GLOBAL EDITION

Statistics for the Life Sciences

FIFTH EDITION

Myra L. Samuels
Jeffrey A. Witmer
Andrew A. Schaffner

ALWAYS LEARNING

PEARSON

Statistics for the Life Sciences

Fifth Edition

Global Edition

Myra L. Samuels

Purdue University

Jeffrey A. Witmer

Oberlin College

Andrew A. Schaffner

California Polytechnic State University, San Luis Obispo

PEARSON

Boston Columbus Indianapolis New York San Francisco Hoboken
Amsterdam Cape Town Dubai London Madrid Milan Munich
Paris Montréal Toronto Delhi Mexico City São Paulo
Sydney Hong Kong Seoul Singapore Taipei Tokyo

Editor in Chief: Deirdre Lynch
Editorial Assistant: Justin Billing
Assistant Acquisitions Editor, Global Edition: Murchana Borthakur
Associate Project Editor, Global Edition: Binita Roy
Program Manager: Tatiana Anacki
Program Team Lead: Marianne Stepanian
Project Team Lead: Christina Lepre
Media Producer: Jean Choe
Senior Marketing Manager: Jeff Weidenaar
Marketing Assistant: Brooke Smith
Senior Author Support/Technology Specialist: Joe Vetere
Rights and Permissions Advisor: Diahanne Lucas
Procurement Specialist: Carol Melville
Senior Manufacturing Controller, Production, Global Edition: Trudy Kimber
Design Manager: Beth Paquin
Cover Design: Lumina Datamatics
Production Management/Composition: Sherrill Redd/iEnergizer Aptara®, Ltd.
Cover Image: © Holly Miller-Pollack/Shutterstock

Acknowledgements of third party content appear on page 636, which constitutes an extension of this copyright page.

Pearson Education Limited
Edinburgh Gate
Harlow
Essex CM20 2JE
England

and Associated Companies throughout the world

Visit us on the World Wide Web at:
www.pearsonglobaleditions.com

Authorized adaptation from the United States edition, entitled Statistics for the Life Sciences, 5th edition, ISBN 978-0-321-98958-1, by Myra L. Samuels, Jeffrey A. Witmer, and Andrew A. Schaffner, published by Pearson Education © 2016.

ISBN 10: 1-292-10181-4
ISBN 13: 978-1-292-10181-1

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

10 9 8 7 6 5 4 3 2 1

Typeset in 9 New Aster LT Std by iEnergizer Aptara®, Ltd.

Printed and bound in Malaysia.

Contents

*Indicates optional chapters

**Selected Chapter Appendices, Chapter References and Selected Chapter Tables can be found on www.pearsonglobaleditions.com/Samuels

Preface

Statistics for the Life Sciences is an introductory text in statistics, specifically addressed to students specializing in the life sciences. Its primary aims are (1) to show students how statistical reasoning is used in biological, medical, and agricultural research; (2) to enable students to confidently carry out simple statistical analyses and to interpret the results; and (3) to raise students' awareness of basic statistical issues such as randomization, confounding, and the role of independent replication.

Style and Approach

The style of *Statistics for the Life Sciences* is informal and uses only minimal mathematical notation. There are no prerequisites except elementary algebra; anyone who can read a biology or chemistry textbook can read this text. It is suitable for use by graduate or undergraduate students in biology, agronomy, medical and health sciences, nutrition, pharmacy, animal science, physical education, forestry, and other life sciences.

Use of Real Data Real examples are more interesting and often more enlightening than artificial ones. *Statistics for the Life Sciences* includes hundreds of examples and exercises that use real data, representing a wide variety of research in the life sciences. Each example has been chosen to illustrate a particular statistical issue. The exercises have been designed to reduce computational effort and focus students' attention on concepts and interpretations.

Emphasis on Ideas The text emphasizes statistical ideas rather than computations or mathematical formulations. Probability theory is included only to support statistical concepts. The text stresses interpretation throughout the discussion of descriptive and inferential statistics. By means of salient examples, we show why it is important that an analysis be appropriate for the research question to be answered, for the statistical design of the study, and for the nature of the underlying distributions. We help the student avoid the common blunder of confusing statistical nonsignificance with practical insignificance and encourage the student to use confidence intervals to assess the magnitude of an effect. The student is led to recognize the impact on real research of design concepts such as random sampling, randomization, efficiency, and the control of extraneous variation by blocking or adjustment. Numerous exercises amplify and reinforce the student's grasp of these ideas.

The Role of Technology The analysis of research data is usually carried out with the aid of a computer. Computer-generated graphs are shown at several places in the text. However, in studying statistics it is desirable for the student to gain experience working directly with data, using paper and pencil and a hand-held calculator, as well as a computer. This experience will help the student appreciate the nature and purpose of the statistical computations. The student is thus prepared to make intelligent use of the computer—to give it appropriate instructions and properly interpret the output. Accordingly, most of the exercises in this text are intended for hand calculation. However, electronic data files are provided

at www.pearsonglobaleditions.com/Samuels for many of the exercises, so that a computer can be used if desired. Selected exercises are identified as **Computer Problems** to be completed with use of a computer. (Typically, the computer exercises require calculations that would be unduly burdensome if carried out by hand.)

Organization

This text is organized to permit coverage in one semester of the maximum number of important statistical ideas, including power, multiple inference, and the basic principles of design. By including or excluding optional sections, the instructor can also use the text for a one-quarter course or a two-quarter course. It is suitable for a terminal course or for the first course of a sequence.

The following is a brief outline of the text.

Unit I: Data and Distributions

Chapter 1: Introduction. The nature and impact of variability in biological data. The hazards of observational studies, in contrast with experiments. Random sampling.

Chapter 2: Description of distributions. Frequency distributions, descriptive statistics, the concept of population versus sample.

Chapters 3, 4, and 5: Theoretical preparation. Probability, binomial and normal distributions, sampling distributions.

Unit II: Inference for Means

Chapter 6: Confidence intervals for a single mean and for a difference in means.

Chapter 7: Hypothesis testing, with emphasis on the *t* test. The randomization test, the Wilcoxon-Mann-Whitney test.

Chapter 8: Inference for paired samples. Confidence interval, *t* test, sign test, and Wilcoxon signed-rank test.

Unit III: Inference for Categorical Data

Chapter 9: Inference for a single proportion. Confidence intervals and the chi-square goodness-of-fit test.

Chapter 10: Relationships in categorical data. Conditional probability, contingency tables. Optional sections cover Fisher's exact test, McNemar's test, and odds ratios.

Unit IV: Modeling Relationships

Chapter 11: Analysis of variance. One-way layout, multiple comparison procedures, one-way blocked ANOVA, two-way ANOVA. Contrasts and multiple comparisons are included in optional sections.

Chapter 12: Correlation and regression. Descriptive and inferential aspects of correlation and simple linear regression and the relationship between them.

Chapter 13: A summary of inference methods.

Most sections within each chapter conclude with section-specific exercises. Chapters and units conclude with supplementary exercises that provide opportunities for students to practice integrating the breadth of methods presented within the chapter or across the entire unit. Selected statistical tables are provided at the back of the book; other tables are available at www.pearsonglobaleditions.com/Samuels.

The tables of critical values are especially easy to use because they follow mutually consistent layouts and so are used in essentially the same way.

Optional appendices at the back of the book and available online at www.pearsonglobaleditions.com/Samuels give the interested student a deeper look into such matters as how the Wilcoxon-Mann-Whitney null distribution is calculated.

Changes to the Fifth Edition

- Chapters are grouped by unit, and feature Unit Highlights with reflections, summaries, and additional examples and exercises at the end of each unit that often require connecting ideas from multiple chapters.
- We added material on randomization-based inference to introduce or motivate most inference procedures presented in this text. There are now presentations of randomization methods at the beginnings of Chapters 7, 8, 10, 11, and 12.
- New exercises have been added throughout the text. Many exercises from the previous edition that involved calculation and reading tables have been updated to exercises that require interpretation of computer output.
- We replaced many older examples throughout the text with examples from current research from a variety life science disciplines.
- Chapter notes have been updated to include references to new examples. These are now available online at www.pearsonglobaleditions.com/Samuels with some selected notes remaining in print.

Instructor Supplements

Instructor's Solutions Manual (downloadable) (ISBN-13: 978-1-292-10183-5; ISBN-10: 1-292-10183-0) Solutions to all exercises are available as a downloadable manual from Pearson Education's online catalog at www.pearsonglobaleditions.com/Samuels. Careful attention has been paid to ensure that all methods of solution and notation are consistent with those used in the core text.

PowerPoint Slides (downloadable) (ISBN-13: 978-1-292-10184-2; ISBN-10: 1-292-10184-9) Selected figures and tables from throughout the textbook are available as downloadable PowerPoint slides for use in creating custom PowerPoint lecture presentations. These slides are available for download at www.pearsonglobaleditions.com/Samuels.

Student Supplements

Data Sets The larger data sets used in examples and exercises in the book are available as .csv files at www.pearsonglobaleditions.com/Samuels

StatCrunch™ StatCrunch is powerful web-based statistical software that allows users to perform complex analyses, share data sets, and generate compelling reports of their data. The vibrant online community offers tens of thousands of shared data sets for students to analyze.

- **Collect.** Users can upload their own data to StatCrunch or search a large library of publicly shared data sets, spanning almost any topic of interest. Also, an online survey tool allows users to quickly collect data via web-based surveys.
- **Crunch.** A full range of numerical and graphical methods allows users to analyze and gain insights from any data set. Interactive graphics help users understand statistical concepts and are available for export to enrich reports with visual representations of data.
- **Communicate.** Reporting options help users create a wide variety of visually appealing representations of their data.

StatCrunch access is available to qualified adopters. StatCrunch Mobile is now available—just visit www.statcrunch.com/mobile from the browser on your smartphone or tablet. For more information, visit our website at www.StatCrunch.com, or contact your Pearson representative.

Acknowledgments for the Fifth Edition

The fifth edition of *Statistics for the Life Science* retains the style and spirit of the writing of Myra Samuels. Prior to her tragic death from cancer, Myra wrote the first edition of the text, based on her experience both as a teacher of statistics and as a statistical consultant. We hope that the book retains her vision.

Many researchers have contributed sets of data to the text, which have enriched the text considerably. We have benefited from countless conversations over the years with David Moore, Dick Scheaffer, Murray Clayton, Alan Agresti, Don Bentley, George Cobb, and many others who have our thanks.

We are grateful for the sound editorial guidance and encouragement of Katherine Roz. We are also grateful for adopters of the earlier editions, particularly Robert Wolf and Jeff May, whose suggestions led to improvements in the current edition. Finally, we express our gratitude to the reviewers of this edition:

Jeffrey Schmidt (University of Wisconsin-Parkside), Liansheng Tang (George Mason University), Tim Hanson (University of South Carolina), Mohammed Kazemi (University of North Carolina–Charlotte), Kyoungmi Kim (University of California, Davis), and Leslie Hendrix (University of South Carolina)

Special Thanks

To Merrilee, for her steadfast support.
JAW

To Michelle, for her patience and encouragement, and for my sons, Ganden and Tashi, for their curiosity and interest in learning something new every day.
AAS

Pearson wishes to thank and acknowledge the following people for their work on the Global Edition:

Contributor

C. V. Vinay, *JSS Academy of Technical Education*

Dilip Nath, *Gauhati University*

Reviewers

D. V. Chandrashekhar, *Vivekananda Institute of Technology*

Sunil Jacob John, *National Institute of Technology Calicut*

D. V. Jayalakshmamma, *Vemana Institute of Technology*

Introduction

Chapter 1

OBJECTIVES

In this chapter we will look at a series of examples of areas in the life sciences in which statistics is used, with the goal of understanding the scope of the field of statistics. We will also

- explain how experiments differ from observational studies.
- discuss the concepts of placebo effect, blinding, and confounding.
- discuss the role of random sampling in statistics.

1.1 Statistics and the Life Sciences

Researchers in the life sciences carry out investigations in various settings: in the clinic, in the laboratory, in the greenhouse, in the field. Generally, the resulting data exhibit some *variability*. For instance, patients given the same drug respond somewhat differently; cell cultures prepared identically develop somewhat differently; adjacent plots of genetically identical wheat plants yield somewhat different amounts of grain. Often the degree of variability is substantial even when experimental conditions are held as constant as possible.

The challenge to the life scientist is to discern the patterns that may be more or less obscured by the variability of responses in living systems. The scientist must try to distinguish the "signal" from the "noise."

Statistics is the science of understanding data and of making decisions in the face of variability and uncertainty. The discipline of statistics has evolved in response to the needs of scientists and others whose data exhibit variability. The concepts and methods of statistics enable the investigator to describe variability and to plan research so as to take variability into account (i.e., to make the "signal" strong in comparison to the background "noise" in data that are collected). Statistical methods are used to analyze data so as to extract the maximum information and also to quantify the reliability of that information.

We begin with some examples that illustrate the degree of variability found in biological data and the ways in which variability poses a challenge to the biological researcher. We will briefly consider examples that illustrate some of the statistical issues that arise in life sciences research and indicate where in this book the issues are addressed.

The first two examples provide a contrast between an experiment that showed no variability and another that showed considerable variability.

Example 1.1.1

Vaccine for Anthrax Anthrax is a serious disease of sheep and cattle. In 1881, Louis Pasteur conducted a famous experiment to demonstrate the effect of his vaccine against anthrax. A group of 24 sheep were vaccinated; another group of 24 unvaccinated sheep served as controls. Then, all 48 animals were inoculated with a virulent culture of anthrax bacillus. Table 1.1.1 shows the results.[1] The data of Table 1.1.1 show no variability; all the vaccinated animals survived and all the unvaccinated animals died. ■

Table 1.1.1 Response of sheep to anthrax

Response	Treatment	
	Vaccinated	Not vaccinated
Died of anthrax	0	24
Survived	24	0
Total	24	24
Percent survival	100%	0%

Example 1.1.2

Bacteria and Cancer To study the effect of bacteria on tumor development, researchers used a strain of mice with a naturally high incidence of liver tumors. One group of mice were maintained entirely germ free, while another group were exposed to the intestinal bacteria *Escherichia coli*. The incidence of liver tumors is shown in Table 1.1.2.[2]

Table 1.1.2 Incidence of liver tumors in mice

Response	Treatment	
	E. coli	Germ free
Liver tumors	8	19
No liver tumors	5	30
Total	13	49
Percent with liver tumors	62%	39%

In contrast to Table 1.1.1, the data of Table 1.1.2 show variability; mice given the same treatment did not all respond the same way. Because of this variability, the results in Table 1.1.2 are equivocal; the data suggest that exposure to *E. coli* increases the risk of liver tumors, but the possibility remains that the observed difference in percentages (62% versus 39%) might reflect only chance variation rather than an effect of *E. coli.* If the experiment were replicated with different animals, the percentages might change substantially.

One way to explore what might happen if the experiment were replicated is to simulate the experiment, which could be done as follows. Take 62 cards and write "liver tumors" on 27 (= 8 + 19) of them and "no liver tumors" on the other 35 (= 5 + 30). Shuffle the cards and randomly deal 13 cards into one stack (to correspond to the *E. coli* mice) and 49 cards into a second stack. Next, count the number of cards in the "*E. coli* stack" that have the words "liver tumors" on them—to correspond to mice exposed to *E. coli* who develop liver tumors—and record whether this number is greater than or equal to 8. This process represents distributing 27 cases of liver tumors to two groups of mice (*E. coli* and germ free) randomly, with *E. coli* mice no more likely, nor any less likely, than germ-free mice to end up with liver tumors.

If we repeat this process many times (say, 10,000 times, with the aid of a computer in place of a physical deck of cards), it turns out that roughly 12% of the time we get 8 or more *E. coli* mice with liver tumors. Since something that happens 12% of the time is not terribly surprising, Table 1.1.2 does not provide significant evidence that exposure to *E. coli* increases the incidence of liver tumors. ■

In Chapter 10 we will discuss statistical techniques for evaluating data such as those in Tables 1.1.1 and 1.1.2. Of course, in some experiments variability is minimal and the message in the data stands out clearly without any special statistical analysis. It is worth noting, however, that absence of variability is itself an experimental result that must be justified by sufficient data. For instance, because Pasteur's anthrax data (Table 1.1.1) show no variability at all, it is intuitively plausible to conclude that the data provide "solid" evidence for the efficacy of the vaccination. But note that this conclusion involves a judgment; consider how much *less* "solid" the evidence would be if Pasteur had included only 3 animals in each group, rather than 24. Statistical analyses can be used to make such a judgment, that is, to determine if the variability is indeed negligible. Thus, a statistical view can be helpful even in the absence of variability.

The next two examples illustrate additional questions that a statistical approach can help to answer.

Example 1.1.3

Flooding and ATP In an experiment on root metabolism, a plant physiologist grew birch tree seedlings in the greenhouse. He flooded four seedlings with water for one day and kept four others as controls. He then harvested the seedlings and analyzed the roots for adenosine triphosphate (ATP). The measured amounts of ATP (nmoles per mg tissue) are given in Table 1.1.3 and displayed in Figure 1.1.1.[3]

Table 1.1.3 ATP concentration in birch tree roots (nmol/mg)

Flooded	Control
1.45	1.70
1.19	2.04
1.05	1.49
1.07	1.91

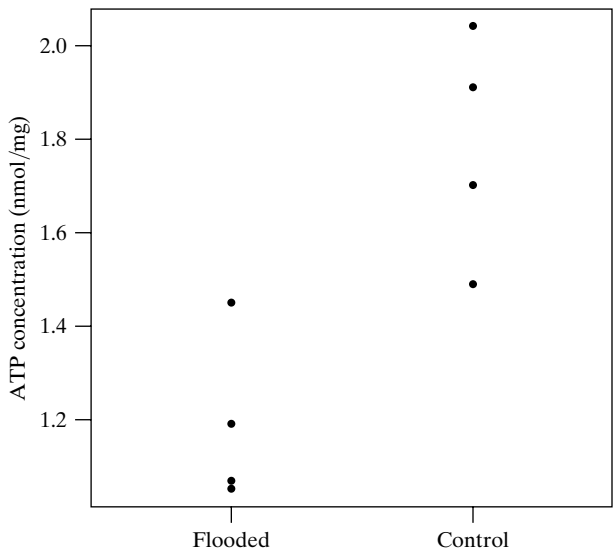


Figure 1.1.1 ATP concentration in birch tree roots

The data of Table 1.1.3 raise several questions: How should one summarize the ATP values in each experimental condition? How much information do the data provide about the effect of flooding? How confident can one be that the reduced ATP in the flooded group is really a response to flooding rather than just random variation? What size experiment would be required in order to firmly corroborate the apparent effect seen in these data?

Chapters 2, 6, and 7 address questions like those posed in Example 1.1.3. One question that we can address here is whether the data in Table 1.1.3 are consistent with the claim that flooding has no effect on ATP concentration, or instead provide significant evidence that flooding affects ATP concentrations. If the claim of no effect is true, then should we be surprised to see that all four of the flooded observations are smaller than each of the control observations? Might this happen by chance alone? If we wrote each of the numbers 1.05, 1.07, 1.19, 1.45, 1.49, 1.91, 1.70, and 2.04 on cards, shuffled the eight cards, and randomly dealt them into two piles, what is the chance that the four smallest numbers would end up in one pile and the four largest numbers in the other pile? It turns out that we could expect this to happen 1 time in 35 random shufflings, so "chance alone" would only create the kind of imbalance seen in Figure 1.1.1 about 2.9% of the time (since $1/35 = 0.029$). Thus, we have some evidence that flooding has an effect on ATP concentration. We will develop this idea more fully in Chapter 7.

Example 1.1.4

MAO and Schizophrenia Monoamine oxidase (MAO) is an enzyme that is thought to play a role in the regulation of behavior. To see whether different categories of patients with schizophrenia have different levels of MAO activity, researchers collected blood specimens from 42 patients and measured the MAO activity in the platelets. The results are given in Table 1.1.4 and displayed in Figure 1.1.2. (Values are expressed as nmol benzylaldehyde product per 10^8 platelets per hour.[4]) Note that it is much easier to get a feeling for the data by looking at the graph (Figure 1.1.2) than it is to read through the data in the table. The use of graphical displays of data is a very important part of data analysis.

Table 1.1.4 MAO activity in patients with schizophrenia

Diagnosis	MAO activity				
I:	6.8	4.1	7.3	14.2	18.8
Chronic	9.9	7.4	11.9	5.2	7.8
undifferentiated	7.8	8.7	12.7	14.5	10.7
schizophrenia	8.4	9.7	10.6		
(18 patients)					
II:	7.8	4.4	11.4	3.1	4.3
Undifferentiated	10.1	1.5	7.4	5.2	10.0
with paranoid	3.7	5.5	8.5	7.7	6.8
features	3.1				
(16 patients)					
III:	6.4	10.8	1.1	2.9	4.5
Paranoid	5.8	9.4	6.8		
schizophrenia					
(8 patients)					

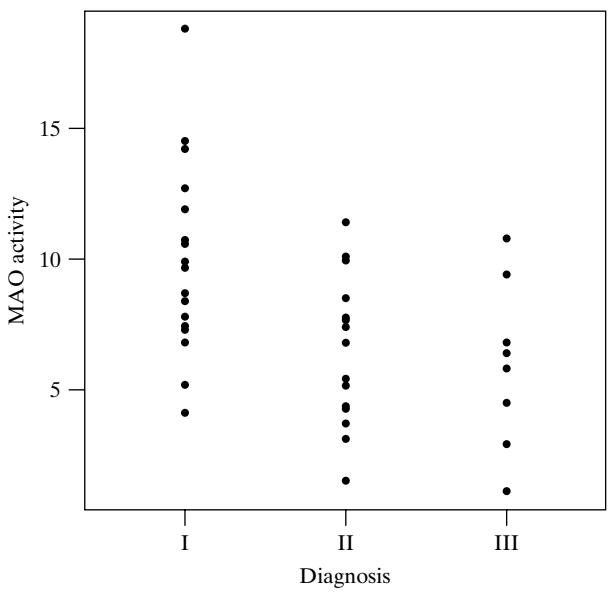


Figure 1.1.2 MAO activity in patients with schizophrenia

To analyze the MAO data, one would naturally want to make comparisons among the three groups of patients, to describe the reliability of those comparisons, and to characterize the variability within the groups. To go beyond the data to a biological interpretation, one must also consider more subtle issues, such as the

following: How were the patients selected? Were they chosen from a common hospital population, or were the three groups obtained at different times or places? Were precautions taken so that the person measuring the MAO was unaware of the patient's diagnosis? Did the investigators consider various ways of subdividing the patients before choosing the particular diagnostic categories used in Table 1.1.4? At first glance, these questions may seem irrelevant—can we not let the measurements speak for themselves? We will see, however, that the proper interpretation of data always requires careful consideration of how the data were obtained.

Sections 1.2 and 1.3. as well as Chapters 2 and 8, include discussions of selection of experimental subjects and of guarding against unconscious investigator bias. In Chapter 11 we will show how sifting through a data set in search of patterns can lead to serious misinterpretations and we will give guidelines for avoiding the pitfalls in such searches.

The next example shows how the effects of variability can distort the results of an experiment and how this distortion can be minimized by careful design of the experiment.

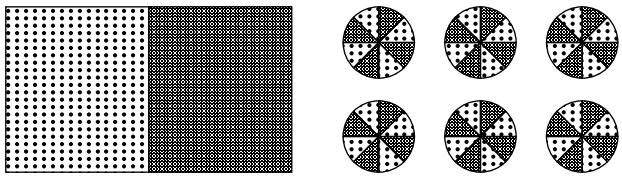
Example 1.1.5

Food Choice by Insect Larvae The clover root curculio, *Sitona hispidulus*, is a root-feeding pest of alfalfa. An entomologist conducted an experiment to study food choice by *Sitona* larvae. She wished to investigate whether larvae would preferentially choose alfalfa roots that were nodulated (their natural state) over roots whose nodulation had been suppressed. Larvae were released in a dish where both nodulated and nonnodulated roots were available. After 24 hours, the investigator counted the larvae that had clearly made a choice between root types. The results are shown in Table 1.1.5.[5]

The data in Table 1.1.5 appear to suggest rather strongly that *Sitona* larvae prefer nodulated roots. But our description of the experiment has obscured an important point—we have not stated how the roots were arranged. To see the relevance of the arrangement, suppose the experimenter had used only one dish, placing all the nodulated roots on one side of the dish and all the nonnodulated roots on the other side, as shown in Figure 1.1.3(a), and had then released 120 larvae in the center of the dish. This experimental arrangement would be seriously deficient, because the data of Table 1.1.5 would then permit several competing interpretations—for instance, (a) perhaps the larvae really do prefer nodulated roots; or (b) perhaps the two sides of the dish were at slightly different temperatures and the larvae were responding to temperature rather than nodulation; or (c) perhaps one larva chose the nodulated roots just by chance and the other larvae followed its trail. Because of these possibilities the experimental arrangement shown in Figure 1.1.3(a) can yield only weak information about larval food preference.

Table 1.1.5 Food choice by *Sitona* larvae

Choice	Number of larvae
Chose nodulated roots	46
Chose nonnodulated roots	12
Other (no choice, died, lost)	62
Total	120



(a) (b)

Figure 1.1.3 Possible arrangements of food choice experiment. The dark-shaded areas contain nodulated roots and the light-shaded areas contain nonnodulated roots.
(a) A poor arrangement.
(b) A good arrangement.

The experiment was actually arranged as in Figure 1.1.3(b), using six dishes with nodulated and nonnodulated roots arranged in a symmetric pattern. Twenty larvae were released into the center of each dish. This arrangement avoids the pitfalls of the arrangement in Figure 1.1.3(a). Because of the alternating regions of nodulated and nonnodulated roots, any fluctuation in environmental conditions (such as temperature) would tend to affect the two root types equally. By using several dishes, the experimenter has generated data that can be interpreted even if the larvae do tend to follow each other. To analyze the experiment properly, we would need to know the results in each dish; the condensed summary in Table 1.1.5 is not adequate. ■

In Chapter 11 we will describe various ways of arranging experimental material in space and time so as to yield the most informative experiment, as well as how to analyze the data to extract as much information as possible and yet resist the temptation to overinterpret patterns that may represent only random variation.

The following example is a study of the relationship between two measured quantities.

Example 1.1.6

Body Size and Energy Expenditure How much food does a person need? To investigate the dependence of nutritional requirements on body size, researchers used underwater weighing techniques to determine the fat-free body mass for each of seven men. They also measured the total 24-hour energy expenditure during conditions of quiet sedentary activity; this was repeated twice for each subject. The results are shown in Table 1.1.6 and plotted in Figure 1.1.4.[6]

Table 1.1.6 Fat-free mass and energy expenditure

Subject	Fat-free mass (kg)	24-hour energy expenditure (kcal)	
1	49.3	1,851	1,936
2	59.3	2,209	1,891
3	68.3	2,283	2,423
4	48.1	1,885	1,791
5	57.6	1,929	1,967
6	78.1	2,490	2,567
7	76.1	2,484	2,653

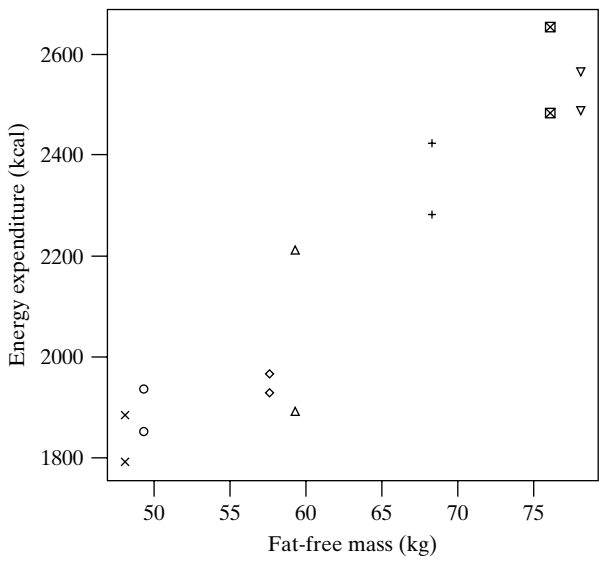


Figure 1.1.4 Fat-free mass and energy expenditure in seven men. Each man is represented by a different symbol.

A primary goal in the analysis of these data would be to describe the relationship between fat-free mass and energy expenditure—to characterize not only the overall trend of the relationship, but also the degree of scatter or variability in the relationship. (Note also that, to analyze the data, one needs to decide how to handle the duplicate observations on each subject.) ■

The focus of Example 1.1.6 is on the relationship between two variables: fat-free mass and energy expenditure. Chapter 12 deals with methods for describing such relationships, and also for quantifying the reliability of the descriptions.

A LOOK AHEAD

Where appropriate, statisticians make use of the computer as a tool in data analysis; computer-generated output and statistical graphics appear throughout this book. The computer is a powerful tool, but it must be used with caution. Using the computer to perform calculations allows us to concentrate on concepts. The danger when using a computer in statistics is that we will jump straight to the calculations without looking closely at the data and asking the right questions about the data. Our goal is to analyze, understand, and interpret data—which are numbers *in a specific context*—not just to perform calculations.

In order to understand a data set it is necessary to know how and why the data were collected. In addition to considering the most widely used methods in statistical inference, we will consider issues in data collection and experimental design. Together, these topics should provide the reader with the background needed to read the scientific literature and to design and analyze simple research projects.

The preceding examples illustrate the kind of data to be considered in this book. In fact, each of the examples will reappear as an exercise or example in an appropriate chapter. As the examples show, research in the life sciences is usually concerned with the comparison of two or more groups of observations, or with the relationship between two or more variables. We will begin our study of statistics by focusing on a simpler situation—observations of a *single* variable for a *single* group. Many of the basic ideas of statistics will be introduced in this oversimplified context. Two-group comparisons and more complicated analyses will then be discussed in Chapter 7 and later chapters.

1.2 Types of Evidence

Researchers gather information and make inferences about the state of nature in a variety of settings. Much of statistics deals with the *analysis* of data, but statistical considerations often play a key role in the planning and *design* of a scientific investigation. We begin with examples of the three major kinds of evidence that one encounters.

Example 1.2.1

Lightning and Deafness On 15 July 1911, 65-year-old Mrs. Jane Decker was struck by lightning while in her house. She had been deaf since birth, but after being struck, she recovered her hearing, which led to a headline in the *New York Times*, "Lightning Cures Deafness."[7] Is this compelling evidence that lightning is a cure for deafness? Could this event have been a coincidence? Are there other explanations for her cure? ■

The evidence discussed in Example 1.2.1 is **anecdotal evidence**. An anecdote is a short story or an example of an interesting event, in this case, of lightning curing deafness. The accumulation of anecdotes often leads to conjecture and to scientific investigation, but it is predictable pattern, not anecdote, that establishes a scientific theory.

Example 1.2.2

Sexual Orientation Some research has suggested that there is a genetic basis for sexual orientation. One such study involved measuring the midsagittal area of the anterior commissure (AC) of the brain for 30 homosexual men, 30 heterosexual men, and 30 heterosexual women. The researchers found that the AC tends to be larger in heterosexual women than in heterosexual men and that it is even larger in homosexual men. These data are summarized in Table 1.2.1 and are shown graphically in Figure 1.2.1.

Table 1.2.1 Midsagittal area of the anterior commissure (mm^2)

Group	Average midsagittal area (mm^2) of the anterior commissure
Homosexual men	14.20
Heterosexual men	10.61
Heterosexual women	12.03

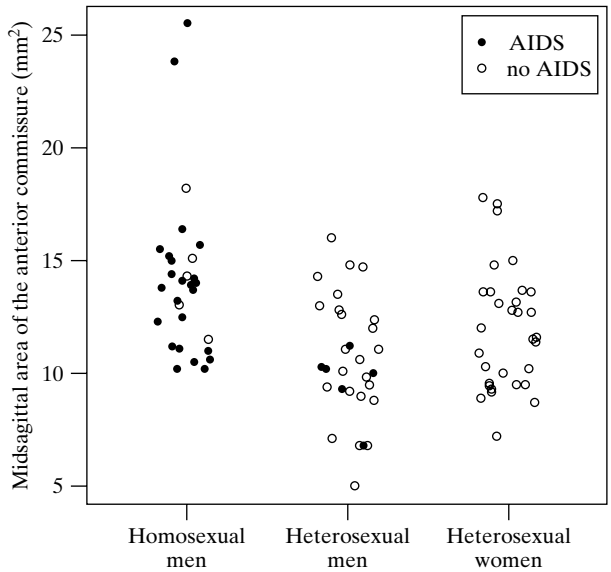


Figure 1.2.1 Midsagittal area of the anterior commissure (mm^2)

The data suggest that the size of the AC in homosexual men is more like that of heterosexual women than that of heterosexual men. When analyzing these data, we should take into account two things. (1) The measurements for two of the homosexual men were much larger than any of the other measurements; sometimes one or two such outliers can have a big impact on the conclusions of a study. (2) Twenty-four of the 30 homosexual men had AIDS, as opposed to 6 of the 30 heterosexual men; if AIDS affects the size of the anterior commissure, then this factor could account for some of the difference between the two groups of men.[8] ■

Example 1.2.2 presents an **observational study**. In an observational study the researcher systematically collects data from subjects, but only as an observer and not as someone who is manipulating conditions. By systematically examining all the data that arise in observational studies, one can guard against selectively viewing and reporting only evidence that supports a previous view. However, observational studies can be misleading due to *confounding variables*. In Example 1.2.2 we noted that having AIDS may affect the size of the anterior commissure. We would say that the effect of AIDS is confounded with the effect of sexual orientation in this study.

Note that the *context* in which the data arose is of central importance in statistics. This is quite clear in Example 1.2.2. The numbers themselves can be used to compute averages or to make graphs, like Figure 1.2.1, but if we are to understand what the data have to say, we must have an understanding of the context in which they arose. This context tells us to be on the alert for the effects that other factors, such as the impact of AIDS, may have on the size of the anterior commissure. Data analysis without reference to context is meaningless.

Example 1.2.3

Health and Marriage A study conducted in Finland found that people who were married at midlife were less likely to develop cognitive impairment (particularly Alzheimer's disease) later in life.[9] However, from an observational study such as this we don't know whether marriage *prevents* later problems or whether persons who are likely to develop cognitive problems are less likely to get married. ■

Example 1.2.4

Toxicity in Dogs Before new drugs are given to human subjects, it is common practice to first test them in dogs or other animals. In part of one study, a new investigational drug was given to eight male and eight female dogs at doses of 8 mg/kg and 25 mg/kg. Within each sex, the two doses were assigned at random to the eight dogs. Many "endpoints" were measured, such as cholesterol, sodium, glucose, and so on, from blood samples, in order to screen for toxicity problems in the dogs before starting studies on humans. One endpoint was alkaline phosphatase level (or APL, measured in U/l). The data are shown in Table 1.2.2 and plotted in Figure 1.2.2.[10]

Table 1.2.2 Alkaline phosphatase level (U/l)

Dose (mg/kg)	Male	Female
8	171	150
	154	127
	104	152
	143	105
Average	**143**	**133.5**
25	80	101
	149	113
	138	161
	131	197
Average	**124.5**	**143**

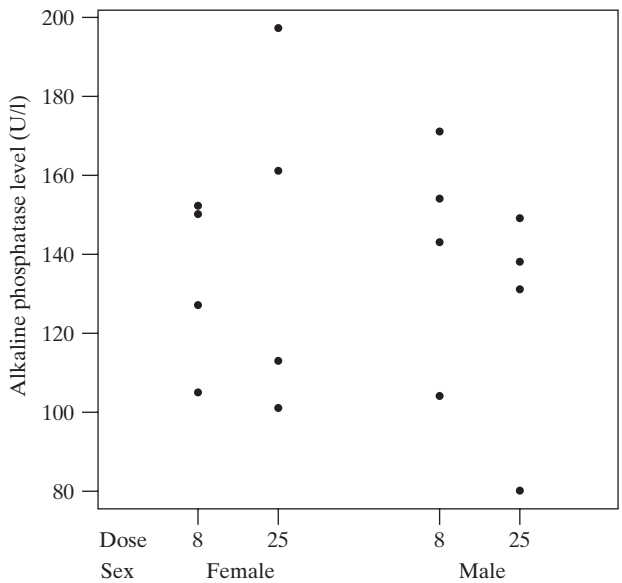


Figure 1.2.2 Alkaline phosphatase level in dogs

The design of this experiment allows for the investigation of the interaction between two factors: sex of the dog and dose. These factors interacted in the following sense: For females, the effect of increasing the dose from 8 to 25 mg/kg was positive, although small (the average APL increased from 133.5 to 143 U/l), but for males the effect of increasing the dose from 8 to 25 mg/kg was negative (the average APL dropped from 143 to 124.5 U/l). Techniques for studying such interactions will be considered in Chapter 11. ■

Example 1.2.4 presents an **experiment**, in that the researchers imposed the conditions—in this case, doses of a drug—on the subjects (the dogs). By randomly assigning treatments (drug doses) to subjects (dogs), we can get around the problem of confounding that complicates observational studies and limits the conclusions that we can reach from them. Randomized experiments are considered the "gold standard" in scientific investigation, but they can also be plagued by difficulties.

Often human subjects in experiments are given a **placebo**—an inert substance, such as a sugar pill. It is well known that people often exhibit a *placebo response*; that is, they tend to respond favorably to *any* treatment, even if it is only inert. This psychological effect can be quite powerful. Research has shown that placebos are effective for roughly one-third of people who are in pain; that is, one-third of pain sufferers report their pain ending after being giving a "painkiller" that is, in fact, an inert pill. For diseases such as bronchial asthma, angina pectoris (recurrent chest pain caused by decreased blood flow to the heart), and ulcers, the use of placebos has been shown to produce clinically beneficial results in over 60% of patients.[11] Of course, if a placebo control is used, then the subjects must not be told which group they are in—the group getting the active treatment or the group getting the placebo.

Example 1.2.5

Autism Autism is a serious condition in which children withdraw from normal social interactions and sometimes engage in aggressive or repetitive behavior. In 1997, an autistic child responded remarkably well to the digestive enzyme secretin. This led to an experiment (a "clinical trial") in which secretin was compared to a placebo. In this experiment, children who were given secretin improved considerably. However, the children given the placebo also improved considerably. There was no statistically significant difference between the two groups. Thus, the favorable response in the secretin group was considered to be only a "placebo response," meaning, unfortunately, that secretin was not found to be beneficial (beyond inducing a positive response associated simply with taking a substance as part of an experiment).[12] ■

The word *placebo* means "I shall please." The word *nocebo* ("I shall harm") is sometimes used to describe adverse reactions to perceived, but nonexistent, risks. The following example illustrates the strength that psychological effects can have.

Example 1.2.6

Bronchial Asthma A group of patients suffering from bronchial asthma were given a substance that they were told was a chest-constricting chemical. After being given this substance, several of the patients experienced bronchial spasms. However, during part of the experiment, the patients were given a substance that they were told would alleviate their symptoms. In this case, bronchial spasms were prevented. In reality, the second substance was identical to the first substance: Both were distilled water. It appears that it was the power of suggestion that brought on the bronchial spasms; the same power of suggestion prevented spasms.[13] ■

Similar to placebo treatment is *sham* treatment, which can be used on animals as well as humans. An example of sham treatment is injecting control animals with an inert substance such as saline. In some studies of surgical treatments, control animals (even, occasionally, humans) are given a "mock" surgery.

Example 1.2.7

Renal Denervation A surgical procedure called "renal denervation" was developed to help people with hypertension who do not respond to medication. An early study suggested that renal denervation (which uses radiotherapy to destroy some nerves in arteries feeding the kidney) reduces blood pressure. In that experiment, patients who received surgery had an average improvement in systolic blood pressure of 33 mmHg more than did control patients who received no surgery. Later an experiment was conducted in which patients were randomly assigned to one of two groups. Patients in

the treatment group received the renal denervation surgery. Patients in the control group received a sham operation in which a catheter was inserted, as in the real operation, but 20 minutes later the catheter was removed *without* radiotherapy being used. These patients had no way of knowing that their operation was a sham. The rates of improvement in the two groups of patients were nearly identical.[14] ■

BLINDING

In experiments on humans, particularly those that involve the use of placebos, **blinding** is often used. This means that the treatment assignment is kept secret from the experimental subject. The purpose of blinding the subject is to minimize the extent to which his or her expectations influence the results of the experiment. If subjects exhibit a psychological reaction to getting a medication, that placebo response will tend to balance out between the two groups so that any difference between the groups can be attributed to the effect of the active treatment.

In many experiments the persons who evaluate the responses of the subjects are also kept blind; that is, during the experiment they are kept ignorant of the treatment assignment. Consider, for instance, the following:

> In a study to compare two treatments for lung cancer, a radiologist reads X-rays to evaluate each patient's progress. The X-ray films are coded so that the radiologist cannot tell which treatment each patient received.
>
> Mice are fed one of three diets; the effects on their liver are assayed by a research assistant who does not know which diet each mouse received.

Of course, *someone* needs to keep track of which subject is in which group, but that person should not be the one who measures the response variable. The most obvious reason for blinding the person making the evaluations is to reduce the possibility of subjective bias influencing the observation process itself: Someone who *expects* or *wants* certain results may unconsciously influence those results. Such bias can enter even apparently "objective" measurements through subtle variation in dissection techniques, titration procedures, and so on.

In medical studies of human beings, blinding often serves additional purposes. For one thing, a patient must be asked whether he or she consents to participate in a medical study. Suppose the physician who asks the question already knows which treatment the patient will receive. By discouraging certain patients and encouraging others, the physician can (consciously or unconsciously) create noncomparable treatment groups. The effect of such biased assignment can be surprisingly large, and it has been noted that it generally favors the "new" or "experimental" treatment.[15] Another reason for blinding in medical studies is that a physician may (consciously or unconsciously) provide more psychological encouragement, or even better care, to the patients who are receiving the treatment that the physician regards as superior.

An experiment in which both the subjects and the persons making the evaluations of the response are blinded is called a **double-blind** experiment. The first mammary artery ligation experiment described in Example 1.2.7 was conducted as a double-blind experiment.

THE NEED FOR CONTROL GROUPS

Example 1.2.8

Clofibrate An experiment was conducted in which subjects were given the drug clofibrate, which was intended to lower cholesterol and reduce the chance of death from coronary disease. The researchers noted that many of the subjects did not take all the medication that the experimental protocol called for them to take. They

calculated the percentage of the prescribed capsules that each subject took and divided the subjects into two groups according to whether or not the subjects took at least 80% of the capsules they were given. Table 1.2.3 shows that the 5-year mortality rate for those who took at least 80% of their capsules was much lower than the corresponding rate for subjects who took fewer than 80% of the capsules. On the surface, this suggests that taking the medication lowers the chance of death. However, there was a placebo control group in the experiment and many of the placebo subjects took fewer than 80% of their capsules. The mortality rates for the two placebo groups—those who adhered to the protocol and those who did not—are quite similar to the rates for the clofibrate groups.

Table 1.2.3 Mortality rates for the clofibrate experiment

	Clofibrate		Placebo	
Adherence	n	5-year mortality	n	5-year mortality
≥80%	708	15.0%	1813	15.1%
<80%	357	24.6%	882	28.2%

The clofibrate experiment seems to indicate that there are two kinds of subjects: those who adhere to the protocol and those who do not. The first group had a much lower mortality rate than the second group. This might be due simply to better health habits among people who show stronger adherence to a scientific protocol for 5 years than among people who only adhere weakly, if at all. A further conclusion from the experiment is that clofibrate does not appear to be any more effective than placebo in reducing the death rate. Were it not for the presence of the placebo control group, the researchers might well have drawn the wrong conclusion from the study and attributed the lower death rate among strong adherers to clofibrate itself, rather than to other confounded effects that make the strong adherers different from the nonadherers.[16] ■

Example 1.2.9

The Common Cold Many years ago, investigators invited university students who believed themselves to be particularly susceptible to the common cold to be part of an experiment. Volunteers were randomly assigned to either the treatment group, in which case they took capsules of an experimental vaccine, or to the control group, in which case they were told that they were taking a vaccine, but in fact were given a placebo—capsules that looked like the vaccine capsules but that contained lactose in place of the vaccine.[17] As shown in Table 1.2.4, both groups reported having dramatically fewer colds during the study than they had had in the previous year. The average number of colds per person dropped 70% in the treatment group. This would have been startling evidence that the vaccine had an effect, except that the corresponding drop in the control group was 69%. ■

Table 1.2.4 Number of colds in cold-vaccine experiment

	Vaccine	Placebo
n	201	203
Average number of colds		
Previous year (from memory)	5.6	5.2
Current year	1.7	1.6
% reduction	70%	69%

We can attribute much of the large drop in colds in Example 1.2.9 to the placebo effect. However, another statistical concern is **panel bias**, which is bias attributable to the study having influenced the behavior of the subjects—that is, people who know they are being studied often change their behavior. The students in this study reported from memory the number of colds they had suffered in the previous year. The fact that they were part of a study might have influenced their behavior so that they were less likely to catch a cold during the study. Being in a study might also have affected the way in which they defined having a cold—during the study, they were "instructed to report to the health service whenever a cold developed"—so that some illness may have gone unreported during the study. (How sick do you have to be before you classify yourself as having a cold?)

Example 1.2.10

Diet and Cancer Prevention A diet that is high in fruits and vegetables may yield many health benefits, but how can we be sure? During the 1990s, the medical community believed that such a diet would reduce the risk of cancer. This belief was based on comparisons from **case-control studies**. In such studies patients with cancer were matched with "control subjects"—persons of the same age, race, sex, and so on—who did not have cancer; then the diets of the two groups were compared, and it was found that the control patients ate more fruits and vegetables than did the cancer patients. This would seem to indicate that cancer rates go down as consumption of fruits and vegetables goes up. The use of case-control studies is quite sensible because it allows researchers to make comparisons (e.g., of diets, etc.) while taking into consideration important characteristics such as age.

Nonetheless, a case-control study is not perfect. Not all people agree to be interviewed and to complete health information surveys, and these individuals thus might be excluded from a case-control study. People who agree to be interviewed about their health are generally more healthy than those who decline to participate. In addition to eating more fruits and vegetables than the average person, they are also less likely to smoke and more likely to exercise.[18] Thus, even though case-control studies took into consideration age, race, and other characteristics, they overstated the benefits of fruits and vegetables. The observed benefits are likely also the result of other healthy lifestyle factors.* Drawing a cause–effect conclusion that fruit and vegetable consumption protects against cancer is dangerous. ■

HISTORICAL CONTROLS

Researchers may be particularly reluctant to use randomized allocation in medical experiments on human beings. Suppose, for instance, that researchers want to evaluate a promising new treatment for a certain illness. It can be argued that it would be unethical to withhold the treatment from any patients, and that therefore all current patients should receive the new treatment. But then who would serve as a control group? One possibility is to use historical controls—that is, previous patients with the same illness who were treated with another therapy. One difficulty with historical controls is that there is often a tendency for later patients to show a better response—even to the same therapy—than earlier patients with the same diagnosis. This tendency has been confirmed, for instance, by comparing experiments conducted at the same medical centers in different years.[19] One major reason for the tendency is that the overall characteristics of the patient population may change with time. For

*A more informative kind of study is a prospective study or cohort study in which people with varying diets are followed over time to see how many of them develop cancer; however, such a study can be difficult to carry out.

instance, because diagnostic techniques tend to improve, patients with a given diagnosis (say, breast cancer) in 2001 may have a better chance of recovery (even with the same treatment) than those with the same diagnosis in 1991 because they were diagnosed earlier in the course of the disease. This is one reason that patients diagnosed with kidney cancer in 1995 had a 61% chance of surviving for at least 5 years but those with the same diagnosis in 2005 had a 75% 5-year survival rate.[20]

Medical researchers do not agree on the validity and value of historical controls. The following example illustrates the importance of this controversial issue.

Example 1.2.11

Coronary Artery Disease Disease of the coronary arteries is often treated by surgery (such as bypass surgery), but it can also be treated with drugs only. Many studies have attempted to evaluate the effectiveness of surgical treatment for this common disease. In a review of 29 of these studies, each study was classified as to whether it used randomized controls or historical controls; the conclusions of the 29 studies are summarized in Table 1.2.5.[21]

Table 1.2.5 Coronary artery disease studies

	Conclusion about effectiveness of surgery		
Type of controls	Effective	Not effective	Total number of studies
Randomized	1	7	8
Historical	16	5	21

It would appear from Table 1.2.5 that enthusiasm for surgery is much more common among researchers who use historical controls than among those who use randomized controls. ■

Example 1.2.12

Healthcare Trials A medical intervention, such as a new surgical procedure or drug, will often be used at one time in a nonrandomized clinical trial and at another time in a clinical trial of patients with the same condition who are assigned to groups randomly. Nonrandomized trials, which include the use of historical controls, tend to overstate the effectiveness of interventions. One analysis of many pairs of studies found that the nonrandomized trial showed a larger intervention effect than the corresponding randomized trial 22 times out of 26 comparisons; see Table 1.2.6.[22] Researchers concluded that overestimates of effectiveness are "due to poorer prognosis in non-randomly selected control groups compared with randomly selected control groups."[23] That is, if you give a new drug to relatively healthy patients and compare them to very sick patients taking the standard drug, the new drug is going to look better than it really is.

Even when randomization is used, trials may or may not be run double-blind. A review of 250 controlled trials found that trials that were not run double-blind produced significantly larger estimates of treatment effects than did trials that were double-blind.[24] ■

Table 1.2.6 Randomized versus nonrandomized trials

	Larger estimate of effect of the (common) intervention		
	Not randomized	Randomized	Total
Number of studies	22	4	26

Proponents of the use of historical controls argue that statistical adjustment can provide meaningful comparison between a current group of patients and a group of historical controls; for instance, if the current patients are younger than the historical controls, then the data can be analyzed in a way that adjusts, or corrects, for the effect of age. Critics reply that such adjustment may be grossly inadequate.

The concept of historical controls is not limited to medical studies. The issue arises whenever a researcher compares current data with past data. Whether the data are from the lab, the field, or the clinic, the researcher must confront the question: Can the past and current results be meaningfully compared? One should always at least ask whether the experimental material, and/or the environmental conditions, may have changed enough over time to distort the comparison.

Exercises 1.2.1–1.2.10

1.2.1 Fluoridation of drinking water has long been a controversial issue in the United States. One of the first communities to add fluoride to their water was Newburgh, New York. In March 1944, a plan was announced to begin to add fluoride to the Newburgh water supply on April 1 of that year. During the month of April, citizens of Newburgh complained of digestive problems, which were attributed to the fluoridation of the water. However, there had been a delay in the installation of the fluoridation equipment so that fluoridation did not begin until May 2.[25] Explain how the placebo effect/nocebo effect is related to this example.

1.2.2 Olestra is a no-calorie, no-fat additive that is used in the production of some potato chips. After the Food and Drug Administration approved the use of olestra, some consumers complained that olestra caused stomach cramps and diarrhea. A randomized, double-blind experiment was conducted in which some subjects were given bags of potato chips made with olestra and other subjects were given ordinary potato chips. In the olestra group, 38% of the subjects reported having gastrointestinal symptoms. However, in the group given regular potato chips the corresponding percentage was 37%. (The two percentages are not statistically significantly different.)[26] Explain how the placebo effect/nocebo effect is related to this example. Also explain why it was important for this experiment to be double-blind.

1.2.3 **(Hypothetical)** In a study of acupuncture, patients with headaches are randomly divided into two groups. One group is given acupuncture and the other group is given aspirin. The acupuncturist evaluates the effectiveness of the acupuncture and compares it to the results from the aspirin group. Explain how lack of blinding biases the experiment in favor of acupuncture.

1.2.4 Randomized, controlled experiments have found that vitamin C is not effective in treating terminal cancer patients.[27] However, a 1976 research paper reported that terminal cancer patients given vitamin C survived much longer than did historical controls. The patients treated with vitamin C were selected by surgeons from a group of cancer patients in a hospital.[28] Explain how this experiment was biased in favor of vitamin C.

1.2.5 On 3 November 2009, the blog lifehacker.com contained a posting by an individual with chronic toenail fungus. He remarked that after many years of suffering and trying all sorts of cures, he resorted to sanding his toenail as thin as he could tolerate, followed by daily application of vinegar and hydrogen-peroxide-soaked bandaids on his toenail. He repeated the vinegar peroxide bandaging for 100 days. After this time his nail grew out and the fungus was gone. Using the language of statistics, what kind of evidence is this? Is this convincing evidence that this procedure is an effective cure of toenail fungus?

1.2.6 For each of the following cases [(a) (b)],

(I) state whether the study should be observational or experimental.

(II) state whether the study should be run blind, double-blind, or neither. If the study should be run blind or double-blind, who should be blinded?

(a) An investigation of whether taking aspirin reduces one's chance of having a heart attack.

(b) An investigation of whether babies born into poor families (family income below $25,000) are more likely to weigh less than 5.5 pounds at birth than babies born into wealthy families (family income above $65,000).

1.2.7 For each of the following cases [(a) and (b)],

(I) state whether the study should be observational or experimental.

(II) state whether the study should be run blind, double-blind, or neither. If the study should be run blind or double-blind, who should be blinded?

(a) An investigation of whether the size of the midsagittal plane of the anterior commissure

(a part of the brain) of a man is related to the sexual orientation of the man.

(b) An investigation of whether drinking more than 1 liter of water per day helps with weight loss for people who are trying to lose weight.

1.2.8 (Hypothetical) In order to assess the effectiveness of a new fertilizer, researchers applied the fertilizer to the tomato plants on the west side of a garden but did not fertilize the plants on the east side of the garden. They later measured the weights of the tomatoes produced by each plant and found that the fertilized plants grew larger tomatoes than did the nonfertilized plants. They concluded that the fertilizer works.

(a) Was this an experiment or an observational study? Why?

(b) This study is seriously flawed. Use the language of statistics to explain the flaw and how this affects the validity of the conclusion reached by the researchers.

(c) Could this study have used the concept of blinding (i.e., does the word "blind" apply to this study)? If so, how? Could it have been double-blind? If so, how?

1.2.9 Reseachers studied 1,718 persons over age 65 living in North Carolina. They found that those who attended religious services regularly were more likely to have strong immune systems (as determined by the blood levels of the protein interleukin-6) than those who didn't.[29] Does this mean that attending religious services improves one's health? Why or why not?

1.2.10 Researchers studied 300,818 golfers in Sweden and found that the "standardized mortality ratios" for golfers, adjusting for age, sex, and socioeconomic status, were lower than for nongolfers, meaning that golfers tend to live longer.[30] Does this mean that playing golf improves one's health? Why or why not?

1.3 Random Sampling

In order to address research questions with data, we first must consider how those data are to be gathered. How we gather our data has tremendous implications on our choice of analysis methods and even on the validity of our studies. In this section we will examine some common types of data-gathering methods with special emphasis on the **simple random sample**.

SAMPLES AND POPULATIONS

Before gathering data, we first consider the scope of our study by identifying the **population**. The population consists of all subjects/animals/specimens/plants, and so on, of interest. The following are all examples of populations:

- All birch tree seedlings in Florida
- All raccoons in Montaña de Oro State Park
- All people with schizophrenia in the United States
- All 100-ml water specimens in Chorro Creek

Typically we are unable to observe the entire population; therefore, we must be content with gathering data from a subset of the population, a **sample** of size n. From this sample we make inferences about the population as a whole (see Figure 1.3.1). The following are all examples of samples:

- A selection of eight ($n = 8$) Florida birch seedlings grown in a greenhouse.

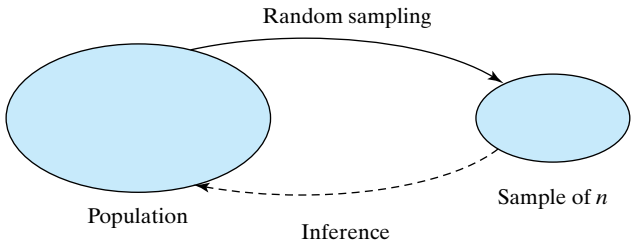


Figure 1.3.1 Sampling from a population

- Thirteen ($n = 13$) raccoons captured in traps at the Montaña de Oro campground.
- Forty-two ($n = 42$) patients with schizophrenia who respond to an advertisement in a U.S. newspaper.
- Ten ($n = 10$) 100-ml vials of water collected one day at 10 locations along Chorro Creek.

Remark There is some potential for confusion between the statistical meaning of the term *sample* and the sense in which this word is sometimes used in biology. If a biologist draws blood from 20 people and measures the glucose concentration in each, she might say she has 20 samples of blood. However, the statistician says she has *one* sample of 20 glucose measurements; the sample size is $n = 20$. In the interest of clarity, throughout this book we will use the term *specimen* where a biologist might prefer *sample*. So we would speak of glucose measurements on a sample of 20 specimens of blood.

Ideally our sample will be a representative subset of the population; however, unless we are careful, we may end up obtaining a **biased** sample. A biased sample systematically overestimates or systematically underestimates a characteristic of the population. For example, consider the raccoons from the sample described previously that are captured in traps at a campground. These raccoons may systematically differ from the population; they may be larger (from having ample access to food from dumpsters and campers), less timid (from being around people who feed them), and may be even longer lived than the general population of raccoons in the entire park.

One method to ensure that samples will be (in the long run) representative of the population is to use random sampling.

DEFINITION OF A SIMPLE RANDOM SAMPLE

Informally, the process of obtaining a simple random sample can be visualized in terms of labeled tickets, such as those used in a lottery or raffle. Suppose that each member of the population (e.g., raccoon, patient, plant) is represented by one ticket, and that the tickets are placed in a large box and thoroughly mixed. Then n tickets are drawn from the box by a blindfolded assistant, with new mixing after each ticket is removed. These n tickets constitute the sample. (Equivalently, we may visualize that n assistants reach in the box simultaneously, each assistant drawing one ticket.)

More abstractly, we may define random sampling as follows.

A Simple Random Sample

A *simple random sample* of n items is a sample in which (a) every member of the population has the same chance of being included in the sample, and (b) the members of the sample are chosen independently of each other. [Requirement (b) means that the chance of a given member of the population being chosen does not depend on which other members are chosen.]*

*Technically, requirement (b) is that every pair of members of the population has the same chance of being selected for the sample, every group of 3 members of the population has the same chance of being selected for the sample, and so on. In contrast to this, suppose we had a population with 30 persons in it and we wrote the names of 3 persons on each of 10 tickets. We could then choose one ticket in order to get a sample of size $n = 3$, but this would not be a simple random sample, since the pair (1,2) could end up in the sample but the pair (1,4) could not. Here the selections of members of the sample are not independent of each other. (This kind of sampling is known as "cluster sampling," with 10 clusters of size 3.) If the population is infinite, then the technical definition that all subsets of a given size are equally likely to be selected as part of the sample is equivalent to the requirement that the members of the sample are chosen independently.

Simple random sampling can be thought of in other, equivalent, ways. We may envision the sample members being chosen one at a time from the population; under simple random sampling, at each stage of the drawing, every remaining member of the population is equally likely to be the next one chosen. Another view is to consider the totality of possible samples of size n. If all possible samples are equally likely to be obtained, then the process gives a simple random sample.

EMPLOYING RANDOMNESS

When conducting statistical investigations, we will need to make use of randomness. As previously discussed, we obtain simple random samples randomly—every member of the population has the same chance of being selected. In Chapter 7 we shall discuss experiments in which we wish to compare the effects of different treatments on members of a sample. To conduct these experiments we will have to assign the treatments to subjects randomly—so that every subject has the same chance of receiving treatment A as they do treatment B.

Unfortunately, as a practical matter, humans are not very capable of mentally employing randomness. We are unable to eliminate unconscious bias that often leads us to systematically exclude or include certain individuals in our sample (or at least decrease or increase the chance of choosing certain individuals). For this reason, we must use external resources for selecting individuals when we want a random sample: mechanical devices such as dice, coins, and lottery tickets; electronic devices that produce random digits such as computers and calculators; or tables of random digits such as Table 1 in the back of this book. Although straightforward, using mechanical devices such as tickets in a box is impractical, so we will focus on the use of random digits for sample selection.

HOW TO CHOOSE A RANDOM SAMPLE

The following is a simple procedure for choosing a random sample of n items from a finite population of items.

(a) Create the **sampling frame**: a list of all members of the population with unique identification numbers for each member. All identification numbers must have the same number of digits; for instance, if the population contains 75 items, the identification numbers could be 01, 02, . . . , 75.

(b) Read numbers from Table 1, a calculator, or computer. Reject any numbers that do not correspond to any population member. (For example, if the population has 75 items that have been assigned identification numbers 01, 02, . . . , 75, then skip over the numbers 76, 77, . . . , 99, and 00.) Continue until n numbers have been acquired. (Ignore any repeated occurrence of the same number.)

(c) The population members with the chosen identification numbers constitute the sample.

The following example illustrates this procedure.

Example 1.3.1

Suppose we are to choose a random sample of size 6 from a population of 75 members. Label the population members 01, 02, . . . , 75. Use Table 1, a calculator, or a computer to generate a string of random digits.* For example, our calculator might produce the following string:

8 3 8 7 1 7 9 4 0 1 6 2 5 3 4 5 9 7 5 3 9 8 2 2

*Most calculators generate random numbers expressed as decimal numbers between 0 and 1; to convert these to random digits, simply ignore the leading zero and decimal and read the digits that follow the decimal. To generate a long string of random digits, simply call the random number function on the calculator repeatedly.

As we examine two-digit pairs of numbers, we ignore numbers greater than 75 as well as any pairs that identify a previously chosen individual.

~~83~~ ~~87~~ 17 ~~94~~ 01 62 53 45 ~~97~~ ~~53~~ ~~98~~ 22

Thus, the population members with the following identification numbers will constitute the sample: 17, 01, 62, 53, 45, 22. ■

Remark In calling the digits in Table 1 or your calculator or computer *random* digits, we are using the term *random* loosely. Strictly speaking, random digits are digits produced by a random *process*—for example, tossing a 10-sided die. The digits in Table 1 or in your calculator or computer are actually *pseudorandom* digits; they are generated by a deterministic (although possibly very complex) process that is designed to produce sequences of digits that mimic randomly generated sequences.

Remark If the population is large, then computer software can be quite helpful in generating a sample. If you need a random sample of size 15 from a population with 2,500 members, have the computer (or calculator) generate 15 random numbers between 1 and 2,500. (If there are duplicates in the set of 15, then go back and get more random numbers.)

PRACTICAL CONCERNS WHEN RANDOM SAMPLING

In many cases, obtaining a proper simple random sample is difficult or impossible. For example, to obtain a random sample of raccoons from Montaña de Oro State Park, one would first have to create the sampling frame, which provides a unique number for each raccoon in the park. Then, after generating the list of random numbers to identify our sample, one would have to capture those particular raccoons. This is likely an impossible task.

In practice, when it is possible to obtain a proper random sample, one should. When a proper random sample is impractical, it is important to take all precautions to ensure that the subjects in the study may be viewed *as if* they were obtained by random sampling from some population. That is, the sample should be comprised of individuals that all have the same chance of being selected from the population, and the individuals should be chosen independently. To do this, the first step is to define the population. The next step is to scrutinize the procedure by which the observational units are selected and to ask: Could the *observations* have been chosen at random? With the raccoon example, this might mean that we first define the population of raccoons by creating a sharp geographic boundary based on raccoon habitat and place traps at randomly chosen locations within the population habitat using a variety of baits and trap sizes. (We could use random numbers to generate latitude and longitude coordinates within the population habitat.) Although still less than ideal (some raccoons might be trap shy, and baby raccoons may not enter the traps at all), this is certainly better than simply capturing raccoons at one nonrandomly chosen atypical location (e.g., the campground) within the park. Presumably, the vast majority of raccoons now have the same chance of being trapped (i.e., equally likely to be selected), and capturing one raccoon has little or no bearing on the capture of any other (i.e., they can be considered to be independently chosen). Thus, it seems reasonable to treat the observations as if they were chosen at random.

NONSIMPLE RANDOM SAMPLING METHODS

There are other kinds of sampling that are random in a sense, but that are not simple. Two common nonsimple random sampling techniques are the **random cluster sample**

and **stratified random sample**. To illustrate the concept of a cluster sample, consider a modification to the lottery method of generating a simple random sample. With cluster sampling, rather than assigning a unique ticket (or ID number) for each member of the population, IDs are assigned to entire groups of individuals. As tickets are drawn from the box, entire groups of individuals are selected for the sample as in the following example and Figure 1.3.2.

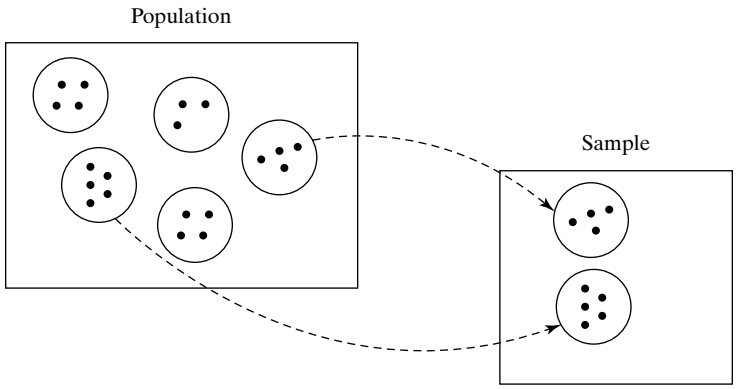


Figure 1.3.2 Random cluster sampling. The dots represent individuals within the population that are grouped into clusters (circles). Individuals in entire clusters are sampled from the population to form the sample.

Example 1.3.2

La Graciosa Thistle The La Graciosa thistle (*Cirsium loncholepis*) is an endangered plant native to the Guadalupe Dunes on the central coast of California. In a seed germination study, 30 plants were randomly chosen from the population of plants in the Guadalupe Dunes and all seeds from the 30 plants were harvested. The seeds form a cluster sample from the population of all La Graciosa thistle seeds in Guadalupe while the individual plants were used to identify the clusters.[31]

A stratified random sample is chosen by first dividing the population into **strata**—homogeneous collections of individuals. Then, many simple random samples are taken—one within each stratum—and combined to comprise the sample (see Figure 1.3.3). The following is an example of a stratified random sample.

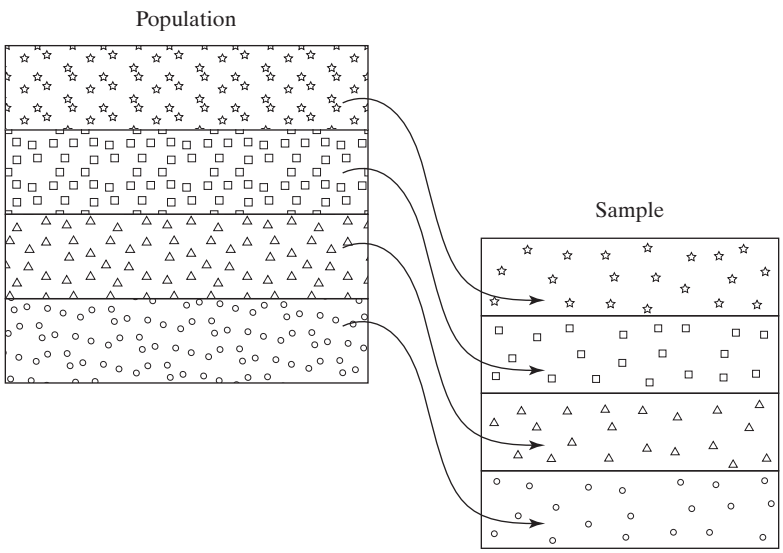


Figure 1.3.3 Stratified random sampling. The dots represent individuals within the population that are grouped into strata. Individuals from each stratum are randomly sampled and combined to form the sample.

Example 1.3.3

Sand Crabs In a study of parasitism of sand crabs (*Emerita analoga*), researchers obtained a stratified random sample of crabs by dividing a beach into 5-meter strips parallel to the water's edge. These strips were chosen as the strata because crab parasite loads may differ systematically based on the distance to the water's edge, thus making the parasite load for crabs within each stratum more similar than loads across strata. The first stratum was the 5-meter strip of beach just under the water's edge parallel to the shoreline. The second stratum was the 5-meter strip of beach just above the shoreline, followed by the third and fourth strata—the next two 5-meter strips above the shoreline. Within each strata, 25 crabs were randomly sampled, yielding a total sample size of 100 crabs.[32] ■

The majority of statistical methods discussed in this textbook will assume we are working with data gathered from a simple random sample. A sample chosen by simple random sampling is often called a *random sample.* But note that it is actually the *process* of sampling rather than the sample itself that is defined as random; randomness is not a property of the particular sample that happens to be chosen.

SAMPLING ERROR

How can we provide a rationale for inference from a limited sample to a much larger population? The approach of statistical theory is to refer to an idealized model of the sample–population relationship. In this model, which is called the **random sampling model**, the sample is chosen from the population by random sampling. The model is represented schematically in Figure 1.3.1.

The random sampling model is useful because it provides a basis for answering the question, How representative (of the population) is a sample likely to be? The model can be used to determine how much an inference might be influenced by chance, or "luck of the draw." More explicitly, a randomly chosen sample will usually not exactly resemble the population from which it was drawn. The discrepancy between the sample and the population is called **chance error due to sampling** or **sampling error**. We will see in later chapters how statistical theory derived from the random sampling model enables us to set limits on the likely amount of error due to sampling in an experiment. The quantification of such error is a major contribution that statistical theory has made to scientific thinking.

Because our samples are chosen randomly, there will always be sampling error present. If we sample nonrandomly, however, we may exacerbate the sampling error in unpredictable ways such as by introducing **sampling bias**, which is a systematic tendency for some individuals of the population to be selected more readily than others. The following two examples illustrate sampling bias.

Example 1.3.4

Lengths of Fish A biologist plans to study the distribution of body length in a certain population of fish in the Chesapeake Bay. The sample will be collected using a fishing net. Smaller fish can more easily slip through the holes in the net. Thus, smaller fish are less likely to be caught than larger ones, so the sampling procedure is biased. ■

Example 1.3.5

Sizes of Nerve Cells A neuroanatomist plans to measure the sizes of individual nerve cells in cat brain tissue. In examining a tissue specimen, the investigator must decide which of the hundreds of cells in the specimen should be selected for measurement. Some of the nerve cells are incomplete because the microtome cut through them when the tissue was sectioned. If the size measurement can be made only on

complete cells, a bias arises because the smaller cells had a greater chance of being missed by the microtome blade. ■

When the sampling procedure is biased, the sample may not accurately represent the population, because it is systematically distorted. For instance, in Example 1.3.4 smaller fish will tend to be underrepresented in the sample, so the length of the fish in the sample will tend to be larger than those in the population.

The following example illustrates a kind of nonrandomness that is different from bias.

Example 1.3.6

Sucrose in Beet Roots An agronomist plans to sample beet roots from a field in order to measure their sucrose content. Suppose she were to take all her specimens from a randomly selected small area of the field. This sampling procedure would not be biased but would tend to produce *too homogeneous* a sample, because environmental variation across the field would not be reflected in the sample. ■

Example 1.3.6 illustrates an important principle that is sometimes overlooked in the analysis of data: In order to check applicability of the random sampling model, one needs to ask not only whether the sampling procedure might be biased, but also whether the sampling procedure will adequately reflect the variability inherent in the population. Faulty information about variability can distort scientific conclusions just as seriously as bias can.

We now consider some examples where the random sampling model might reasonably be applied.

Example 1.3.7

Fungus Resistance in Corn A certain variety of corn is resistant to fungus disease. To study the inheritance of this resistance, an agronomist crossed the resistant variety with a nonresistant variety and measured the degree of resistance in the progeny plants. The actual progeny in the experiment can be regarded as a random sample from a conceptual population of all *potential* progeny of that particular cross. ■

When the purpose of a study is to *compare* two or more experimental conditions, a very narrow definition of the population may be satisfactory, as illustrated in the next example.

Example 1.3.8

Nitrite Metabolism To study the conversion of nitrite to nitrate in the blood, researchers injected four New Zealand White rabbits with a solution of radioactively labeled nitrite molecules. Ten minutes after injection, they measured for each rabbit the percentage of the nitrite that had been converted to nitrate.[33] Although the four animals were not literally chosen at random from a specified population, it might be reasonable, nevertheless, to view the measurements of nitrite metabolism as a random sample from similar measurements made on all New Zealand White rabbits. (This formulation assumes that age and sex are irrelevant to nitrite metabolism.) ■

Example 1.3.9

Treatment of Ulcerative Colitis A medical team conducted a study of two therapies, A and B, for treatment of ulcerative colitis. All the patients in the study were referral patients in a clinic in a large city. Each patient was observed for satisfactory "response" to therapy. In applying the random sampling model, the researchers might want to make an inference to the population of all ulcerative colitis patients in urban referral clinics. First, consider inference about the actual probabilities of response; such an inference would be valid if the probability of response to each therapy is the same at

all urban referral clinics. However, this assumption might be somewhat questionable, and the investigators might believe that the population should be defined very narrowly—for instance, as "the type of ulcerative colitis patients who are referred to this clinic." Even such a narrow population can be of interest in a comparative study. For instance, if treatment A is better than treatment B for the narrow population, it might be reasonable to infer that A would be better than B for a broader population (even if the actual response probabilities might be different in the broader population). In fact, it might even be argued that the broad population should include all ulcerative colitis patients, not merely those in urban referral clinics. ■

It often happens in research that, for practical reasons, the population actually studied is narrower than the population that is of real interest. In order to apply the kind of rationale illustrated in Example 1.3.9, one must argue that the results in the narrowly defined population (or, at least, some aspects of those results) can meaningfully be extrapolated to the population of interest. This extrapolation is not a *statistical* inference; it must be defended on biological, not statistical, grounds.

In Section 2.8 we will say more about the connection between samples and populations as we further develop the concept of statistical inference.

NONSAMPLING ERRORS

In addition to sampling errors, other concerns can arise in statistical studies. A **nonsampling error** is an error that is not caused by the sampling method; that is, a nonsampling error is one that would have arisen even if the researcher had a census of the entire population. For example, the way in which questions are worded can greatly influence how people answer them, as Example 1.3.10 shows.

Example 1.3.10

Abortion Funding In 1991, the U.S. Supreme Court made a controversial ruling upholding a ban on abortion counseling in federally financed family-planning clinics. Shortly after the ruling, a sample of 1,000 people were asked, "As you may know, the U.S. Supreme Court recently ruled that the federal government is not required to use taxpayer funds for family planning programs to perform, counsel, or refer for abortion as a method of family planning. In general, do you favor or oppose this ruling?" In the sample, 48% favored the ruling, 48% were opposed, and 4% had no opinion.

A separate opinion poll conducted at nearly the same time, but by a different polling organization, asked over 1,200 people, "Do you favor or oppose that Supreme Court decision preventing clinic doctors and medical personnel from discussing abortion in family-planning clinics that receive federal funds?" In this sample, 33% favored the decision and 65% opposed it.[34] The difference in the percentages favoring the opinion is too large to be attributed to chance error in the sampling. It seems that the way in which the question was worded had a strong impact on the respondents. ■

Another type of nonsampling error is **nonresponse bias**, which is bias caused by persons not responding to some of the questions in a survey or not returning a written survey. It is common to have only one-third of those receiving a survey in the mail complete the survey and return it to the researchers. (We consider the people receiving the survey to be part of the sample, even if some of them don't complete the entire survey, or even return the survey at all.) If the people who respond are unlike those who choose not to respond—and this is often the case, since people with strong feelings about an issue tend to complete a questionnaire, while others will ignore it—then the data collected will not accurately represent the population.

Example 1.3.11

HIV Testing A sample of 949 men were asked if they would submit to an HIV test of their blood. Of the 782 who agreed to be tested, 8 (1.02%) were found to be HIV positive. However, some of the men refused to be tested. The health researchers conducting the study had access to serum specimens that had been taken earlier from these 167 men and found that 9 of them (5.4%) were HIV positive.[35] Thus, those who refused to be tested were much more likely to have HIV than those who agreed to be tested. An estimate of the HIV rate based only on persons who agree to be tested is likely to substantially underestimate the true prevalence. ■

There are other cases in which an experimenter is faced with the vexing problem of **missing data**—that is, observations that were planned but could not be made. In addition to nonresponse, this can arise because experimental animals or plants die, because equipment malfunctions, or because human subjects fail to return for a follow-up observation.

A common approach to the problem of missing data is to simply use the remaining data and ignore the fact that some observations are missing. This approach is temptingly simple but must be used with extreme caution, because comparisons based on the remaining data may be seriously biased. For instance, if observations on some experimental mice are missing because the mice died of causes related to the treatment they received, it is obviously not valid to simply compare the mice that survived. As another example, if patients drop out of a medical study because they think their treatment is not working, then analysis of the remaining patients could produce a greatly distorted picture.

Naturally, it is best to make every effort to avoid missing data. But if data are missing, it is crucial that the possible reasons for the omissions be considered in interpreting and reporting the results.

Data can also be misleading if there is bias in how the data are collected. People have difficulty remembering the dates on which events happen and they tend to give unreliable answers if asked a question such as "How many times per week do you exercise?" They may also be biased as they make observations, as the following example shows.

Example 1.3.12

Sugar and Hyperactivity Mothers who thought that their young sons were "sugar sensitive" were randomly divided into two groups. Those in the first group were told that their sons had been given a large dose of sugar, whereas those in the second group were told that their sons had been given a placebo. In fact, all the boys had been given the placebo. Nonetheless, the mothers in the first group rated their sons to be much more hyperactive during a 25-minute study period than did the mothers in the second group.[36] Neutral measurements found that boys in the first group were actually a bit *less* active than those in the second group. Numerous other studies have failed to find a link between sugar consumption and activity in children, despite the widespread belief that sugar causes hyperactive behavior. It seems that the expectations that these mothers had colored their observations.[37] ■

Exercises 1.3.1–1.3.7

1.3.1 In each of the following studies, identify which sampling technique best describes the way the data were collected (or could be treated as if they were collected): simple random sampling, random cluster sampling, or stratified random sampling. For cluster samples identify the clusters, and for stratified samples identify the strata.

(a) All 257 leukemia patients from three randomly chosen pediatric clinics in the United States were enrolled in a clinical trial for a new drug.

(b) A total of twelve 10-g soil specimens were collected from random locations on a farm to study physical and chemical soil profiles.

(c) In a pollution study three 100-ml air specimens were collected at each of four specific altitudes (100 m, 500 m, 1000 m, 2000 m) for a total of twelve 100-ml specimens.

(d) A total of 20 individual grapes were picked, one from each of 20 random vines in a vineyard, to evaluate readiness for harvest.

(e) Twenty-four dogs (eight randomly chosen small breed, eight randomly chosen medium breed, and eight randomly chosen large breed) were enrolled in an experiment to evaluate a new training program.

1.3.2 For each of the following studies, identify the source(s) of sampling bias and describe (i) how it might affect the study conclusions and (ii) how you might alter the sampling method to avoid the bias.

(a) Eight hundred volunteers were recruited from nightclubs to enroll in an experiment to evaluate a new treatment for social anxiety.

(b) In a water pollution study, water specimens were collected from a stream on 15 rainy days.

(c) To study the size (radius) distribution of scrub oaks (shrubby oak trees), 20 oak trees were selected by using random latitude/longitude coordinates. If the random coordinate fell within the canopy of a tree, the tree was selected; if not, another random location was generated.

1.3.3 For each of the following studies, identify the source(s) of sampling bias and describe (i) how it might affect the study conclusions and (ii) how you might alter the sampling method to avoid the bias.

(a) To study the size distribution of rock cod (*Epinephelus puscus*) off the coast of southeastern Australia, scientists recorded the lengths and weights for all cod captured by a commercial fishing vessel on one day (using standard hook-and-line fishing methods).

(b) A nutritionist is interested in the eating habits of college students and observes what each student who enters a dining hall between 8:00 A.M. and 8:30 A.M. chooses for breakfast on a Monday morning.

(c) To study how fast an experimental painkiller relieves headache pain residents of a nursing home who complain of headaches are given the painkiller and are later asked how quickly their headaches subsided.

1.3.4 (A fun activity) Write the digits 1, 2, 3, 4 in order on an index card. Bring this card to a busy place (e.g., dining hall, library, university union) and ask at least 30 people to look at the card and select one of the digits at random in their head. Record their responses.

(a) If people can think "randomly," about what fraction of the people should respond with the digit 1? 2? 3? 4?

(b) What fraction of those surveyed responded with the digit 1? 2? 3? 4?

(c) Do the results suggest anything about people's ability to choose randomly?

1.3.5 Consider a population consisting of 600 individuals with unique IDs: 001, 002, . . . , 600. Use the following string of random digits to select a simple random sample of 5 individuals. List the IDs of the individuals selected for your sample.

7281218764421215937878035472165968 51

1.3.6 (Sampling exercise) Refer to the collection of 100 ellipses shown in the accompanying figure, which can be thought of as representing a natural population of the mythical organism *C. ellipticus*. The ellipses have been given identification numbers 00, 01, . . . , 99 for convenience in sampling. Certain individuals of *C. ellipticus* are mutants and have two tail bristles.

(a) Use your *judgment* to choose a sample of size 10 from the population that you think is representative of the entire population. Note the number of mutants in the sample.

(b) Use *random digits* (from Table 1 or your calculator or computer) to choose a random sample of size 10 from the population and note the number of mutants in the sample.

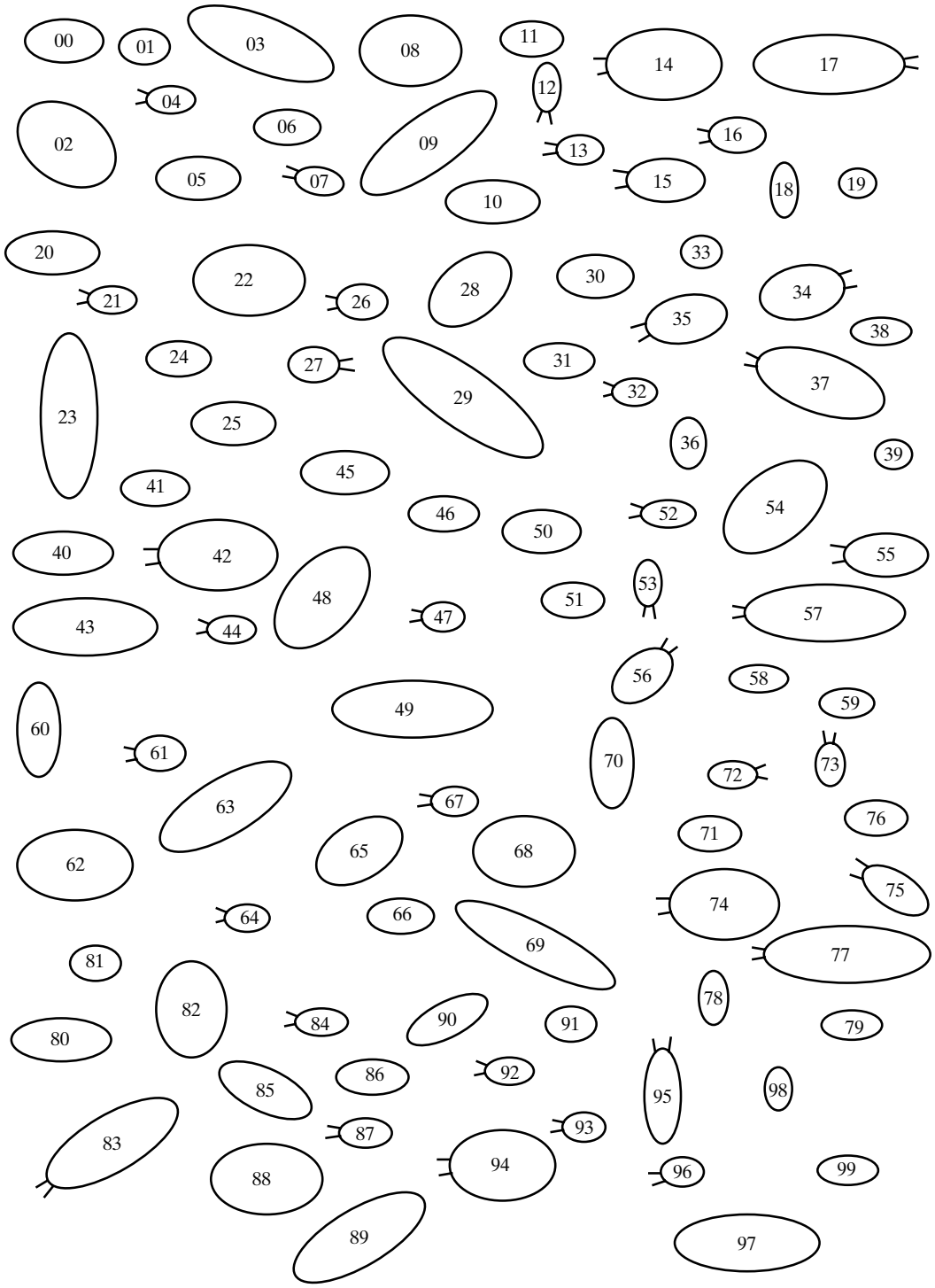
1.3.7 (Sampling exercise) Refer to the collection of 100 ellipses.

(a) Use random digits (from Table 1 or your calculator or computer) to choose a random sample of size 5 from the population and note the number of mutants in the sample.

(b) Repeat part (a) nine more times, for a total of 10 samples. (Some of the 10 samples may overlap.)

To facilitate pooling of results from the entire class, report your results in the following format:

Number of mutants	Nonmutants	Frequency (no. of samples)
0	5	
1	4	
2	3	
3	2	
4	1	
5	0	
		Total: 10


00
01
03
08
11
14
17
04
12
02
06
09
13
16
05
07
15
18
19
10
20
33
22
28
30
34
21
26
35
38
24
27
31
37
29
32
23
25
36
39
45
41
46
52
54
50
40
42
55
48
53
51
57
43
44
47
56
58
49
59
60
61
70
72
73
63
67
76
71
62
65
68
75
74
64
66
69
77
81
82
78
84
90
91
79
80
86
92
85
95
98
83
87
93
94
96
99
88
89
97

Description of Samples and Populations

Chapter 2

OBJECTIVES

In this chapter we will study how to describe data. In particular, we will

- show how frequency distributions are used to make bar charts and histograms.
- compare the mean and median as measures of center.
- demonstrate how to construct and read a variety of graphics including dotplots, boxplots, and scatterplots.
- compare several measures of variability with emphasis on the standard deviation.
- examine how transformations of variables affect distributions.
- consider the relationship between populations and samples.

2.1 Introduction

Statistics is the science of analyzing and learning from data. In this section we introduce some terminology and notation for dealing with data.

VARIABLES

We begin with the concept of a **variable**. A variable is a characteristic of a person or a thing that can be assigned a number or a category. For example, blood type (A, B, AB, O) and age are two variables we might measure on a person.

Blood type is an example of a **categorical variable***: A categorical variable is a variable that records which of several categories a person or thing is in. Examples of categorical variables are

Blood type of a person: A, B, AB, O
Sex of a fish: male, female
Color of a flower: red, pink, white
Shape of a seed: wrinkled, smooth

Age is an example of a **numeric variable**, that is, a variable that records the amount of something. A **continuous variable** is a numeric variable that is measured on a continuous scale. Examples of continuous variables are

Weight of a baby
Cholesterol concentration in a blood specimen
Optical density of a solution

A variable such as weight is continuous because, in principle, two weights can be arbitrarily close together. Some types of numeric variables are not continuous but fall on a discrete scale, with spaces between the possible values. A **discrete variable** is a numeric variable for which we can list the possible values. For example, the number of eggs in a bird's nest is a discrete variable because only the values $0, 1, 2, 3, \ldots,$ are possible. Other examples of discrete variables are

Number of bacteria colonies in a petri dish
Number of cancerous lymph nodes detected in a patient
Length of a DNA segment in basepairs

*For some categorical variables, the categories can be arrayed in a meaningful rank order. Such a variable is said to be **ordinal**. For example, the response of a patient to therapy might be none, partial, or complete.

The distinction between continuous and discrete variables is not a rigid one. After all, physical measurements are always rounded off. We may measure the weight of a steer to the nearest kilogram, of a rat to the nearest gram, or of an insect to the nearest milligram. The scale of the actual measurements is always discrete, strictly speaking. The continuous scale can be thought of as an approximation to the actual scale of measurement.

OBSERVATIONAL UNITS

When we collect a sample of n persons or things and measure one or more variables on them, we call these persons or things **observational units** or cases. The following are some examples of samples.

Sample	Variable	Observational unit
150 babies born in a certain hospital	Birthweight (kg)	A baby
73 *Cecropia* moths caught in a trap	Sex	A moth
81 plants that are a progeny of a single parental cross	Flower color	A plant
Bacterial colonies in each of six petri dishes	Number of colonies	A petri dish

NOTATION FOR VARIABLES AND OBSERVATIONS

We will adopt a notational convention to distinguish between a variable and an observed value of that variable. We will denote variables by uppercase letters such as Y. We will denote the observations themselves (that is, the data) by lowercase letters such as y. Thus, we distinguish, for example, between $Y =$ birthweight (the variable) and $y = 7.9$ lb (the observation). This distinction will be helpful in explaining some fundamental ideas concerning variability.

Exercises 2.1.1–2.1.5

For each of the following settings in Exercises 2.1.1–2.1.5, (i) identify the variable(s) in the study, (ii) for each variable tell the type of variable (e.g., categorical and ordinal, discrete, etc.), (iii) identify the observational unit (the thing sampled), and (iv) determine the sample size.

2.1.1

(a) A paleontologist measured the width (in mm) of the last upper molar in 36 specimens of the extinct mammal *Acropithecus rigidus*.

(b) The birthweight, date of birth, and the mother's race were recorded for each of 65 babies.

2.1.2

(a) A physician measured the height and weight of each of 37 children.

(b) During a blood drive, a blood bank offered to check the cholesterol of anyone who donated blood. A total of 129 persons donated blood. For each of them, the blood type and cholesterol levels were recorded.

2.1.3

(a) A biologist measured the number of leaves on each of 25 plants.

(b) A physician recorded the number of seizures that each of 20 patients with severe epilepsy had during an eight-week period.

2.1.4

(a) A conservationist recorded the weather (clear, partly cloudy, cloudy, rainy) and number of cars parked at noon at a trailhead on each of 18 days.

(b) An enologist measured the pH and residual sugar content (g/l) of seven barrels of wine.

2.1.5

(a) A biologist measured the body mass (g) and sex of each of 123 blue jays.

(b) A biologist measured the lifespan (in days), the thorax length (in mm), and the percent of time spent sleeping for each of 125 fruit flies.

2.2 Frequency Distributions

A first step toward understanding a set of data on a given variable is to explore the data and describe the data in summary form. In this chapter we discuss three mutually complementary aspects of data description: frequency distributions, measures of center, and measures of dispersion. These tell us about the shape, center, and spread of the data.

A **frequency distribution** is simply a display of the **frequency**, or number of occurrences, of each value in the data set. The information can be presented in tabular form or, more vividly, with a graph. A **bar chart** is a graph of categorical data showing the number of observations in each category. Here are two examples of frequency distributions for categorical data.

Example 2.2.1

Color of Poinsettias Poinsettias can be red, pink, or white. In one investigation of the hereditary mechanism controlling the color, 182 progeny of a certain parental cross were categorized by color.[1] The bar graph in Figure 2.2.1 is a visual display of the results given in Table 2.2.1. ■

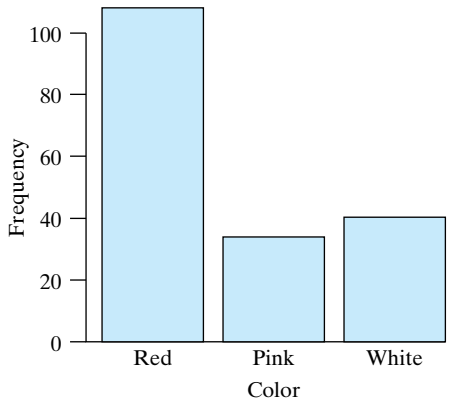


Figure 2.2.1 Bar chart of color of 182 poinsettias

Table 2.2.1 Color of 182 poinsettias

Color	Frequency (number of plants)
Red	108
Pink	34
White	40
Total	182

Example 2.2.2

School Bags and Neck Pain Physiologists in Australia were concerned that carrying a school bag loaded with heavy books was a cause of neck pain in adolescents, so they asked a sample of 585 teenage girls how often they get neck pain when carrying their school bag (never, almost never, sometimes, often, always). A summary of the results reported to them is given in Table 2.2.2 and displayed as a bar graph in Figure 2.2.2(a).[2] As the variable incidence is an ordinal categorical variable, our tables and graphs should respect the natural ordering. Figure 2.2.2(b) shows the same data but with the categories in alphabetical order (a default setting for much software), which obscures the information in the data. ■

Table 2.2.2 Neck pain associated with carrying a school bag

Incidence	Frequency (number of girls)
Never	179
Almost never	159
Sometimes	173
Often	64
Always	10
Total	585

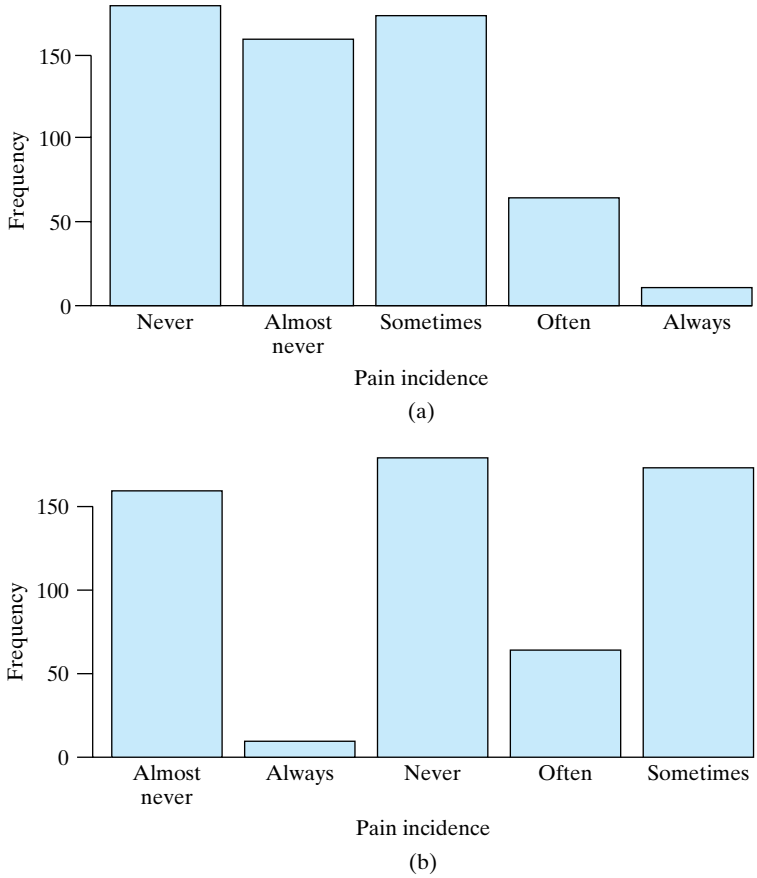


Figure 2.2.2 (a) Bar chart of incidence of neck pain reported by 585 adolescents; (b) the same data but with the categories in alphabetical order

A **dotplot** is a simple graph that can be used to show the distribution of a numeric variable when the sample size is small. To make a dotplot, we draw a number line covering the range of the data and then put a dot above the number line for each observation, as the following example shows.

Example 2.2.3

Infant Mortality Table 2.2.3 shows the infant mortality rate (infant deaths per 1,000 live births) in each of seven countries in South Asia, as of 2013.[3] The distribution is shown in Figure 2.2.3.

Table 2.2.3 Infant mortality in seven South Asian countries

Country	Infant mortality rate (deaths per 1,000 live births)
Bangladesh	47.3
Bhutan	40.0
India	44.6
Maldives	25.5
Nepal	41.8
Pakistan	59.4
Sri Lanka	9.2

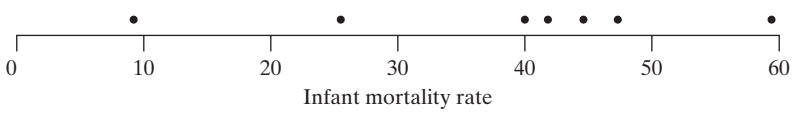


Figure 2.2.3 Dotplot of infant mortality in seven South Asian countries

When two or more observations take on the same value, we stack the dots in a dotplot on top of each other. This gives an effect similar to the effect of the bars in a bar chart. If we create bars in place of the stacks of dots, we then have a **histogram**. A histogram is like a bar chart, except that a histogram displays a numeric variable, which means that there is a natural order and scale for the variable. In a bar chart the amount of space between the bars (if any) is arbitrary, since the data being displayed are categorical. In a histogram the scale of the variable determines the placement of the bars. The following example shows a dotplot and a histogram for a frequency distribution.

Example 2.2.4

Litter Size of Sows A group of thirty-six 2-year-old sows of the same breed ($\frac{3}{4}$ Duroc, $\frac{1}{4}$ Yorkshire) were bred to Yorkshire boars. The number of piglets surviving to 21 days of age was recorded for each sow.[4] The results are given in Table 2.2.4 and displayed as a dotplot in Figure 2.2.4 and as a histogram in Figure 2.2.5. ■

Table 2.2.4 Number of surviving piglets of 36 sows

Number of piglets	Frequency (number of sows)
5	1
6	0
7	2
8	3
9	3
10	9
11	8
12	5
13	3
14	2
Total	36

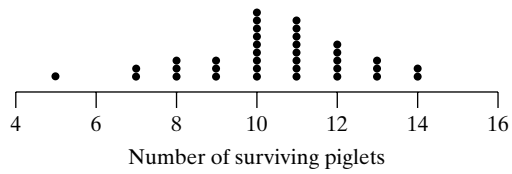


Figure 2.2.4 Dotplot of number of surviving piglets of 36 sows

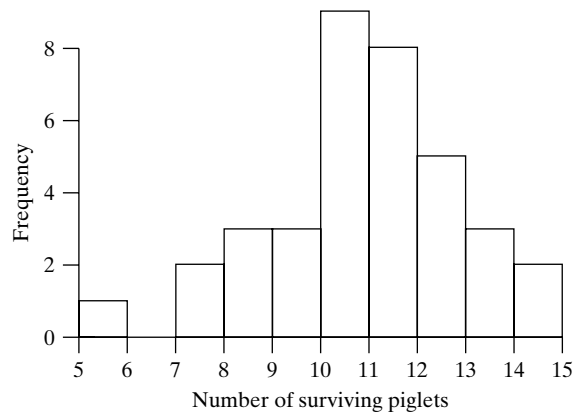


Figure 2.2.5 Histogram of number of surviving piglets of 36 sows

RELATIVE FREQUENCY

The frequency scale is often replaced by a **relative frequency** scale:

$$\text{Relative frequency} = \frac{\text{Frequency}}{n}$$

The relative frequency scale is useful if several data sets of different sizes (n's) are to be displayed together for comparison. As another option, a relative frequency can be expressed as a percentage frequency. The shape of the display is not affected by the choice of frequency scale, as the following example shows.

Example 2.2.5

Color of Poinsettias The poinsettia color distribution of Example 2.2.1 is expressed as frequency, relative frequency, and percent frequency in Table 2.2.5 and Figure 2.2.6.

Table 2.2.5 Color of 182 poinsettias

Color	Frequency	Relative frequency	Percent frequency
Red	108	.59	59
Pink	34	.19	19
White	40	.22	22
Total	182	1.00	100

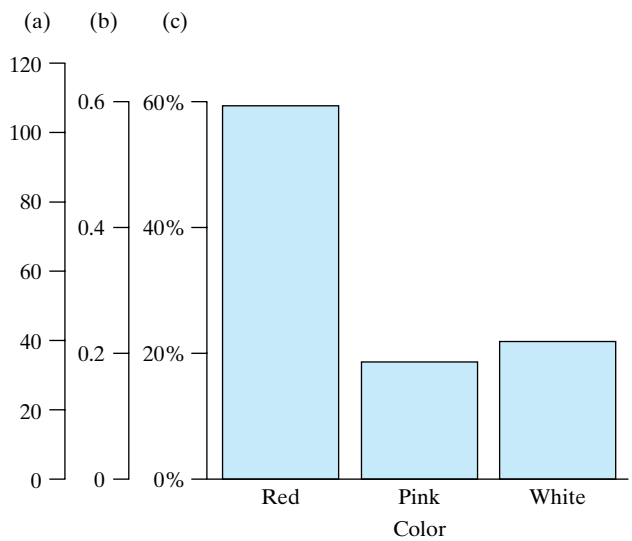


Figure 2.2.6 Bar chart of poinsettia colors on three scales:
(a) Frequency
(b) Relative frequency
(c) Percent frequency

GROUPED FREQUENCY DISTRIBUTIONS

In the preceding examples, simple ungrouped frequency distributions provided concise summaries of the data. For many data sets, it is necessary to group the data in order to condense the information adequately. (This is usually the case with continuous variables.) The following example shows a grouped frequency distribution.

Example 2.2.6

Serum CK Creatine phosphokinase (CK) is an enzyme related to muscle and brain function. As part of a study to determine the natural variation in CK concentration, blood was drawn from 36 male volunteers. Their serum concentrations of CK (measured in U/l) are given in Table 2.2.6.[5] Table 2.2.7 shows these data grouped into **classes**. For instance, the frequency of the class [20,40) (all values in the interval $20 \le y < 40$) is 1, which means that one CK value fell in this range. The grouped frequency distribution is displayed as a histogram in Figure 2.2.7.

Table 2.2.6 Serum CK values for 36 men

121	82	100	151	68	58
95	145	64	201	101	163
84	57	139	60	78	94
119	104	110	113	118	203
62	83	67	93	92	110
25	123	70	48	95	42

Table 2.2.7 Frequency distribution of serum CK values for 36 men

Serum CK (U/l)	Frequency (number of men)
[20,40)	1
[40,60)	4
[60,80)	7
[80,100)	8
[100,120)	8
[120,140)	3
[140,160)	2
[160,180)	1
[180,200)	0
[200,220)	2
Total	36

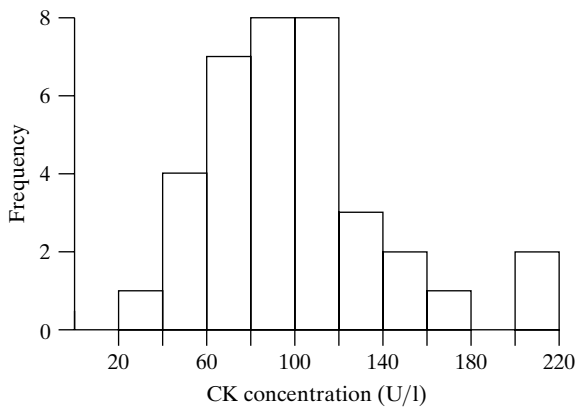
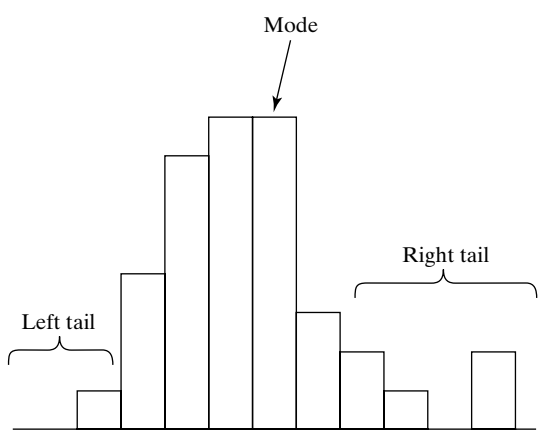


Figure 2.2.7 Histogram of serum CK concentrations for 36 men

A grouped frequency distribution should display the essential features of the data. For instance, the histogram of Figure 2.2.7 shows that the average CK value is about 100 U/l, with the majority of the values falling between 60 and 140 U/l. In addition, the histogram shows the *shape* of the distribution. Note that the CK values are piled up around a central peak, or **mode**. On either side of this mode, the frequencies decline and ultimately form the **tails** of the distribution. These shape features are labeled in Figure 2.2.8. The CK distribution is not symmetric but is a bit **skewed to the right**, which means that the right tail is more stretched out than the left.*

Figure 2.2.8 Shape features of the CK distribution



*To help remember which tail of a skewed distribution is the longer tail, think of skew as stretch. Which side of the distribution is more stretched away from the center? A distribution that is skewed to the right is one in which the right tail stretches out more than the left.

When making a histogram, we need to decide how many classes to have and how wide the classes should be. If we use computer software to generate a histogram, the program will choose the number of classes and the class width for us, but most software allows the user to change the number of classes and to specify the class width. If a data set is large and is quite spread out, it is a good idea to look at more than one histogram of the data, as is done in Example 2.2.7.

Example 2.2.7

Heights of Students A sample of 510 college students were asked how tall they were. Note that they were not measured; rather, they just reported their heights.[6] Figure 2.2.9 shows the distribution of the self-reported values, using 7 classes and a class width of 3 (inches). By using only 7 classes, the distribution appears to be reasonably symmetric, with a single peak around 66 inches.

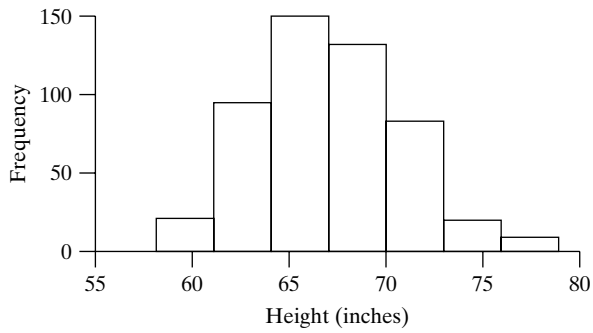


Figure 2.2.9 Heights of students, using 7 classes (class width = 3)

Figure 2.2.10 shows the height data, but in a histogram that uses 18 classes and a class width of 1.1. This view of the data shows two modes—one for women and one for men.

Figure 2.2.11 shows the height data again, this time using 37 classes, each of width 0.5. Using such a large number of classes makes the distribution look jagged. In this case, we see an alternating pattern between classes with lots of observations and classes with few observations. In the middle of the distribution we see that there were many students who reported a height of 63 inches, few who reported a height of 63.5 inches, many who reported a height of 64 inches, and so on. It seems that most students round off to the nearest inch!

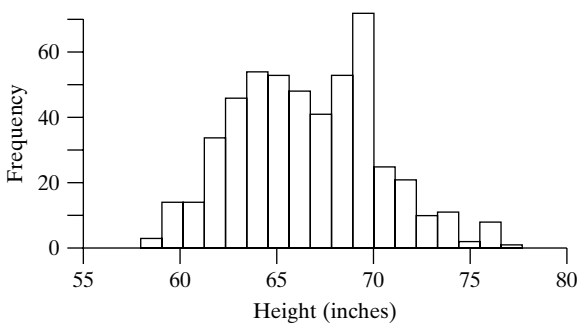


Figure 2.2.10 Heights of students, using 18 classes (class width = 1.1)

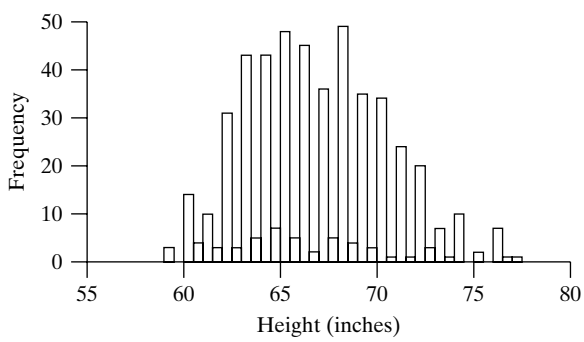


Figure 2.2.11 Heights of students, using 37 classes (class width = 0.5)

INTERPRETING AREAS IN A HISTOGRAM

A histogram can be looked at in two ways. The tops of the bars sketch out the shape of the distribution. But the *areas* within the bars also have a meaning. The area of each bar is proportional to the corresponding frequency. Consequently, the

area of one or several bars can be interpreted as expressing the number of observations in the classes represented by the bars. For example, Figure 2.2.12 shows a histogram of the CK distribution of Example 2.2.6. The shaded area is 42% of the total area in all the bars. Accordingly, 42% of the CK values are in the corresponding classes; that is, 15 of 36 or 42% of the values are between 60 U/I and 100 U/l.*

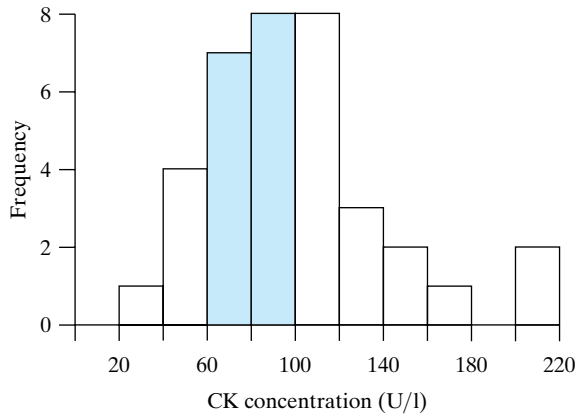


Figure 2.2.12 Histogram of CK distribution. The shaded area is 42% of the total area and represents 42% of the observations.

The area interpretation of histograms is a simple but important idea. In our later work with distributions we will find the idea to be indispensable.

SHAPES OF DISTRIBUTIONS

When discussing a set of data, we want to describe the shape, center, and spread of the distribution. In this section we concentrate on the shapes of frequency distributions and illustrate some of the diversity of distributions encountered in the life sciences. The shape of a distribution can be indicated by a smooth curve that approximates the histogram, as shown in Figure 2.2.13.

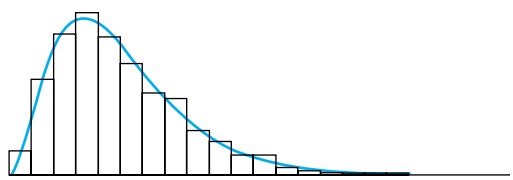


Figure 2.2.13 Approximation of a histogram by a smooth curve

Some distributional shapes are shown in Figure 2.2.14. A common shape for biological data is **unimodal** (has one mode) and is somewhat skewed to the right, as in (c). Approximately bell-shaped distributions, as in (a), also occur. Sometimes a distribution is symmetric but differs from a bell in having long tails; an exaggerated version is shown in (b). Left-skewed (d) and exponential (e) shapes are less common. **Bimodality** (two modes), as in (f), can indicate the existence of two distinct subgroups of observational units.

Notice that the shape characteristics we are emphasizing, such as number of modes and degree of symmetry, are *scale free;* that is, they are not affected by the arbitrary choices of vertical and horizontal scale in plotting the distribution. By contrast, a characteristic such as whether the distribution appears short and fat, or tall and skinny, is affected by how the distribution is plotted and so is not an inherent feature of the biological variable.

*Strictly speaking, between 60 U/l and 99 U/l, inclusive.

The following three examples illustrate biological frequency distributions with various shapes. In the first example, the shape provides evidence that the distribution is in fact biological rather than nonbiological.

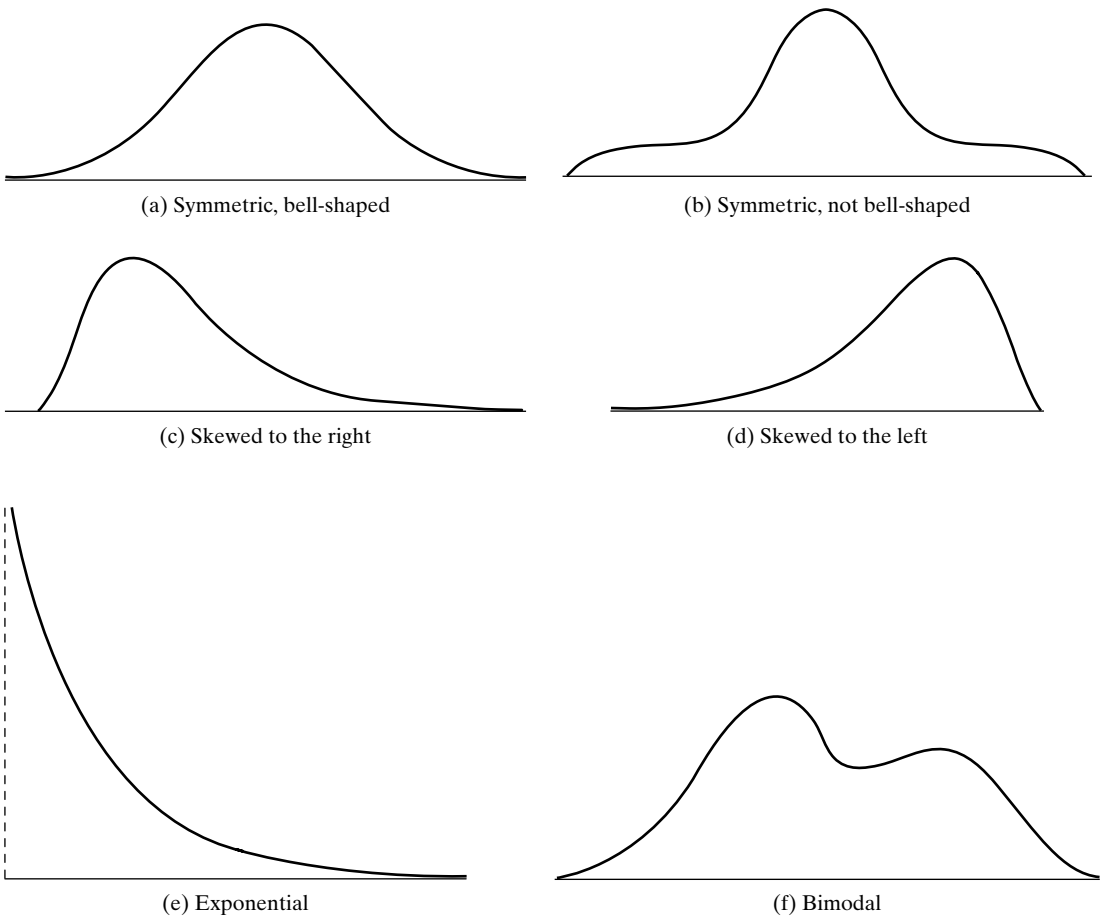
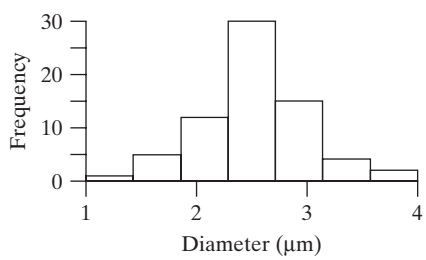


Figure 2.2.14 Shapes of distributions

Example 2.2.8

Microfossils In 1977, paleontologists discovered microscopic fossil structures, resembling algae, in rocks 3.5 billion years old. A central question was whether these structures were biological in origin. One line of argument focused on their size distribution, which is shown in Figure 2.2.15. This distribution, with its unimodal and rather symmetric shape, resembles that of known microbial populations, but not that of known nonbiological structures.[7]

Figure 2.2.15 Sizes of microfossils



Example 2.2.9

Cell Firing Times A neurobiologist observed discharges from rat muscle cells grown in culture together with nerve cells. The time intervals between 308 successive discharges were distributed as shown in Figure 2.2.16. Note the exponential shape of the distribution.[8]

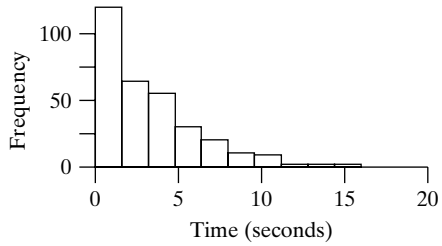


Figure 2.2.16 Time intervals between electrical discharges in rat muscle cells

Example 2.2.10

Brain Weight In 1888, P. Topinard published data on the brain weights of hundreds of French men and women. The data for males and females are shown in Figure 2.2.17(a) and (b). The male distribution is fairly symmetric and bell shaped; the female distribution is somewhat skewed to the right. Part (c) of the figure shows the brain weight distribution for males and females combined. This combined distribution is slightly bimodal.[9]

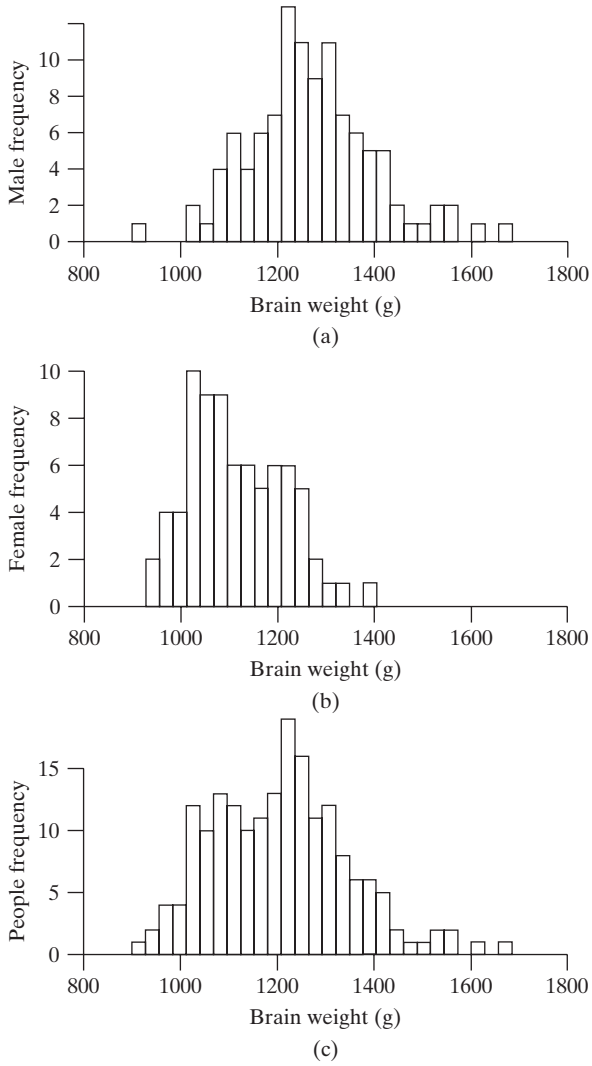


Figure 2.2.17 Brain weights

SOURCES OF VARIATION

In interpreting biological data, it is helpful to be aware of sources of variability. The variation among observations in a data set often reflects the combined effects of several underlying factors. The following two examples illustrate such situations.

Example 2.2.11

Weights of Seeds In a classic experiment to distinguish environmental from genetic influence, a geneticist weighed seeds of the princess bean *Phaseolus vulgaris*. Figure 2.2.18 shows the weight distributions of (a) 5,494 seeds from a commercial seed lot, and (b) 712 seeds from a highly inbred line that was derived from a single seed from the original lot. The variability in (a) is due to both environmental and genetic factors; in (b), because the plants are nearly genetically identical, the variation in weights is due largely to environmental influence.[10] Thus, there is less variability in the inbred line.

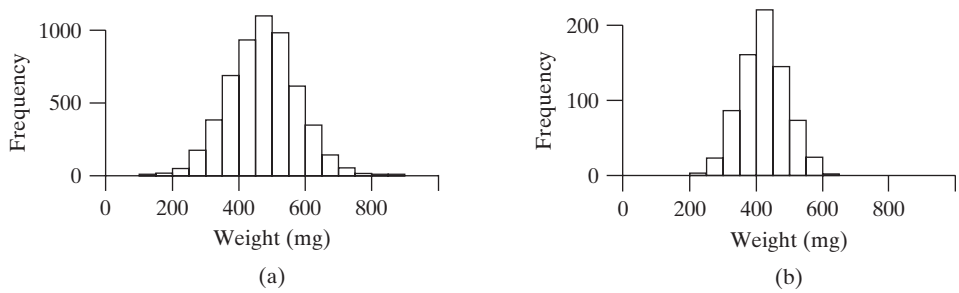


Figure 2.2.18 Weights of princess bean seeds: (a) from an open-bred population; (b) from an inbred line

Example 2.2.12

Serum ALT Alanine aminotransferase (ALT) is an enzyme found in most human tissues. Part (a) of Figure 2.2.19 shows the serum ALT concentrations for 129 adult volunteers. The following are potential sources of variability among the measurements:

1. Interindividual
   (a) Genetic
   (b) Environmental
2. Intraindividual
   (a) Biological: changes over time
   (b) Analytical: imprecision in assay

The effect of the last source—analytical variation—can be seen in part (b) of Figure 2.2.19, which shows the frequency distribution of 109 assays of the *same* specimen of serum; the figure shows that the ALT assay is fairly imprecise.[11]

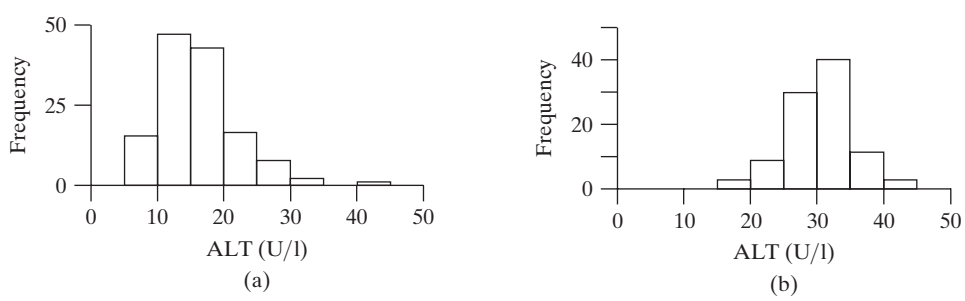


Figure 2.2.19 Distribution of serum ALT measurements (a) for 129 volunteers; (b) for 109 assays of the same specimen

Exercises 2.2.1–2.2.9

2.2.1 A paleontologist measured the width (in mm) of the last upper molar in 36 specimens of the extinct mammal *Acropithecus rigidus*. The results were as follows:[12]

6.1	5.7	6.0	6.5	6.0	5.7
6.1	5.8	5.9	6.1	6.2	6.0
6.3	6.2	6.1	6.2	6.0	5.7
6.2	5.8	5.7	6.3	6.2	5.7
6.2	6.1	5.9	6.5	5.4	6.7
5.9	6.1	5.9	5.9	6.1	6.1

(a) Construct a frequency distribution and display it as a table and as a histogram.

(b) Describe the shape of the distribution.

2.2.2 In a study of schizophrenia, researchers measured the activity of the enzyme monoamine oxidase (MAO) in the blood platelets of 18 patients. The results (expressed as nmoles benzylaldehyde product per 108 platelets) were as follows:[13]

6.8	8.4	8.7	11.9	14.2	18.8
9.9	4.1	9.7	12.7	5.2	7.8
7.8	7.4	7.3	10.6	14.5	10.7

Construct a dotplot of the data.

2.2.3 Consider the data presented in Exercise 2.2.2. Construct a frequency distribution and display it as a table and as a histogram.

2.2.4 A dendritic tree is a branched structure that emanates from the body of a nerve cell. As part of a study of brain development, 36 nerve cells were taken from the brains of newborn guinea pigs. The investigators counted the number of dendritic branch segments emanating from each nerve cell. The numbers were as follows:[14]

23	30	54	28	31	29	34	35	30
27	21	43	51	35	51	49	35	24
26	29	21	29	37	27	28	33	33
23	37	27	40	48	41	20	30	57

Construct a dotplot of the data.

2.2.5 Consider the data presented in Exercise 2.2.4. Construct a frequency distribution and display it as a table and as a histogram.

2.2.6 The total amount of protein produced by a dairy cow can be estimated from periodic testing of her milk. The following are the total annual protein production values (lb) for twenty-eight 2-year-old Holstein cows. Diet, milking procedures, and other conditions were the same for all the animals.[15]

425	481	477	434	410	397	438
545	528	496	502	529	500	465
539	408	513	496	477	445	546
471	495	445	565	499	508	426

Construct a frequency distribution and display it as a table and as a histogram.

2.2.7 For each of 31 healthy dogs, a veterinarian measured the glucose concentration in the anterior chamber of the right eye and also in the blood serum. The following data are the anterior chamber glucose measurements, expressed as a percentage of the blood glucose.[16]

81	85	93	93	99	76	75	84
78	84	81	82	89	81	96	82
74	70	84	86	80	70	131	75
88	102	115	89	82	79	106	

Construct a frequency distribution and display it as a table and as a histogram.

2.2.8 Agronomists measured the yield of a variety of hybrid corn in 16 locations in Illinois. The data, in bushels per acre, were[17]

241	230	207	219	266	167
204	144	178	158	153	
187	181	196	149	183	

(a) Construct a dotplot of the data.

(b) Describe the shape of the distribution.

2.2.9 (Computer problem) Trypanosomes are parasites that cause disease in humans and animals. In an early study of trypanosome morphology, researchers measured the lengths of 500 individual trypanosomes taken from the blood of a rat. The results are summarized in the accompanying frequency distribution.[18]

Length (μm)	Frequency (number of individuals)	Length (μm)	Frequency (number of individuals)
15	1	27	36
16	3	28	41
17	21	29	48
18	27	30	28
19	23	31	43
20	15	32	27
21	10	33	23
22	15	34	10
23	19	35	4
24	21	36	5
25	34	37	1
26	44	38	1

(a) Construct a histogram of the data using 24 classes (i.e., one class for each integer length, from 15 to 38).

(b) What feature of the histogram suggests the interpretation that the 500 individuals are a mixture of two distinct types?

(c) Construct a histogram of the data using only 6 classes. Discuss how this histogram gives a qualitatively different impression than the histogram from part (a).

2.3 Descriptive Statistics: Measures of Center

For categorical data, the frequency distribution provides a concise and complete summary of a sample. For numeric variables, the frequency distribution can usefully be supplemented by a few numerical measures. A numerical measure calculated from sample data is called a **statistic.*** **Descriptive statistics** are statistics that describe a set of data. Usually the descriptive statistics for a sample are calculated in order to provide information about a population of interest (see Section 2.8). In this section we discuss measures of the center of the data. There are several different ways to define the "center" or "typical value" of the observations in a sample. We will consider the two most widely used measures of center: the median and the mean.

THE MEDIAN

Perhaps the simplest measure of the center of a data set is the sample **median**. The sample median is the value that most nearly lies in the middle of the sample—it is the data value that splits the ordered data into two equal halves. To find the median, first arrange the observations in increasing order. In the array of ordered observations, the median is the middle value (if n is odd) or midway between the two middle values (if n is even). We denote the median of the sample by the symbol $\tilde{y}$ (read "y-tilde"). Example 2.3.1 illustrates these definitions.

Example 2.3.1

Weight Gain of Lambs The following are the 2-week weight gains (lb) of six young lambs of the same breed that had been raised on the same diet:[19]

$$11 \quad 13 \quad 19 \quad 2 \quad 10 \quad 1$$

The ordered observations are

$$1 \quad 2 \quad 10 \quad 11 \quad 13 \quad 19$$

The median weight gain is

$$\tilde{y} = \frac{10 + 11}{2} = 10.5 \text{ lb}$$

The median divides the sorted data into two equal pieces (the same number of observations fall above and below the median). Figure 2.3.1 shows a dotplot of the lamb weight-gain data, along with the location of $\tilde{y}$.

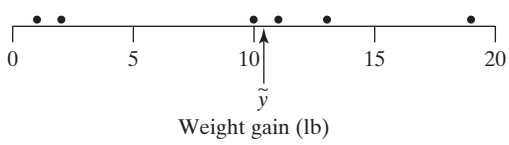


Figure 2.3.1 Plot of the lamb weight-gain data

*Numerical measures based on the entire population are called **parameters**, which are discussed in greater detail in Section 2.8.

Example 2.3.2

Weight Gain of Lambs Suppose the sample contained one more lamb, with the seven ranked observations as follows:

$$1 \quad 2 \quad 10 \quad 10 \quad 11 \quad 13 \quad 19$$

For this sample, the median weight gain is

$$\tilde{y} = 10 \text{ lb}$$

(Notice that in this example there are two lambs whose weight gain is equal to the median. The fourth observation—the second 10—is the median.) ■

A more formal way to define the median is in terms of rank position in the ordered array (counting the smallest observation as rank 1, the next as 2, and so on). The rank position of the median is equal to

$$(0.5)(n + 1)$$

Thus, if $n = 7$, we calculate $(0.5)(n + 1) = 4$, so that the median is the fourth largest observation; if $n = 6$, we have $(0.5)(n + 1) = 3.5$, so that the median is midway between the third and fourth largest observations. Note that the formula $(0.5)(n + 1)$ does not give the median, it gives the location of the median within the ordered list of the data.

THE MEAN

The most familiar measure of center is the ordinary average or **mean** (sometimes called the arithmetic mean). The mean of a sample (or "the sample mean") is the sum of the observations divided by the number of observations. If we denote a variable by Y, then we denote the observations in a sample by $y_1, y_2, \ldots, y_n$ and we denote the mean of the sample by the symbol $\bar{y}$ (read "y-bar"). Example 2.3.3 illustrates this notation.

Example 2.3.3

Weight Gain of Lambs The following are the data from Example 2.3.1:

$$11 \quad 13 \quad 19 \quad 2 \quad 10 \quad 1$$

Here $y_1 = 11, y_2 = 13$, and so on, and $y_6 = 1$. The sum of the observations is $11 + 13 + \cdots + 1 = 56$. We can write this using "summation notation" as $\sum_{i=1}^{n} y_i = 56$. The symbol $\sum_{i=1}^{n} y_i$ means to "add up the y_i's." Thus, when $n = 6$, $\sum_{i=1}^{n} y_i = y_1 + y_2 + y_3 + y_4 + y_5 + y_6$. In this case we get $\sum_{i=1}^{n} y_i = 11 + 13 + 19 + 2 + 10 + 1 = 56$.

The mean weight gain of the six lambs in this sample is

$$\begin{aligned} \bar{y} &= \frac{11 + 13 + 19 + 2 + 10 + 1}{6} \\ &= \frac{56}{6} \\ &= 9.33 \text{ lb} \end{aligned}$$

THE SAMPLE MEAN The general definition of the sample mean is

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

where the y_i's are the observations in the sample and n is the sample size (that is, the number of y_i's).

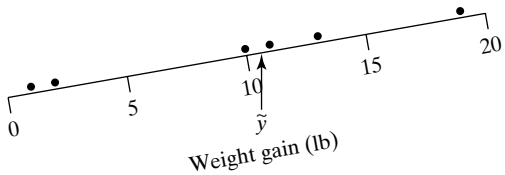


Figure 2.3.2 Plot of the lamb weight-gain data with the sample median as the fulcrum of a balance

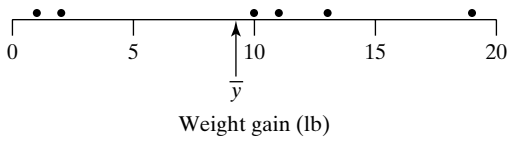


Figure 2.3.3 Plot of the lamb weight-gain data with the sample mean as the fulcrum of a balance

While the median divides the data into two equal pieces (i.e., the same number of observations above and below), the mean is the "point of balance" of the data. Figure 2.3.2 shows a dotplot of the lamb weight-gain data, along with the location of $\tilde{y}$. If the data points were children on a weightless seesaw, then the seesaw would tip if the fulcrum were placed at $\tilde{y}$ despite there being the same number of children on either side. The children on the left side (below $\tilde{y}$) tend to sit further from $\tilde{y}$ than the children on the right (above $\tilde{y}$) causing the seesaw to tip. However, if the fulcrum were placed at $\bar{y}$, the seesaw would exactly balance as in Figure 2.3.3. ■

The difference between a data point and the mean is called a **deviation**: $\text{deviation}_i = y_i - \bar{y}$. The mean has the property that the sum of the deviations from the mean is zero—that is, $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$. In this sense, the mean is a center of the distribution—the positive deviations balance the negative deviations.

Example 2.3.4

Weight Gain of Lambs For the lamb weight-gain data, the deviations are as follows:

$$\text{deviation}_1 = y_1 - \bar{y} = 11 - 9.33 = 1.67$$
$$\text{deviation}_2 = y_2 - \bar{y} = 13 - 9.33 = 3.67$$
$$\text{deviation}_3 = y_3 - \bar{y} = 19 - 9.33 = 9.67$$
$$\text{deviation}_4 = y_4 - \bar{y} = 2 - 9.33 = -7.33$$
$$\text{deviation}_5 = y_5 - \bar{y} = 10 - 9.33 = 0.67$$
$$\text{deviation}_6 = y_6 - \bar{y} = 1 - 9.33 = -8.33$$

The sum of the deviations is $\sum_{i=1}^{n}(y_i - \bar{y}) = 1.67 + 3.67 + 9.67 - 7.33 + 0.67 - 8.33 = 0$. ■

Robustness A statistic is said to be **robust** if the value of the statistic is relatively unaffected by changes in a small portion of the data, even if the changes are dramatic ones. The median is a robust statistic, but the mean is not robust because it can be greatly shifted by changes in even one observation. Example 2.3.5 illustrates this behavior.

Example 2.3.5

Weight Gain of Lambs Recall that for the lamb weight-gain data

$$1 \quad 2 \quad 10 \quad 11 \quad 13 \quad 19$$

we found

$$\bar{y} = 9.33 \text{ and } \tilde{y} = 10.5$$

Suppose now that the observation 19 is changed. How would the mean and median be affected? You can visualize the effect by imagining moving the right-hand dot in Figure 2.3.3. Clearly the mean could change a great deal; the median would not be affected. For instance,

If the 19 is changed to 14, the mean becomes 8.5 and the median does not change.

If the 19 is changed to 29, the mean becomes 11 and the median does not change.

These changes are not wild ones; that is, the changed samples might well have arisen from the same feeding experiment. Of course, a huge change, such as changing the 19 to 100, would shift the mean very drastically. Note that it would not shift the median at all. ■

VISUALIZING THE MEAN AND MEDIAN

We can visualize the mean and the median in relation to the histogram of a distribution. The median divides the area under the histogram roughly in half because it divides the observations roughly in half ["roughly" because some observations may be tied at the median, as in Example 2.3.3(b), and because the observations within each class are not uniformly distributed across the class]. The mean can be visualized as the point of balance of the histogram: If the histogram were made out of plywood, it would balance if supported at the mean.

If the frequency distribution is symmetric, the mean and the median are equal and fall in the center of the distribution. If the frequency distribution is skewed, both measures are pulled toward the longer tail, but the mean is usually pulled farther than the median. The effect of skewness is illustrated by the following example.

Example 2.3.6

Cricket Singing Times Male Mormon crickets *(Anabrus simplex)* sing to attract mates. A field researcher measured the duration of 51 unsuccessful songs—that is, the time until the singing male gave up and left his perch.[20] Figure 2.3.4 shows the histogram of the 51 singing times. Table 2.3.1 gives the raw data. The median is 3.7 min and the mean is 4.3 min. The discrepancy between these measures is due largely to the long straggly tail of the distribution; the few unusually long singing times influence the mean, but not the median. ■

Table 2.3.1 Fifty-one cricket singing times (min)

4.3	3.9	17.4	2.3	0.8	1.5	0.7	3.7
24.1	9.4	5.6	3.7	5.2	3.9	4.2	3.5
6.6	6.2	2.0	0.8	2.0	3.7	4.7	
7.3	1.6	3.8	0.5	0.7	4.5	2.2	
4.0	6.5	1.2	4.5	1.7	1.8	1.4	
2.6	0.2	0.7	11.5	5.0	1.2	14.1	
4.0	2.7	1.6	3.5	2.8	0.7	8.6	

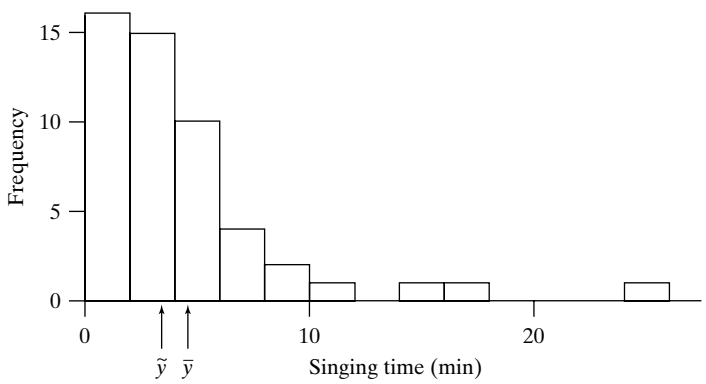


Figure 2.3.4 Histogram of cricket singing times

MEAN VERSUS MEDIAN

Both the mean and the median are usually reasonable measures of the center of a data set. The mean is related to the sum; for example, if the mean weight gain of 100 lambs is 9 lb, then the total weight gain is 900 lb, and this total may be of primary interest since it translates more or less directly into profit for the farmer. In some

situations the mean makes very little sense. Suppose, for example, that the observations are survival times of cancer patients on a certain treatment protocol, and that most patients survive less than 1 year, while a few respond well and survive for 5 or even 10 years. In this case, the mean survival time might be greater than the survival time of most patients; the median would more nearly represent the experience of a "typical" patient. Note also that the mean survival time cannot be computed until the last patient has died; the median does not share this disadvantage. Situations in which the median can readily be computed, but the mean cannot, are not uncommon in bioassay, survival, and toxicity studies.

We have noted that the median is more robust than the mean. If a data set contains a few observations rather distant from the main body of the data—that is, a long, straggly tail—then the mean may be unduly influenced by these few unusual observations. Thus, the "tail" may "wag the dog"—an undesirable situation. In such cases, the robustness of the median may be advantageous.

An advantage of the mean is that in some circumstances it is more efficient than the median. Efficiency is a technical notion in statistical theory; roughly speaking, a method is efficient if it takes full advantage of all the information in the data. Partly because of its efficiency, the mean has played a major role in classical methods in statistics.

Exercises 2.3.1–2.3.14

2.3.1 Invent a sample of size 5 for which the sample mean is 20 and not all the observations are equal.

2.3.2 Invent a sample of size 5 for which the sample mean is 20 and the sample median is 15.

2.3.3 A researcher applied the carcinogenic (cancer-causing) compound benzo(a)pyrene to the skin of five mice, and measured the concentration in the liver tissue after 48 hours. The results (nmol/gm) were as follows:[21]

6.3 5.9 7.0 6.9 5.9

Determine the mean and the median.

2.3.4 Consider the data from Exercise 2.3.3. Do the calculated mean and median support the claim that, in general, liver tissue concentration after 48 hours differs from 6.3 nmol/gm?

2.3.5 Six men with high serum cholesterol participated in a study to evaluate the effects of diet on cholesterol level. At the beginning of the study their serum cholesterol levels (mg/dl) were as follows: [22]

366 327 274 292 274 230

Determine the mean and the median.

2.3.6 Consider the data from Exercise 2.3.5. Suppose an additional observation equal to 400 were added to the sample. What would be the mean and the median of the seven observations?

2.3.7 The weight gains of beef steers were measured over a 140-day test period. The average daily gains (lb/day) of 9 steers on the same diet were as follows:[23]

3.89 3.51 3.97 3.31 3.21
3.36 3.67 3.24 3.27

Determine the mean and median.

2.3.8 Consider the data from Exercise 2.3.7. Are the calculated mean and median consistent with the claim that, in general, steers gain 3.5 lb/day? Are they consistent with a claim of 4.0 lb/day?

2.3.9 Consider the data from Exercise 2.3.7. Suppose an additional observation equal to 2.46 were added to the sample. What would be the mean and the median of the 10 observations?

2.3.10 As part of a classic experiment on mutations, 10 aliquots of identical size were taken from the same culture of the bacterium *E. coli*. For each aliquot, the number of bacteria resistant to a certain virus was determined. The results were as follows:[24]

14 15 13 21 15
14 26 16 20 13

(a) Construct a frequency distribution of these data and display it as a histogram.

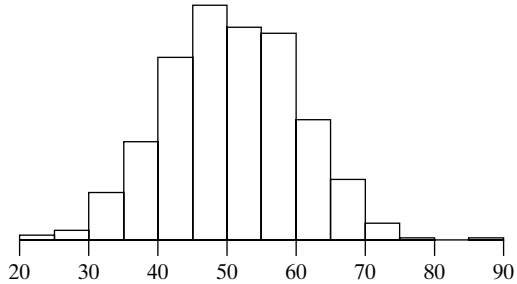
(b) Determine the mean and the median of the data and mark their locations on the histogram.

2.3.11 The accompanying table gives the litter size (number of piglets surviving to 21 days) for each of 36 sows (as in Example 2.2.4). Determine the median litter size. (*Hint*: Note that there is one 5, but there are two 7's, three 8's, etc.)

Number of piglets	Frequency (Number of sows)
5	1
6	0
7	2
8	3
9	3
10	9
11	8
12	5
13	3
14	2
Total	36

2.3.12 Consider the data from Exercise 2.3.11. Determine the mean of the 36 observations. (*Hint*: Note that there is one 5 but there are two 7's, three 8's, etc. Thus, $\sum y_i = 5 + 7 + 7 + 8 + 8 + 8 + \cdots = 5 + 2(7) + 3(8) + \cdots$)

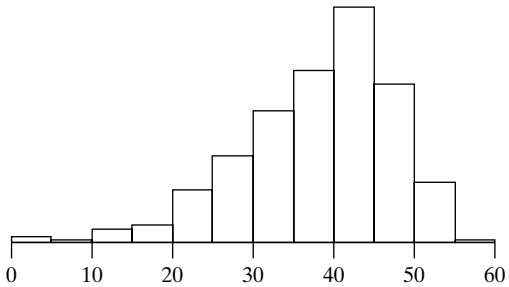
2.3.13 Here is a histogram.



(a) Estimate the median of the distribution.

(b) Estimate the mean of the distribution.

2.3.14 Here is a histogram.



(a) Estimate the median of the distribution.

(b) Estimate the mean of the distribution.

2.4 Boxplots

One of the most efficient graphics, both for examining a single distribution and for making comparisons between distributions, is known as a boxplot, which is the topic of this section. Before discussing boxplots, however, we need to discuss quartiles.

QUARTILES AND THE INTERQUARTILE RANGE

The median of a distribution splits the distribution into two parts, a lower part and an upper part. The **quartiles** of a distribution divide each of these parts in half, thereby dividing the distribution into four quarters. The **first quartile**, denoted by Q_1, is the median of the data values in the lower half of the data set. The **third quartile**, denoted by Q_3, is the median of the data values in the upper half of the data set.* The following example illustrates these definitions.

*Some authors use other definitions of quartiles, as does some computer software. A common alternative definition is to say that the first quartile has rank position $(0.25)(n + 1)$ and that the third quartile has rank position $(0.75)(n + 1)$. Thus, if $n = 10$, the first quartile would have rank position $(0.25)(11) = 2.75$—that is, to find the first quartile we would have to interpolate between the second and third largest observations. If n is large, then there is little practical difference between the definitions that various authors use.

Example 2.4.1

Blood Pressure The systolic blood pressures (mm Hg) of seven middle-aged men were as follows:[25]

151 124 132 170 146 124 113

Putting these values in rank order, the sample is

113 124 124 132 146 151 170

The median is the fourth largest observation, which is 132. There are three data points in the lower part of the distribution: 113, 124, and 124. The median of these three values is 124. Thus, the first quartile, Q_1, is 124.

Likewise, there are three data points in the upper part of the distribution: 146, 151 and 170. The median of these three values is 151. Thus, the third quartile, Q_3, is 151.

113	124	124	132	146	151	170
	↑		⋮		↑	
	first quartile		median		third quartile	
	Q_1				Q_3	

Note that the median is not included in either the lower part or the upper part of the distribution. If the sample size, n, is even, then exactly one-half of the observations are in the lower part of the distribution and one-half are in the upper part.

The **interquartile range** is the difference between the first and third quartiles and is abbreviated as **IQR**: IQR $= Q_3 - Q_1$. For the blood pressure data in Example 2.4.1, the IQR is $151 - 124 = 27$. Note that the IQR is a *number,* not an interval; the IQR measures the spread of the middle 50% of the distribution.

Example 2.4.2

Pulse The pulses of 12 college students were measured.[26] Here are the data, arranged in order, with the position of the median indicated by a dashed line:

62 64 68 70 70 74 ⋮ 74 76 76 78 78 80

The median is $\frac{74 + 74}{2} = 74$. There are six observations in the lower part of the distribution: 62, 64, 68, 70, 70, 74. Thus, the first quartile is the average of the third and fourth largest data values:

$$Q_1 = \frac{68 + 70}{2} = 69$$

There are six observations in the upper part of the distribution: 74, 76, 76, 78, 78, 80. Thus, the third quartile is the average of the ninth and tenth largest data values (the third and fourth values in the upper part of the distribution):

$$Q_3 = \frac{76 + 78}{2} = 77$$

Thus, the interquartile range is

$$\text{IQR} = 77 - 69 = 8$$

We have

$$\begin{array}{ccccccccccccc} 62 & 64 & 68 & 70 & 70 & 74 & \vdots\ 74 & 76 & 76 & 78 & 78 & 80 \\ & & \uparrow & & & \text{median} & & & & \uparrow & & \\ & & \text{first quartile} & & & & & & & \text{third quartile} & & \\ & & Q_1 & & & & & & & Q_3 & & \end{array}$$

The minimum pulse value is 62 and the maximum is 80.

The minimum, the maximum, the median, and the quartiles, taken together, are referred to as the **five-number summary** of the data.

OUTLIERS

Sometimes a data point differs so much from the rest of the data that it doesn't seem to belong with the other data. Such a point is called an **outlier**. An outlier might occur because of a recording error or typographical error when the data are recorded, because of an equipment failure during an experiment, or for many other reasons. Outliers are the most interesting points in a data set. Sometimes outliers tell us about a problem with the experimental protocol (e.g., an equipment failure, a failure of a patient to take his or her medication consistently during a medical trial). At other times an outlier might alert us to the fact that a special circumstance has happened (e.g., an abnormally high or low value on a medical test could indicate the presence of a disease in a patient).

People often use the term "outlier" informally. There is, however, a common definition of "outlier" in statistical practice. To give a definition of outlier, we first discuss what are known as fences. The **lower fence** of a distribution is

$$\text{lower fence} = Q_1 - 1.5 \times \text{IQR}$$

The **upper fence** of a distribution is

$$\text{upper fence} = Q_3 + 1.5 \times \text{IQR}$$

Note that the fences need not be data values; indeed, there might be no data near the fences. The fences just locate limits within the sample distribution. These limits give us a way to define outliers. *An outlier is a data point that falls outside of the fences*. That is, if

$$\text{data point} < Q_1 - 1.5 \times \text{IQR}$$

or

$$\text{data point} > Q_3 + 1.5 \times \text{IQR}$$

then we call the point an outlier.

Example 2.4.3

Pulse In Example 2.4.2 we saw that $Q_1 = 69$, $Q_3 = 77$, and IQR $= 8$. Thus, the lower fence is $69 - 1.5 \times 8 = 69 - 12 = 57$. Any point less than 57 would be an outlier. The upper fence is $77 + 1.5 \times 8 = 77 + 12 = 89$. Any point greater than

89 would be an outlier. Since there are no points less than 57 or greater than 89, there are no outliers in this data set. ■

Example 2.4.4

Radish Growth in Light A common biology experiment involves growing radish seedlings under various conditions. In one experiment students grew 14 radish seedlings in constant light. The observations, in order, are

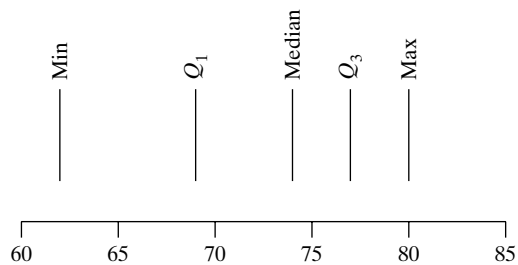
$$3 \quad 5 \quad 5 \quad 7 \quad 7 \quad 8 \quad 9 \quad \underset{\text{median}}{\vdots} \quad 10 \quad 10 \quad 10 \quad 10 \quad 14 \quad 20 \quad 21$$

first quartile Q_1 third quartile Q_3

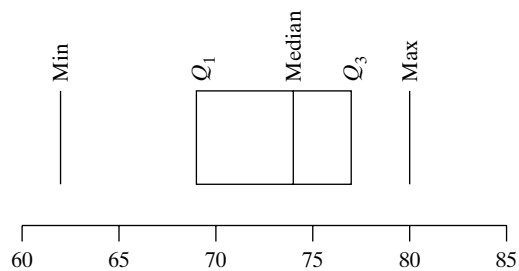
Thus, the median is $\frac{9 + 10}{2} = 9.5$, Q_1 is 7, and Q_3 is 10. The interquartile range is $\text{IQR} = 10 - 7 = 3$. The lower fence is $7 - 1.5 \times 3 = 7 - 4.5 = 2.5$, so any point less than 2.5 would be an outlier. The upper fence is $10 + 1.5 \times 3 = 10 + 4.5 = 14.5$, so any point greater than 14.5 is an outlier. Thus, the two largest observations in this data set are outliers: 20 and 21. ■

BOXPLOTS FOR DATA WITH NO OUTLIERS

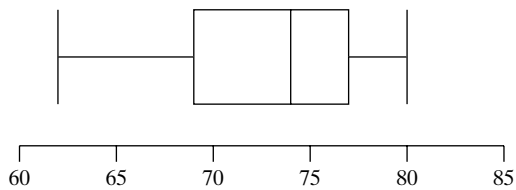
A **boxplot** is a visual representation of the five-number summary. To make a boxplot for a data set with no outliers, we first make a number line; then we mark the positions minimum, Q_1, the median, Q_3, and the maximum:



Next, we make a box connecting the quartiles:



Note that the interquartile range is equal to the length of the box. Finally, provided there are no outliers* we extend "whiskers" from Q_1 down to the minimum and from Q_3 up to the maximum:



A boxplot gives a quick visual summary of the distribution. We can immediately see where the center of the data is from the line within the box that locates the median. We see the spread of the total distribution, from the minimum up to the maximum, as well as the spread of the middle half of the distribution—the interquartile range—from the length of the box. The boxplot also gives an indication of the shape of the distribution; the preceding boxplot has a long lower whisker, indicating that the distribution is skewed to the left. Example 2.4.5 shows a boxplot for data from a radish growth experiment that had no outliers.†

Example 2.4.5

Radish Growth In another version of the experiment in Example 2.4.4, a moist paper towel is put into a plastic bag. About one third of the way from the bottom of the bag a seam of staples was created; the radish seeds were placed along the seam. One group of students kept their radish seed bags in total darkness for 3 days and then measured the length, in mm, of each radish shoot at the end of the 3 days. They collected 14 observations; the data are shown in Table 2.4.1.[27]

Table 2.4.1 Radish growth, in mm, after three days in total darkness

15	20	11	30	33
20	29	35	8	10
22	37	15	25	

Here are the data in order from smallest to largest:

8 10 11 **15** 15 20 **20** ¦ **22** 25 29 **30** 33 35 37

↑ first quartile Q_1

median

↑ third quartile Q_3

The quartiles are $Q_1 = 15$ and $Q_3 = 30$. The median, $\tilde{y} = 21$, is the average of the two middle values of 20 and 22. Figure 2.4.1 shows a boxplot of the same data. ■

*We will consider situations with outliers after the next example.

†This and subsequent boxplots in our text are slightly stylized. Different computer packages present the plot somewhat differently, but all boxplots have the same basic five-number summary.

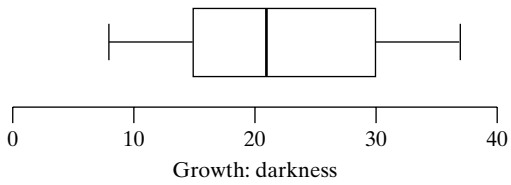


Figure 2.4.1 Boxplot of data on radish growth in darkness

BOXPLOTS FOR DATA WITH OUTLIERS

If there are outliers in the upper part of the distribution, then we can identify them with dots (or other plotting symbols) on the boxplot. We then extend a whisker from Q_3 up to the largest data point that is *not* an outlier. Likewise, if there are outliers in the lower part of the distribution, we identify them with dots and extend a whisker from Q_1 down to the smallest observation that is not an outlier. Figure 2.4.2 shows the distribution of radish seedlings grown under constant light. The area between the lower and upper fences is white, while the outlying region is blue.

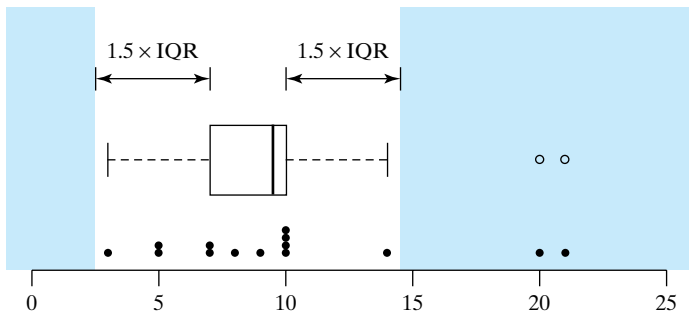


Figure 2.4.2 Dotplot and boxplot of data on radish growth in constant light. The points in the blue region are outliers.

Figure 2.4.3 shows a boxplot of the data on radish seedlings grown in constant light.*

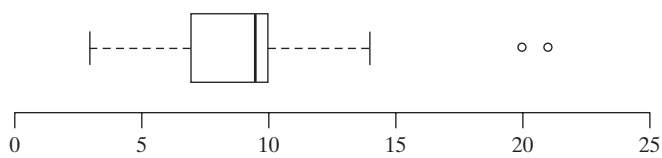


Figure 2.4.3 Boxplot of data on radish growth in constant light

The method we have defined for identifying outliers allows the bulk of the data to determine how extreme an observation must be before we consider it to be an

*Most computer software has options that can alter how outliers are determined and displayed.

outlier, since the quartiles and the IQR are determined from the data themselves. Thus, a point that is an outlier in one data set might not be an outlier in another data set. We label a point as an outlier if it is unusual relative to the inherent variability in the entire data set.

After an outlier has been identified, people are often tempted to remove the outlier from the data set. In general this is not a good idea. If we can identify that an outlier occurred due to an equipment error, for example, then we have good reason to remove the outlier before analyzing the rest of the data. However, quite often outliers appear in data sets without any identifiable, external reason for them. In such cases, we simply proceed with our analysis, aware that there is an outlier present. In some cases, we might want to calculate the mean, for example, with and without the outlier and then report both calculations to show the effect of the outlier in the overall analysis. This is preferable to removing the outlier, which obscures the fact that there was an unusual data point present.

Exercises 2.4.1–2.4.8

2.4.1 Here are the data from Exercise 2.3.10 on the number of virus-resistant bacteria in each of 10 aliquots:

14	15	13	21	15
14	26	16	20	13

(a) Determine the median and the quartiles.

(b) Determine the interquartile range.

(c) How large would an observation in this data set have to be in order to be an outlier?

2.4.2 Here are the 18 measurements of MAO activity reported in Exercise 2.2.2:

6.8	8.4	8.7	11.9	14.2	18.8
9.9	4.1	9.7	12.7	5.2	7.8
7.8	7.4	7.3	10.6	14.5	10.7

(a) Determine the median and the quartiles.

(b) Determine the interquartile range.

(c) How large would an observation in this data set have to be in order to be an outlier?

(d) Construct a boxplot of the data.

2.4.3 In a study of milk production in sheep (for use in making cheese), a researcher measured the 3-month milk yield for each of 11 ewes. The yields (liters) were as follows:[28]

56.5	89.8	110.1	65.6	63.7	82.6
75.1	91.5	102.9	44.4	108.1	

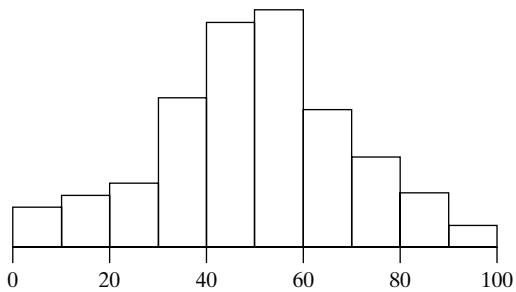
(a) Determine the median and the quartiles.

(b) Determine the interquartile range.

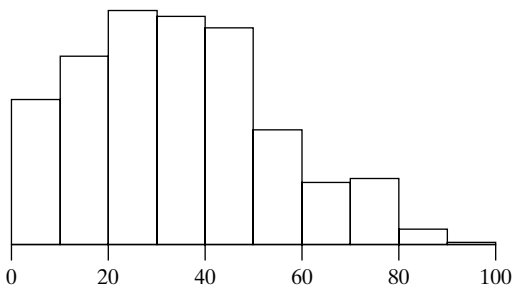
(c) Construct a boxplot of the data.

2.4.4 For each of the following histograms, use the histogram to estimate the median and the quartiles; then construct a boxplot for the distribution.

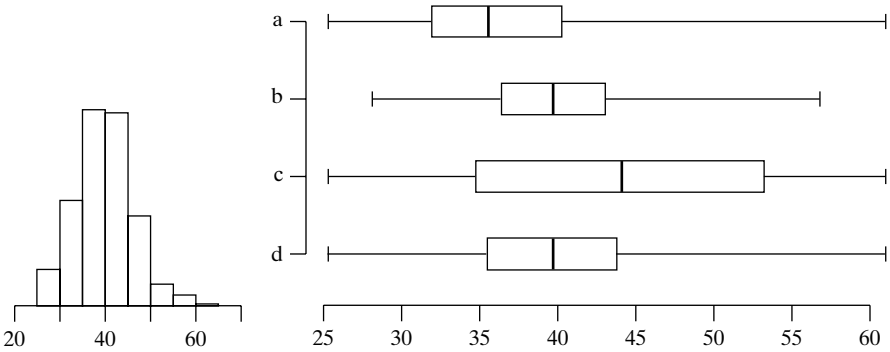
(a)



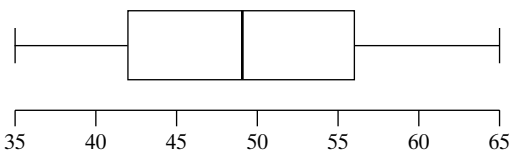
(b)



2.4.5 The following histogram shows the same data that are shown in one of the four boxplots. Which boxplot goes with the histogram? Explain your answer.



2.4.6 The following boxplot shows the five-number summary for a data set. For these data the minimum is 35, Q_1 is 42, the median is 49, Q_3 is 56, and the maximum is 65. Is it possible that no observation in the data set equals 42? Explain your answer.



2.4.7 Statistics software can be used to find the five-number summary of a data set. Here is an example of MINITAB's descriptive statistics summary for a variable stored in column 1 (C1) of MINITAB's worksheet.

```
Variable  N   Mean   Median  TrMean  StDev  SEMean
   C1    75  119.94  118.40  119.98   9.98    1.15

Variable  Min     Max     Q1      Q3
   C1    95.16  145.11  113.59  127.42
```

(a) Use the MINITAB output to calculate the interquartile range.

(b) Are there any outliers in this set of data?

2.4.8 Consider the data from Exercise 2.4.7. Use the five-number summary that is given to create a boxplot of the data.

2.5 Relationships between Variables

In the previous sections we have studied **univariate** summaries of both numeric and categorical variables. A univariate summary is a graphical or numeric summary of a single variable.

The histogram, boxplot, sample mean, and median are all examples of univariate summaries for numeric data. The bar chart, frequency, and relative frequency tables are examples of univariate summaries for categorical data. In this section we present some common **bivariate** graphical summaries used to examine the *relationship* between pairs of variables.

CATEGORICAL–CATEGORICAL RELATIONSHIPS

To understand the relationship between two categorical variables, we first summarize the data in a **bivariate frequency table**. Unlike the frequency table presented in Section 2.2 (a univariate table), the bivariate frequency table has both rows and columns—one dimension for each variable. The choice of which variable to list with the rows and which to list with the columns is arbitrary. The following example considers the relationship between two categorical variables: *E. Coli* Source and Sampling Location.

Example 2.5.1

***E. Coli* Watershed Contamination** In an effort to determine if there are differences in the primary sources of fecal contamination at different locations in the Morro Bay watershed, $n = 623$ water specimens were collected at three primary locations that feed into Morro Bay: Chorro Creek ($n_1 = 241$), Los Osos Creek ($n_2 = 256$), and Baywood Seeps ($n_3 = 126$).[29] DNA fingerprinting techniques were used to determine the intestinal origin of the dominant *E. coli* strain in each water specimen. *E. coli* origins were classified into the following five categories: bird, domestic pet (e.g., cat or dog), farm animal (e.g., horse, cow, pig), human, or other terrestrial mammal (e.g., fox, mouse, coyote . . .). Thus, each water specimen had *two* categorical variables measured: location (Chorro, Los Osos, or Baywood) and *E. coli* source (bird, . . . , terrestrial mammal). Table 2.5.1 presents a frequency table of the data. ■

Table 2.5.1 Frequency table of *E. coli* source by location

	E. Coli Source					
Location	Bird	Domestic pet	Farm animal	Human	Terrestrial mammal	Total
Chorro Creek	46	29	106	38	22	**241**
Los Osos Creek	79	56	32	63	26	**256**
Baywood Seeps	35	23	0	60	8	**126**
Total	**160**	**108**	**138**	**161**	**56**	**623**

While Table 2.5.1 provides a concise summary of the data, it is difficult to discover any patterns in the data. Examining relative frequencies (row or column proportions) often helps us make meaningful comparisons as seen in the following example.

Example 2.5.2

***E. Coli* Watershed Contamination** Are domestic pets more of an *E. coli* problem (i.e., source) at Chorro Creek or Baywood Seeps? Table 2.5.1 shows that the domestic pet *E. coli* source count at Chorro (29) is higher than Baywood (23), so at first glance it seems that pets are more problematic at Chorro. However, as more water specimens were collected at Chorro ($n_1 = 241$) than Baywood ($n_2 = 126$), the relative frequency of domestic pet source *E. coli* is actually lower at Chorro ($29/241 = 0.120$) than Baywood ($23/126 = 0.183$). Table 2.5.2 displays row percentages and thus facilitates comparisons of *E. coli* sources among the locations. (Note that column percentages would not be meaningful in this context since the water was sampled by location and not by *E. coli* source.) ■

Table 2.5.2 Bivariate relative frequency table (row percentages) of *E. coli* source by location

	E. Coli Source					
Location	Bird	Domestic pet	Farm animal	Human	Terrestrial mammal	Total
Chorro Creek	19.1	12.0	44.0	15.8	9.1	**100**
Los Osos Creek	30.9	21.9	12.5	24.6	10.2	**100**
Baywood Seeps	27.8	18.3	0.0	47.6	6.3	**100**
All locations	**25.7**	**17.3**	**22.2**	**25.8**	**9.0**	**100**

To visualize the data in Tables 2.5.1 and 2.5.2, we can examine **stacked bar charts**. With a stacked frequency bar chart, the overall height of each bar reflects the sample size for a level of the X categorical variable (e.g., location), while the height or thickness of a slice that makes up a bar represents the count of the Y categorical variable (e.g., *E. coli* source) for that level of X. Figure 2.5.1 displays a stacked bar chart for the *E. coli* watershed count data in Table 2.5.1.

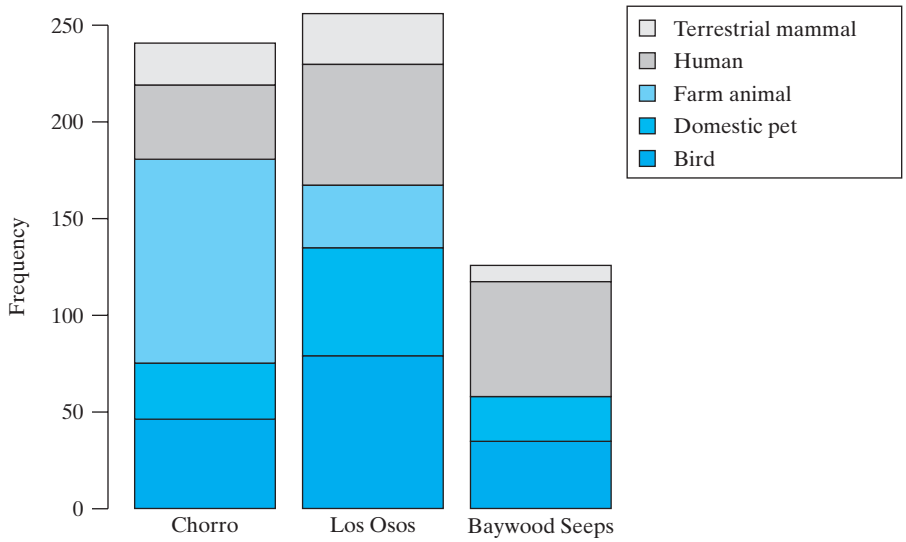


Figure 2.5.1 Stacked frequency chart of *E. coli* source by location

Like the frequency table, the stacked frequency bar chart is not conducive to making comparisons across the three locations as the sample sizes differ for these locations. (This graph does help highlight the difference in sample sizes; for example, it is very clear that many fewer water specimens were collected at Baywood Seeps.) A chart that better displays the distribution of one categorical variable across levels of another is a **stacked relative frequency** (or percentage) bar chart, which graphs the summaries from a bivariate relative frequency table such as Table 2.5.2. Figure 2.5.2 provides an example using the *E. coli* watershed contamination data. This plot normalizes the bars of Figure 2.5.1 to have the same height (100%) to facilitate comparisons across the three locations.

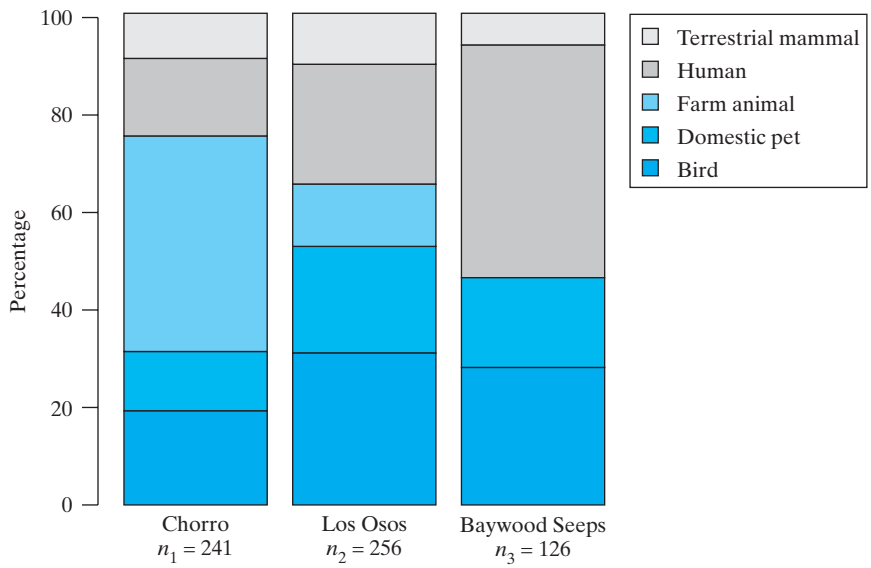


Figure 2.5.2 Stacked relative frequency (percentage) chart of *E. coli* source by location

Figure 2.5.2 makes it very easy to see that farm animals are the largest contributors of *E. coli* to Chorro Creek while humans are primarily responsible for the pollution at Baywood Seeps. The distribution of the slices in the three bars appears quite different, suggesting that the distribution of *E. coli* sources is not the same at the three locations. In Chapter 10 we will learn how to determine if these apparent differences are large enough to be compelling evidence for real differences in the distribution of *E. coli* source by location, or whether they are likely due to chance variation.

NUMERIC–CATEGORICAL RELATIONSHIPS

In Section 2.4 we learned that boxplots are graphs based on only five numbers: the minimum, first quartile, median, third quartile, and maximum. They are appealing plots because they are very simple and uncluttered, yet contain easy to read information about center, spread, skewness, and even outliers of a data set. By displaying **side-by-side boxplots** on the same graph, we are able to compare numeric data among several groups. We now consider an extension of the radish shoot growth problem in Example 2.4.3.

Example 2.5.3

Radish Growth Does light exposure alter initial radish shoot growth? The complete radish growth experiment of Examples 2.4.4 and 2.4.5 actually involved a total of 42 radish seeds randomly divided to receive one of three lighting conditions for germination (14 seeds in each lighting condition): 24-hour light, diurnal light (12 hours of light and 12 hours of darkness each day), and 24 hours of darkness. At the end of 3 days, shoot length was measured (mm). Thus, each shoot has two variables that are measured in this study: the categorical variable lighting condition (light, diurnal, dark) and the numeric variable sprout length (mm). Figure 2.5.3 displays side-by-side boxplots of the data. The boxplots make it very easy to compare the growth under the three conditions: It appears that light inhibits shoot growth. Are the observed differences in growth among the lighting conditions just due to chance variation, or is light really altering growth? We will learn how to numerically measure the strength of this evidence and answer this question in Chapters 7 and 11. ■

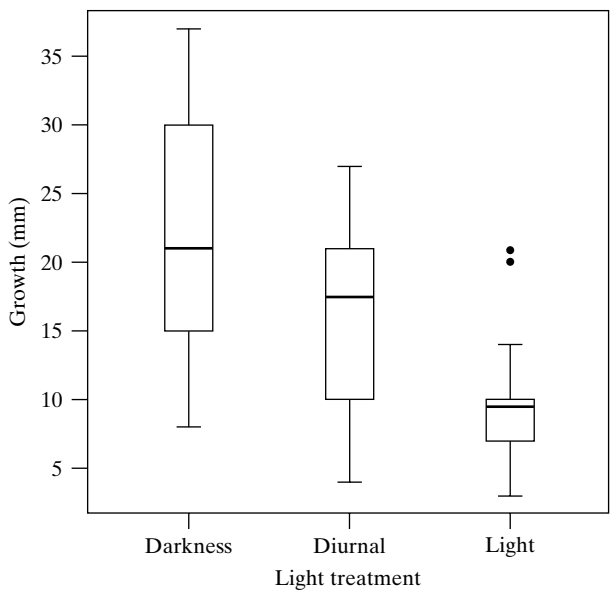


Figure 2.5.3 Side-by-side boxplots of radish growth under three conditions: constant darkness, half light–half darkness, and constant light

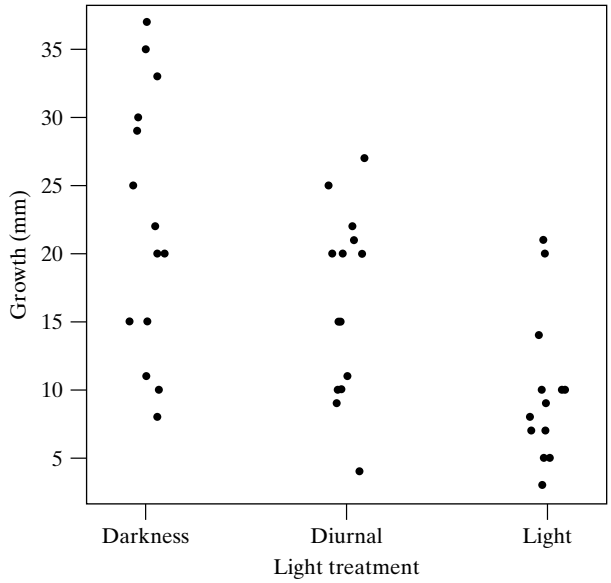


Figure 2.5.4 Side-by-side jittered dotplots of radish growth under three conditions: constant darkness, half light–half darkness, and constant light

For smaller data sets, we also may consider side-by-side dotplots of the data. Figure 2.5.4 displays a jittered side-by-side dotplot of the radish growth data of Example 2.5.3. The "jitter" is a common software option that adds horizontal scatter to the plot, helping to reduce the overlap of the dots. Choosing between side-by-side boxplots and dotplots is matter of personal preference. A good rule of thumb is to choose the plot that accurately reflects patterns in the data in the cleanest (least ink on the paper) way possible. For the radish growth example, the boxplot enables a very clean comparison of the growth under the three light treatments without hiding any information revealed by the dotplot.

NUMERIC–NUMERIC RELATIONSHIPS

Each of the previous examples considered comparing the distribution of one variable (either categorical or numeric) among several groups (i.e., across levels of a categorical variable). In the next example we illustrate the **scatterplot** as a tool to examine the relationship between two numeric variables, X and Y. A scatterplot plots each observed (x,y) pair as a dot on the x–y plane.

Example 2.5.4

Whale Selenium Can metal concentration in marine mammal teeth be used as a bioindicator for body burden? Selenium (Se) is an essential element that has been shown to play an important role in protecting marine mammals against the toxic effects of mercury (Hg) and other metals. Twenty beluga whales (*Delphinapterus leucas*) were harvested from the Mackenzie Delta, Northwest Territories, as part of an annual traditional Inuit hunt.[30] Each whale yielded two numeric measurements: Tooth Se (μg/g) and Liver Se (ng/g). Selenium concentrations for the whales are listed in Table 2.5.3. Liver Se concentration (Y) is graphed against Tooth Se concentration (X) in the scatterplot of Figure 2.5.5. ■

Table 2.5.3 Liver and tooth selenium concentrations of 20 belugas

Whale	Liver Se (μg/g)	Tooth Se (ng/g)	Whale	Liver Se (μg/g)	Tooth Se (ng/g)
1	6.23	140.16	11	15.28	112.63
2	6.79	133.32	12	18.68	245.07
3	7.92	135.34	13	22.08	140.48
4	8.02	127.82	14	27.55	177.93
5	9.34	108.67	15	32.83	160.73
6	10.00	146.22	16	36.04	227.60
7	10.57	131.18	17	37.74	177.69
8	11.04	145.51	18	40.00	174.23
9	12.36	163.24	19	41.23	206.30
10	14.53	136.55	20	45.47	141.31

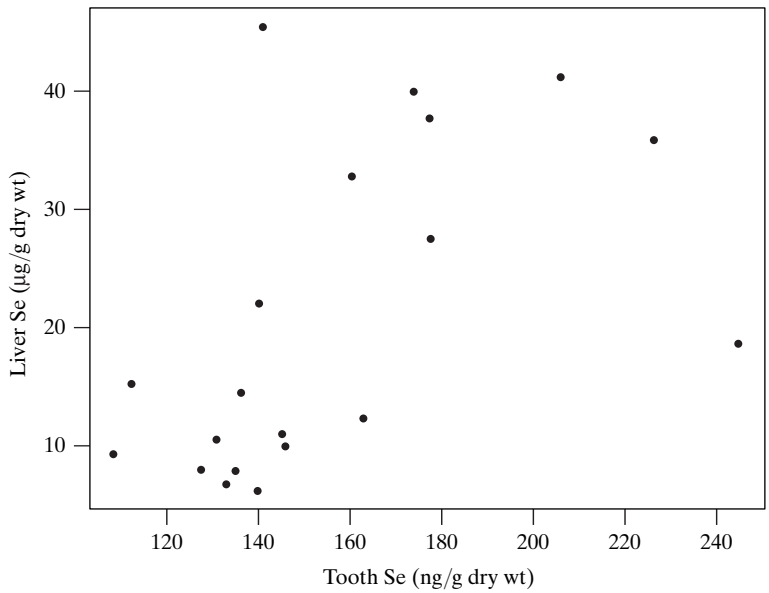


Figure 2.5.5 Scatterplot of liver selenium concentration against tooth selenium concentration for 20 belugas

Scatterplots are helpful in revealing relationships between numeric variables. In Figure 2.5.6 two lines have been added to the whale selenium scatterplot of Figure 2.5.5 to highlight the increasing trend in the data: Tooth Se concentration tends to increase with liver Se concentration. The dashed line is called a **lowess smooth**, whereas the straight solid line is called a **regression line**. Many software packages allow one to easily add these lines to a scatterplot. The lowess smooth is particularly helpful in visualizing curved or nonlinear relationships in data, while the regression line is used to highlight a linear trend. Generally speaking, we would choose only one of these to display on our graph. In this case, since the pattern is fairly linear (the lowess smooth is fairly straight), we would choose the solid regression line. In Chapter 12 we will learn how to identify the equation of the regression line that best summarizes the data and determine if the apparent trend in the data is likely to be just due to chance or if there is evidence for a real relationship between X and Y.