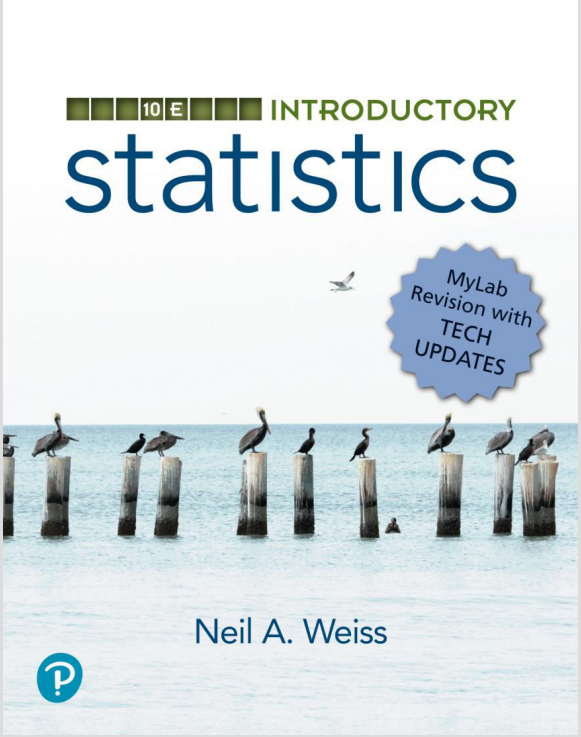
Chapter 1

The Nature of Statistics



Section 1.1

Statistics Basics

Definition 1.1

Descriptive Statistics

Descriptive Statistics consists of methods for organizing and summarizing information.

Descriptive statistics includes the construction of graphs, charts, and tables and the calculation of various descriptive measures such as averages, measures of variation, and percentiles.

Example 1.1

The 1948 Baseball Season. In 1948, the Washington Senators played 153 games, winning 56 and losing 97. They finished seventh in the American League and were led in hitting by Bud Stewart, whose batting average was .279.

The work of baseball statisticians is an illustration of *descriptive statistics*.

Definition 1.2

Population and Sample

Population: The collection of all individuals or items under consideration in a statistical study.

Sample: That part of the population from which information is obtained.

Example 1.2

Political polling provides an example of **inferential statistics**. Interviewing everyone of voting age in the United States on their voting preferences would be expensive and unrealistic. Statisticians who want to gauge the sentiment of the entire **population** of U.S. voters can afford to interview only a carefully chosen group of a few thousand voters. This group is called a **sample** of the **population**.

Definition 1.3

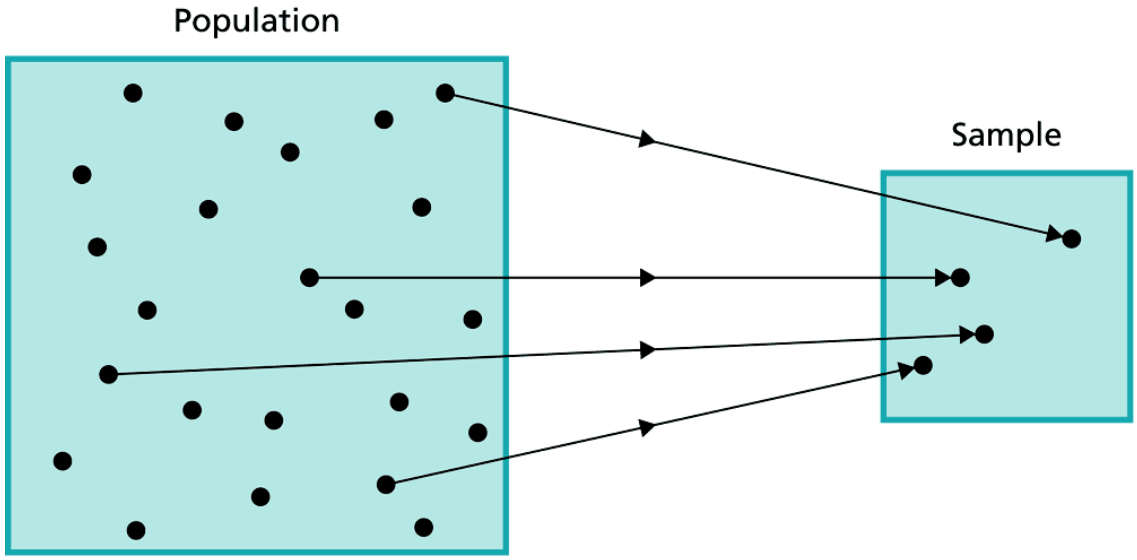
Inferential Statistics

Inferential statistics consists of methods for drawing and measuring the reliability of conclusions about a population based on information obtained from a sample of the population.

Statisticians analyze the information obtained from a **sample** of the voting **population** to make **inferences** (draw conclusions) about the preferences of the entire voting **population**. Inferential statistics provides methods for drawing such conclusions.

Figure 1.1

Relationship between population and sample



Example 1.3 Classifying Statistical Studies

The 1948 Presidential Election Table 1.1 displays the voting results for the 1948 presidential election

Ticket	Votes	Percentage
Truman–Barkley (Democratic)	24,179,345	49.7
Dewey–Warren (Republican)	21,991,291	45.2
Thurmond–Wright (States Rights)	1,176,125	2.4
Wallace–Taylor (Progressive)	1,157,326	2.4
Thomas–Smith (Socialist)	139,572	0.3

Classification This study is descriptive. It is a summary of the votes cast by U.S. voters in the 1948 presidential election. No inferences are made.

Section 1.2

Simple Random Sampling

Definition 1.4

Simple Random Sampling; Simple Random Sample

Simple random sampling: A sampling procedure for which each possible sample of a given size is equally likely to be the one obtained.

Simple random sample: A sample obtained by simple random sampling.

There are two types of **simple random sampling**. One is simple random sampling **with replacement (SRSWR)**, whereby a member of the population can be selected more than once; the other is **simple random sampling without replacement (SRS)**, whereby a member of the population can be selected at most once.

Random-Number Tables

Obtaining a simple random sample by picking slips of paper out of a box is usually impractical, especially when the population is large. Fortunately, we can use several practical procedures to get simple random samples. One common method involves a **table of random numbers** – a table of randomly chosen digits, as illustrated in Table 1.5.

Table 1.5

Random numbers

Line number	Column number									
	00–09		*10–19*		*20–29*		*30–39*		*40–49*	
00	15544	80712	97742	21500	97081	42451	50623	56071	28882	28739
01	01011	21285	04729	39986	73150	31548	30168	76189	56996	19210
02	47435	53308	40718	29050	74858	64517	93573	51058	68501	42723
03	91312	75137	86274	59834	69844	19853	06917	17413	44474	86530
04	12775	08768	80791	16298	22934	09630	98862	39746	64623	32768
05	31466	43761	94872	92230	52367	13205	38634	55882	77518	36252
06	09300	43847	40881	51243	97810	18903	53914	31688	06220	40422
07	73582	13810	57784	72454	68997	72229	30340	08844	53924	89630
08	11092	81392	58189	22697	41063	09451	09789	00637	06450	85990
09	93322	98567	00116	35605	66790	52965	62877	21740	56476	49296
10	80134	12484	67089	08674	70753	90959	45842	59844	45214	36505
11	97888	31797	95037	84400	76041	96668	75920	68482	56855	97417
12	92612	27082	59459	69380	98654	20407	88151	56263	27126	63797
13	72744	45586	43279	44218	83638	05422	00995	70217	78925	39097
14	96256	70653	45285	26293	78305	80252	03625	40159	68760	84716
15	07851	47452	66742	83331	54701	06573	98169	37499	67756	68301
16	25594	41552	96475	56151	02089	33748	65289	89956	89559	33687
17	65358	15155	59374	80940	03411	94656	69440	47156	77115	99463
18	09402	31008	53424	21928	02198	61201	02457	87214	59750	51330
19	97424	90765	01634	37328	41243	33564	17884	94747	93650	77668

Random-Number Generators

Nowadays, statisticians prefer statistical software packages or graphing calculators, rather than random-number tables, to obtain simple random samples. The built-in programs for doing so are called **random-number generators**. When using random-number generators, be aware of whether they provide samples with replacement or samples without replacement.

Section 1.3
Other Sampling Designs

Procedure 1.1

Systematic Random Sampling

Step 1 Divide the population size by the sample size and round the result down to the nearest whole number, m.

Step 2 Use a random-number table or a similar device to obtain a number, k, between 1 and m.

Step 3 Select for the sample those members of the population that are numbered $k, k + m, k + 2m, \ldots$.

Procedure 1.2

Cluster Sampling

Step 1 **Divide the population into groups (clusters).**

Step 2 **Obtain a simple random sample of the clusters.**

Step 3 **Use all the members of the clusters obtained in Step 2 as the sample.**

Procedure 1.3

Stratified Random Sampling with Proportional Allocation

Step 1 **Divide the population into subpopulations (strata).**

Step 2 **From each stratum, obtain a simple random sample of size proportional to the size of the stratum; that is, the sample size for a stratum equals the total sample size times the stratum size divided by the population size.**

Step 3 **Use all the members obtained in Step 2 as the sample.**

Section 1.4

Experimental Designs

Definition 1.5

Experimental Units; Subjects

In a designed experiment, the individuals or items on which the experiment is performed are called **experimental units.** When the experimental units are humans, the term **subject** is often used in place of experimental unit.

Folic Acid and Birth Defects. For the study, the doctors enrolled 4753 women prior to conception, and divided them randomly into two groups. One group took daily multivitamins containing 0.8 mg of folic acid, whereas the other group received only trace elements. In the language of experimental design, each woman in the folic acid study is an **experimental unit**, or a **subject**.

Key Fact 1.1

Principles of Experimental Design

The following principles of experimental design enable a researcher to conclude that differences in the results of an experiment not reasonably attributable to chance are likely caused by the treatments.

- **Control:** Two or more treatments should be compared.
- **Randomization:** The experimental units should be randomly divided into groups to avoid unintentional selection bias in constituting the groups.
- **Replication:** A sufficient number of experimental units should be used to ensure that randomization creates groups that resemble each other closely and to increase the chances of detecting any differences among the treatments.

Folic Acid and Birth Defects

· ***Control***: The doctors compared the rate of major birth defects for the women who took folic acid to that for the women who took only trace elements.

· ***Randomization***: The women were divided randomly into two groups to avoid unintentional selection bias.

· ***Replication***: A large number of women were recruited for the study to make it likely that the two groups created by randomization would be similar and also to increase the chances of detecting any effect due to the folic acid.

Folic Acid and Birth Defects

One of the most common experimental situations involves a specified treatment and *placebo*, an inert or innocuous medical substance. Technically, both the specified treatment and placebo are treatments. The group receiving the specified treatment is called the **treatment group**, and the group receiving placebo is called the **control group**. In the folic acid study, the women who took folic acid constituted the **treatment group** and those who took only trace elements constituted the **control group**.

Definition 1.6

Response Variable, Factors, Levels, and Treatments

Response variable: The characteristic of the experimental outcome that is to be measured or observed.

Factor: A variable whose effect on the response variable is of interest in the experiment.

Levels: The possible values of a factor.

Treatment: Each experimental condition. For one-factor experiments, the treatments are the levels of the single factor. For multifactor experiments, each treatment is a combination of levels of the factors.

Example 1.15 Experimental Design

Weight Gain of Golden Torch Cacti

The Golden Torch Cactus (*Trichocereus spachianus*), a cactus native to Argentina, has excellent landscape potential. W. Feldman and F. Crosswhite, two researchers at the Boyce Thompson Southwestern Arboretum, investigated the optimal method for producing these cacti. The researchers examined, among other things, the effects of a hydrophilic polymer and irrigation regime on weight gain. Hydrophilic polymers are used as soil additives to keep moisture in the root zone. For this study, the researchers chose Broadleaf P-4 polyacrylamide, abbreviated P4. The hydrophilic polymer was either used or not used, and five irrigation regimes were employed: none, light, medium, heavy, and very heavy.

Example 1.15 Experimental Design

Weight Gain of Golden Torch Cacti

Identify the

a. experimental units.
b. response variable.
c. factors.
d. levels of each factor.
e. treatments.

Example 1.15 Experimental Design
Weight Gain of Golden Torch Cacti

Solution

a. The experimental units are the cacti used in the study.
b. The response variable is weight gain.
c. The factors are hydrophilic polymer and irrigation regime.
d. Hydrophilic polymer has two levels: with and without. Irrigation regime has five levels: none, light, medium, heavy, and very heavy.
e. Each treatment is a combination of a level of hydrophilic polymer and a level of irrigation regime. Table 1.8 depicts the 10 treatments for this experiment. In the table, we abbreviated "very heavy" as "Xheavy."

Table 1.8

Schematic for the 10 treatments in the cactus study

		Irrigation regime				
		None	Light	Medium	Heavy	Xheavy
Polymer	No P4	No water No P4 (Treatment 1)	Light water No P4 (Treatment 2)	Medium water No P4 (Treatment 3)	Heavy water No P4 (Treatment 4)	Xheavy water No P4 (Treatment 5)
	With P4	No water With P4 (Treatment 6)	Light water With P4 (Treatment 7)	Medium water With P4 (Treatment 8)	Heavy water With P4 (Treatment 9)	Xheavy water With P4 (Treatment 10)

Definition 1.7

Completely Randomized Design

In a **completely randomized design,** all the experimental units are assigned randomly among all the treatments.

Once we have chosen the treatments, we must decide how the experimental units are to be assigned to the treatments (or vice versa). The women in the folic acid study were randomly divided into two groups; one group received folic acid and the other only trace elements. In the cactus study, 40 cacti were divided randomly into 10 groups of four cacti each and then each group was assigned a different treatment from among the 10 depicted in Table 1.8. Both of these experiments used a **completely randomized design**.

Definition 1.8

Randomized Block Design

In a **randomized block design,** the experimental units are assigned randomly among all the treatments separately within each block.

Although the completely randomized design is commonly used and simple, it is not always the best design. Several alternatives to that design exist. For instance, in a **randomized block design**, experimental units that are similar in ways that are expected to affect the response variable are grouped in **blocks**. Then the random assignment of experimental units to the treatments is made block by block.

Example 1.16 Statistical Designs

Golf Ball Driving Distances

Suppose we want to compare the driving distances for five different brands of golf ball. For 40 golfers, discuss a method of comparison based on

a. a completely randomized design.
b. a randomized block design.

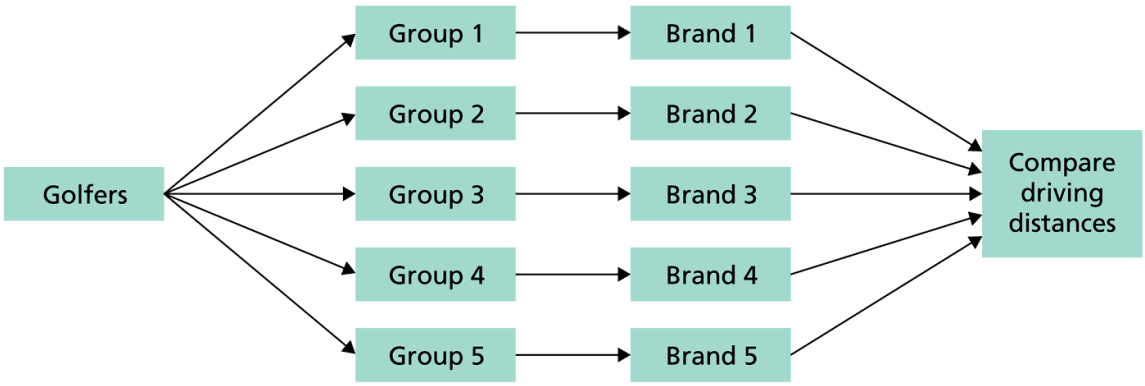
Solution

Here the experimental units are the golfers, the response variable is driving distance, the factor is brand of golf ball, and the levels (and treatments) are the five brands.

a. For a completely randomized design, we would randomly divide the 40 golfers into five groups of 8 golfers each and then randomly assign each group to drive a different brand of ball, as illustrated in Fig.1.5.

Figure 1.5

Completely randomized design for golf ball experiment



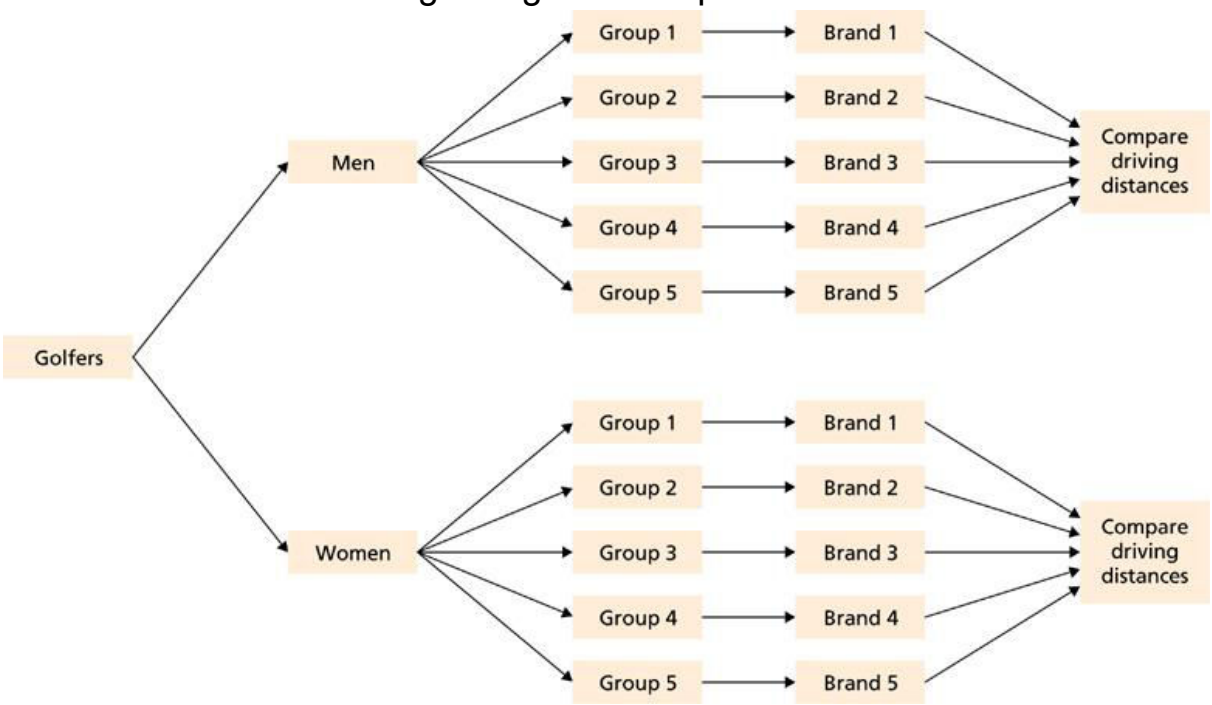
Example 1.16 Statistical Designs

Golf Ball Driving Distances

b. Because driving distance is affected by gender, using a randomized block design that blocks by gender is probably a better approach. We could do so by using 20 men golfers and 20 women golfers. We would randomly divide the 20 men into five groups of 4 men each and then randomly assign each group to drive a different brand of ball, as shown in Fig.1.6. Likewise, we would randomly divide the 20 women into five groups of 4 women each and then randomly assign each group to drive a different brand of ball, as also shown in Fig.1.6.

Figure 1.6

Randomized block design for golf ball experiment

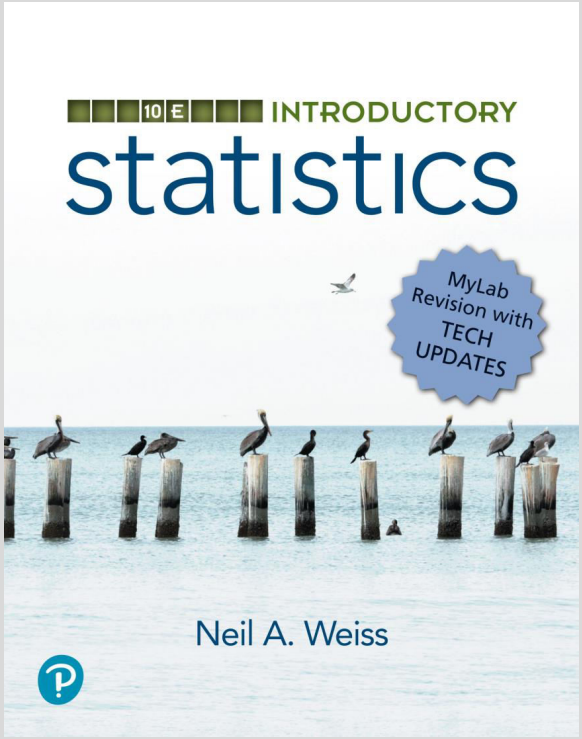


By blocking, we can isolate and remove the variation in driving distances between men and women and thereby make it easier to detect any differences in driving distances among the five brands of golf ball. Additionally, blocking permits us to analyze separately the differences in driving distances among the five brands for men and women.

As illustrated in Example 1.16, blocking can isolate and remove systematic differences among blocks, thereby making any differences among treatments easier to detect. Blocking also makes possible the separate analysis of treatment effects on each block.

Chapter 2

Organizing Data



Chapter 2

Organizing Data

Section 2.1

Variables and Data

Definition 2.1

Variables

Variable: A characteristic that varies from one person or thing to another.

Qualitative variable: A nonumerically valued variable.

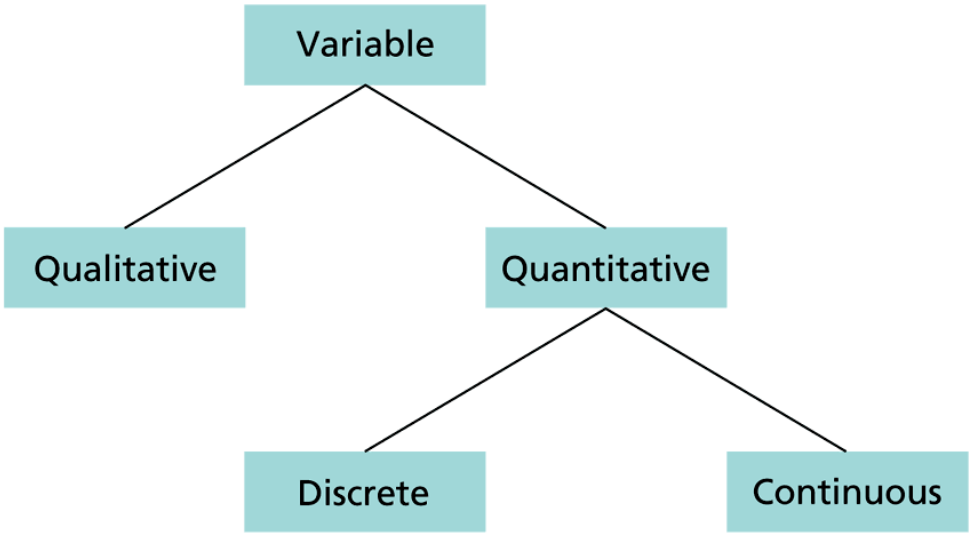
Quantitative variable: A numerically valued variable.

Discrete variable: A quantitative variable whose possible values can be listed. In particular, a quantitative variable with only a finite number of possible values is a discrete variable.

Continuous variable: A quantitative variable whose possible values form some interval of numbers.

Figure 2.1

Types of variables



Definition 2.2

Data

Data: Values of a variable.

Qualitative data: Values of a qualitative variable.

Quantitative data: Values of a quantitative variable.

Discrete data: Values of a discrete variable.

Continuous data: Values of a continuous variable.

Section 2.2

Organizing Qualitative Data

Definition 2.3

Frequency Distribution of Qualitative Data

A **frequency distribution** of qualitative data is a listing of the distinct values and their frequencies.

Procedure 2.1

To Construct a Frequency Distribution of Qualitative Data

Step 1 List the distinct values of the observations in the data set in the first column of a table.

Step 2 For each observation, place a tally mark in the second column of the table in the row of the appropriate distinct value.

Step 3 Count the tallies for each distinct value and record the totals in the third column of the table.

Table 2.1

Political party affiliations of the students in introductory statistics

Democratic	Other	Democratic	Other	Democratic
Republican	Republican	Other	Other	Republican
Republican	Republican	Republican	Democratic	Republican
Republican	Democratic	Democratic	Other	Republican
Democratic	Democratic	Republican	Democratic	Democratic
Republican	Republican	Other	Other	Democratic
Republican	Democratic	Republican	Other	Other
Republican	Republican	Republican	Democratic	Republican

Table 2.2

Table for constructing a frequency distribution for the political party affiliation data in Table 2.1

Party	Tally	Frequency
Democratic	𝍸 𝍸 |||	13
Republican	𝍸 𝍸 𝍸 |||	18
Other	𝍸 ||||	9
		40

Definition 2.4

Relative-Frequency Distribution of Qualitative Data

A **relative-frequency distribution** of qualitative data is a listing of the distinct values and their relative frequencies.

Procedure 2.2

To Construct a Relative-Frequency Distribution of Qualitative Data

Step 1 Obtain a frequency distribution of the data.

Step 2 Divide each frequency by the total number of observations.

Table 2.3

Relative-frequency distribution for the political party affiliation data in Table 2.1

Party	Relative frequency	
Democratic	0.325	← *13/40*
Republican	0.450	← *18/40*
Other	0.225	← *9/40*
	1.000	

Definition 2.5

Pie Chart

A **pie chart** is a disk divided into wedge-shaped pieces proportional to the relative frequencies of the qualitative data.

Procedure 2.3

To Construct a Pie Chart

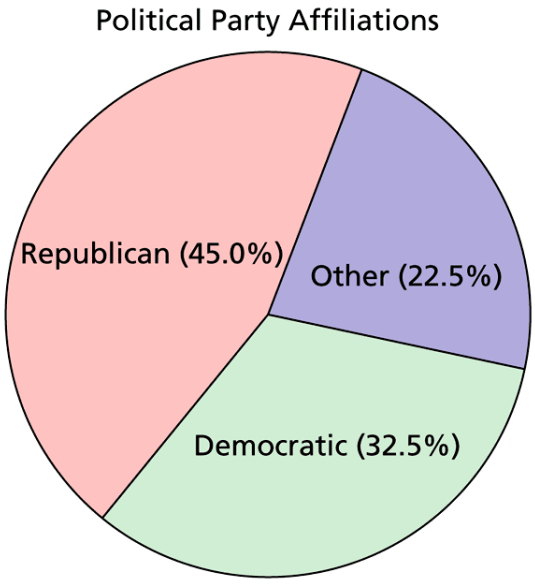
Step 1 Obtain a relative-frequency distribution of the data by applying Procedure 2.2.

Step 2 Divide a disk into wedge-shaped pieces proportional to the relative frequencies.

Step 3 Label the slices with the distinct values and their relative frequencies.

Figure 2.2

Pie chart of the political party affiliation data in Table 2.1



Definition 2.6

Bar Chart

A **bar chart** displays the distinct values of the qualitative data on a horizontal axis and the relative frequencies (or frequencies or percents) of those values on a vertical axis. The relative frequency of each distinct value is represented by a vertical bar whose height is equal to the relative frequency of that value. The bars should be positioned so that they do not touch each other.

Procedure 2.4

To Construct a Bar Chart

Step 1 Obtain a relative-frequency distribution of the data by applying Procedure 2.2.

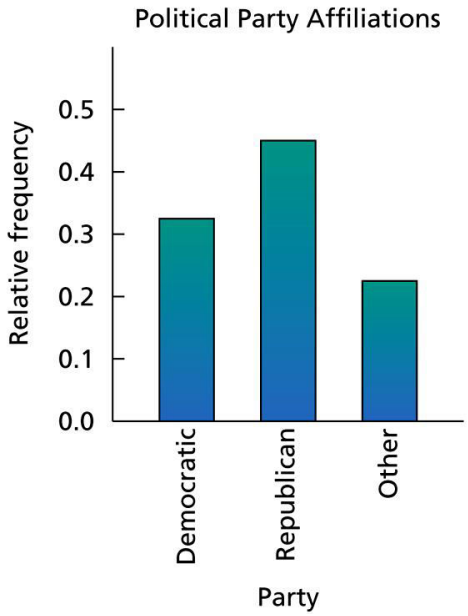
Step 2 Draw a horizontal axis on which to place the bars and a vertical axis on which to display the relative frequencies.

Step 3 For each distinct value, construct a vertical bar whose height equals the relative frequency of that value.

Step 4 Label the bars with the distinct values, the horizontal axis with the name of the variable, and the vertical axis with "Relative frequency."

Figure 2.3

Bar chart of the political party affiliation data in Table 2.1



Section 2.3

Organizing Quantitative Data

Table 2.4

Number of TV sets in each of 50 randomly selected households.

1	1	1	2	6	3	3	4	2	4
3	2	1	5	2	1	3	6	2	2
3	1	1	4	3	2	2	2	2	3
0	3	1	2	1	2	3	1	1	3
3	2	1	2	1	1	3	1	5	1

Table 2.5

Frequency and relative-frequency distributions, using single-value grouping, for the number-of-TVs data in Table 2.4

Number of TVs	Frequency	Relative frequency
0	1	0.02
1	16	0.32
2	14	0.28
3	12	0.24
4	3	0.06
5	2	0.04
6	2	0.04
	50	1.00

Table 2.6

Days to maturity for 40 short-term investments

70	64	99	55	64	89	87	65
62	38	67	70	60	69	78	39
75	56	71	51	99	68	95	86
57	53	47	50	55	81	80	98
51	36	63	66	85	79	83	70

Table 2.7

Frequency and relative-frequency distributions, using limit grouping, for the days-to-maturity data in Table 2.6

Days to maturity	Tally	Frequency	Relative frequency
30–39	III	3	0.075
40–49	I	1	0.025
50–59	𝍸 III	8	0.200
60–69	𝍸 𝍸	10	0.250
70–79	𝍸 II	7	0.175
80–89	𝍸 II	7	0.175
90–99	IIII	4	0.100
		40	1.000

Definition 2.7

Terms Used in Limit Grouping

Lower class limit: The smallest value that could go in a class.

Upper class limit: The largest value that could go in a class.

Class width: The difference between the lower limit of a class and the lower limit of the next-higher class.

Class mark: The average of the two class limits of a class.

Definition 2.8

Terms Used in Cutpoint Grouping

Lower class cutpoint: The smallest value that could go in a class.

Upper class cutpoint: The largest value that could go in the next-higher class (equivalent to the lower cutpoint of the next-higher class).

Class width: The difference between the cutpoints of a class.

Class midpoint: The average of the two cutpoints of a class.

Choosing the Grouping Method

Grouping method	When to use
Single-value grouping	Use with discrete data in which there are only a small number of distinct values.
Limit grouping	Use when the data are expressed as whole numbers and there are too many distinct values to employ single-value grouping.
Cutpoint grouping	Use when the data are continuous and are expressed with decimals.

Definition 2.9

Histogram

A **histogram** displays the classes of the quantitative data on a horizontal axis and the frequencies (relative frequencies, percents) of those classes on a vertical axis. The frequency (relative frequency, percent) of each class is represented by a vertical bar whose height is equal to the frequency (relative frequency, percent) of that class. The bars should be positioned so that they touch each other.

- For single-value grouping, we use the distinct values of the observations to label the bars, with each such value centered under its bar.
- For limit grouping or cutpoint grouping, we use the lower class limits (or, equivalently, lower class cutpoints) to label the bars. Note: Some statisticians and technologies use class marks or class midpoints centered under the bars.

Procedure 2.5

To Construct a Histogram

Step 1 Obtain a frequency (relative-frequency, percent) distribution of the data.

Step 2 Draw a horizontal axis on which to place the bars and a vertical axis on which to display the frequencies (relative frequencies, percents).

Step 3 For each class, construct a vertical bar whose height equals the frequency (relative frequency, percent) of that class.

Step 4 Label the bars with the classes, as explained in Definition 2.9, the horizontal axis with the name of the variable, and the vertical axis with "Frequency" ("Relative frequency," "Percent").

Figure 2.4

Single-value grouping. Number of TVs per household:
(a) frequency histogram; (b) relative-frequency histogram

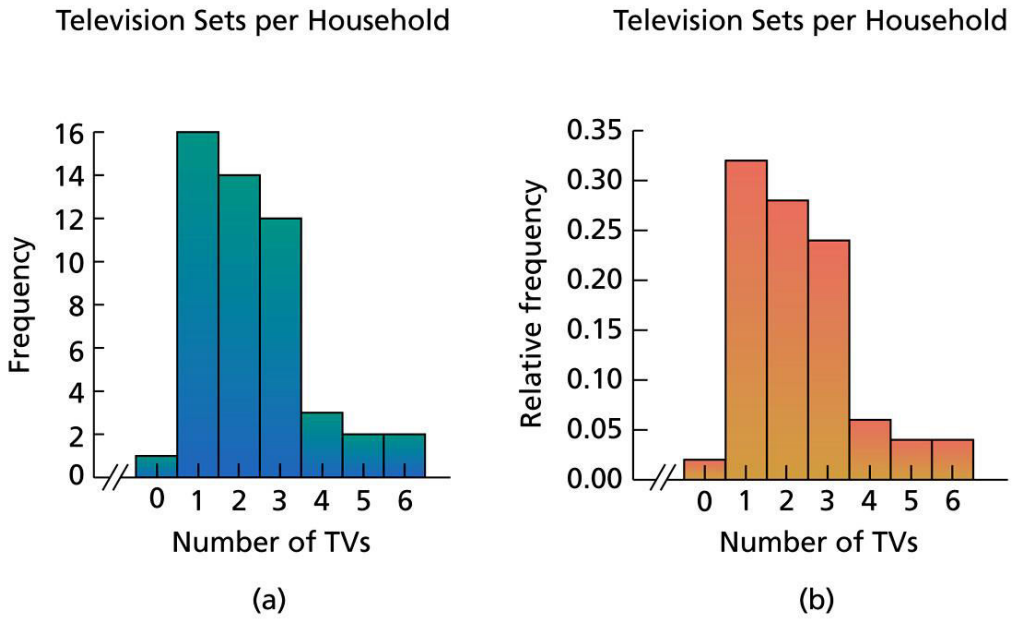


Figure 2.5

Limit grouping. Days to maturity: (a) frequency histogram; (b) relative-frequency histogram

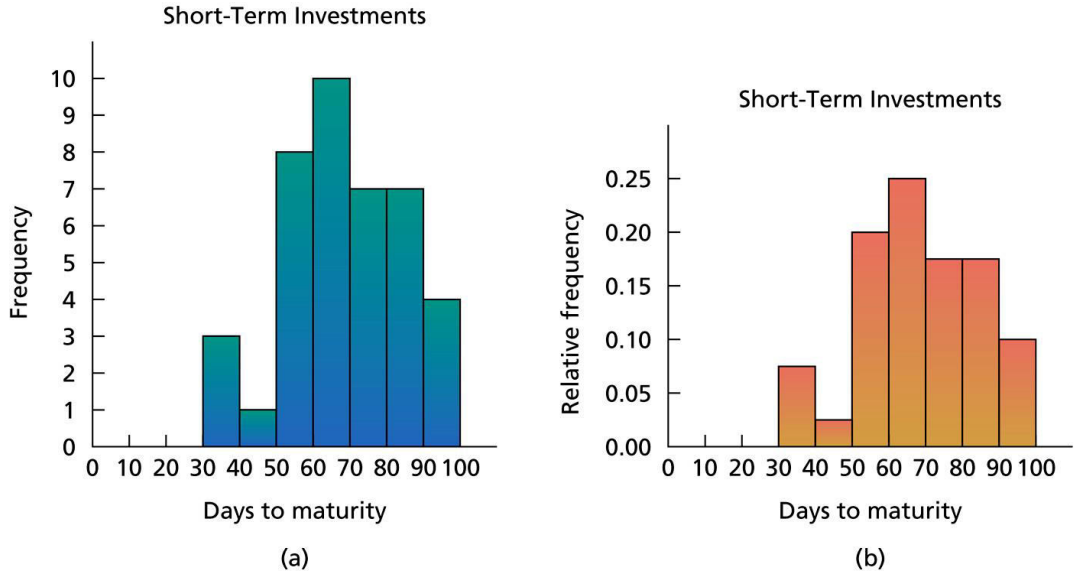
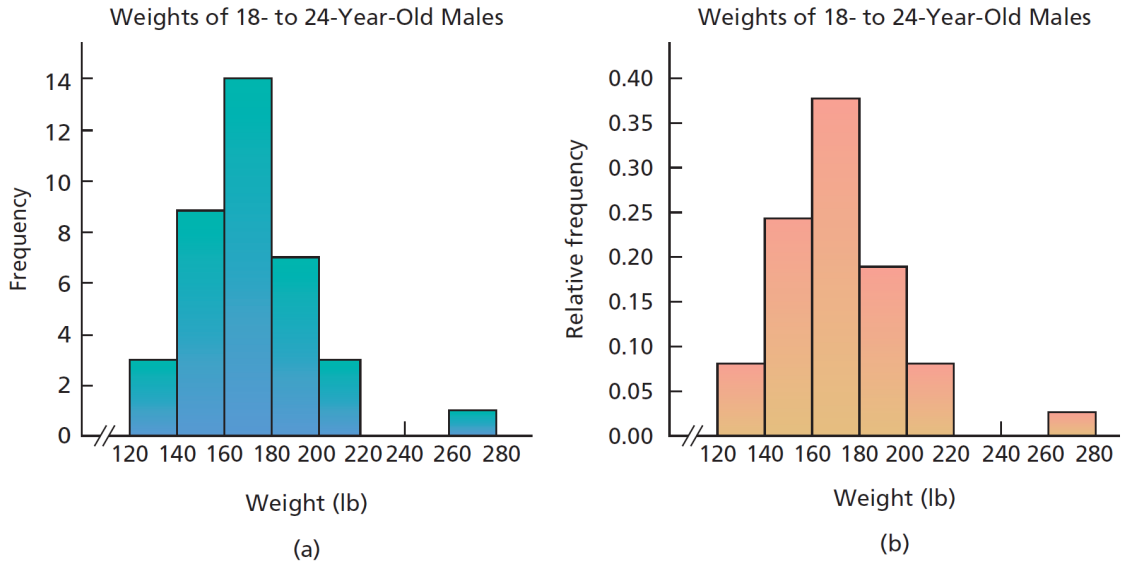


Figure 2.6

Cutpoint grouping. Weight of 18- to 24-year old males: (a) frequency histogram; (b) relative-frequency histogram



Definition 2.10

Dotplot

A **dotplot** is a graph in which each observation is plotted as a dot at an appropriate place above a horizontal axis. Observations having equal values are stacked vertically.

Procedure 2.6

To Construct a Dotplot

Step 1 Draw a horizontal axis that displays the possible values of the quantitative data.

Step 2 Record each observation by placing a dot over the appropriate value on the horizontal axis.

Step 3 Label the horizontal axis with the name of the variable.

Table 2.11 & Figure 2.7

Prices, in dollars, of 16 DVD players

210	219	214	197
224	219	199	199
208	209	215	199
212	212	219	210

Definition 2.11

Stem-and-Leaf Diagrams

In a **stem-and-leaf diagram** (or **stemplot**), each observation Is separated into two parts, namely, a **stem**—consisting of all but the rightmost digit– and a **leaf**, the rightmost digit.

Procedure 2.7

To Construct a Stem-and-Leaf Diagram

Step 1 Think of each observation as a stem—consisting of all but the rightmost digit—and a leaf, the rightmost digit.

Step 2 Write the stems from smallest to largest in a vertical column to the left of a vertical rule.

Step 3 Write each leaf to the right of the vertical rule in the row that contains the appropriate stem.

Step 4 Arrange the leaves in each row in ascending order.

Table 2.12 & Figure 2.8

Days to maturity for
40 short-term investments

70	64	99	55	64	89	87	65
62	38	67	70	60	69	78	39
75	56	71	51	99	68	95	86
57	53	47	50	55	81	80	98
51	36	63	66	85	79	83	70

Constructing a stem-and-leaf diagram
for the days-to-maturity data

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

(a)

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

(b)

Table 2.13 & Figure 2.9

Cholesterol levels
for 20 high-level patients

210	209	212	208
217	207	210	203
208	210	210	199
215	221	213	218
202	218	200	214

Stem-and-leaf diagram for cholesterol levels:
(a) one line per stem; (b) two lines per stem

19	9
20	0 2 3 7 8 8 9
21	0 0 0 0 2 3 4 5 7 8 8
22	1

(a)

19	
19	9
20	0 2 3
20	7 8 8 9
21	0 0 0 0 2 3 4
21	5 7 8 8
22	1
22	

(b)

Section 2.4

Distribution Shapes

Definition 2.12

Distribution of a Data Set

The **distribution of a data set** is a table, graph, or formula that provides the values of the observations and how often they occur.

Figure 2.10

Relative-frequency histogram and approximating smooth curve for the distribution of heights

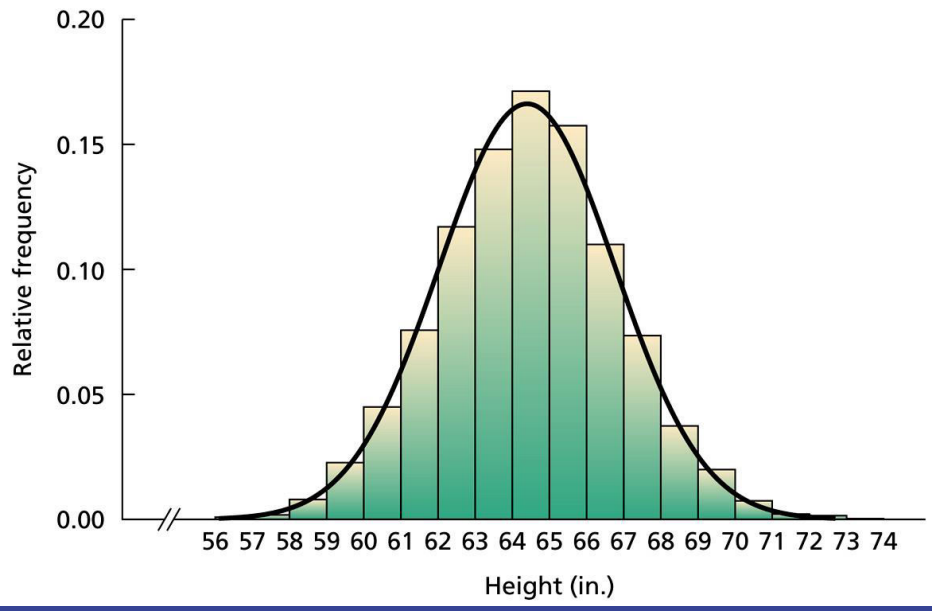
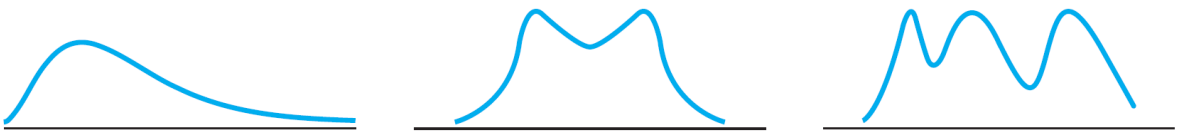


Figure 2.11

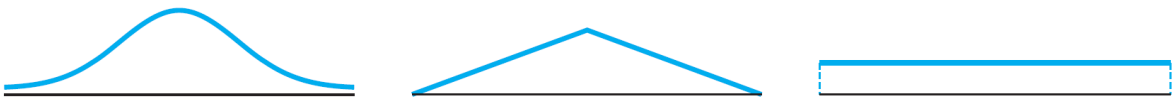
Examples of (a) unimodal, (b) bimodal, and (c) multimodal distributions



(a) Unimodal (b) Bimodal (c) Multimodal

Figure 2.12

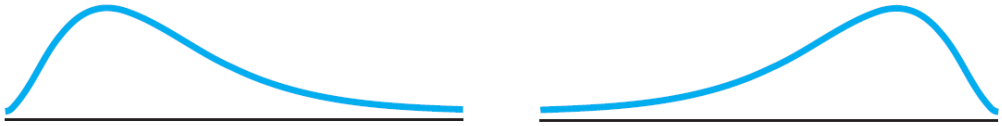
Examples of symmetric distributions: (a) bell shaped, (b) triangular, and (c) uniform



(a) Bell shaped (b) Triangular (c) Uniform (or rectangular)

Figure 2.13

Generic skewed distributions: (a) right skewed (b) left skewed



(a) Right skewed

(b) Left skewed

Figure 2.14

Reverse-J-shaped distribution

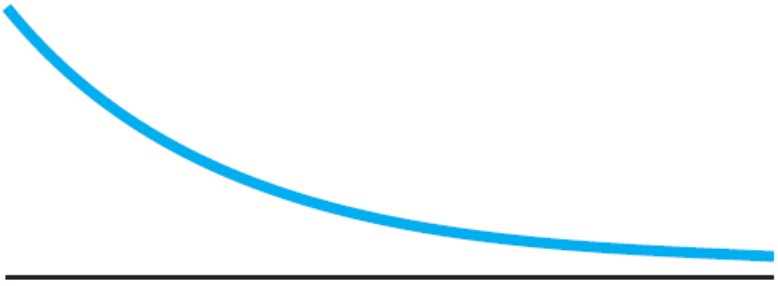
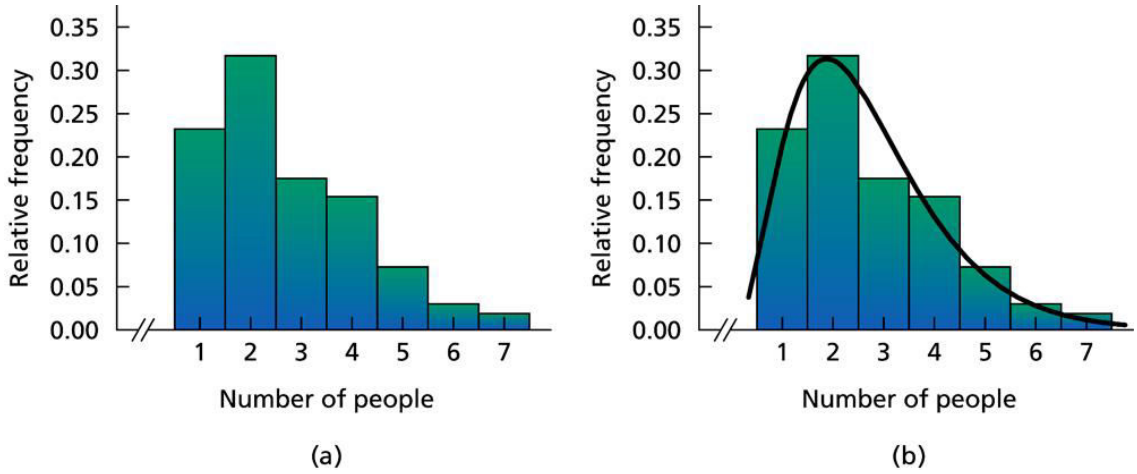


Figure 2.15

Relative-frequency histogram for household size



Definition 2.13

Population and Sample Data

Population data: The values of a variable for the entire population.

Sample data: The values of a variable for a sample of the population.

Definition 2.14

Population and Sample Distributions; Distribution of a Variable

The distribution of population data is called the **population distribution,** or the **distribution of the variable.**

The distribution of sample data is called a **sample distribution.**

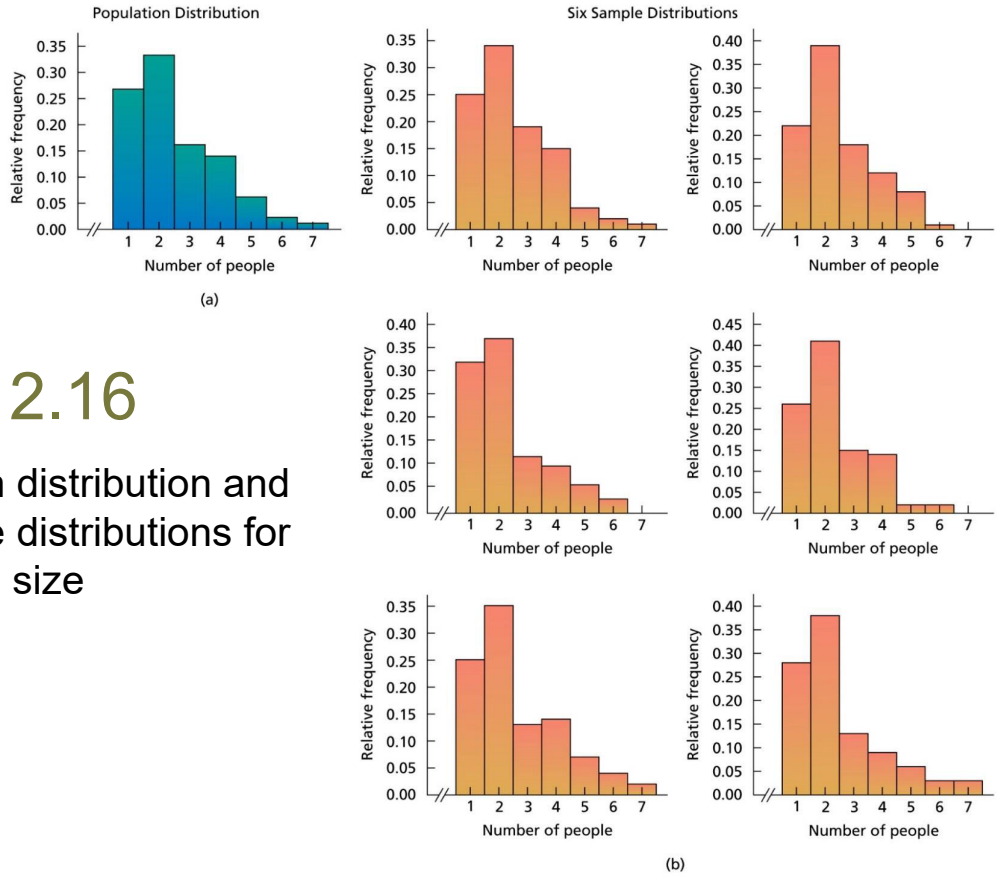


Figure 2.16

Population distribution and six sample distributions for household size

Key Fact 2.1

Population and Sample Distributions

For a simple random sample, the sample distribution approximates the population distribution (i.e., the distribution of the variable under consideration). The larger the sample size, the better the approximation tends to be.

Section 2.5

Misleading Graphs

Figure 2.17

Unemployment rates: (a) truncated graph; (b) nontruncated graph

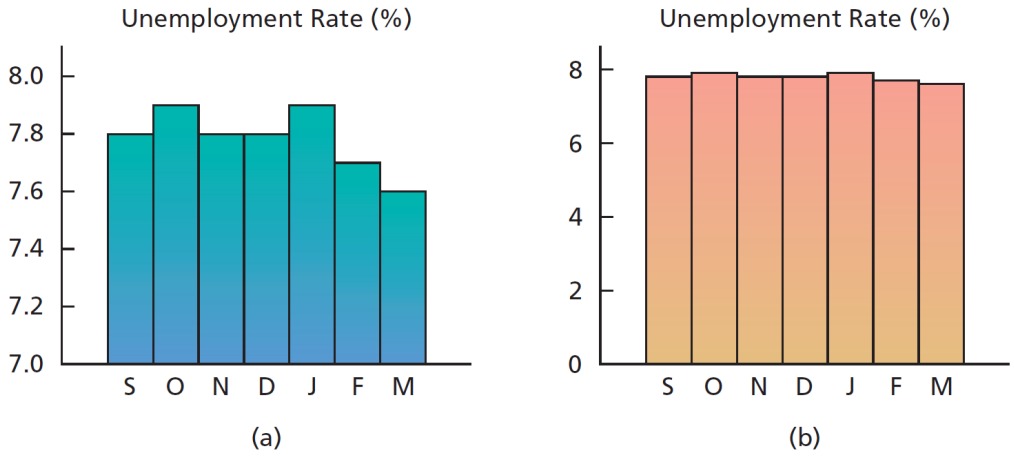
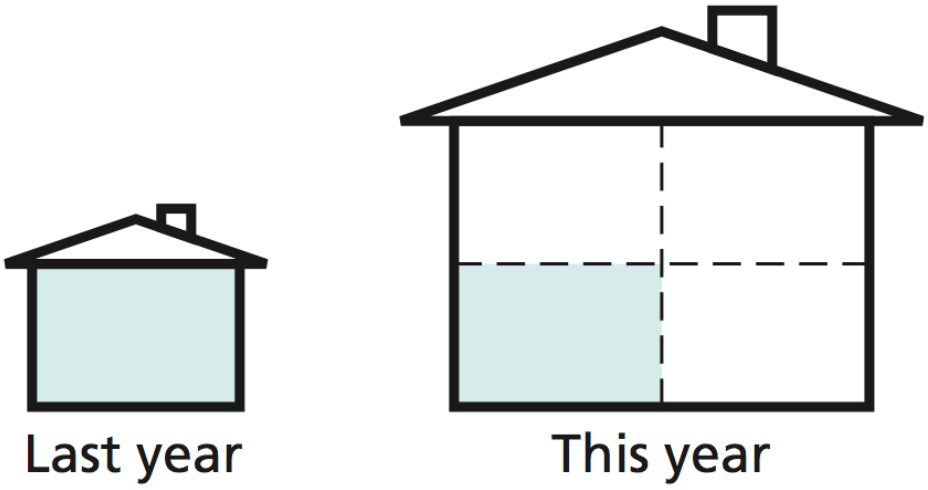


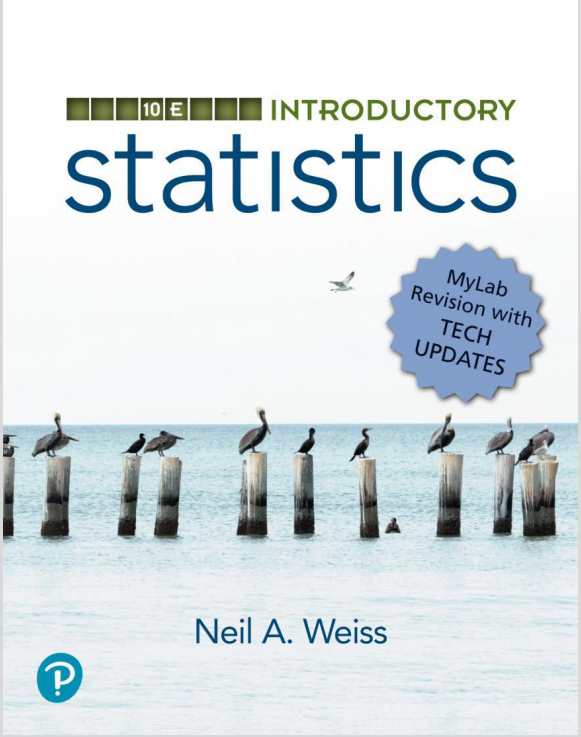
Figure 2.19

Improper scaling: Number of homes this year will be double last year, so the developer doubled the width and height, which makes it look like four times the number of homes will be built.



Chapter 3

Descriptive Measures



Chapter 3

Descriptive Measures

Section 3.1

Measures of Center

Definition 3.1

Mean of a Data Set

The **mean** of a data set is the sum of the observations divided by the number of observations.

Definition 3.2

Median of a Data Set

Arrange the data in increasing order.

- If the number of observations is odd, then the **median** is the observation exactly in the middle of the ordered list.
- If the number of observations is even, then the **median** is the mean of the two middle observations in the ordered list.

In both cases, if we let n denote the number of observations, then the median is at position $(n + 1) / 2$ in the ordered list.

Definition 3.3

Mode of a Data Set

Find the frequency of each value in the data set.

- If no value occurs more than once, then the data set has *no mode.*
- Otherwise, any value that occurs with the greatest frequency is a **mode** of the data set.

Tables 3.1, 3.2 & 3.4

Data Set I

$300	300	300	940	300
300	400	300	400	
450	800	450	1050	

Data Set II

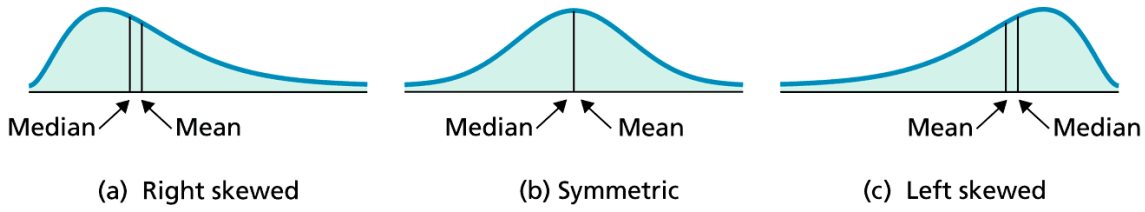
$300	300	940	450	400
400	300	300	1050	300

Means, medians, and modes of salaries in Data Set I and Data Set II

Measure of center	Definition	Data Set I	Data Set II
Mean	$\frac{\text{Sum of observations}}{\text{Number of observations}}$	$483.85	$474.00
Median	Middle value in ordered list	$400.00	$350.00
Mode	Most frequent value	$300.00	$300.00

Figure 3.1

Relative positions of the mean and median for (a) right-skewed, (b) symmetric, and (c) left-skewed distributions



(a) Right skewed (b) Symmetric (c) Left skewed

Definition 3.4

Sample Mean

For a variable x, the mean of the observations for a sample is called a **sample mean** and is denoted $\bar{x}$. Symbolically,

$$\bar{x} = \frac{\Sigma x_i}{n},$$

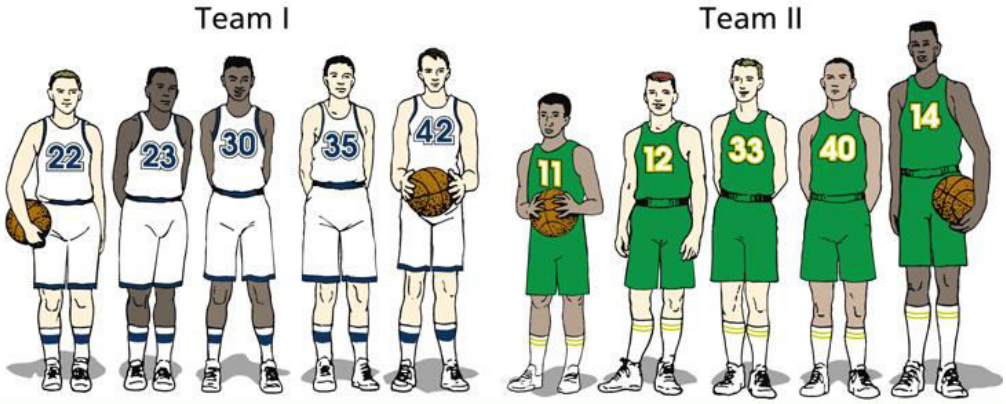
where n is the sample size.

Section 3.2

Measures of Variation

Figure 3.2

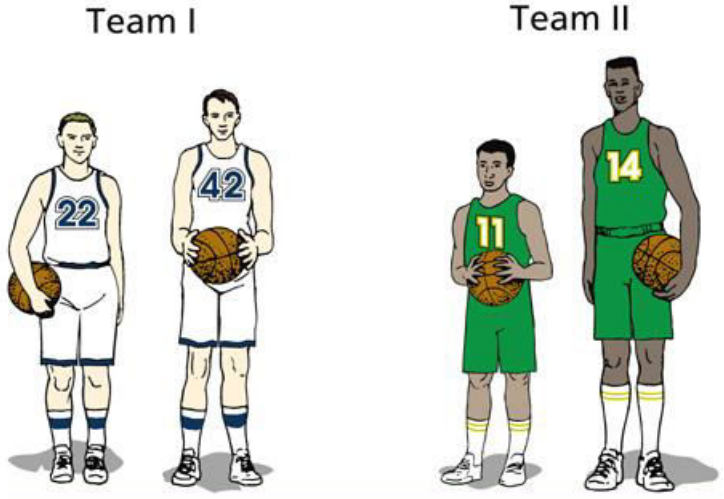
Five starting players on two basketball teams



Feet and inches	6'	6'1"	6'4"	6'4"	6'6"	5'7"	6'	6'4"	6'4"	7'
Inches	72	73	76	76	78	67	72	76	76	84

Figure 3.3

Shortest and tallest starting players on the teams



Feet and inches	6'	6'6"	5'7"	7'
Inches	72	78	67	84

Definition 3.5

Range of a Data Set

The **range** of a data set is given by the formula

$$\text{Range} = \text{Max} - \text{Min},$$

where Max and Min denote the maximum and minimum observations, respectively.

Definition 3.6

Sample Standard Deviation

For a variable x, the standard deviation of the observations for a sample is called a **sample standard deviation.** It is denoted s_x or, when no confusion will arise, simply s. We have

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}},$$

where n is the sample size and $\bar{x}$ is the sample mean.

Key Fact 3.1

Variation and the Standard Deviation

The more variation that there is in a data set, the larger is its standard deviation.

Formula 3.1

Computing Formula for a Sample Standard Deviation

A sample standard deviation can be computed using the formula

$$s = \sqrt{\frac{\Sigma x_i^2 - (\Sigma x_i)^2/n}{n-1}},$$

where n is the sample size.

Tables 3.10 & 3.11

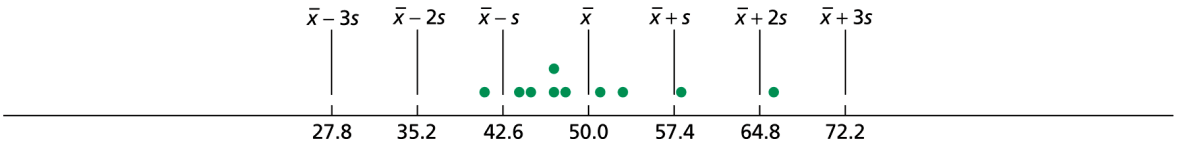
Data sets that have different variation

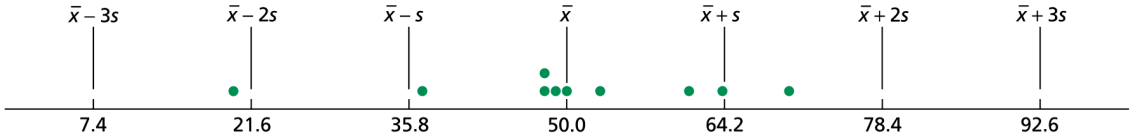
Data Set I	41	44	45	47	47	48	51	53	58	66
Data Set II	20	37	48	48	49	50	53	61	64	70

Means and standard deviations of the data sets in Table 3.10

Data Set I	Data Set II
$\bar{x} = 50.0$ $s = 7.4$	$\bar{x} = 50.0$ $s = 14.2$

Figures 3.5 and 3.6





Key Fact 3.2

Three-Standard-Deviations Rule

Almost all the observations in any data set lie within three standard deviations to either side of the mean.

Section 3.3

Chebyshev's Rule and the Empirical Rule

Key Fact 3.3

Chebyshev's Rule

For any quantitative data set and any real number k greater than or equal to 1, at least $1 - 1/k^2$ of the observations lie within k standard deviations to either side of the mean, that is, between $\bar{x} - k \cdot s$ and $\bar{x} + k \cdot s$.

Key Fact 3.4

Empirical Rule

For any quantitative data set with roughly a bell-shaped distribution, the following properties hold.

Property 1: Approximately 68% of the observations lie within one standard deviation to either side of the mean, that is, between $\bar{x} - s$ and $\bar{x} + s$.

Property 2: Approximately 95% of the observations lie within two standard deviations to either side of the mean, that is, between $\bar{x} - 2s$ and $\bar{x} + 2s$.

Property 3: Approximately 99.7% of the observations lie within three standard deviations to either side of the mean, that is, between $\bar{x} - 3s$ and $\bar{x} + 3s$.

These three properties are illustrated together in Fig. 3.9.

Figure 3.9

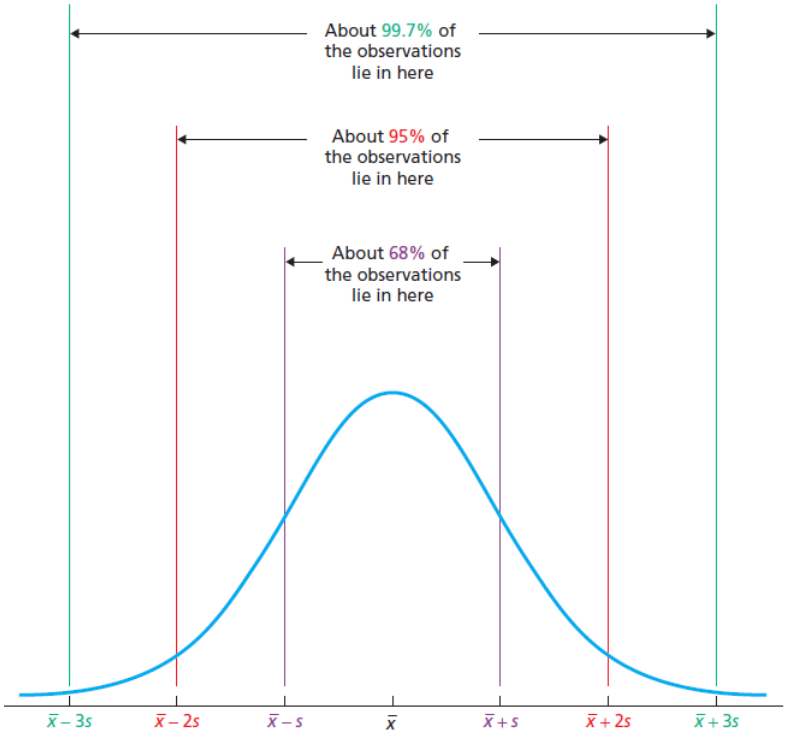


Table 3.12

PCB concentrations, in parts per million, of 60 pelican eggs

139	166	175	260	204	138	316	396	46	218
173	220	147	216	216	177	246	296	188	89
198	122	250	256	261	132	212	171	164	199
214	177	205	208	320	191	305	230	204	143
175	119	216	185	236	356	289	324	109	265
193	203	214	150	229	236	144	232	87	237

Figure 3.10

Histogram of the PCB-concentration data

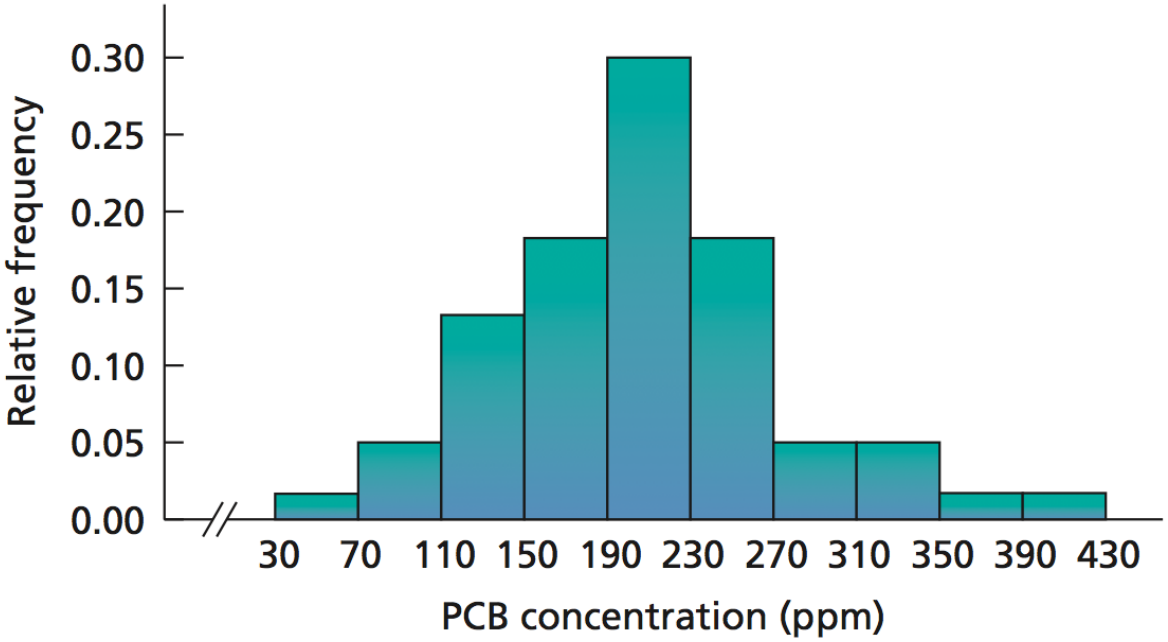
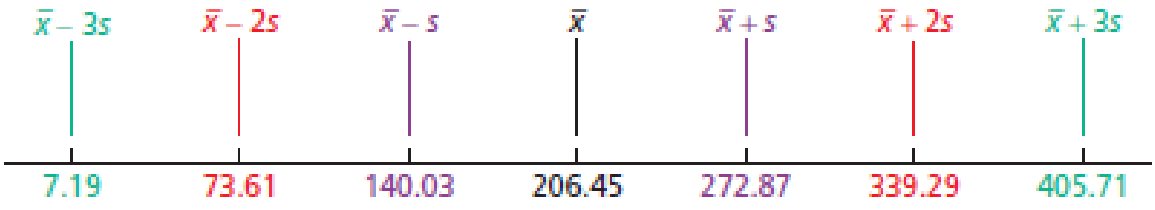


Figure 3.10

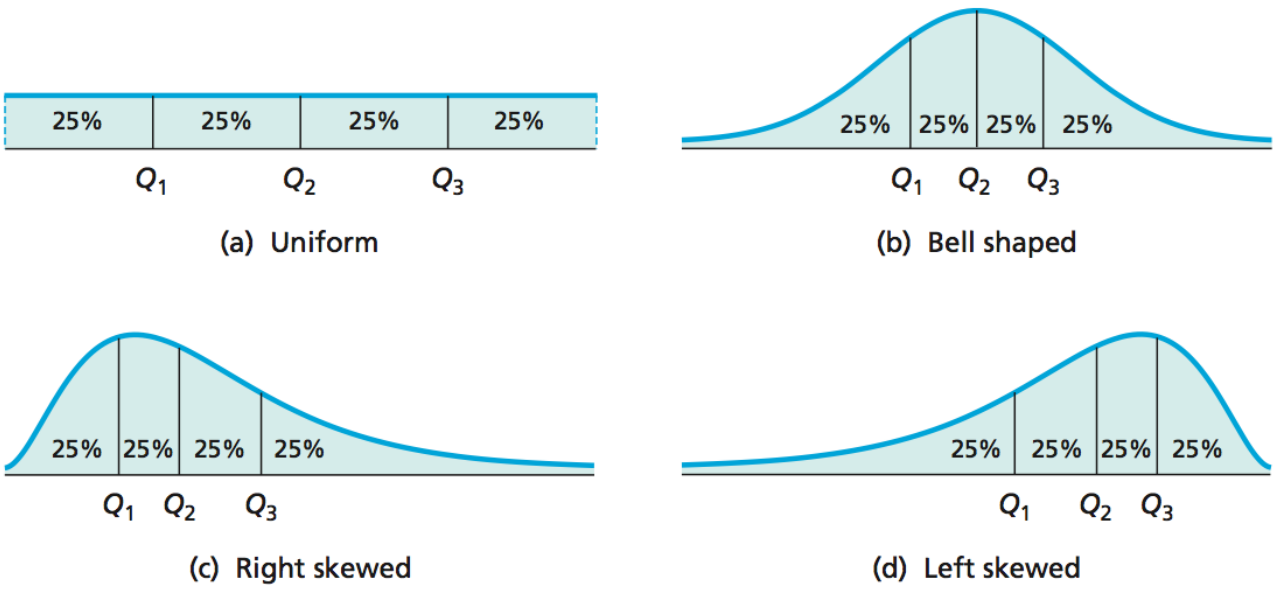
The mean and one, two, and three standard deviations to either side of the mean for the PCB-concentration data



Section 3.4
The Five-Number Summary; Boxplots

Figure 3.12

Quartiles for (a) uniform, (b) bell-shaped, (c) right-skewed, and (d) left-skewed distributions



Definition 3.7

Quartiles

First, arrange the data in increasing order. Next, determine the median. Then, divide the (ordered) data set into two halves, a bottom half and a top half; if the number of observations is odd, include the median in both halves.

- The **first quartile** (Q_1) is the median of the bottom half of the data set.
- The **second quartile** (Q_2) is the median of the entire data set.
- The **third quartile** (Q_3) is the median of the top half of the data set.

Procedure 3.1

To Determine the Quartiles

Step 1 Arrange the data in increasing order.

Step 2 Find the median of the entire data set. This value is the second quartile, Q_2.

Step 3 Divide the ordered data set into two halves, a bottom half and a top half; if the number of observations is odd, include the median in both halves.

Step 4 Find the median of the bottom half of the data set. This value is the first quartile, Q_1.

Step 5 Find the median of the top half of the data set. This value is the third quartile, Q_3.

Step 6 Summarize the results.

Definition 3.8

Interquartile Range

The **interquartile range,** or **IQR,** is the difference between the first and third quartiles; that is, IQR = $Q_3 - Q_1$.

Definition 3.9

Five-Number Summary

The **five-number summary** of a data set is

Min, Q_1, Q_2, Q_3, Max.

Definition 3.10

Lower and Upper Limits

The **lower limit** and **upper limit** of a data set are

$$\text{Lower limit} = Q_1 - 1.5 \cdot \text{IQR};$$
$$\text{Upper limit} = Q_3 + 1.5 \cdot \text{IQR}.$$

Procedure 3.2

To Construct a Boxplot

Step 1 **Determine the quartiles.**

Step 2 **Determine potential outliers and the adjacent values.**

Step 3 **Draw a horizontal axis on which the numbers obtained in Steps 1 and 2 can be located. Above this axis, mark the quartiles and the adjacent values with vertical lines.**

Step 4 **Connect the quartiles to make a box, and then connect the box to the adjacent values with lines.**

Step 5 **Plot each potential outlier with an asterisk.**

Figure 3.14

Constructing a boxplot for TV viewing times in Table 3.13

25	41	27	32	43
66	35	31	15	5
34	26	32	38	16
30	38	30	20	21

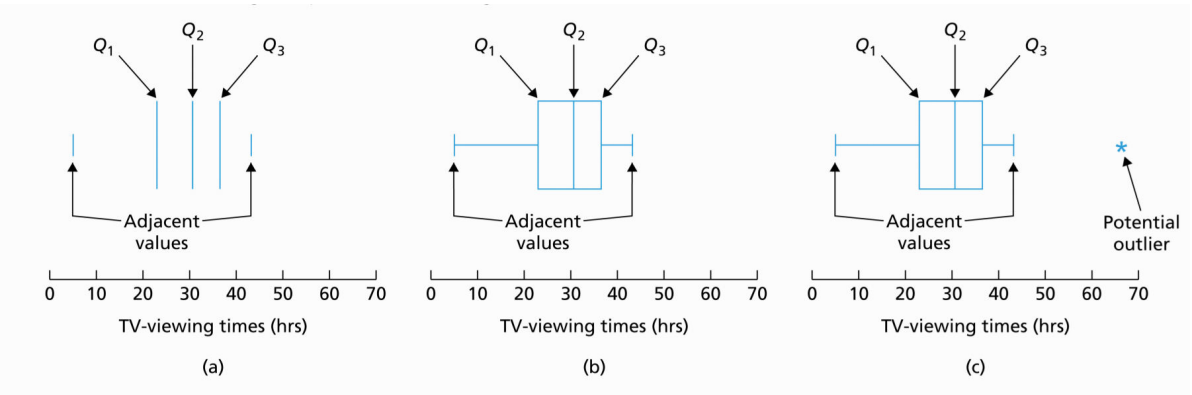


Figure 3.15

Boxplots for the data in Table 3.15

Runners			Others			
7.3	6.7	8.7	24.0	19.9	7.5	18.4
3.0	5.1	8.8	28.0	29.4	20.3	19.0
7.8	3.8	6.2	9.3	18.1	22.8	24.2
5.4	6.4	6.3	9.6	19.4	16.3	16.3
3.7	7.5	4.6	12.4	5.2	12.2	15.6

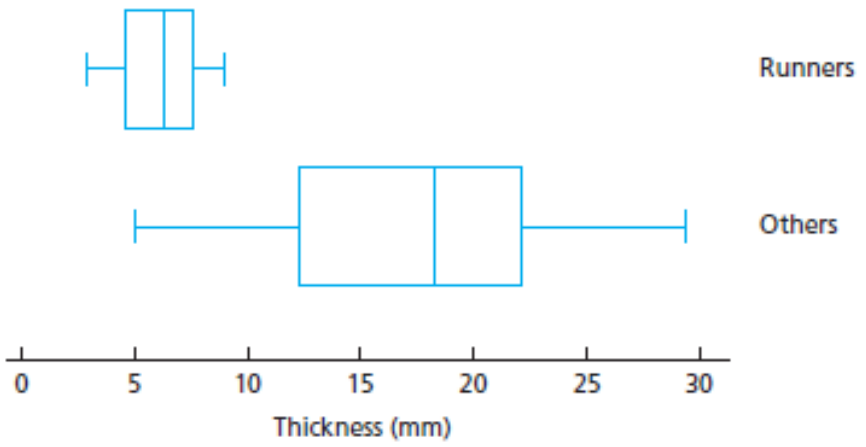
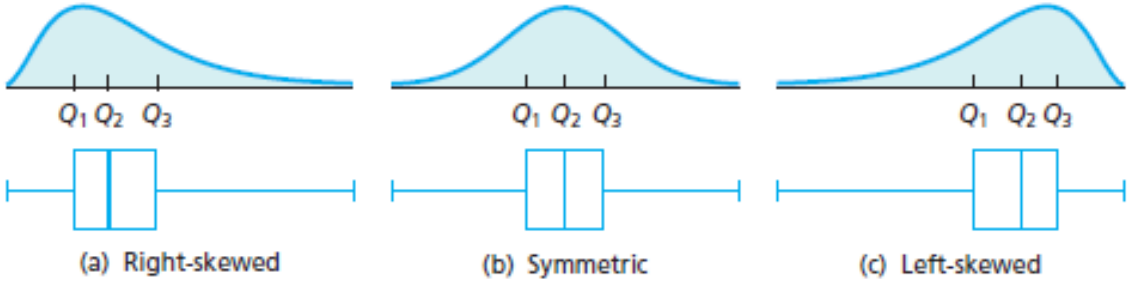


Figure 3.16

Boxplots for (a) right-skewed, (b) symmetric, and (d) left-skewed distributions



Section 3.5

Descriptive Measures for Populations; Use of Samples

Definition 3.11

Population Mean (Mean of a Variable)

For a variable x, the mean of all possible observations for the entire population is called the **population mean** or **mean of the variable x.** It is denoted μ_x or, when no confusion will arise, simply μ. For a finite population,

$$\mu = \frac{\Sigma x_i}{N},$$

where N is the population size.

Definition 3.12

Population Standard Deviation (Standard Deviation of a Variable)

For a variable x, the standard deviation of all possible observations for the entire population is called the **population standard deviation** or **standard deviation of the variable x.** It is denoted σ_x or, when no confusion will arise, simply σ. For a finite population, the defining formula is

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}},$$

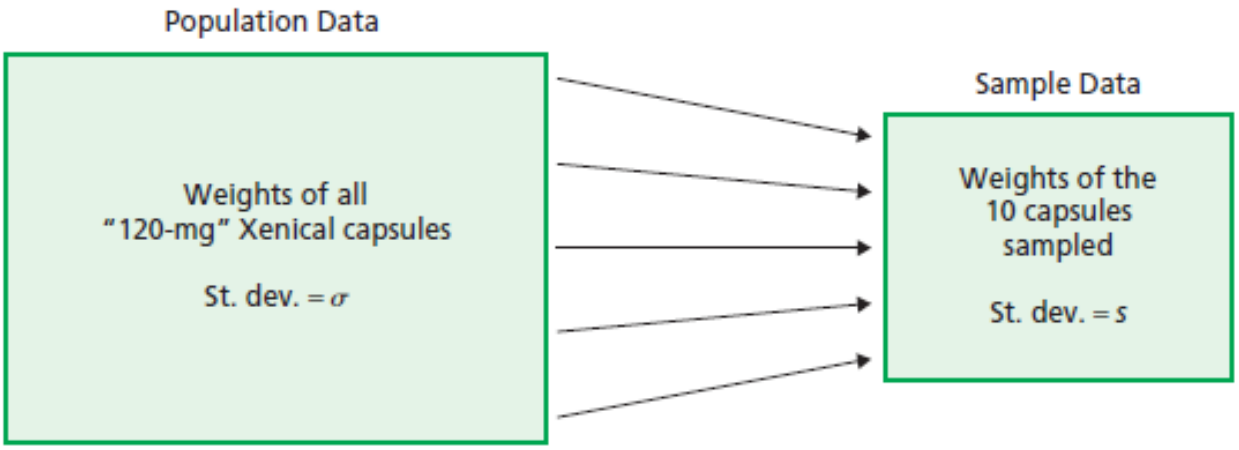
where N is the population size.

The population standard deviation can also be found from the computing formula

$$\sigma = \sqrt{\frac{\Sigma x_i^2}{N} - \mu^2}.$$

Figure 3.18

Population and sample for bolt diameters



Definition 3.13

Parameter and Statistic

Parameter: A descriptive measure for a population.

Statistic: A descriptive measure for a sample.

Definition 3.14 & 3.15

Standardized Variable

For a variable x, the variable

$$z = \frac{x - \mu}{\sigma}$$

is called the **standardized version** of x or the **standardized variable** corresponding to the variable x.

z-Score

For an observed value of a variable x, the corresponding value of the standardized variable z is called the ***z-score*** of the observation. The term **standard score** is often used instead of *z-score.*

