Calculating Descriptive Statistics

- Measures of Central Tendency
- Measures of Variability
- Using the Mean and Standard Deviation Together
- Measures of Relative Position
- Measures of Association Between Two Variables

- Reading:
  - Chapter 3 (except Section 3.4) and Chapter 2 (Section 2.6)
- Optional:
  - Wikipedia article on Bessel's correction (mathematically challenging, but section on Sources of Bias explains intuitively correction $n - 1$ in the sample variance)
  - Note on skewed distribution on Canvas

Sample vs. Population Measures

Recall…

If the measures are computed for data from a **sample**, they are called *sample statistics*

- Most times denoted by Latin letters (x, y, A, B, …)

If the measures are computed for data from a **population**, they are called *population parameters*

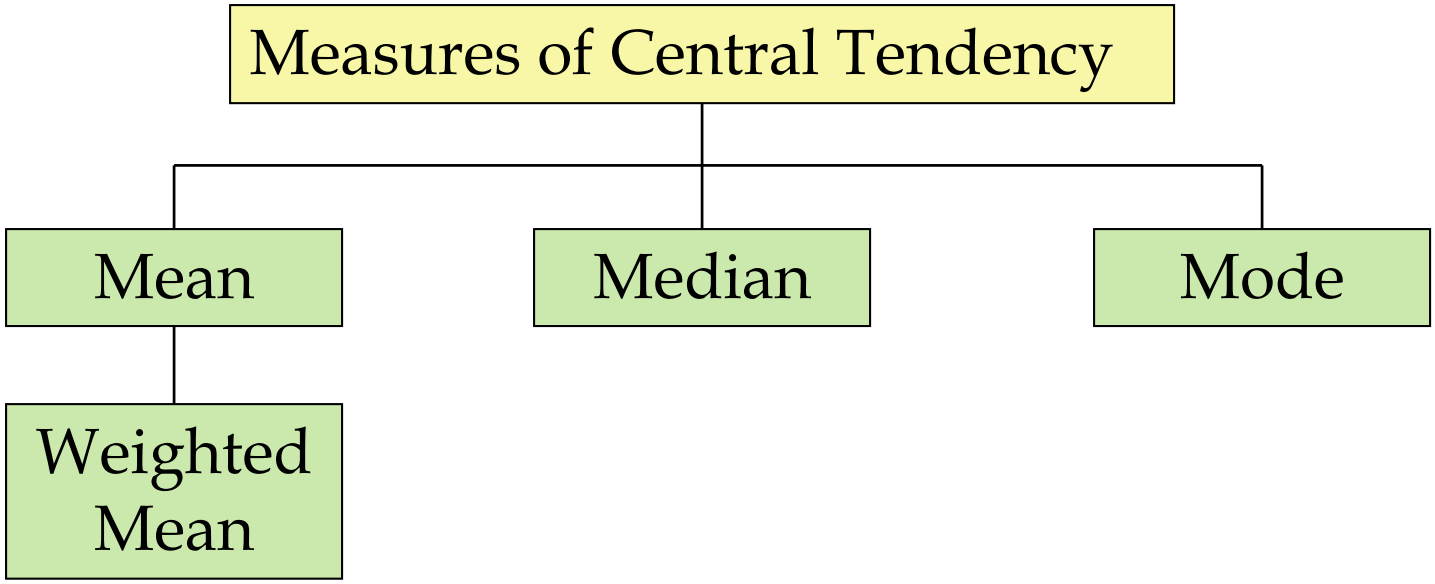
- Most times denoted by Greek letters (μ, σ, θ, …)

A sample statistic is referred to as the *point estimator* of the corresponding population parameter:

- E.g. sample mean is a point estimator of the population mean

Measures of Central Tendency

Central tendency is a single value used to describe the center point of a data set



The Mean

The **mean**, or the **average**, is the most common measure of central tendency

- The mean of a data set is the average of all the data values
- Calculate the *sample mean* by adding all the values in a data set and then dividing the result by the number of observations **in the sample**
- The *population mean*: similar to the sample mean but is based on the all observations **in the population**

The Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where $\bar{x}$ = the **sample** mean

x_i = the values in the sample

$\sum_{i=1}^{n} x_i$ = the sum of all the data values

n = the number of data values **in the sample**

Pronounced "x-bar"

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Sample size

Observed values

The Population Mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

where μ = the **population** mean
(the Greek letter "*mu*")

N = the number of data values
in the population

Calculating The Mean

Example:

Suppose a sample of size $n = 5$ gives the following values:

6.2 7.1 4.8 9 3.3

The sample mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{6.2 + 7.1 + 4.8 + 9 + 3.3}{5} = \frac{30.4}{5} = 6.08$$

Note: units = units of measurement of the original data

The Weighted Mean

A **weighted mean** allows us to assign more weight to certain values and less weight to others

Think of computing GPA: the weights are the number of credit hours earned for each grade

The Weighted Mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

where w_i = the weight for each data value x_i

$\sum_{i=1}^{n} w_i$ = the sum of all the weights

The Weighted Mean: Example

Suppose you have summarized the number of days you worked each week for a sample of n = 20 weeks into the table and are asked to find the mean number of days you work every week:

# of days worked	Frequency
3	4
4	7
5	6
6	3

- Find the sample mean using direct calculation (add 3 four times, 4 seven times and so on and divide it by n = 20) or using weighted mean formula:
  - The frequencies represent the weights in the formula

The Weighted Mean: Example

# of days worked	Frequency
3	4
4	7
5	6
6	3

Values, x_i *Weights,* w_i

The weighted mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{4 \cdot 3 + 7 \cdot 4 + 6 \cdot 5 + 3 \cdot 6}{4 + 7 + 6 + 3}$$

$$= \frac{12 + 28 + 30 + 18}{20} = \frac{88}{20} = 4.4 \text{ days}$$

In words, you worked 4.4 days a week on average

Advantages and Disadvantages of Using the Mean to Summarize Data

Advantages

- Simple to calculate
- Summarizes the data with a single value

Disadvantages

- With only a summary value we lose information about the original data
  - Sample 1 with $n = 3$: 999, 1000, 1001 $\longrightarrow$ $\bar{x} = 1000$
  - Sample 2 with $n = 3$: 0, 1000, 2000 $\longrightarrow$ $\bar{x} = 1000$
  - Just knowing the mean does not help us know what the underlying data looks like
- The value of the mean is sensitive to **outliers** (values that are much higher or lower than most of the data)

The Median

The **median** is the value in the data set for which half the observations are higher and half are lower

Think of a value in the middle when the data items are arranged in ascending order

Rule of Thumb (to find the median):

1. When there are an **odd number** of data values, the median is always the middle value (in the **SORTED** data set)
2. When there are an **even number** of data values, the median is an average between the two middle values (in the **SORTED** data set)

The Median: Examples

Data 1: 27 21 27 34 45 50 28

Sample size: $n = 7$ (odd number!)

1. Sort the data: 21 27 27 28 34 45 50
2. The median is in the fourth position of the **sorted** data

21 27 27 28 34 45 50

Data 2: 157 145 170 204 209 182

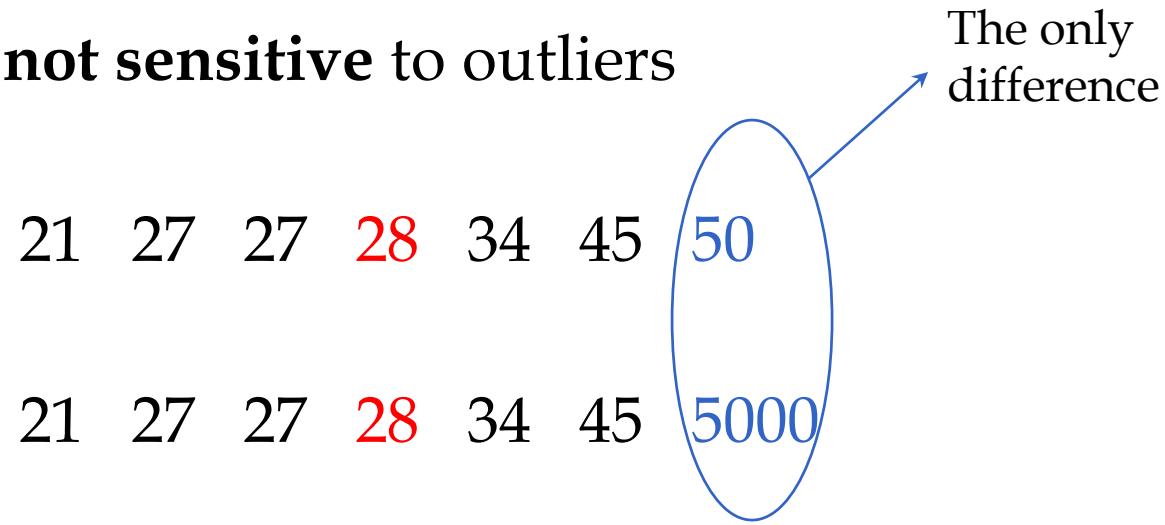
Sample size: $n = 6$ (even number!)

1. Sort the data: 145 157 170 182 204 209
2. Median is an average of the values in the third and fourth positions of the **sorted** data: (170 + 182)/2 = 176

145 157 170 182 204 209

The Median: Useful Notes

The median is **not sensitive** to outliers



The median is 28 in both datasets!

The Mode

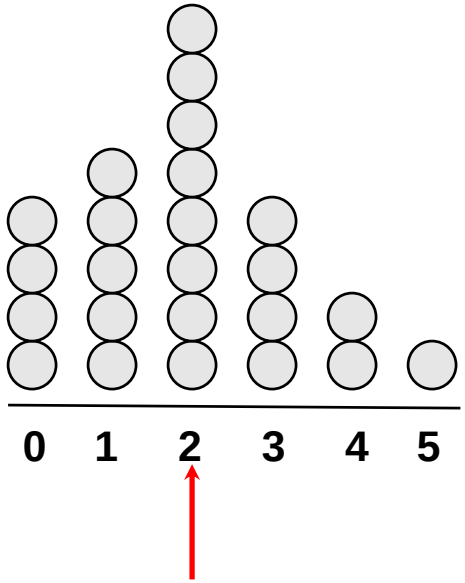
The **mode** is the value that appears most often in a data set (value that occurs with the highest frequency)

- If no data value or category repeats more than once, then we say that the mode does not exist
- More than one mode can exist if two or more values tie for most frequent
  - If the data have exactly two modes, the data are bimodal
  - If the data have more than two modes, the data are multimodal

The Mode: Example (Quantitative Data)

Number of children per family in a sample of 24 families:

0,0,0,0,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,4,5

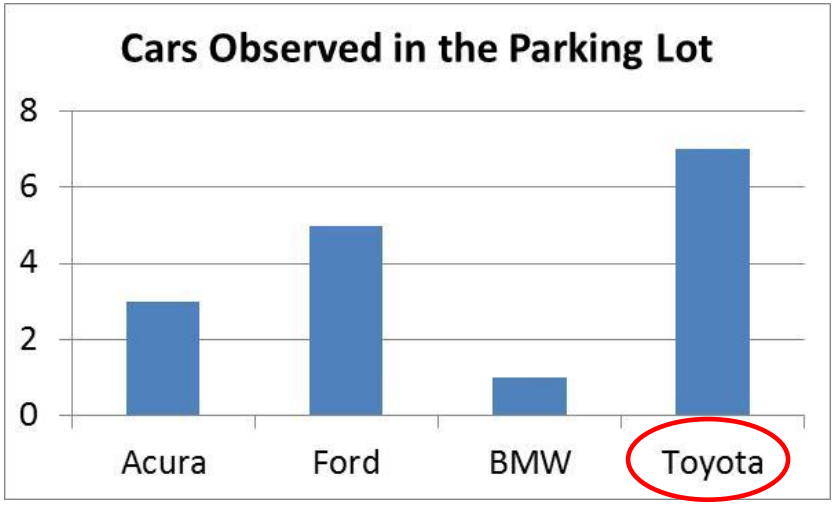


Number of children	Frequency
0	4
1	5
2	8
3	4
4	2
5	1

The value that appears most often is 2 (occurs 8 times), so the mode = 2 children

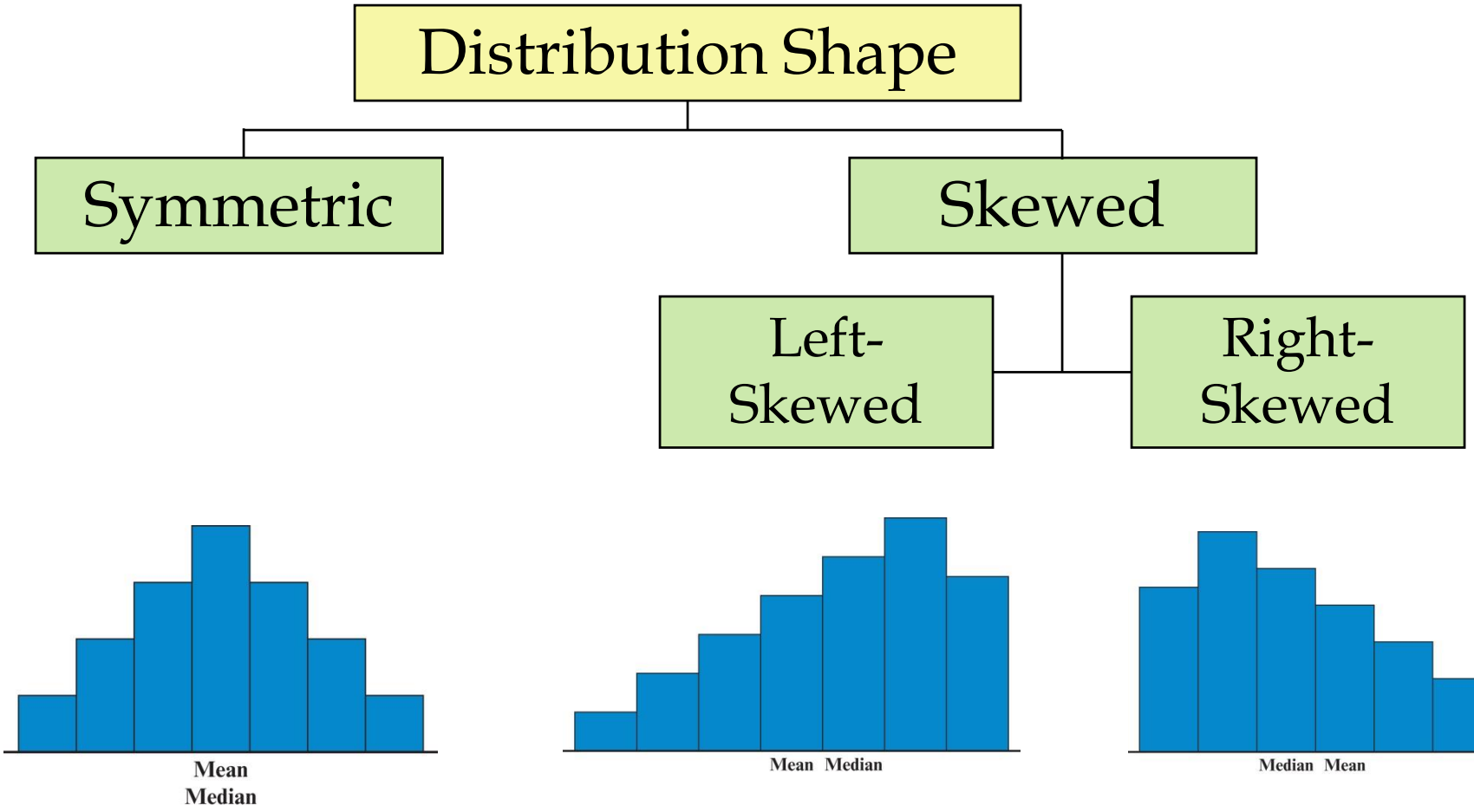
The Mode: Example (Qualitative Data)

Car Model	# of Cars
Acura	3
Ford	5
BMW	1
Toyota	7



The car that appears most often is Toyota (occurs 7 times), so the mode is the *Toyota model*

The Shapes of Frequency Distributions



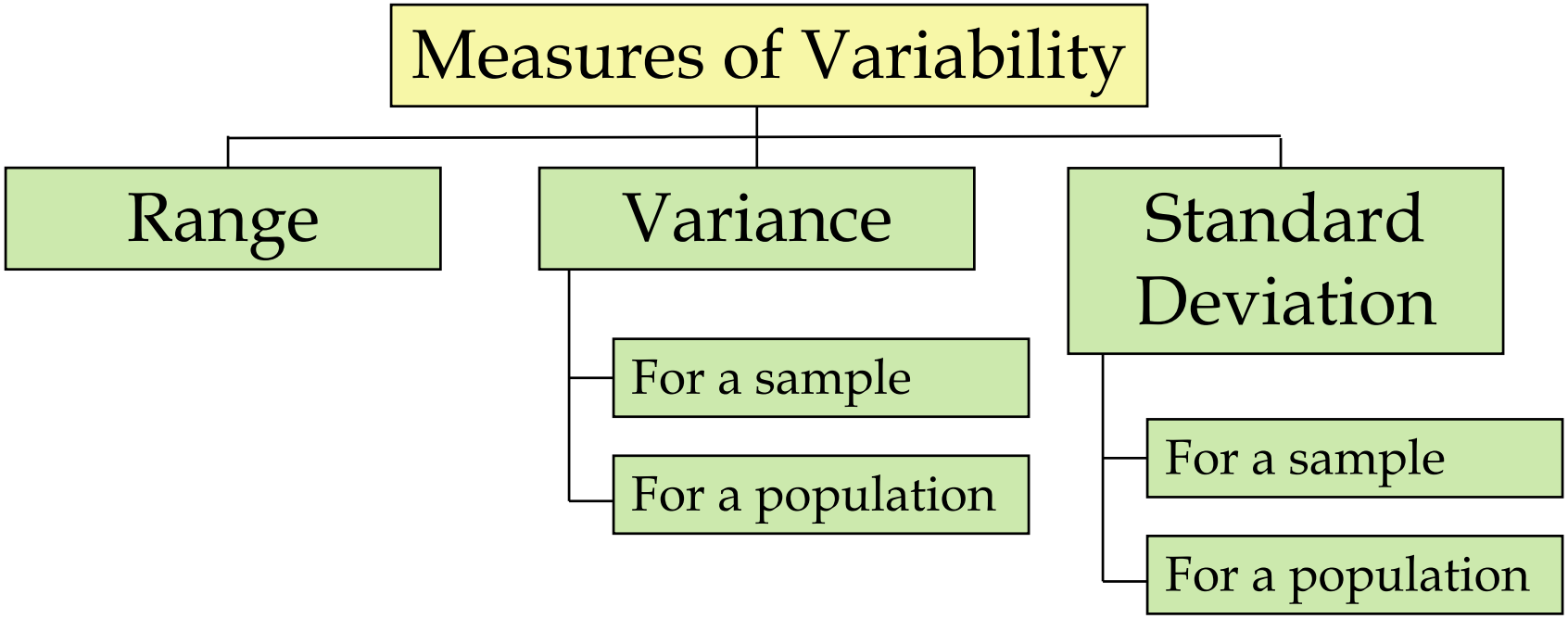
Mean = Median

Mean < Median

Mean > Median

Measures of Variability

Measures of variability show how much spread is present in the data



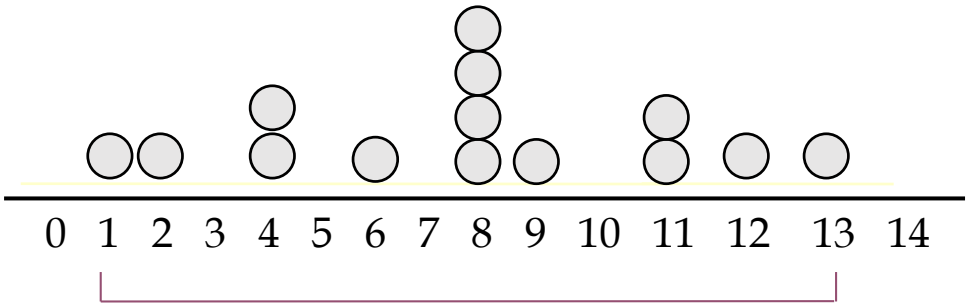
The Range

The simplest measure of variation

Difference between the highest value and the lowest value in a data set:

Range = Highest value – Lowest Value

Example: 1, 2, 4, 4, 6, 8, 8, 8, 8, 9, 11, 11, 12, 13



Range = 13 - 1 = 12

The Range

Advantages

- Easy to calculate and understand

Disadvantages

- Based on two numbers in the data set and ignores the way in which the data are distributed
- Sensitive to outliers:

Example:

1, 2, 4, 4, 6, 8, 8, 8, 8, 9, 11, 11, 12, 13 ⟶ Range = 12

1, 2, 4, 4, 6, 8, 8, 8, 8, 9, 11, 11, 12, 1000 ⟶ Range = 999

Change in one value causes a dramatic change in the range!

⇒ The range does not accurately reflect the overall variability of the data

The Sample Variance

The **sample variance** is denoted by s^2 and is the average of the squared differences between each data value and the mean

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

where $\bar{x}$ = the sample mean

n = sample size

$(x_i - \bar{x})$ = the difference between each data value and the sample mean

Note: units = squared units of measurement of the original data

The Sample Standard Deviation (SD)

The **sample standard deviation** is the square root of the sample variance

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

Note: units = units of measurement of the original data

Population Variance and Standard Deviation

Used when the data set represents an entire population rather than a sample from a population

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

Population Standard Deviation $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$

where μ = the **population** mean

N = **population** size

$(x_i - \mu)$ = difference b/w each data value and the population mean

The Coefficient of Variation

The standard deviation is affected by the scale of the data

- When sample means are different, comparing SDs can be misleading

The **coefficient of variation, *CV***, measures the SD in terms of its percentage of the mean and indicates how large the SD is in relation to the mean

- A high *CV* indicates high variability relative to the mean
- A low *CV* indicates low variability relative to the mean

The Coefficient of Variation

The **sample coefficient of variation**:

$$CV = \frac{s}{\bar{x}} \times 100$$

where s = the sample SD

$\bar{x}$ = the sample mean

The **population coefficient of variation**:

$$CV = \frac{\sigma}{\mu} \times 100$$

where σ = the population SD

μ = the population mean

Coefficient of Variation, Example

Stock A:

Average price last year = $50

Standard deviation = $4

Stock B:

Average price last year = $100

Standard deviation = $5

Coefficient of Variation:

Stock A:

$$CV = \frac{s}{\bar{x}} \times 100 = \frac{\$4}{\$50} \times 100 = 8\%$$

Stock B:

$$CV = \frac{s}{\bar{x}} \times 100 = \frac{\$5}{\$100} \times 100 = 5\%$$

Although stock B had a larger SD, the CV is smaller for this stock meaning that stock B had more consistent prices

The *z*-Score

The ***z*-score** identifies the number of standard deviations a particular value is from the mean of its distribution

A *z*-score has no units

The *z*-score is - **zero** for values equal to the mean

- **positive** for values above the mean

- **negative** for values below the mean

A data value that has a *z*-scores above +3 or below -3 is categorized as an **outlier** (has a value far from the mean)

The z-Score

The **sample** z-score

$$z = \frac{x - \bar{x}}{s}$$

where s = the sample SD

$\bar{x}$ = the sample mean

x = the data value of interest

The **population** z-score

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

where σ = the population SD

μ = the population mean

x = the data value of interest

z-Score: Example

Table 3.14 | **Calories for Various Fast-Food Hamburgers**

HAMBURGER TYPE	RESTAURANT	CALORIES
Cheeseburger	McDonald's	300
Single with Everything	Wendy's	430
Big Mac	McDonald's	540
Whopper	Burger King	670
Bacon Cheeseburger	Sonic	780
Baconator	Wendy's	840
Triple Whopper with Cheese	Burger King	1,230
2/3 lb. Monster Thickburger	Hardee's	1,420
	Sample Mean	**776.3**
	Sample Standard Deviation	**385.1**

Question: How far is 670 from the sample mean of 776.3 (in standard deviation increments)?

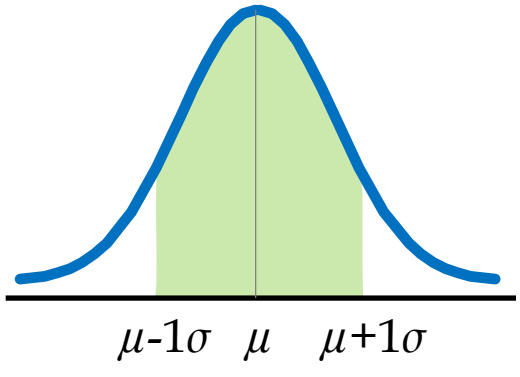
$$z = \frac{x - \bar{x}}{s} = \frac{670 - 776.3}{385.1} = -0.276$$

Answer: 670 is 0.276 standard deviations below the mean

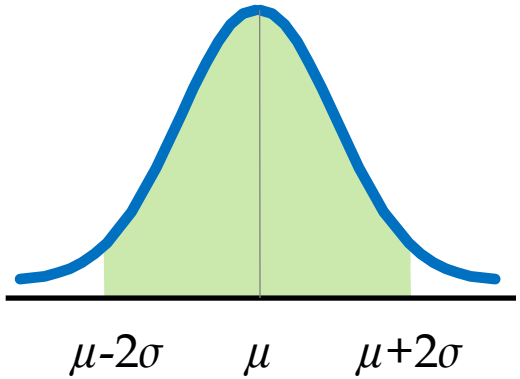
The Empirical Rule

According to the **empirical rule,** if a distribution follows a bell-shaped, symmetrical curve centered around the mean, we would expect:

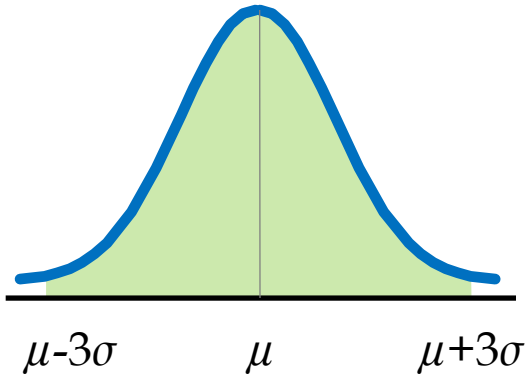
Approximately **68%** of the values to fall within ± 1 standard deviations from the mean



Approximately **95%** of the values to fall within ± 2 standard deviations from the mean

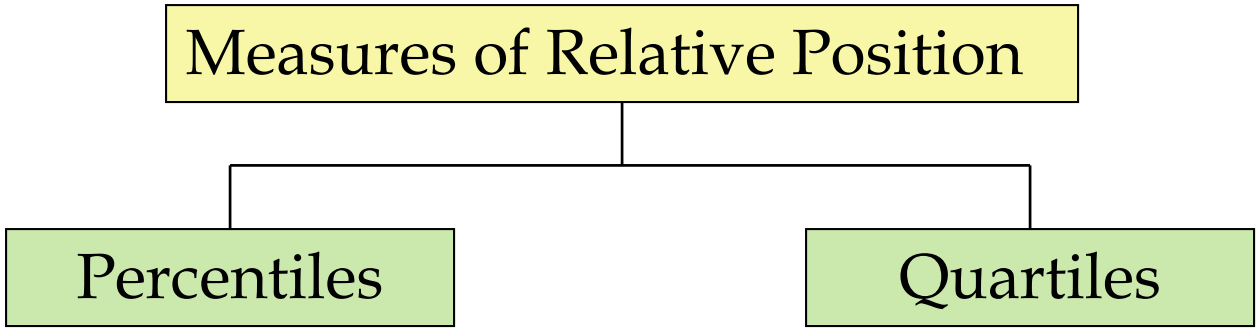


Approximately **99.7%** of the values to fall within ± 3 standard deviations from the mean



Measures of Relative Position

Measures of relative position compare the position of one value in relation to other values in the data set



Percentiles

The p^{th} percentile divides a data set into two parts:

- Approximately p percent of the observations have values less than the pth percentile
- Approximately $(100 - p)$ percent of the observations have values greater than the pth percentile

For example: suppose we have miles per gallon (MPG) recorded for a sample of 12 cars. Based on this data we found that 60^{th} percentile = 31.1 MPG

$\Rightarrow$ 60% of cars in the sample have MPG < 31.1

Quartiles

Quartiles split the ranked data into 4 equal groups:

- The first quartile (Q_1) is the value that constitutes the 25^{th} percentile
- The second quartile (Q_2) is the value that constitutes the 50^{th} percentile
  - Second quartile (the 50^{th} percentile) = Median
- The third quartile (Q_3) is the value that constitutes the 75^{th} percentile

Excel Exercise >>

Descriptive Statistics for Characterizing Two Variables

In many instances, we are interested in the relationship between two variables. For example:

- Does immigration cause lower wages?
- Does advertising increase sales?
- Does income vary with education?

Descriptive Statistics to Summarize the Relationship Between Two Variables:

- Scatter Plots (Chapter 2, Section 2.6)
- Numerical Measures of Association Between Two Variables (Chapter 3, Section 3.6)

Scatter Plots

A **scatter plot** is a graphical tool used to determine if two variables are related. Each point represents a pair of known values of the two variables for one observation

In the relationship, we usually distinguish between dependent and independent variables

The **dependent** variable:

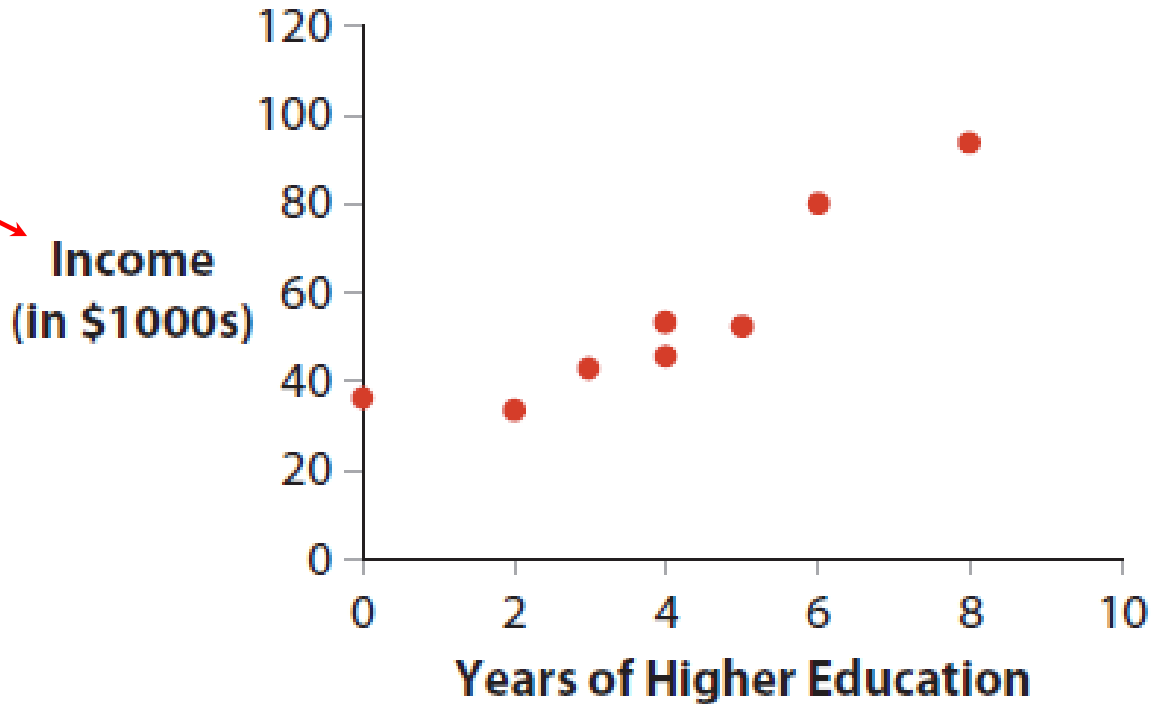
- influenced by changes in the independent variable;
- denoted by *y*;
- placed on the vertical axis.

The **independent** variable:

- used to explain changes in the dependent variable;
- denoted by *x*;
- placed on the horizontal axis.

Scatter Plots

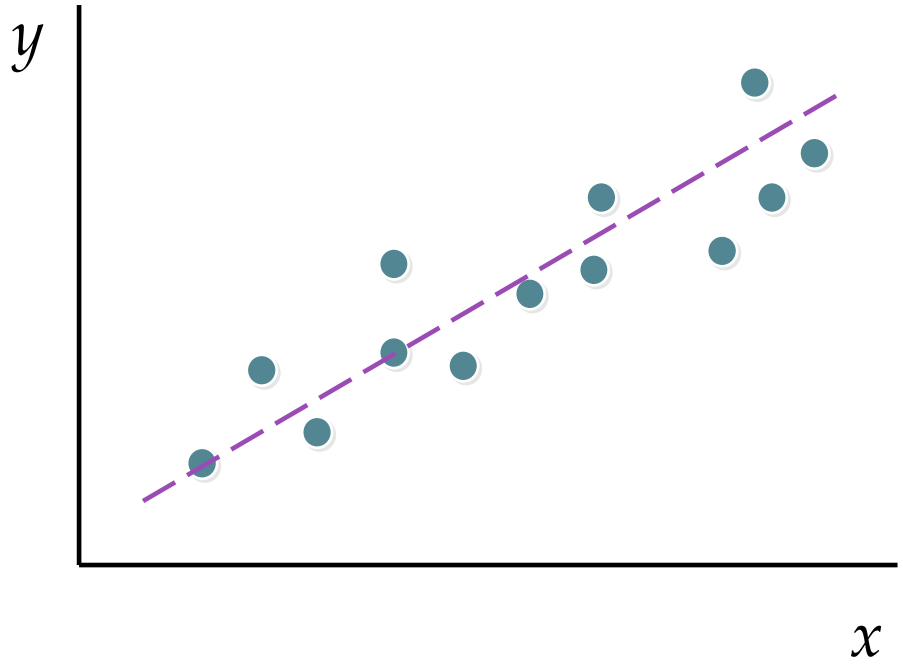
Dependent variable (y-axis)



Independent variable (x-axis)

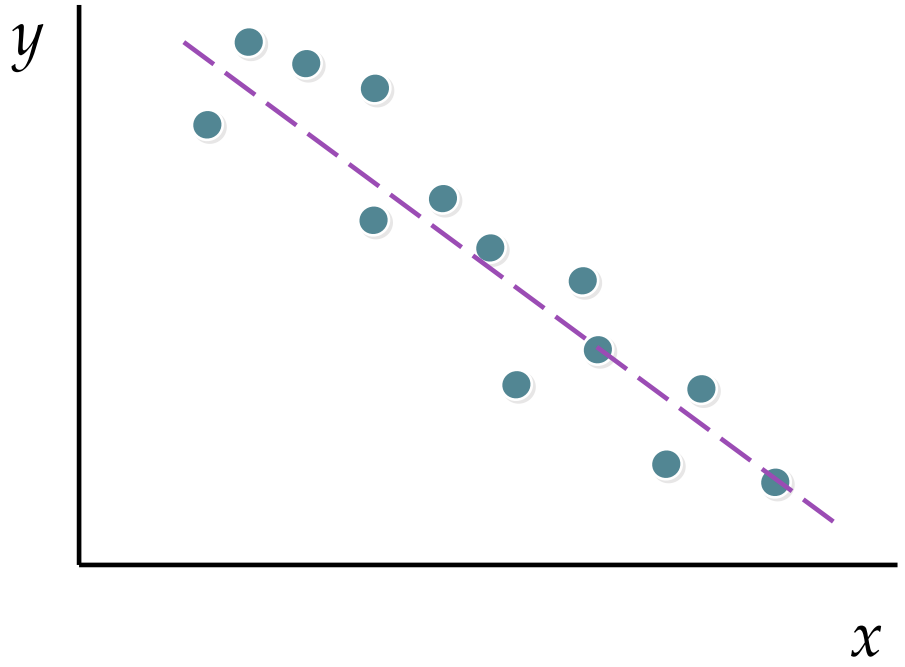
Scatter Plots

Positive Relationship: points are clustered together along a *line* with a *positive* slope



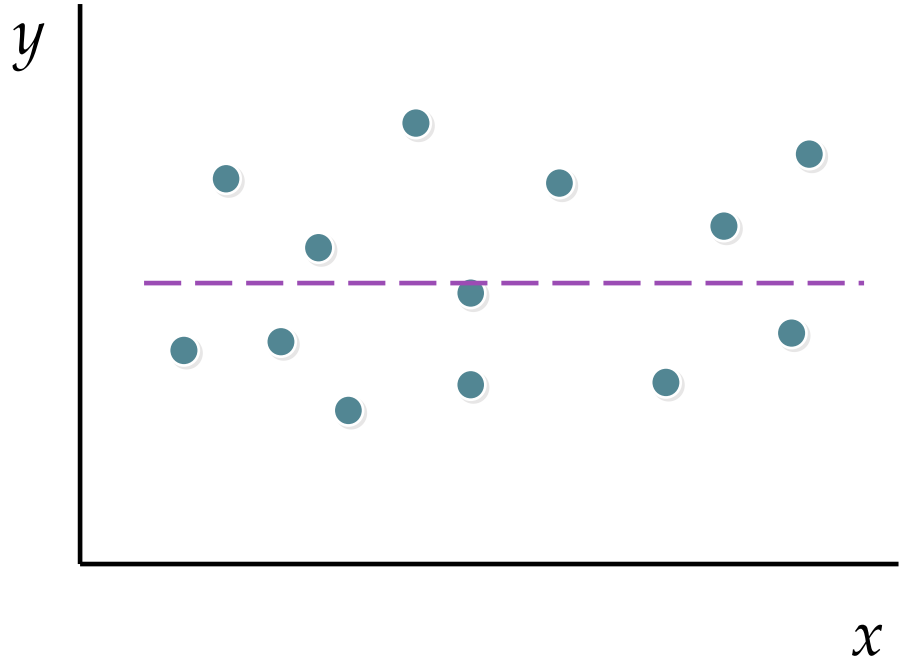
Scatter Plots

Negative Relationship : points are clustered together along a *line* with a *negative* slope

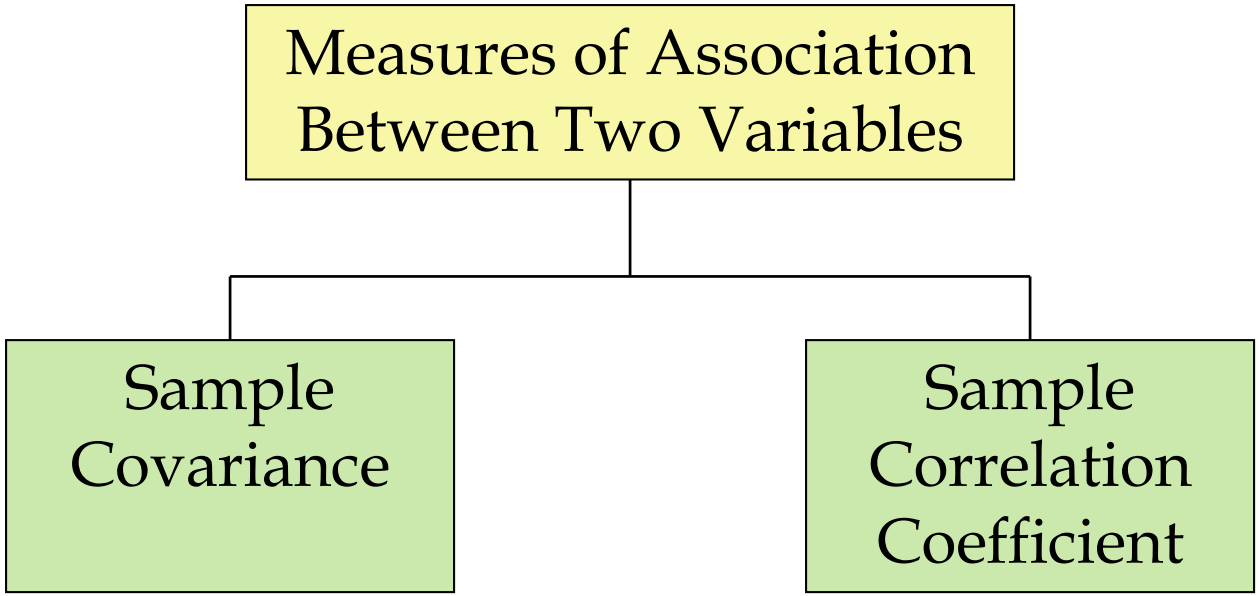


Scatter Plots

No Relationship: data are randomly scattered with no discernible pattern



Measures of Association Between Two Variables



Sample Covariance

The **sample covariance**, s_{xy}, measures the **direction** of the linear relationship between two variables

- A relationship is linear if the scatter plot has a straight-line pattern

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

where $\bar{x}$ = the sample mean of the x variable

$\bar{y}$ = the sample mean of the y variable

$(x_i - \bar{x})$ = the difference between each data value and the sample mean for the x variable

$(y_i - \bar{y})$ = the difference between each data value and the sample mean for the y variable

n = the sample size

Sample Covariance

The sample covariance, s_{xy}, measures the **direction** of the linear relationship between two variables:

- A positive value implies a positive linear relationship

  (as one variable **increases**, the second variable also tends to **increase**)

- A negative value implies a negative linear relationship

  (as one variable **increases**, the second variable tends to **decrease**)

- The covariance is zero if y and x have no linear relationship
  - Note that covariance is sensitive to the units of measurement. Therefore, values close to zero may not indicate a weak relationship between variables.

Correlation Coefficient

The **sample correlation coefficient**, r_{xy}, measures both the **strength** and **direction** of the linear relationship between two variables

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where s_{xy} = the sample covariance between variables x and y
s_x = the sample standard deviation for the x variable
s_y = the sample standard deviation for the y variable

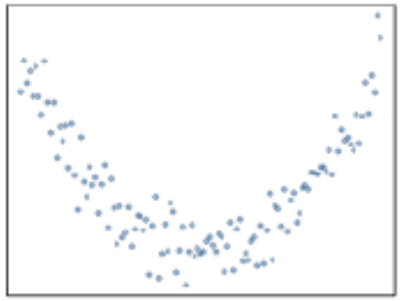
The Correlation Coefficient

The sample correlation coefficient, r_{xy}, indicates both the **strength** is and **direction** of the linear relationship between the independent and dependent variables:

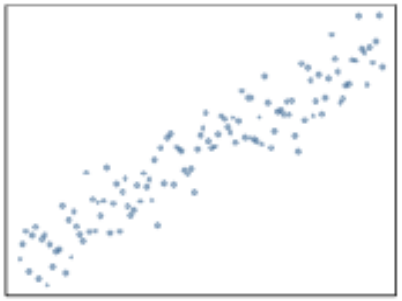
- The values of *r* range from -1.0, a **strong** negative relationship, to +1.0, a **strong** positive relationship
- When $r = 0$, there no linear relationship between variables x and y

Practice Question, I

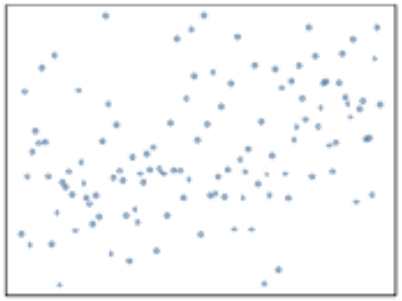
Which of the following shows high correlation?



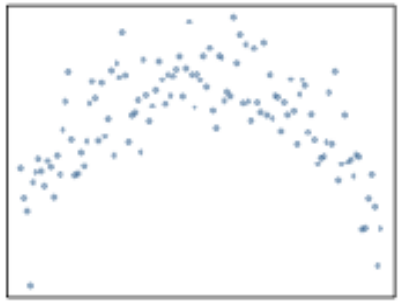
(1)



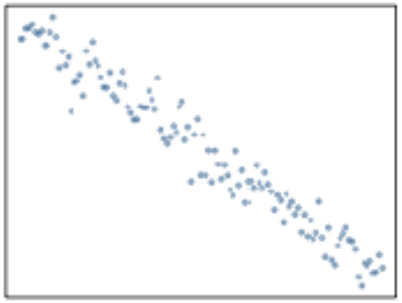
(2)



(3)



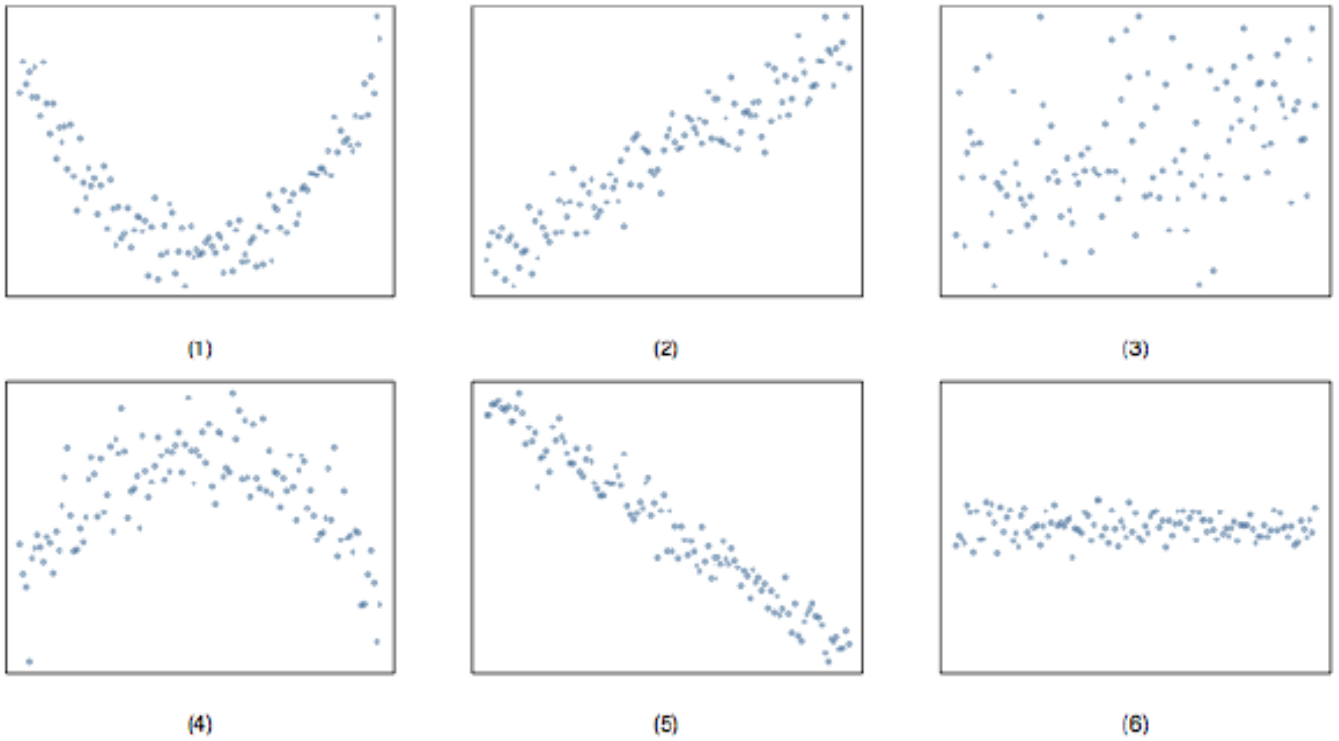
(4)



(5)

Practice Question, II

Which of the following has correlation $r = 0.93$?



Excel Exercise >>

Excel Time: Exercise 3.42 (Modified)

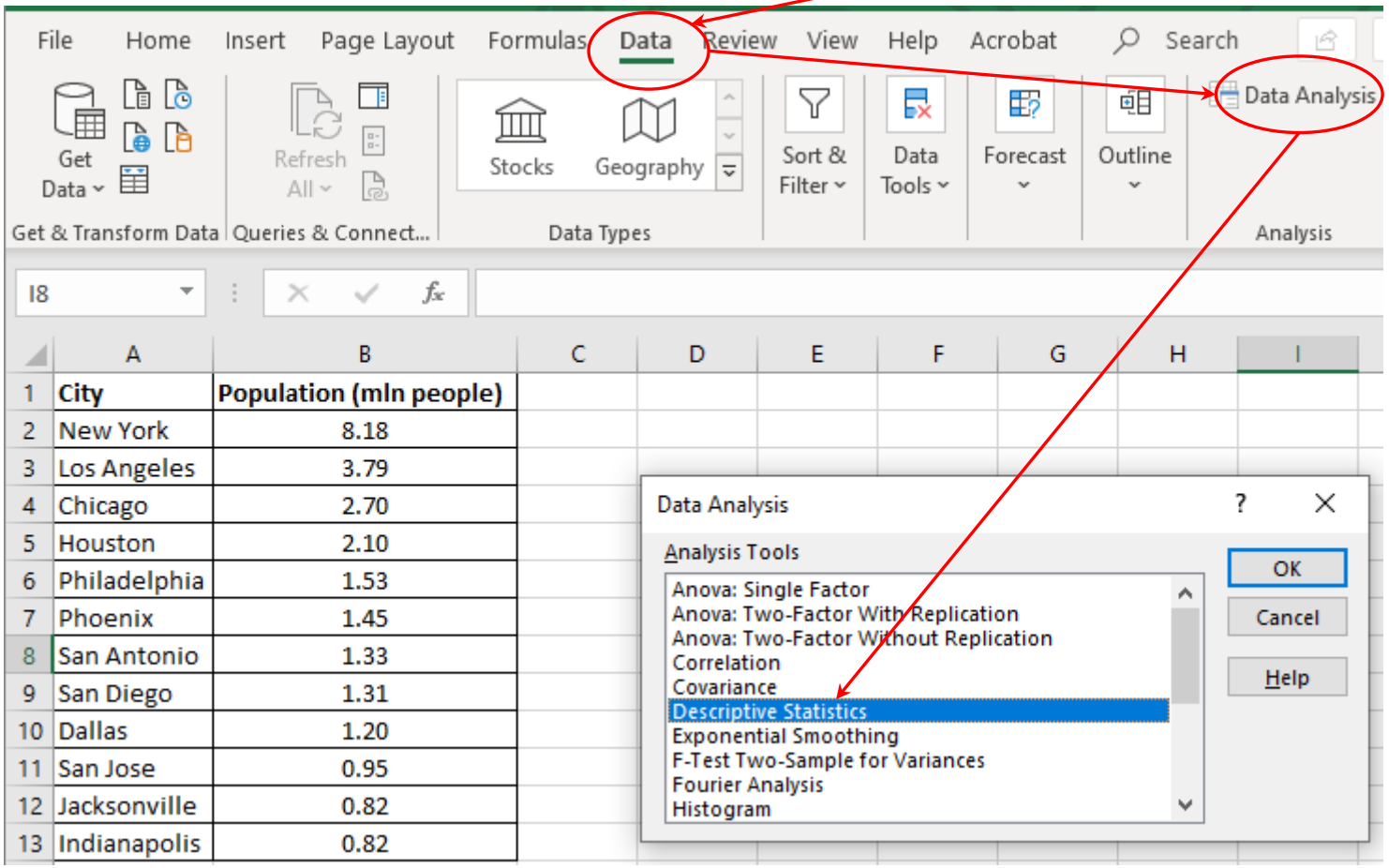
Excel file **city_populations.xlsx (***Excel Files* → *Ch 03***)** lists 12 largest U.S. cities from the 2010 Census. Using the data in the file:

1. Determine descriptive statistics using Data Analysis add-in.
2. Using Excel functions, find descriptive statistics for this sample: mean, median, mode, range, variance, standard deviation. Provide units of measurement for these values.
3. Describe the shape of this distribution.
4. Additionally, find 60th and 85th percentile, 1st quartile and the coefficient of variation. Interpret calculated values.
5. Calculate the *z*-scores for the first and the last city in the table.
6. Are there any outliers in the sample?
7. What measure of central tendency would best describe this data?

Back <<

Excel Time: Data Analysis for Descriptive Statistics

Descriptive statistics for the sample can be obtained using **Excel functions** or the **Data Analysis** add-ins

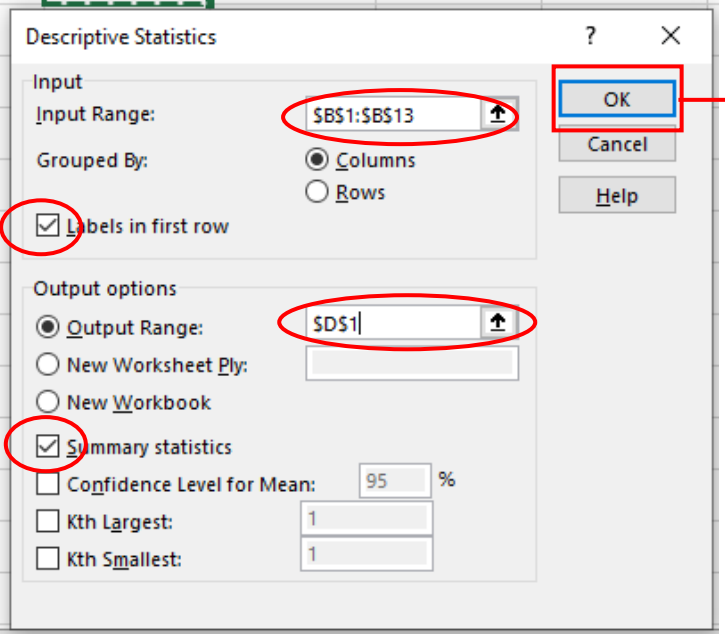


Choose the "Descriptive Statistics" option…

Excel Time: Data Analysis for Descriptive Statistics

Descriptive Statistics output:

	A	B
1	**City**	**Population (mln people)**
2	New York	8.18
3	Los Angeles	3.79
4	Chicago	2.70
5	Houston	2.10
6	Philadelphia	1.53
7	Phoenix	1.45
8	San Antonio	1.33
9	San Diego	1.31
10	Dallas	1.20
11	San Jose	0.95
12	Jacksonville	0.82
13	Indianapolis	0.82



D	E
Population (mln people)	
Mean	2.18167
Standard Error	0.59972
Median	1.39
Mode	0.82
Standard Deviation	2.0775
Sample Variance	4.31602
Kurtosis	7.07345
Skewness	2.55869
Range	7.36
Minimum	0.82
Maximum	8.18
Sum	26.18
Count	12

Note: numbers may be reported in exponential notation (e.g. 1E+5)

Excel Time: Functions to Calculate the Mean, Median and Mode

Descriptive statistics for the sample can be obtained using **Excel functions** or the **Data Analysis** add-ins

	A	B	C	D	E	F
1	**City**	**Population (mln people)**		**Statistics**	**Value**	**Formula**
2	New York	8.18		**Mean**	2.18	=AVERAGE(B2:B13)
3	Los Angeles	3.79		**Median**	1.39	=MEDIAN(B2:B13)
4	Chicago	2.70		**Mode**	0.82	=MODE.SNGL(B2:B13)

Note: Not all the data is displayed on the screenshot

- If there is no mode, Excel displays "#N/A"
- If there is more than one mode, Excel will incorrectly display only one mode, so use caution
- Excel will not identify the mode for qualitative data (though it can be identified)

Excel Time: Functions to Calculate the Variance and the SD

The Excel functions for the **sample** variance and SD are:

=VAR.S(*data values*)

=STDEV.S(*data values*)

the letter *S* in VAR.S and STDEV.S indicates "sample"

The functions for the **population** variance and SD are:

=VAR.P(*data values*)

=STDEV.P(*data values*)

the letter *P* in VAR.P and STDEV.P indicates "population"

	A	B	C	D	E	F
1	**City**	**Population (mln people)**		**Statistics**	**Value**	**Formula**
2	New York	8.18		**Sample Standard Deviation**	2.08	=STDEV.S(B2:B13)
3	Los Angeles	3.79		**Sample Variance**	4.32	=VAR.S(B2:B13)

Excel Time: Percentiles

Excel calculates percentiles using the **PERCENTILE.EXC ()** function:

=PERCENTILE.EXC(*array*, *k*)

where:

array = The data range of interest

k = The percentile of interest between 0 and 1 exclusive

Excel uses a sophisticated technique to calculate percentiles. If you calculate percentile by hand using the methodology from the textbook, your percentile may differ from that that calculated by Excel (not necessarily)

Excel Time: Quartiles

Quartiles can be found in Excel with the **QUARTILE.EXC()** function

=QUARTILE.EXC(*array*, *quart*)

where: *array* = The data range of interest

quart = 1, 2, or 3 (for the first, second, or third quartile)

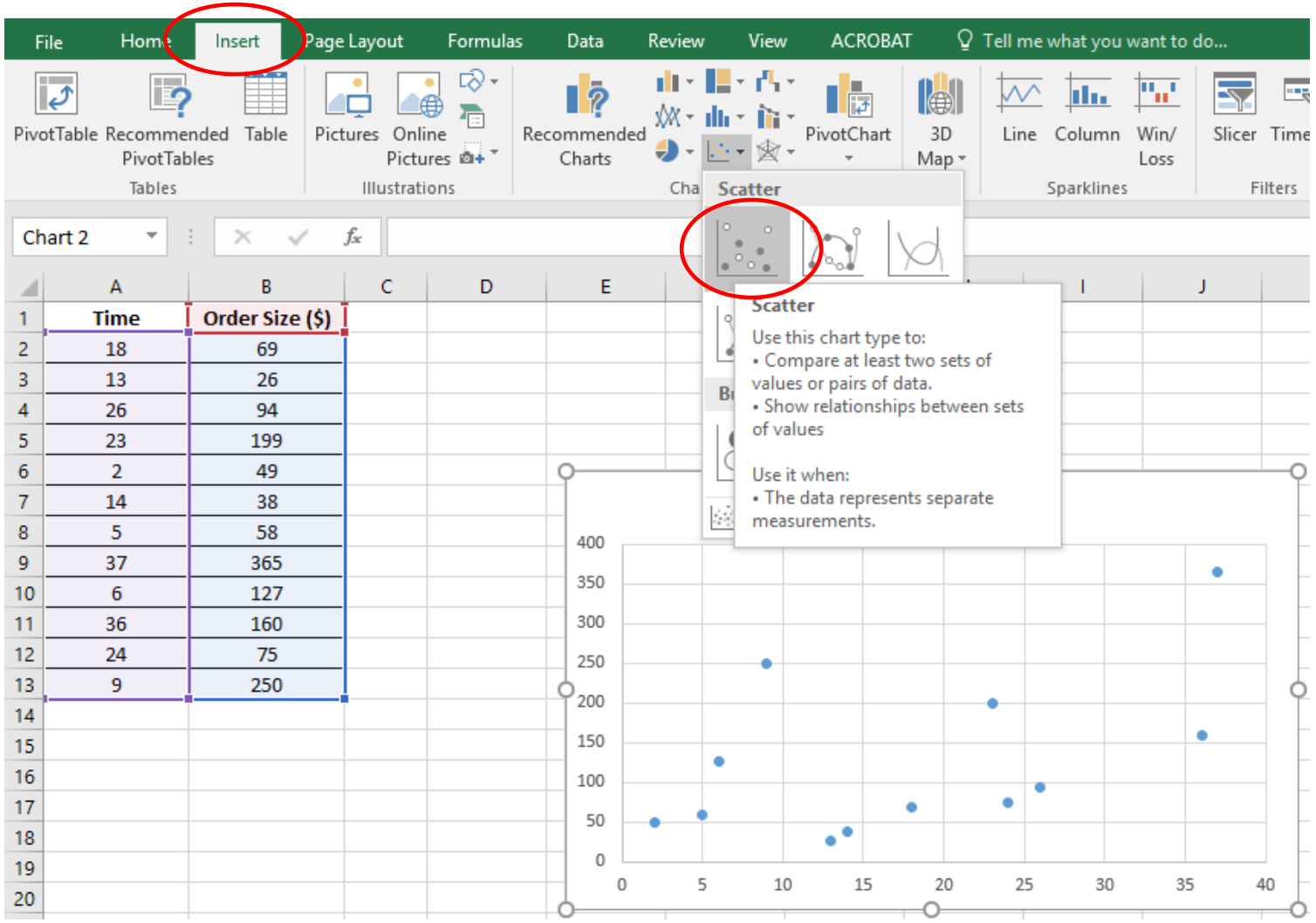
Excel Time: Exercise 2.28 (Modified)

Suppose the Internet retailer Buy.com would like to investigate the relationship between the amount of time in minutes a purchaser spends on its Web site and the amount of money he or she spends on an order. The Excel file **Buy_com.xlsx** (*Excel Files → Ch 02*) shows the data from a sample of 12 customers.

1. Construct a *scatter plot* using these data. Describe the relationship between the variables of interest.
2. Find the sample covariance and the correlation coefficient.
3. What can you conclude based on these values?

Back <<

Excel Time: Constructing a Scatter Plot



Excel Time: Functions to Calculate the Covariance and the Correlation

The Excel functions for the **sample** and **population** covariance are:

=COVARIANCE.**S** (*data array* 1, *data array* 2)

=COVARIANCE.**P** (*data array* 1, *data array* 2)

As with other formulas, the letter *S* indicates "sample" and the letter *P* indicates "population"

The function **CORREL** is used to calculate the correlation coefficient:

=CORREL (*data array* 1, *data array* 2)

22		Value	Formula
23	Covariance	617.8	=COVARIANCE.S(A2:A13,B2:B13)
24	Correlation	0.5	=CORREL(A2:A13,B2:B13)

Exercise 3.64 (Extra Practice)

The data in the file **market_cap.xlsx** (*Excel Files* → *Ch 03*) lists the 20 largest companies in the world in 2017 ranked their market capitalization (in billions of dollars). A company's market capitalization is defined as the number if its shares multiplied by the price per share.

1. Using Excel functions, find descriptive statistics for these data: mean, median, mode, range, variance, standard deviation, coefficient of variation, 30th and 80th percentile, and 3d quartile. Interpret these values and provide units of measurement.
2. Describe the shape of this distribution.
3. Calculate the z-score for Amazon and Bank of America. Interpret your findings.

Exercise 3.63 (Extra Practice)

Find the data on the number of wins per season for the New England Patriots and Detroit Lions NFL teams from 2008 through 2017 in the Excel file **NFL_wins.xlsx** (*Excel Files* → *Ch 03*).

Which team experienced more variability with games won per season?

Hint: to answer this question think about statistics you would want to compute

Exercise 3.57 (Extra Practice)

A country's fertility rate can have a major long-term impact on its economic health. Low fertility rates eventually cause the average age of the population to skew higher, making it more difficult to fund programs such as social security. The following data represents a sample showing the number of children from 10 families:

1 5 3 1 2 2 1 4 0 1

1. Calculate the range.
2. Calculate the variance.
3. Calculate the standard deviation.

Exercise 3.78 (Extra Practice)

Fair Isaac, the company that developed the current credit score model used by most lenders today, would like to examine the relationship between the age and credit score of an individual. Data in the Excel file **Fair_Isaac.xlsx** (*Excel Files* → *Ch 03*) records the credit scores and ages of 10 randomly selected people.

a. Calculate the sample covariance.

b. Calculate the correlation coefficient.

c. Describe the relationship between variables.