CHAPTER 3

Data Description

Objectives

After completing this chapter, you should be able to

1. Summarize data, using measures of central tendency, such as the mean, median, mode, and midrange.
2. Describe data, using measures of variation, such as the range, variance, and standard deviation.
3. Identify the position of a data value in a data set, using various measures of position, such as percentiles, deciles, and quartiles.
4. Use the techniques of exploratory data analysis, including boxplots and five-number summaries, to discover various aspects of data.

Outline

Introduction

3–1 Measures of Central Tendency

3–2 Measures of Variation

3–3 Measures of Position

3–4 Exploratory Data Analysis

Summary

Introduction

Traditional Statistics

- ***Average***
- ***Variation***
- ***Position***

3.1 Measures of Central Tendency

- A **statistic** is a characteristic or measure obtained by using the data values from a sample.
- A **parameter** is a characteristic or measure obtained by using all the data values for a specific population.

Measures of Central Tendency

General Rounding Rule

The basic rounding rule is that rounding should not be done until the final answer is calculated. Use of parentheses on calculators or use of spreadsheets help to avoid early rounding error.

Measures of Central Tendency

What Do We Mean By **Average**?

- Mean
- Median
- Mode
- Midrange
- Weighted Mean

Measures of Central Tendency: Mean

Result of the division

- The **mean** is the quotient of the sum of the values and the total number of values.
- The symbol $\bar{X}$ is used for sample mean.

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

- For a population, the Greek letter μ (mu) is used for the mean.

$$\mu = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\sum X}{N}$$

Chapter 3
Data Description

Section 3-1

Example 3-1

Page #106

Example 3-1: Days Off per Year

The data represent the number of days off per year for a sample of individuals selected from nine different countries. Find the mean.

20, 26, 40, 36, 23, 42, 35, 24, 30 n = 9

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

$$\bar{X} = \frac{20 + 26 + 40 + 36 + 23 + 42 + 35 + 24 + 30}{9} = \frac{276}{9} = 30.7$$

The mean number of days off is 30.7 days / years.

Rounding Rule: Mean

The mean should be rounded to one more decimal place than occurs in the raw data.

The mean, in most cases, is not an actual data value.

Measures of Central Tendency: Mean for Grouped Data

- The mean for grouped data is calculated by multiplying the frequencies and midpoints of the classes.

$$\overline{X} = \frac{\sum f \cdot X_m}{n}$$

Chapter 3
Data Description

Section 3-1

Example 3-3

Page #107

Example 3-3: Miles Run

Below is a frequency distribution of miles run per week. Find the mean.

Class Boundaries	Frequency
5.5 - 10.5	1
10.5 - 15.5	2
15.5 - 20.5	3
20.5 - 25.5	5
25.5 - 30.5	4
30.5 - 35.5	3
35.5 - 40.5	2
	$\Sigma f = 20$

Example 3-3: Miles Run

Class Boundaries	Frequency, f	Midpoint, X_m	$f \cdot X_m$
5.5 - 10.5	1	8	8
10.5 - 15.5	2	13	26
15.5 - 20.5	3	18	54
20.5 - 25.5	5	23	115
25.5 - 30.5	4	28	112
30.5 - 35.5	3	33	99
35.5 - 40.5	2	38	76
	$\Sigma f = 20$		$\Sigma f \cdot X_m = 490$

$$\bar{X} = \frac{\sum f \cdot X_m}{n} = \frac{490}{20} = \boxed{24.5 \text{ miles}}$$

Measures of Central Tendency: Median

- The **median** is the midpoint of the data array. The symbol for the median is MD.
- The median will be one of the data values if there is an odd number of values.
- The median will be the average of two data values if there is an even number of values.

Chapter 3
Data Description

Section 3-1

Example 3-4

Page #110

Example 3-4: Hotel Rooms

The number of rooms in the seven hotels in downtown Pittsburgh is 713, 300, 618, 595, 311, 401, and 292. Find the median.

Sort in ascending order. (data array)

292, 300, 311, 401, 596, 618, 713

Select the middle value.

MD = 401

The median is 401 rooms.
(one of the data values)

Chapter 3
Data Description

Section 3-1

Example 3-6

Page #111

Example 3-6: Tornadoes in the U.S.

The number of tornadoes that have occurred in the United States over an 8-year period follows. Find the median.

684, 764, 656, 702, 856, 1133, 1132, 1303

Find the average of the two middle values.

656, 684, 702, 764, 856, 1132, 1133, 1303 (data array)

$$\text{MD} = \frac{764 + 856}{2} = \frac{1620}{2} = 810$$

The median number of tornadoes is 810.

Measures of Central Tendency: Mode

- The **mode** is the value that occurs most often in a data set.
- It is sometimes said to be the most typical case.
- There may be no mode, one mode (unimodal), two modes (bimodal), or many modes (multimodal).

Chapter 3
Data Description

Section 3-1

Example 3-9

Page #111

Example 3-9: NFL Signing Bonuses

Find the mode of the signing bonuses of eight NFL players for a specific year. The bonuses in millions of dollars are

18.0, 14.0, 34.5, 10, 11.3, 10, 12.4, 10

You may find it easier to sort first.

10, 10, 10, 11.3, 12.4, 14.0, 18.0, 34.5 (data array)

↑ ↑ ↑

Select the value that occurs the most.

The mode is 10 million dollars.

Chapter 3
Data Description

Section 3-1

Example 3-10

Page #111

Example 3-10: Coal Employees in PA

Find the mode for the number of coal employees per county for 10 selected counties in southwestern Pennsylvania.

110, 731, 1031, 84, 20, 118, 1162, 1977, 103, 752 (data array?)

No value occurs more than once.

There is no mode.
Which is different than mode = 0.

Chapter 3
Data Description

Section 3-1

Example 3-11

Page #111

Example 3-11: Licensed Nuclear Reactors

The data show the number of licensed nuclear reactors in the United States for a recent 15-year period. Find the mode.

104 104 104 104 104 107 109 109 109 110

109 111 112 111 109

104 (5 times) and 109 (5 times) both occur the most. The data set is said to be bimodal.

The modes are 104 and 109.

Chapter 3
Data Description

Section 3-1

Example 3-12

Page #111

Example 3-12: Miles Run per Week

Find the modal class for the frequency distribution of miles that 20 runners ran in one week.

Class boundaries	Frequency
5.5 – 10.5	1
10.5 – 15.5	2
15.5 – 20.5	3
20.5 – 25.5	5
25.5 – 30.5	4
30.5 – 35.5	3
35.5 – 40.5	2

The modal class is 20.5 – 25.5.

The mode, the midpoint of the modal class, is 23 miles per week.

Measures of Central Tendency: Midrange

- The **midrange** is the average of the lowest and highest values in a data set.

$$MR = \frac{Lowest + Highest}{2}$$

Chapter 3
Data Description

Section 3-1

Example 3-15

Page #114

Example 3-15: Water-Line Breaks

In the last two winter seasons, the city of Brownsville, Minnesota, reported these numbers of water-line breaks per month. Find the midrange.

2, 3, 6, 8, 4, 1 (data array?)

$$\text{MR} = \frac{1+8}{2} = \frac{9}{2} = 4.5$$

The midrange is 4.5.

Measures of Central Tendency: Weighted Mean

- Find the **weighted mean** of a variable by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights.

$$\bar{X} = \frac{w_1 X_1 + w_2 X_2 + \cdots + w_n X_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum wX}{\sum w}$$

Chapter 3
Data Description

Section 3-1

Example 3-17

Page #115

Example 3-17: Grade Point Average

A student received the following grades. Find the corresponding GPA.

Course	Credits, *w*	Grade, *X*
English Composition	3	A (4 points)
Introduction to Psychology	3	C (2 points)
Biology	4	B (3 points)
Physical Education	2	D (1 point)

$$\bar{X} = \frac{\sum wX}{\sum w} = \frac{3\cdot 4 + 3\cdot 2 + 4\cdot 3 + 2\cdot 1}{3+3+4+2} = \frac{32}{12} = 2.7$$

The grade point average is 2.7.

Properties of the Mean

- Uses all data values.
- Varies less than the median or mode
- Used in computing other statistics, such as the variance
- Unique, usually not one of the data values
- Cannot be used with open-ended classes
- Affected by extremely high or low values, called outliers

Properties of the Median

- Gives the midpoint
- Used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution.
- Can be used for an open-ended distribution.
- Affected less than the mean by extremely high or extremely low values. (Therefore, it is more appropriate measure for central tendency with extreme value in data than mean)

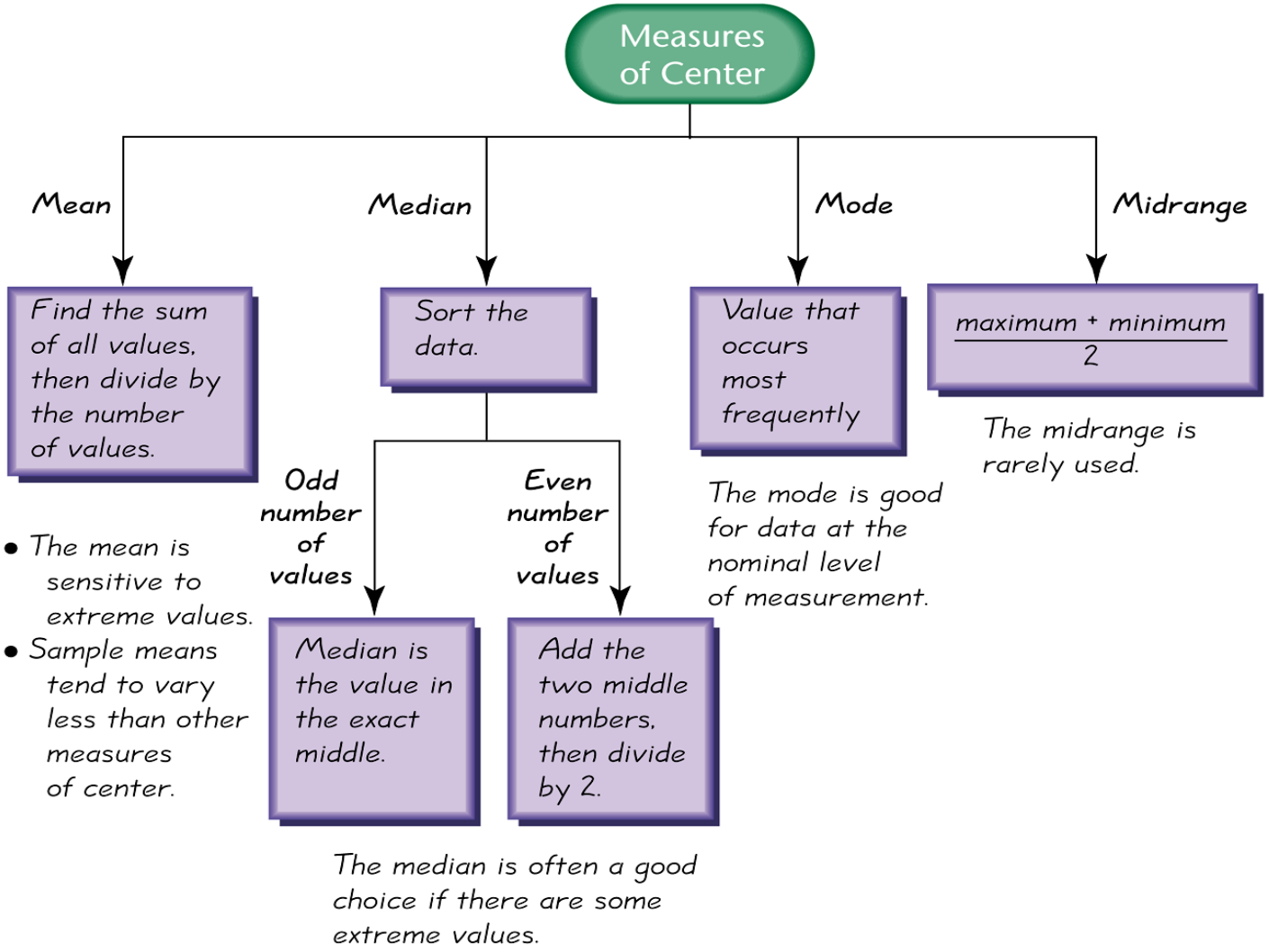
Properties of the Mode

- Used when the most typical case is desired
- Easiest average to compute
- Can be used with nominal data
- Not always unique or may not exist

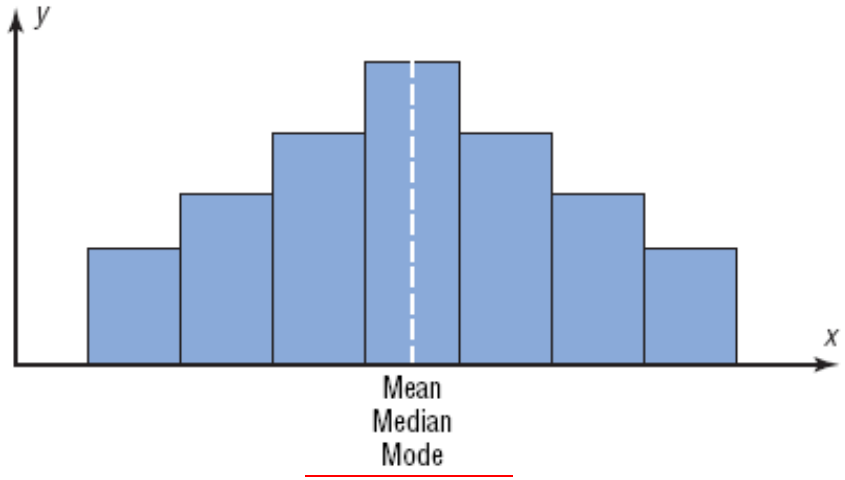
Properties of the Midrange

- Easy to compute.
- Gives the midpoint.
- Affected by extremely high or low values in a data set

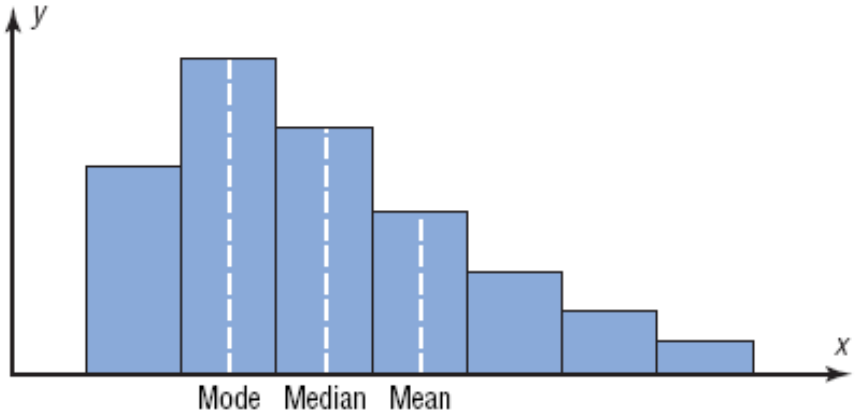
Best Measure of Center



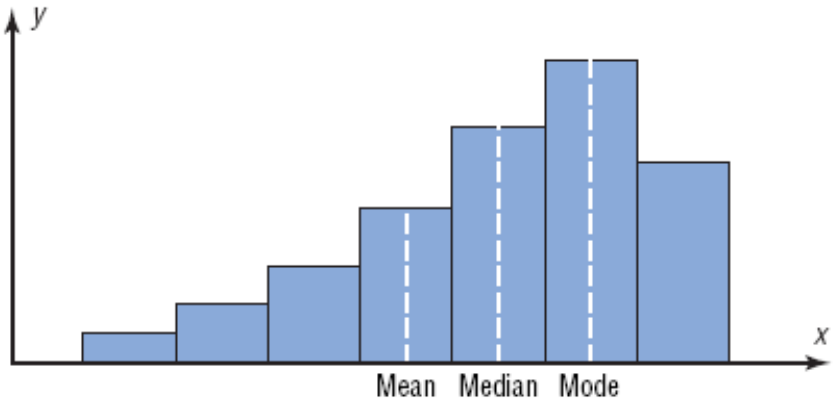
Distributions



Normal distribution



(a) Positively skewed or right-skewed



Negative skewed or left skewed

3-2 Measures of Variation

How Can We Measure Variability?

- Range
- Variance
- Standard Deviation
- Coefficient of Variation
- Chebyshev's Theorem
- Empirical Rule (Normal)

Measures of Variation: Range

- The **range** is the difference between the highest and lowest values in a data set.

$$R = Highest - Lowest$$

Chapter 3
Data Description

Section 3-2

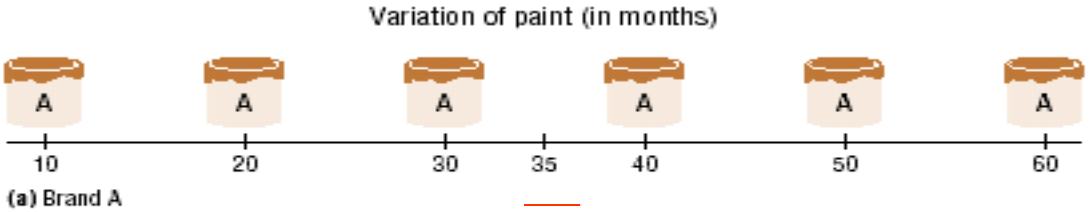
Example 3-18/19

Page #123

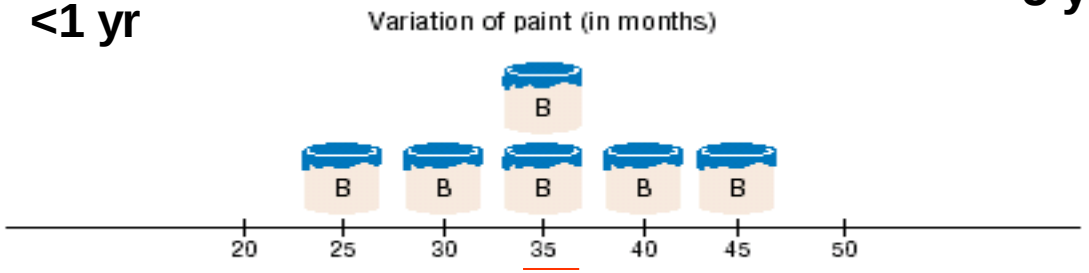
Example 3-18/19: Outdoor Paint

Two experimental brands of outdoor paint are tested to see how long each will last before fading. Six cans of each brand constitute a small population. The results (in months) are shown. Find the mean and range of each group.

Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25



<1 yr **5 yrs**



~2 yrs **<4 yrs**

Example 3-18: Outdoor Paint

Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25

Brand A:

$$\mu = \frac{\sum X}{N} = \frac{210}{6} = \boxed{35}$$

$$R = 60 - 10 = \boxed{50}$$

Brand B:

$$\mu = \frac{\sum X}{N} = \frac{210}{6} = \boxed{35}$$

$$R = 45 - 25 = \boxed{20}$$

The average for both brands is the same, but the range for Brand A is much greater than the range for Brand B.

Which brand would you buy?

Measures of Variation: Variance & Standard Deviation

- The **variance** is the average of the squares of the distance each value is from the mean.
- The **standard deviation** is the square root of the variance.
- The standard deviation is a measure of how spread out your data are.

•Uses of the Variance and Standard Deviation

- To determine the spread of the data.
- To determine the consistency of a variable.
- To determine the number of data values that fall within a specified interval in a distribution (Chebyshev's Theorem).
- Used in inferential statistics.

Measures of Variation: Variance & Standard Deviation (Population Theoretical Model)

- The **population variance** is

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

- The **population standard deviation** is

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

Chapter 3
Data Description

Section 3-2

Example 3-21

Page #125

Example 3-21: Outdoor Paint

Find the variance and standard deviation for the data set for Brand A paint. 10, 60, 50, 30, 40, 20

Months, X	μ	$X - \mu$	$(X - \mu)^2$
10	35	-25	625
60	35	25	625
50	35	15	225
30	35	-5	25
40	35	5	25
20	35	-15	225
			1750

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

$$= \frac{1750}{6}$$

$$= \boxed{291.7}$$

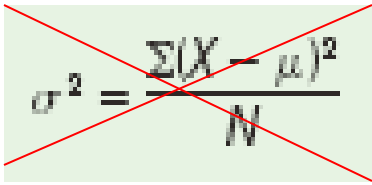
$$\sigma = \sqrt{\frac{1750}{6}}$$

$$= \boxed{17.1}$$

Measures of Variation: Variance & Standard Deviation (Sample Theoretical Model)

- The **sample variance** is

$$s^2 = \frac{\sum (X - \bar{X})^2}{n-1}$$



- The **sample standard deviation** is

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

Unbiased estimator: when the population is large and the sample is small (usually less than 30), the variance computed by this formula usually underestimates the population variance. Therefore, instead of dividing by n, find the variance of the sample by dividing by $n - 1$, giving a slightly larger value and an *unbiased* estimate of the population variance.

Bluman, Ch

Measures of Variation: Variance & Standard Deviation (Sample Computational Model)

- Is mathematically equivalent to the theoretical formula.
- Saves time when calculating by hand
- Does not use the mean
- Is more accurate when the mean has been rounded.

Measures of Variation: Variance & Standard Deviation (Sample Computational Model)

- The **sample variance** is

$$s^2 = \frac{n\sum X^2 - \left(\sum X\right)^2}{n(n-1)}.$$

Vs. Sample theoretical model

$$s^2 = \frac{\sum\left(X - \bar{X}\right)^2}{n-1}$$

- The **sample standard deviation** is

$$s = \sqrt{s^2}$$

Chapter 3
Data Description

Section 3-2

Example 3-23

Page #129

Example 3-23: European Auto Sales

Find the variance and standard deviation for the amount of European auto sales for a sample of 6 years. The data are in millions of dollars.

11.2, 11.9, 12.0, 12.8, 13.4, 14.3

	X	X^2
	11.2	125.44
	11.9	141.61
	12.0	144.00
	12.8	163.84
	13.4	179.56
	14.3	204.49
Σ	75.6	958.94

$$s^2 = \frac{n\sum X^2 - \left(\sum X\right)^2}{n(n-1)}$$

$$s^2 = \frac{6(958.94) - (75.6)^2}{6(5)}$$

$$s^2 = \left(6 \cdot 958.94 - 75.6^2\right) / (6 \cdot 5)$$

$$s^2 = 1.28$$

$$s = 1.13$$

Measures of Variation: Coefficient of Variation

The **coefficient of variation** is the standard deviation divided by the mean, expressed as a percentage.

For samples,	For populations,
$CVar = \frac{s}{\bar{X}} \cdot 100\%$	$CVar = \frac{\sigma}{\mu} \cdot 100\%$

Use *CVar* to compare standard deviations when the units are different.

Chapter 3
Data Description

Section 3-2

Example 3-25

Page #132

Example 3-25: Sales of Automobiles

The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is $5225, and the standard deviation is $773. Compare the variations of the two.

$$CVAR = \frac{s}{\bar{X}} \cdot 100\%$$

$$CVar = \frac{5}{87} \cdot 100\% = 5.7\% \qquad \text{Sales}$$

$$CVar = \frac{773}{5225} \cdot 100\% = 14.8\% \qquad \text{Commissions}$$

Commissions are more variable than sales.

Measures of Variation: Range Rule of Thumb

The **Range Rule of Thumb** approximates the standard deviation as

$$s \approx \frac{Range}{4}$$

when the distribution is unimodal and approximately symmetric.

Measures of Variation: Range Rule of Thumb

Use $\bar{X} - 2s$ to approximate the lowest value and $\bar{X} + 2s$ to approximate the highest value in a data set.

Example: $\bar{X} = 10,\ Range = 12$

$$s \approx \frac{12}{4} = 3$$

$$LOW \approx 10 - 2(3) = \boxed{4}$$

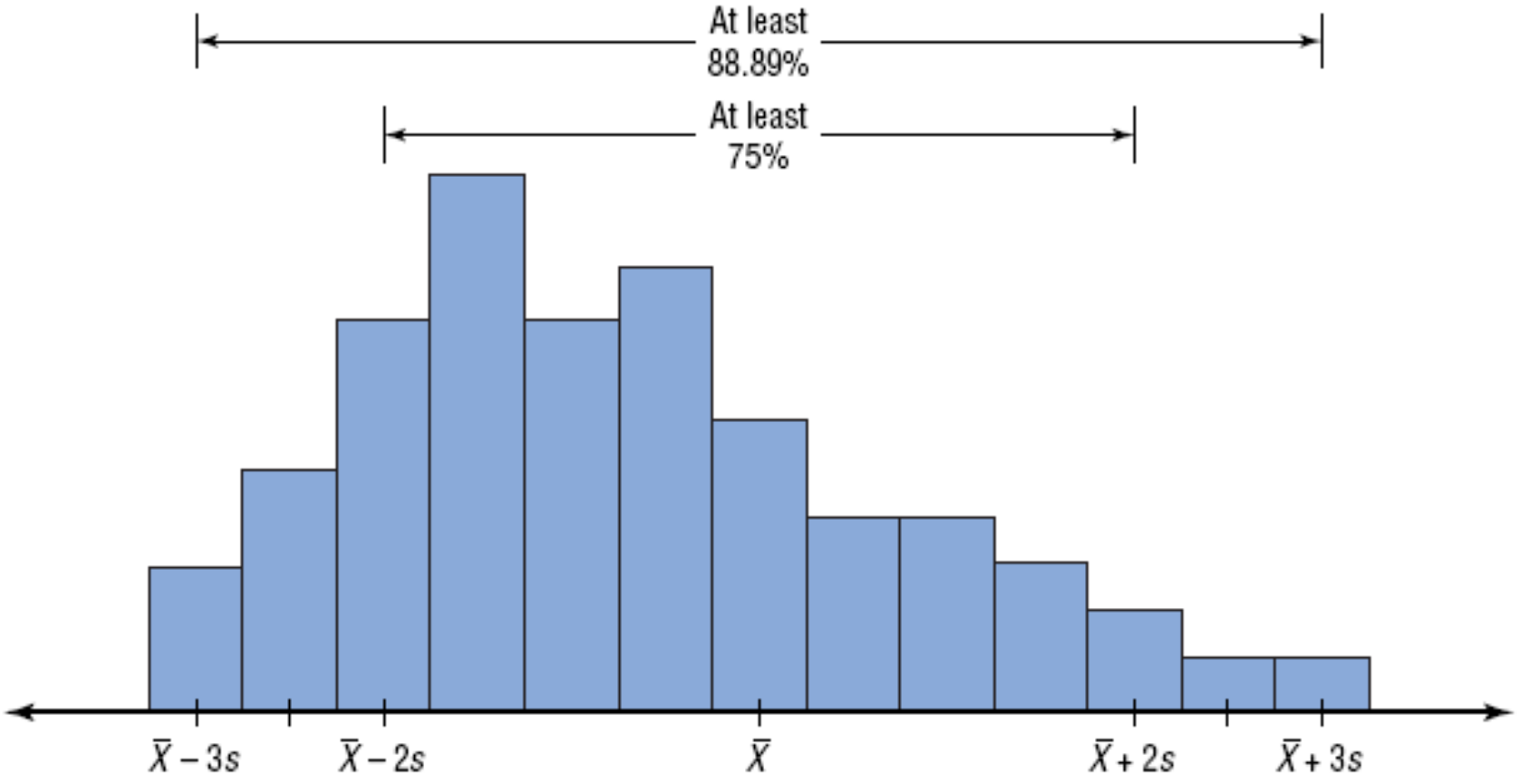
$$HIGH \approx 10 + 2(3) = \boxed{16}$$

Measures of Variation: Chebyshev's Theorem

The proportion of values from any data set that fall within k standard deviations of the mean will be at least $1-1/k^2$, where k is a number greater than 1 (k is not necessarily an integer).

# of standard deviations, k	Minimum Proportion within k standard deviations	Minimum Percentage within k standard deviations
2	$1-1/2^2=3/4$	75%
3	$1-1/3^2=8/9$	88.89%
4	$1-1/4^2=15/16$	93.75%

Measures of Variation: Chebyshev's Theorem (Any distribution)



Chapter 3
Data Description

Section 3-2

Example 3-27

Page #135

Example 3-27: Prices of Homes

The mean price of houses in a certain neighborhood is $50,000, and the standard deviation is $10,000. Find the price range for which at least 75% of the houses will sell.

Chebyshev's Theorem states that at least 75% of a data set will fall within 2 standard deviations of the mean.

$50,000 – 2(10,000) = 30,000$

$50,000 + 2(10,000) = 70,000$

At least 75% of all homes sold in the area will have a price range from $30,000 and $70,000.

Chapter 3
Data Description

Section 3-2

Example 3-28

Page #135

Example 3-28: Travel Allowances

A survey of local companies found that the mean amount of travel allowance for executives was \$0.25 per mile. The standard deviation was 0.02. Using Chebyshev's theorem, find the minimum percentage of the data values that will fall between \$0.20 and \$0.30.

$$(.30-.25)/.02=2.5$$
$$(.25-.20)/.02=2.5$$
$$k=2.5$$

$$1-1/k^2=1-1/2.5^2$$
$$=0.84$$

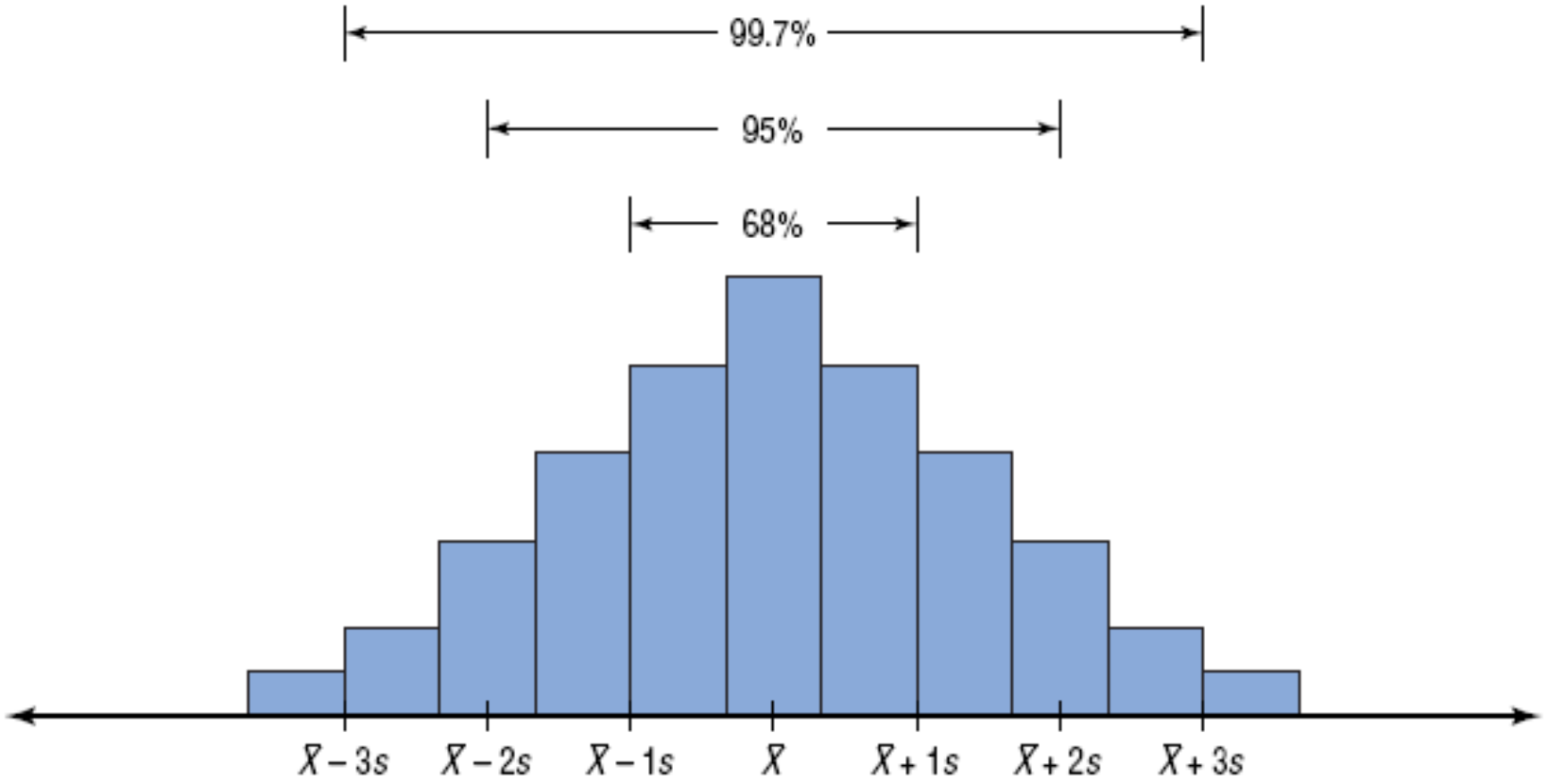
At least 84% of the data values will fall between \$0.20 and \$0.30.

Measures of Variation: Empirical Rule (Normal)

The percentage of values from a data set that fall within k standard deviations of the mean in a normal (bell-shaped) distribution is listed below.

# of standard deviations, k	Proportion within k standard deviations	Vs. Chebyshev's Theorem
1	68%	
2	95%	75%
3	99.7%	88.89%

Measures of Variation: Empirical Rule (Normal)



3-3 Measures of Position

- Z-score
- Percentile
- Quartile
- Outlier

Measures of Position: Z-score

- A **z-score** or **standard score** for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation.

Sample: $$z = \frac{X - \bar{X}}{s}$$ Population: $$z = \frac{X - \mu}{\sigma}$$

- A z-score represents the number of standard deviations a value is above or below the mean.

Chapter 3
Data Description

Section 3-3

Example 3-29

Page #142

Example 3-29: Test Scores

A student scored 65 on a calculus test that had a mean of 50 and a standard deviation of 10; she scored 30 on a history test with a mean of 25 and a standard deviation of 5. Compare her relative positions on the two tests.

$$z = \frac{X - \bar{X}}{s} = \frac{65 - 50}{10} = 1.5 \quad \text{Calculus}$$

1.5 sd above mean

$$z = \frac{X - \bar{X}}{s} = \frac{30 - 25}{5} = 1.0 \quad \text{History}$$

1.0 sd above mean

She has a higher relative position in the Calculus class.

Measures of Position: Percentiles

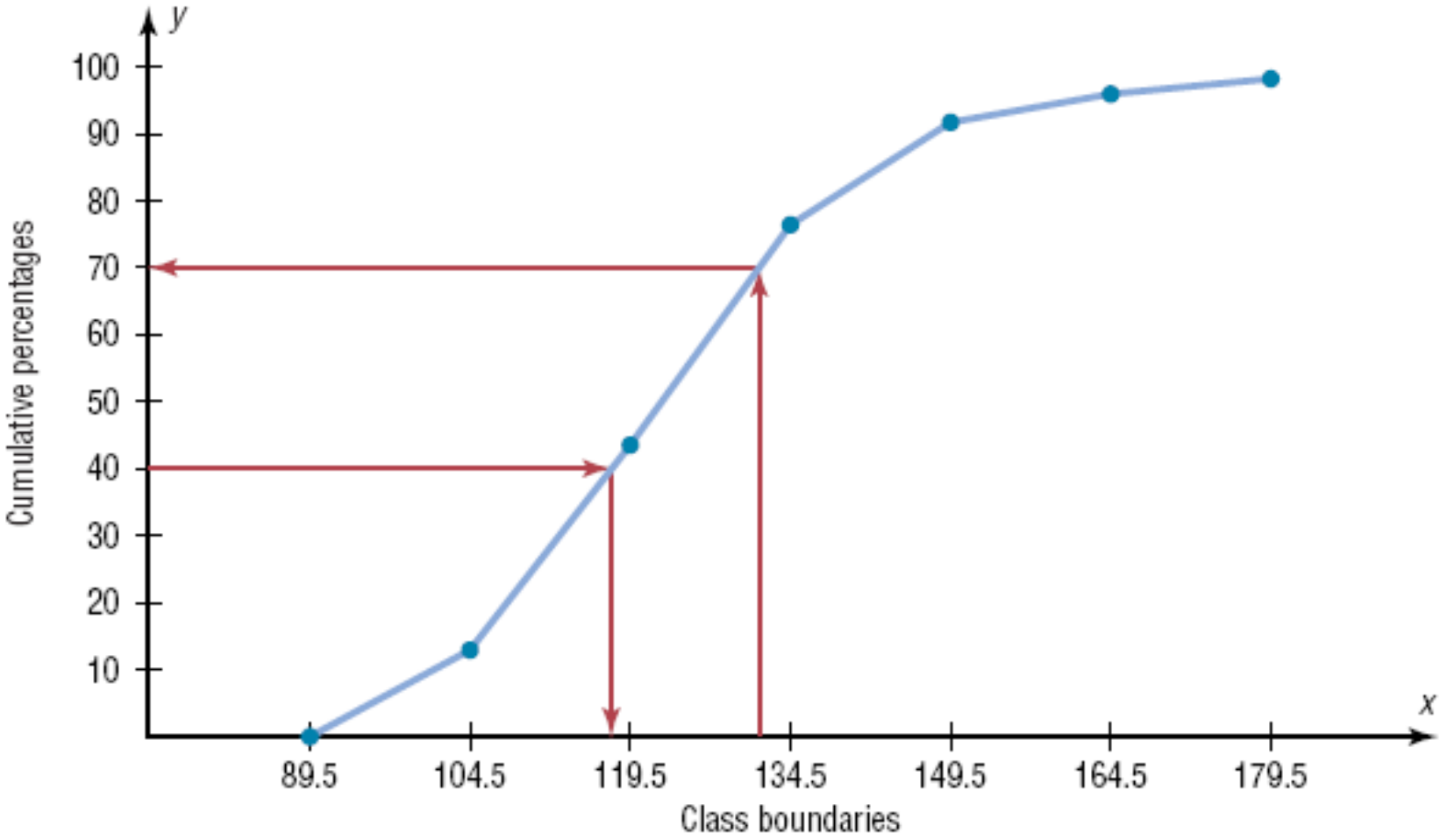
- **Percentiles** separate the data set into 100 equal groups.
- A percentile rank for a datum represents the percentage of data values below the datum.

$$Percentile_{(p)} = \frac{\left(\#\text{ of values below } X\right) + 0.5}{\text{total \# of values}_{(n)}} \cdot 100\%$$

$$c = \frac{n \cdot p}{100}$$

c is the position of the data value in the given data set.

Measures of Position: Example of a Percentile Graph



Chapter 3
Data Description

Section 3-3

Example 3-32

Page #147

Example 3-32: Test Scores

A teacher gives a 20-point test to 10 students. Find the percentile rank of a score of 12.

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

Sort in ascending order.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20 (data array), n = 10

6 values

$$Percentile = \frac{(\#\text{ of values below } X) + 0.5}{\text{total \# of values}} \cdot 100\%$$

$$= \frac{6 + 0.5}{10} \cdot 100\%$$

$$= 65\%$$

A student whose score was 12 did better than 65% of the class.

Chapter 3
Data Description

Section 3-3

Example 3-34

Page #148

Example 3-34: Test Scores

A teacher gives a 20-point test to 10 students. Find the value corresponding to the 25th percentile.

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

Sort in ascending order.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20 (data array), n = 10

3th value 6th value 7th value

$$c = \frac{n \cdot p}{100} = \frac{10 \cdot 25}{100} = 2.5 \approx 3$$

c is the position of the data value in the given data set.
If c is not a whole number, round it up to the next whole number.

If c = a whole number, use the half way between the cth and (c+1)st values: e.g. c = 6, (10 + 12) / 2 = 11

The value 5 (the third number) corresponds to the 25th percentile.

Measures of Position: Quartiles and Deciles

- **Deciles** separate the data set into 10 equal groups. $D_1=P_{10}$, $D_4=P_{40}$
- **Quartiles** separate the data set into 4 equal groups. $Q_1=P_{25}$, $Q_2=MD$, $Q_3=P_{75}$
- Q_2 = median(Low,High)
  Q_1 = median(Low,Q_2)
  Q_3 = median(Q_2,High)
- The **Interquartile Range**, $IQR = Q_3 - Q_1$.

Chapter 3
Data Description

Section 3-3

Example 3-36

Page #150

Example 3-36: Quartiles

Find Q_1, Q_2, and Q_3 for the data set.

15, 13, 6, 5, 12, 50, 22, 18

Sort in ascending order.

5, 6, 12, 13, 15, 18, 22, 50 (data array), n = 8

$$Q_2 = median(Low, High) = \frac{13+15}{2} = \boxed{14}$$

$$Q_1 = median(Low, MD) = \frac{6+12}{2} = \boxed{9}$$

$$Q_3 = median(MD, High) = \frac{18+22}{2} = \boxed{20}$$

Measures of Position: Outliers

- An **outlier** is an extremely high or low data value when compared with the rest of the data values.

  e.g. 1, 2, 3, 3, 7, 11, 18, 30, 61 n = 9

- A data value less than $Q_1 - 1.5(IQR)$ or greater than $Q_3 + 1.5(IQR)$ can be considered an outlier.

e.g. **Q2: 7, Q1: (2+3) / 2 = 2.5, Q3: (18+30) / 2 = 24**

IQR = Q3 – Q1 = 24 – 2.5 = 21.5

L: Q1 – 1.5(IQR) = -29.75

H: Q3 + 1.5(IQR) = 56.25

61 (> 56.25) is an outlier.

3.4 Exploratory Data Analysis

- The **Five-Number Summary** is composed of the following numbers:

  Lowest , Q_1, MD, Q_3, Highest

  (Minimum) (Maximum)

- The Five-Number Summary can be graphically represented using a **Boxplot**.

Procedure Table

Constructing Boxplots

1. Find the five-number summary.
2. Draw a horizontal axis with a scale that includes the maximum and minimum data values.
3. Draw a box with vertical sides through Q_1 *and* Q_3*, and draw a vertical line* though the median.
4. Draw a line from the minimum data value to the left side of the box and a line from the maximum data value to the right side of the box.

Chapter 3
Data Description

Section 3-4

Example 3-38

Page #163

Example 3-38: Meteorites

The number of meteorites found in 10 U.S. states is shown. Construct a boxplot for the data.

89, 47, 164, 296, 30, 215, 138, 78, 48, 39

30, 39, 47, 48, 78, 89, 138, 164, 215, 296 (data array)

↑ Lowest ↑ Q_1 ↑ MD ↑ Q_3 ↑ Highest

Five-Number Summary: 30-47-83.5-164-296