression analysis (Section 11.8)
- Interval estimation for correlation coefficients (Section 11.11.3)
- Partial and multiple correlation (Section 11.12)
- Intraclass correlation coefficient (Section 11.13)
- Expanded discussion of multiple logistic regression including methods for prediction and assessment of goodness of fit (Sections 11.14.4 and 11.14.5)
- A new chapter on inference for person-time data (Chapter 13), including
- Measures of effect for person-time data (Section 13.1)
- Inference for stratified person-time data (Section 13.3)
- Power and sample size estimation for person-time data (Section 13.4)
- Testing for trend with incidence rate data (Section 13.5)
- Estimation of survival curves with the Kaplan-Meier estimator (Section 13.7)

The new sections and the expanded sections for this edition have been indicated by an asterisk in the Contents.

The Exercises

There are a total of 1600 exercises in the Fourth Edition (compared with 1300 in the Third Edition). Students have indicated that they would like to see more completely solved problems. As a result, 639 of the problems have been moved to a Study Guide to accompany the text given with complete solutions. 95 of these problems are given in a Miscellaneous Problems section and are randomly ordered so that they are not tied to a specific chapter in the book. This gives the student additional practice in determining "what method to use in what situation." The remaining 544 problems are related to specific chapters in the text. All problems have complete solutions. Approximately 900 problems remain in the text, including all data-set based problems. Brief solutions are given to 300 of these problems in the Answer section, and are indicated by an asterisk (*) in the problem section of each chapter.

A Method of Computation

The method of handling computations in this edition of the book has also changed. All intermediate results are carried to full precision (10+ significant digits) even though they are presented with fewer significant digits (usually 2–3) in the text. Thus, intermediate results may seem to be inconsistent with final results in some instances, although this is not the case. This method allows for greater accuracy of final results and is the reason why there are slightly different results given for many calculations versus the previous editions, where intermediate results were carried to the same precision as shown in the text.

Organization

Fundamentals of Biostatistics, fourth edition, is organized as follows:

Chapter 1 is an *introductory chapter* giving an outline of the development of an actual medical study I was involved with. It provides a unique sense of the role of biostatistics in the medical research process.

Chapter 2 concerns *descriptive statistics* and presents all the major numeric and graphic tools used for displaying medical data. This chapter is especially important for both consumers and producers of medical literature, since much of the actual communication of information is accomplished via descriptive material.

Chapters 3 through 5 discuss *probability*. The basic principles of probability are developed, and the most common probability distributions, such as the binomial and normal distributions, are introduced. These distributions are used extensively in the later chapters of the book.

Chapters 6 through 10 cover some of the basic methods of *statistical inference*.

Chapter 6 introduces the concept of drawing random samples from populations. The difficult notion of a sampling distribution is also developed, including an introduction to the most common sampling distributions, such as the t and chi-square distributions. The basic methods of *estimation* are also presented, including an extensive discussion of confidence intervals.

Chapters 7 and 8 contain the basic principles of *hypothesis testing*. The most elementary hypothesis tests for normally distributed data, such as the t test, are also fully discussed for one- and two-sample problems.

Chapter 9 introduces the basic principles of the *analysis of variance* (ANOVA). The one-way analysis of variance fixed and random effects models are discussed as well as the analysis of data obtained using crossover designs.

Chapter 10 contains the basic concepts of *hypothesis testing* as applied to categorical data, including some of the most widely used statistical procedures, such as the chi-square test and Fisher's exact test.

Chapter 11 develops the principles of *regression analysis*. The case of simple linear regression is thoroughly covered, and extensions are provided for the multiple regression case. Important sections on goodness-of-fit of regression models are also included. Multiple logistic regression is also discussed.

Chapter 12 covers the basic principles of *nonparametric statistics*. The assumptions of normality are relaxed, and distribution-free analogues are developed for the tests in Chapters 7, 8, 9, and 11.

Chapter 13 introduces methods of analysis for person-time data. Included are methods for incidence rate data, as well as methods of survival analysis including the Kaplan-Meier survival curve estimator, the log rank test, and the Cox proportional hazards model.

The elements of study design are also discussed, including the concepts of matching, cohort studies, case-control studies, retrospective studies, prospective studies, and the sensitivity, specificity, and predictive value of screening tests. These designs are presented in the context of actual samples. In addition, specific sections on sample size estimation are provided for different statistical situations in Chapters 7, 8, 9, and 10.

A flowchart of appropriate methods of statistical inference on pages 671–675 provides an easy reference to the methods developed in this book. This flowchart is referred to at the end of each of Chapters 7 through 13 to give the student some perspective on how the methods in a particular chapter fit in with the overall collection of statistical methods introduced in this book.

In addition, an index summarizing all examples and problems used in this book is provided, grouped by *medical specialty*.

Acknowledgments

I am indebted to Debra Sheldon, Marie Sheehan, and Harry Taplin, who have been invaluable in helping to type this manuscript. I am indebted to those who reviewed the manuscript, among them: Stuart J. Anderson, University of Pittsburgh; Kenneth J. Koehler, Iowa State University; Donald J. Slymen, San Diego State University; and Craig D. Turnbull, University of North Carolina, Chapel Hill. I wish to thank Alex Kugushev, Jennie Burger, George Calmenson, and Linda Purrington, who were instrumental in providing editorial advice and in the preparation of the manuscript. I am indebted to my many colleagues at the Channing Laboratory, most notably Edward Kass, Frank Speizer, Charles Hennekens, Frank Polk, Ira Tager, Jerome Klein, James Taylor, Stephen Zinner, Scott Weiss, Frank Sacks, Walter Willett, Alvaro Munoz, Graham Colditz, and Susan Hankinson, and to my other colleagues at the Harvard Medical School, most notably Frederick Mosteller, Eliot Berson, Robert Ackerman, Mark Abelson, Arthur Garvey, Leo Chylack, Eugene Braunwald, and Arthur Dempster, who provided the inspiration for writing this book. Finally, I wish to acknowledge Leslie Miller, Andrea Wagner, Loren Fishman, and Roberta Shapiro, without whose clinical help the current edition of this book would not have been possible.

Bernard Rosner
Boston

CONTENTS

*asterisks indicate new section or subsection for the Fourth Edition

CHAPTER ONE

GENERAL OVERVIEW

Statistics is the science whereby inferences are made about specific random phenomena on the basis of relatively limited sample material. The field of statistics can be subdivided into two main areas: mathematical statistics and applied statistics. **Mathematical statistics** concerns the development of new methods of statistical inference and requires detailed knowledge of abstract mathematics for its implementation. **Applied statistics** concerns the application of the methods of mathematical statistics to specific subject areas, such as economics, psychology, and public health. **Biostatistics** is the branch of applied statistics that concerns the application of statistical methods to medical and biological problems.

A good way to learn about biostatistics and its role in the research process is to follow the flow of a research study from its inception at the planning stage to its completion, which usually occurs when a manuscript reporting the results of the study is published. As an example, I will describe one such study in which I participated.

A friend called one morning and in the course of our conversation mentioned that he had recently used a new, automated blood-pressure device of the type seen in many banks, hotels, and department stores. The machine had read his average diastolic blood pressure on several occasions as 115 mm Hg; the highest reading was 130 mm Hg. I was horrified to hear of his experience, since if these readings were true, my friend might be in imminent danger of having a stroke or developing some other serious cardiovascular disease. I referred him to a clinical colleague of mine who, using a standard blood-pressure cuff, measured my friend's diastolic blood pressure as 90 mm Hg. The contrast in the readings aroused my interest, and I began to jot down the readings on the digital display every time I passed the machine at my local bank. I got the distinct impression that a large percentage of the reported readings were in the hypertensive range. Although one would expect that hypertensives would be more likely to use such a machine, I still believed that blood-pressure readings obtained with the machine might not be comparable with those obtained using standard methods of blood-pressure measurement. I spoke to Dr. B. Frank Polk about my suspicion and succeeded in interesting him in a small-scale evaluation of such machines. We decided to send a human observer who was well trained in blood-pressure measurement techniques to several of these machines. He would offer to pay subjects 50¢ for the cost of using the machine if they would agree to fill out a short questionnaire and have their blood pressure measured by both a human observer and the machine.

At this stage we had to make several important decisions, each of which would prove vital to the success of the study. The decisions were based on the following questions:

(1) How many machines should we test?

(2) How many people should we test at each machine?

(3) In what order should the measurements be taken—should the human observer or the machine be used first? Ideally, we would have preferred to avoid this problem by taking both the human and machine readings simultaneously, but this procedure was logistically impossible.

(4) What other data should we collect on the questionnaire that might influence the comparison between methods?

(5) How should the data be recorded to facilitate their computerization at a later date?

(6) How should the accuracy of the computerized data be checked?

We resolved these problems as follows:

(1) and (2) We decided to test more than one machine (four to be exact), since we were not sure if the machines were comparable in quality. However, we wanted to sample enough subjects from each machine so that we would have an accurate comparison of the standard and automated methods for each machine. We tried to predict how large a discrepancy there might be between the two methods. Using the methods of sample-size determination discussed in this book, we calculated that we would need 100 subjects at each site to have an accurate comparison.

(3) We then had to decide in what order the measurements should be taken for each person. According to some reports, one problem that occurs with repeated blood-pressure measurements is that people tense up at the initial measurement, yielding higher blood pressure than at subsequent repeated measurements. Thus, we would not always want to use the automated or manual method first, since the effect of the method would get confused with the order-of-measurement effect. A conventional technique that we used here was to **randomize** the order in which the measurements were taken, so that for any person it was equally likely that the machine or the human observer would take the first measurement. This random pattern could be implemented by flipping a coin or, more likely, by using a table of **random numbers** as appears in Table 4 of the Appendix.

(4) We felt that the major extraneous factor that might influence the results would be body size, since we might have more difficulty getting accurate readings from people with fatter arms than from those with leaner arms. We also wanted to get some idea of the type of people who use these machines; so we asked questions about age, sex, and previous hypertensive history.

(5) To record the data, we developed a coding form that could be filled out on site and from which data could be easily entered on a computer terminal for subsequent analysis. Each person in the study was assigned an identification (ID) number by which the computer could uniquely identify that person. The data on the coding forms were then keyed and verified. That is, the same form was entered twice, and a comparison was made between the two records to make sure they were the same. If the records were not the same, the form was reentered.

(6) After data entry we ran some editing programs to ensure that the data were accurate. Checking each item on each form was impossible because of the large amount of data. Alternatively, we checked that the values for individual variables were within specified ranges and printed out aberrant values for manual checking. For example, we checked that all blood-pressure readings were at least 50 and no more than 300 and printed out all readings that fell outside this range.

After completing the data-collection, data-entry, and data-editing phases, we were ready to look at the results of the study. The first step in this process is to get a general feel for the data by summarizing the information in the form of several descriptive

statistics. This descriptive material can be numerical or graphical. If numerical, it can be in the form of a few summary statistics, which can be presented in tabular form or, alternatively, in the form of a **frequency distribution**, which lists each value in the data and how frequently it occurs. If graphical, the data are summarized pictorially and can be presented in one or more figures. The appropriate type of descriptive material will vary with the type of distribution considered. If the distribution is **continuous**, that is, if there are essentially an infinite number of possible values, as would be the case for blood pressure, then means and standard deviations might be the appropriate descriptive statistics. However, if the distribution is **discrete**, that is, if there are only a few possible values, as would be the case for sex, then percentages of people taking on each value would be the appropriate descriptive measure. In some cases both types of descriptive statistics are used for continuous distributions by condensing the range of possible values into a few groups and giving the percentage of people that fall into each group (e.g., the percentages of people that have blood pressures between 120 and 129 mm Hg and between 130 and 139 mm Hg, etc.).

In this study we decided first to look at mean blood pressure for each method at each of the four sites. Table 1.1 summarizes this information [1].

TABLE 1.1
Mean blood pressures and differences between machine and human readings at four locations

		Systolic blood pressure (mm Hg)					
		Machine		Human		Difference	
Location	Number of people	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
A	98	142.5	21.0	142.0	18.1	0.5	11.2
B	84	134.1	22.5	133.6	23.2	0.5	12.1
C	98	147.9	20.3	133.9	18.3	14.0	11.7
D	62	135.4	16.7	128.5	19.0	6.9	13.6

Source: By permission of the American Heart Association, Inc.

You might notice from this table that we did not obtain meaningful data from all the 100 people interviewed at each site, since we could not obtain valid readings from the machine for many of the people. This type of missing-data problem is very common in biostatistics and should be anticipated at the planning stage when deciding on sample sizes (which was not done in this study).

Our next step in the study was to determine whether the apparent differences in blood pressure between machine and human measurements at two of the locations (C, D) were "real" in some sense or were "due to chance." This type of question falls into the area of **inferential statistics**. We realized that although there was a 14-mm Hg difference in mean systolic blood pressure between the two methods for the 98 people we interviewed at location C, this difference might not hold up if we interviewed 98 other people at a different time, and we wanted to have some idea as to the **error in the estimate** of 14 mm Hg. In statistical jargon this group of 98 people represents a **sample** from the **population** of all people who use that machine. We were interested in the population and we wished to use the sample to help us learn something about

the population. In particular, we wanted to know how different the **estimated** mean difference of 14 mm Hg in our sample was likely to be from the **true** mean difference in the population of all people who might use this machine. More specifically, we wanted to know if it was still possible that there was no underlying difference between the two methods and that our results were due to chance. The 14-mm Hg difference in our group of 98 people is referred to as an **estimator** of the true mean difference (d) in the population. The problem of inferring characteristics of a population from a sample is the central concern of statistical inference and is a major topic in this text. To accomplish this aim, we needed to develop a **probability model**, which would tell us how likely it is that we would obtain a 14-mm Hg difference between the two methods in a sample of 98 people if there were no real difference between the two methods over the entire population of users of the machine. If this probability were sufficiently small, then we would begin to believe that a real difference existed between the two methods. In this particular case, using a probability model based on the t distribution, we were able to conclude that this probability was less than 1 in 1000 for each of machines C and D. This probability was sufficiently small for us to conclude that there was a real difference between the automatic and manual methods of taking blood pressure for two of the four machines tested.

We used a statistical package to perform the preceding data analyses. A package is a collection of statistical programs that describe data and perform various statistical tests on the data. Currently the most widely used statistical packages include SAS, SPSS^X, BMDP, and Minitab.

The final step in this study, after completing the data analysis, was to compile the results in the form of a publishable manuscript. Inevitably, because of space considerations, much of the material developed during the data-analysis phase was weeded out and only the essential items were presented for publication.

The review of this study should give you some idea of what medical research is about and what the role of biostatistics is in this process. The material in this text parallels the description of the data-analysis phase of the study described. Chapter 2 summarizes different types of descriptive statistics. In Chapters 3 through 5, some basic principles of probability and various probability models for use in later discussions of inferential statistics are presented. In Chapters 6 through 13, the major topics of inferential statistics as used in biomedical practice are discussed. Issues of study design or data collection are brought up only as they relate to other topics discussed in the text.

Reference

[1] Polk, B. F., Rosner, B., Feudo, R., & Vandenburgh, M. (1980). An evaluation of the Vita-Stat automatic blood pressure measuring device. *Hypertension*, *2*(2), 221–227.

CHAPTER TWO

DESCRIPTIVE STATISTICS

SECTION 2.1 Introduction

The first step in looking at data is to describe the data at hand in some concise way. In smaller studies this step can be accomplished by listing each data point. In general, however, this procedure is tedious or impossible and, even if it were possible, would not give an overall picture of what the data look like.

EXAMPLE 2.1 **Cancer, Nutrition** Some investigators have proposed that consumption of vitamin A prevents cancer. To test this theory, a dietary questionnaire to collect data on vitamin-A consumption among 200 hospitalized cancer cases and 200 controls might be used. The controls would be matched on age and sex to the cancer cases and would be in the hospital at the same time for an unrelated disease. What should be done with these data after they are collected? ■■■

Before any formal attempt to answer this question can be made, the vitamin-A consumption among cases and controls must be described. Consider Figure 2.1. The **bar graphs** show visually that the controls have a higher vitamin-A consumption than the cases do, particularly in doses higher than the recommended daily allowance (RDA).

EXAMPLE 2.2 **Pulmonary Disease** Medical researchers have often suspected that passive smokers—people who themselves do not smoke but who live or work in an environment where others smoke—might have impaired pulmonary function as a result. In 1980 a research group in San Diego published results indicating that passive smokers did indeed have significantly lower pulmonary function than comparable nonsmokers who did not work in smoky environments [1]. As supporting evidence, the authors measured the carbon-monoxide (CO) concentrations in the working environments of passive smokers and of nonsmokers (where no smoking was permitted in the workplace) to see if the relative CO concentration changed over the course of the day. These results are displayed in the form of a **scatter plot** in Figure 2.2. ■■■

Figure 2.2 clearly shows that the CO concentrations in the two working environments are about the same early in the day but diverge widely in the middle of the day and then converge again after the working day is over at 7 P.M.

Graphic displays illustrate the important role of descriptive statistics, which is to quickly display data to give the researcher a clue as to the principal trends in the data and suggest hints as to where a more detailed look at the data, using the methods of inferential statistics, might be worthwhile. Descriptive statistics are also crucially important in conveying the final results of studies in written publications. Unless it is one of their primary interests, most readers will not have time to critically evaluate the work of others but will be influenced mainly by the descriptive statistics presented.