# 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.

Finding mean, median and mode of price of 7 round trip flights from Chicago to Mexico.

The ordered data set is

388 397 397 (427) 432 782 872

mean price of flights $\bar{x} = \frac{\Sigma x_i}{n} = \frac{3675}{7} = \$527.85$

Since the number of observations is odd, the median price is the single middle value = $427

Since 397 occurs twice (most often), the mode of the flight prices is $397

If the data set had 8 observations

388 397 397 (427 432) 782 872 892

mean = $\frac{4587}{8} = \$573.37$

median is the average of the two middle values

$\frac{427+432}{2} = \$429.50$

mode is $397

# 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

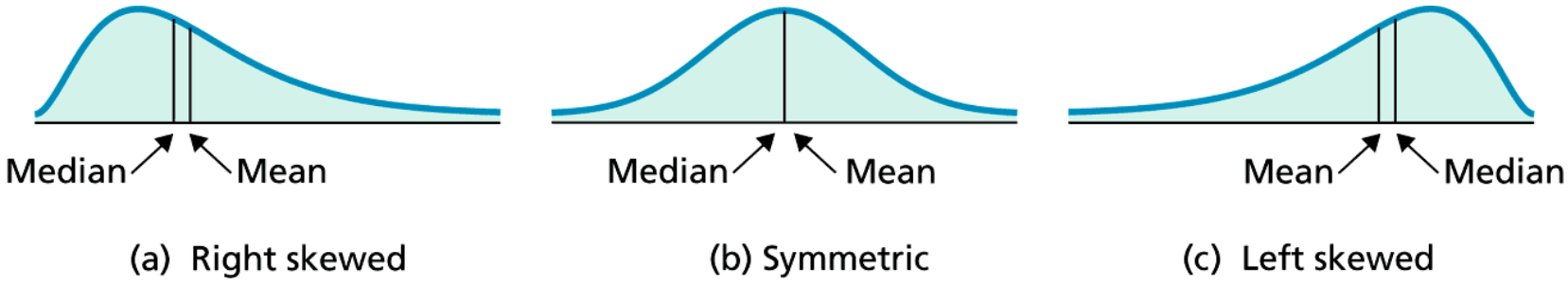
| | | | | |
|---|---|---|---|---|
| $300 | 300 | 940 | 450 | 400 |
| 400 | 300 | 300 | 1050 | 300 |

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



# 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.

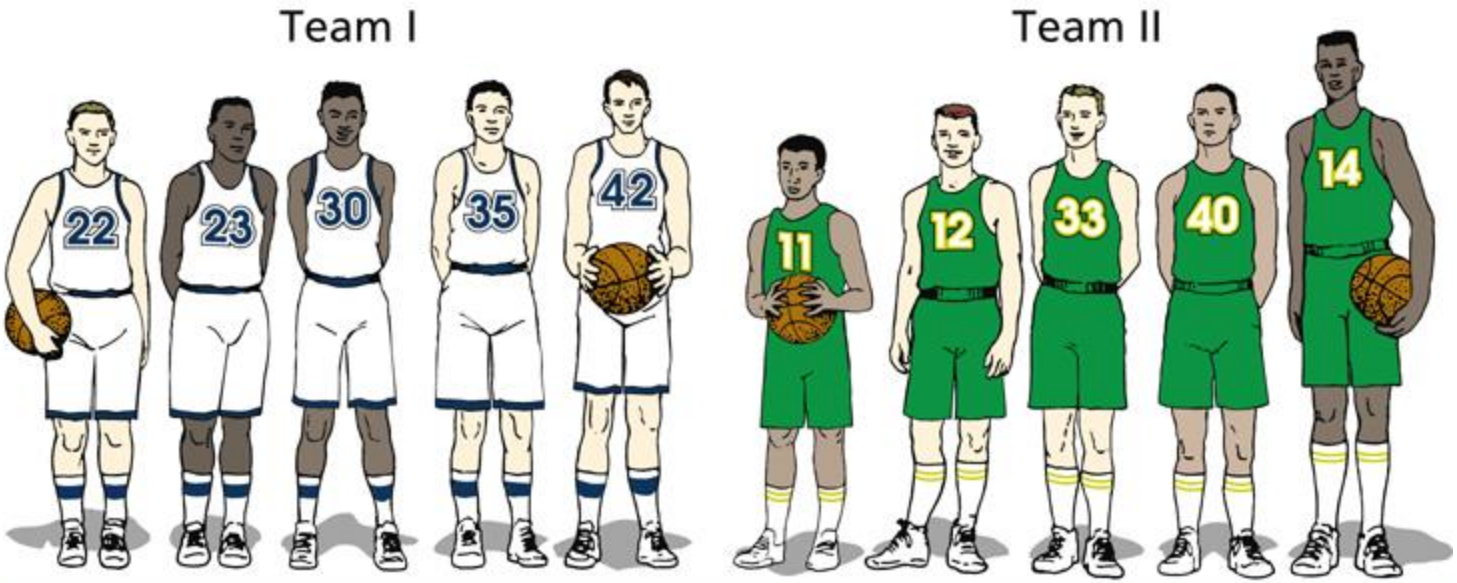
# Section 3.2

## Measures of Variation

# Figure 3.2

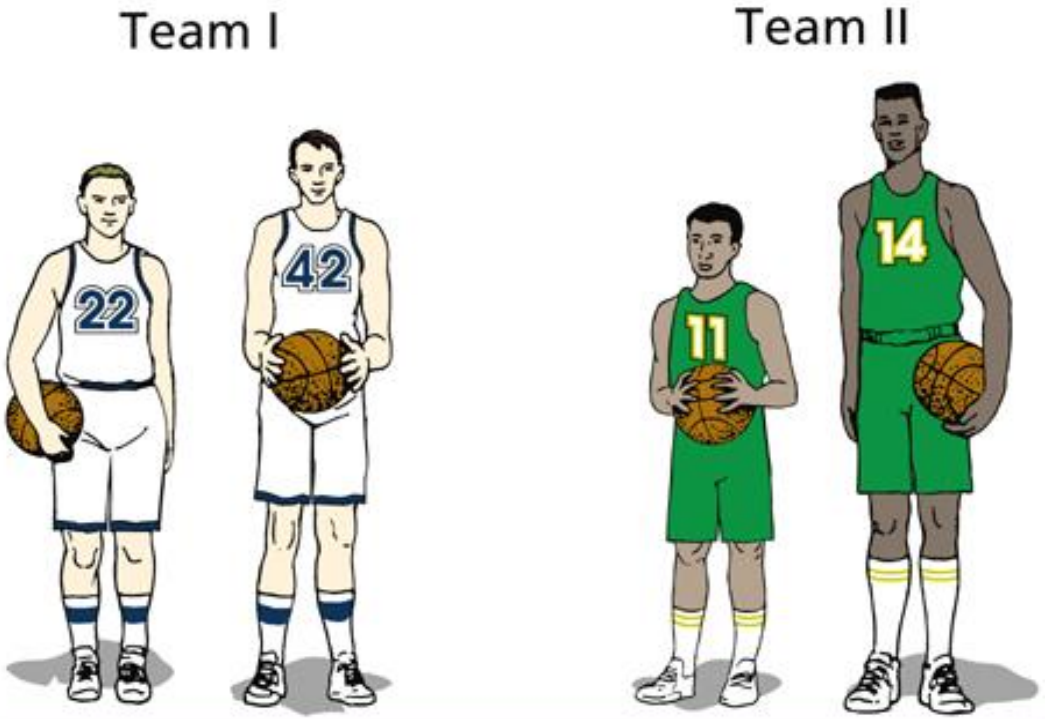
Five starting players on two basketball teams




| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 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




| | Team I | | Team II | |
|---|---|---|---|---|
| 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.

The Sample Standard Deviation is a measure of variation that takes into account all the observations in a data set.

It indicates how far on average the observations are from the mean

Computation:

1. Find the mean of the sample $\bar{x}$
2. Compute the deviation of each observation from the mean, i.e $x_i - \bar{x}$
3. Square each deviation $(x_i - \bar{x})^2$
4. Add the squared deviations from the mean
   $= \Sigma(x_i - \bar{x})^2$ = Sum of squared deviations
5. Compute $s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$ where '$s^2$' is called the sample variance (units are square of the original units)
6. Compute $s = \sqrt{s^2} = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$ where '$s$' is called the sample standard deviation expressed in the original units.

Computation of Sample variance and Sample Standard deviation

Data from Team 1 → heights of 5 players in inches

72, 73, 76, 76, 78

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{375}{5} = 75 \text{ inches}$$

| height $x_i$ | deviation from mean $(x_i - \bar{x})$ | squared deviation from mean |
|---|---|---|
| 72 | $72 - 75 = -3$ | 9 |
| 73 | $73 - 75 = -2$ | 4 |
| 76 | $76 - 75 = 1$ | 1 |
| 76 | $76 - 75 = 1$ | 1 |
| 78 | $78 - 75 = 3$ | 9 |
| | $\Sigma(x_i - \bar{x}) = 0$ | $\Sigma(x_i - \bar{x})^2 = 24$ |

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

$$s = \sqrt{6} = 2.4 \text{ inches}$$

On average, the heights of players on Team 1 vary from the mean height of 75 inches by about 2.4 inches.

# 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 $\boldsymbol{s_x}$ or, when no confusion will arise, simply $\boldsymbol{s}$. 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

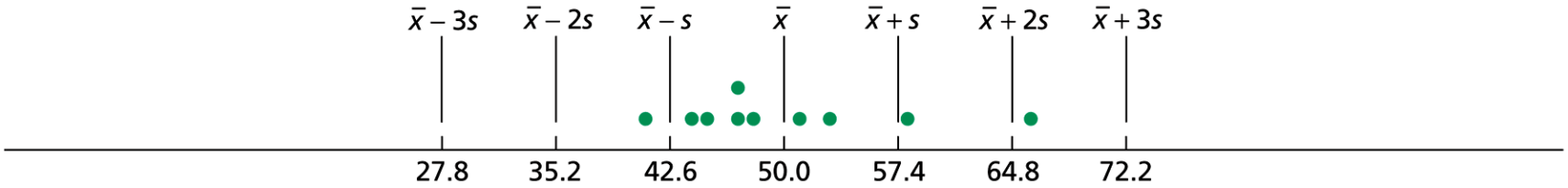
Data sets that have different variation

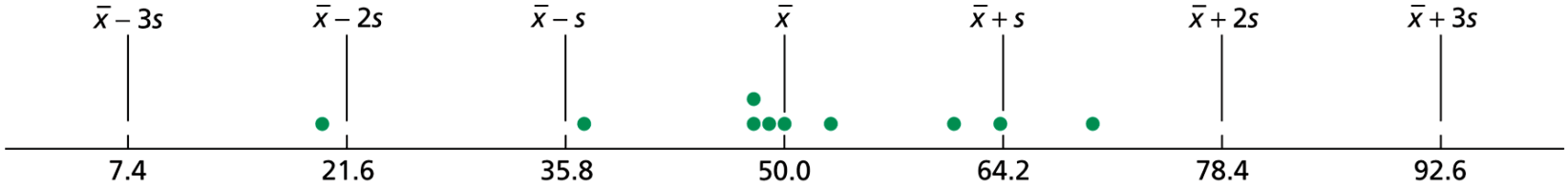
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Data Set I** | 41 | 44 | 45 | 47 | 47 | 48 | 51 | 53 | 58 | 66 |
| **Data Set II** | 20 | 37 | 48 | 48 | 49 | 50 | 53 | 61 | 64 | 70 |

Means and standard deviations of the data sets in Table 3.10

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

# 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 $\bar{x} + k \cdot s$.

Example: The body mass index (BMI) of 75 randomly selected U.S. adults has a mean of 26.0 and a s.d of 5.0

$n = 75$

$\bar{x} = 26$

$s = 5$

Using Chebyshev's Rule

for $k=2$, $1-\frac{1}{k^2} = 1-\frac{1}{4} = 0.75 \Rightarrow$ approx$_{\text{at least}}$ 75% of adults have BMI between $\bar{x}-2s$ and $\bar{x}+2s$ i.e between $26-2(5)=16$ and $26+2(5)=36$

for $k=3$, $1-\frac{1}{k^2} = 1-\frac{1}{9} = 0.89 \Rightarrow$ approx$_{\text{at least}}$ 89% of adults have BMI between $\bar{x}-3s$ and $\bar{x}+3s$ i.e between $26-15=11$ and $26+15=41$

Also 75% of $n=75$ is $0.75(75) = 56.25 \simeq$ at least 56 of 75 people have bmi between 16 and 36 and 89% of $n=75$ is $0.89(75)=67$ i.e at least 67 of 75 people have bmi between 11 and 41.

# Key Fact 3.4

## Empirical Rule

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

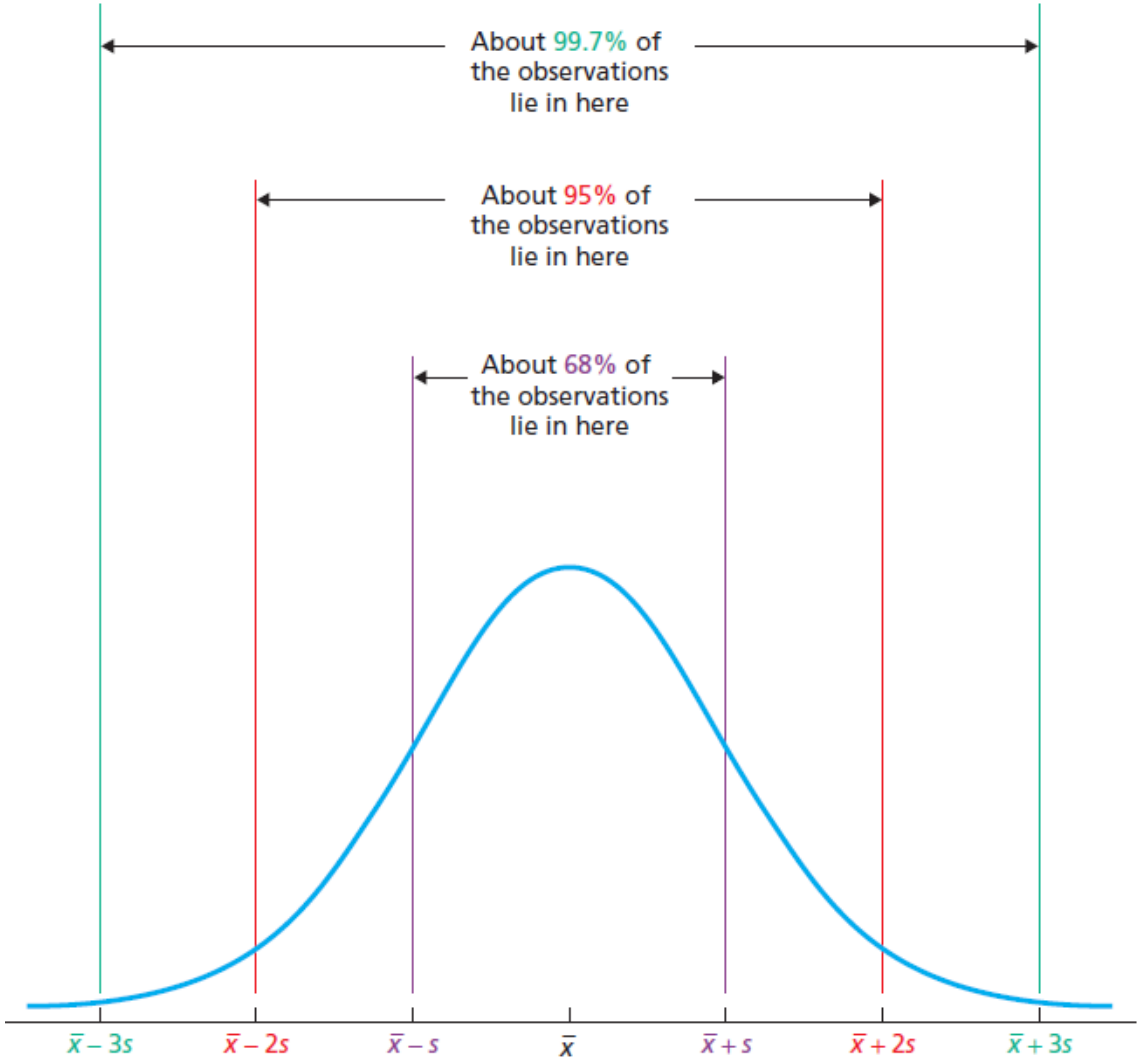
**Property 1:** Approximately 68% of the observations lie within one standard deviation to either side of the mean, that is, between $\bar{x} - s$ and $\bar{x} + s$.

**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



If the distribution of BMI is symmetric, the Empirical Rule can be used.

for $k=2$, approx 95% of people have BMI between 16 and 36 or approx $0.95(75) = 71.25$, so approx 71 of US adults out of 75 in sample have bmi between 16 and 36.

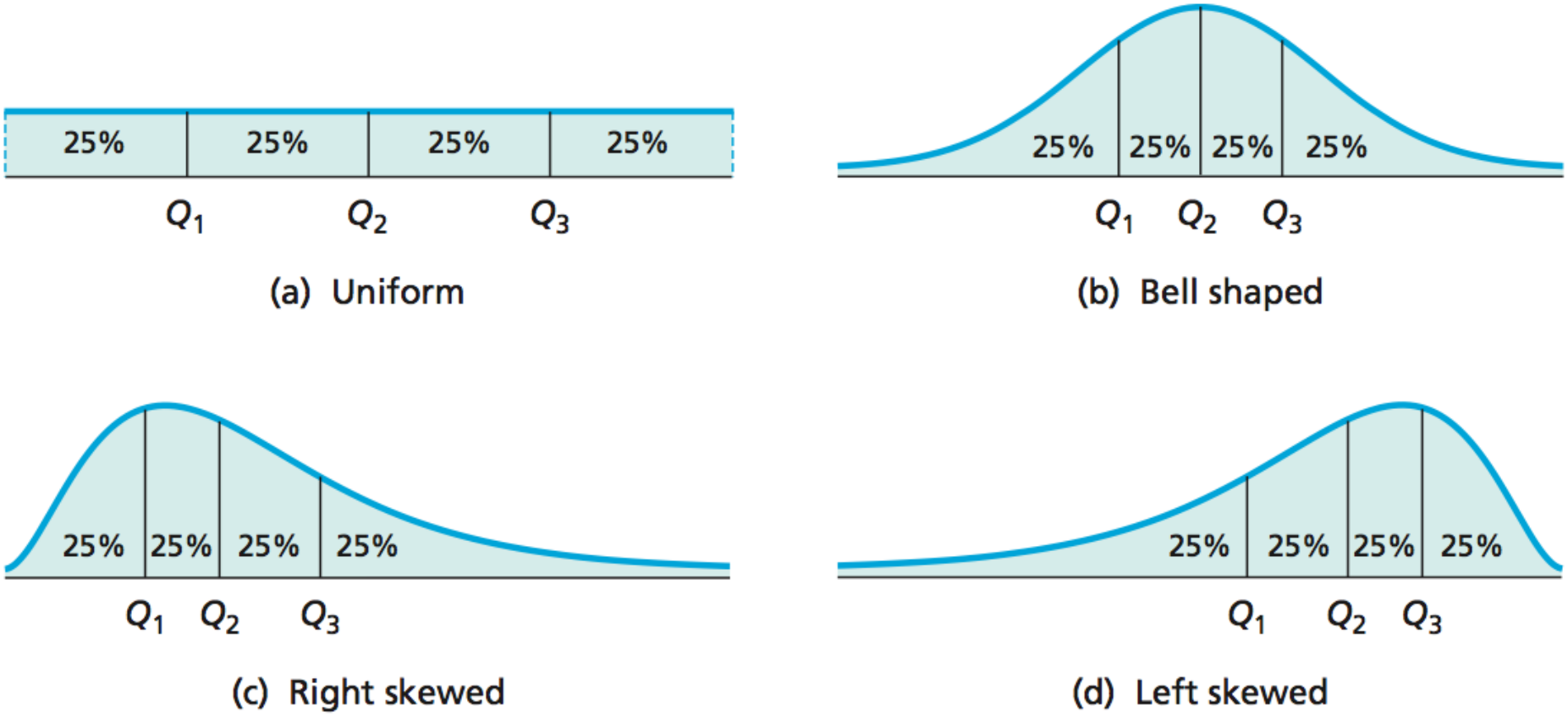
Similarly for $k=3$, approx 99.7% of people have BMI between 11 and 41 or approx $0.997(75) = 74.775$, so approx 75 of 75 adults in sample have BMI between 11 and 41.

# 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, exclude 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

$$\text{Min}, Q_1, Q_2, Q_3, \text{Max}.$$

## Computation of Quartiles of 20 TV viewing times

data: Weekly TV viewing times in hours

$n = 20$ ; ordered values are:

5, 15, 16, 20, (21, 25), 26, 27, 30, 30 → $n=10$ lower half

31, 32, 32, 34, (35, 38), 38, 41, 43, 66 → $n=10$ upper half

median is average of 10th and 11th values $\frac{30+31}{2} = \frac{30.5}{\text{hours}} = Q_2$

$Q_1$ = 1st quartile $= \frac{21+25}{2} = 23$ hours → indicating that 25% of viewing times per wk are less than 23 hours

$Q_3 = \frac{35+38}{2} = 36.5$ hours → indicating that 25% of viewing times are more than 36.5 hours

IQR = Interquartile range $= Q_3 - Q_1 = 36.5 - 23 = 13.5$ hours which is the range of the middle 50% of TV viewing times

OR the middle 50% of the data set (TV viewing times) varies between 23 and 36.5 hours

5 # summary is $\min, Q_1, Q_2, Q_3, \max = 5, 23, 30.5, 36.5, 66$

37, 38, 39, 41, 41, 41, 42, 44, 45, 47

n = 10

Q2 = (41+41)/2 = 41

Q1= 39

Q3 = 44

37, 38, 39, 41, 41, 41, 42, 44, 45

n = 9

Q2 = 41

Q1 = (38+39)/2 = 38.5

Q3 = (42+44)/2 = 43

Finding Quartiles

# Definition 3.10

**Lower and Upper Limits**

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

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

Outliers are extremely low or high observations far removed from the actual data set.

Observations that lie below the lower limit or above the upper limit are potential outliers, where

Lower limit $= Q_1 - 1.5\,IQR$

& upper limit $= Q_3 + 1.5\,IQR$

In the TV viewing example,

lower limit $= 23 - 1.5(13.5) = 2.75$ hrs

upper limit $= 36.5 + 1.5(13.5) = 56.75$ hrs

The data value 66 in the data set lies outside (above) the upper limit and is a potential outlier.

The adjacent values of the data set are the most extreme values that are not potential outliers and in the TV viewing example, the values are 5 and 43.

# 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.

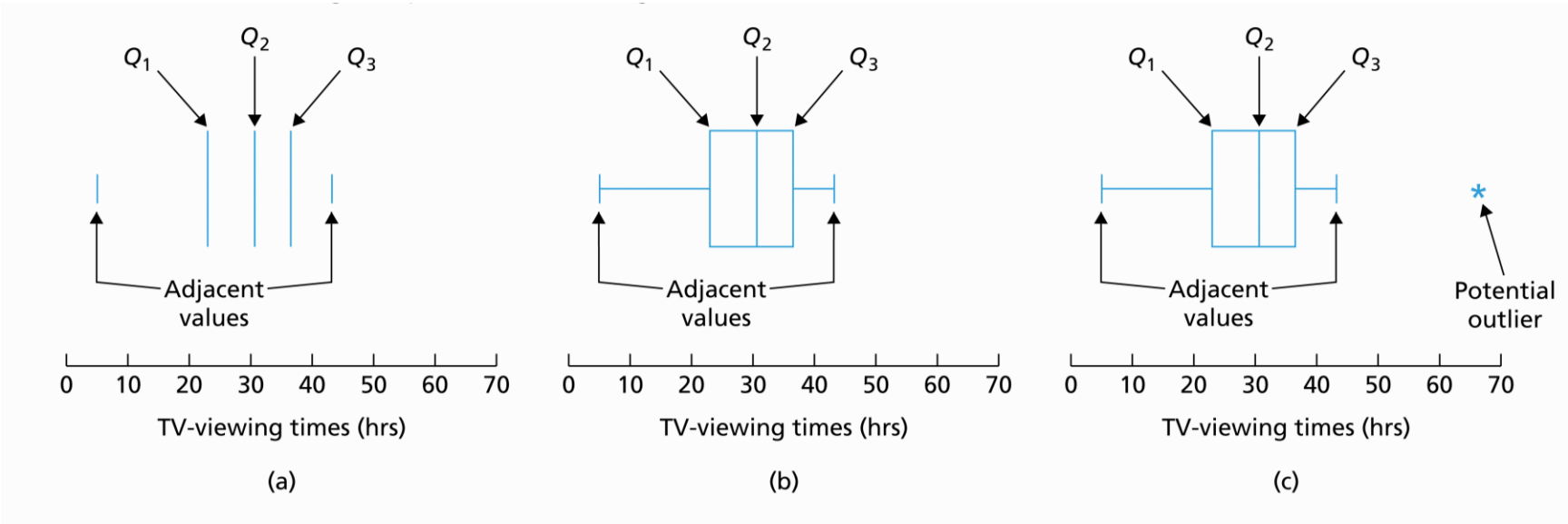
**Step 4** Connect the quartiles to make a box, and then connect the box to the adjacent values with lines.

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

# Figure 3.14

Constructing a boxplot for TV viewing times in Table 3.13

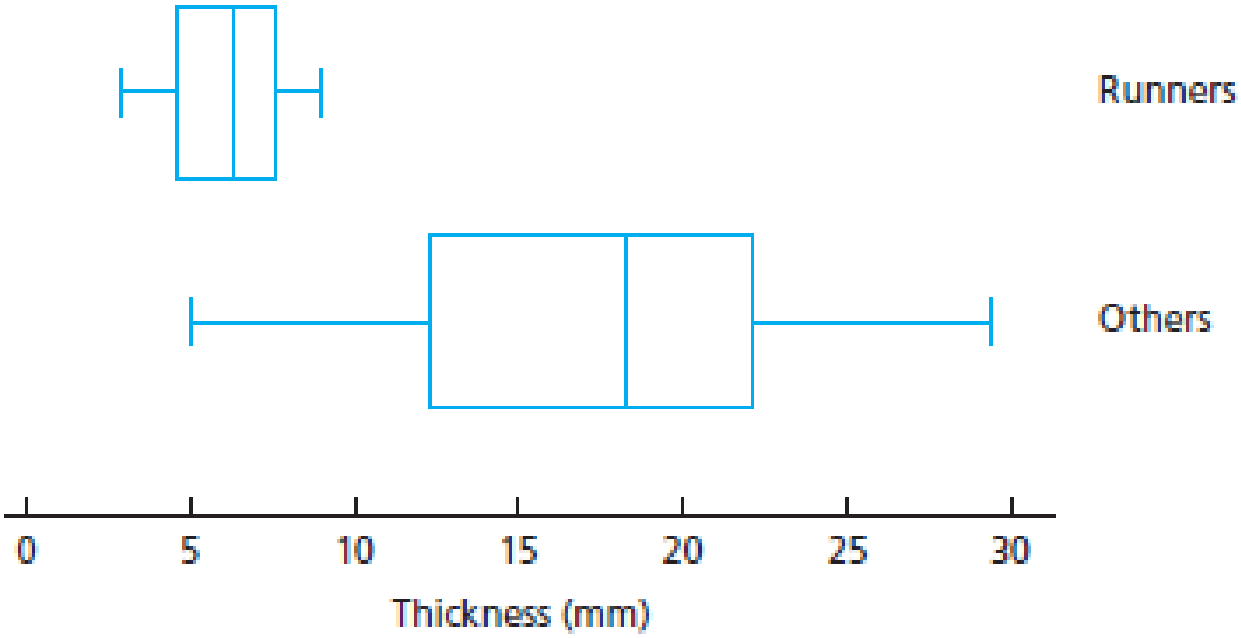
| | | | | |
|---|---|---|---|---|
| 25 | 41 | 27 | 32 | 43 |
| 66 | 35 | 31 | 15 | 5 |
| 34 | 26 | 32 | 38 | 16 |
| 30 | 38 | 30 | 20 | 21 |



# Figure 3.15

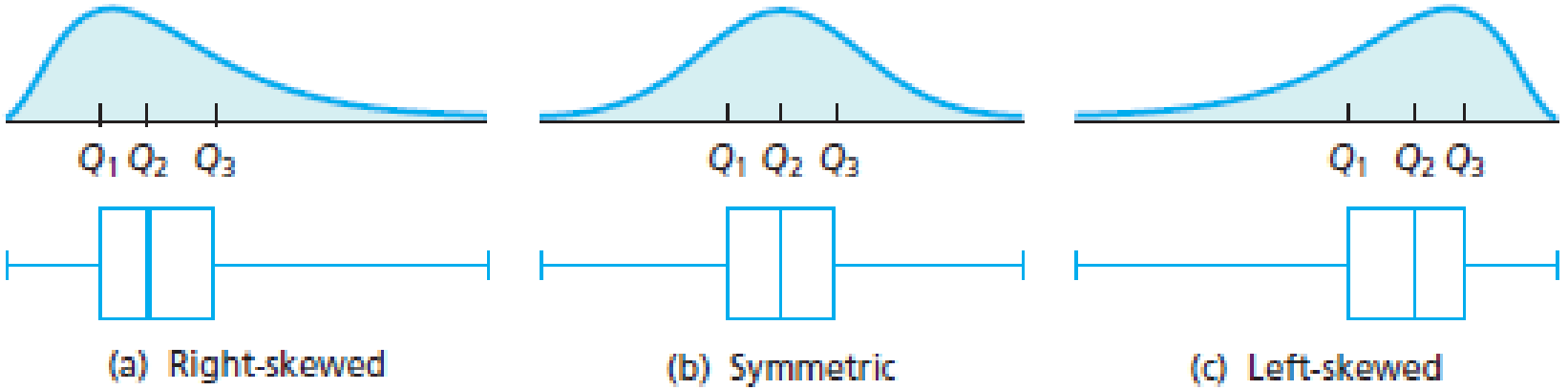
Boxplots for the data in Table 3.15

| 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 $\mu_x$ or, when no confusion will arise, simply $\mu$. 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 $\sigma_x$ or, when no confusion will arise, simply $\sigma$. For a finite population, the defining formula is

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

where $N$ is the population size.

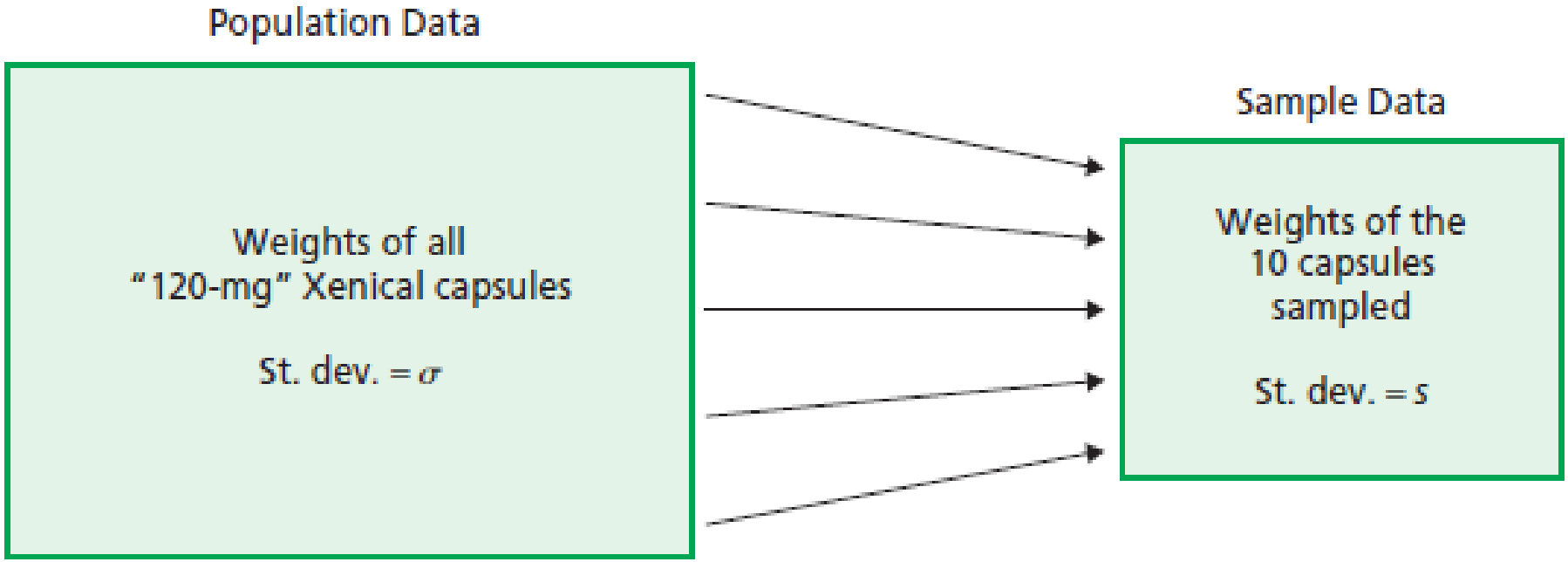
The population standard deviation can also be found from the computing formula

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

# 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.*

The Z-score of an observation can also be used as a rough measure of its relative standing among all observations in a dataset.

Example: A statistics exam has a $\mu = 63$ and $\sigma = 7$ and a Biology exam has grades with a $\mu = 23$ and $\sigma = 3.9$
A student scored 60 on the Statistics exam and a 22 on the Biology exam. In which exam did he do better?

Compute the two Z-scores and compare

Stats exam

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

$$= \frac{60 - 63}{7}$$

$$= -0.43$$

Bio exam

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

$$= \frac{22 - 23}{3.9}$$

$$= -0.25$$

→ Student scored below the mean in both exams with 0.43 standard deviations below the mean for Stats exam and 0.25 s.d below the mean for the Bio exam.

So he/she fared better in the Bio exam.