



Chapter 3.

Numerical Descriptive Measures

The Measures

- ▶ Measures of Central Tendency
- ▶ Measures of Variations

Measures of Central Tendency

- ▶ Arithmetic Mean
- ▶ Median
- ▶ Mode
- ▶ Geometric Mean

The Mean

Arithmetic Mean

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

X_i is the observation number i from a sample of size n .

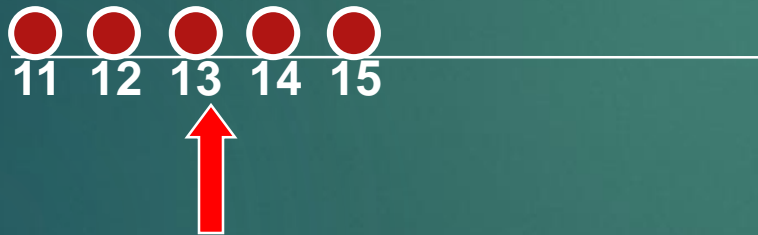
The Sigma Operator (Σ) sums up all the values observed in the sample; i.e., $\Sigma X_i = X_1 + X_2 + X_3 + \dots + X_n$

Arithmetic Mean:

The Importance of an Outlier

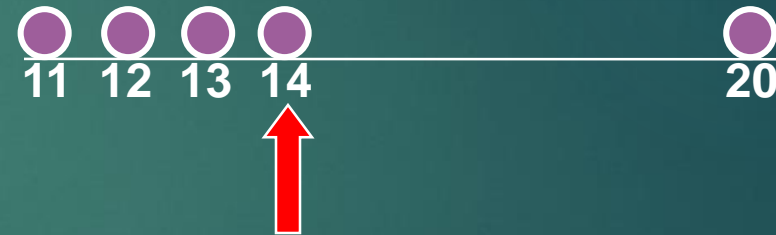
Data Set No. 1: [11, 12, 15, 14, 13]

Data Set No. 2: [11, 12, 14, 13, 20]



Mean = 13

$$\frac{11+12+13+14+15}{5} = \frac{65}{5} = 13$$



Mean = 14

$$\frac{11+12+13+14+20}{5} = \frac{70}{5} = 14$$

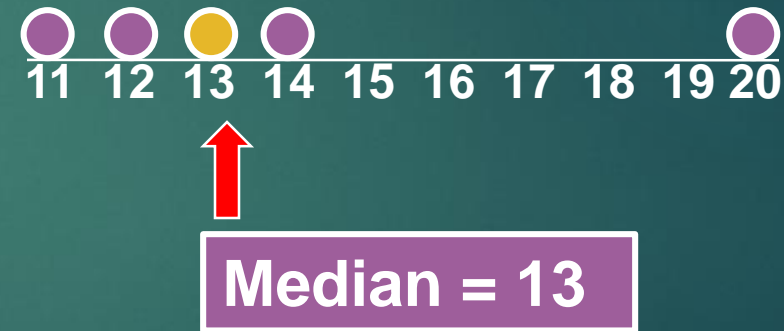
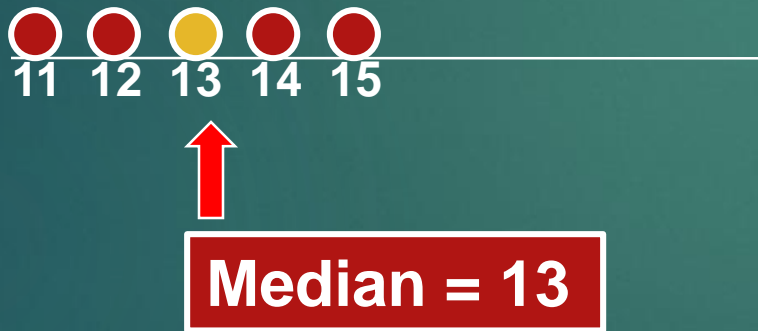
The Median

The Median:

- ▶ In an ordered array, the median is the **middle number** (50% above, 50% below)

Data Set No. 1: [11, 12, 15, 14, 13]

Data Set No. 2: [11, 12, 14, 13, 20]



- ▶ Median is less sensitive than the mean to extreme values

Locating the Median

- ▶ The location of the median when the values are in numerical order (smallest to largest):

$$\text{Median position} = \frac{n+1}{2} \text{ position in the ordered data}$$

- ▶ If the number of values is odd, the median is the middle number (Think about the median in: 1, 2, 3, 4, and 5.)
- ▶ If the number of values is even, the median is the average of the two middle numbers (Think about the median in: 1, 2, 3, 4, 5, and 6)

Note that $\frac{n+1}{2}$ is **not** the *value* of the median, only the *position* of the median in the ranked data

The Median

- Identify the Median in the following Ordered Arrays:

21, 39, 35, 39, 39, 40, 43, 44, 44

- How about the Median in this array?

21, 39, 35, 39, 39, 40, 43, 44, 44, 52

The Median

- Identify the Median in the following Ordered Arrays:

21, 39, 35, 39, 39, 40, 43, 44, 44

- How about the Median in this array?

21, 39, 35, 39, 39 -- 40, 43, 44, 44, 52

39.5

$$\text{Median} = 0.5 * (39 + 40) = 39.5$$

StatTalk



- ▶ The difference between Average, Mean, and Median.
- ▶ Median is useful in describing the data, whereas Mean is useful in making decisions.

NOTE: You can find StatTalk videos in the Multimedia Library on your MyStatLab account. Go to the MyStatLab entry for this course, click on “Multimedia Library,” choose “Select All” under “Media Type,” and select “All Chapter” under “Chapter.” Then, click on “Find Now,” and scroll towards the bottom of the page to find StatTalk Videos

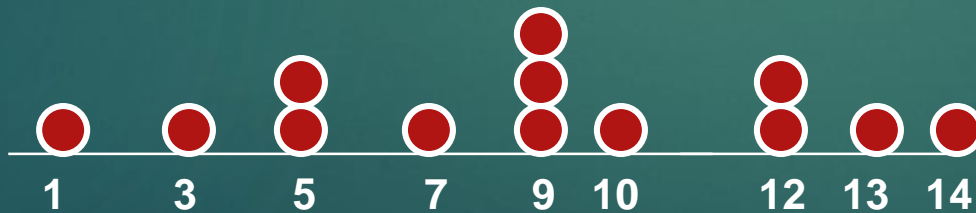
The Mode

The Mode

- ▶ Mode is the value that occurs most often in the data
 - ▶ Not affected by extreme values
 - ▶ Used for either numerical or categorical data
 - ▶ There may be no mode
 - ▶ There may be several modes

Data Set No. 1: [1, 5, 5, 3, 7, 9, 12, 13, 12, 9, 14, 10, 9]

Data Set No. 2: [6, 0, 3, 2, 5, 1]



Mode = 9



No Mode

Mean, Median, and Mode

House Prices:

HN1: \$2,000,000

HN2: \$ 500,000

HN3: \$ 300,000

HN4: \$ 100,000

HN5: \$ 100,000

- Compute the Mean, and identify the Median and Mode.

House Prices:

\$2,000,000
\$ 500,000
\$ 300,000
\$ 100,000
\$ 100,000

Total \$ 3,000,000

- Mean = $\$3,000,000 / 5 = \$600,000$
- Median = \$300,000
middle value of ranked data
- Mode = \$100,000
most frequent value

The Geometric Mean

Geometric Mean

$$\overline{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n}$$

X_1 is the observation at the first time period (e.g., first month),
 X_2 is the observation at the second time period (e.g., second month) ... **X_n** is the observation at the n -th time period (e.g., the n -th month)

The **Geometric Mean** is useful to compute the **Rate of Change** (i.e., the rate by which a variable changes over time).
This mean is the n -th root of the product of n values.

Geometric Mean Rate of Return

$$\bar{R}_G = [(1 + R_1) \times (1 + R_2) \times \cdots \times (1 + R_n)]^{1/n} - 1$$

R_1 is the **rate of return** at the first time period (e.g., first month),
 R_2 is the **rate of return** at the second time period (e.g., second month) ... **R_n** is the **rate of return** at the n -th time period (e.g., the n -th month)

The **Geometric Mean Rate of Return** is useful to compute the **Average Percentage Return of an Investment** (i.e., the rate by which an investment changes over time).

This mean is the n -th root of the product of n values.

Example

An investment of \$100,000 declined to \$50,000 at the end of year one and rebounded to \$100,000 at end of year two:

$$X_1 = \$100,000 \quad X_2 = \$50,000 \quad X_3 = \$100,000$$

50% decrease

100% increase

The overall two-year return is zero, since it started and ended at the same level.

$$X_1 = \$100,000 \quad X_2 = \$50,000 \quad X_3 = \$100,000$$

50% decrease

100% increase

Arithmetic mean rate of return:

$$\overline{X} = \frac{(-.5) + (1)}{2} = .25 = 25\%$$

Misleading result

Geometric mean rate of return:

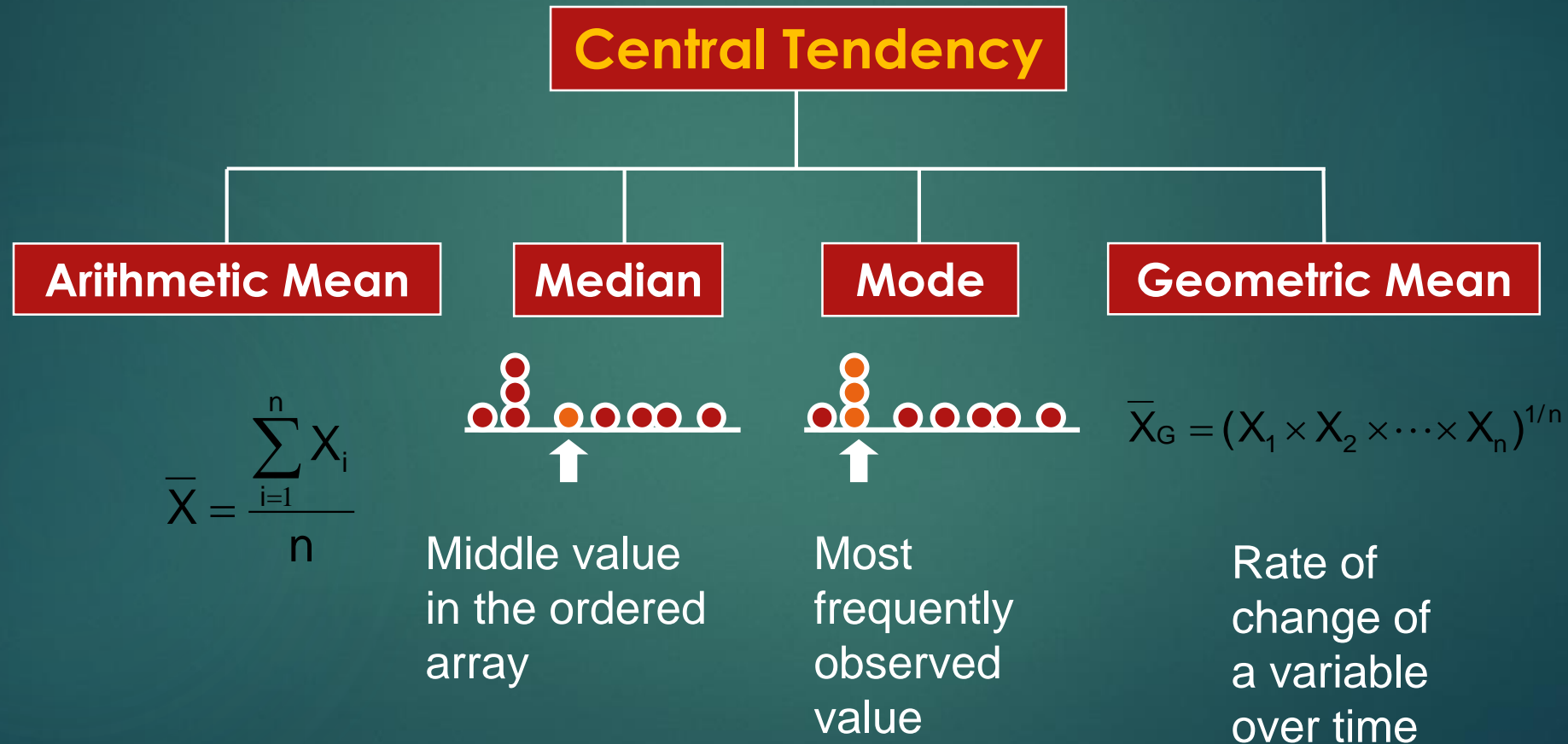
$$\begin{aligned}\overline{R}_G &= [(1 + R_1) \times (1 + R_2) \times \cdots \times (1 + R_n)]^{1/n} - 1 \\ &= [(1 + (-.5)) \times (1 + (1))]^{1/2} - 1 \\ &= [(.50) \times (2)]^{1/2} - 1 = 1^{1/2} - 1 = 0\%\end{aligned}$$

More
representative
result

Measures of Central Tendency:

Summary

DCOVA

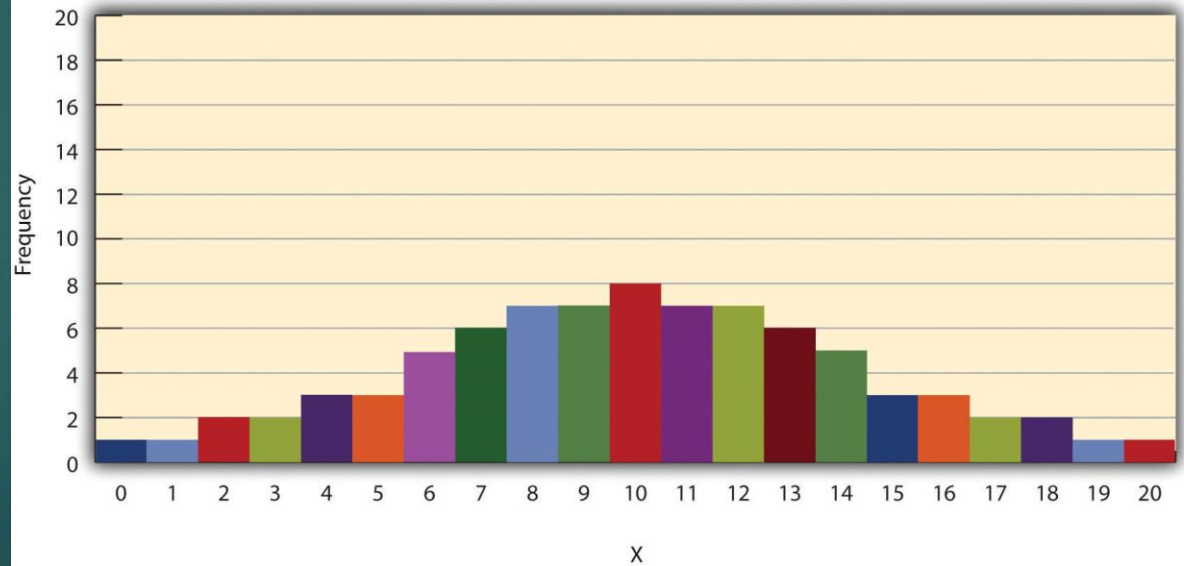
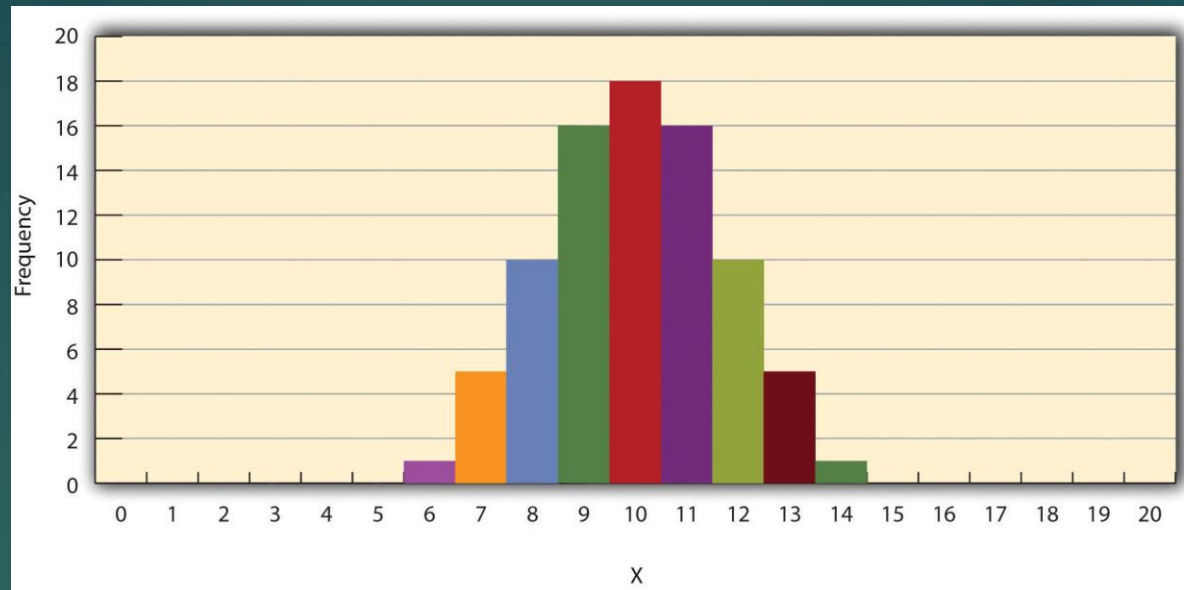


Measures of Variations

- ▶ Range
- ▶ Variance
- ▶ Standard Deviation
- ▶ Coefficient of Variation

Dispersion

- Two variables may have identical measures of central tendency, but different distribution.
- The dispersion of the data matters.



Source: Psychology Research Methods: Core Skills and Concepts

The Range

The Range

- Simplest measure of variation
- Difference between the largest and the smallest values:

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

Data Set: [1, 4, 10, 10, 7, 4, 14, 2, 11, 13, 15, 13, 10]

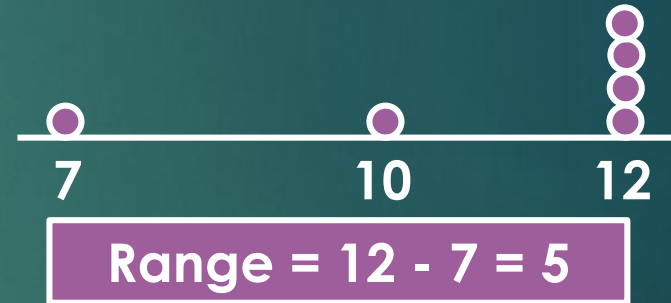
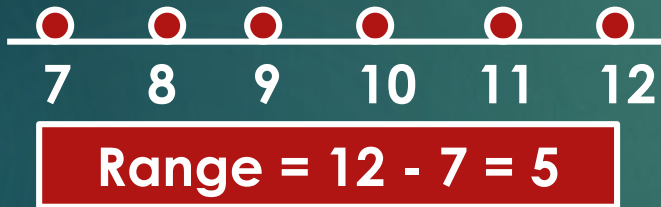


$$\text{Range} = 15 - 1 = 14$$

The Range:

The shortcomings

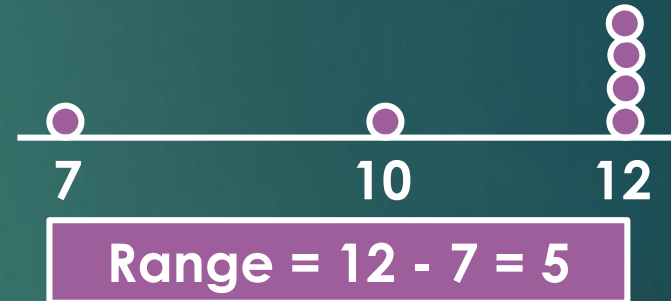
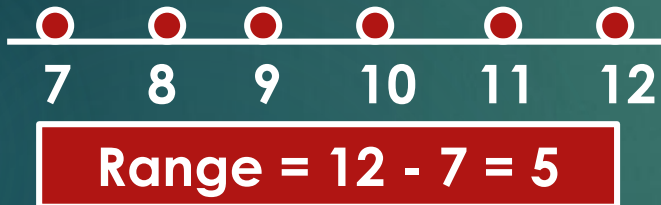
- ▶ The Range measures the Total Spread
- ▶ The way that the data is distributed cannot be described well when we use the Range.



The Range:

The shortcomings

- ▶ The Range measures the Total Spread
- ▶ The way that the data is distributed cannot be described well when we use the Range.



- ▶ It is also very sensitive to the outliers.

Data Set A: [1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 5]

$$\text{Range} = 5 - 1 = 4$$

Data Set B: [1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 150]

$$\text{Range} = 150 - 1 = 149$$

What did we learn?

- ▶ Measures of dispersion should take into account the data distribution.
- ▶ They should reflect the **scatter around the mean**.

The Variance

The Variance

- ▶ The Variance is a measure of the average scatter around the mean.
- ▶ What do we need?
 - ▶ We need to measure the scatter around the mean
 - ▶ We need to consider the average scatter
- ▶ Let's work with a simple data set:

DATA: $[X_1, X_2, X_3, X_4, X_5]$

Scatter around the mean

DATA: $[X_1, X_2, X_3, X_4, X_5]$



Mean = $(X_1 + X_2 + X_3 + X_4 + X_5) / 5$

- ▶ The scatter around the mean for each data value:
 $X_1 - \text{Mean}; X_2 - \text{Mean}; X_3 - \text{Mean}; X_4 - \text{Mean}; X_5 - \text{Mean}$
- ▶ The sum of the above scatters, however, is equal to zero.

Scatter around the mean

DATA: $[X_1, X_2, X_3, X_4, X_5]$



Mean = $(X_1 + X_2 + X_3 + X_4 + X_5) / 5$

- ▶ The scatter around the mean for each data value:
 $X_1 - \text{Mean}; X_2 - \text{Mean}; X_3 - \text{Mean}; X_4 - \text{Mean}; X_5 - \text{Mean}$
- ▶ The sum of the above scatters, however, is equal to zero. Why?

$$Y = X_1 - \text{Mean} + X_2 - \text{Mean} + X_3 - \text{Mean} + X_4 - \text{Mean} + X_5 - \text{Mean}$$

Scatter around the mean

DATA: $[X_1, X_2, X_3, X_4, X_5]$



Mean = $(X_1 + X_2 + X_3 + X_4 + X_5) / 5$

- ▶ The scatter around the mean for each data value:

$X_1 - \text{Mean}; X_2 - \text{Mean}; X_3 - \text{Mean}; X_4 - \text{Mean}; X_5 - \text{Mean}$

- ▶ The sum of the above scatters, however, is equal to zero. Why?

$Y = X_1 - \text{Mean} + X_2 - \text{Mean} + X_3 - \text{Mean} + X_4 - \text{Mean} + X_5 - \text{Mean}$

$= X_1 + X_2 + X_3 + X_4 + X_5 - 5 * \text{Mean}$

Scatter around the mean

DATA: $[X_1, X_2, X_3, X_4, X_5]$



Mean = $(X_1 + X_2 + X_3 + X_4 + X_5) / 5$

- ▶ The scatter around the mean for each data value:

$X_1 - \text{Mean}; X_2 - \text{Mean}; X_3 - \text{Mean}; X_4 - \text{Mean}; X_5 - \text{Mean}$

- ▶ The sum of the above scatters, however, is equal to zero. Why?

$Y = X_1 - \text{Mean} + X_2 - \text{Mean} + X_3 - \text{Mean} + X_4 - \text{Mean} + X_5 - \text{Mean}$

$= X_1 + X_2 + X_3 + X_4 + X_5 - 5 * \text{Mean}$

$= X_1 + X_2 + X_3 + X_4 + X_5 - 5 * (X_1 + X_2 + X_3 + X_4 + X_5) / 5$

Scatter around the mean

DATA: $[X_1, X_2, X_3, X_4, X_5]$



Mean = $(X_1 + X_2 + X_3 + X_4 + X_5) / 5$

- ▶ The scatter around the mean for each data value:
 $X_1 - \text{Mean}; X_2 - \text{Mean}; X_3 - \text{Mean}; X_4 - \text{Mean}; X_5 - \text{Mean}$
- ▶ The sum of the above scatters, however, is equal to zero. Why?

$$Y = X_1 - \text{Mean} + X_2 - \text{Mean} + X_3 - \text{Mean} + X_4 - \text{Mean} + X_5 - \text{Mean}$$

$$= X_1 + X_2 + X_3 + X_4 + X_5 - 5 * \text{Mean}$$

$$= X_1 + X_2 + X_3 + X_4 + X_5 - 5 * (X_1 + X_2 + X_3 + X_4 + X_5) / 5$$

$$= X_1 + X_2 + X_3 + X_4 + X_5 - (X_1 + X_2 + X_3 + X_4 + X_5) = 0$$

Scatter around the mean

DATA: $[X_1, X_2, X_3, X_4, X_5]$



Mean = $(X_1 + X_2 + X_3 + X_4 + X_5) / 5$

- ▶ The scatter around the mean for each data value:
 $X_1 - \text{Mean}; X_2 - \text{Mean}; X_3 - \text{Mean}; X_4 - \text{Mean}; X_5 - \text{Mean}$
- ▶ **The sum of the square** of the above scatters is often non-zero.
 $SS = (X_1 - \text{Mean})^2 + (X_2 - \text{Mean})^2 + (X_3 - \text{Mean})^2 + (X_4 - \text{Mean})^2 + (X_5 - \text{Mean})^2$
- ▶ Thus, we use the sum of squares (SS) as our measure for the scatter around the mean.

Scatter around the mean

DATA: $[X_1, X_2, X_3, X_4, X_5]$



Mean = $(X_1 + X_2 + X_3 + X_4 + X_5) / 5$



$$SS = (X_1 - \text{Mean})^2 + (X_2 - \text{Mean})^2 + (X_3 - \text{Mean})^2 + (X_4 - \text{Mean})^2 + (X_5 - \text{Mean})^2$$

Computing the average scatter

DATA: $[X_1, X_2, X_3, X_4, X_5]$

Mean = $(X_1 + X_2 + X_3 + X_4 + X_5) / 5$

$$SS = (X_1 - \text{Mean})^2 + (X_2 - \text{Mean})^2 + (X_3 - \text{Mean})^2 + (X_4 - \text{Mean})^2 + (X_5 - \text{Mean})^2$$

$$\text{VARIANCE} = S^2 = SS / (n - 1)$$

The Sample Variance

► Average (approximately) of squared deviations of values from the mean

► Sample variance:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Where

\bar{x} = arithmetic mean

n = sample size

x_i = i^{th} value of the variable X

The Sample Variance

- ▶ Average (approximately) of squared deviations of values from the mean

Why “approximately”?

- ▶ Sample variance:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Where

\bar{x} = arithmetic mean

n = sample size

x_i = i^{th} value of the variable X

The Sample Variance

► Average (approximately) of squared deviations of values from the mean

► Sample variance:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Why “approximately”?

It has to do with inference. We will discuss this in more details later. But, in general, we are able to get a better estimate of POPULATION VARIANCE when we divide the SS by $n-1$ rather than n .

Where

\bar{x} = arithmetic mean

n = sample size

x_i = i^{th} value of the variable X

The Standard Deviation

The Sample Standard Deviation

DCOVA

- ▶ Standard Deviation is simply the square root of the Variance
- ▶ Most commonly used measure of variation
- ▶ Has the same units as the original data

- ▶ Sample standard deviation:

$$S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$



Example:

The Variations of the Calories in the Cereals

$$\begin{aligned} n &= 7 \\ \bar{X} &= 130 \end{aligned}$$

Calories	Step 1: $(X_i - \bar{X})$	Step 2: $(X_i - \bar{X})^2$
80	-50	2,500
100	-30	900
100	-30	900
110	-20	400
130	0	0
190	60	3,600
200	70	4,900
Step 3: Sum		13,200
Step 4: Divide by $(n - 1)$		2,220

V

$$\begin{aligned} S^2 &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} \\ &= \frac{(80 - 130)^2 + (100 - 130)^2 + \dots + (200 - 130)^2}{7 - 1} \\ &= \frac{13,200}{6} \\ &= 2,200 \end{aligned}$$

SD

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}} = \sqrt{2,200} = 46.9042$$

Measures of Variation:

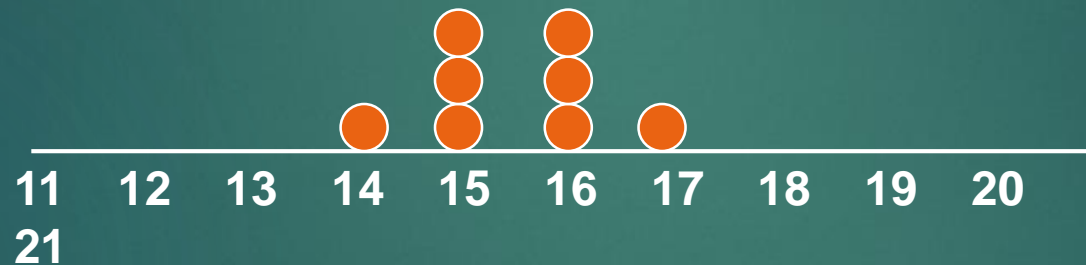
Comparing Standard Deviations

Data A



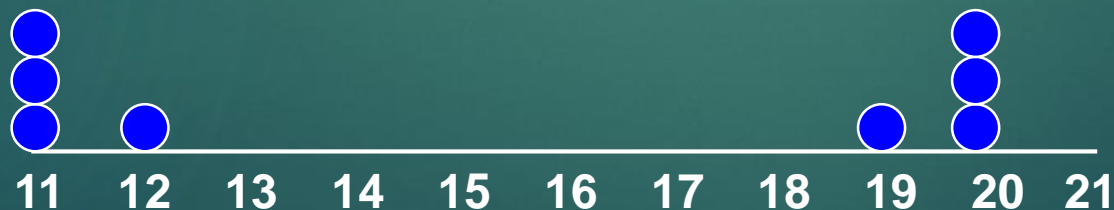
Mean = 15.5
 $S = 3.338$

Data B



Mean = 15.5
 $S = 0.926$

Data C



Mean = 15.5
 $S = 4.567$

Standard Deviation Applications:

Coefficient of Variation

- ▶ Coefficient of Variation (CV) measures the **RELATIVE Variations**:

$$CV = \left(\frac{S}{\bar{X}} \right) \cdot 100\%$$

- ▶ Unlike SD, CV is unit-less (always in percentage)

► Stock A:

► Average price last year = \$50

► Standard deviation = \$5

$$CV_A = \left(\frac{S}{\bar{X}} \right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

► Stock B:

► Average price last year = \$100

► Standard deviation = \$5

$$CV_B = \left(\frac{S}{\bar{X}} \right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Both stocks have the same standard deviation, but stock B is less variable relative to its price

Standard Deviation Applications:

The Z-score

- ▶ Given the mean and the standard deviation, each data point has a Z-score, which shows its relative dispersion from the mean

$$Z = \frac{X - \bar{X}}{S}$$

- ▶ The Z-score is used to identify the outliers
- ▶ A data value is considered an extreme outlier if its Z-score is less than -3.0 or greater than +3.0.

- ▶ Suppose the mean math SAT score is 490, with a standard deviation of 100.
- ▶ Let's compute the Z-score for a test score of 620.

$$Z = \frac{X - \bar{X}}{S} = \frac{620 - 490}{100} = \frac{130}{100} = 1.3$$

A score of 620 is 1.3 standard deviations above the mean and would not be considered an outlier.

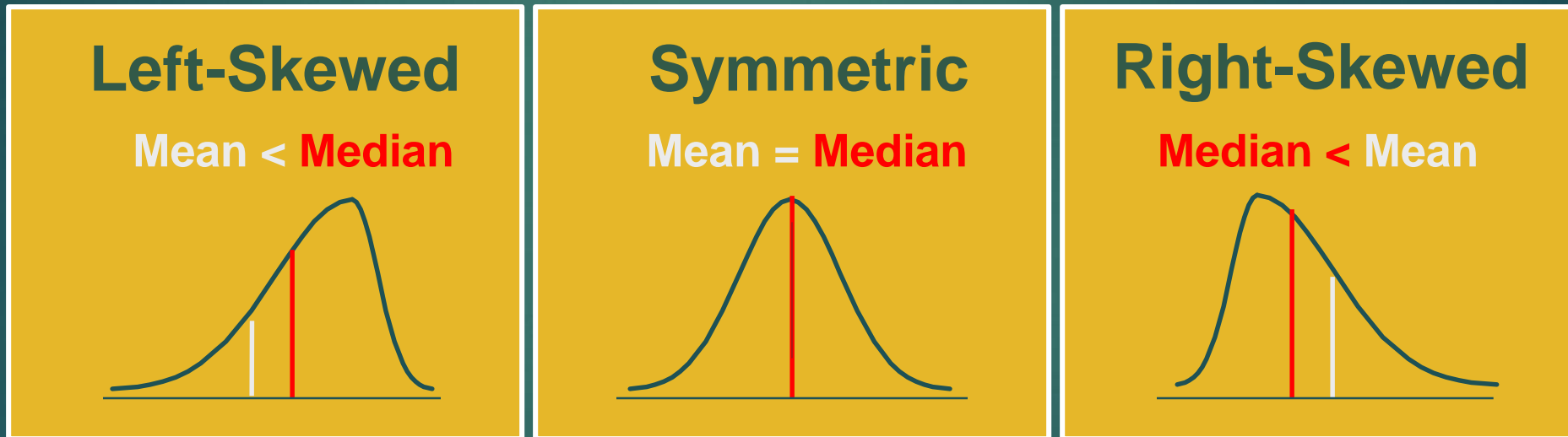
Shape of a Distribution

- ▶ Describes how data are distributed
- ▶ Two useful shape related statistics are:
 - ▶ **Skewness**
 - ▶ Measures the extent to which data values are **not symmetrical**
 - ▶ **Kurtosis**
 - ▶ Kurtosis affects the “**peakedness**” of the curve of the distribution—that is, how sharply the curve rises approaching the center of the distribution

Shape of a Distribution

Skewness

- Measures the extent to which data is not symmetrical



Skewness
Statistic

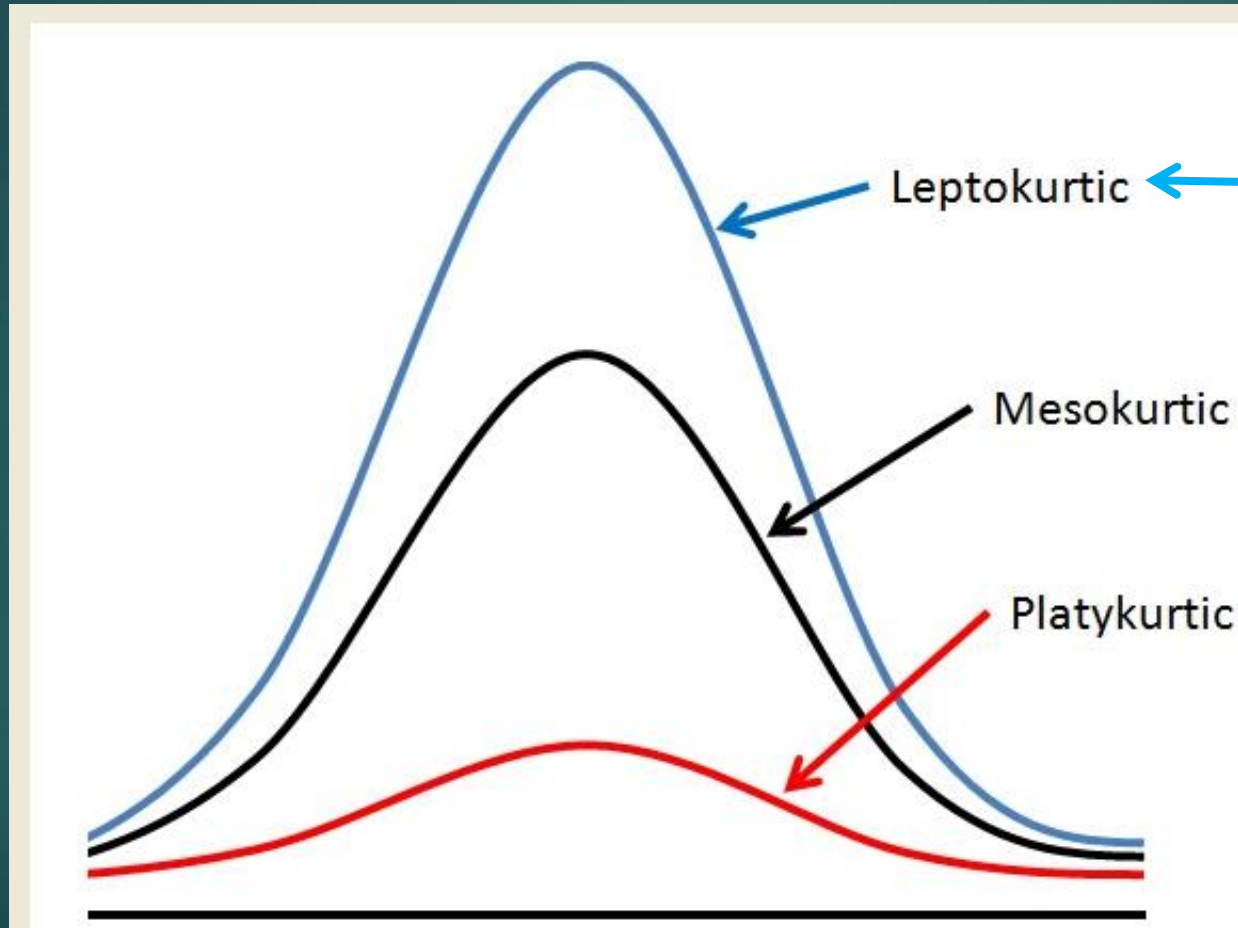
< 0

0

> 0

Shape of a Distribution

Kurtosis



**Sharper Peak
Than Bell-Shaped
(Kurtosis > 0)**

**Bell-Shaped
Normal
(Kurtosis $= 0$)**

**Flatter Than
Bell-Shaped
(Kurtosis < 0)**

Other Measures

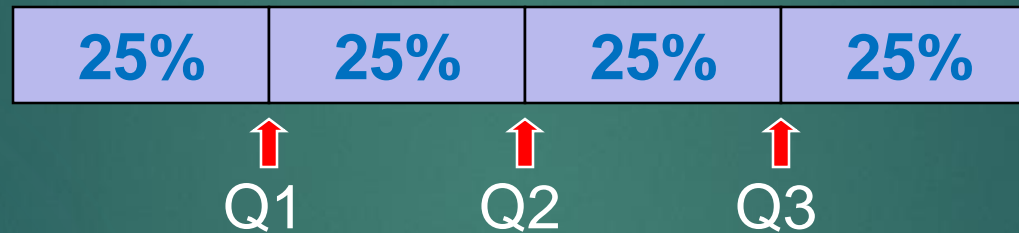
Variance and Standard Deviation are useful variations of dispersion when we use the Mean as our measure for central tendency.

What are the useful dispersion variations when we use the Median? Let's start with the Quartiles.

The Quartiles

Quartile Measures

- ▶ Quartiles split the ranked data into 4 segments with an equal number of values per segment



- The **first quartile**, Q_1 , is **the value** for which 25% of the observations are smaller and 75% are larger
- The **second quartile**, Q_2 , is the same as the **median**: 50% of the observations are smaller and 50% are larger
- The **third quartile**, Q_3 , is **the value** for which 75% of the observations are smaller and 25% are larger

Quartile Measures:

Locating Quartiles

Find a quartile by determining the value in the appropriate position in the ranked data, where

First quartile position: $Q_1 = (n+1)/4$ ranked value

Second quartile position: $Q_2 = (n+1)/2$ ranked value

Third quartile position: $Q_3 = 3*(n+1)/4$ ranked value

where n is the number of observed values

Quartile Measures:

Calculation Rules for the Position

- ▶ When calculating the **ranked position** use the following rules
 - ▶ If the result is a whole number then it is the ranked position to use
 - ▶ If the result is a fractional half (e.g. 2.5, 7.5, 8.5, etc.) then average the two corresponding data values.
 - ▶ If the result is not a whole number or a fractional half then round the result to the nearest integer to find the ranked position.

Example:

Locating Quartiles

Data in Ordered Array: [11 12 13 16 16 17 18 21 22]

Quartile Measures:

Locating Quartiles

Data in Ordered Array: [11 12 13 16 16 17 18 21 22]

(n = 9)

Q_1 is in the $(9+1)/4 = 2.5$ position of the ranked data,

so $Q_1 = (12+13)/2 = 12.5$

Q_2 is in the $(9+1)/2 = 5^{\text{th}}$ position of the ranked data,

so $Q_2 = \text{median} = 16$

Q_3 is in the $3(9+1)/4 = 7.5$ position of the ranked data,

so $Q_3 = (18+21)/2 = 19.5$

The Interquartile Range

Quartile Measures:

The Interquartile Range (IQR)

- ▶ The **IQR** is $Q_3 - Q_1$ and measures the **spread in the middle 50%** of the data
- ▶ The IQR is also called the **mid-spread** because it covers the middle 50% of the data
- ▶ The IQR is a measure of variability that is not influenced by outliers or extreme values (i.e., it is a resistant measures)


Calculating The Interquartile Range

Median				
X_{minimum}	Q_1	Q_2	Q_3	X_{maximum}
12	30	45	57	70



A horizontal double-headed arrow is drawn below the table, spanning from the column for Q_1 (value 30) to the column for Q_3 (value 57). Vertical lines extend from the Q_1 and Q_3 cells down to the ends of the arrow.

Interquartile range = $57 - 30 = 27$



The Five Number Summary And The Box Plot

The Five Number Summary

The five numbers that help describe the center, spread and shape of data are:

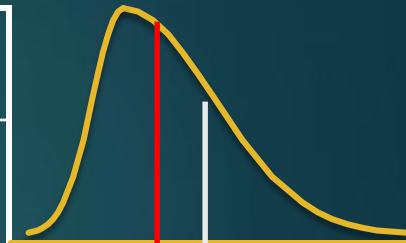
- X_{smallest}
- First Quartile (Q_1)
- Median (Q_2)
- Third Quartile (Q_3)
- X_{largest}

Relationships Among the Five-number Summary



Mean < Median

Left-Skewed	Symmetric	Right-Skewed
Median – X_{smallest} >	Median – X_{smallest} \approx	Median – X_{smallest} <
X_{largest} – Median	X_{largest} – Median	X_{largest} – Median
Q_1 – X_{smallest} >	Q_1 – X_{smallest} \approx	Q_1 – X_{smallest} <
X_{largest} – Q_3	X_{largest} – Q_3	X_{largest} – Q_3
Median – Q_1 >	Median – Q_1 \approx	Median – Q_1 <
Q_3 – Median	Q_3 – Median	Q_3 – Median



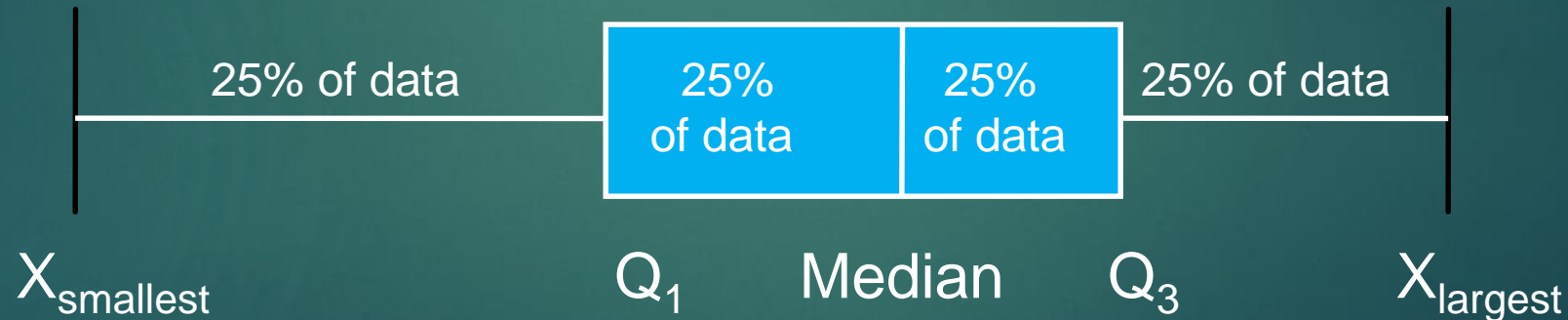
Median < Mean

Five Number Summary and The Boxplot

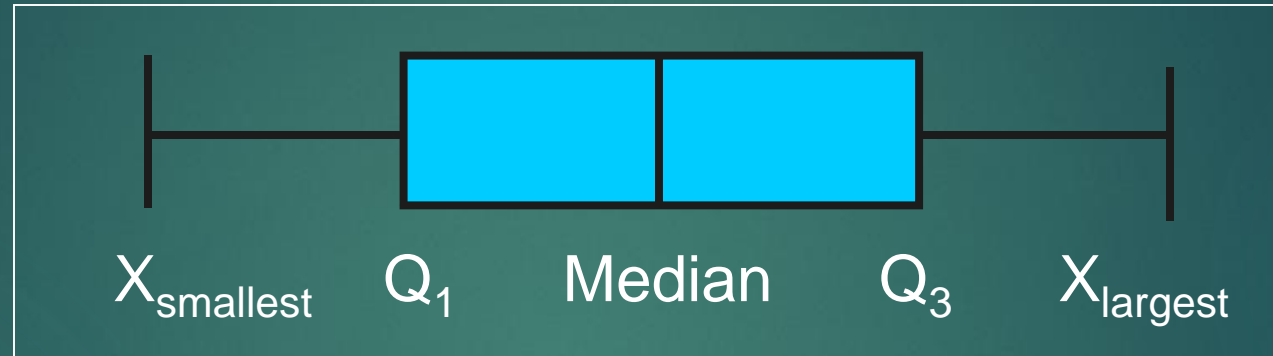
DCOVA

- **The Boxplot:** A Graphical display of the data based on the five-number summary

Example:



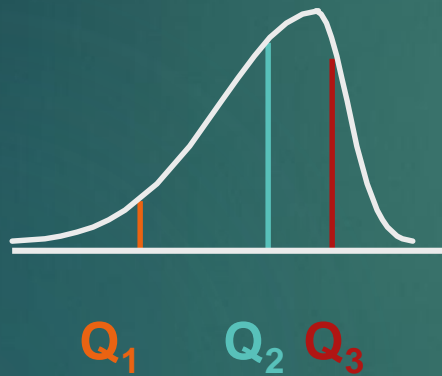
Example: A Symmetric Boxplot



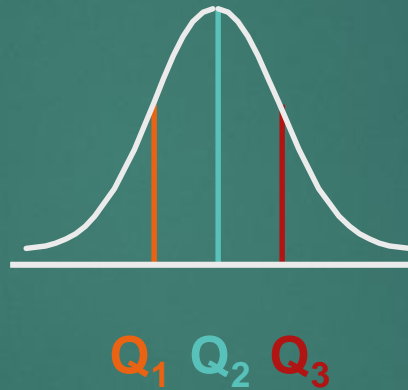
Unlike a Frequency Distribution plot (e.g. Polygon), the Box plot only shows where the **smallest** and the **largest** values as well as the **quartiles** are located – and does not include the frequencies

Example: Box Plots and Freq. Distributions

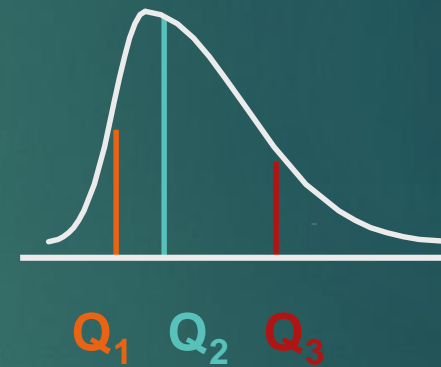
Left-Skewed




Symmetric



Right-Skewed



- 
- ▶ Assignments:
Section 3.1, 3.2, and 3.3.



The Numerical Descriptive Measures for the Population

Numerical Descriptive Measures for a Population

- Descriptive statistics discussed previously described a *sample*, not the *population*.
- Summary measures describing a population, called **parameters**, are denoted with Greek letters.
- Important population parameters are the population mean, variance, and standard deviation.

Numerical Descriptive Measures for a Population:

The mean μ (read as “mu”)

$$\mu = \frac{\sum_{i=1}^N X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

Where μ = population mean

N = population size

X_i = i^{th} value of the variable X

Numerical Descriptive Measures For A Population:

The Variance σ^2 (read as “sigma-square”)

The Standard Deviation σ (read as “sigma-square”)

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}}$$

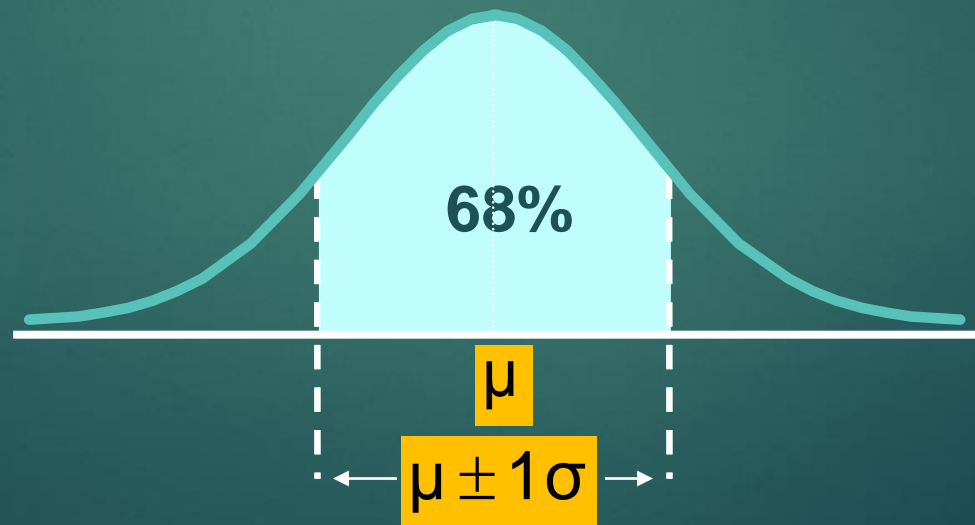
Where μ = population mean

N = population size

X_i = i^{th} value of the variable X

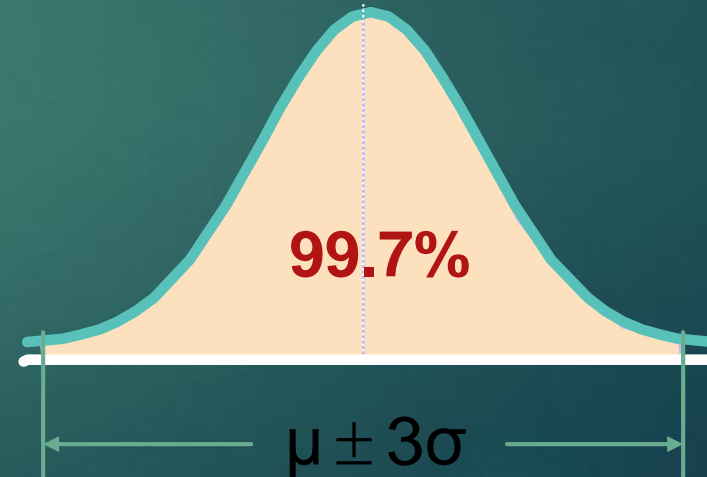
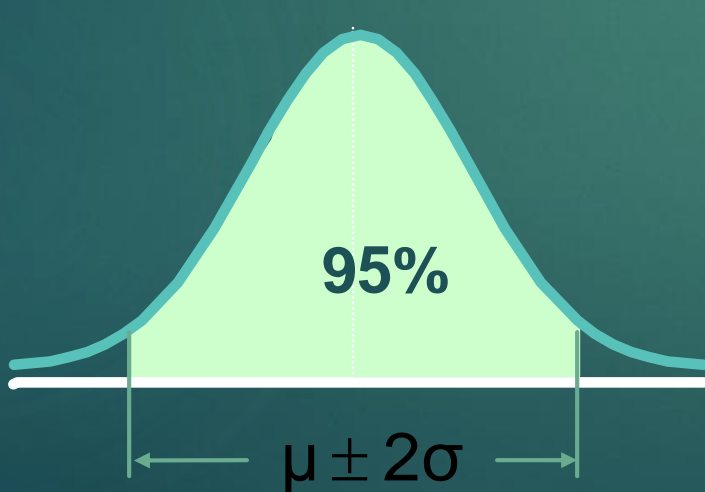
The Empirical Rule

- ▶ The empirical rule approximates the variation of data in a bell-shaped distribution
- ▶ Approximately 68% of the data in a bell shaped distribution is within 1 standard deviation of the mean; i.e. $\mu \pm 1\sigma$



The Empirical Rule (cont'd)

- ▶ Approximately **95%** of the data in a bell-shaped distribution lies **within two standard deviations of the mean**, or $\mu \pm 2\sigma$
- ▶ Approximately **99.7%** of the data in a bell-shaped distribution lies **within three standard deviations of the mean**, or $\mu \pm 3\sigma$



Using the Empirical Rule

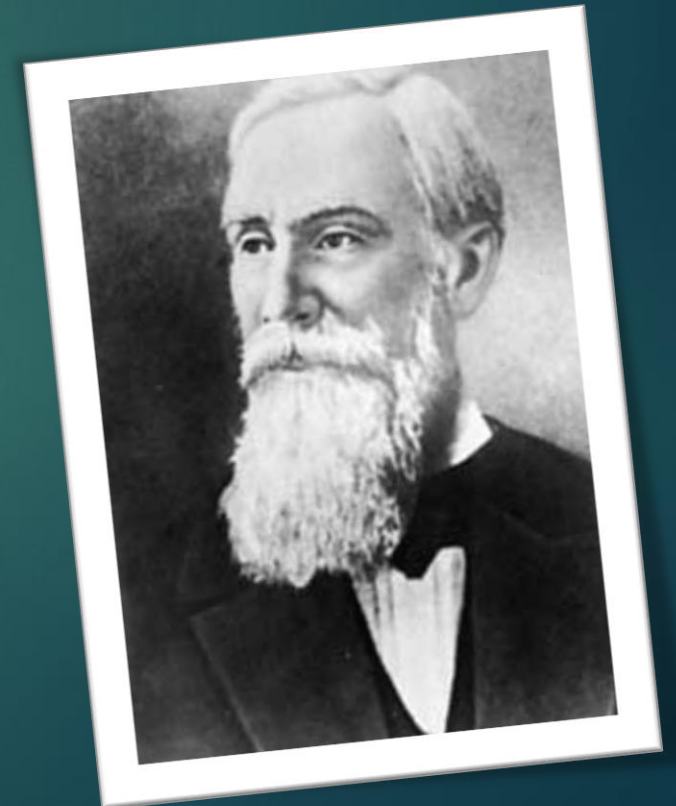
- Suppose that the variable Math SAT scores is bell-shaped with a **mean** of **500** and a **standard deviation** of **90**. Then,
 - **68%** of all test takers scored between **410** and **590** : (500 ± 90) .
 - **95%** of all test takers scored between **320** and **680**: (500 ± 180) .
 - **99.7%** of all test takers scored between **230** and **770**: (500 ± 270) .

The Chebyshev Rule

- ▶ At least $(1 - 1/k^2) \times 100\%$ of the values (for $k > 1$) fall within k standard deviations of the mean regardless of how the data are distributed

▶ Examples:

At least	Within
$(1 - 1/2^2) \times 100\% = 75\%$	$k=2 \quad (\mu \pm 2\sigma)$
$(1 - 1/3^2) \times 100\% = 88.89\%$	$k=3 \quad (\mu \pm 3\sigma)$





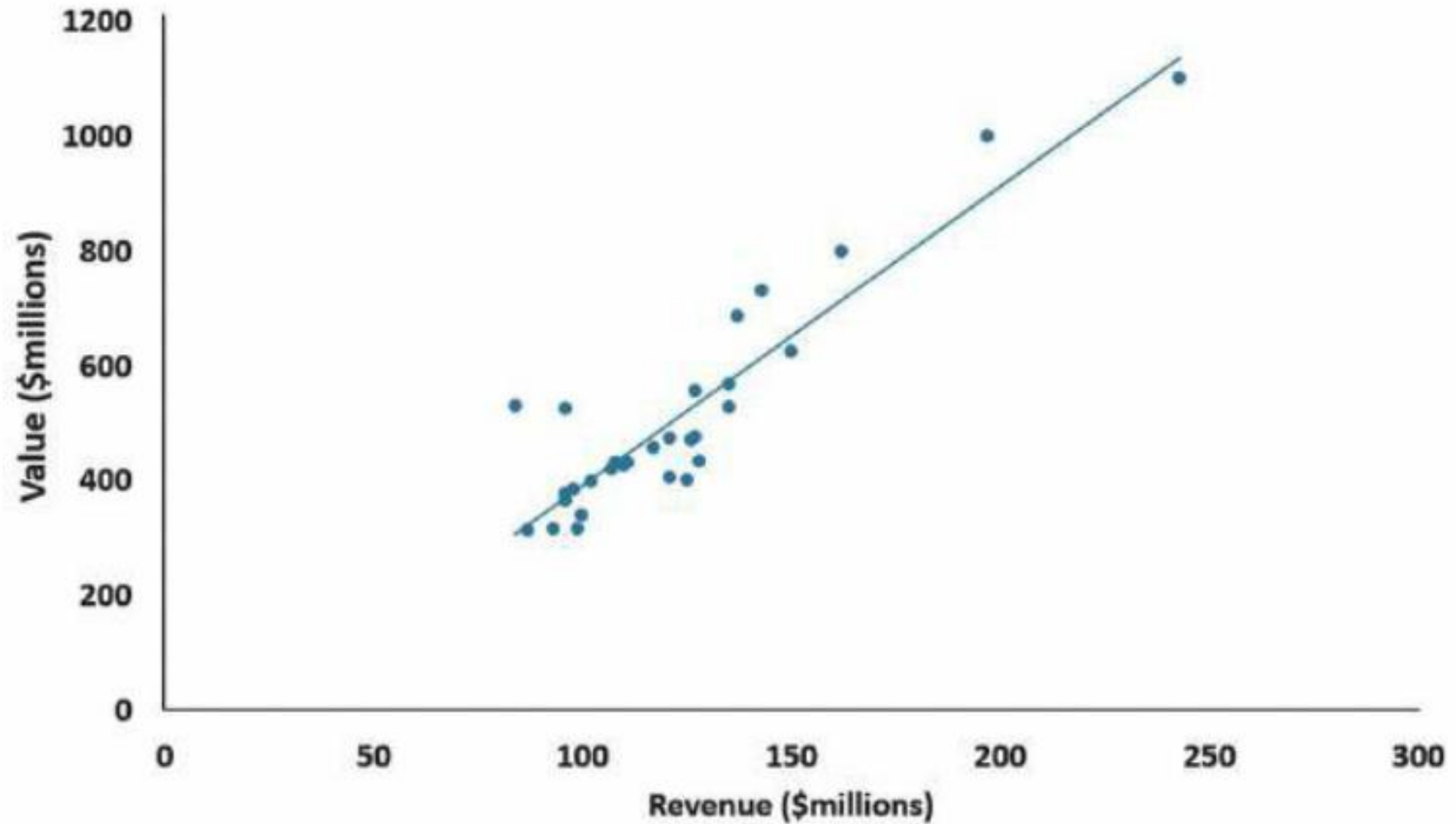
Measuring the Relationship btw. Two Numerical Variables

Detecting a Relationship

Team Code	Revenue (\$millions)	Value (\$millions)	Team Code	Revenue (\$millions)	Value (\$millions)	Team Code	Revenue (\$millions)	Value (\$millions)
ATL	99	316	HOU	135	568	OKC	127	475
BOS	143	730	IND	98	383	ORL	126	470
BRK	84	530	LAC	108	430	PHI	107	418
CHA	93	315	LAL	197	1,000	PHX	121	474
CHI	162	800	MEM	96	377	POR	117	457
CLE	128	434	MIA	150	625	SAC	96	525
DAL	137	685	MIL	87	312	SAS	135	527
DEN	110	427	MIN	96	364	TOR	121	405
DET	125	400	NOH	100	340	UTA	111	432
GSW	127	555	NYK	243	1,100	WAS	102	397

Source: Data extracted from www.forbes.com/nba-valuations.

Scatter Plot of Revenue and Value for NBA Teams



The Covariance

- ▶ The covariance measures the strength of the linear relationship between **two numerical variables** (X & Y)
- ▶ The **sample covariance**:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

- ▶ Only concerned with the strength of the relationship
- ▶ No causal effect is implied

Interpreting Covariance

DCOVA

► Covariance between two variables:

$\text{cov}(X,Y) > 0$: X and Y tend to move in the **same** direction

$\text{cov}(X,Y) < 0$: X and Y tend to move in **opposite** directions

$\text{cov}(X,Y) = 0$: X and Y are **independent**

► The covariance has a major flaw: It is not possible to determine the relative strength of the relationship from the size of the covariance

Coefficient of Correlation

- ▶ Measures the relative strength of the linear relationship between two numerical variables
- ▶ Sample coefficient of correlation:

$$r = \frac{\text{cov}(X, Y)}{S_X S_Y}$$

where

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

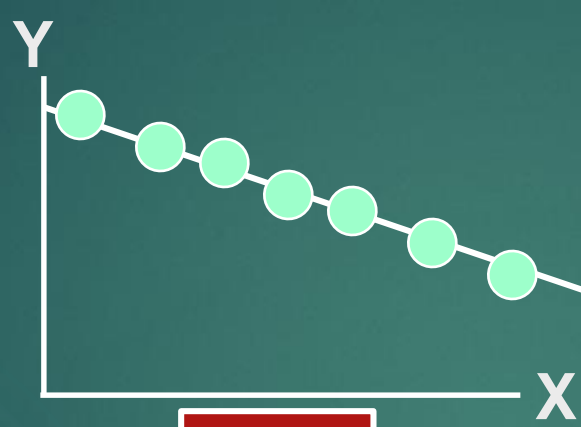
$$S_X = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

$$S_Y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}}$$

Features of the Coefficient of Correlation

- ▶ The **population coefficient** of correlation is referred as ρ .
- ▶ The **sample coefficient** of correlation is referred to as r .
- ▶ Either ρ or r have the following features:
 - ▶ Unit free
 - ▶ Range between **-1** and **1**
 - ▶ The closer to **-1**, the stronger the negative linear relationship
 - ▶ The closer to **1**, the stronger the positive linear relationship
 - ▶ The closer to **0**, the weaker the linear relationship

Scatter Plots of Sample Data with Various Coefficients of Correlation



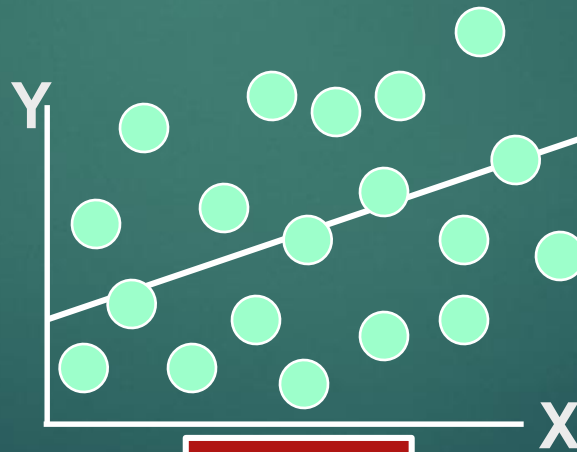
$$r = -1$$



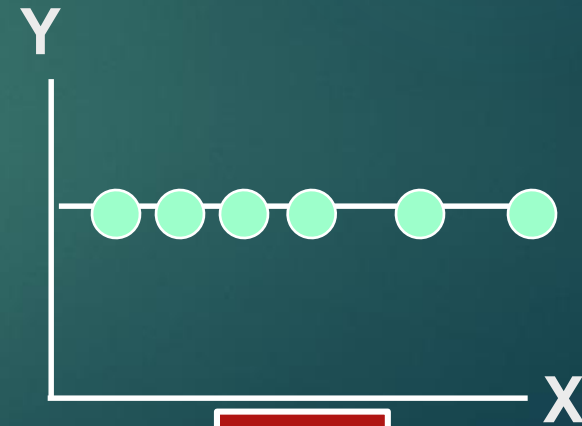
$$r = -.6$$




$$r = +1$$



$$r = +.3$$



$$r = 0$$

- 
- ▶ Assignments:
Section 3.4 and 3.5.
 - ▶ Reading:
Section 3.6