

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 3.4

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,

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

where n is the sample size.

Population Mean or Sample Mean

Population mean: $\mu = \frac{\sum X_i}{N} = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N}$

Sample mean: $\bar{X} = \frac{\sum X_i}{n} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$

where μ is the population mean.

- N and n is the total number of observations in the population and sample respectively.
- X is a particular value.
- \sum indicates the operation of adding.

Characteristics of the Mean

The **arithmetic mean** is the most widely used measure of location. Commonly called ‘the mean’.

- It is calculated by summing all values in the data set and dividing the sum by the number of values in the data set (the average of a group of numbers).
- Applicable for interval and ratio data.
- Not applicable for nominal or ordinal data.
- Affected by each value in the data set, including extreme values.

EXAMPLE

A **statistic** is a measurable characteristic of a sample.

Example 2: A sample of five executives received the following bonus last year (\$000):

14.0, 15.0, 17.0, 16.0, 15.0

$$\bar{X} = \frac{\Sigma X}{n} = \frac{14.0 + \dots + 15.0}{5} = \frac{77}{5} = 15.4$$

EXAMPLE Children of Diabetic Mothers

The paper “Correlations Between the Intrauterine Metabolic Environment and Blood Pressure in Adolescent Offspring of Diabetic Mothers” (*Journal of Pediatrics*, Vol. 136, Issue 5, pp. 587–592) by N. Cho et al. presented findings of research on children of diabetic mothers. Past studies showed that maternal diabetes results in obesity, blood pressure, and glucose tolerance complications in the offspring.

Table 3.5 presents the arterial blood pressures, in millimeters of mercury (mm Hg), for a sample of 16 children of diabetic mothers. Determine the sample mean of these arterial blood pressures.

EXAMPLE Children of Diabetic Mothers

TABLE 3.5

Arterial blood pressures
of 16 children of diabetic mothers

81.6	84.1	87.6	82.8
82.0	88.9	86.7	96.4
84.6	104.9	90.8	94.0
69.4	78.9	75.2	91.0

EXAMPLE Children of Diabetic Mothers

Solution Let x denote the variable “arterial blood pressure.” We want to find the mean, \bar{X} , of the 16 observations of x shown in Table 3.5. The sum of those observations is $\bar{X} = \sum_{k=1}^n x_k / n$. The sample size (or number of observations) is 16, so $n = 16$. Thus,

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1378.9}{16} = 86.18.$$

Hence the mean arterial blood pressure of the sample of 16 children of diabetic mothers is 86.18 mm Hg.

Summary of the Arithmetic Mean:

- Every set of interval-level and ratio-level data has a mean.
- All the values are included in computing the mean.
- A set of data has a unique mean.
- The mean is affected by unusually large or small data values.
- The arithmetic mean is the only measure of central tendency where the sum of the deviations of each value from the mean is zero.

An Example to illustrate the last point

Consider the set of values: 3, 8, and 4. The **mean** is 5.

Illustrating the fifth property:

$$\Sigma(X - \bar{X}) = [(3 - 5) + (8 - 5) + (4 - 5)] = 0$$

Median

Middle value in an ordered array of numbers.

Applicable for ordinal, interval, and ratio data

Not applicable for nominal data

Unaffected by extremely large and extremely small values.

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.

Median: Example with an Odd Number of Terms

Ordered Array

3, 4, 5, 7, 8, 9, 11, 14, 15, 16, 16, 17, 19, 19, 20, 21, 22

There are 17 terms in the ordered array.

Position of median = $(n+1)/2 = (17+1)/2 = 9$

The median is the 9th term, 15.

If the 22 is replaced by 100, the median is 15.

If the 3 is replaced by -103, the median is 15.

Median: Example with an Even Number of Terms

Ordered Array

3, 4, 5, 7, 8, 9, 11, 14, 15, 16, 16, 17, 19, 19, 20, 21

- There are 16 terms in the ordered array.
- Position of median = $(n+1)/2 = (16+1)/2 = 8.5$
- The median is between the 8th and 9th terms, 14.5.
- If the 21 is replaced by 100, the median is 14.5.
- If the 3 is replaced by -88, the median is 14.5.

Median is a Resistant Measure

A **resistant measure** *is not sensitive to the influence of a few extreme observations.*

The **median is a resistant measure** of center, but the mean is not. A *trimmed mean* can improve the resistance of the mean: removing a percentage of the smallest and largest observations before computing the mean gives a **trimmed mean**.

Mode

Applicable to all levels of data measurement (nominal, ordinal, interval, and ratio)

A data set Might have many values that have the same frequency, in such a case, we might have:

- Bimodal -- Data sets that have two modes
- Multimodal -- Data sets that contain more than two modes

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.

Mode -- Example

The mode is 44.

There are more 44s
than any other value.

35	41	44	45
37	41	44	46
37	43	44	46
39	43	44	46
40	43	44	46
40	43	45	48

Example 3.1 Weekly Salaries

Professor Hassett spent one summer working for a small mathematical consulting firm. The firm employed a few senior consultants, who made between \$800 and \$1050 per week; a few junior consultants, who made between \$400 and \$450 per week; and several clerical workers, who made \$300 per week.

The firm required more employees during the first half of the summer than the second half. Tables 3.1 and 3.2 list typical weekly earnings for the two halves of the summer.

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

Solution for example:

Solution: To find the median of Data Set I, we first arrange the data in increasing order:

300 300 300 300 300 300 **400** 400 450 450 800 940 1050

The number of observations is 13, so $(n + 1)/2 = (13 + 1)/2 = 7$. Consequently, the median is the seventh observation in the ordered list, which is 400 (shown in boldface).

Solution for example:

To find the median of Data Set II, again arrange the data in increasing order:

300 300 300 300 **300** **400** 400 450 940 1050

The number of observations is 10, so $(n + 1)/2 = (10 + 1)/2 = 5.5$. Consequently, the median is halfway between the fifth and sixth observations (shown in boldface) in the ordered list, which is 350.

Solution for example:

TABLE 3.3

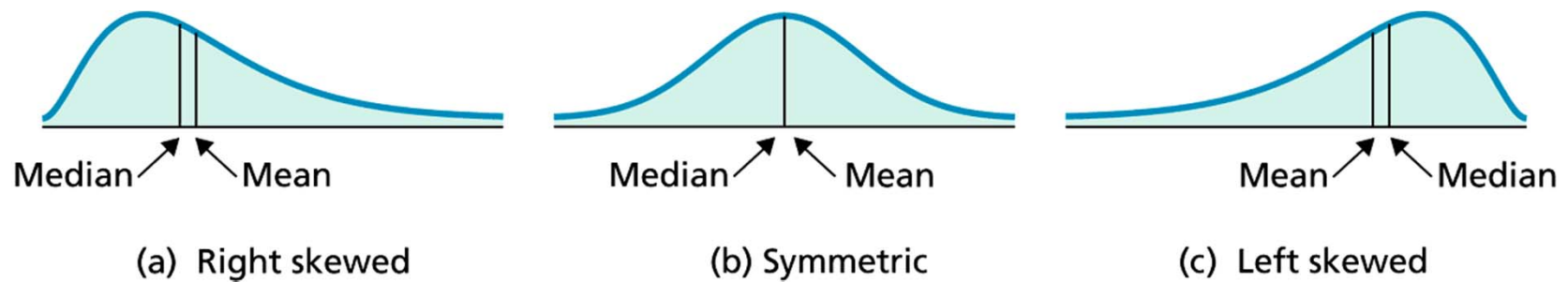
Frequency distribution for Data Set I

Salary	Frequency
300	6
400	2
450	2
800	1
940	1
1050	1

To find the mode of Data Set I, we count the frequency of each number and find that here mode of the set is 300.

Figure 3.1

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



EXAMPLE 3.4 Selecting an Appropriate Measure of Center

A student takes four exams in a biology class. His grades are 88, 75, 95, and 100. Which measure of center is the student likely to report?

Ans.: Chances are that the student would report the mean of his scores, which is 89.5. The mean is probably the most suitable measure of center for the student to use because it takes into account the numerical value of each score and therefore indicates his overall performance.

EXAMPLE 3.4 Selecting an Appropriate Measure of Center

An Governmental agency publishes data on resale prices of U.S. homes. Which measure of center is most appropriate for such resale prices?

Ans.: The most appropriate measure of center for resale home prices is the median because it is aimed at finding the center of the data on resale home prices and because it is not strongly affected by the relatively few homes with extremely high resale prices. Thus the median provides a better indication of the “typical” resale price than either the mean or the mode.

EXAMPLE 3.4 Selecting an Appropriate Measure of Center

The 2009 Boston Marathon had two categories of official finishers: male and female, of which there were 13,547 and 9,302, respectively. Which measure of center should be used here?

Ans.: The only suitable measure of center for these data is the mode, which is “male”. Each observation in this data set is either “male” or “female.” There is no way to compute a mean or median for such data. Of the mean, median, and mode, the mode is the only measure of center that can be used for qualitative data.

Section 3.2

Measures of Variation



Measures of Dispersion: variation or spread

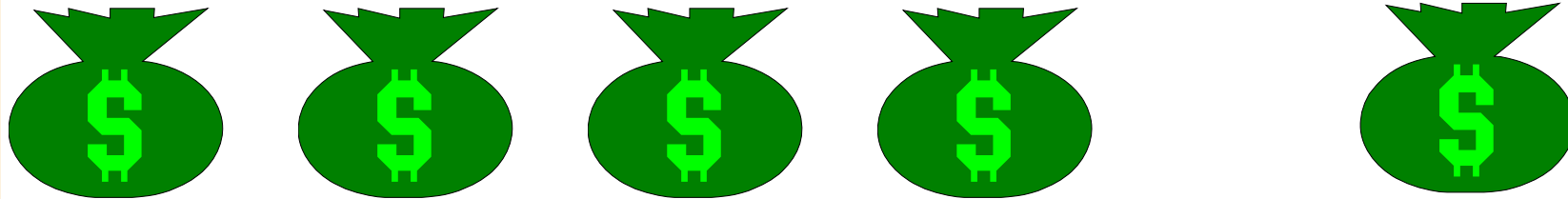
Measures of central tendency alone cannot completely characterize a set of data. Two very different data sets may have similar measures of central tendency.

Measures of dispersion are used to describe the spread, or variability, of a distribution

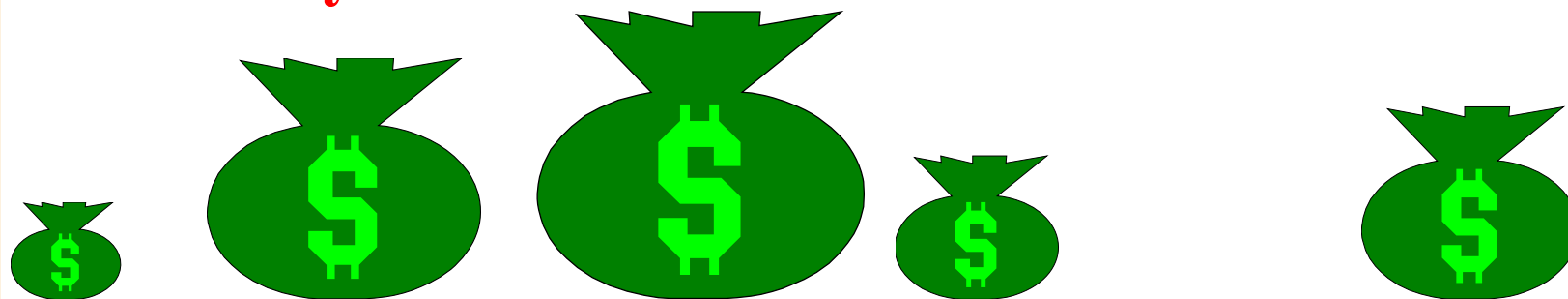
Common measures of dispersion: range, variance, and standard deviation

Example about Variability

No Variability in Cash Flow

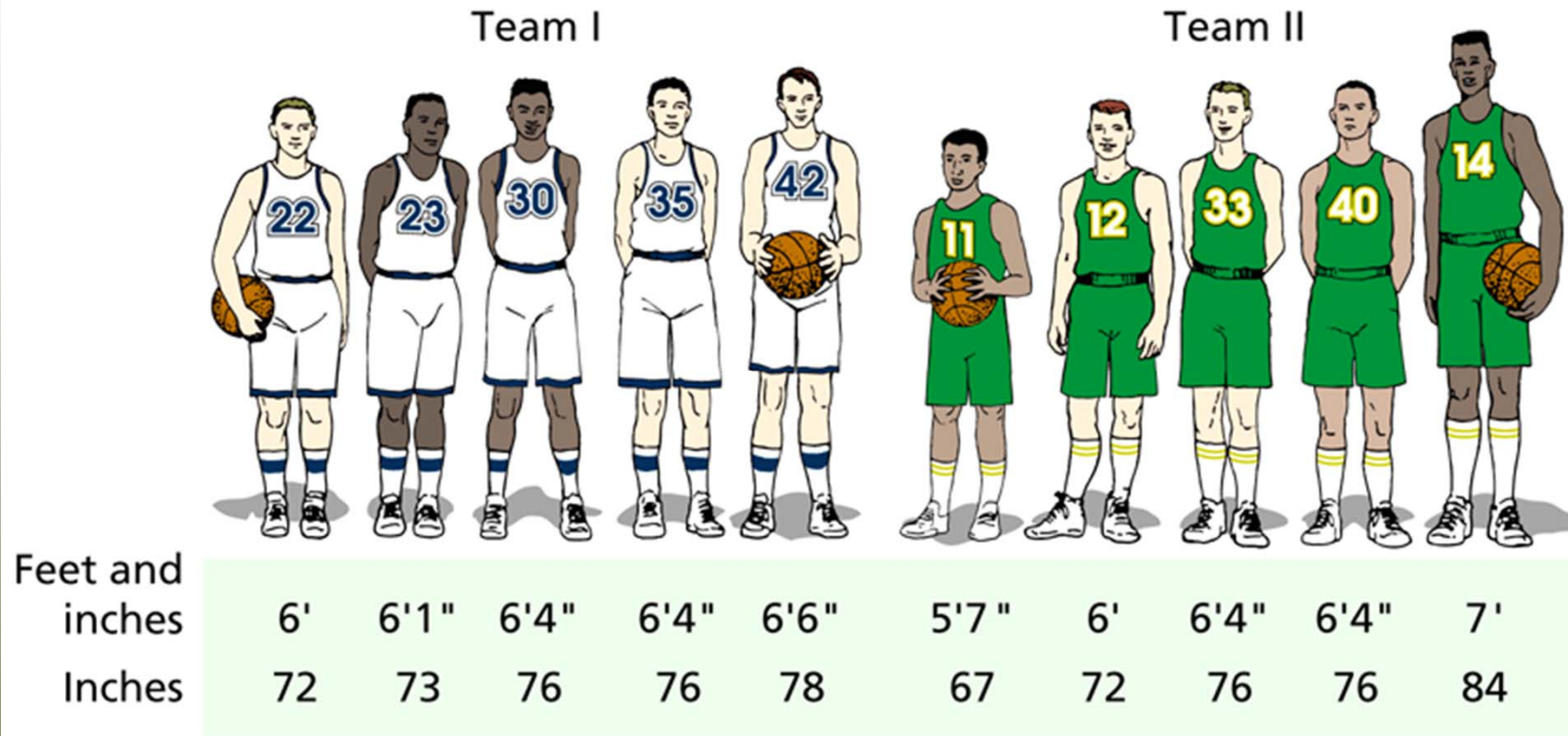


Variability in Cash Flow



Example about Variability

Five starting players on two basketball teams



Team 1: $72 + 73 + 76 + 76 + 78 = 375$, mean = $375/5 = 75$

Team 2: $67 + 72 + 76 + 76 + 84 = 375$, mean = $375/5 = 75$

Implies no difference?

Definition 3.5

Range of a Data Set

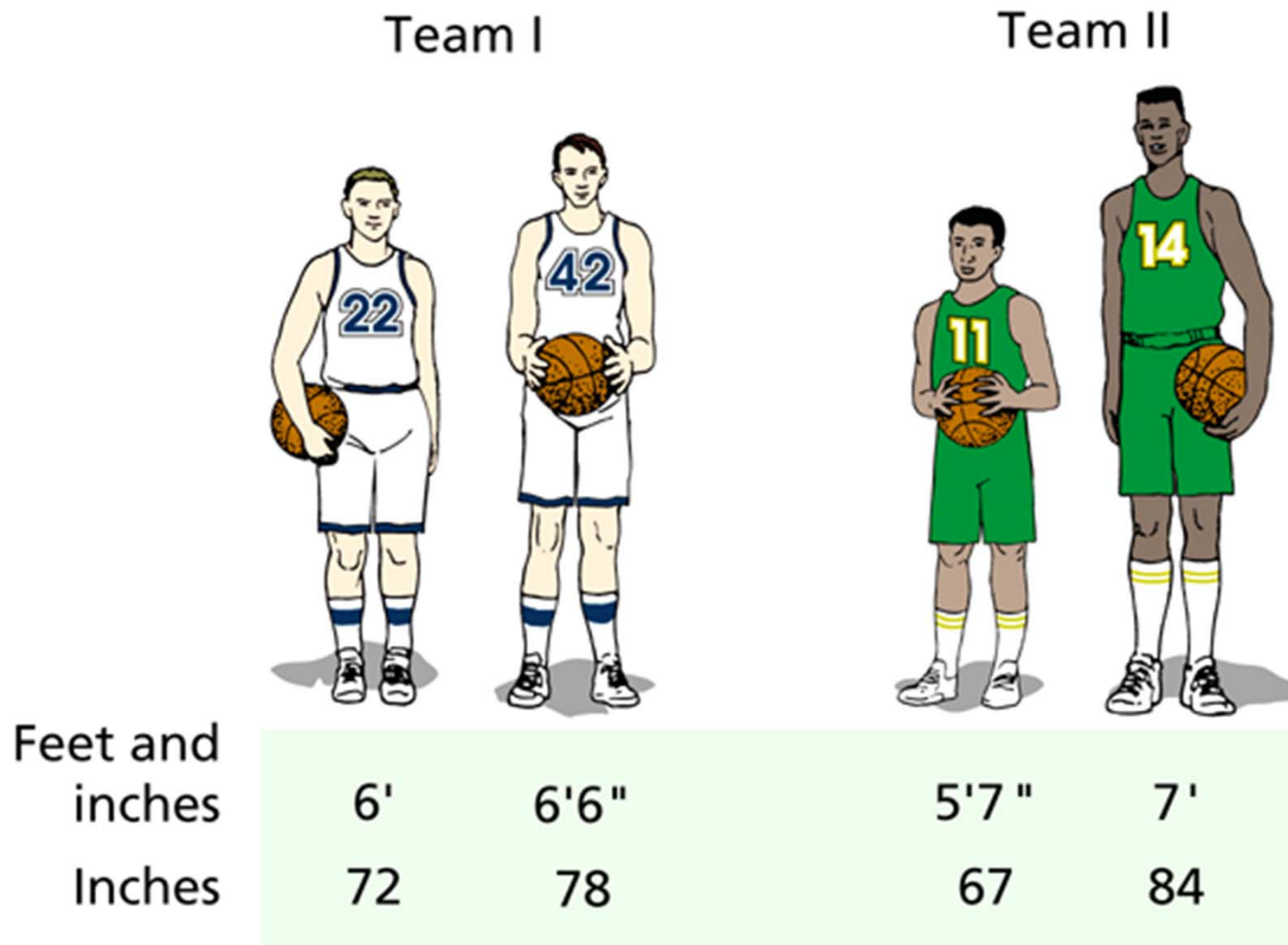
The **range** of a data set is given by the formula

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

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

Figure 3.3

Shortest and tallest starting players on the teams



Range

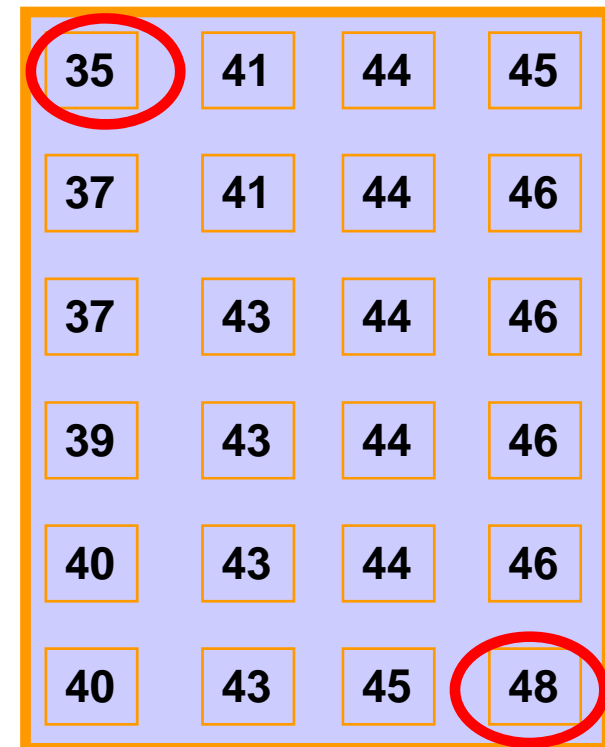
The difference between the largest and the smallest values in a set of data

Simple to compute.

Ignores all data points except the two extremes.

Example:

$$\begin{aligned}\text{Range} &= \text{Largest} - \text{Smallest} \\ &= 48 - 35 = 13\end{aligned}$$



A 6x4 grid of numbers. The smallest value, 35, is circled in red in the top-left cell. The largest value, 48, is circled in red in the bottom-right cell. All other cells are outlined in orange.

35	41	44	45
37	41	44	46
37	43	44	46
39	43	44	46
40	43	44	46
40	43	45	48

More Example

The weights of a sample of crates containing books for the bookstore (in pounds) are:

103, 97, 101, 106, 103

Find the range:

$$\text{Range} = 106 - 97 = 9$$

The Sample Standard Deviation

In contrast to the range, the standard deviation takes into account all the observations.

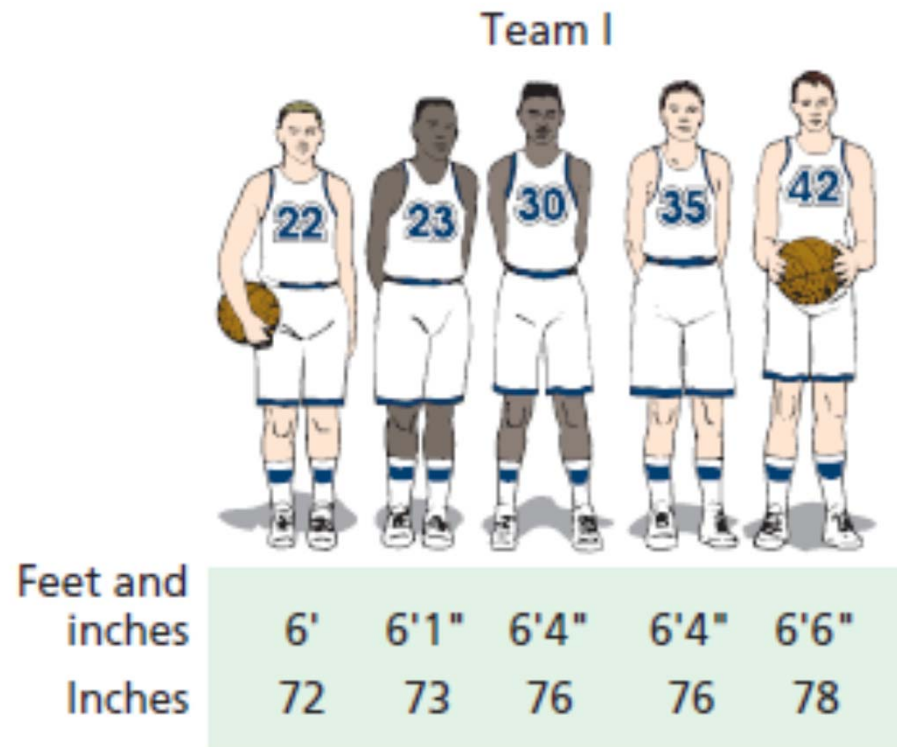
It is the preferred measure of variation when the mean is used as the measure of center.

Roughly speaking, the standard deviation measures variation by indicating how far, on average, the observations are from the mean.

The first step in computing a sample standard deviation is to find the deviations from the mean, that is, how far each observation is from the mean.

EXAMPLE 3.8 The Deviations From the Mean

Heights of Starting Players The heights, in inches, of the five starting players on Team I are 72, 73, 76, 76, and 78, as we saw in Fig. 3.2. Find the deviations from the mean.



$$\bar{x} = 75$$

EXAMPLE 3.8 The Deviations From the Mean

To find the deviation from the mean for an observation x_i , we subtract the mean from it; that is, we compute $x_i - \bar{x}$. For instance, the deviation from the mean for the height of 72 inches is $x_1 - \bar{x} = 72 - 75 = -3$. The deviations from the mean for all five observations are given in the second column of Table 3.6 and are represented by arrows in Fig. 3.4

EXAMPLE 3.8 The Deviations From the Mean

by arrows in Fig. 3.4.

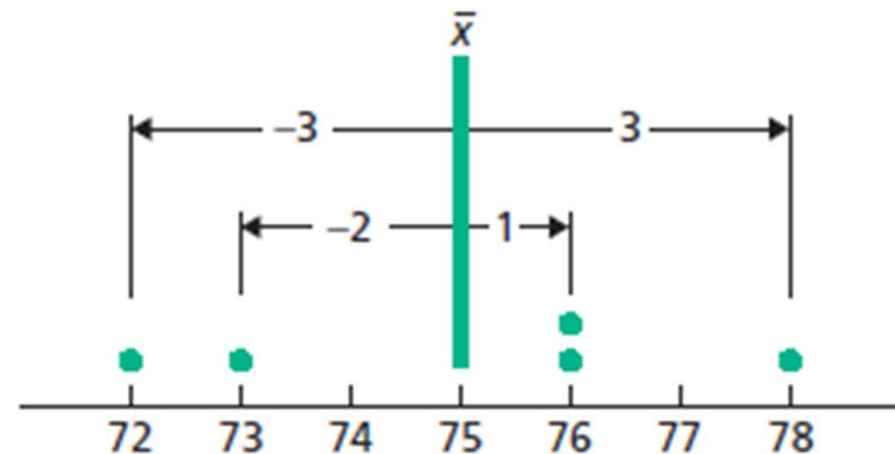
TABLE 3.6
Deviations from the mean

Height x	Deviation from mean $x - \bar{x}$
72	-3
73	-2
76	1
76	1
78	3

Total sum
= 0!!!

FIGURE 3.4

Observations (shown by dots) and deviations from the mean (shown by arrows)



The Sample Standard Deviation

In contrast to the range, the standard deviation takes into account all the observations.

It is the preferred measure of variation when the mean is used as the measure of center.

Roughly speaking, the standard deviation measures variation by indicating how far, on average, the observations are from the mean. The more variation that there is in a data set, the larger is its standard deviation.

The first step in computing a sample standard deviation is to find the deviations from the mean, that is, how far each observation is from the mean.

Definition 3.6

Sample Standard Deviation

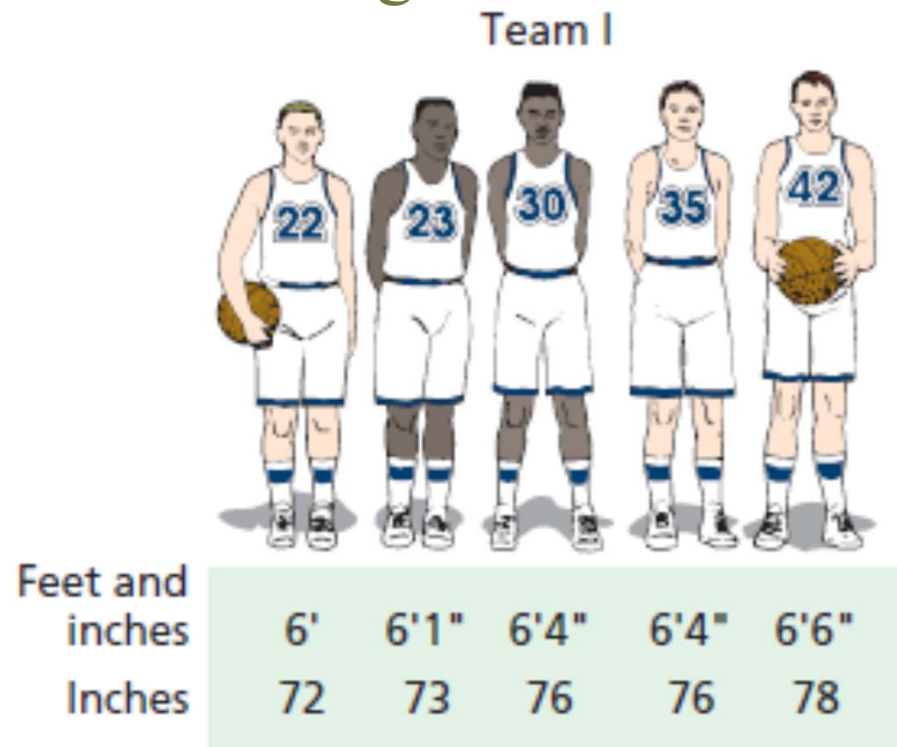
For a variable x , the standard deviation of the observations for a sample is called a **sample standard deviation**. It is denoted s_x or, when no confusion will arise, simply s . We have

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

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

EXAMPLE 3.9 The Sum of Squared Deviations

Heights of Starting Players The heights, in inches, of the five starting players on Team I are 72, 73, 76, 76, and 78, as we saw in Fig. 3.2. Compute the sum of squared deviations for the heights of the starting players on Team I.



$$\bar{x} = 75$$

EXAMPLE 3.9 The Sum of Squared Deviations

TABLE 3.7

Table for computing the sum of squared deviations for the heights of Team I

Height x	Deviation from mean $x - \bar{x}$	Squared deviation $(x - \bar{x})^2$
72	-3	9
73	-2	4
76	1	1
76	1	1
78	3	9
		24

sum of
squared
deviations

EXAMPLE 3.10 The **Sample Variance**

Heights of Starting Players Determine the sample variance of the heights of the starting players on Team I.

Solution From Example 3.9, the sum of squared deviations is 24 inches². Because $n = 5$,

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{24}{5 - 1} = 6.$$

The sample variance is 6 inches².

EXAMPLE 3.11 The **Sample Standard Deviation**

Heights of Starting Players Determine the sample standard deviation of the heights of the starting players on Team I.

Solution From Example 3.10, the sample variance is 6 inches². Thus the sample standard deviation is

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}} = \sqrt{6} = 2.4 \text{ inches (rounded).}$$

Interpretation Roughly speaking, on average, the heights of the players on Team I vary from the mean height of 75 inches by about 2.4 inches.

A Proof for the **Computing Formula** for s

Formula 3.1

Computing Formula for a Sample Standard Deviation

A sample standard deviation can be computed using the formula

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

where n is the sample size.

Rounding Basics for calculating *s*

Rounding Rule:

- 1) Do not perform any rounding until the computation is complete; otherwise, substantial round-off error can result.
- 2) To round final answers that contain units to one more decimal place than the raw data. Although we usually abide by this convention, occasionally we vary from it for instructional reasons.

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
$\bar{x} = 50.0$ $s = 7.4$	$\bar{x} = 50.0$ $s = 14.2$

Figure 3.5 and Figure 3.6

FIGURE 3.5 Data Set I; $\bar{x} = 50$, $s = 7.4$

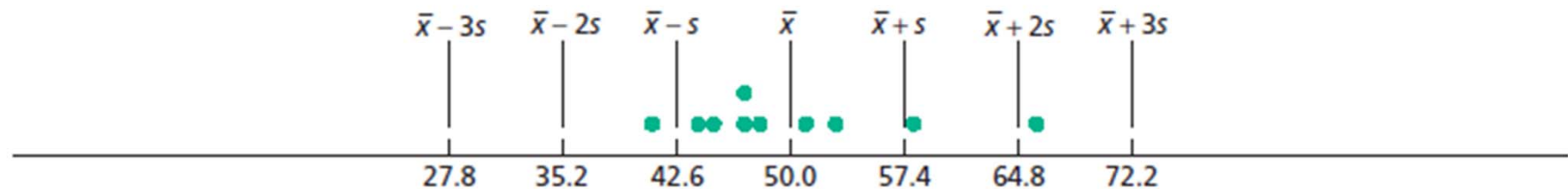
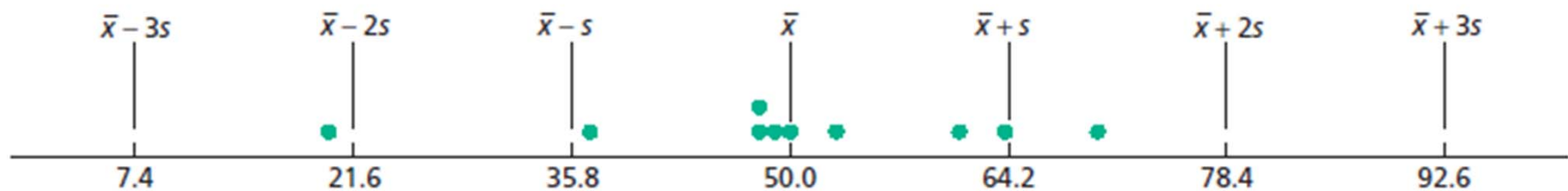


FIGURE 3.6 Data Set II; $\bar{x} = 50$, $s = 14.2$



Data Set II has more variation than Data Set I.

Section 3.3

The Five-Number Summary; Boxplots



Percentiles

Measures of central tendency that divide a group of data into 100 parts.

At least $n\%$ of the data lie below the n^{th} percentile, and at most $(100 - n)\%$ of the data lie above the n^{th} percentile

Example: 90th percentile indicates that at least 90% of the data lie below it, and at most 10% of the data lie above it.

The median and the 50^{th} percentile have the same value.

Applicable for ordinal, interval, and ratio data.

Not applicable for nominal data.

Percentiles: Computational Procedure

Sort the data into an *ascending* ordered array (from small to large).

Calculate the percentile location:

$$i = \frac{P}{100}(n)$$

Determine the percentile's location and its value.

If i is a whole number, the percentile is the average of the values at the i and $(i+1)$ positions.

If i is not a whole number, the percentile is at the $(i+1)$ position in the ordered array.

Percentiles: Example

Raw Data: 14, 12, 19, 23, 5, 13, 28, 17

Ordered Array: 5, 12, 13, 14, 17, 19, 23, 28

Location of 30th percentile:

$$i = \frac{30}{100}(8) = 2.4$$

The location index, i , is not a whole number; $i+1 = 2.4+1=3.4$; the whole number portion is 3; the 30th percentile is at the 3rd location of the array; the 30th percentile is 13.

Quartiles

Quartiles are the most commonly used percentiles. A data set has three quartiles, which we denote Q_1 , Q_2 , and Q_3 . Roughly speaking, the **first quartile**, Q_1 , is the number that divides the bottom 25% of the data from the top 75%; the **second quartile**, Q_2 , is the median, which, as you know, is the number that divides the bottom 50% of the data from the top 50%; and the **third quartile**, Q_3 , is the number that divides the bottom 75% of the data from the top 25%. Note that the first and third quartiles are the 25th and 75th percentiles, respectively.

Quartiles

Measures of central tendency that divide a group of data into four subgroups

Q_1 : 25% of the data set is below the first quartile

Q_2 : 50% of the data set is below the second quartile

Q_3 : 75% of the data set is below the third quartile

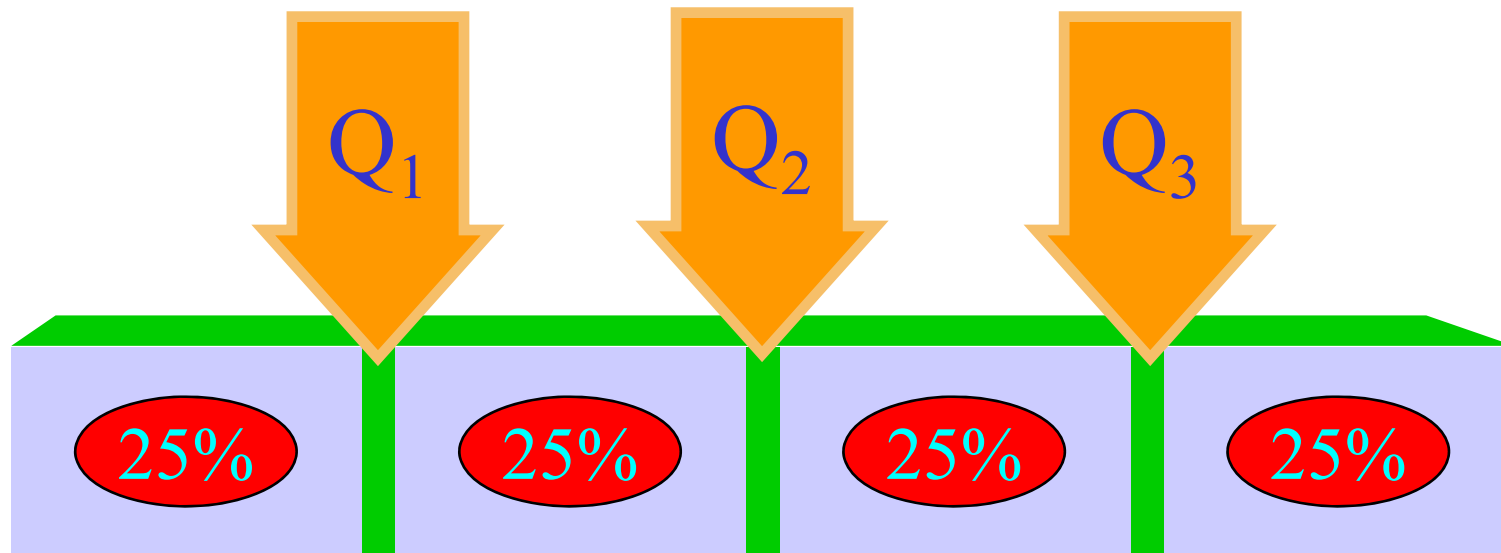
Q_1 is equal to the 25th percentile

Q_2 is located at 50th percentile and equals the median

Q_3 is equal to the 75th percentile

Quartile values are not necessarily members of the data set

Quartiles



Quartiles: Example

Ordered array: 106, 109, 114, 116, 121, 122, 125, 129

$$Q_1 \quad i = \frac{25}{100}(8) = 2 \quad Q_1 = \frac{109 + 114}{2} = 111.5$$

$$Q_2: \quad i = \frac{50}{100}(8) = 4 \quad Q_2 = \frac{116 + 121}{2} = 118.5$$

$$Q_3: \quad i = \frac{75}{100}(8) = 6 \quad Q_3 = \frac{122 + 125}{2} = 123.5$$

Definition 3.7

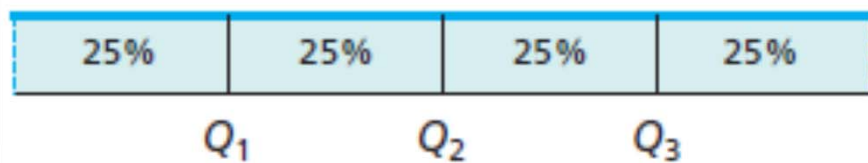
Quartiles

Arrange the data in increasing order and determine the median.

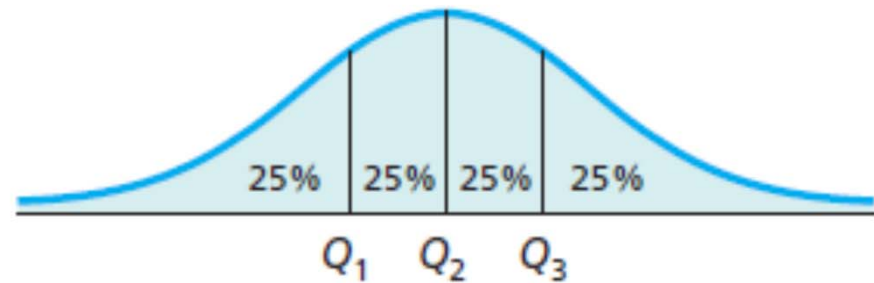
- The **first quartile** is the median of the part of the entire data set that lies at or below the median of the entire data set.
- The **second quartile** is the median of the entire data set.
- The **third quartile** is the median of the part of the entire data set that lies at or above the median of the entire data set.

FIGURE 3.7

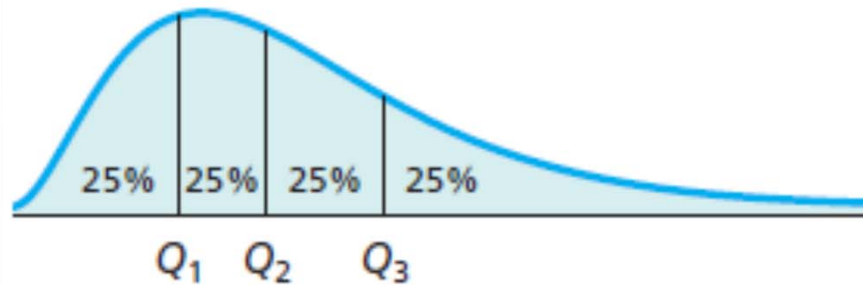
Figure 3.7 depicts the quartiles for uniform, bell-shaped, right-skewed, and left-skewed distributions.



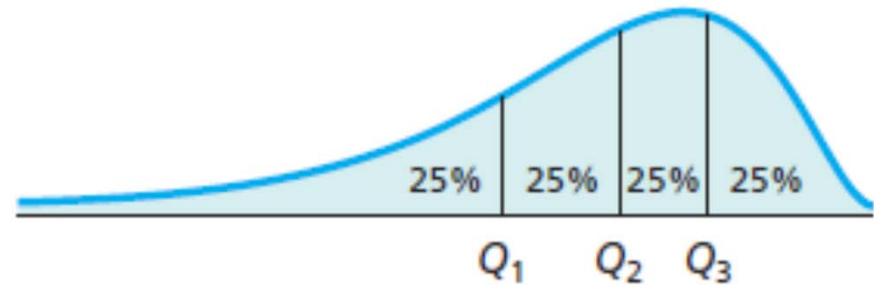
(a) Uniform



(b) Bell shaped



(c) Right skewed



(d) Left skewed

The Interquartile Range

Because quartiles are used to define the interquartile range, it is the preferred measure of variation when the median is used as the measure of center. Like the median, the interquartile range is a resistant measure.

Definition 3.8

Interquartile Range

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

EXAMPLE 3.15 Weekly TV-Viewing Times

The A. C. Nielsen Company publishes information on the TV-viewing habits of Americans in *Nielsen Report on Television*. A sample of 20 people yielded the weekly viewing times, in hours, displayed in Table 3.12.

1. Determine and interpret the quartiles for these data.
2. Find the IQR for the TV-viewing-time data given in Table 3.12.

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

EXAMPLE 3.15 Part 1

Solution First, we arrange the data in Table 3.12 in increasing order:

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

Next, we determine the median of the entire data set. The number of observations is 20, so the median is at position $(20 + 1)/2 = 10.5$, halfway between the tenth and eleventh observations (shown in boldface) in the ordered list. Thus, the median of the entire data set is $(30 + 31)/2 = 30.5$.

EXAMPLE 3.15 Part 1

Because the median of the entire data set is 30.5, the part of the entire data set that **lies at or below** the median of the entire data set is

5 15 16 20 **21 25** 26 27 30 30

This data set has 10 observations, so its median is at position $(10 + 1)/2 = 5.5$, halfway between the fifth and sixth observations (shown in boldface) in the ordered list. Thus the median of this data set and hence the first quartile is $(21 + 25)/2 = 23$; that is, **$Q_1 = 23$** .

The second quartile is the median of the entire data set, or 30.5. Therefore, we have **$Q_2 = 30.5$** .

EXAMPLE 3.15 Part 1

Because the median of the entire data set is 30.5, the part of the entire data set that **lies at or above** the median of the entire data set is

31 32 32 34 **35 38** 38 41 43 66

This data set has 10 observations, so its median is at position $(10 + 1)/2 = 5.5$, halfway between the fifth and sixth observations (shown in boldface) in the ordered list. Thus the median of this data set and hence the third quartile is $(35 + 38)/2 = 36.5$; that is, $Q_3 = 36.5$.

In summary, the three quartiles for the TV-viewing times in Table 3.12 are $Q_1 = 23$ hours, $Q_2 = 30.5$ hours, and $Q_3 = 36.5$ hours.

EXAMPLE 3.15 Part 1

In summary, the three quartiles for the TV-viewing times in Table 3.12 are $Q_1 = 23$ hours, $Q_2 = 30.5$ hours, and $Q_3 = 36.5$ hours.

Interpretation We see that 25% of the TV-viewing times are less than 23 hours, 25% are between 23 hours and 30.5 hours, 25% are between 30.5 hours and 36.5 hours, and 25% are greater than 36.5 hours.

EXAMPLE 3.15 Part 2

Solution As we discovered in Example 3.15, the first and third quartiles are 23 and 36.5, respectively. Therefore, $\text{IQR} = Q_3 - Q_1 = 36.5 - 23 = 13.5$ hours.

Interpretation The middle 50% of the TV-viewing times are spread out over a 13.5-hour interval, roughly.

Definition 3.9

Five-Number Summary

The **five-number summary** of a data set is
Min, Q_1 , Q_2 , Q_3 , Max.

EXAMPLE 3.17 Weekly TV-Viewing Times

3. Find and interpret the **five-number summary** for the TV-viewing-time data given in Table 3.12 on page 116.

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

EXAMPLE 3.17 Part 3

Solution We have $\text{Min} = 5$ and $\text{Max} = 66$. Furthermore, as we showed earlier, $Q_1 = 23$, $Q_2 = 30.5$, and $Q_3 = 36.5$. Consequently, the five-number summary of the data on TV-viewing times **is 5, 23, 30.5, 36.5, and 66** hours.

The variations of the four quarters of the TV-viewing-time data are therefore 18, 7.5, 6, and 29.5 hours, respectively.

Interpretation There is less variation in the middle two quarters of the TV-viewing times than in the first and fourth quarters, and the fourth quarter has the greatest variation of all.

Outliers

Outliers: observations that fall well outside the overall pattern of the data.

It may be the result of a measurement or recording error, an observation from a different population, or an unusual extreme observation.

Note that an extreme observation need not be an outlier; it may instead be an indication of skewness.

Definition 3.10: Identify potential outliers

Lower and Upper Limits

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

$$\text{Lower limit} = Q_1 - 1.5 \cdot \text{IQR};$$

$$\text{Upper limit} = Q_3 + 1.5 \cdot \text{IQR}.$$

EXAMPLE 3.18 Weekly TV-Viewing Times

For the TV-viewing-time data in Table 3.12 on page 116:

4. obtain the lower and upper limits.
5. determine potential outliers, if any.

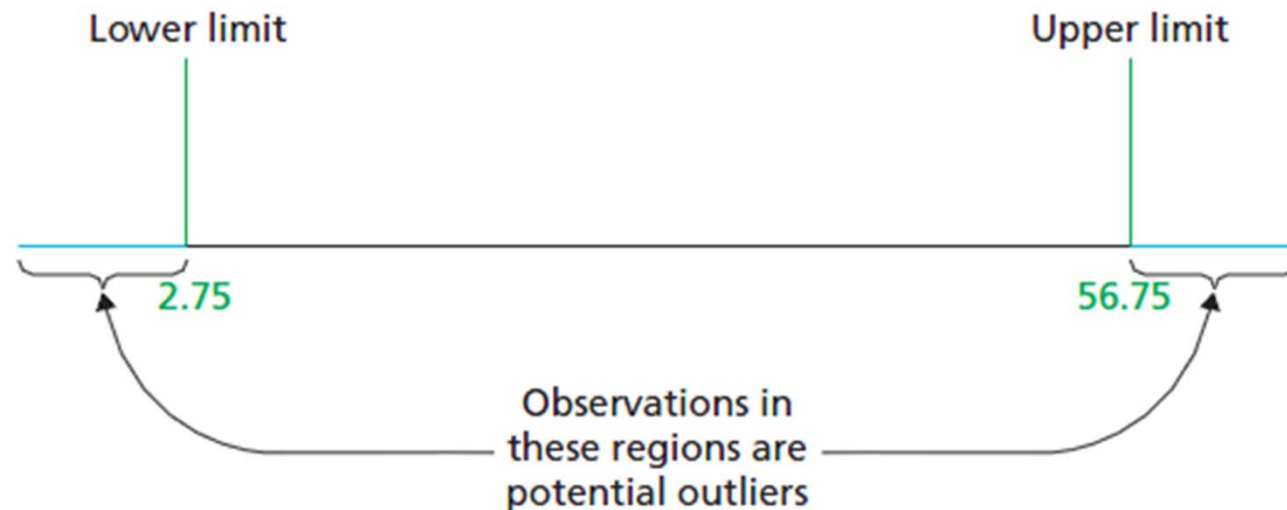
EXAMPLE 3.18 Weekly TV-Viewing Times

Solution: (4) As before, $Q_1 = 23$, $Q_3 = 36.5$, and

IQR = 13.5. Therefore

Lower limit = $Q_1 - 1.5 \times \text{IQR} = 23 - 1.5 \times 13.5 = 2.75$
hours;

Upper limit = $Q_3 + 1.5 \times \text{IQR} = 36.5 + 1.5 \times 13.5 = 56.75$
hours.



EXAMPLE 3.18 Weekly TV-Viewing Times

Solution: (5) The ordered list of the entire data set on page 116 reveals one observation, 66, that lies outside the lower and upper limits; in particular, above the upper limit. Consequently, 66 is a potential outlier.

Procedure 3.1

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.

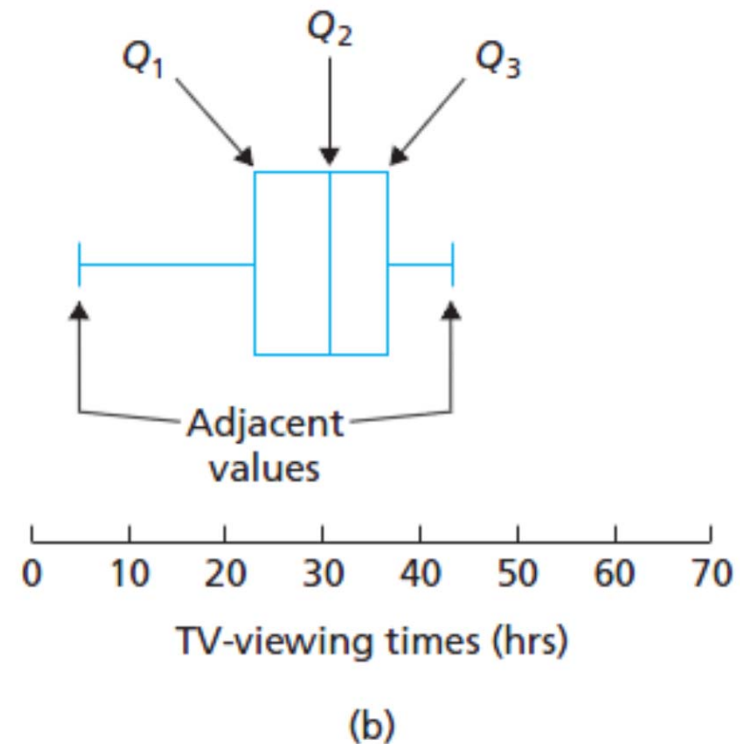
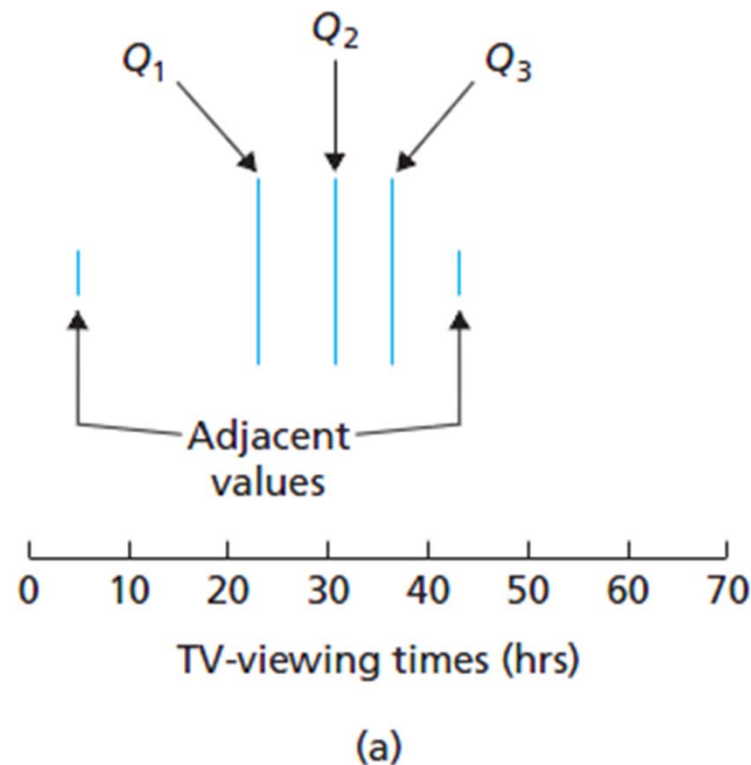
Step 5 Plot each potential outlier with an asterisk.

EXAMPLE 3.19 Weekly TV-Viewing Times

6. For the weekly TV-viewing times' sample of 20 people given in Table 3.12 on page 116. Construct a boxplot for these data. Determine potential outliers, if any.

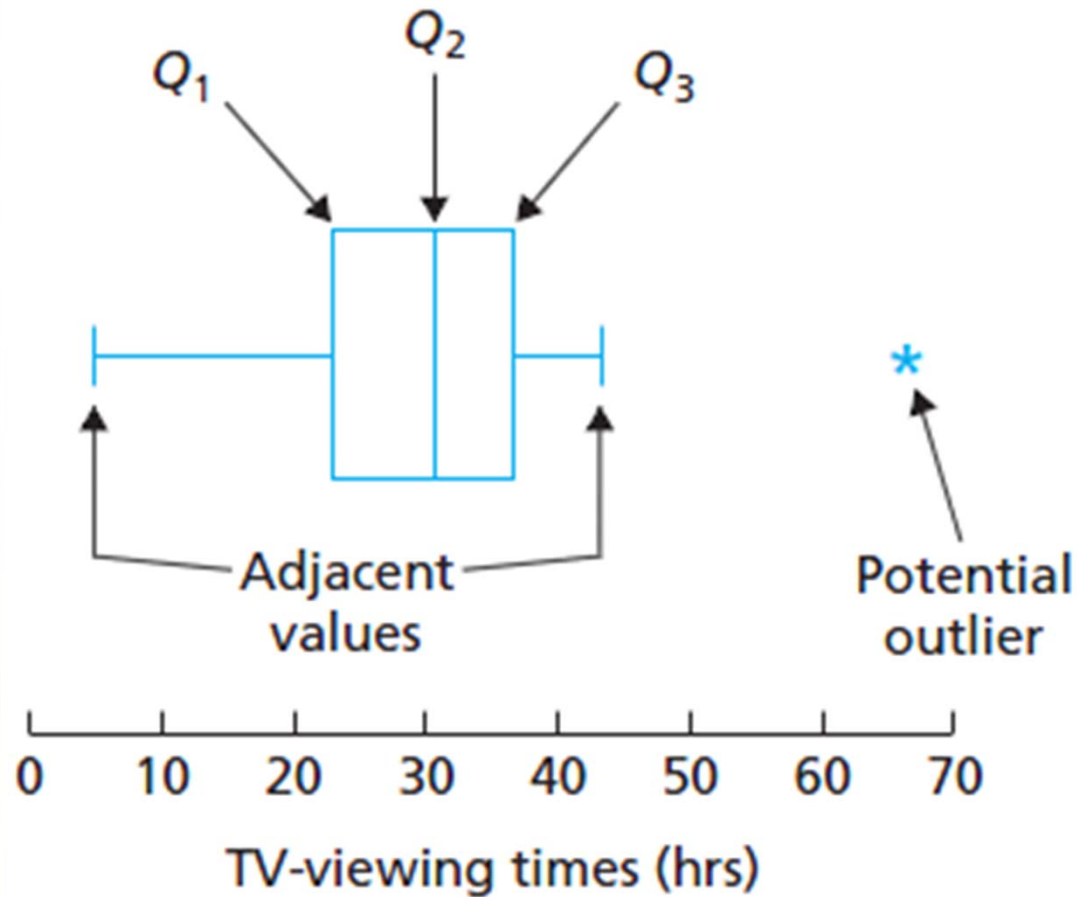
EXAMPLE 3.19 Weekly TV-Viewing Times

Solution: (6) FIGURE 3.9 Constructing a boxplot for the TV-viewing times



EXAMPLE 3.19 Weekly TV-Viewing Times

Solution: (6) FIGURE 3.9 Constructing a boxplot for the TV-viewing times



Interpretation There is less variation in the middle two quarters of the TV-viewing times than in the first and fourth quarters, and the fourth quarter has the greatest variation of all.

(c)

EXAMPLE 3.20 Comparing Data Sets by Using Boxplots

Skinfold Thickness A study was conducted to determine whether elite distance runners are actually thinner than other people. The researchers measured skinfold thickness, an indirect indicator of body fat, of samples of runners and non-runners in the same age group. The sample data, in millimeters (mm), presented in Table 3.13 are based on their results. Use boxplots to compare these two data sets, paying special attention to center and variation.

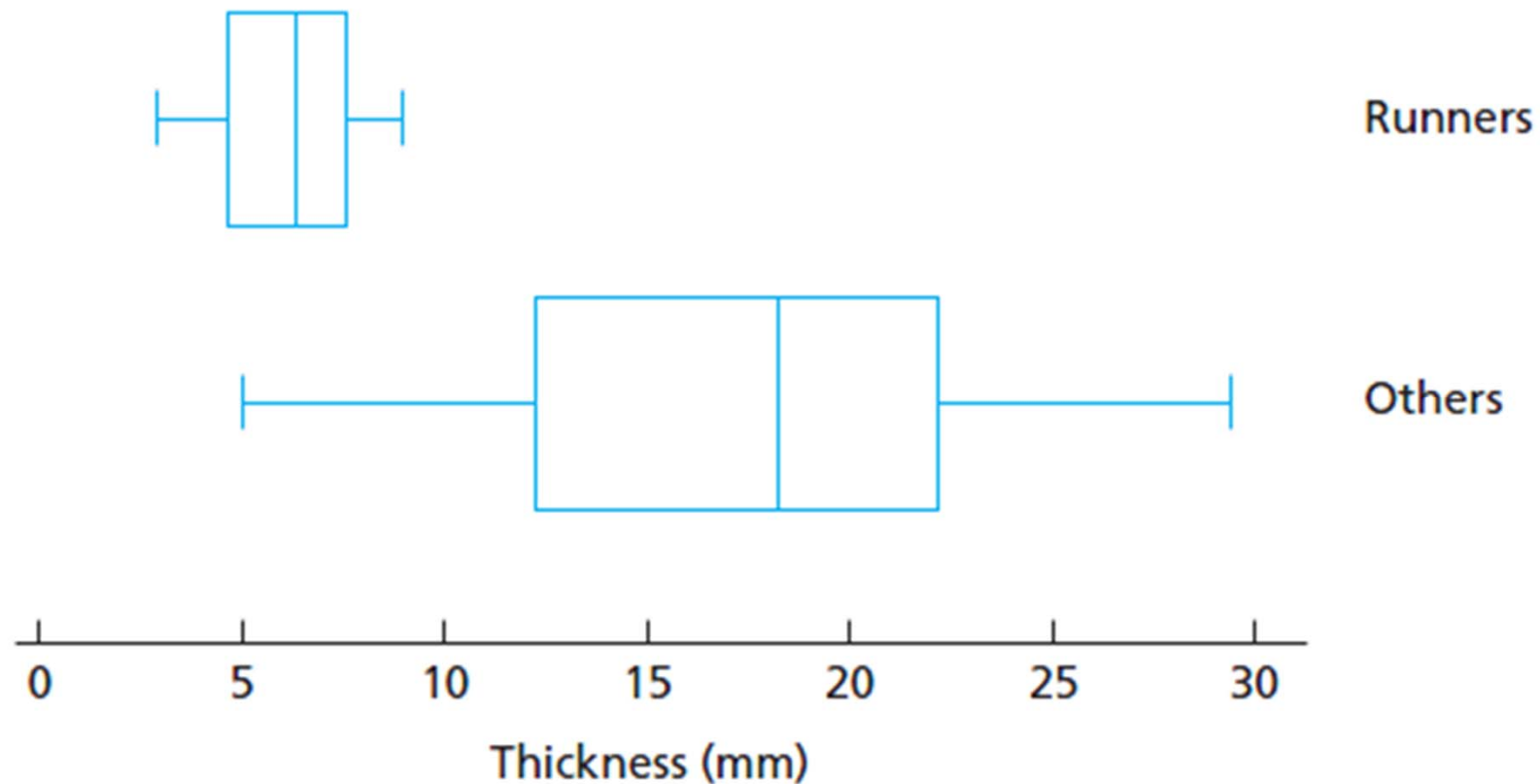
EXAMPLE 3.20 Comparing Data Sets by Using Boxplots

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

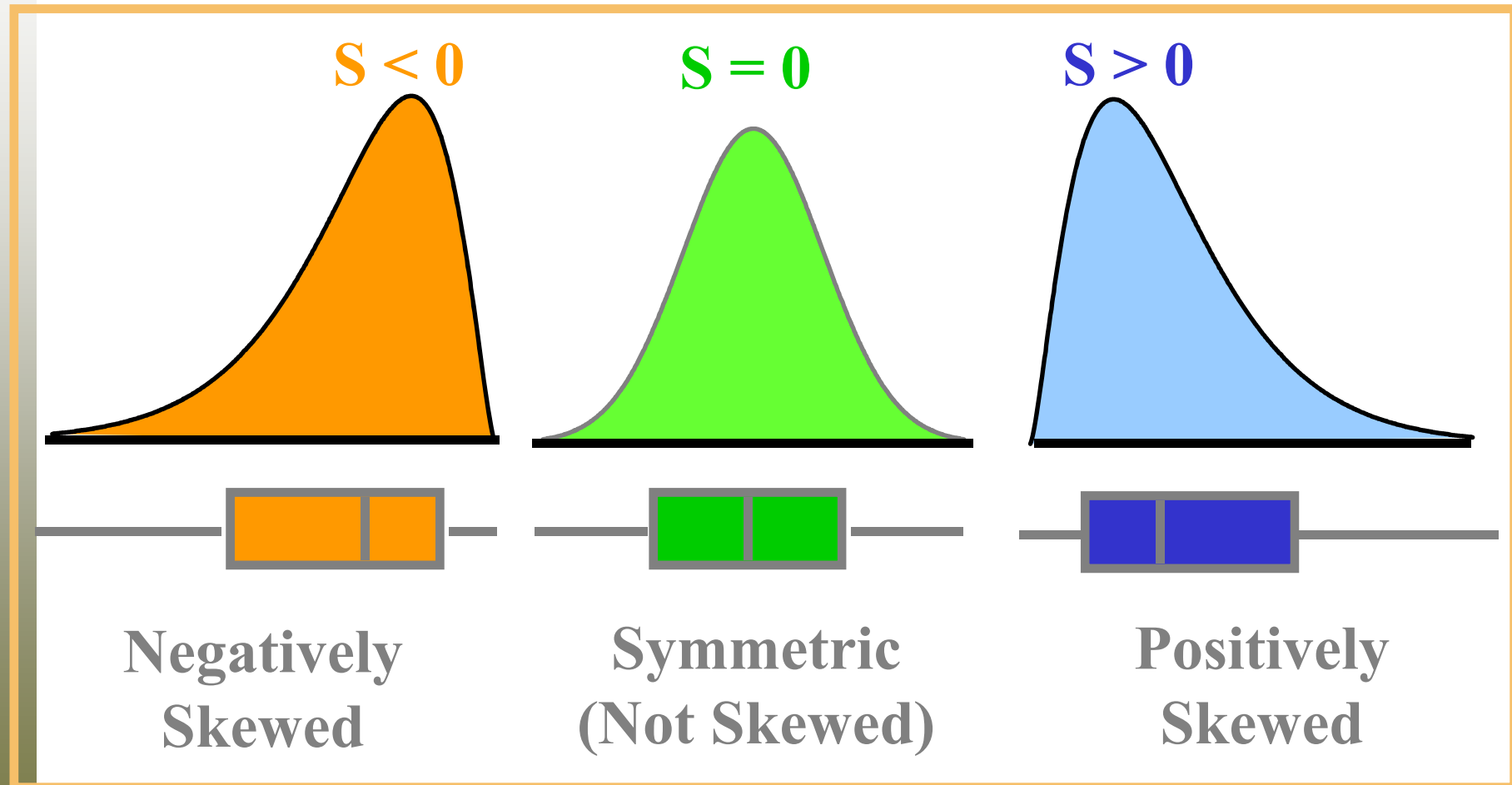
TABLE 3.13 Skinfold thickness (mm) for samples of elite runners and others

EXAMPLE 3.19 Weekly TV-Viewing Times

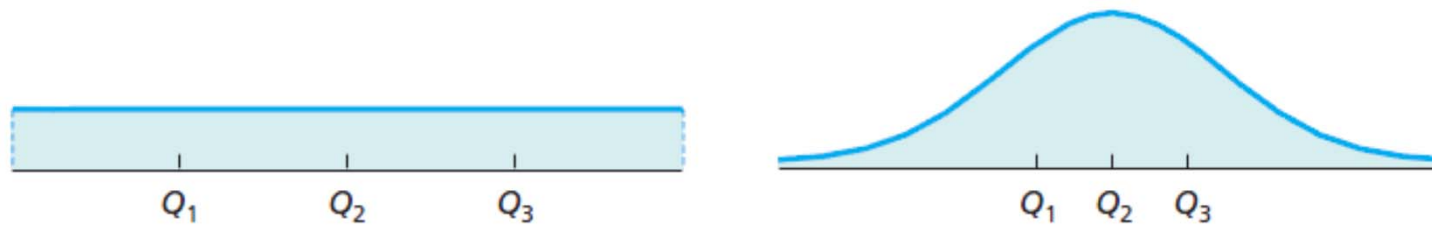
Solution: Figure 3.10 displays boxplots for the two data sets, **using the same scale.**



Boxplots (Box and Whisker Plots) and Skewness

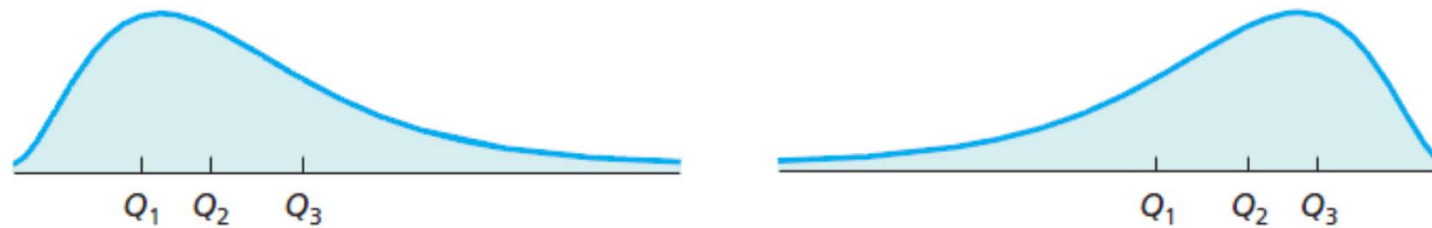


Boxplots (Box and Whisker Plots) and Shape of the data distribution



(a) Uniform

(b) Bell shaped



(c) Right skewed

(d) Left skewed

Section 3.4

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,

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

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

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

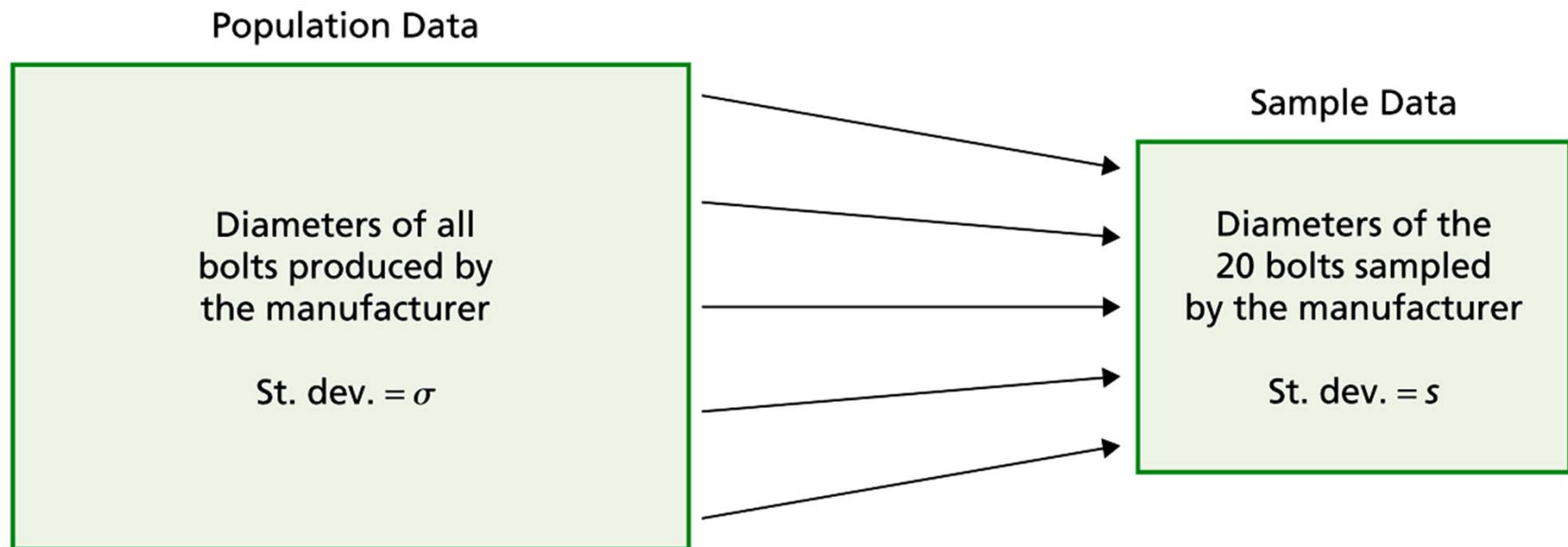
where N is the population size.

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

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

Figure 3.13 & Definition 3.13

Population and sample for bolt diameters



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*.