

### 3. NUMERICAL DESCRIPTIVE MEASURES

Mean, median, mode, variance, IQR, etc

# Objective

- In chapter 2, we have organized our data, and visually inspect them using various graphs and plots.
- In this chapter, we would like to describe or summarize the data numerically.
- We will only cover ungrouped data, as it is more common and relevant for this course.

# Measures of center (ungrouped data)

Mean, median, mode

# Measures of center

- Measure of center – a value that represents a typical, or central, entry of a data set.
- Most common measures of central tendency:
  - ▣ Mean
  - ▣ Median
  - ▣ Mode

# Mean

- The sum of all the data entries divided by the number of entries.

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

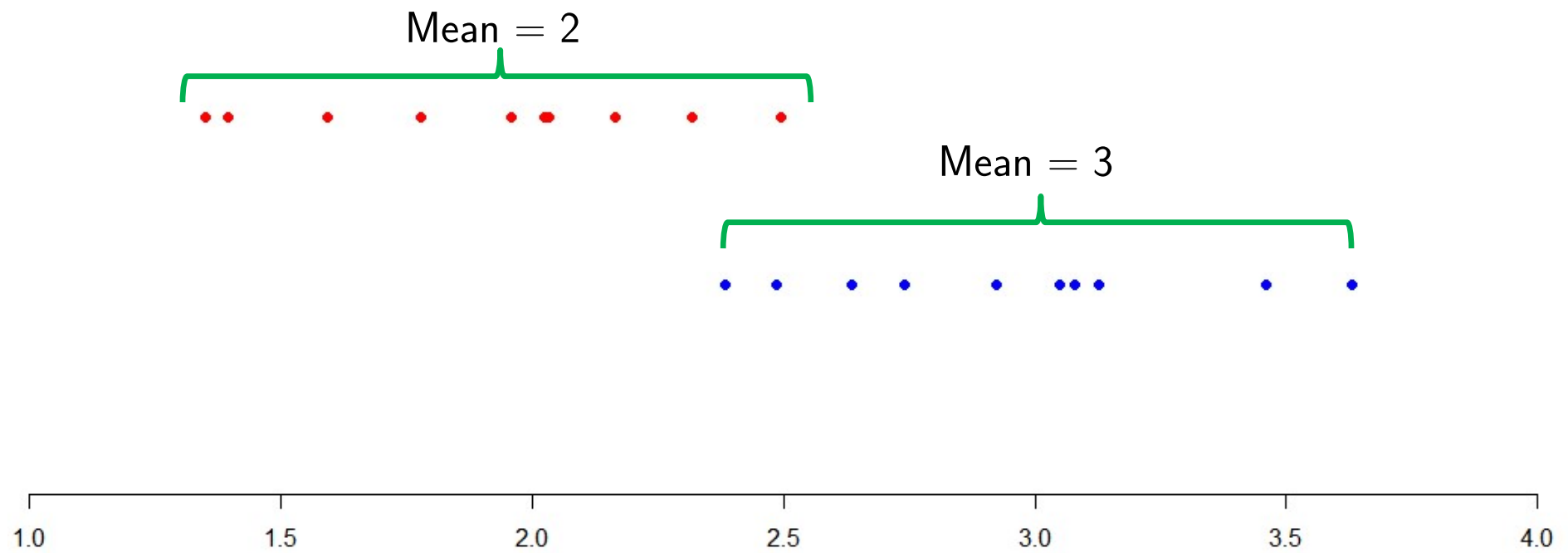
- Sigma notation:  $\Sigma x$  = add all of the data entries ( $x$ ) in the data set.
- Suppose  $N$  is the population size, and  $n$  is the sample size.
- **Population** mean:

$$\mu = \frac{\Sigma x}{N}$$

- **Sample** mean:

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

# Mean



# Example – finding mean

## EXAMPLE 3–1 2014 Profits of 10 U.S. Companies

Table 3.1 lists the total profits (in million dollars) of 10 U.S. companies for the year 2014 (www.fortune.com).

**Table 3.1 2014 Profits of 10 U.S. Companies**

Company	Profits (million of dollars)
Apple	37,037
AT&T	18,249
Bank of America	11,431
Exxon Mobil	32,580
General Motors	5346
General Electric	13,057
Hewlett-Packard	5113
Home Depot	5385
IBM	16,483
Wal-Mart	16,022

Find the mean of the 2014 profits for these 10 companies.

# Example – finding mean

- Sample size:  $n = 10$
- Denote the variable Profits as  $x$
- Sum of all profits:

$$\begin{aligned}\sum x &= 37,037 + 18,249 + 11,431 + 32,580 + 5346 \\ &\quad + 13,057 + 5113 + 5385 + 16,483 + 16,022 \\ &= 160,703\end{aligned}$$

- Sample mean:

$$\bar{x} = \frac{\sum x}{n} = \frac{160,703}{10} = 16070.3$$

- The mean profits is \$16070.3 million.



# Example – finding mean

## **EXAMPLE 3–2** Ages of Employees of a Company

The following are the ages (in years) of all eight employees of a small company:

53      32      61      27      39      44      49      57

Find the mean age of these employees.

# Example – finding mean

□ Population size,  $N = 8$

□ Sum of all ages:

$$\sum x = 53 + 32 + 61 + 27 + 39 + 44 + 49 + 57 = 362$$

□ Population mean:

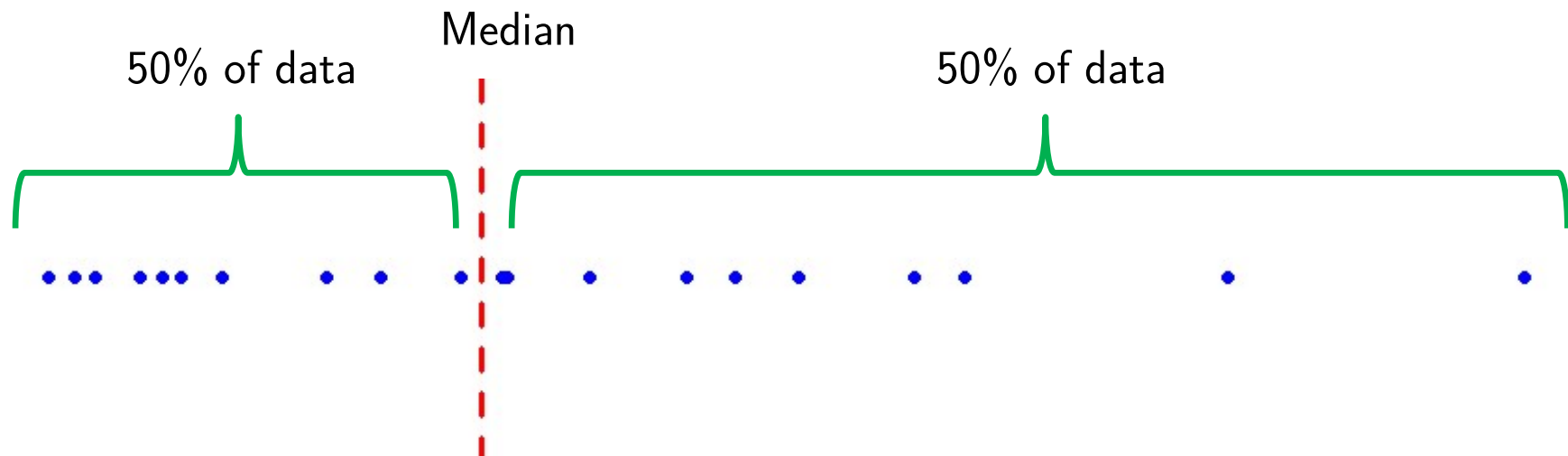
$$\mu = \frac{\sum x}{N} = \frac{362}{8} = 45.25$$

□ Mean age of all eight employees of the company is 45.25 years.

# Median

- The value that **lies in the middle** of the data when the data set is ordered.
- Calculation:
  1. Rank the given data set in increasing order.
  2. Find the value that divides the ranked data in two equal size parts.
- If the data set has an
  - ▣ odd number of entries: median is the middle data entry.
  - ▣ even number of entries: median is the mean of the two middle data entries.

# Median



# Example – finding median

## EXAMPLE 3–4 Compensations of Female CEOs

Table 3.2 lists the 2014 compensations of female CEOs of 11 American companies (*USA TODAY*, May 1, 2015). (The compensation of Carol Meyrowitz of TJX is for the fiscal year ending in January 2015.)

**Table 3.2** Compensations of 11 Female CEOs

Company & CEO	2014 Compensation (millions of dollars)
General Dynamics, Phebe Novakovic	19.3
GM, Mary Barra	16.2
Hewlett-Packard, Meg Whitman	19.6
IBM, Virginia Rometty	19.3
Lockheed Martin, Marillyn Hewson	33.7
Mondelez, Irene Rosenfeld	21.0
PepsiCo, Indra Nooyi	22.5
Sempra, Debra Reed	16.9
TJX, Carol Meyrowitz	28.7
Yahoo, Marissa Mayer	42.1
Xerox, Ursula Burns	22.2

Find the median for these data.

# Example – finding median

- Ranked data in increasing order:

16.2 16.9 19.3 19.3 19.6 21.0 22.2 22.5 28.7 33.7 42.1

- Find the middle value:

16.2 16.9 19.3 19.3 19.6 **21.0** 22.2 22.5 28.7 33.7 42.1

└──────────┘ └──────────┘

5 entries 5 entries

- Median = \$21.0 million

# Example – finding median

## EXAMPLE 3–5 Cell Phone Minutes Used

The following data give the cell phone minutes used last month by 12 randomly selected persons.

230   2053   160   397   510   380   263   3864   184   201   326   721

Find the median for these data.

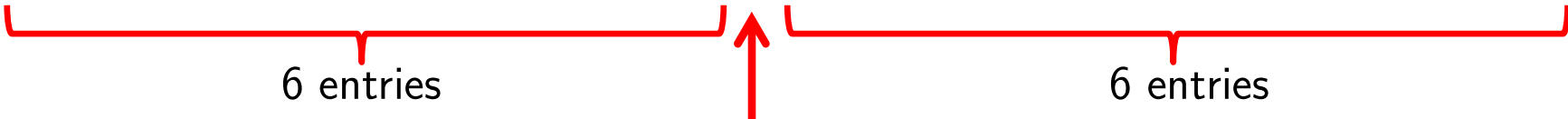
# Example – finding median

- Ranked data in increasing order:

160 184 201 230 263 326 380 397 510 721 2053 3864

- Find the middle value:

160 184 201 230 263 326 380 397 510 721 2053 3864



- Calculate median:

$$\text{Median} = \frac{326 + 380}{2} = 353$$

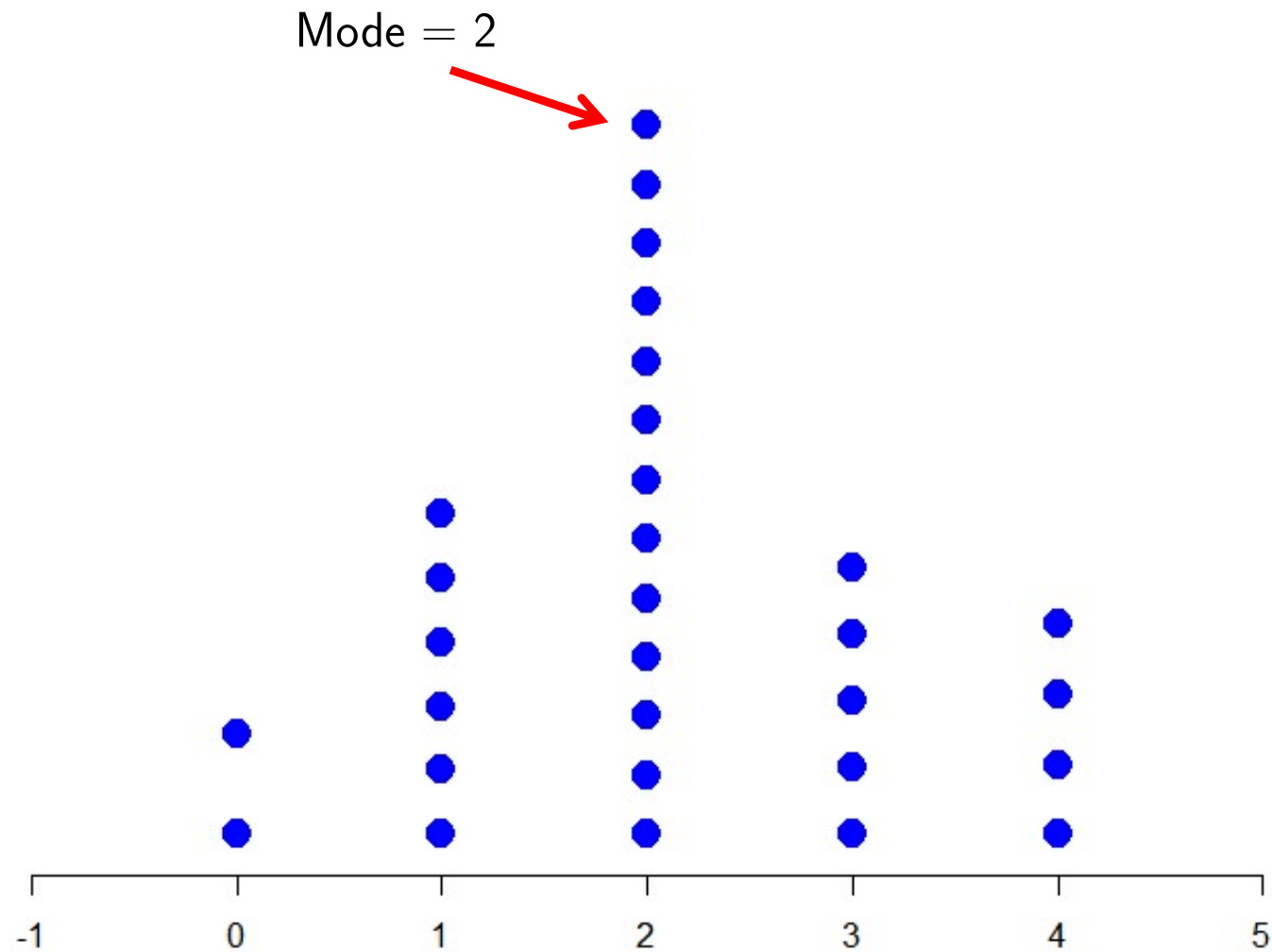
- Median is 353 minutes.



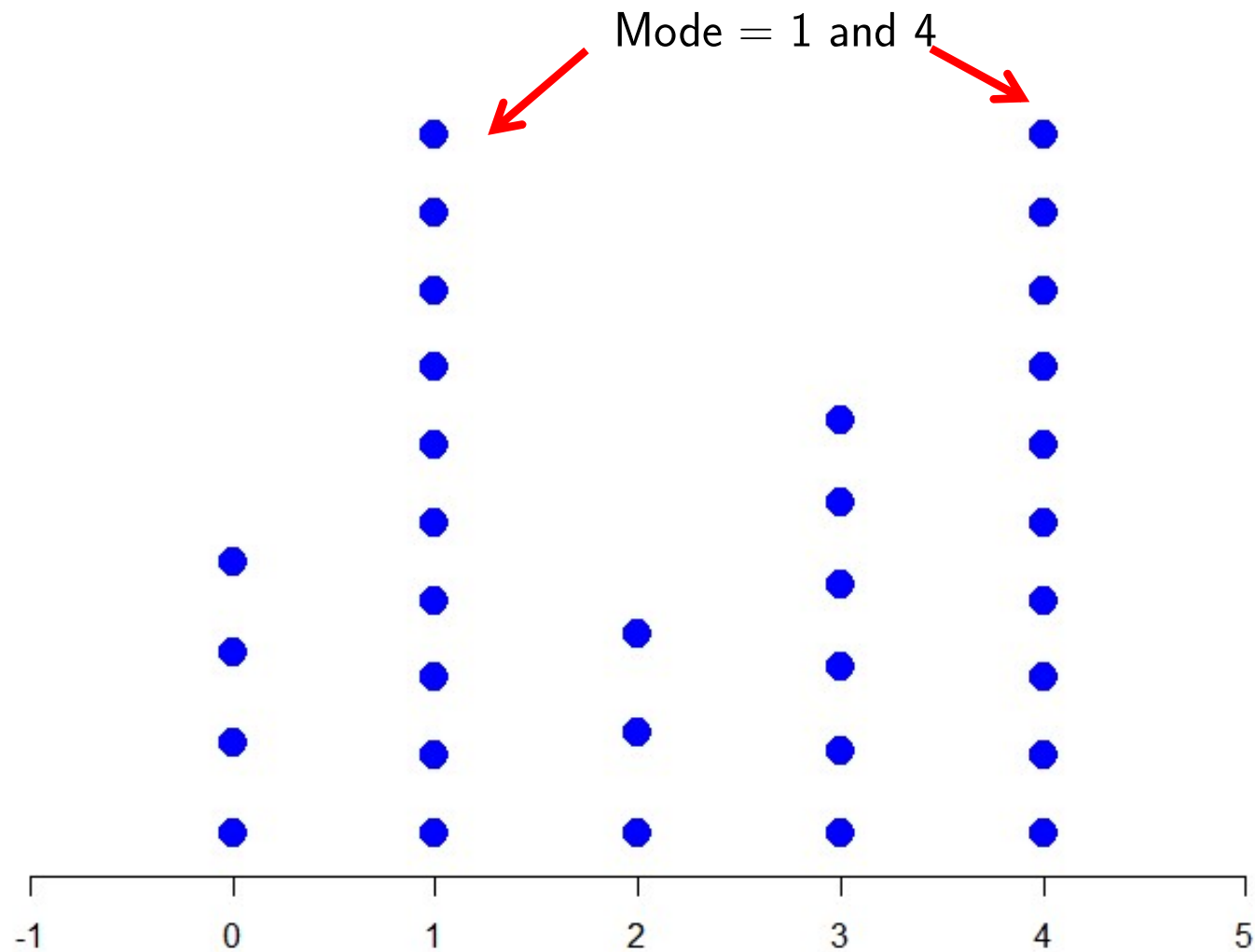
# Mode

- The data entry that occurs with the **highest frequency**.
- If no entry is repeated the data set has no mode.
- If two entries occur with the same greatest frequency, each entry is a mode (bimodal).

# Mode



# Mode



# Example – finding mode

## EXAMPLE 3–6 Speeds of Cars

The following data give the speeds (in miles per hour) of eight cars that were stopped on I-95 for speeding violations.

77      82      74      81      79      84      74      78

Find the mode.

- Only 74 occurs twice, while the other values occur only once.
- Therefore, Mode = 74.

# Example – finding mode

## EXAMPLE 3–7 Incomes of Families

Last year's incomes of five randomly selected families were \$76,150, \$95,750, \$124,985, \$87,490, and \$53,740. Find the mode.

- All values occur once.
- There is no mode in the data.

# Example – finding mode

## EXAMPLE 3–8 Commuting Times of Employees

A small company has 12 employees. Their commuting times (rounded to the nearest minute) from home to work are 23, 36, 14, 23, 47, 32, 8, 14, 26, 31, 18, and 28, respectively. Find the mode for these data.

- 14 and 23 occur twice in the data, while the other values occur once.
- This data has two modes: 14 and 23

# Example – finding mode

## EXAMPLE 3–10 Status of Students

The status of five students who are members of the student senate at a college are senior, sophomore, senior, junior, and senior, respectively. Find the mode.

- Senior occurs 3 times, whereas sophomore and junior occur once.
- Senior is the mode for this data set.
- Since this is a qualitative variable, we cannot calculate the mean and median.

# Exercise

population dataset

**3.18** The following table gives the number of major penalties for each of the 15 teams in the Eastern Conference of the National Hockey League during the 2008–09 season (*NHL*, 2009). A major penalty is subject to 5 minutes in the penalty box for a player.

Team	Number of Major Penalties
Philadelphia	65
Columbus	59
Boston	53
Pittsburgh	51
New York Rangers	50
Tampa Bay	40
Nashville	39
Florida	38
Ottawa	35
Washington	35
Montreal	34
Atlanta	31
New York Islanders	29
Buffalo	26
Toronto	25

mean = 40.67  
median = 38  
mode = 35

Compute the mean and median for the data on major penalties. Do these data have a mode? Why or why not?





# Exercise

sample dataset

**3.23** The following data represent the numbers of tornadoes that touched down during 1950 to 1994 in the 12 states that had the most tornadoes during this period (*Storm Prediction Center*, 2009). The data for these states are given in the following order: CO, FL, IA, IL, KS, LA, MO, MS, NE, OK, SD, TX.

1113 2009 1374 1137 2110 1086 1166 1039 1673 2300 1139 5490

- Calculate the mean and median for these data.
- Identify the outlier in this data set. Drop the outlier and recalculate the mean and median. Which of these two summary measures changes by a larger amount when you drop the outlier?
- Which is the better summary measure for these data, the mean or the median? Explain.

a) mean = 1803  
Ranked data = 1039 1086 1113 1137 1139 1166 1374 1673 2009 2110 2300 5490  
median =  $(1166 + 1374) / 2 = 1270$

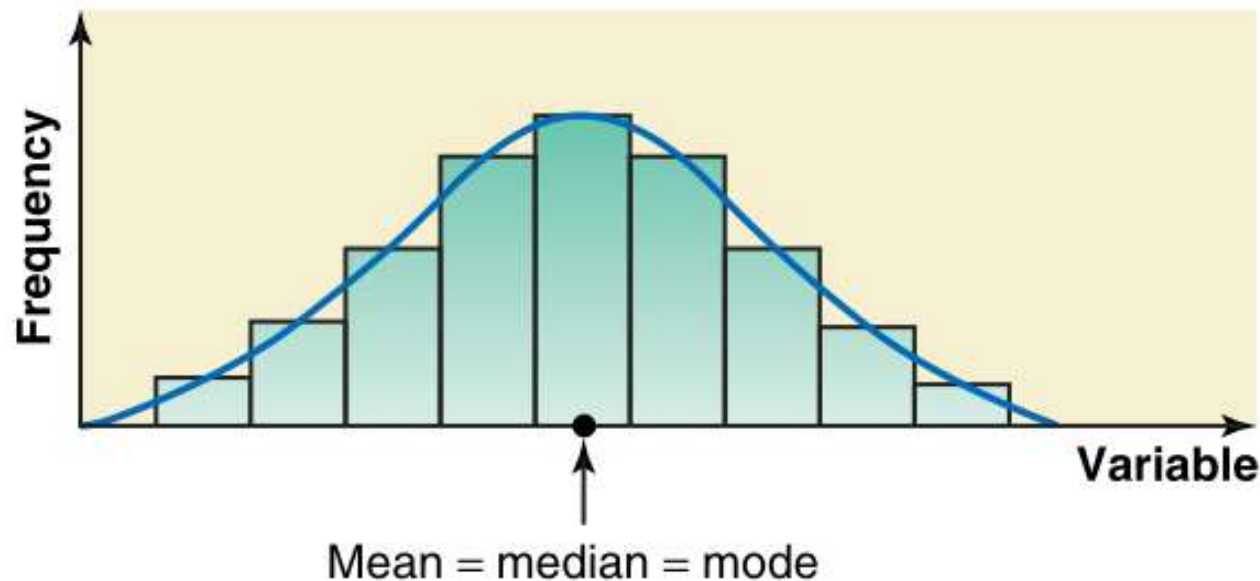
b) Identify outlier = 5490 can be considered outlier  
New mean =  $16146 / 11 = 1467.82$   
New median = 1166  
Mean changes by a larger amount when the outlier is dropped

c) Median is better summary measure because the data contains outlier



# Relationships Between Mean, Median, and Mode

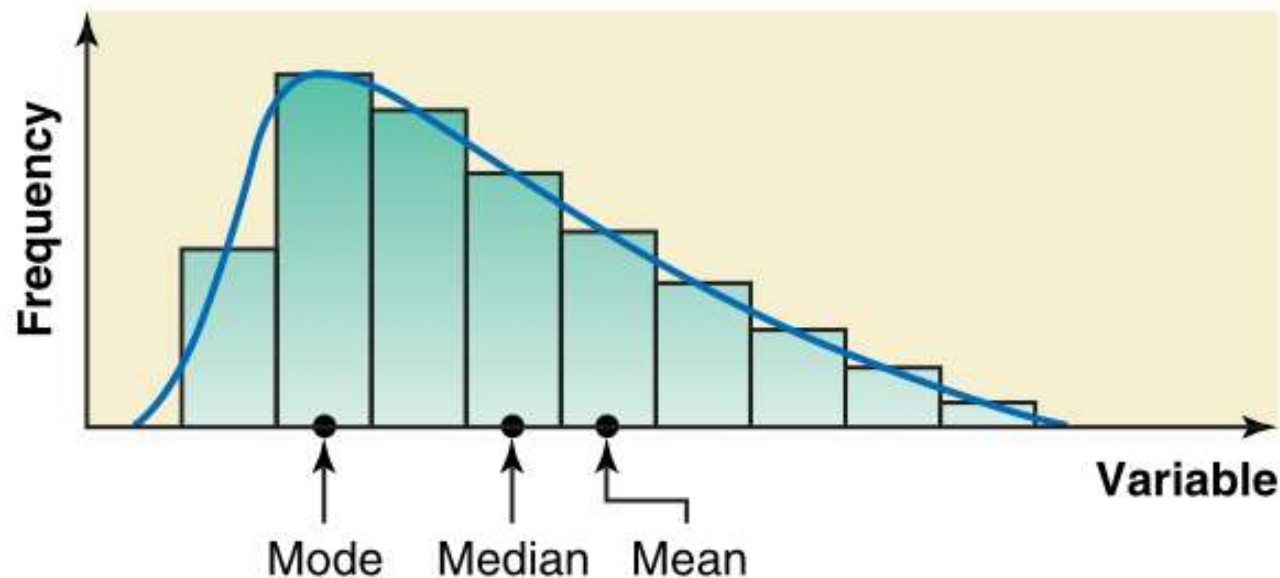
- Knowing the mean, median and, mode can give us some idea about the shape of a frequency distribution curve.
- Symmetric: Mean = median = mode



# Relationships Between Mean, Median, and Mode

- Skewed to the right (or positively skewed):

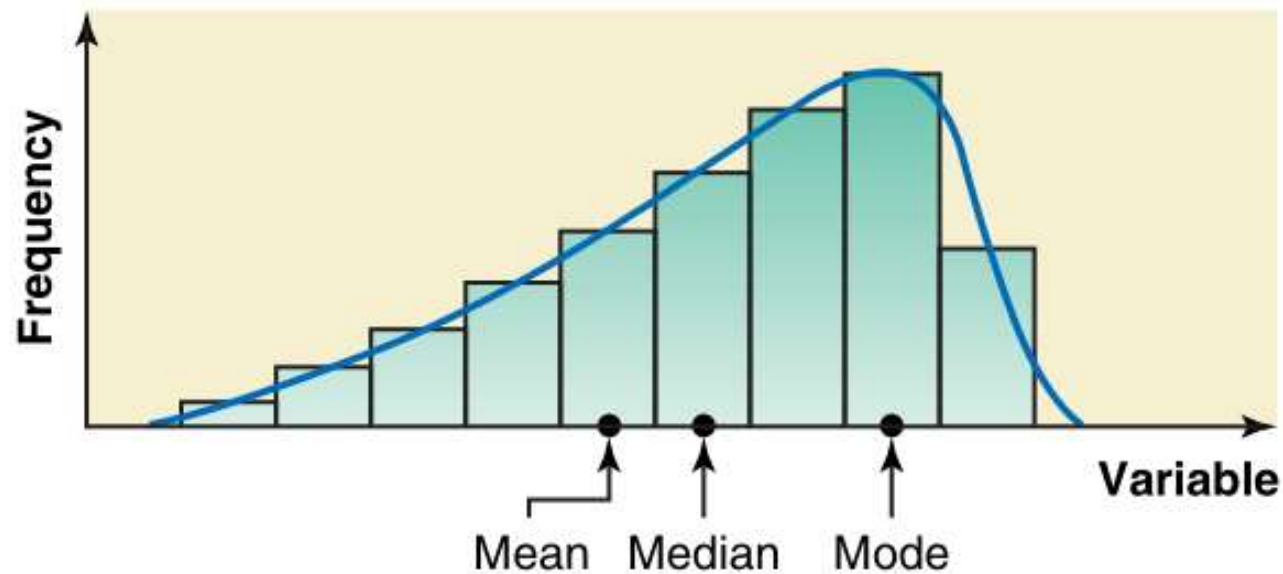
Mode < median < mean



# Relationships Between Mean, Median, and Mode

- Skewed to the left (or negatively skewed):

Mean < median < mode



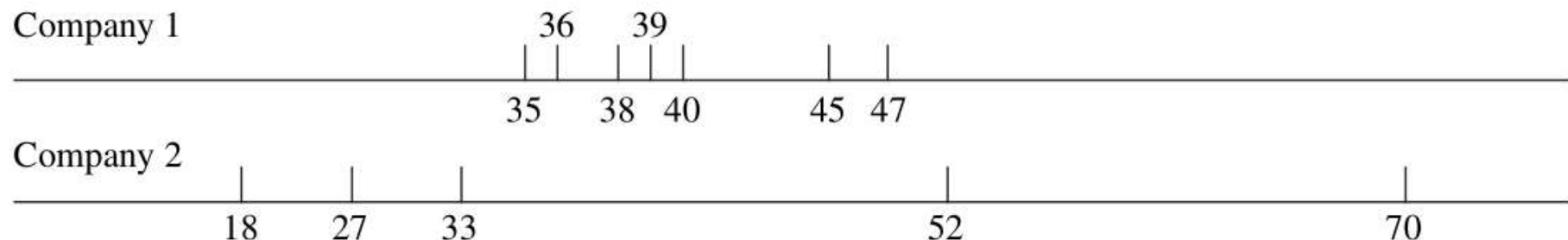
# Measures of dispersion (ungrouped data)

Range, variance, standard deviation

# Measures of dispersion

- The measures of center, such as the mean, median, and mode, do not reveal the whole picture of the distribution of a data set.
- Eg: 

Company 1:	47	38	35	40	36	45	39
Company 2:	70	33	18	52	27		



- ▣ Same mean (40) for both companies.
- ▣ But company 2 has larger variation.



# Range

- The difference between the largest and smallest data entries in the set.

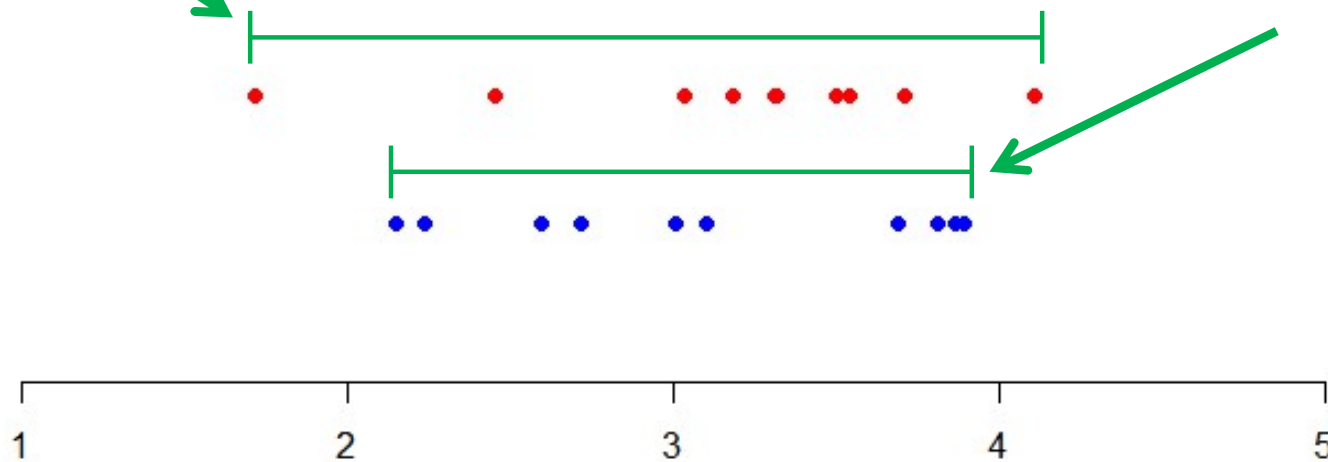
$$\text{Range} = \text{Largest value} - \text{Smallest value}$$

- The larger the range, the more spread the data is.
- If data contains outlier, range is not suitable for measuring dispersion.

# Range

Range = 2.4

Range = 1.7



# Example – finding range

## EXAMPLE 3–13 Total Areas of Four States

Table 3.4 gives the total areas in square miles of the four western South-Central states of the United States.

Table 3.4

State	Total Area (square miles)
Arkansas	53,182
Louisiana	49,651
Oklahoma	69,903
Texas	267,277

Find the range for this data set.

- Largest value = 267,277
- Smallest value = 49,651
- Range =  $267,277 - 49,651 = 217,626$

# Variance and standard deviation

- Standard deviation is the most used measure of dispersion.
- The value tells how closely the observations are to the mean.
- Standard deviation is the square root of variance
- **Population** variance and **population** standard deviation:

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

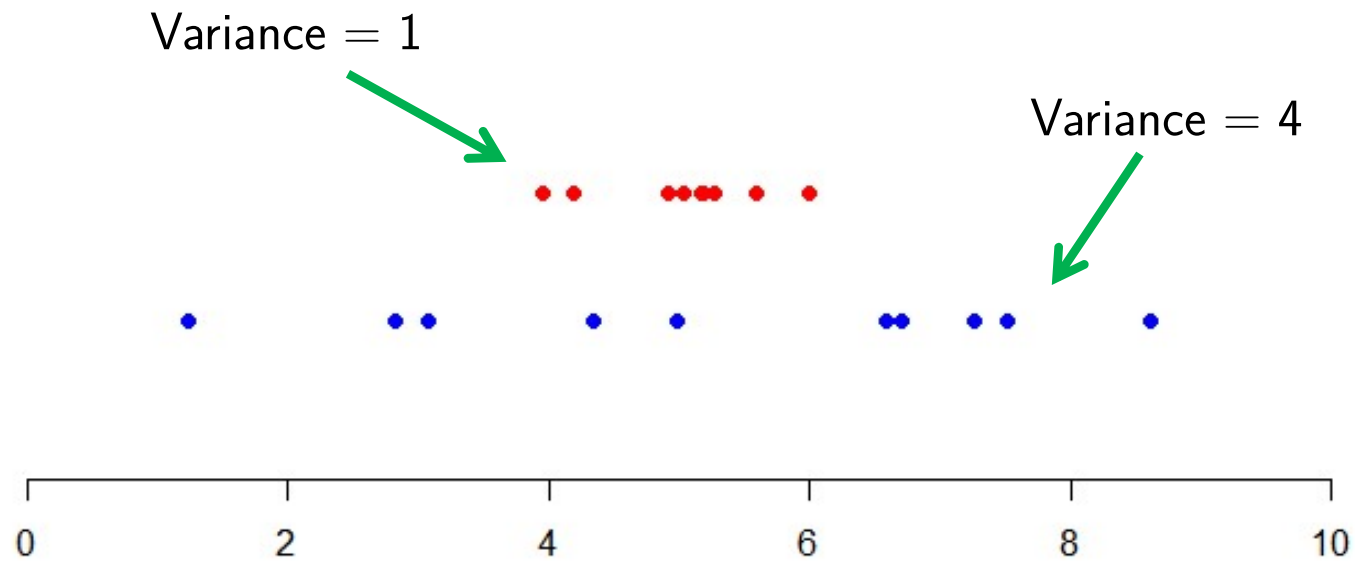
- **Sample** variance and **sample** standard deviation:

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

# Variance and standard deviation

- If observations are close to the mean, then  $x - \mu$  or  $x - \bar{x}$  is small in magnitude.
- This leads to small variance and standard deviation.
- The higher the variance or standard deviation, the more spread the observations are.

# Variance and standard deviation



□ Mean = 5 for both data

# Alternative formulas

- Alternative formulas for **population** variance and **population** standard deviation:

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

- Alternative formulas for **sample** variance and **sample** standard deviation:

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

# Important note

- $\sum x^2$  is **not the same** as  $(\sum x)^2$
- For example, let's say the observations are 1, 2, and 3.
- $\sum x^2 = 1^2 + 2^2 + 3^2 = 1 + 4 + 9 = 14$
- $(\sum x)^2 = (1 + 2 + 3)^2 = 6^2 = 36$



# Example – finding variance and standard deviation

## EXAMPLE 3–14 Compensations of Female CEOs

Refer to the 2014 compensations of 11 female CEOs of American companies given in Example 3–4. The table from that example is reproduced below.

Company & CEO	2014 Compensation (millions of dollars)
General Dynamics, Phebe Novakovic	19.3
GM, Mary Barra	16.2
Hewlett-Packard, Meg Whitman	19.6
IBM, Virginia Rometty	19.3
Lockheed Martin, Marillyn Hewson	33.7
Mondelez, Irene Rosenfeld	21.0
PepsiCo, Indra Nooyi	22.5
Sempra, Debra Reed	16.9
TJX, Carol Meyrowitz	28.7
Yahoo, Marissa Mayer	42.1
Xerox, Ursula Burns	22.2

[Sample dataset](#)

Find the variance and standard deviation for these data.

# Example – finding variance and standard deviation

$x$	$x^2$
19.3	372.49
16.2	262.44
19.6	384.16
19.3	372.49
33.7	1135.69
21.0	441
22.5	506.25
16.9	285.61
28.7	823.69
42.1	1772.41
22.2	492.84
$\sum x = 261.5$	$\sum x^2 = 6849.07$

$$n = 11$$

$$\begin{aligned}
 s^2 &= \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n - 1} \\
 &= \frac{6849.07 - \frac{(261.5)^2}{11}}{11 - 1} \\
 &= 63.250
 \end{aligned}$$

$$s = \sqrt{63.25} = 7.953$$

- The table on the left is not necessary.
- You can simply calculate using calculator
 
$$\sum x = 19.3 + 16.2 + \dots + 22.2 = 261.5$$

$$\sum x^2 = 19.3^2 + 16.2^2 + \dots + 22.2^2 = 6849.07$$
- Then apply the formula.

# Exercise

**3.32** The following data set belongs to a population:

5      -7      2      0      -9      16      10      7

$N = 8$

Calculate the range, variance, and standard deviation.

Ranked data = -9 -7 0 2 5 7 10 16

Range =  $16 - (-9) = 25$

Variance = 7.842

Standard deviation = 61.5

# Exercise

**3.47** The following data give the numbers of pieces of junk mail received by 10 families during the past month.

41      33      28      21      29      19      14      31      39      36

Find the range, variance, and standard deviation.

Range = 27

Variance =

Standard deviation =

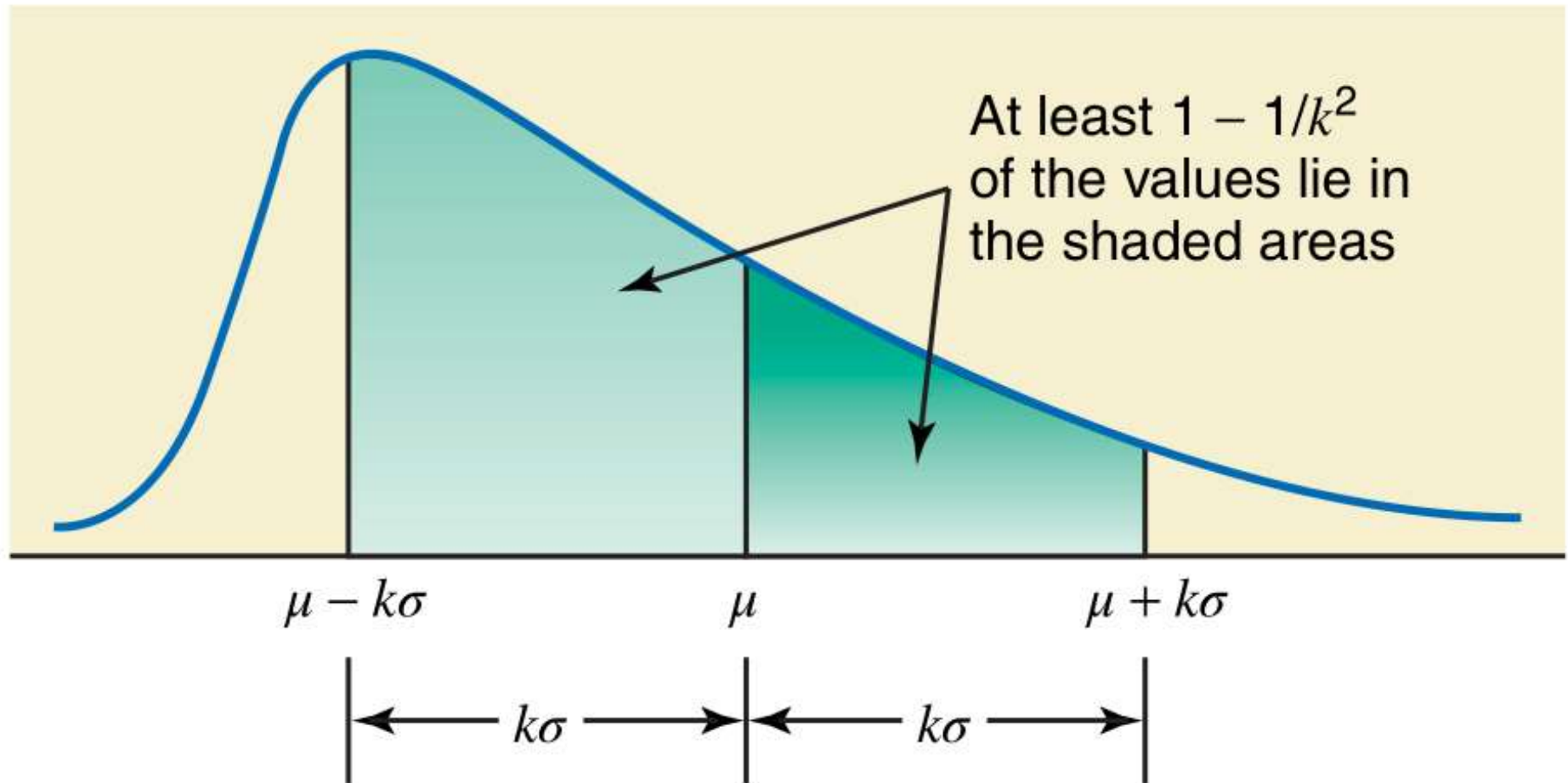
# Use of standard deviation

Chebyshev's Theorem, empirical rule

# Chebyshev's Theorem

- By using the mean and standard deviation, we can approximate the total observations that fall within a given interval.
- Chebyshev's Theorem:
  - ▣ For any number  $k$  greater than 1, at least  $100 \left(1 - \frac{1}{k^2}\right) \%$  of the data values lie within  $k$  standard deviations of the mean.
- It gives a lower bound of how much data is between the  $k$  standard deviations of the mean

# Chebyshev's Theorem



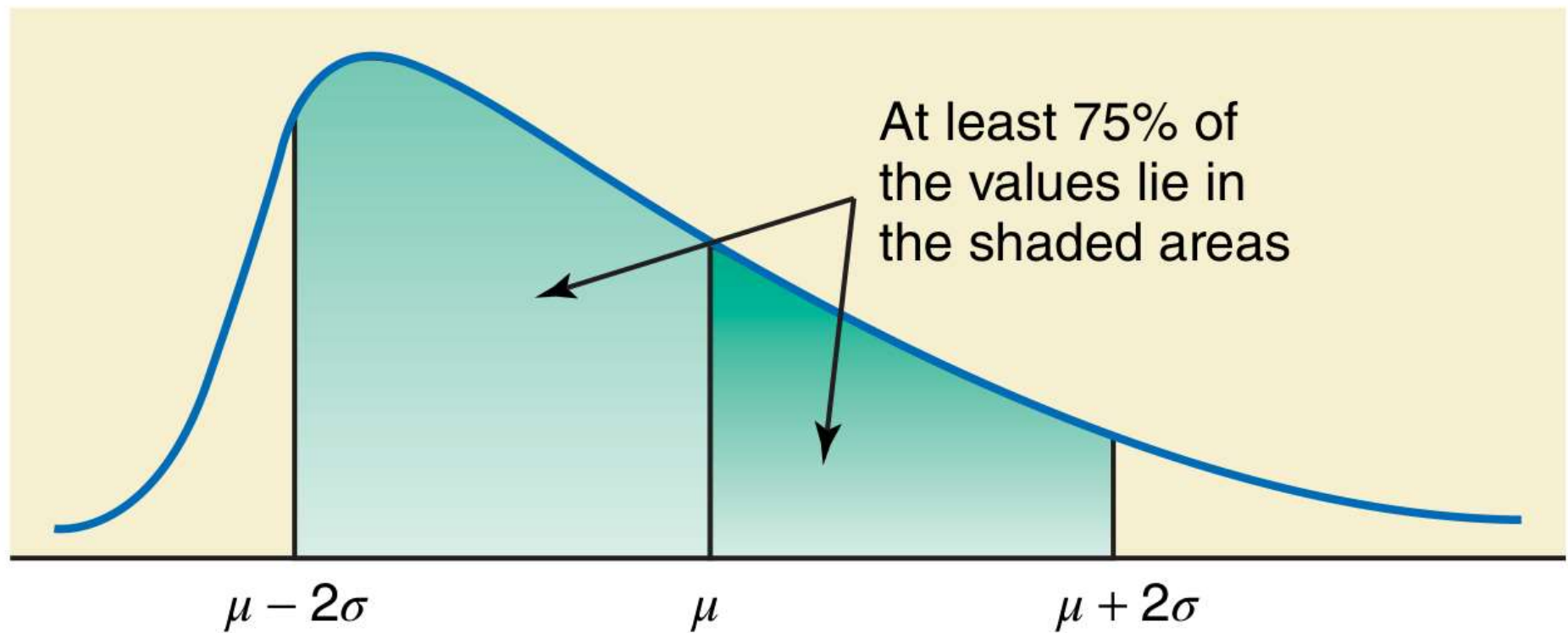
# Chebyshev's Theorem

**Table 3.14** Areas Under the Distribution Curve Using Chebyshev's Theorem

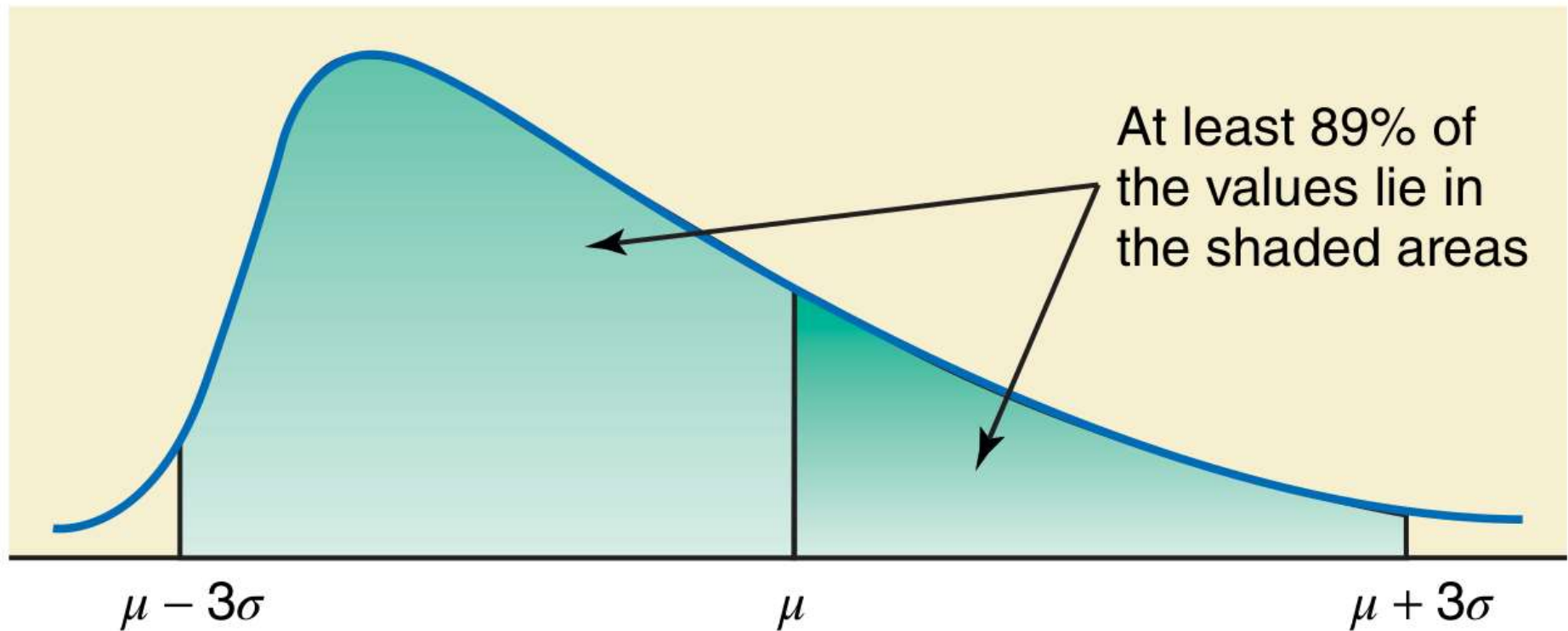
$k$	Interval	$1 - \frac{1}{k^2}$	Minimum Area Within $k$ Standard Deviations
1.5	$\mu \pm 1.5\sigma$	$1 - \frac{1}{1.5^2} = 1 - .44 = .56$	56%
2.0	$\mu \pm 2\sigma$	$1 - \frac{1}{2^2} = 1 - .25 = .75$	75%
2.5	$\mu \pm 2.5\sigma$	$1 - \frac{1}{2.5^2} = 1 - .16 = .84$	84%
3.0	$\mu \pm 3\sigma$	$1 - \frac{1}{3.0^2} = 1 - .11 = .89$	89%



# Chebyshev's Theorem



# Chebyshev's Theorem



# Example – Chebyshev's Theorem

## **EXAMPLE 3-21**    Blood Pressure of Women

The average systolic blood pressure for 4000 women who were screened for high blood pressure was found to be 187 mm Hg with a standard deviation of 22. Using Chebyshev's theorem, find the minimum percentage of women in this group who have a systolic blood pressure between 143 and 231 mm Hg.

# Example – Chebyshev's Theorem

□  $\mu = 187, \sigma = 22$

□ Determine  $k$ :

$$\begin{array}{c} \overline{\leftarrow 44 \rightarrow} \quad \overline{\leftarrow 44 \rightarrow} \\ 143 \qquad \qquad \mu = 187 \qquad \qquad 231 \end{array}$$

$$\mu - k\sigma = 143$$

$$187 - k\sigma = 143$$

$$k\sigma = 187 - 143 = 44$$

$$k = \frac{44}{22} = 2$$

□ Chebyshev's Theorem:

▣ At least  $100 \left(1 - \frac{1}{k^2}\right) = 100 \left(1 - \frac{1}{4}\right) = 75\%$  of data is between 143 mm Hg and 231 mm Hg.

# Empirical rule

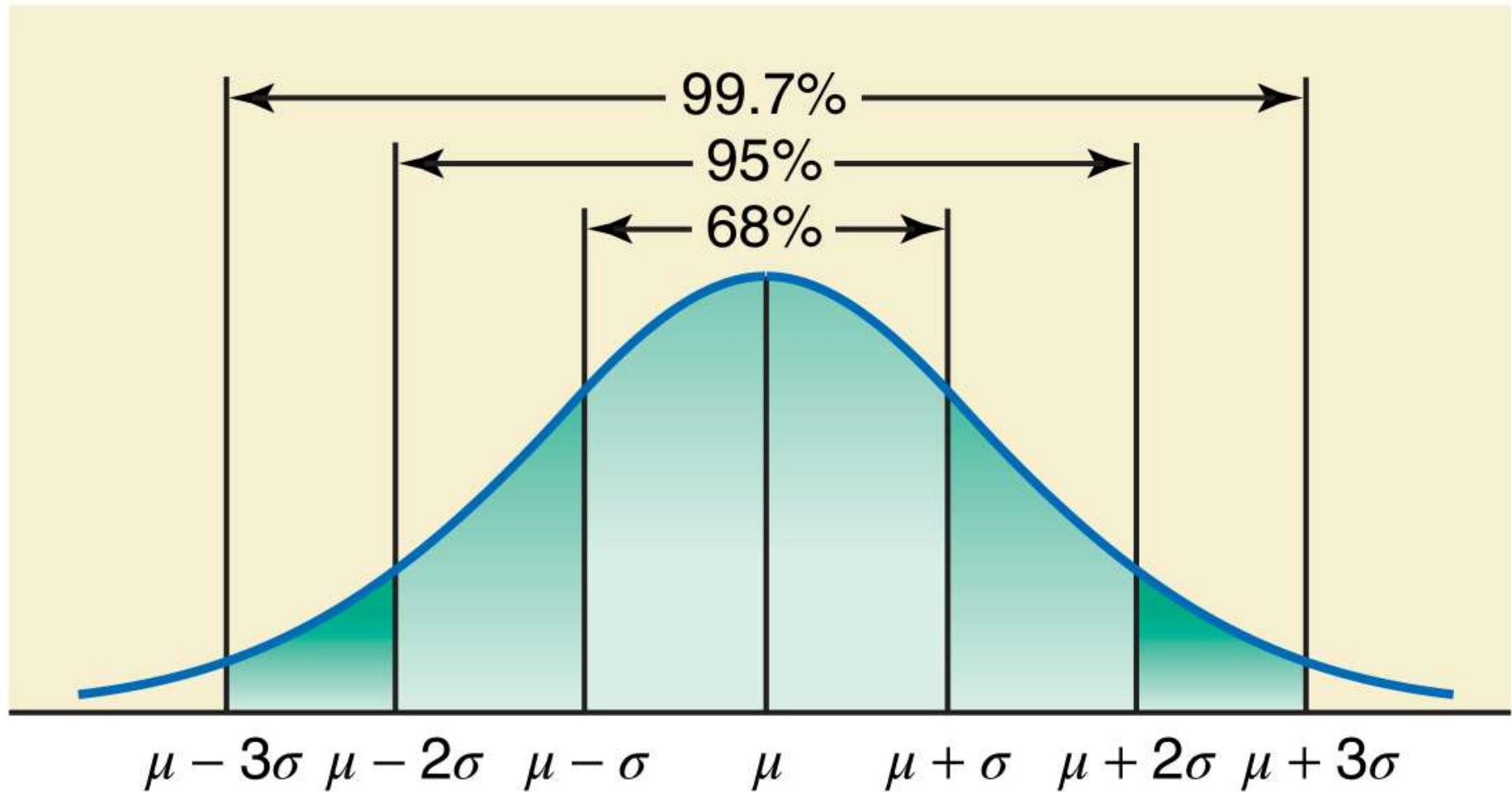
- The Chebyshev's Theorem applies to all distribution.
- But if the distribution has a **bell shape**, we can approximate the proportion of data better.
- **Empirical rule** – for a bell-shaped distribution:
  - ▣ 68% of observations lie within 1 standard deviation of the mean.
  - ▣ 95% of observations lie within 2 standard deviation of the mean.
  - ▣ 99.7% of observations lie within 3 standard deviation of the mean.

# Empirical rule

**Table 3.15**    **Approximate Areas under a Bell-Shaped Distribution Using the Empirical Rule**

Interval	Approximate Area
$\mu \pm 1 \sigma$	68%
$\mu \pm 2 \sigma$	95%
$\mu \pm 3 \sigma$	99.7%

# Empirical rule



# Example – empirical rule

## EXAMPLE 3–22 Age Distribution of Persons

The age distribution of a sample of 5000 persons is bell-shaped with a mean of 40 years and a standard deviation of 12 years. Determine the approximate percentage of people who are 16 to 64 years old.

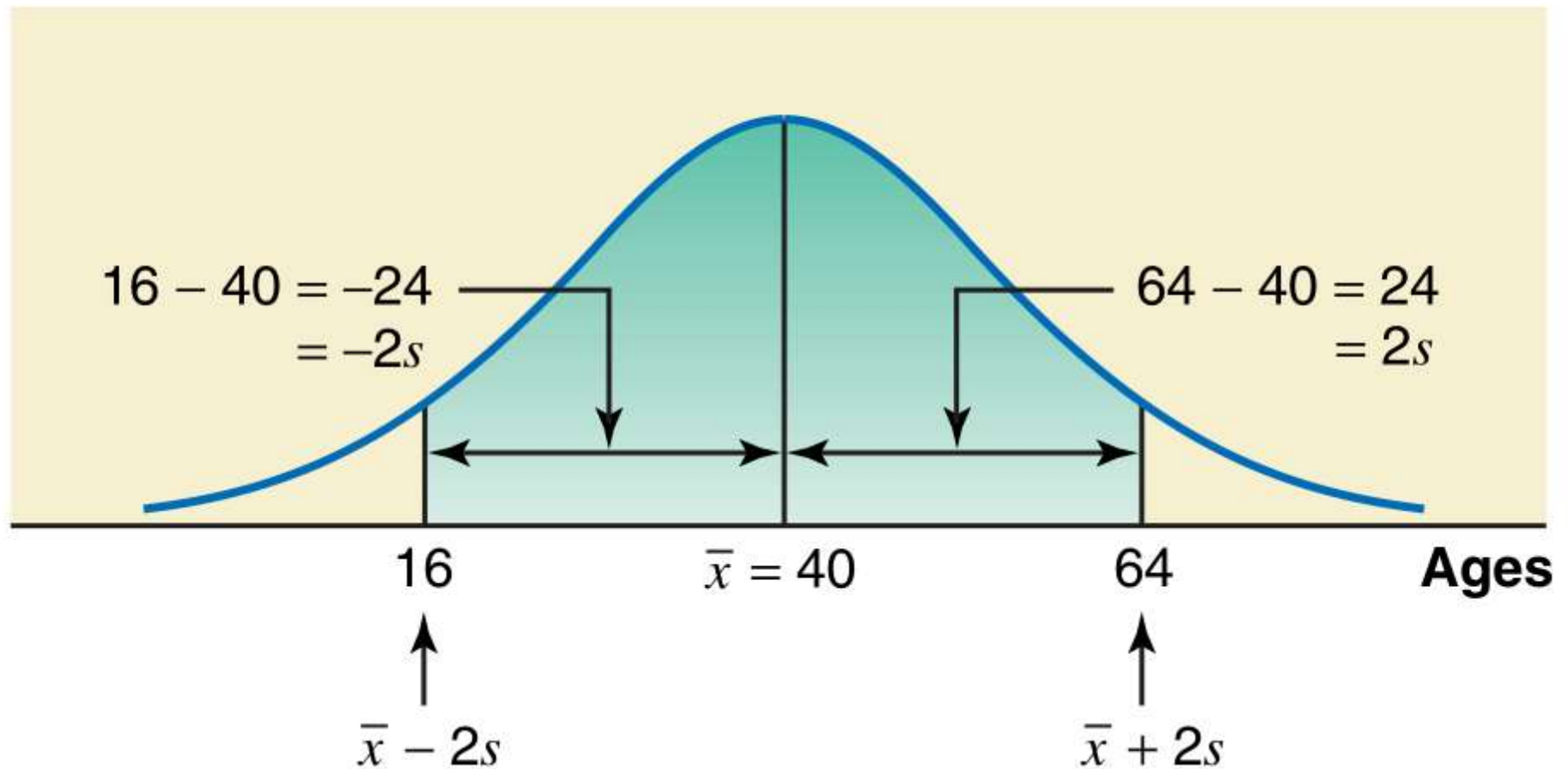
- $\bar{x} = 40, s = 12$
- Determine  $k$ :

$$\begin{aligned}\bar{x} - ks &= 16 \\ 40 - 12k &= 16 \\ k &= \frac{24}{12} = 2\end{aligned}$$

- Empirical rule:
  - ▣ Approximately 95% of the people in the sample is between 16 to 64 years old.



# Example – empirical rule



# Exercise

- 3.81** The mean monthly mortgage paid by all home owners in a town is \$2365 with a standard deviation of \$340.
- a.** Using Chebyshev's theorem, find at least what percentage of all home owners in this town pay a monthly mortgage of
    - i.** \$1685 to \$3045
    - ii.** \$1345 to \$3385
  - \*b.** Using Chebyshev's theorem, find the interval that contains the monthly mortgage payments of at least 84% of all home owners.



# Exercise

- 3.84** The prices of all college textbooks follow a bell-shaped distribution with a mean of \$105 and a standard deviation of \$20.
- a.** Using the empirical rule, find the percentage of all college textbooks with their prices between
    - i.** \$85 and \$125
    - ii.** \$65 and \$145
  - \*b.** Using the empirical rule, find the interval that contains the prices of 99.7% of college textbooks.



# Measures of position

Quartiles, interquartile range, percentiles

# Measures of position

- Determines the position of a single value in relation to other values in a sample or a population data set.
- We will cover quartiles, percentiles, and deciles here. These are known as quantiles.

# Quartile

- **Quartiles** divide the distribution into four equal groups.
- The boundaries for the groups are denoted by  $Q_1$  (first quartile),  $Q_2$  (second quartile),  $Q_3$  (third quartile).



- Eg: Approximately 75% of values in a ranked data set are more than  $Q_1$ .
- Note that the median is equal to  $Q_2$ .



# Quartile

- Finding  $Q_1$ ,  $Q_2$ , and  $Q_3$ :
  - ▣ Rank the data in increasing order.
  - ▣ Find the median ( $Q_2$ ).
  - ▣ For values that are less than the median, find the point that divides the values equally ( $Q_1$ ).
  - ▣ For values that are greater than the median, find the point that divides the values equally ( $Q_3$ ).
  
- Interquartile range (IQR):
  - ▣ Is the difference between the third quartile and the first quartile.
  - ▣  $IQR = Q_3 - Q_1$

# Example – quartile

## EXAMPLE 3–23 Commuting Times for College Students

A sample of 12 commuter students was selected from a college. The following data give the typical one-way commuting times (in minutes) from home to college for these 12 students.

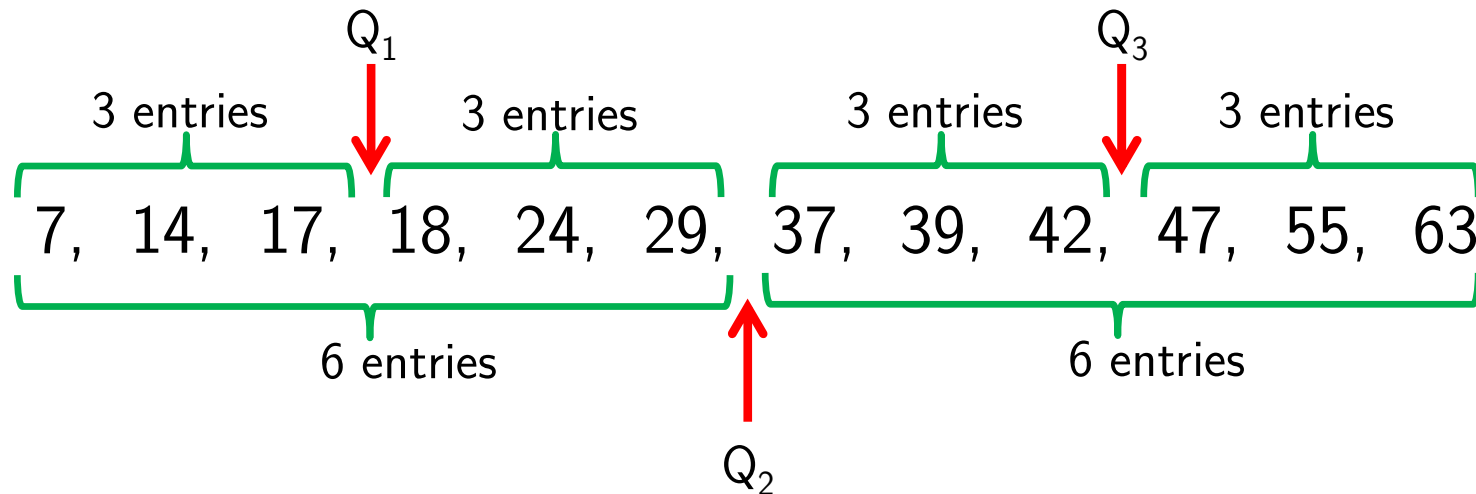
29    14    39    17    7    47    63    37    42    18    24    55

- (a) Find the values of the three quartiles.
- (b) Where does the commuting time of 47 fall in relation to the three quartiles?
- (c) Find the interquartile range.

# Example – quartile

- Find median:
  - ▣ Ranked data: 7, 14, 17, 18, 24, 29, 37, 39, 42, 47, 55, 63
  - ▣  $Q_2 = \frac{29+37}{2} = 33$
- Find  $Q_1$ :
  - ▣ Data less than  $Q_2$ : 7, 14, 17, 18, 24, 29
  - ▣  $Q_1 = \frac{17+18}{2} = 17.5$
- Find  $Q_3$ :
  - ▣ Data more than  $Q_2$ : 37, 39, 42, 47, 55, 63
  - ▣  $Q_3 = \frac{42+47}{2} = 44.5$

# Example – quartile



- 47 minutes is in the top 25%.
- $IQR = Q_3 - Q_1 = 44.5 - 17.5 = 27$  minutes.
- Interpretation:
  - $Q_1$ : 25% of the students takes less than 17.5 minutes commuting time.
  - $Q_3$ : 25% of the students takes more than 44.5 minutes commuting time.

# Exercise

**3.69** The following data give the speeds of 13 cars (in mph) measured by radar, traveling on I-84.

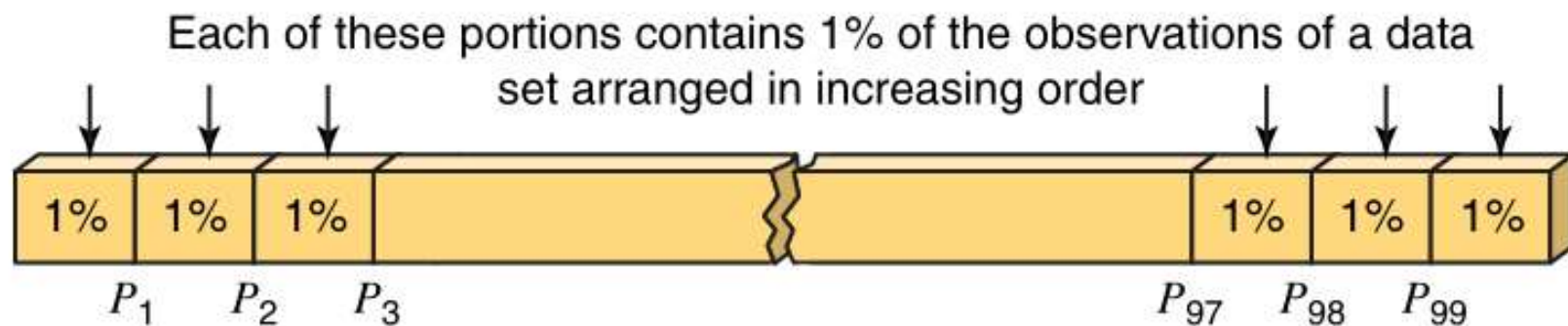
73	75	69	68	78	69	74
76	72	79	68	77	71	

**a.** Find the values of the three quartiles and the interquartile range.

Ranked data: 68 68 69 69 71 72 73 74 75 76 77 78 78  
Q2 = 73  
Q1 = 69  
Q3 = 76.5

# Percentiles

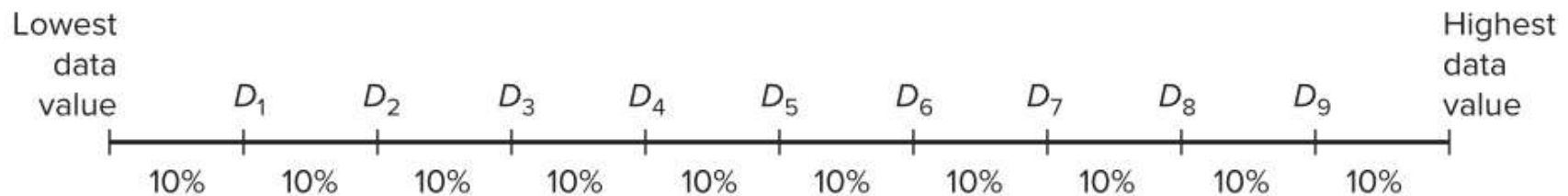
- **Percentiles** divide a ranked data set into 100 equal parts.



- The  $k$ th percentile is denoted by  $P_k$ , where  $k$  is an integer in the range 1 to 99.
- About  $k\%$  of the measurements are smaller than  $P_k$ , and about  $(100 - k)\%$  of the measurements are greater than  $P_k$ .

# Deciles

- **Deciles** divide a ranked data set into 10 equal parts.



- The  $k$ th decile is denoted by  $D_k$ , where  $k$  is an integer in the range 1 to 9.
- Note that  $D_1 = P_{10}$ ,  $D_5 = Q_2 = P_{50}$ , etc.

# Box-and-whisker plot



# Box-and-whisker plot

- Box-and-whisker plot (also called box plot or boxplot) gives a graphic presentation of data using five measures:
  - ▣ Median
  - ▣ First quartile,  $Q_1$
  - ▣ Third quartile,  $Q_3$
  - ▣ The smallest values in the data set between the lower and the upper inner fences.
  - ▣ The largest values in the data set between the lower and the upper inner fences.

# Box-and-whisker plot

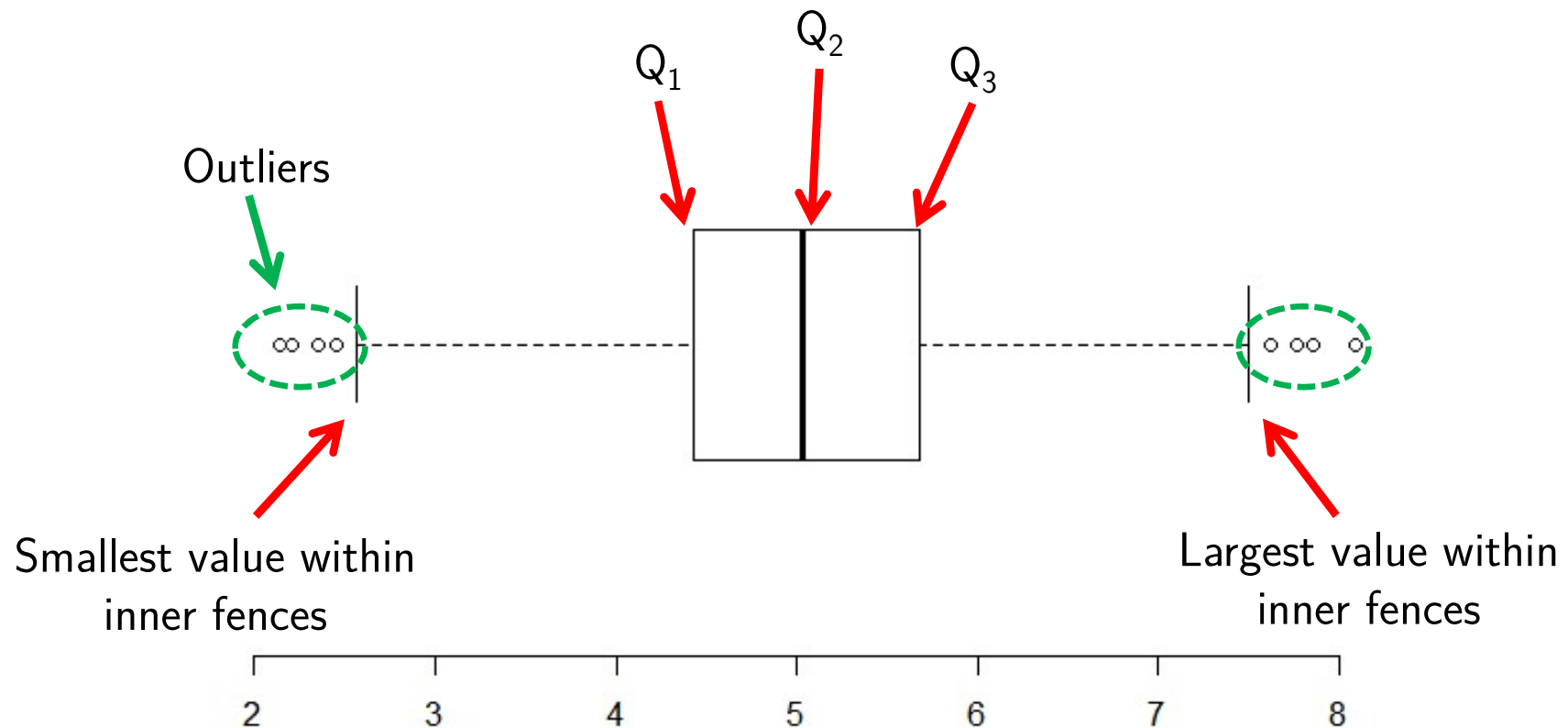
- Inner fences:

- Lower inner fence =  $Q_1 - 1.5 \times \text{IQR}$
- Upper inner fence =  $Q_3 + 1.5 \times \text{IQR}$

- **Outlier:**

- Extreme observations, either too high or too low.
- Using box plot, any observations outside the inner fences are considered outlier.

# Box-and-whisker plot



# Box-and-whisker plot

- Steps:
  - ▣ Find the values of median,  $Q_1$ , and  $Q_3$ .
  - ▣ Calculate the lower and upper inner fences.
  - ▣ Determine the smallest and largest value within the lower and upper inner fences.
  - ▣ Draw a horizontal axis and place the scale on the axis.
  - ▣ Draw a box whose vertical sides go through  $Q_1$  to  $Q_3$ .
  - ▣ Draw a vertical line through the median.
  - ▣ Draw a line from the smallest value within the fences to the left side of the box.
  - ▣ Draw a line from the largest value within the fences to the right side of the box.

# Box-and-whisker plot and skewness

- Symmetric:
  - ▣ The median is at the center of the box
  - ▣ The left and right whiskers have the same length
  
- Skewed to the right:
  - ▣ The median is at the left of the center the box
  - ▣ The right whiskers is longer compared to the left whisker
  
- Skewed to the left:
  - ▣ The median is at the right of the center the box
  - ▣ The left whiskers is longer compared to the right whisker

# Example – box plot

## EXAMPLE 3-27 Incomes of Households

The following data are the incomes (in thousands of dollars) for a sample of 12 households.

75    69    84    112    74    104    81    90    94    144    79    98

Construct a box-and-whisker plot for these data.

Ranked data:

69, 74, 75, 79, 81, 84, 90, 94, 98, 104, 112, 144

q1 = 77  
q2 = 87  
q3 = 101

lower inner fence = 41  
higher inner fence = 137

lowest within inner fence = 69  
highest within inner fence = 112



# Exploratory data analysis (EDA)



# Exploratory data analysis (EDA)

- Exploratory data analysis (EDA) is a process to analyze and summarize data using their main characteristics and data visualization methods.
- It is often done early during the analysis, before further analysis is performed.
- It can help detecting
  - ▣ Outliers and errors
  - ▣ Relationship between variables
  - ▣ Patterns and trend
- The methods on graphing data which we covered in previous topic and the numerical measures covered in this topic can be used for EDA.

# Example – EDA

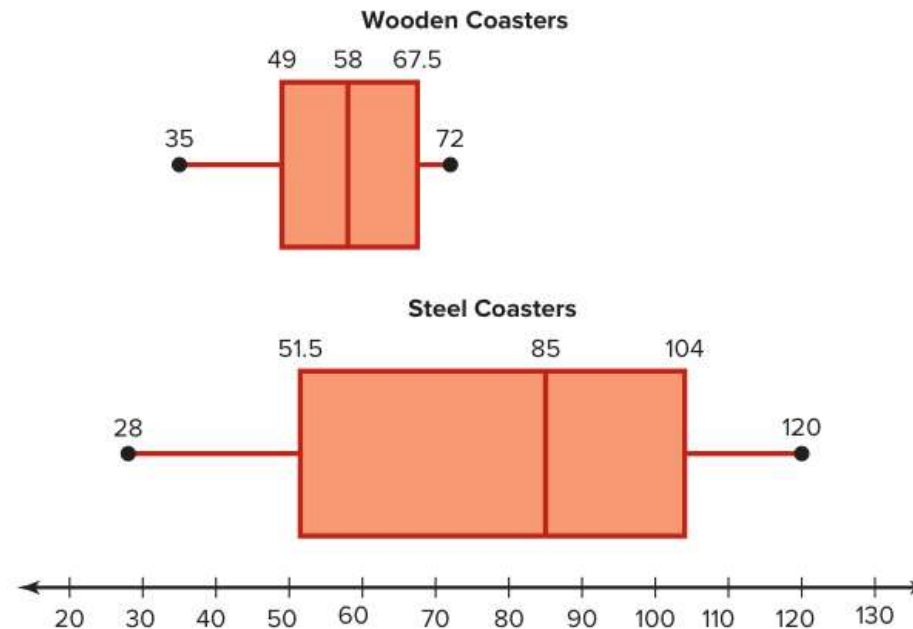
## EXAMPLE 3–38 Speeds of Roller Coasters

The data shown are the speeds in miles per hour of a sample of wooden roller coasters and a sample of steel roller coasters. Compare the distributions by using boxplots.

Wood				Steel			
50	56	60	48	55	70	48	28
35	67	72	68	100	106	102	120

*Source:* UltimateRollerCoaster.com

# Example – EDA



- ❑ Median of the speeds of the steel coasters is much higher than that of the wooden coasters.
- ❑ Interquartile range of the steel coasters is much larger than that of the wooden coasters.
- ❑ The range of the speeds of the steel coasters is larger than that of the wooden coasters.
- ❑ Both the speeds of the wooden and steel coasters are slightly skewed to the left.

# Summary

- Measure of center – mean, median, mode
- Measure of dispersion – range, variance, standard deviation
- Using standard deviation – Chebyshev's Theorem, empirical rule.
- Quartiles, percentiles, deciles.
- Box-and-whisker plot.