

Scientific Research Workshops - A Practical Perspective

Prepared by Dr. Mona Moustafa Elashry

Session 8





مجموعة جامعة
الجميع الذكية
للتعليم القابضة



“Statistical Analysis Part 1- Introduction”

Prepared by
Dr. Mona Moustafa Elashry



Session Outline

- Chapter 1: The Where, why, and How of Data Collection
- Chapter 2 : Graphs, Charts, and Tables – Describing Your Data
- Chapter 3 : Describing Data Using Numerical Measures
- Chapter 9: Introduction to Hypothesis Testing
- Chapter 14: Introduction to Linear Regression and Correlation Analysis
- Chapter 15: Multiple Regression Analysis and Model Building

Chapter 1- The Where, why, and How of Data Collection

1



Chapter Goals

After completing this chapter, you should be able to:

- Describe key data collection methods
- Know key definitions:
 - ◆ Population vs. Sample
 - ◆ Primary vs. Secondary data types
 - ◆ Qualitative vs. Quantitative data
 - ◆ Time Series vs. Cross-Sectional data
- Explain the difference between descriptive and inferential statistics
- Describe different sampling methods

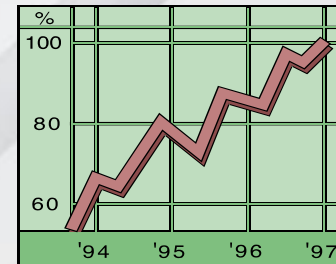
Tools of Business Statistics

- **Descriptive statistics**
 - Collecting, presenting, and describing data
- **Inferential statistics**
 - Drawing conclusions and/or making decisions concerning a population based only on sample data

Descriptive Statistics

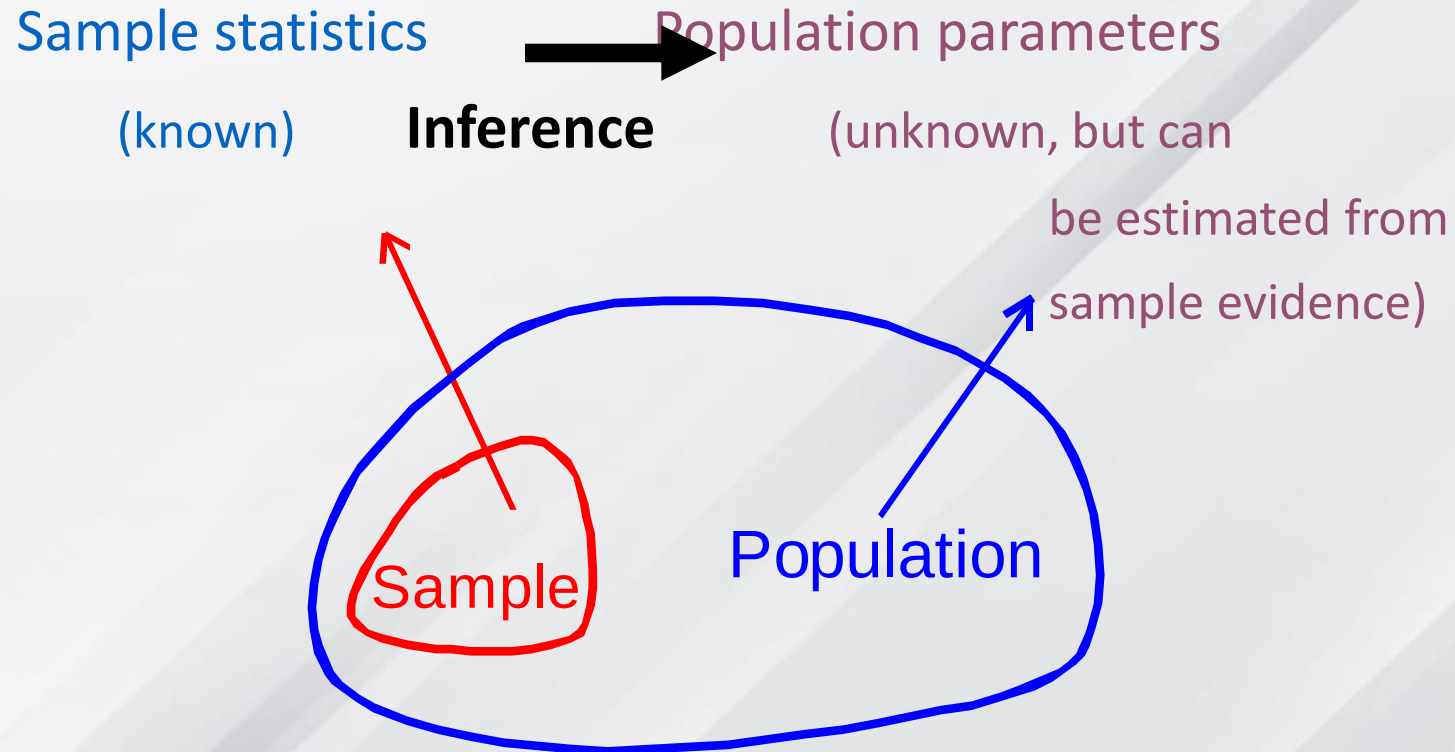
- **Collect data**
 - e.g., Survey, Observation, Experiments
- **Present data**
 - e.g., Charts and graphs
- **Characterize data**
 - e.g., Sample mean =

$$\frac{\sum x_i}{n}$$



Inferential Statistics

- Making statements about a population by examining sample results



Inferential Statistics

Drawing conclusions and/or making decisions concerning a **population based on **sample** results.**

- **Estimation**

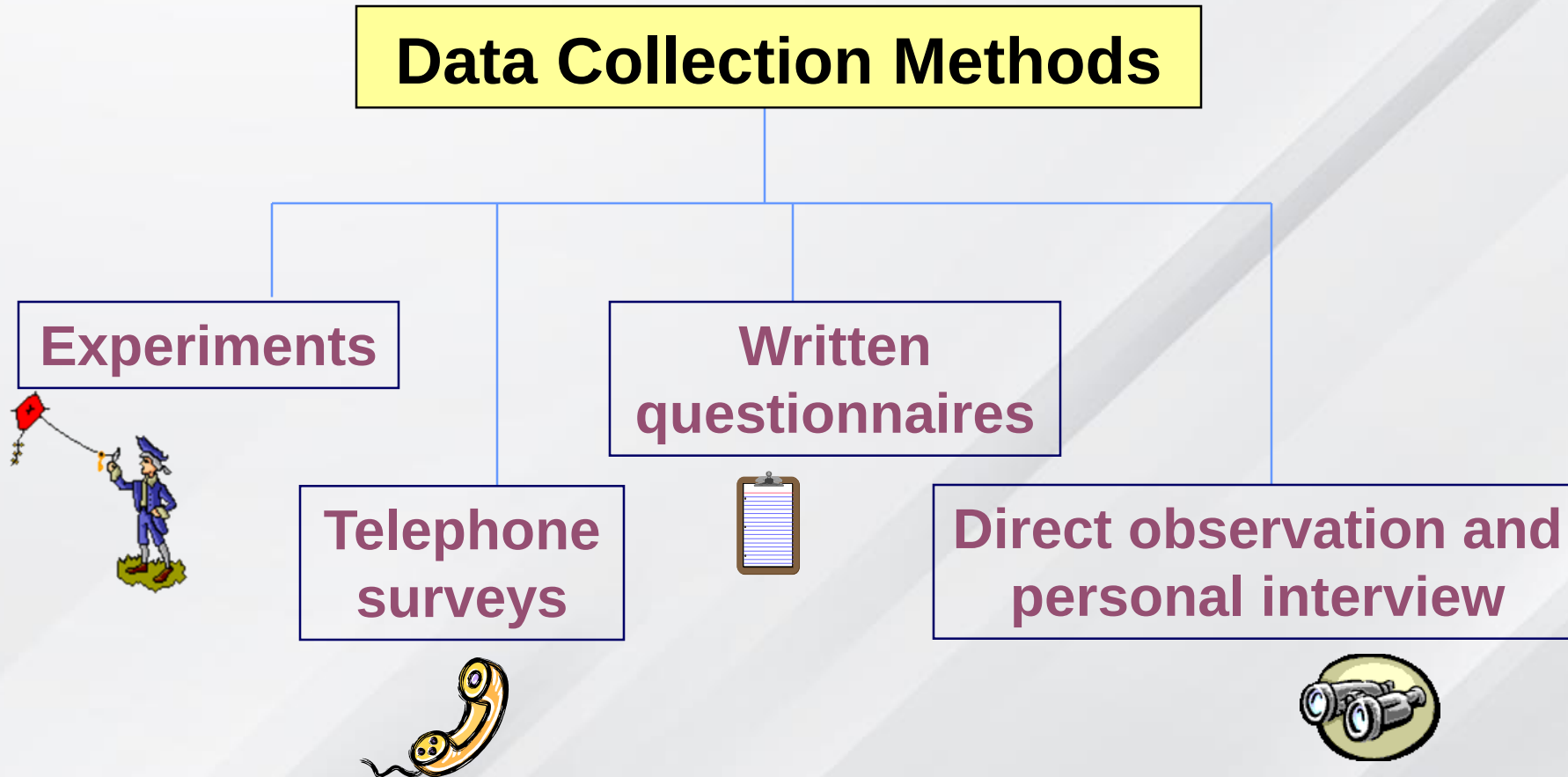
- e.g., Estimate the population mean weight using the sample mean weight

- **Hypothesis Testing**

- e.g., Use sample evidence to test the claim that the population mean weight is 120 pounds



Tools for Collecting Data



Survey Design Steps

- Define the issue
 - what are the purpose and objectives of the survey?
- Define the population of interest
- Develop survey questions
 - make questions clear and unambiguous
 - use universally-accepted definitions
 - limit the number of questions

Survey Design Steps

(continued)

- Pre-test the survey
 - pilot test with a small group of participants
 - assess clarity and length
- Determine the sample size and sampling method
- Select sample and administer the survey

Types of Questions

- **Closed-end Questions**

- Select from a short list of defined choices

Example: Major: ☐ business ☐ liberal arts
☐ science ☐ other

- **Open-end Questions**

- Respondents are free to respond with any value, words, or statement

Example: What did you like best about this course?

- **Demographic Questions**

- Questions about the respondents' personal characteristics

Example: Gender: ☐ Female ☐ Male

Populations and Samples

- A **Population** is the set of all items or individuals of interest

- **Examples:**
 - All likely voters in the next election
 - All parts produced today
 - All sales receipts for November

- A **Sample** is a subset of the population

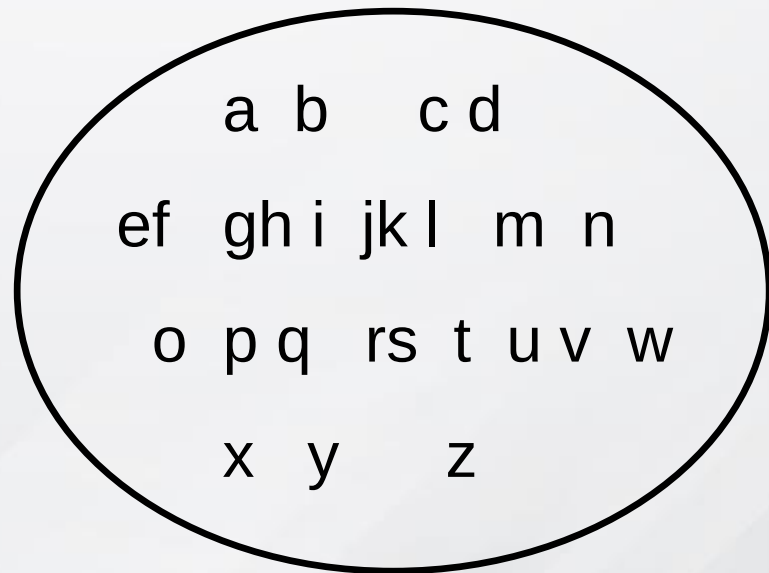
- **Examples:**
 - 1000 voters selected at random for interview
 - A few parts selected for destructive testing
 - Every 100th receipt selected for audit

Key Definitions

- A **population** is the entire collection of things under consideration
 - A **parameter** is a summary measure computed to describe a characteristic of the population
- A **sample** is a portion of the population selected for analysis
 - A **statistic** is a summary measure computed to describe a characteristic of the sample

Population vs. Sample

Population



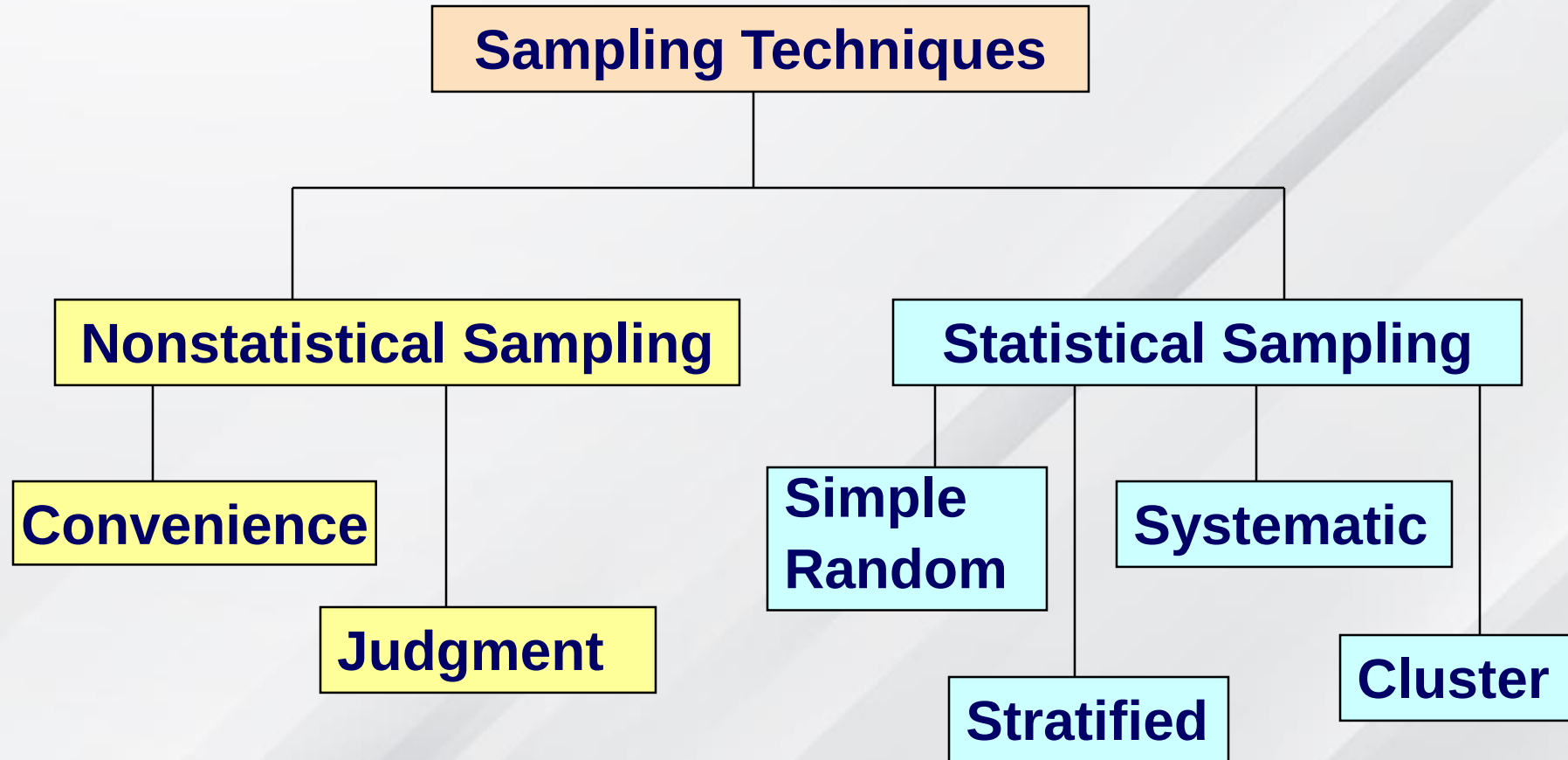
Sample



Why Sample?

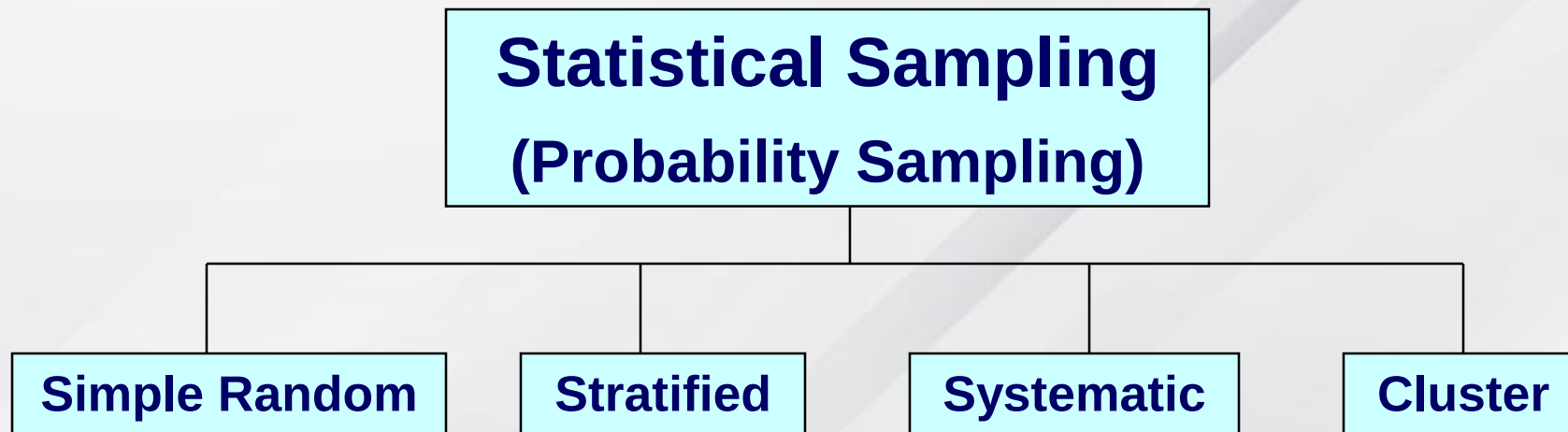
- Less time consuming than a census
- Less costly to administer than a census
- It is possible to obtain statistical results of a sufficiently high precision based on samples.

Sampling Techniques



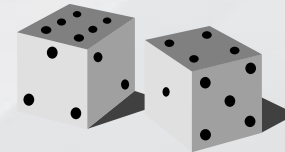
Statistical Sampling

- Items of the sample are chosen based on known or calculable probabilities



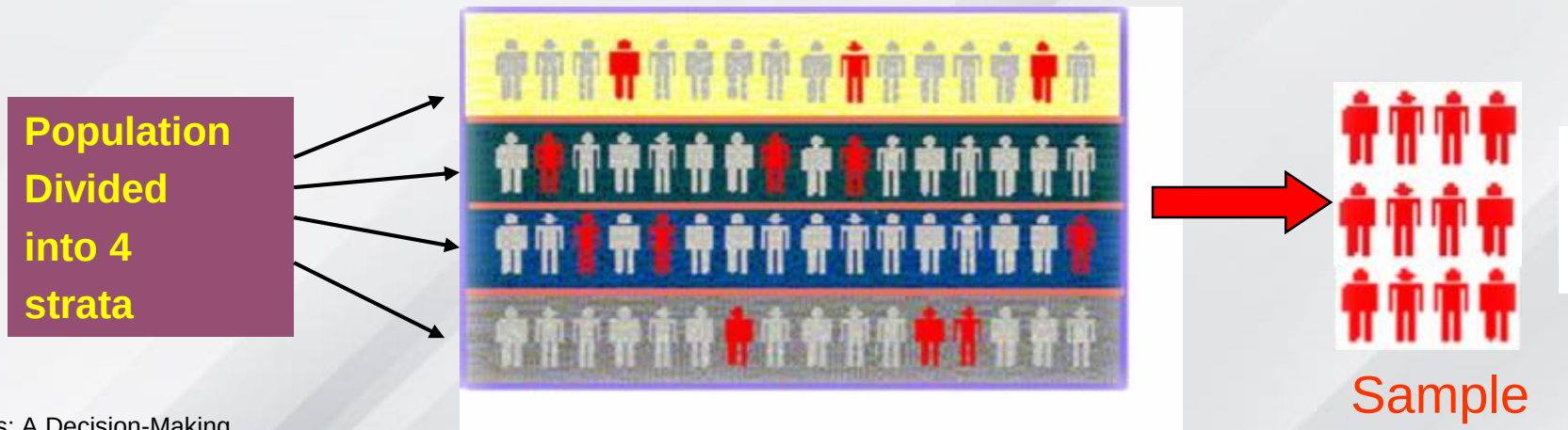
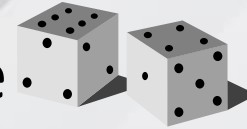
Simple Random Sampling

- Every possible sample of a given size has an **equal chance** of being selected
- Selection may be with replacement or without replacement
- The sample can be obtained using a table of random numbers or computer random number generator



Stratified Random Sampling

- Divide population into subgroups (called *strata*) according to some common characteristic
- Select a simple random sample from each subgroup
- Combine samples from subgroups into one



Systematic Random Sampling

- Decide on sample size: n
- Divide frame of N individuals into groups of k individuals: $k = N/n$
- Randomly select one individual from the 1st group
- Select every k^{th} individual thereafter

$N = 64$
 $n = 8$
 $k = 8$

First Group



Cluster Sampling

- Divide population into several “clusters,” each representative of the population
- Select a simple random sample of clusters
 - All items in the selected clusters can be used, or items can be chosen from a cluster using another probability sampling technique

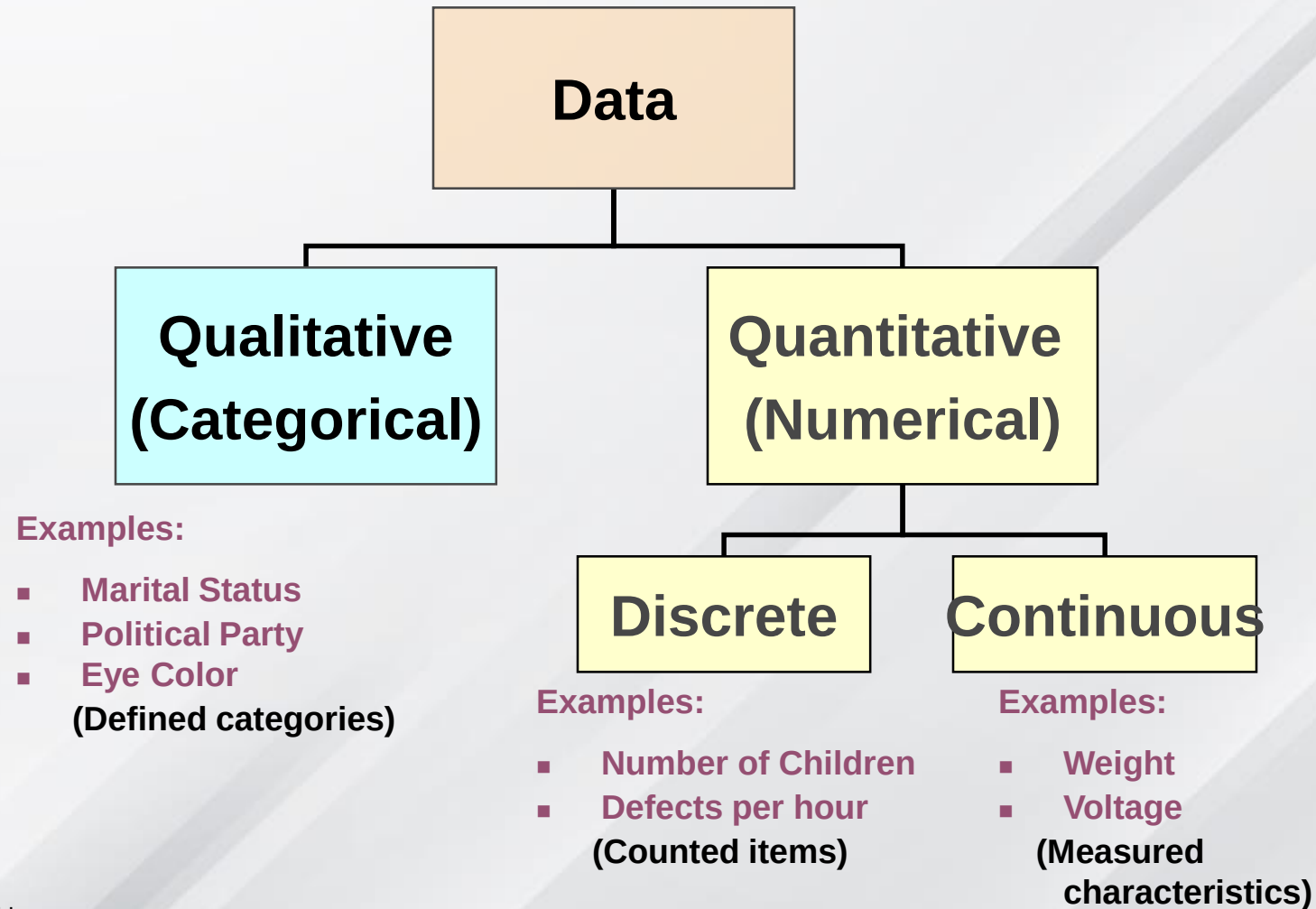


**Population
divided into
16 clusters.**



**Randomly selected
clusters for sample**

Data Types



Data Types

- **Time Series Data**
 - Ordered data values observed over time
- **Cross Section Data**
 - Data values observed at a fixed point in time

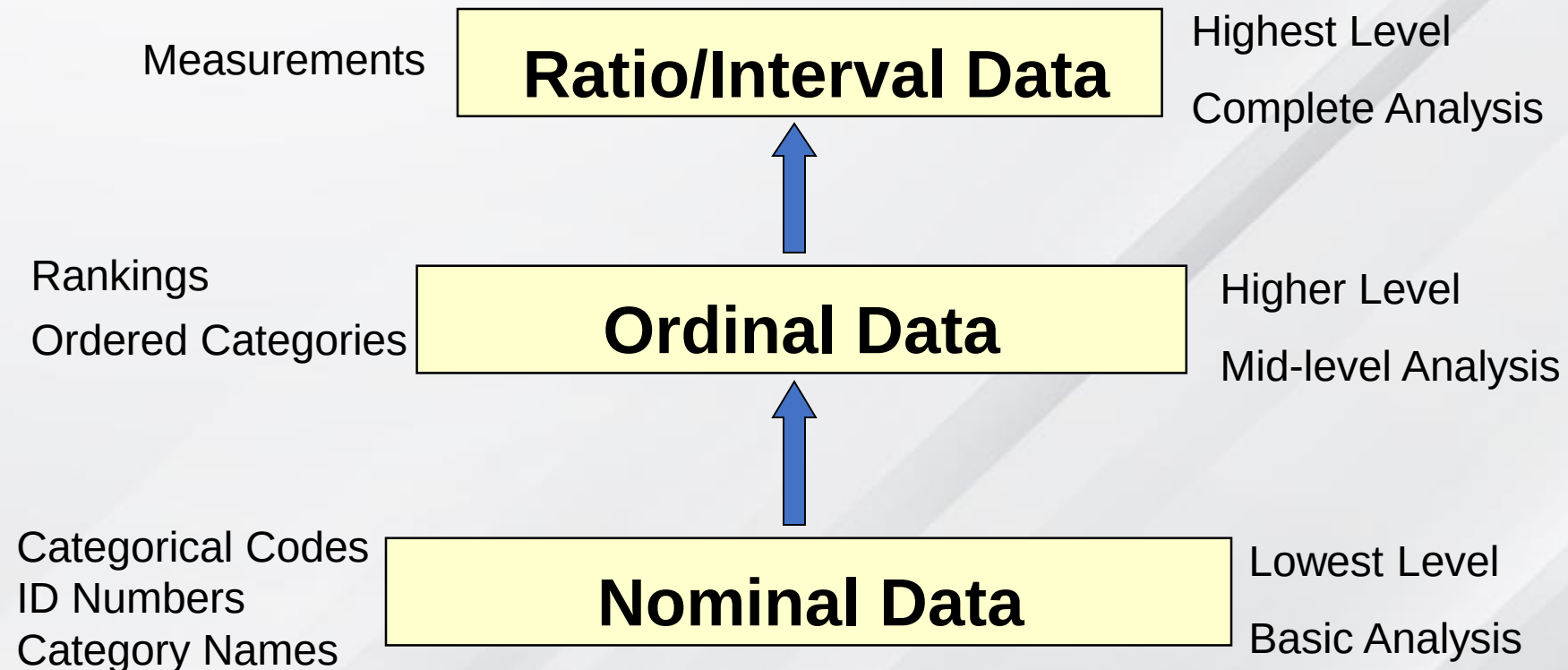
Data Types

	Sales (in \$1000's)			
	2003	2004	2005	2006
Atlanta	435	460	475	490
Boston	320	345	375	395
Cleveland	405	390	410	395
Denver	260	270	285	280

Time Series Data

Cross Section Data

Data Measurement Levels



Chapter 2

Graphs, Charts, and Tables

– Describing Your Data

2



Chapter Goals

After completing this chapter, you should be able to:

- Construct a frequency distribution
- Construct and interpret a histogram

Frequency Distributions

What is a Frequency Distribution?

- A frequency distribution is a **list or a table** ...
- containing the **values of a variable** (or a set of ranges within which the data fall) ...
- and the **corresponding frequencies** with which each value occurs (or frequencies with which data fall within each range)

Why Use Frequency Distributions?

- A frequency distribution is a way to summarize data
- The distribution condenses the raw data into a more useful form...
- and allows for a quick visual interpretation of the data

Frequency Distribution: Discrete Data

- **Discrete data:** possible values are countable

Example: An advertiser asks 200 customers how many days per week they read the daily newspaper.



Number of days read	Frequency
0	44
1	24
2	18
3	16
4	20
5	22
6	26
7	30
Total	200

Relative Frequency

Relative Frequency: What proportion is in each category?

Number of days read	Frequency	Relative Frequency
0	44	.22
1	24	.12
2	18	.09
3	16	.08
4	20	.10
5	22	.11
6	26	.13
7	30	.15
Total	200	1.00

$$\frac{44}{200} = .22$$

22% of the people in the sample report that they read the newspaper 0 days per week



Frequency Distribution: Continuous Data

- **Continuous Data:** may take on any value in some interval

Example: A manufacturer of insulation randomly selects 20 winter days and records the **daily high temperature**

24, 35, 17, 21, 24, 37, 26, 46, 58, 30,
32, 13, 12, 38, 41, 43, 44, 27, 53, 27

(Temperature is a continuous variable because it could be measured to any degree of precision desired)

Grouping Data by Classes

Sort raw data from low to high:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

- Find range: $58 - 12 = 46$
- Select number of classes: 5 (usually between 5 and 20)
- Compute class width: 10 (46/5 then round off)
- Determine class boundaries: 10, 20, 30, 40, 50
(Sometimes class midpoints are reported: 15, 25, 35, 45, 55)
- Count the number of values in each class

Frequency Distribution Example

Data from low to high:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Frequency Distribution		
Class	Frequency	Relative Frequency
10 but under 20	3	.15
20 but under 30	6	.30
30 but under 40	5	.25
40 but under 50	4	.20
50 but under 60	2	.10
Total	20	1.00

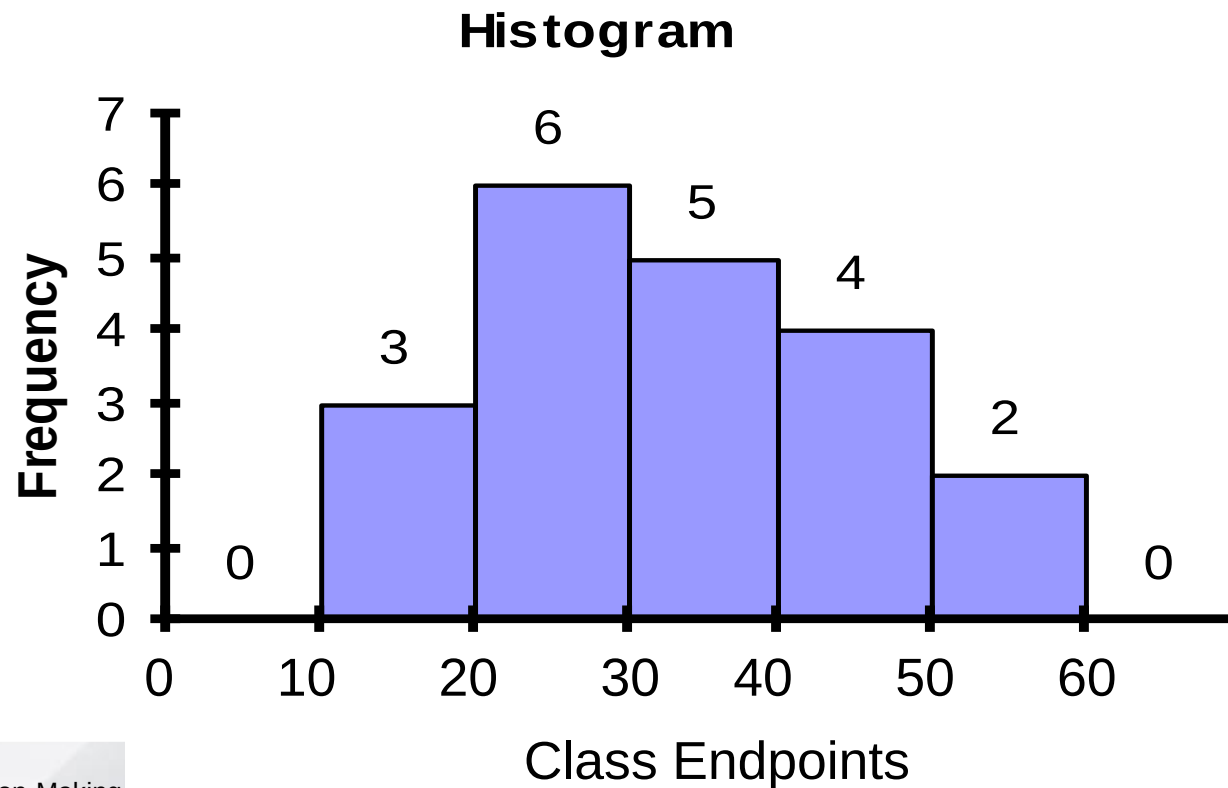
Frequency Histograms

- The **classes** or **intervals** are shown on the **horizontal axis**
- **frequency** is measured on the **vertical axis**
- Bars of the appropriate heights can be used to represent the number of observations within each class
- Such a graph is called a **histogram**

Histogram Example

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58



No gaps
between
bars, since
continuous
data

Chapter 3

Describing Data Using Numerical Measures

3



Chapter Goals

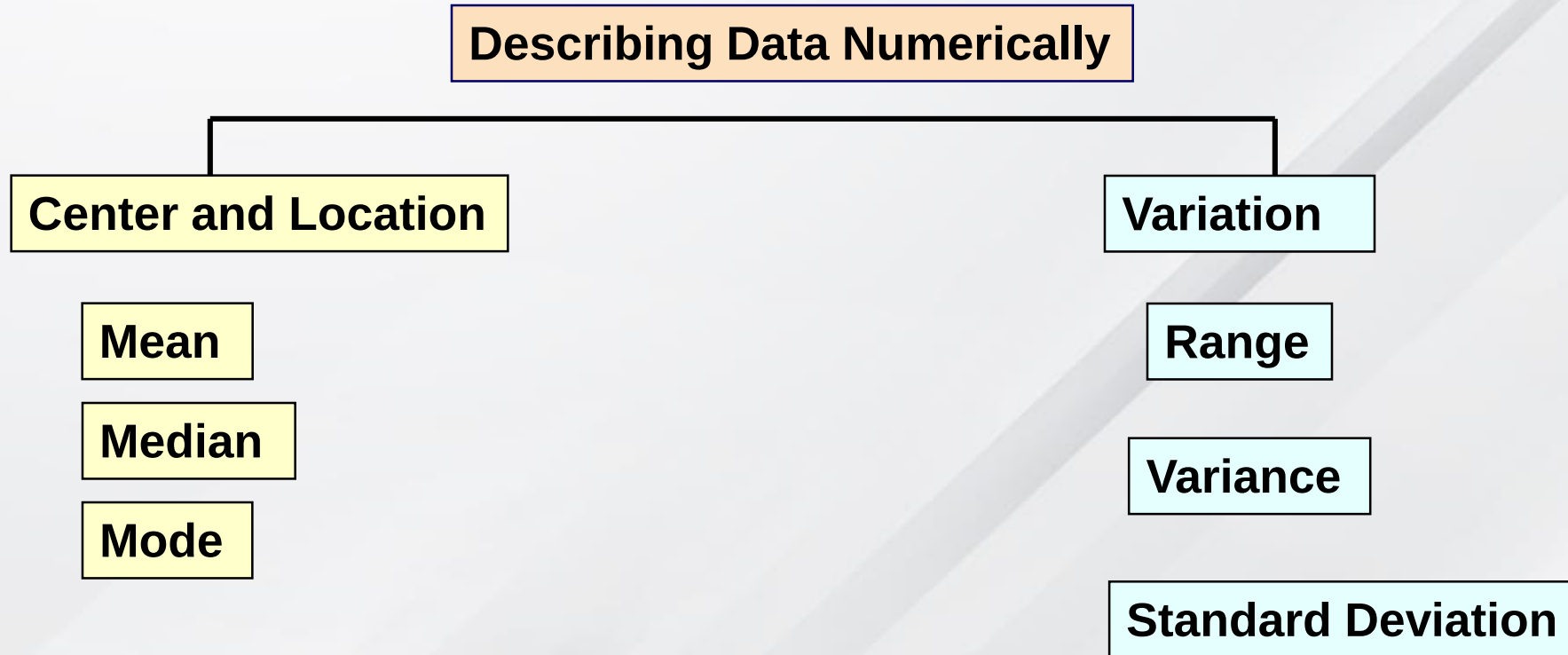
After completing this chapter, you should be able to:

- Compute and interpret the **mean, median, and mode** for a set of data
- Compute the **range, variance, and standard deviation** and know what these values mean

Chapter Topics

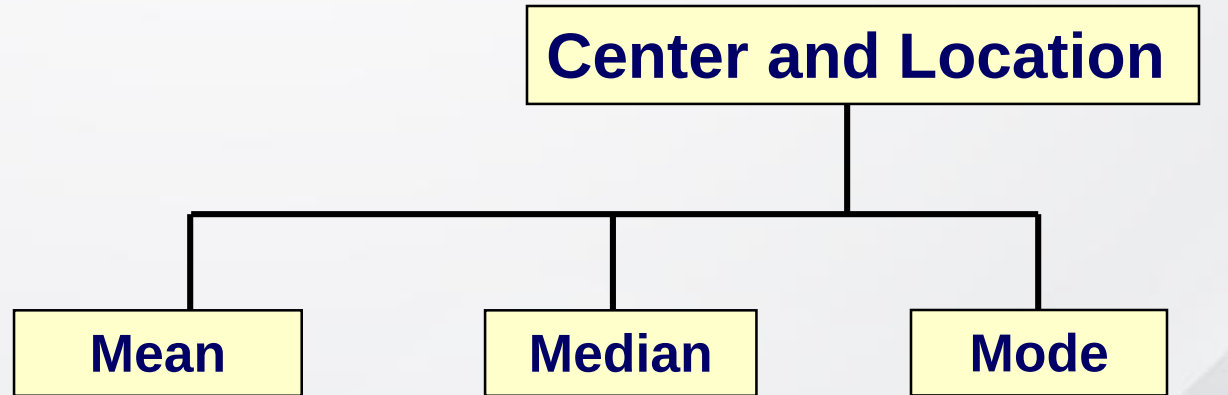
- Measures of Center and Location
 - Mean, median, mode
- Measures of Variation
 - Range, variance and standard deviation

Summary Measures



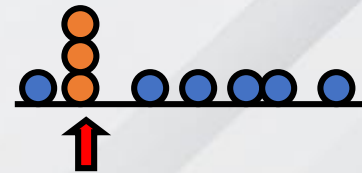
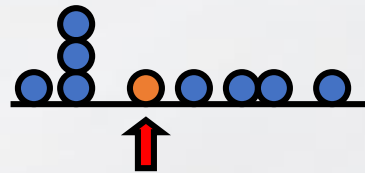
Measures of Center and Location

Overview



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

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



Mean (Arithmetic Average)

- The **Mean** is the arithmetic average of data values
 - Population mean

• Sample mean $\mu = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$

N = Population Size

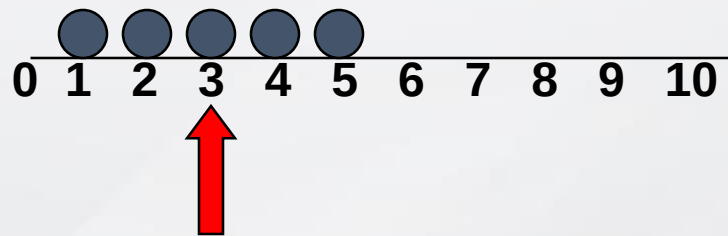
$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$

n = Sample Size

Mean (Arithmetic Average)

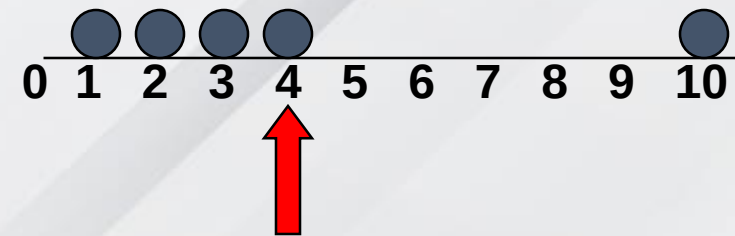
(continued)

- The most common measure of central tendency
- Mean = sum of values divided by the number of values
- Affected by extreme values (outliers)



Mean = 3

$$\frac{1 + 2 + 3 + 4 + 5}{5} = \frac{15}{5} = 3$$

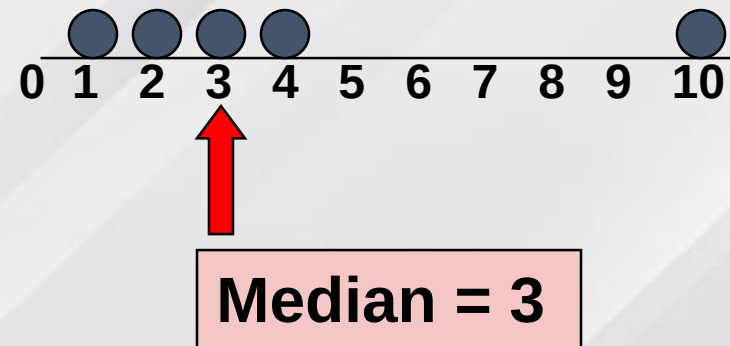
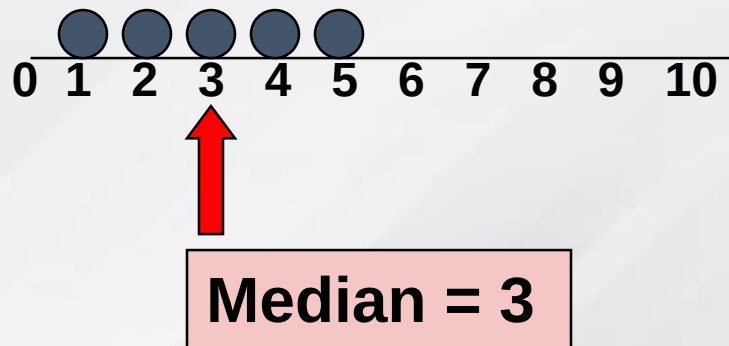


Mean = 4

$$\frac{1 + 2 + 3 + 4 + 10}{5} = \frac{20}{5} = 4$$

Median

- In an ordered array, the median is the “middle” number, i.e., the number that splits the distribution in half
- The median is not affected by extreme values



Median

(continued)

- To find the median, sort the n data values from low to high (sorted data is called a data array)
- Find the value in the $i = (1/2)n$ position
- The i^{th} position is called the Median Index Point
 - If i is not an integer, round up to next highest integer

Median Example

(continued)

Data array:

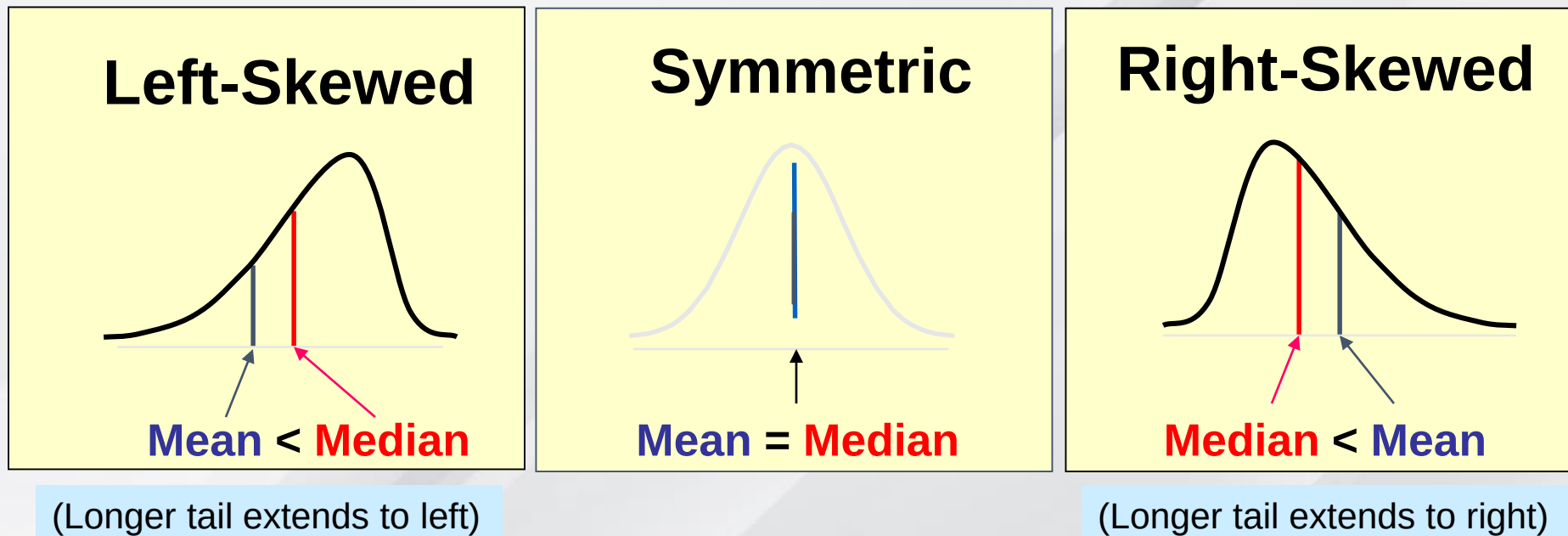
4, 4, 5, 5, 9, 11, 12, 14, 16, 19, 22, 23, 24

- Note that $n = 13$
- Find the $i = (1/2)n$ position:
$$i = (1/2)(13) = 6.5$$
- Since 6.5 is not an integer, round up to 7
- The median is the value in the 7th position:

$$M_d = 12$$

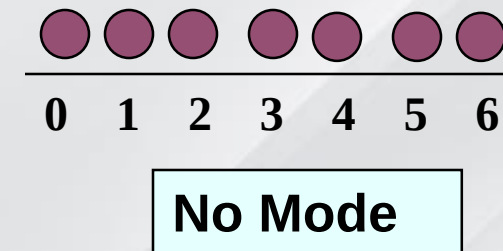
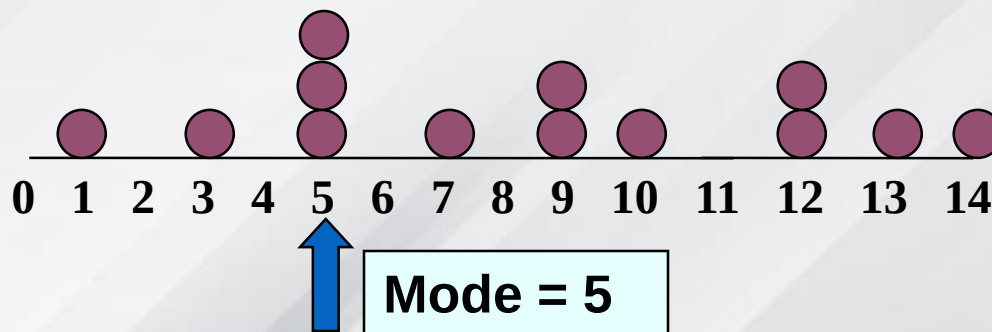
Shape of a Distribution

- Describes how data is distributed
- Symmetric or skewed



Mode

- A measure of location
- The value that occurs **most often**
- Not affected by extreme values
- Used for either numerical or categorical data
- There may be no mode
- There may be several modes

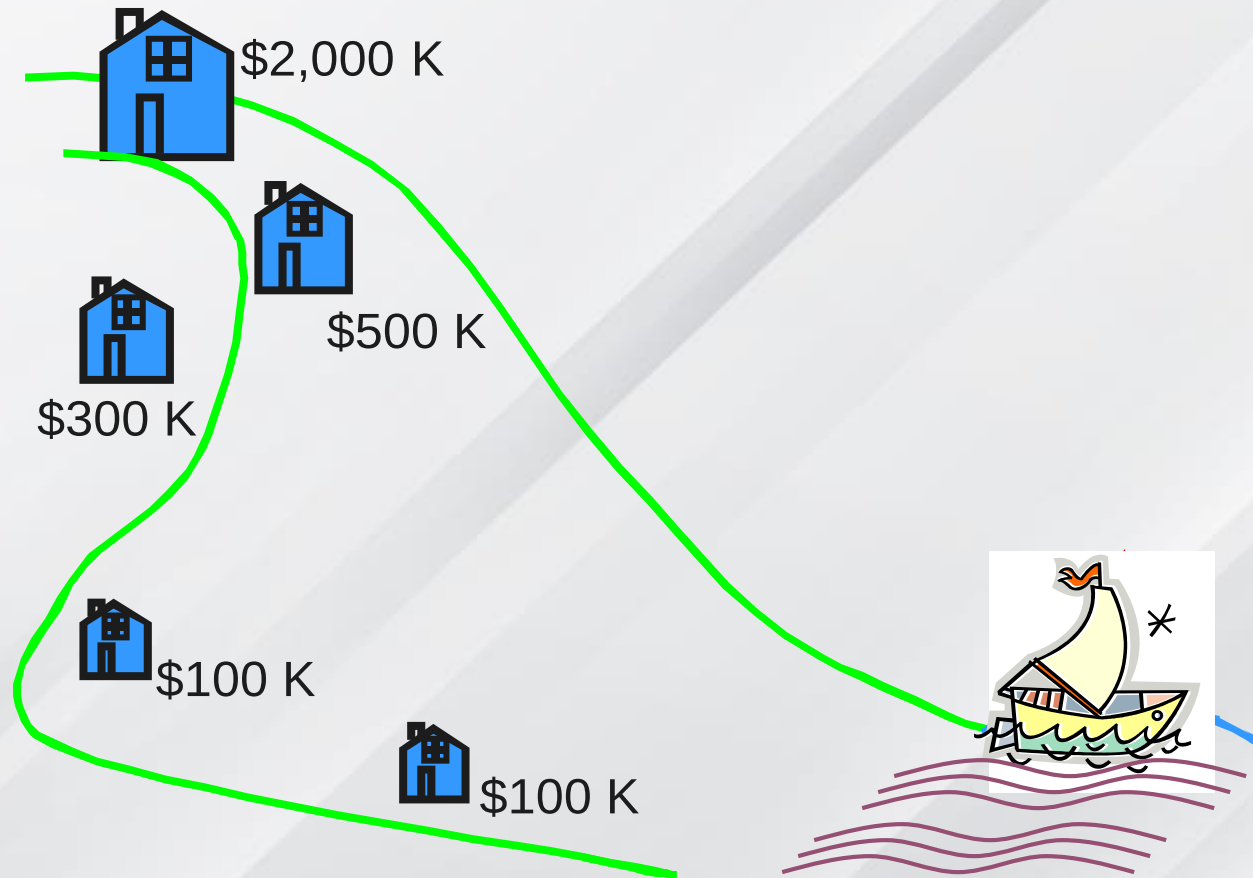


Review Example

- Five houses on a hill by the beach

House Prices:

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



Summary Statistics

House Prices:

\$2,000,000
500,000
300,000
100,000
<u>100,000</u>

Sum 3,000,000

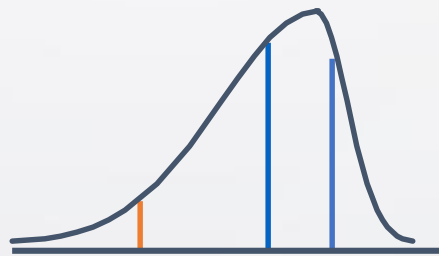
- **Mean:** $(\$3,000,000/5)$
= **\$600,000**
- **Median:** middle value of ranked data
= **\$300,000**
- **Mode:** most frequent value
= **\$100,000**

Which measure of location is the “best”?

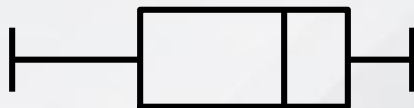
- **Mean** is generally used, unless extreme values (outliers) exist
- Then **Median** is often used, since the median is not sensitive to extreme values.
 - **Example:** Median home prices may be reported for a region – less sensitive to outliers

Distribution Shape and Box and Whisker Plot

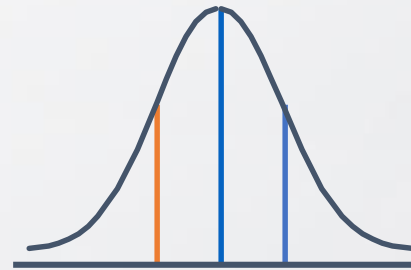
Left-Skewed



Q1 Q2 Q3



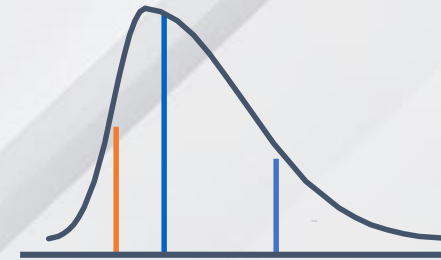
Symmetric



Q1 Q2 Q3



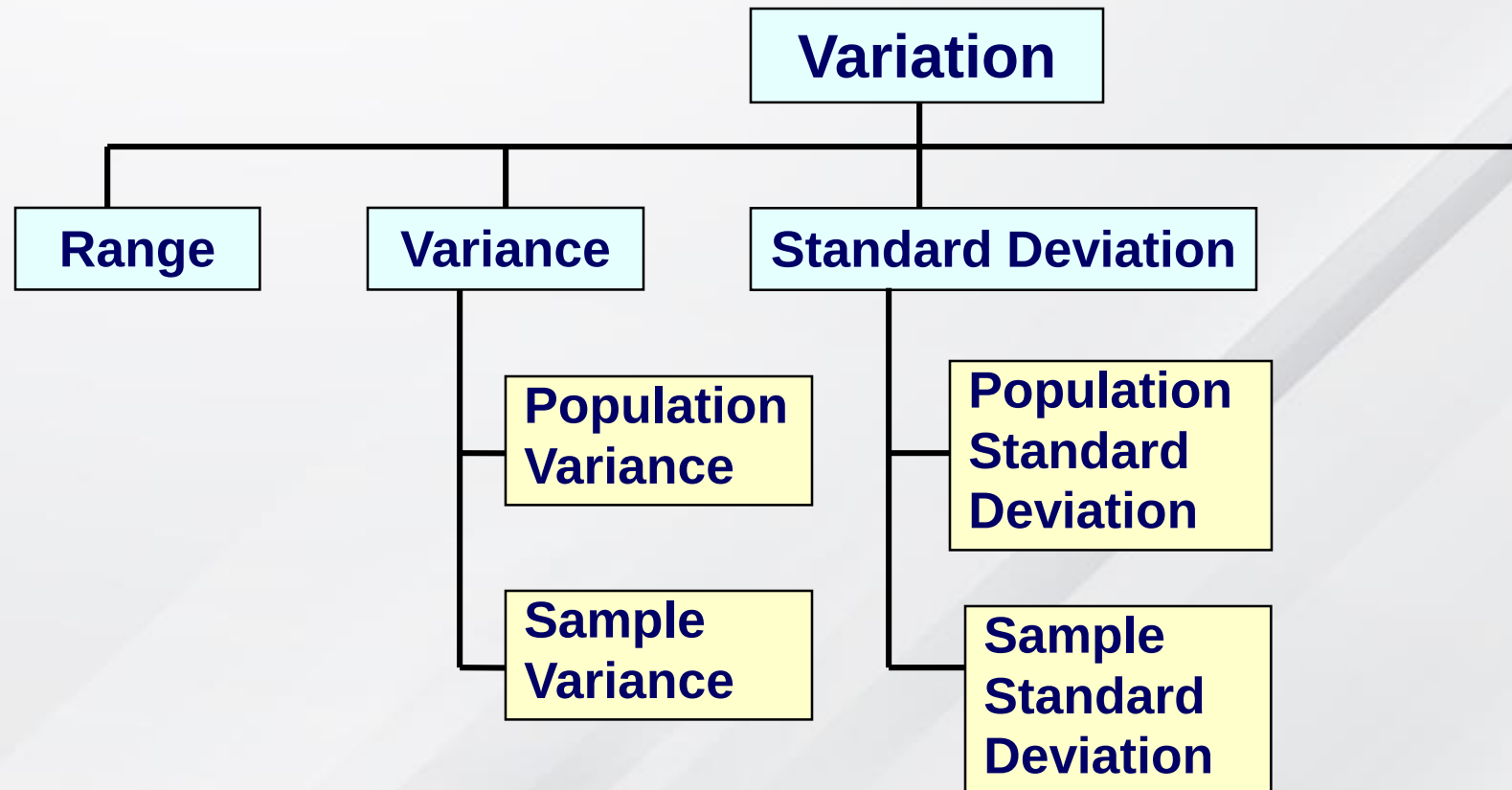
Right-Skewed



Q1 Q2 Q3

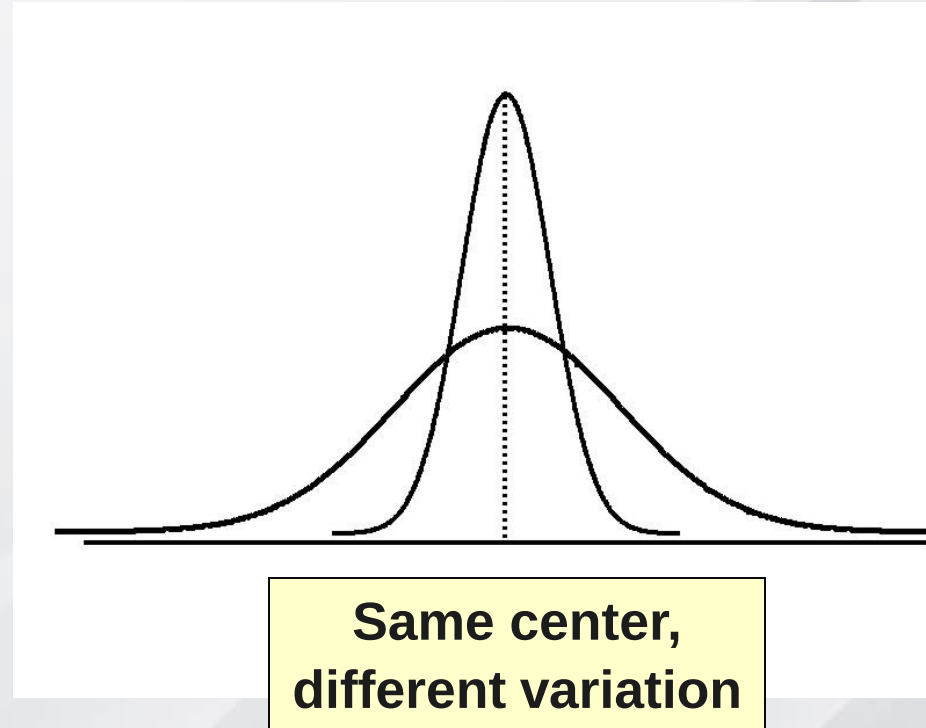


Measures of Variation



Variation

- Measures of variation give information on the **spread** or **variability** of the data values.

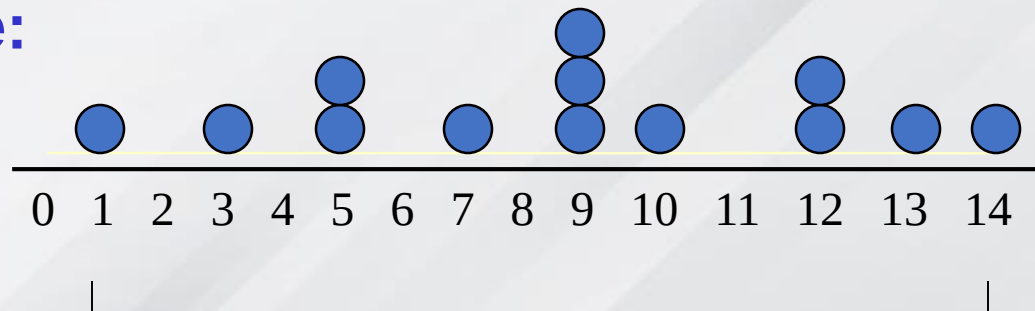


Range

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

$$\text{Range} = x_{\text{maximum}} - x_{\text{minimum}}$$

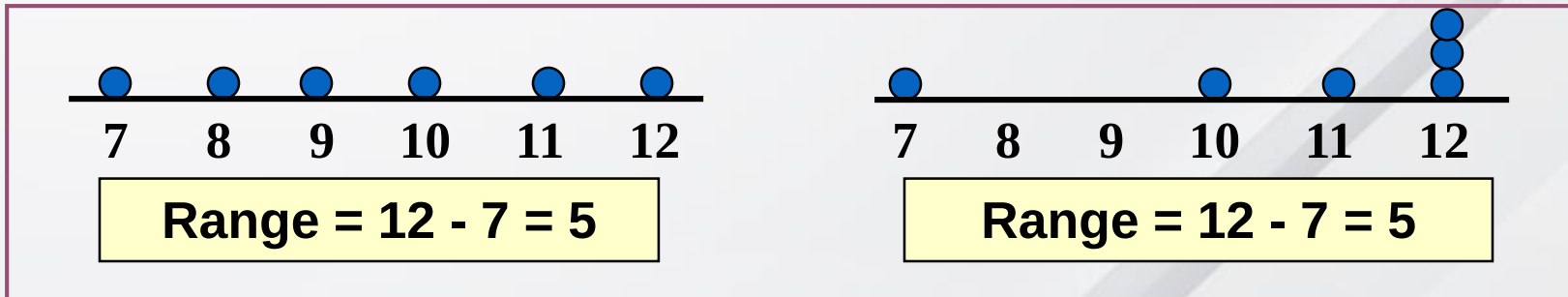
Example:



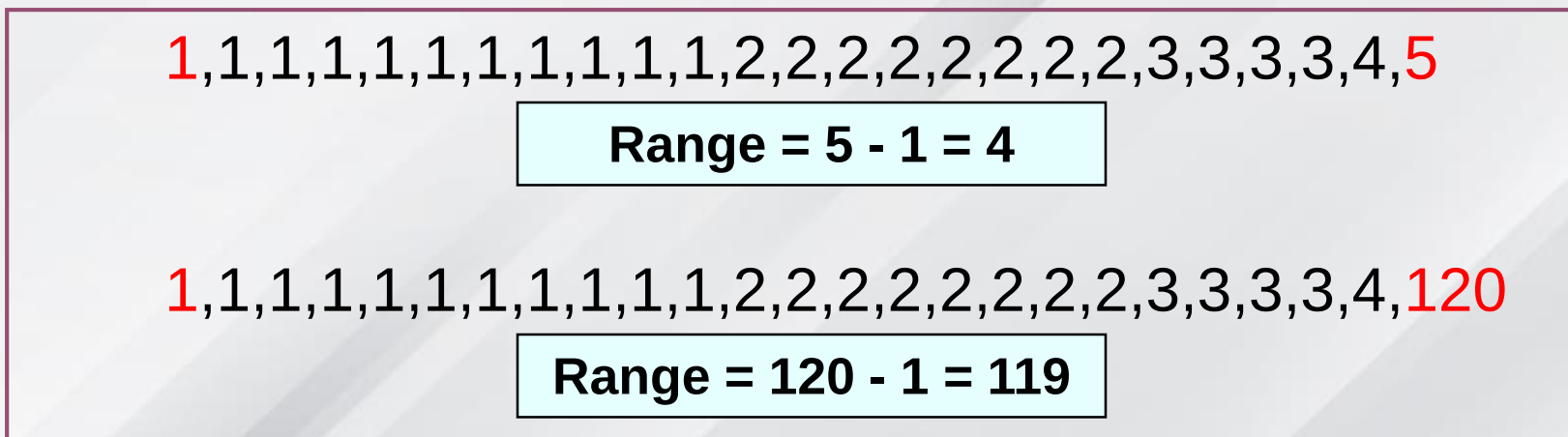
$$\text{Range} = 14 - 1 = 13$$

Disadvantages of the Range

- Ignores the way in which data are distributed



- Sensitive to outliers



Variance

- Average of squared deviations of values from the mean

- **Population** variance:

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

- **Sample** variance:

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

Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

- **Population** standard deviation:

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

- **Sample** standard deviation:

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

Calculation Example: Sample Standard Deviation

Sample

Data (X_i) :

10 12 14 15 17 18 18 24

$n = 8$

Mean = $\bar{x} = 16$

$$s = \sqrt{\frac{(10 - \bar{x})^2 + (12 - \bar{x})^2 + (14 - \bar{x})^2 + \cdots + (24 - \bar{x})^2}{n - 1}}$$

$$= \sqrt{\frac{(10 - 16)^2 + (12 - 16)^2 + (14 - 16)^2 + \cdots + (24 - 16)^2}{8 - 1}}$$

$$= \sqrt{\frac{130}{7}} = \boxed{4.3095}$$

Chapter 9

Introduction to Hypothesis Testing

4



Chapter Goals

After completing this chapter, you should be able to:

- Formulate null and alternative hypotheses for applications involving a single population mean or proportion
- Formulate a decision rule for testing a hypothesis
- Know how to use the test statistic, critical value, and p-value approaches to test the null hypothesis
- Know what Type I and Type II errors are
- Compute the probability of a Type II error

What is a Hypothesis?

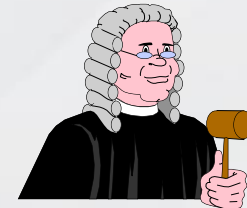
- A hypothesis is a claim (assumption) about a population parameter:



The Null Hypothesis, H_0

(continued)

- Begin with the assumption that the null hypothesis is true
 - Similar to the notion of innocent until proven guilty
- Refers to the **status quo**
- May or may not be rejected



The Alternative Hypothesis, H_A

- Is the opposite of the null hypothesis
 - e.g.: Challenges the status quo
- May or may not be accepted
- Is generally the hypothesis that is believed (or needs to be supported) by the researcher – a **research hypothesis**

Errors in Making Decisions

- **Type I Error**

- Reject a true null hypothesis
- Considered a serious type of error

The probability of Type I Error is α

- Called **level of significance** of the test
- Set by researcher in advance

Errors in Making Decisions

(continued)

- **Type II Error**
 - Fail to reject a false null hypothesis

The probability of Type II Error is β

- β is a calculated value, the formula is discussed later in the chapter

Outcomes and Probabilities

Possible Hypothesis Test Outcomes

	State of Nature	
Decision	H_0 True	H_0 False
Do Not Reject H_0	No error ($1 - \alpha$)	Type II Error (β)
Reject H_0	Type I Error (α)	No Error ($1 - \beta$)

Key:
Outcome
(Probability)

Type I & II Error Relationship

- Type I and Type II errors cannot happen at the same time
 - Type I error can only occur if H_0 is true
 - Type II error can only occur if H_0 is false

Level of Significance, α

- **Defines unlikely values of sample statistic if null hypothesis is true**
 - Defines **rejection region** of the sampling distribution
- Is designated by α , (level of significance)
 - Typical values are .01, .05, or .10
- Is selected by the researcher at the beginning
- Provides the critical value(s) of the test

Process of Hypothesis Testing

- 1. Specify population parameter of interest
- 2. Formulate the null and alternative hypotheses
- 3. Specify the desired significance level, α
- 4. Define the rejection region
- 5. Take a random sample and determine whether or not the sample result is in the rejection region
- 6. Reach a decision and draw a conclusion

p-Value Approach to Testing

- Determine the **p-value** from a table or computer
- Compare the **p-value** with **α**
 - If $\text{p-value} < \alpha$, reject H_0
 - If $\text{p-value} \geq \alpha$, do not reject H_0

p-Value Approach to Testing

(continued)

- p-value: Probability of obtaining a test statistic more extreme ($\leq$ or $\geq$) than the observed sample value **given H_0 is true**
 - Also called observed level of significance
 - Smallest value of α for which H_0 can be rejected

p-value example

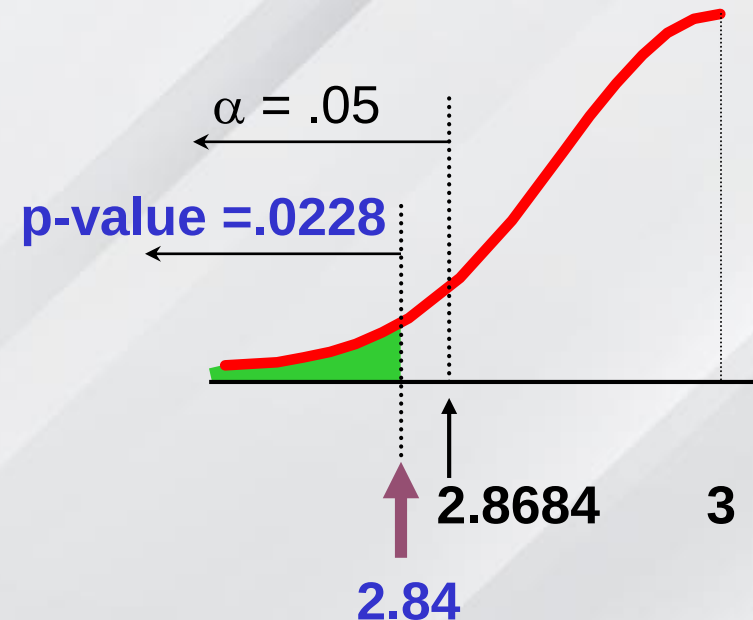
(continued)

- Compare the p-value with α
 - If $\text{p-value} < \alpha$, reject H_0
 - If $\text{p-value} \geq \alpha$, do not reject H_0

Here: p-value = .0228

$\alpha = .05$

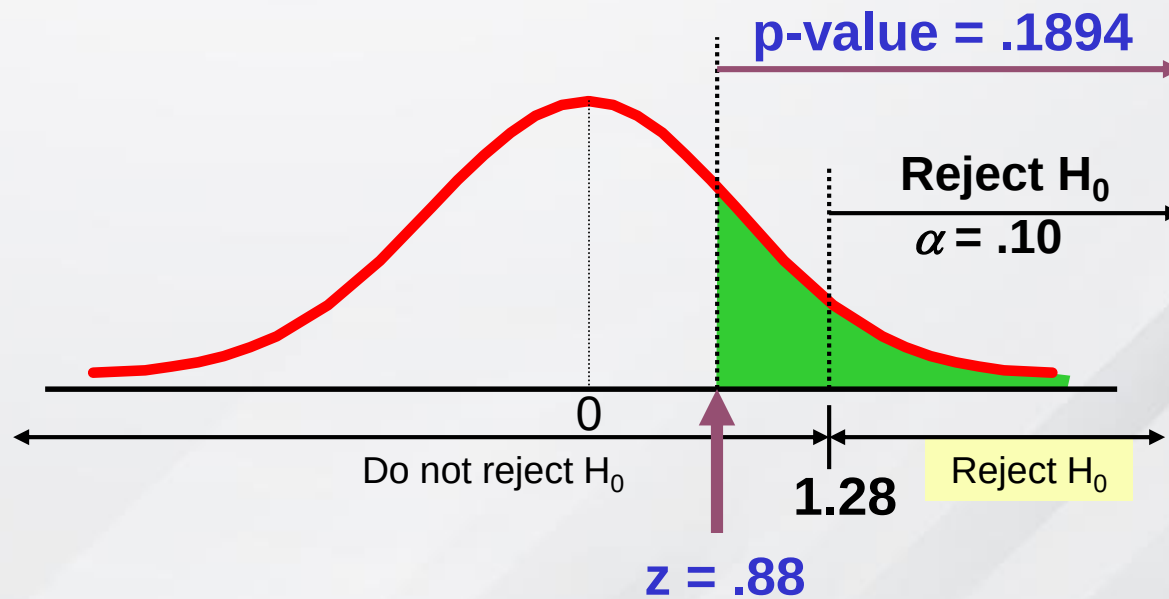
Since $.0228 < .05$, we reject
the null hypothesis



p -Value Solution

(continued)

Calculate the p -value and compare to α



$$\begin{aligned} &P(\bar{x} \geq 53.1 | \mu = 52.0) \\ &= P\left(z \geq \frac{53.1 - 52.0}{10/\sqrt{64}}\right) \\ &= P(z \geq 0.88) = .5 - .3106 \\ &= .1894 \end{aligned}$$

Do not reject H_0 since $p\text{-value} = .1894 > \alpha = .10$

Chapter 14
Introduction to Linear
Regression and Correlation
Analysis

5



Chapter Goals

After completing this chapter, you should be able to:

- Calculate and interpret the simple correlation between two variables
- Determine whether the correlation is significant
- Calculate and interpret the simple linear regression equation for a set of data
- Understand the assumptions behind regression analysis
- Determine whether a regression model is significant

Scatter Plots and Correlation

- A **scatter plot** (or scatter diagram) is used to show the relationship between two variables
- **Correlation** analysis is used to measure strength of the association (linear relationship) between two variables
 - Only concerned with strength of the relationship
 - No causal effect is implied

Scatter Plot Examples

Linear relationships

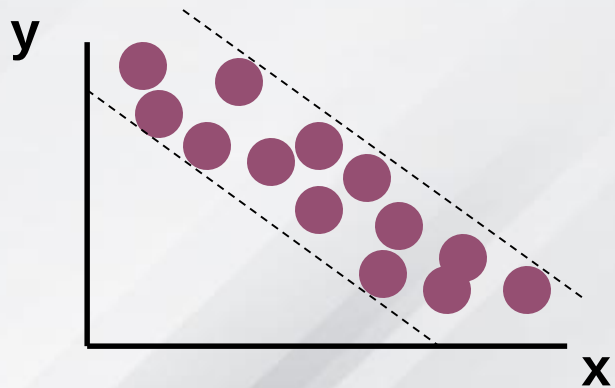
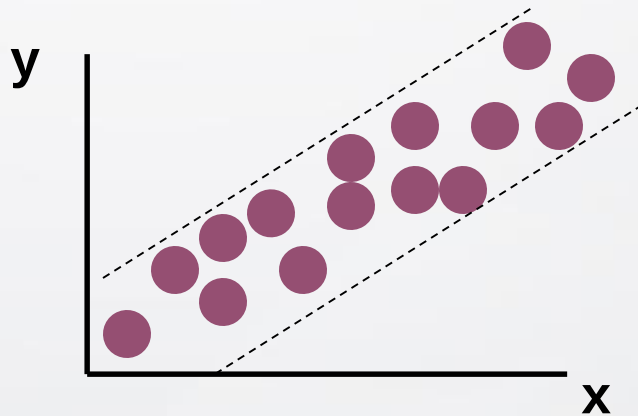


Curvilinear relationships



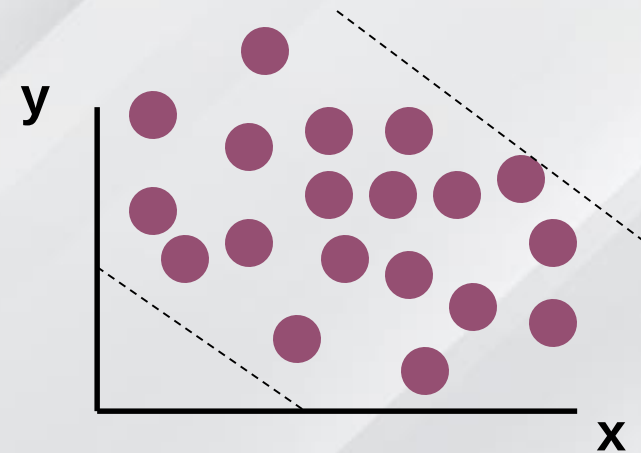
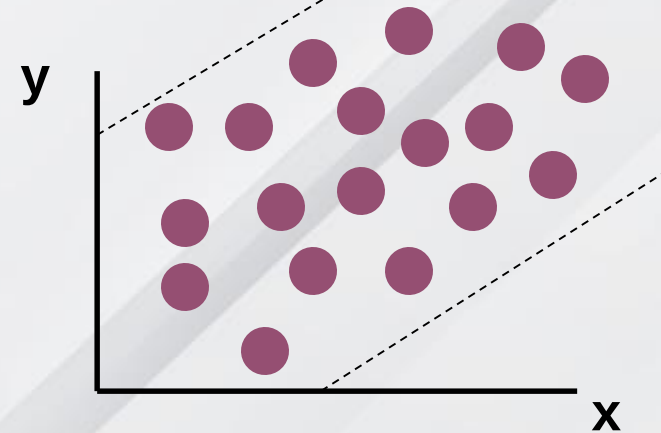
Scatter Plot Examples

Strong relationships



(continued)

Weak relationships



Scatter Plot Examples

(continued)

No relationship



Correlation Coefficient

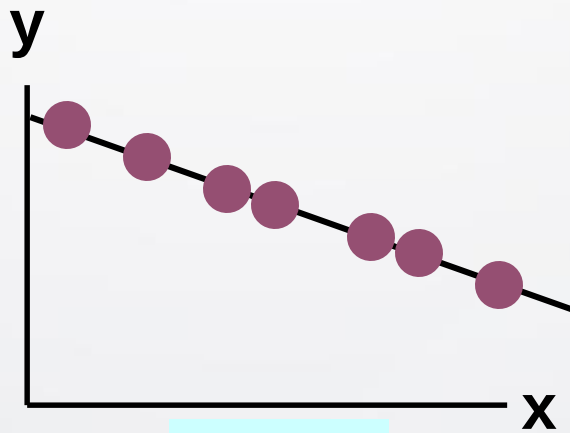
(continued)

- **Correlation** measures the strength of the linear association between two variables
- The **sample correlation coefficient r** is a measure of the strength of the linear relationship between two variables, based on sample observations

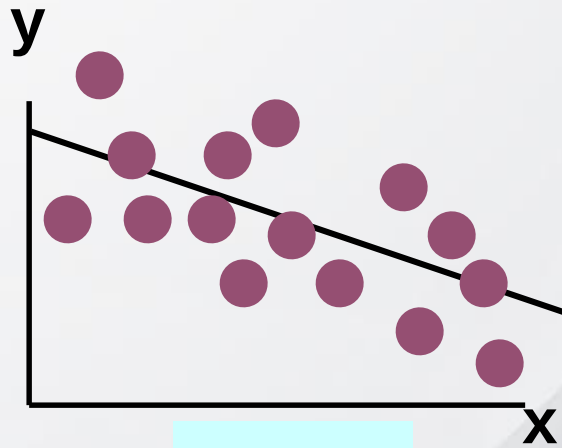
Features of r

- 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

Examples of Approximate r Values



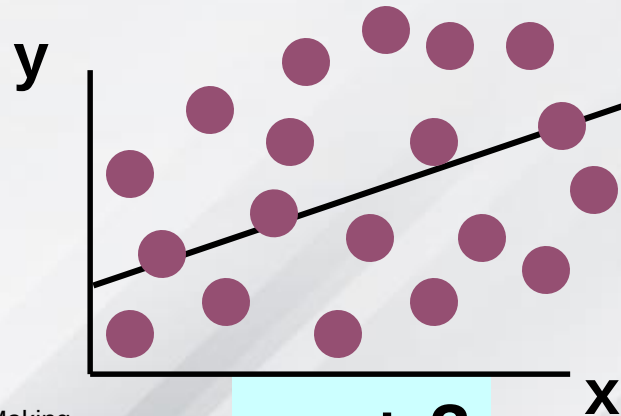
$r = -1$



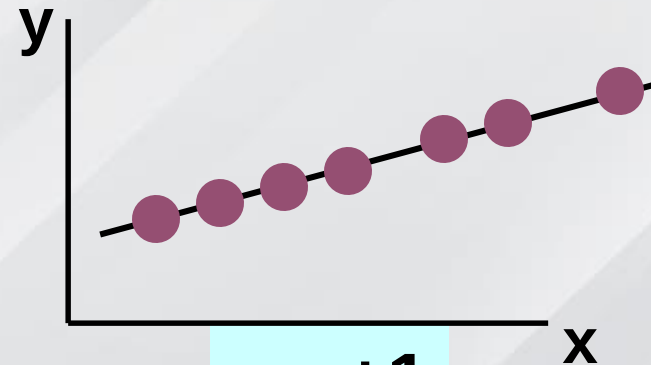
$r = -.6$



$r = 0$



$r = +.3$



$r = +1$

Calculating the Correlation Coefficient

Sample correlation coefficient:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{[\sum (x - \bar{x})^2][\sum (y - \bar{y})^2]}}$$

or the algebraic equivalent:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

where:

r = Sample correlation coefficient

n = Sample size

x = Value of the independent variable

y = Value of the dependent variable

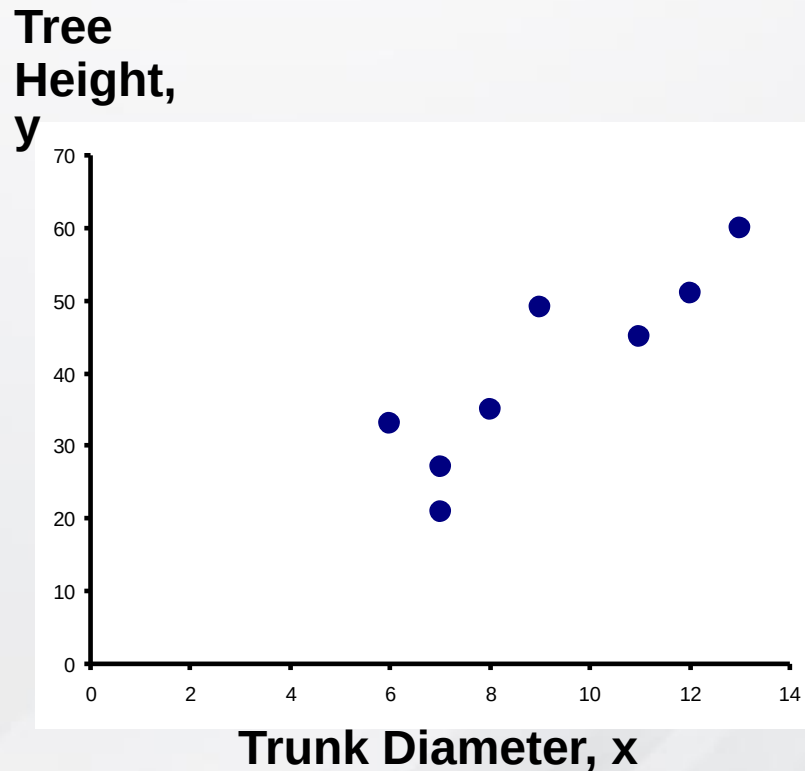
Calculation Example



Tree Height	Trunk Diameter			
y	x	xy	y^2	x^2
35	8	280	1225	64
49	9	441	2401	81
27	7	189	729	49
33	6	198	1089	36
60	13	780	3600	169
21	7	147	441	49
45	11	495	2025	121
51	12	612	2601	144
$\Sigma=321$	$\Sigma=73$	$\Sigma=3142$	$\Sigma=14111$	$\Sigma=713$

Calculation Example

(continued)



$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$
$$= \frac{8(3142) - (73)(321)}{\sqrt{[8(713) - (73)^2][8(14111) - (321)^2]}}$$
$$= 0.886$$

$r = 0.886$ → relatively strong positive linear association between x and y



Excel Output

Excel Correlation Output

Tools / data analysis / correlation...

	Tree Height	Trunk Diameter
Tree Height	1	
Trunk Diameter	0.886231	1

Correlation between
Tree Height and Trunk Diameter



Significance Test for Correlation

- Hypotheses

$$\begin{array}{ll} H_0: \rho = 0 & \text{(no correlation)} \\ H_A: \rho \neq 0 & \text{(correlation exists)} \end{array}$$

The Greek letter ρ (rho) represents the population correlation coefficient

- Test statistic

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

(with $n - 2$ degrees of freedom)



Example: Produce Stores

Is there evidence of a linear relationship between tree height and trunk diameter at the .05 level of significance?

$H_0: \rho = 0$ (No correlation)

$H_1: \rho \neq 0$ (correlation exists)

$\alpha = .05$, $df = 8 - 2 = 6$

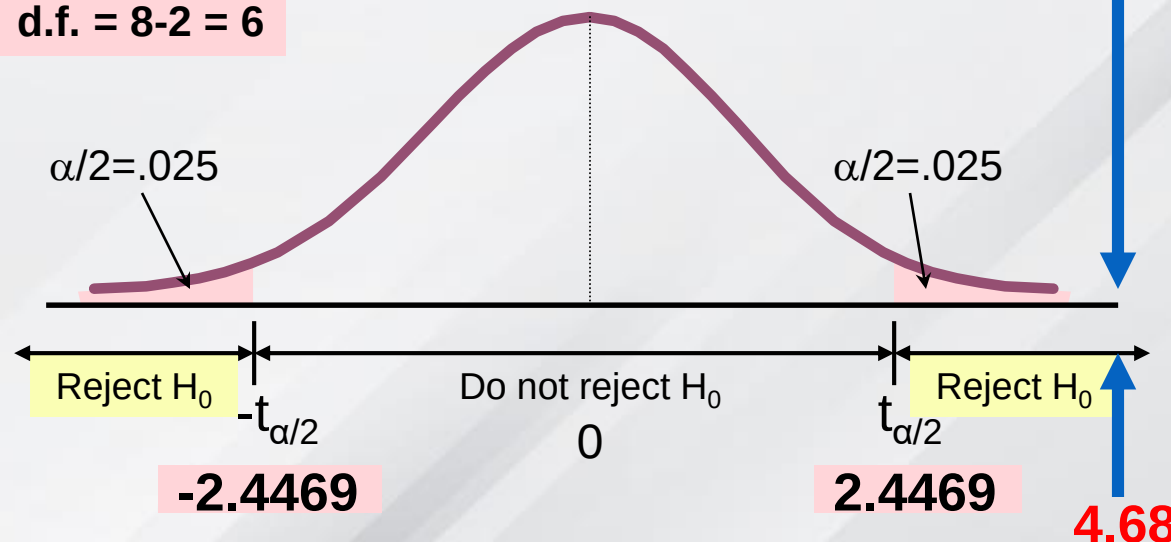
$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.886}{\sqrt{\frac{1-.886^2}{8-2}}} = 4.68$$



Example: Test Solution

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.886}{\sqrt{\frac{1-.886^2}{8-2}}} = 4.68$$

d.f. = 8-2 = 6



Decision:
Reject H₀

Conclusion:
There is sufficient evidence of a linear relationship at the 5% level of significance

Introduction to Regression Analysis

- **Regression analysis** is used to:
 - Predict the value of a dependent variable based on the value of at least one independent variable
 - Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to explain

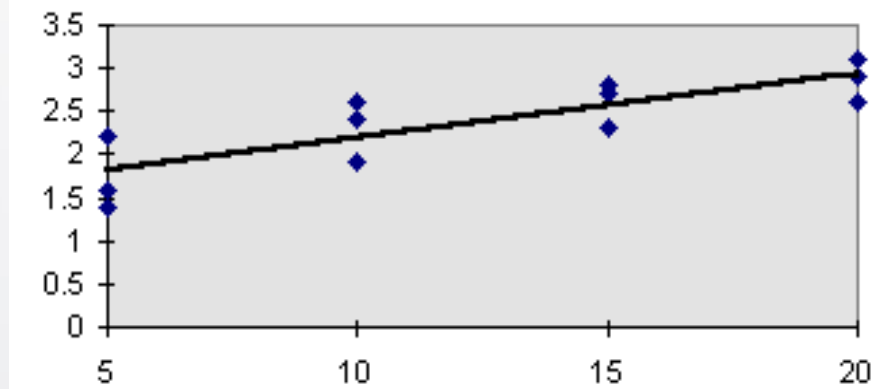
Independent variable: the variable used to explain the dependent variable

Simple Linear Regression Model

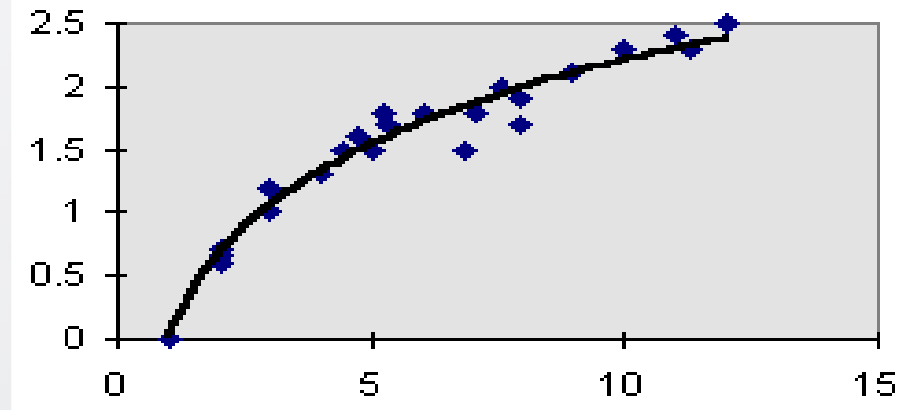
- Only **one independent variable**, x
- Relationship between x and y is described by a linear function
- Changes in y are assumed to be **caused** by changes in x

Types of Regression Models

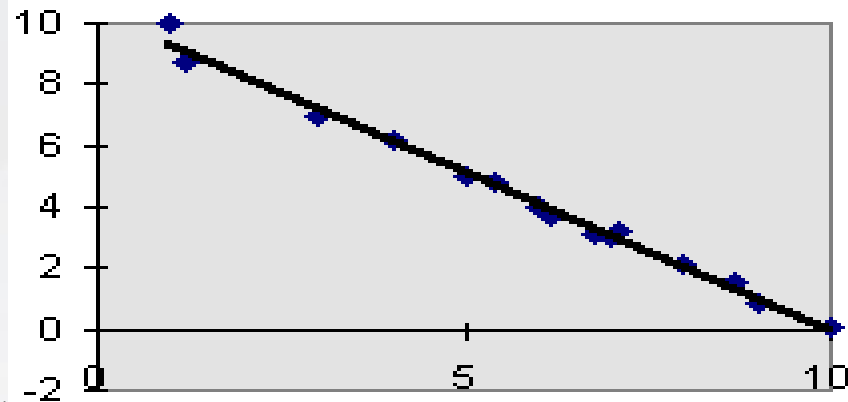
Positive Linear Relationship



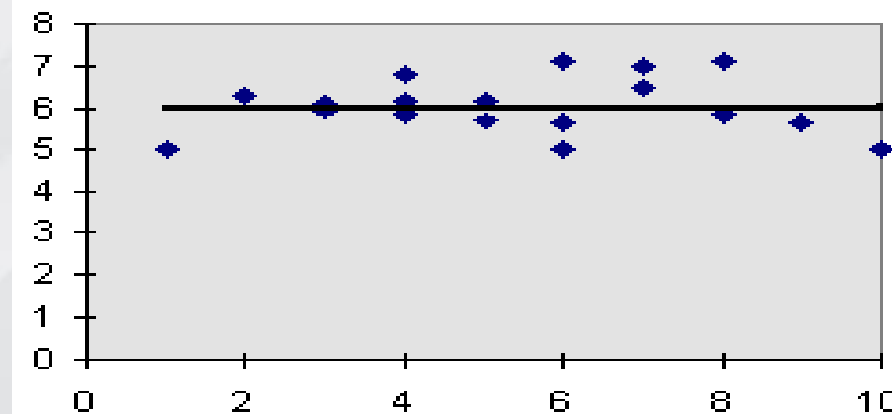
Relationship NOT Linear



Negative Linear Relationship



No Relationship



Population Linear Regression

The population regression model:

The diagram illustrates the population linear regression model equation: $y = \beta_0 + \beta_1 X + \epsilon$. The equation is enclosed in a yellow box. Labels with arrows point to each part of the equation: 'Dependent Variable' points to y ; 'Population y intercept' points to β_0 ; 'Population Slope Coefficient' points to β_1 ; 'Independent Variable' points to X ; and 'Random Error term, or residual' points to ϵ . Below the equation, two purple curly braces group the terms: the first brace under $\beta_0 + \beta_1 X$ is labeled 'Linear component', and the second brace under ϵ is labeled 'Random Error component'.

$$y = \beta_0 + \beta_1 X + \epsilon$$

Labels and components:

- Dependent Variable: y
- Population y intercept: β_0
- Population Slope Coefficient: β_1
- Independent Variable: X
- Random Error term, or residual: ϵ
- Linear component: $\beta_0 + \beta_1 X$
- Random Error component: ϵ

Estimated Regression Model

The sample regression line provides an **estimate** of the population regression line

Estimated
(or predicted)
y value

Estimate of
the regression
intercept

Estimate of the
regression slope

Independent
variable

$$\hat{y}_i = b_0 + b_1 x$$

The individual random error terms e_i have a mean of zero

Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)
- A random sample of 10 houses is selected
 - Dependent variable (y) = house price in \$1000s
 - Independent variable (x) = square feet



Sample Data for House Price Model

House Price in \$1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700



Regression Using Excel

- Data / Data Analysis / Regression

The screenshot shows the Microsoft Excel interface with the 'Data' tab selected in the ribbon. The 'Data Analysis' button is circled in the 'Analysis' group. The 'Regression' dialog box is open, showing the following settings:

- Input Y Range:** \$A\$1:\$A\$11
- Input X Range:** \$B\$1:\$B\$11
- ☒ **Labels**
- ☐ **Constant is Zero**
- ☐ **Confidence Level:** 95 %
- Output options:**
 - ☐ **Output Range:**
 - ☒ **New Worksheet Ply:**
 - ☐ **New Workbook**
- Residuals:**
 - ☐ **Residuals**
 - ☐ **Standardized Residuals**
 - ☐ **Residual Plots**
 - ☐ **Line Fit Plots**
- Normal Probability:**
 - ☐ **Normal Probability Plots**

The background spreadsheet shows the following data:

	A	B
1	House Price	Square Feet
2	245	1400
3	312	1600
4	279	1700
5	308	1875
6	199	1100
7	219	1550
8	405	2350
9	324	2450
10	319	1425
11	255	1700
12		
13		

Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

$$\widehat{\text{house price}} = 98.24833 + 0.10977 (\text{square feet})$$

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580



Interpretation of the Intercept, b_0

$$\widehat{\text{house price}} = 98.24833 + 0.10977 (\text{square feet})$$

- b_0 is the estimated average value of Y when the value of X is zero (if $x = 0$ is in the range of observed x values)
 - Here, no houses had 0 square feet, so $b_0 = 98.24833$ just indicates that, for houses within the range of sizes observed, \$98,248.33 is the portion of the house price not explained by square feet



Interpretation of the Slope Coefficient, b_1

$$\widehat{\text{house price}} = 98.24833 + 0.10977(\text{square feet})$$

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
 - Here, $b_1 = .10977$ tells us that the average value of a house increases by $.10977(\$1000) = \109.77 , on average, for each additional one square foot of size



Coefficient of Determination, R^2

- The **coefficient of determination** is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called **R-squared** and is denoted as R^2

where

$$0 \leq R^2 \leq 1$$

Coefficient of Determination, R^2

(continued)

Coefficient of determination

Note: In the single independent variable case, the coefficient of determination is

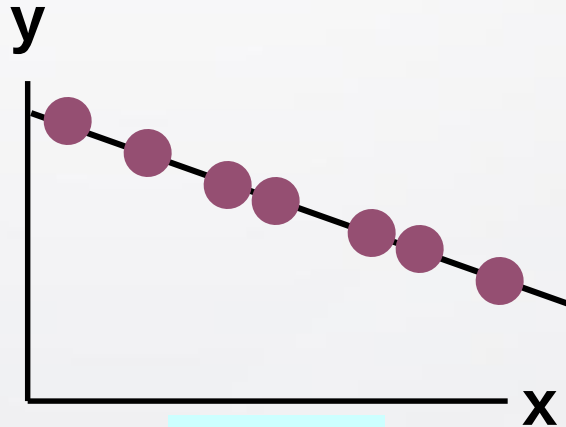
$$R^2 = r^2$$

where:

R^2 = Coefficient of determination

r = Simple correlation coefficient

Examples of Approximate R^2 Values

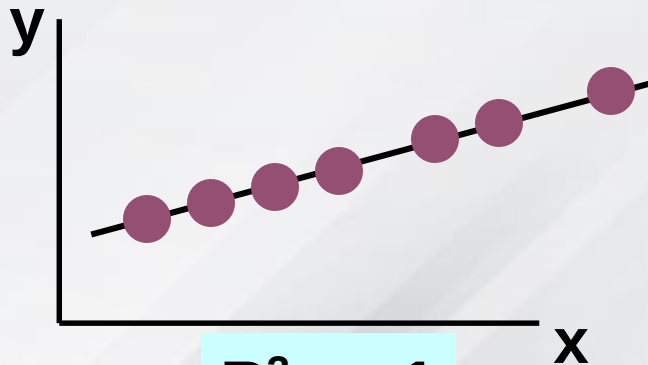


$$R^2 = 1$$

$$R^2 = 1$$

**Perfect linear relationship
between x and y:**

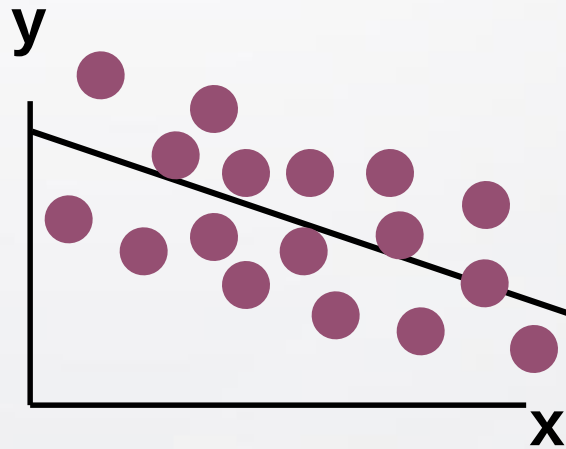
**100% of the variation in y is
explained by variation in x**



$$R^2 = +1$$

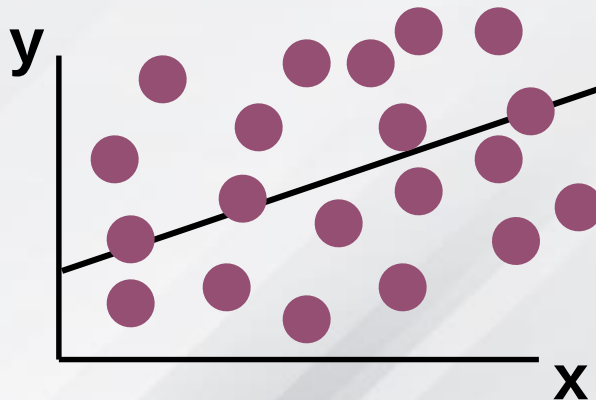
Examples of Approximate R^2 Values

(continued)



$$0 < R^2 < 1$$

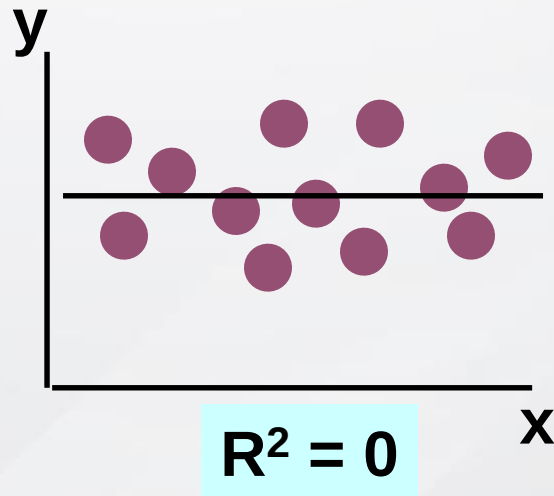
**Weaker linear relationship
between x and y:**



**Some but not all of the
variation in y is explained
by variation in x**

Examples of Approximate R^2 Values

(continued)



$$R^2 = 0$$

**No linear relationship
between x and y:**

**The value of Y does not
depend on x. (None of the
variation in y is explained
by variation in x)**

Chapter 15
Multiple Regression
Analysis
and Model Building

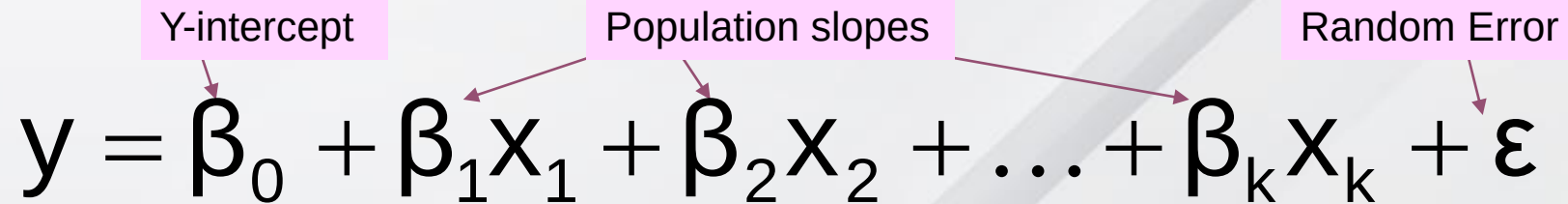
6



The Multiple Regression Model

Idea: Examine the linear relationship between 1 dependent (y) & 2 or more independent variables (x_i)

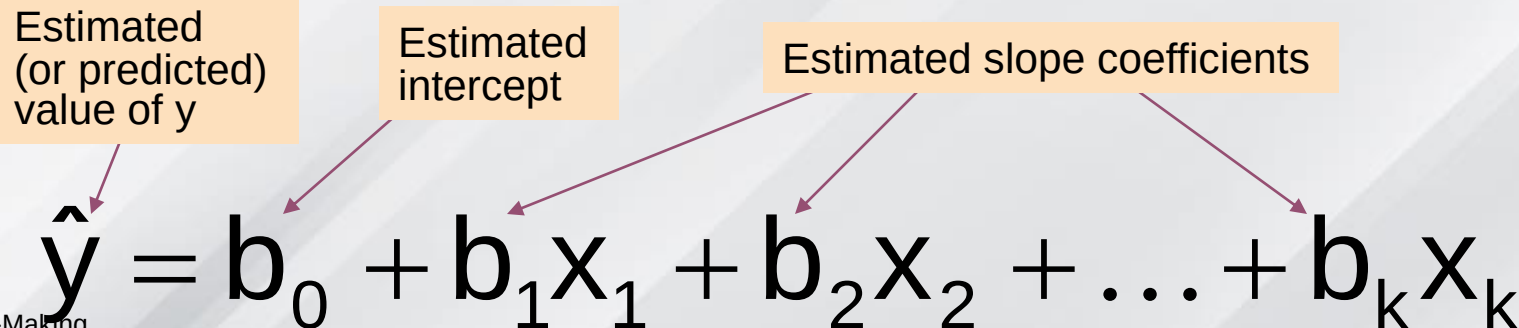
Population model:



The diagram shows the population regression model equation: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$. Three labels with arrows point to specific parts of the equation: 'Y-intercept' points to β_0 , 'Population slopes' points to β_1 , and 'Random Error' points to ε .

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

Estimated multiple regression model:



The diagram shows the estimated multiple regression model equation: $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$. Four labels with arrows point to specific parts of the equation: 'Estimated (or predicted) value of y' points to $\hat{y}$, 'Estimated intercept' points to b_0 , and 'Estimated slope coefficients' points to both b_1 and b_2 .

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

The Correlation Matrix

- Correlation between the dependent variable and selected independent variables can be found using Excel:
 - Formula Tab: Data Analysis / Correlation
- Can check for statistical significance of correlation with a t test

Example

- A distributor of frozen desert pies wants to evaluate factors thought to influence demand
 - Dependent variable: Pie sales (units per week)
 - Independent variables: Price (in \$)
Advertising (\$100's)
- Data are collected for 15 weeks



Pie Sales Model

Week	Pie Sales	Price (\$)	Advertising (\$100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

Multiple regression model:

$$\widehat{\text{Sales}} = b_0 + b_1 (\text{Price}) + b_2 (\text{Advertising})$$

Correlation matrix:

	Pie Sales	Price	Advertising
Pie Sales	1		
Price	-0.44327	1	
Advertising	0.55632	0.03044	1



Pie Sales Correlation Matrix

	Pie Sales	Price	Advertising
Pie Sales	1		
Price	-0.44327	1	
Advertising	0.55632	0.03044	1

- Price vs. Sales : $r = -0.44327$
 - There is a **negative** association between price and sales
- Advertising vs. Sales : $r = 0.55632$
 - There is a **positive** association between advertising and sales



Multiple Regression Output

<i>Regression Statistics</i>						
Multiple R	0.72213	 $\text{Sales} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$				
R Square	0.52148					
Adjusted R Square	0.44172					
Standard Error	47.46341					
Observations	15					
ANOVA	df	SS	MS	F	Significance F	
Regression	2	29460.027	14730.013	6.53861	0.01201	
Residual	12	27033.306	2252.776			
Total	14	56493.333				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

The Multiple Regression Equation

$$\widehat{\text{Sales}} = 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising})$$

where

Sales is in number of pies per week

Price is in \$

Advertising is in \$100's.

$b_1 = -24.975$: sales will decrease, on average, by 24.975 pies per week for each \$1 increase in selling price, net of the effects of changes due to advertising

$b_2 = 74.131$: sales will increase, on average, by 74.131 pies per week for each \$100 increase in advertising, net of the effects of changes due to price



Using The Model to Make Predictions

Predict sales for a week in which the selling price is \$5.50 and advertising is \$350:

$$\begin{aligned}\widehat{\text{Sales}} &= 306.526 - 24.975(\text{Price}) + 74.131(\text{Advertising}) \\ &= 306.526 - 24.975(5.50) + 74.131(3.5) \\ &= 428.62\end{aligned}$$

Predicted sales
is 428.62 pies

Note that Advertising is
in \$100's, so \$350
means that $x_2 = 3.5$

Is the Model Significant?

- F-Test for Overall Significance of the Model
- Shows if there is a linear relationship between all of the x variables considered together and y
- Use F test statistic
- Hypotheses:
 - $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ (no linear relationship)
 - H_A : at least one $\beta_i \neq 0$ (at least one independent variable affects y)

F-Test for Overall Significance

(continued)



Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

ANOVA	df	SS	MS	F	Significance F
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

P-value for the F-Test

F-Test for Overall Significance

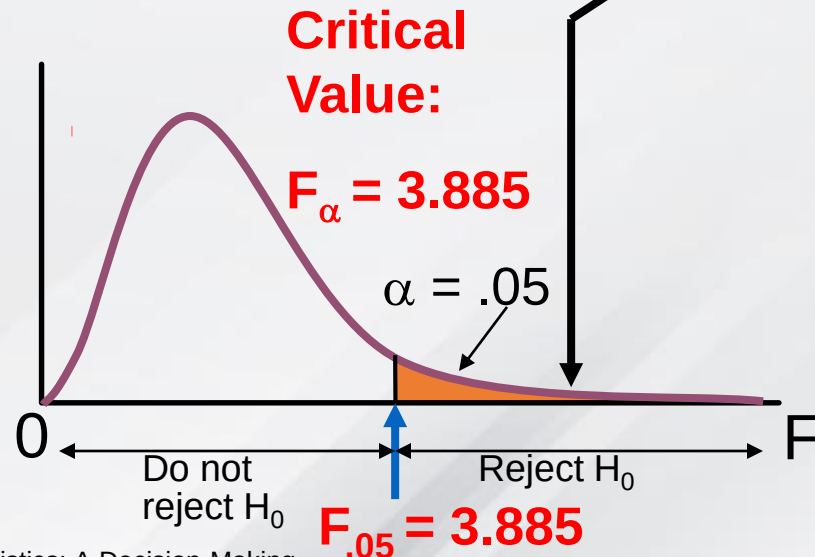
(continued)

$$H_0: \beta_1 = \beta_2 = 0$$

$$H_A: \beta_1 \text{ and } \beta_2 \text{ not both zero}$$

$$\alpha = .05$$

$$df_1 = 2 \quad df_2 = 12$$



Test Statistic:

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

The regression model does explain a significant portion of the variation in pie sales

(There is evidence that at least one independent variable affects y)

Are Individual Variables Significant?

- Use t-tests of individual variable slopes
- Shows if there is a linear relationship between the variable x_i and y
- Hypotheses:
 - $H_0: \beta_i = 0$ (no linear relationship)
 - $H_A: \beta_i \neq 0$ (linear relationship does exist between x_i and y)

Are Individual Variables Significant?

(continued)

Regression Statistics		t-value for Price is $t = -2.306$, with p-value .0398 t-value for Advertising is $t = 2.855$, with p-value .0145				
Multiple R	0.72213					
R Square	0.52148					
Adjusted R Square	0.44172					
Standard Error	47.46341					
Observations	15					
ANOVA		df	SS	MS	F	Significance F
Regression		2	29460.027	14730.013	6.53861	0.01201
Residual		12	27033.306	2252.776		
Total		14	56493.333			
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888



Inferences about the Slope: t Test Example

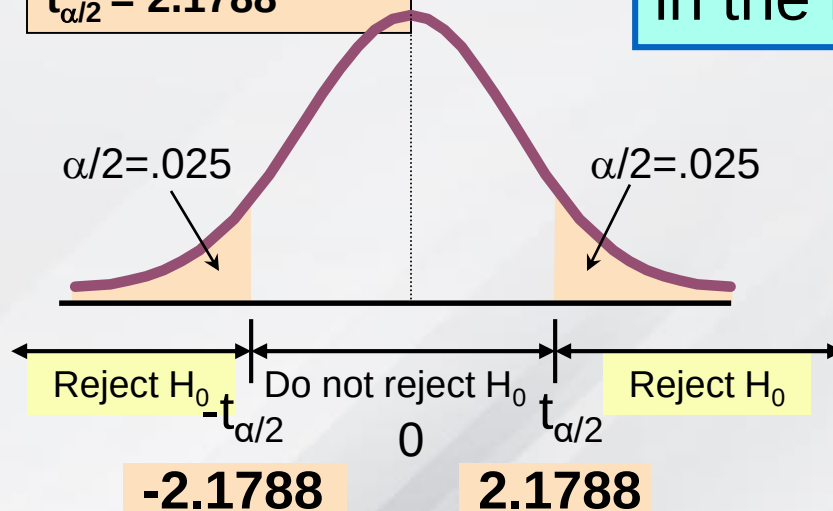
$$H_0: \beta_i = 0$$

$$H_A: \beta_i \neq 0$$

$$\text{d.f.} = 15 - 2 - 1 = 12$$

$$\alpha = .05$$

$$t_{\alpha/2} = 2.1788$$



From Excel output:

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Price	-24.97509	10.83213	-2.30565	0.03979
Advertising	74.13096	25.96732	2.85478	0.01449

The test statistic for each variable falls in the rejection region (p-values < .05)

Decision:

Reject H_0 for each variable

Conclusion:

There is evidence that both Price and Advertising affect pie sales at $\alpha = .05$

Confidence Interval Estimate for the Slope

Confidence interval for the population slope β_1
(the effect of changes in price on pie sales):

	<i>Coefficients</i>	<i>Standard Error</i>	...	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	306.52619	114.25389	...	57.58835	555.46404
Price	-24.97509	10.83213	...	-48.57626	-1.37392
Advertising	74.13096	25.96732	...	17.55303	130.70888

Multicollinearity

- Multicollinearity: High correlation exists between two independent variables
- This means the two variables contribute redundant information to the multiple regression model

Multicollinearity

(continued)

- Including two highly correlated independent variables can adversely affect the regression results
 - No new information provided
 - Can lead to unstable coefficients (large standard error and low t-values)
 - Coefficient signs may not match prior expectations

Some Indications of Severe Multicollinearity

- Incorrect signs on the coefficients
- Large change in the value of a previous coefficient when a new variable is added to the model
- A previously significant variable becomes insignificant when a new independent variable is added
- The estimate of the standard deviation of the model increases when a variable is added to the model

Detect Collinearity (Variance Inflationary Factor)

VIF_j is used to measure collinearity:

$$VIF_j = \frac{1}{1 - R_j^2}$$

R_j^2 is the coefficient of determination when the j^{th} independent variable is regressed against the remaining $k - 1$ independent variables

If $VIF_j \geq 5$, x_j is highly correlated with the other explanatory variables

Detect Collinearity in PHStat

Add-Ins/ PHStat / Regression / Multiple Regression ...

Check the “variance inflationary factor (VIF)” box

Regression Analysis	
Price and all other X	
<i>Regression Statistics</i>	
Multiple R	0.030437581
R Square	0.000926446
Adjusted R Square	-0.075925366
Standard Error	1.21527235
Observations	15
VIF	1.000927305

Output for the pie sales example:

- Since there are only two explanatory variables, only one VIF is reported
 - VIF is < 5
 - There is no evidence of collinearity between Price and Advertising

Course Evaluation





رأيك يهمني
أكتب أسمك + تقييمك لدورة البحث العلمي

<https://padlet.com/monaashry77/padlet-xr709opfeqfrbh7t>



مجموعة جامعة
الجميع الذكية
للتعليم القابضة

MIDOCEAN
UNIVERSITY



Thank You

