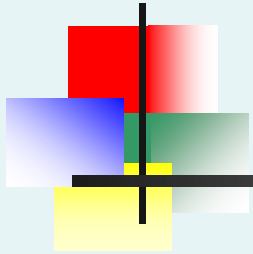


Business Statistics: A First Course Fifth Edition



Chapter 1

Introduction and Data Collection



Learning Objectives

In this chapter you learn:

- How Statistics is used in business
- The sources of data used in business
- The types of data used in business
- The basics of Microsoft Excel
- The basics of Minitab



Why Learn Statistics?

So you are able to make better sense of the ubiquitous use of numbers:

- Business memos
- Business research
- Technical reports
- Technical journals
- Newspaper articles
- Magazine articles



What is statistics?

- A branch of mathematics taking and transforming numbers into useful information for decision makers
- Methods for processing & analyzing numbers
- Methods for helping reduce the uncertainty inherent in decision making



Why Study Statistics?

Decision Makers Use Statistics To:

- Present and describe business data and information properly
- Draw conclusions about large groups of individuals or items, using information collected from subsets of the individuals or items.
- Make reliable forecasts about a business activity
- Improve business processes



Types of Statistics

■ **Statistics**

- The branch of mathematics that transforms data into useful information for decision makers.



Descriptive Statistics

Collecting, summarizing, and describing data



Inferential Statistics

Drawing conclusions and/or making decisions concerning a population based only on sample data

Descriptive Statistics

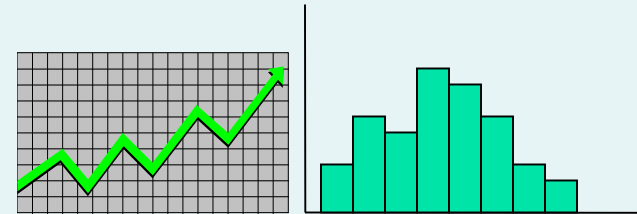
- Collect data

- e.g., Survey



- Present data

- e.g., Tables and graphs



- Characterize data

- e.g., Sample mean = $\frac{\sum X_i}{n}$

Inferential Statistics

- Estimation
 - e.g., Estimate the population mean weight using the sample mean weight
- Hypothesis testing
 - e.g., Test the claim that the population mean weight is 120 pounds



Drawing conclusions about a large group of individuals based on a subset of the large group.



Basic Vocabulary of Statistics

VARIABLE

A **variable** is a characteristic of an item or individual.

DATA

Data are the different values associated with a variable.

OPERATIONAL DEFINITIONS

Data values are meaningless unless their variables have **operational definitions**, universally accepted meanings that are clear to all associated with an analysis.



Basic Vocabulary of Statistics

POPULATION

A **population** consists of all the items or individuals about which you want to draw a conclusion.

SAMPLE

A **sample** is the portion of a population selected for analysis.

PARAMETER

A **parameter** is a numerical measure that describes a characteristic of a population.

STATISTIC

A **statistic** is a numerical measure that describes a characteristic of a sample.



Population vs. Sample

Population



Measures used to describe the population are called **parameters**

Sample



Measures computed from sample data are called **statistics**



Why Collect Data?

- A marketing research analyst needs to assess the effectiveness of a new television advertisement.
- A pharmaceutical manufacturer needs to determine whether a new drug is more effective than those currently in use.
- An operations manager wants to monitor a manufacturing process to find out whether the quality of the product being manufactured is conforming to company standards.
- An auditor wants to review the financial transactions of a company in order to determine whether the company is in compliance with generally accepted accounting principles.



Sources of Data

- **Primary Sources:** The data collector is the one using the data for analysis
 - Data from a political survey
 - Data collected from an experiment
 - Observed data
- **Secondary Sources:** The person performing data analysis is not the data collector
 - Analyzing census data
 - Examining data from print journals or data published on the internet.

Sources of data fall into four categories



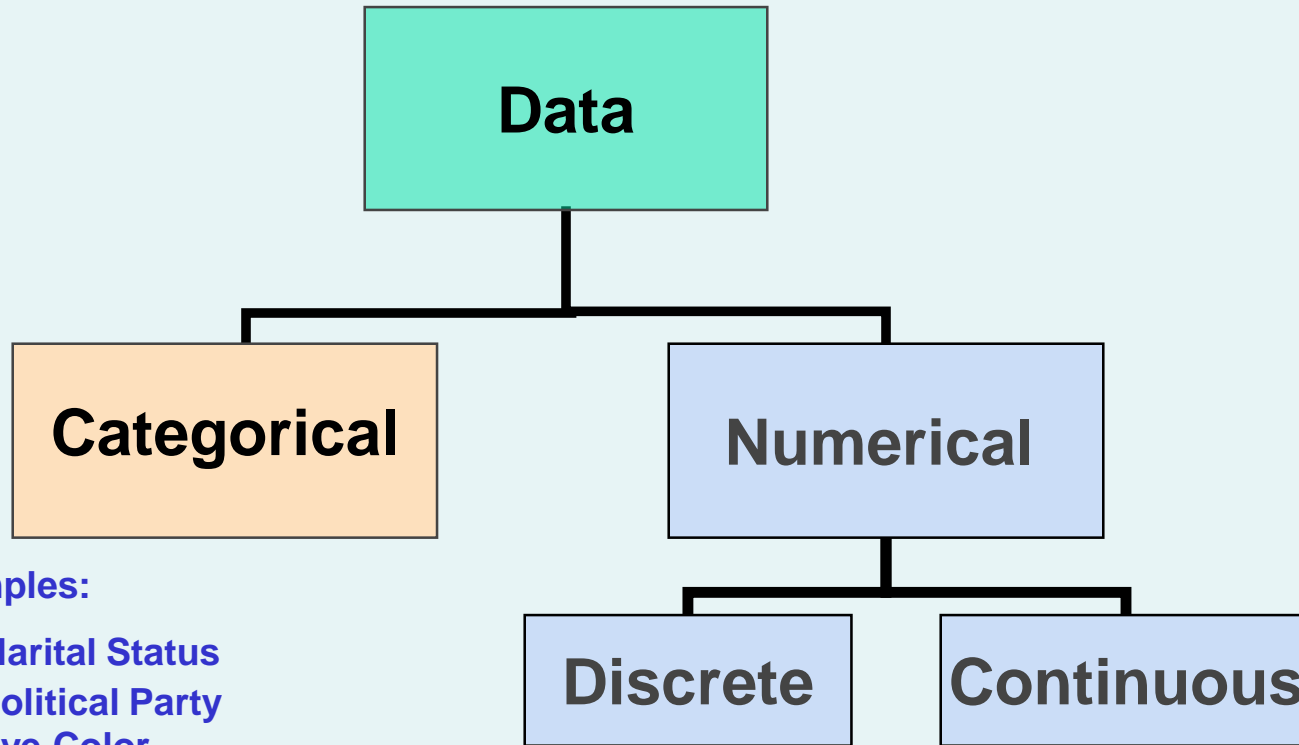
- Data distributed by an organization or an individual
- A designed experiment
- A survey
- An observational study



Types of Variables

- **Categorical** (qualitative) variables have values that can only be placed into categories, such as “yes” and “no.”
- **Numerical** (quantitative) variables have values that represent quantities.

Types of Data



Examples:

- Marital Status
 - Political Party
 - Eye Color
- (Defined categories)

Examples:

- Number of Children
 - Defects per hour
- (Counted items)

Examples:

- Weight
 - Voltage
- (Measured characteristics)

Personal Computer Programs Used For Statistics



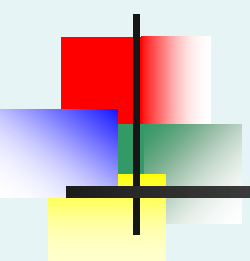
- Minitab
 - A statistical package to perform statistical analysis
 - Designed to perform analysis as accurately as possible
- Microsoft Excel
 - A multi-functional data analysis tool
 - Can perform many functions but none as well as programs that are dedicated to a single function.
- Both Minitab and Excel use worksheets to store data



Minitab & Microsoft Excel Terms

- When you use Minitab or Microsoft Excel, you place the data you have collected in **worksheets**.
- The intersections of the columns and rows of worksheets form boxes called **cells**.
- If you want to refer to a group of cells that forms a contiguous rectangular area, you can use a **cell range**.
- Worksheets exist inside a **workbook in Excel and inside a Project in Minitab**.
- Both worksheets and projects can contain both data, summaries, and charts.

You are using programs properly if you can



- Understand how to operate the program
- Understand the underlying statistical concepts
- Understand how to organize and present information
- Know how to review results for errors
- Make secure and clearly named backups of your work

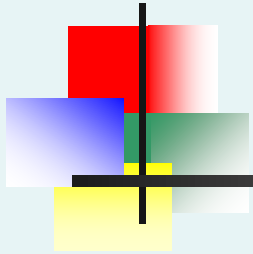


Chapter Summary

In this chapter, we have

- Reviewed why a manager needs to know statistics
- Introduced key definitions:
 - Population vs. Sample
 - Primary vs. Secondary data types
 - Categorical vs. Numerical data
- Examined descriptive vs. inferential statistics
- Reviewed data types
- Discussed Minitab and Microsoft Excel terms

Business Statistics: A First Course Fifth Edition



Chapter 2

Presenting Data in Tables and Charts

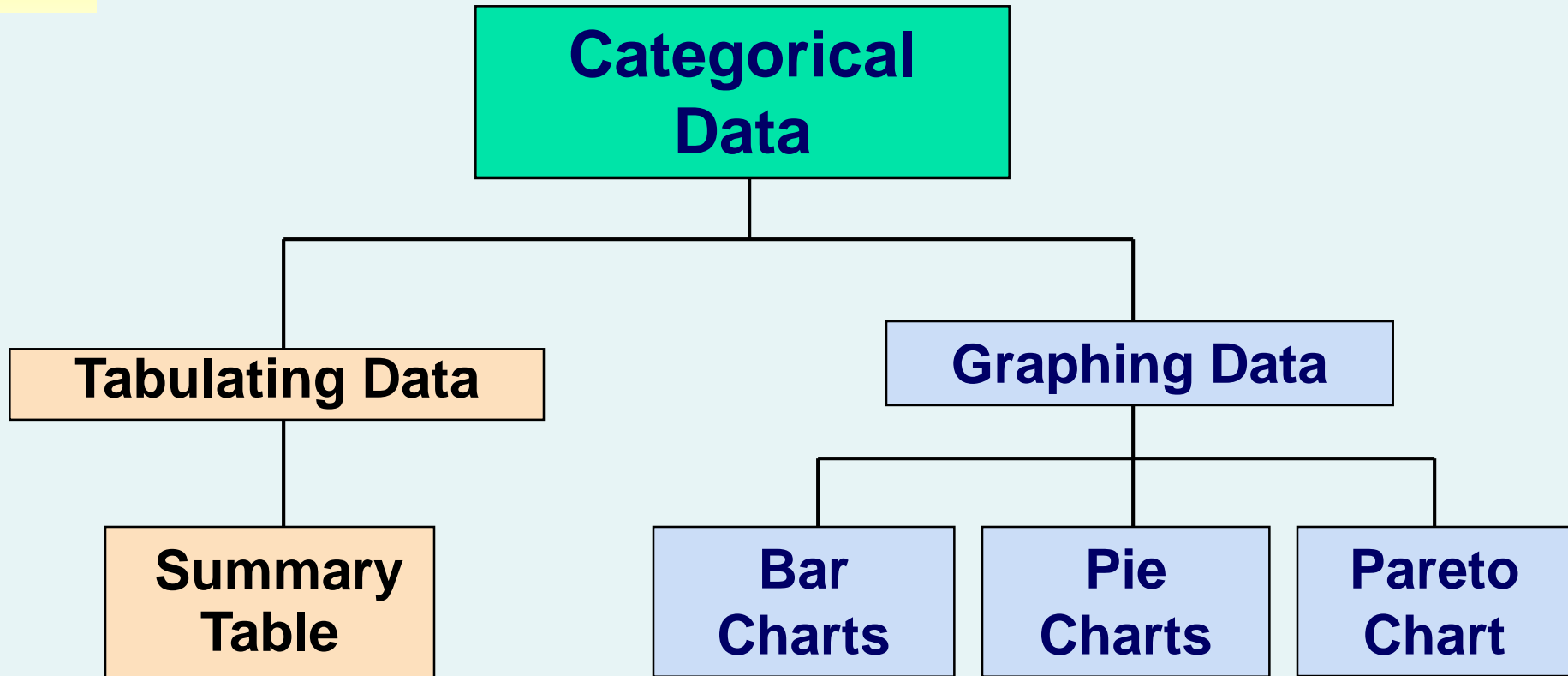


Learning Objectives

In this chapter you learn:

- To develop tables and charts for categorical data
- To develop tables and charts for numerical data
- The principles of properly presenting graphs

Categorical Data Are Summarized By Tables & Graphs





Organizing Categorical Data: Summary Table

- A **summary table** indicates the frequency, amount, or percentage of items in a set of categories so that you can see differences between categories.

| Banking Preference? | Percent |
|---------------------------------|----------------|
| ATM | 16% |
| Automated or live telephone | 2% |
| Drive-through service at branch | 17% |
| In person at branch | 41% |
| Internet | 24% |

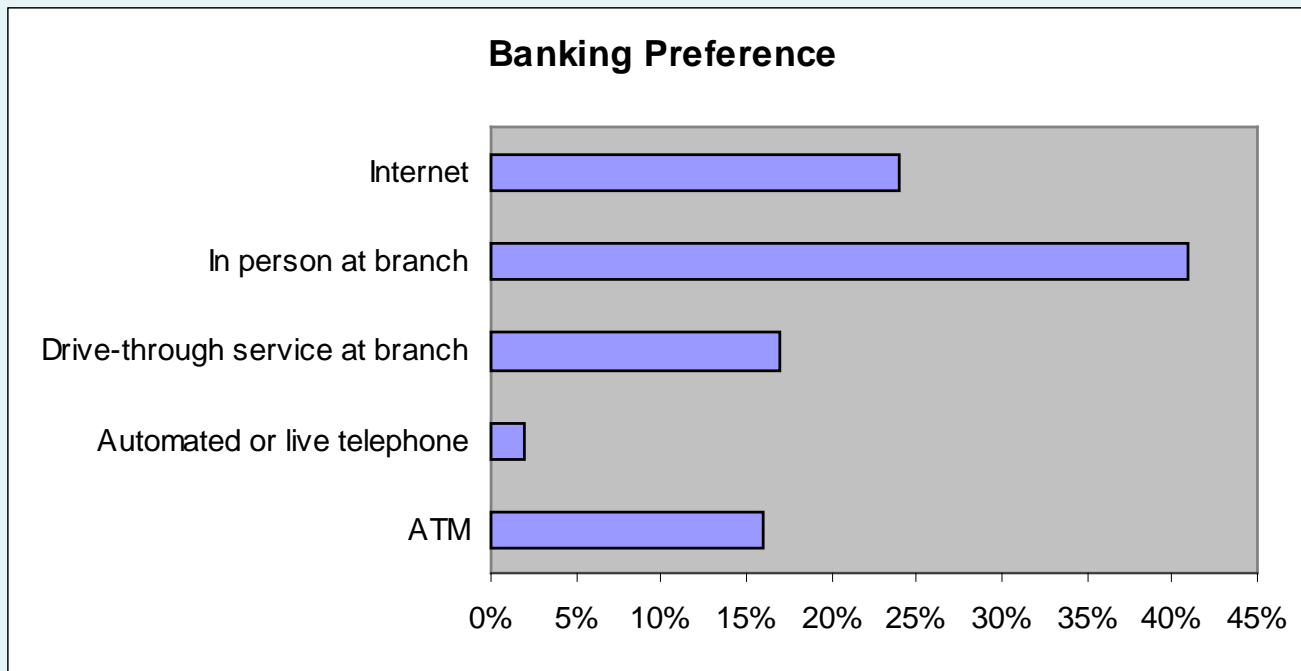


Bar and Pie Charts

- Bar charts and Pie charts are often used for categorical data
- Length of bar or size of pie slice shows the frequency or percentage for each category

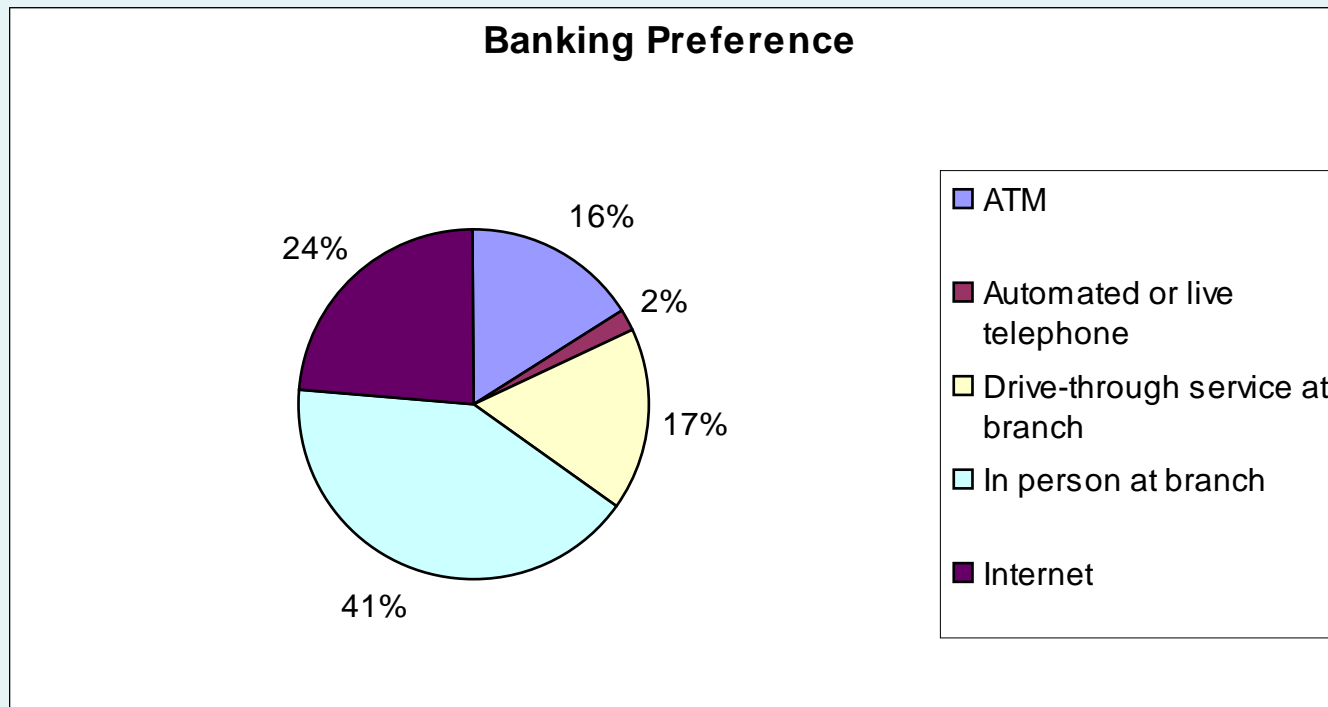
Organizing Categorical Data: Bar Chart

- In a **bar chart**, a bar shows each category, the length of which represents the amount, frequency or percentage of values falling into a category.



Organizing Categorical Data: Pie Chart

- The **pie chart** is a circle broken up into slices that represent categories. The size of each slice of the pie varies according to the percentage in each category.



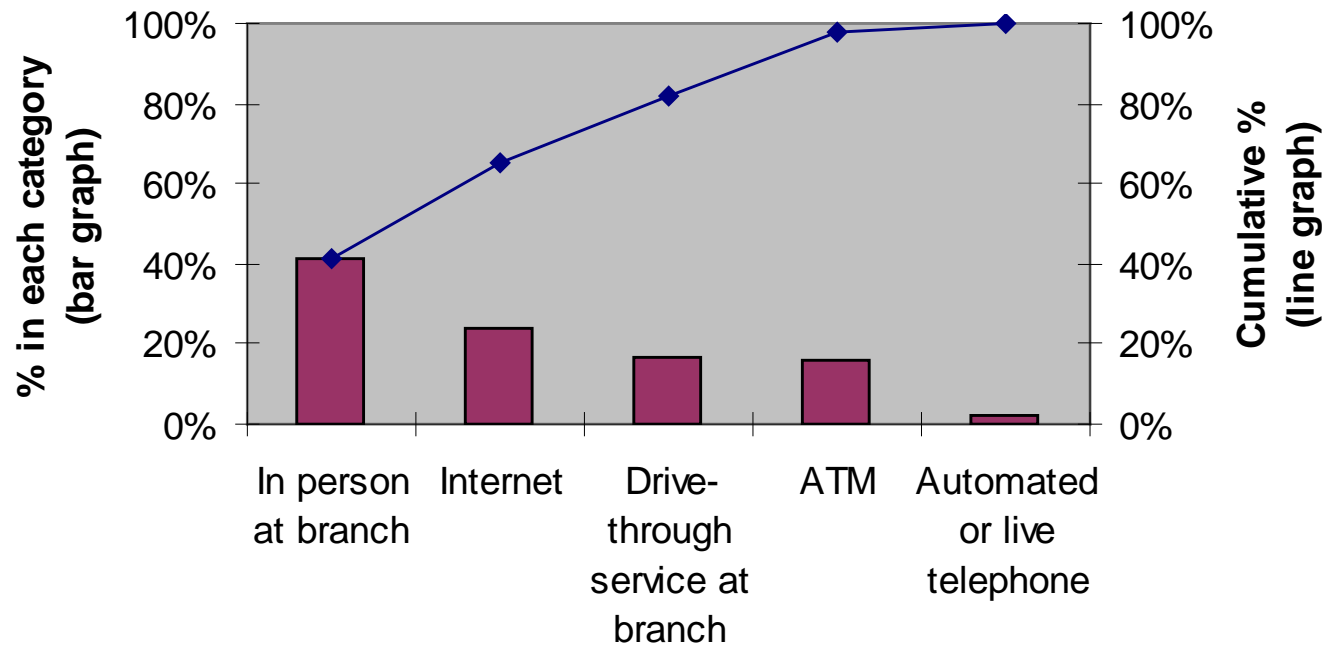
Organizing Categorical Data: Pareto Chart



- Used to portray categorical data (nominal scale)
- A vertical bar chart, where categories are shown in descending order of frequency
- A cumulative polygon is shown in the same graph
- Used to separate the “vital few” from the “trivial many”

Organizing Categorical Data: Pareto Chart

Pareto Chart For Banking Preference



Tables and Charts for Numerical Data

Numerical Data

```
graph TD; A[Numerical Data] --> B[Ordered Array]; A --> C[Frequency Distributions and Cumulative Distributions]; B --> D[Stem-and-Leaf Display]; C --> E[Histogram]; C --> F[Polygon]; C --> G[Ogive];
```

Ordered Array

**Frequency Distributions
and
Cumulative Distributions**

**Stem-and-Leaf
Display**

Histogram

Polygon

Ogive

Organizing Numerical Data: Ordered Array

- An **ordered array** is a sequence of data, in rank order, from the smallest value to the largest value.
- Shows range (minimum value to maximum value)
- May help identify outliers (unusual observations)

| | | | | | | |
|-----------------------------------------------------|-----------------------|----|----|----|----|----|
| Age of Surveyed College Students | Day Students | | | | | |
| | 16 | 17 | 17 | 18 | 18 | 18 |
| | 19 | 19 | 20 | 20 | 21 | 22 |
| | 22 | 25 | 27 | 32 | 38 | 42 |
| | Night Students | | | | | |
| | 18 | 18 | 19 | 19 | 20 | 21 |
| | 23 | 28 | 32 | 33 | 41 | 45 |



Stem-and-Leaf Display

- A simple way to see how the data are distributed and where concentrations of data exist

METHOD: Separate the sorted data series into leading digits (the **stems**) and the trailing digits (the **leaves**)

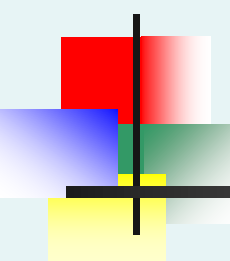
Organizing Numerical Data: Stem and Leaf Display

- A **stem-and-leaf display** organizes data into groups (called stems) so that the values within each group (the leaves) branch out to the right on each row.

Age of College Students

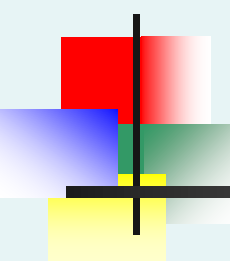
| Age of Surveyed College Students | Day Students | | | | | |
|-------------------------------------------|----------------|----|----|----|----|----|
| | 16 | 17 | 17 | 18 | 18 | 18 |
| | 19 | 19 | 20 | 20 | 21 | 22 |
| | 22 | 25 | 27 | 32 | 38 | 42 |
| | Night Students | | | | | |
| | 18 | 18 | 19 | 19 | 20 | 21 |
| | 23 | 28 | 32 | 33 | 41 | 45 |

| Day Students | | Night Students | |
|--------------|----------|----------------|------|
| Stem | Leaf | Stem | Leaf |
| 1 | 67788899 | 1 | 8899 |
| 2 | 0012257 | 2 | 0138 |
| 3 | 28 | 3 | 23 |
| 4 | 2 | 4 | 15 |



Organizing Numerical Data: Frequency Distribution

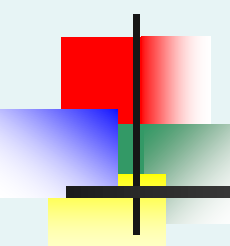
- The **frequency distribution** is a summary table in which the data are arranged into numerically ordered classes.
- You must give attention to selecting the appropriate *number* of **class groupings** for the table, determining a suitable *width* of a class grouping, and establishing the *boundaries* of each class grouping to avoid overlapping.
- The number of classes depends on the number of values in the data. With a larger number of values, typically there are more classes. In general, a frequency distribution should have at least 5 but no more than 15 classes.
- To determine the **width of a class interval**, you divide the **range** (Highest value–Lowest value) of the data by the number of class groupings desired.



Organizing Numerical Data: Frequency Distribution Example

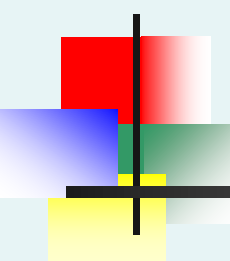
Example: A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

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



Organizing Numerical Data: Frequency Distribution Example

- Sort raw data in ascending order:
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 15)**
- Compute class interval (width): **10 ($46/5$ then round up)**
- Determine class boundaries (limits):
 - **Class 1: 10 to less than 20**
 - **Class 2: 20 to less than 30**
 - **Class 3: 30 to less than 40**
 - **Class 4: 40 to less than 50**
 - **Class 5: 50 to less than 60**
- Compute class midpoints: **15, 25, 35, 45, 55**
- Count observations & assign to classes



Organizing Numerical Data: Frequency Distribution 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

| Class | Frequency | Relative Frequency | Percentage |
|----------------------------|------------------|---------------------------|-------------------|
| 10 but less than 20 | 3 | .15 | 15 |
| 20 but less than 30 | 6 | .30 | 30 |
| 30 but less than 40 | 5 | .25 | 25 |
| 40 but less than 50 | 4 | .20 | 20 |
| 50 but less than 60 | 2 | .10 | 10 |
| Total | 20 | 1.00 | 100 |



Tabulating Numerical Data: Cumulative Frequency

Data in ordered array:

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

| Class | Frequency | Percentage | Cumulative Frequency | Cumulative Percentage |
|----------------------------|------------------|-------------------|-----------------------------|------------------------------|
| 10 but less than 20 | 3 | 15 | 3 | 15 |
| 20 but less than 30 | 6 | 30 | 9 | 45 |
| 30 but less than 40 | 5 | 25 | 14 | 70 |
| 40 but less than 50 | 4 | 20 | 18 | 90 |
| 50 but less than 60 | 2 | 10 | 20 | 100 |
| Total | 20 | 100 | | |



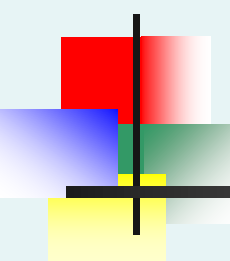
Why Use a Frequency Distribution?

- It condenses the raw data into a more useful form
- It allows for a quick visual interpretation of the data
- It enables the determination of the major characteristics of the data set including where the data are concentrated / clustered



Frequency Distributions: Some Tips

- Different class boundaries may provide different pictures for the same data (especially for smaller data sets)
- Shifts in data concentration may show up when different class boundaries are chosen
- As the size of the data set increases, the impact of alterations in the selection of class boundaries is greatly reduced
- When comparing two or more groups with different sample sizes, you must use either a relative frequency or a percentage distribution



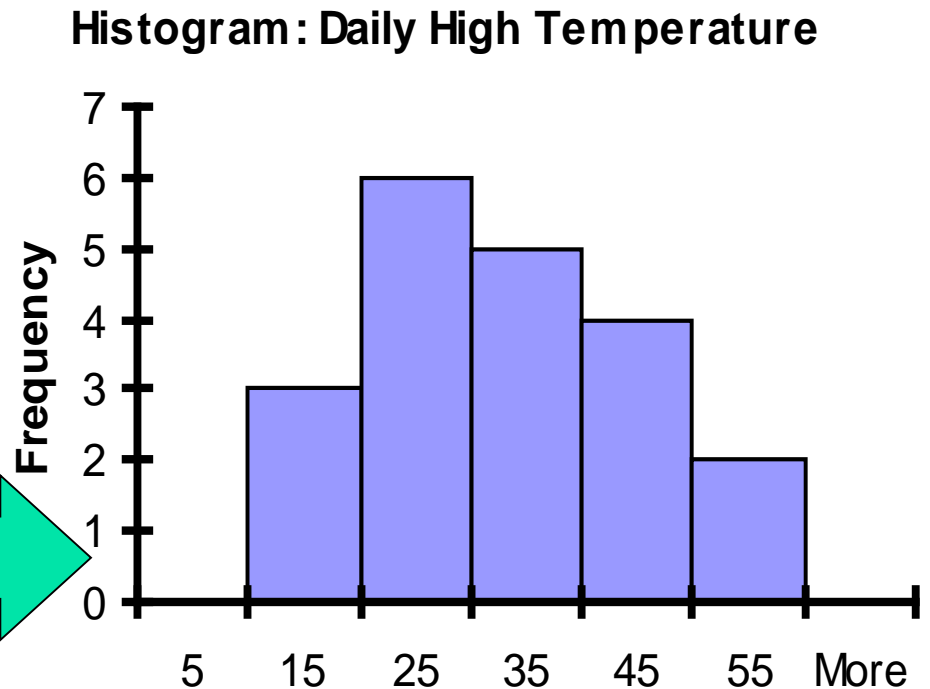
Organizing Numerical Data: The Histogram

- A vertical bar chart of the data in a frequency distribution is called a **histogram**.
- In a histogram there are no gaps between adjacent bars.
- The **class boundaries** (or **class midpoints**) are shown on the horizontal axis.
- The vertical axis is either **frequency**, **relative frequency**, or **percentage**.
- The height of the bars represent the frequency, relative frequency, or percentage.

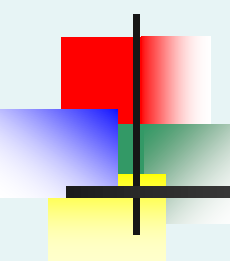
Organizing Numerical Data: The Histogram

| Class | Frequency | Relative Frequency | Percentage |
|---------------------|-----------|--------------------|------------|
| 10 but less than 20 | 3 | .15 | 15 |
| 20 but less than 30 | 6 | .30 | 30 |
| 30 but less than 40 | 5 | .25 | 25 |
| 40 but less than 50 | 4 | .20 | 20 |
| 50 but less than 60 | 2 | .10 | 10 |
| Total | 20 | 1.00 | 100 |

(In a percentage histogram the vertical axis would be defined to show the percentage of observations per class)



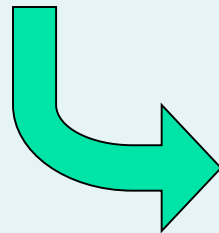
Organizing Numerical Data: The Polygon



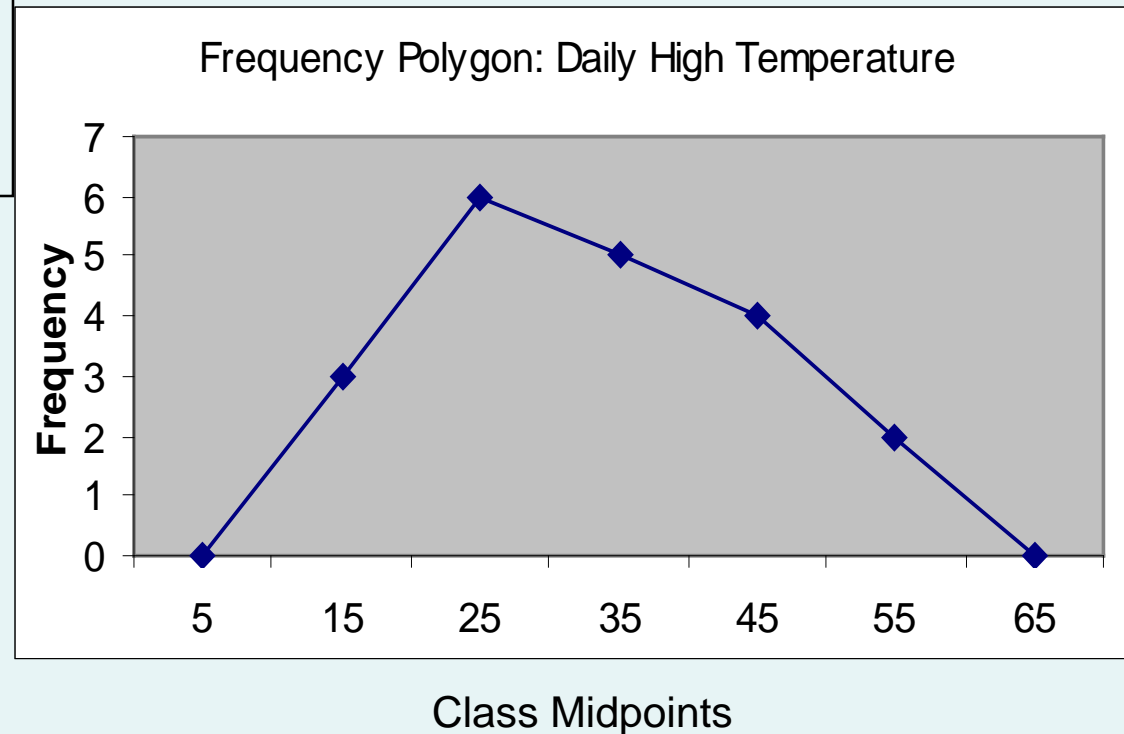
- A **percentage polygon** is formed by having the midpoint of each class represent the data in that class and then connecting the sequence of midpoints at their respective class percentages.
- The **cumulative percentage polygon**, or **ogive**, displays the variable of interest along the X axis, and the cumulative percentages along the Y axis.
- Useful when there are two or more groups to compare.

Graphing Numerical Data: The Frequency Polygon

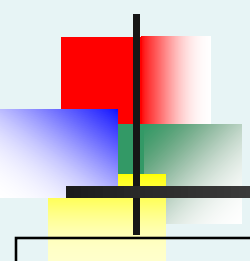
| Class | Class Midpoint | Frequency |
|---------------------|----------------|-----------|
| 10 but less than 20 | 15 | 3 |
| 20 but less than 30 | 25 | 6 |
| 30 but less than 40 | 35 | 5 |
| 40 but less than 50 | 45 | 4 |
| 50 but less than 60 | 55 | 2 |



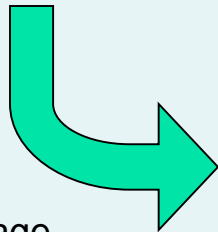
(In a percentage polygon the vertical axis would be defined to show the percentage of observations per class)



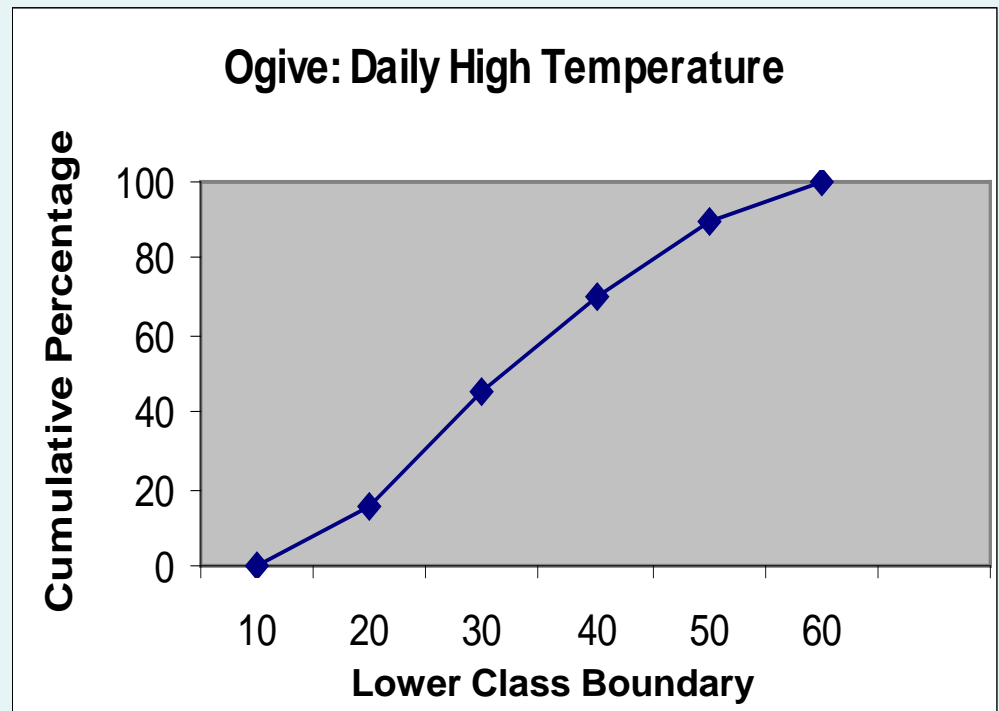
Graphing Cumulative Frequencies: The Ogive (Cumulative % Polygon)



| Class | Lower class boundary | % less than lower boundary |
|---------------------|-----------------------------|-----------------------------------|
| 10 but less than 20 | 10 | 15 |
| 20 but less than 30 | 20 | 45 |
| 30 but less than 40 | 30 | 70 |
| 40 but less than 50 | 40 | 90 |
| 50 but less than 60 | 50 | 100 |



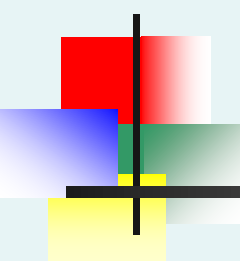
(In an ogive the percentage of the observations less than each lower class boundary are plotted versus the lower class boundaries.)





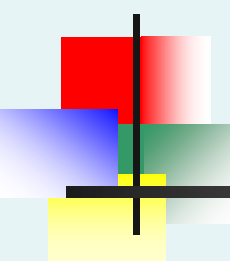
Cross Tabulations

- Used to study patterns that may exist between two or more categorical variables.
- Cross tabulations can be presented in Contingency Tables



Cross Tabulations: The Contingency Table

- A **cross-classification** (or **contingency**) **table** presents the results of two categorical variables. The joint responses are classified so that the categories of one variable are located in the rows and the categories of the other variable are located in the columns.
- The cell is the intersection of the row and column and the value in the cell represents the data corresponding to that specific pairing of row and column categories.



Cross Tabulations: The Contingency Table

A survey was conducted to study the importance of brand name to consumers as compared to a few years ago. The results, classified by gender, were as follows:

| Importance of Brand Name | Male | Female | Total |
|---------------------------------|-------------|---------------|--------------|
| More | 450 | 300 | 750 |
| Equal or Less | 3300 | 3450 | 6750 |
| Total | 3750 | 3750 | 7500 |

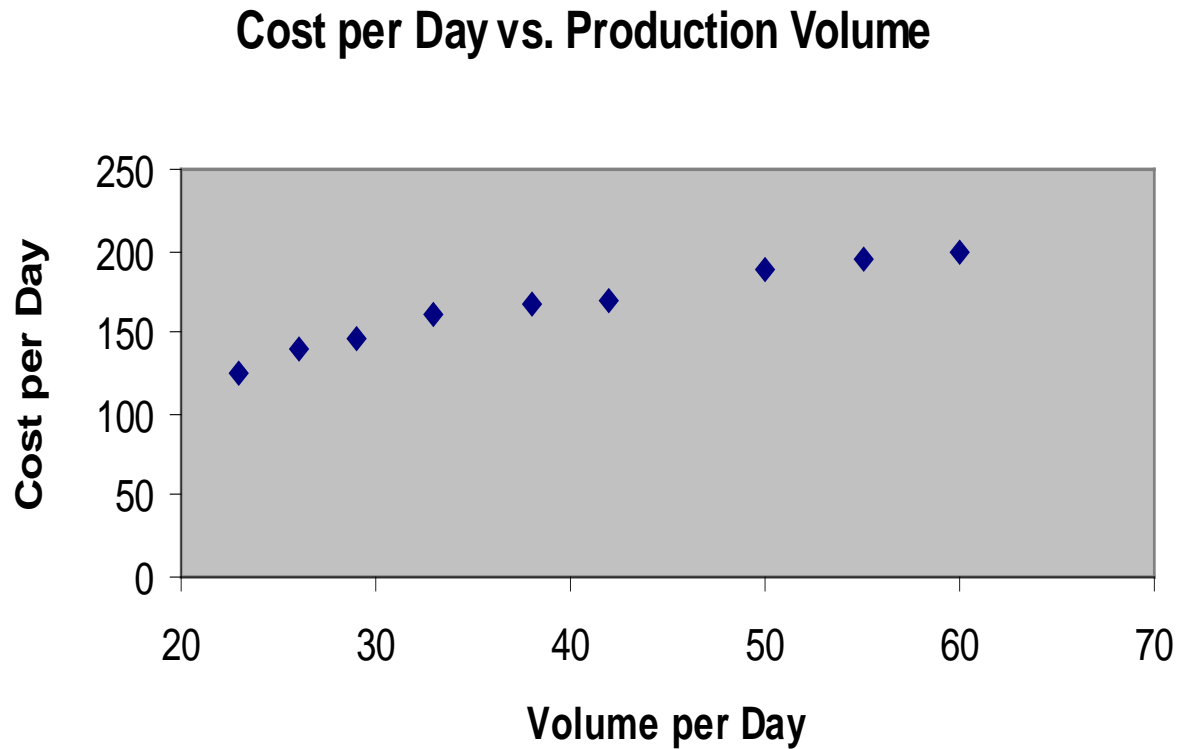


Scatter Plots

- **Scatter plots** are used for numerical data consisting of paired observations taken from two numerical variables
- One variable is measured on the vertical axis and the other variable is measured on the horizontal axis
- Scatter plots are used to examine possible relationships between two numerical variables

Scatter Plot Example

| Volume per day | Cost per day |
|----------------|--------------|
| 23 | 125 |
| 26 | 140 |
| 29 | 146 |
| 33 | 160 |
| 38 | 167 |
| 42 | 170 |
| 50 | 188 |
| 55 | 195 |
| 60 | 200 |



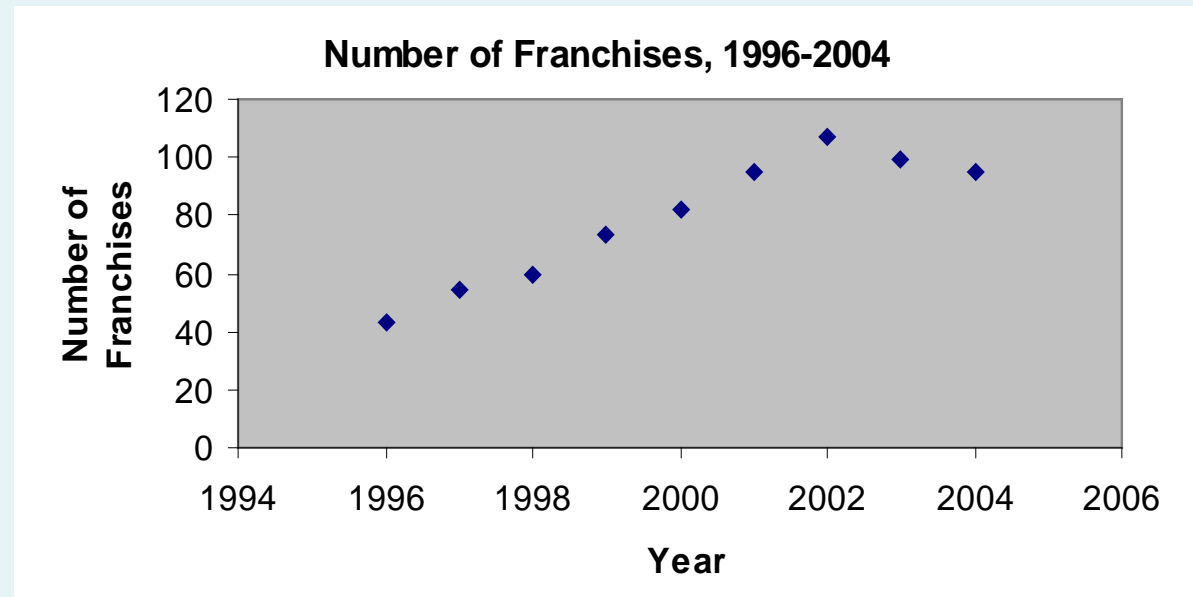


Time Series Plot

- A **Time Series Plot** is used to study patterns in the values of a numeric variable over time
- The Time Series Plot:
 - Numeric variable is measured on the vertical axis and the time period is measured on the horizontal axis

Time Series Plot Example

| Year | Number of Franchises |
|------|----------------------|
| 1996 | 43 |
| 1997 | 54 |
| 1998 | 60 |
| 1999 | 73 |
| 2000 | 82 |
| 2001 | 95 |
| 2002 | 107 |
| 2003 | 99 |
| 2004 | 95 |





Principles of Excellent Graphs

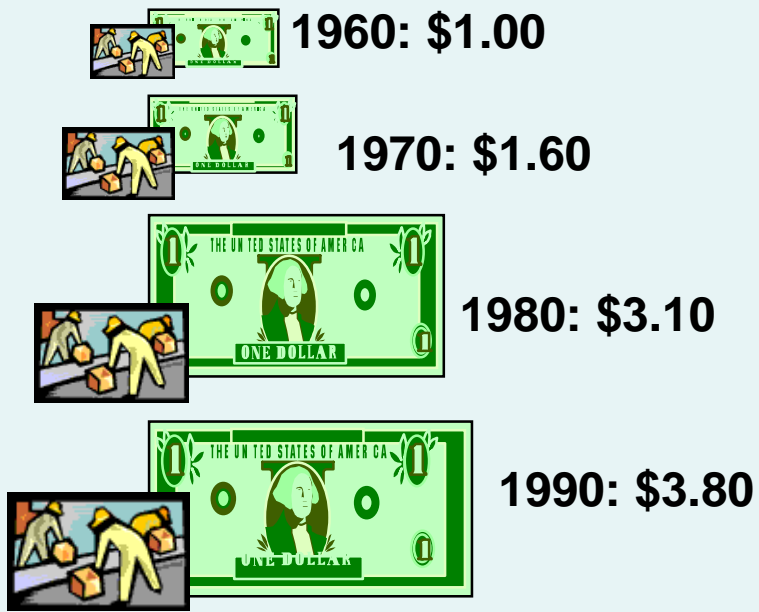
- The graph should not distort the data.
- The graph should not contain unnecessary adornments (sometimes referred to as chart junk).
- The scale on the vertical axis should begin at zero.
- All axes should be properly labeled.
- The graph should contain a title.
- The simplest possible graph should be used for a given set of data.

Graphical Errors: Chart Junk

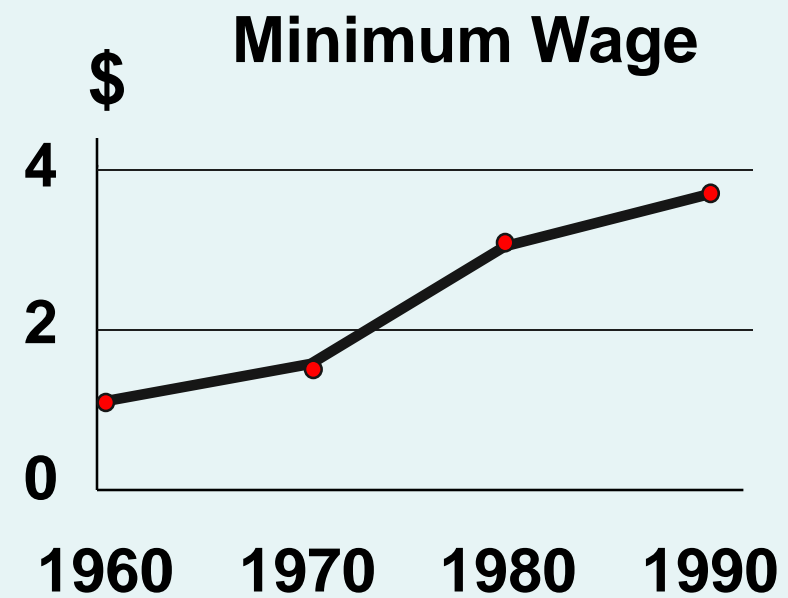


Bad Presentation

Minimum Wage



Good Presentation



Graphical Errors: No Relative Basis



Bad Presentation

A's received by
students.

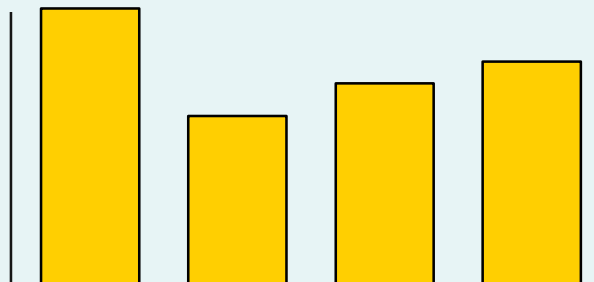
Freq.

300

200

100

0



FR

SO

JR

SR



Good Presentation

A's received by
students.

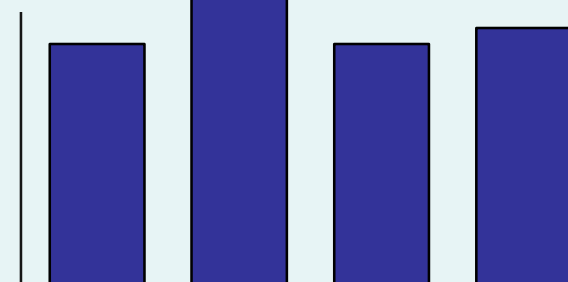
%

30%

20%

10%

0%



FR

SO

JR

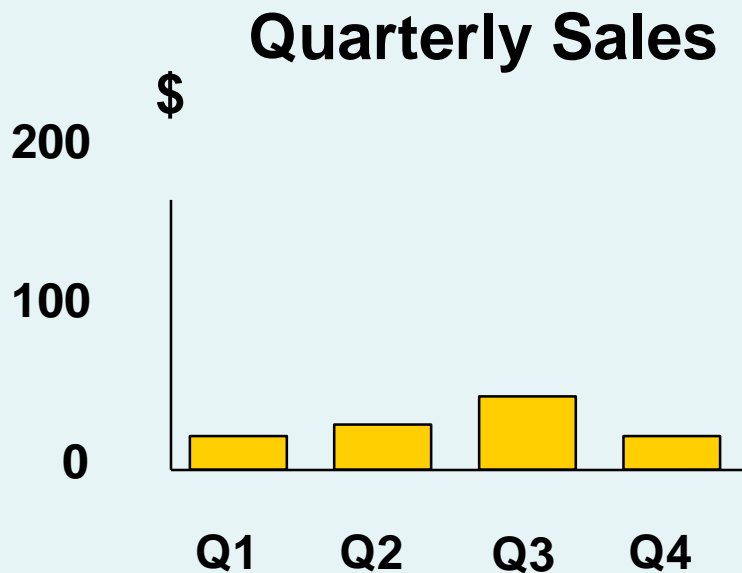
SR

FR = Freshmen, SO = Sophomore, JR = Junior, SR = Senior

Graphical Errors: Compressing the Vertical Axis



Bad Presentation



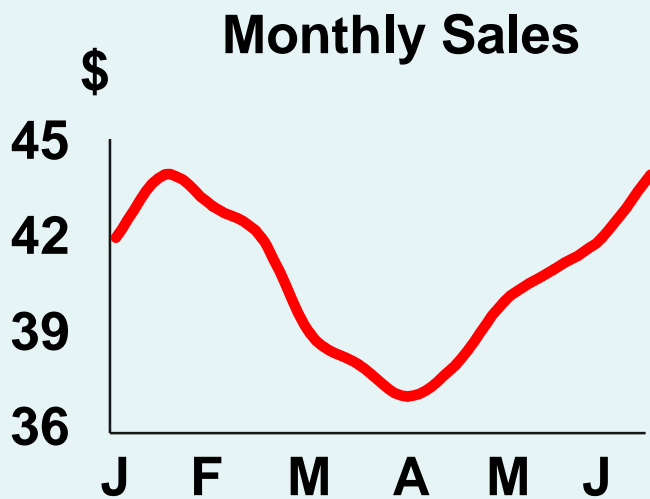
Good Presentation



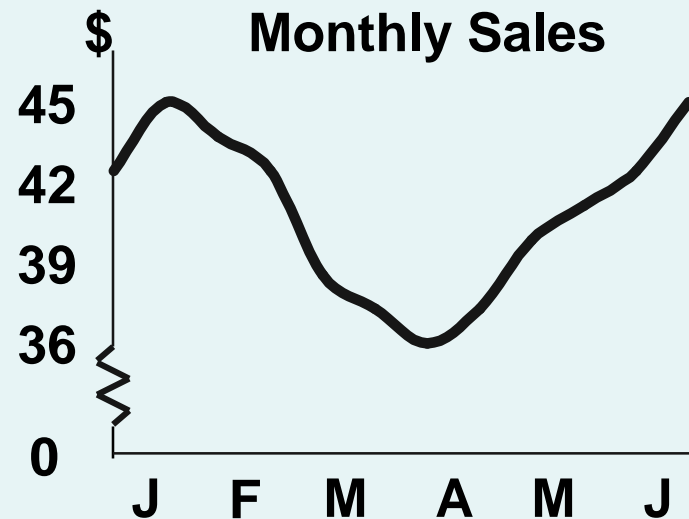
Graphical Errors: No Zero Point on the Vertical Axis



Bad Presentation



✓ Good Presentations



Graphing the first six months of sales

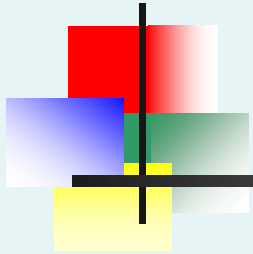


Chapter Summary

In this chapter, we have

- Organized categorical data using the summary table, bar chart, pie chart, and Pareto chart.
- Organized numerical data using the ordered array, stem-and-leaf display, frequency distribution, histogram, polygon, and ogive.
- Examined cross tabulated data using the contingency table.
- Developed scatter plots and time series graphs.
- Examined the do's and don'ts of graphically displaying data.

Business Statistics: A First Course Fifth Edition



Chapter 3

Numerical Descriptive Measures



Learning Objectives

In this chapter, you learn:

- To describe the properties of central tendency, variation, and shape in numerical data
- To calculate descriptive summary measures for a population
- To construct and interpret a boxplot
- To calculate the covariance and the coefficient of correlation



Summary Definitions

- The **central tendency** is the extent to which all the data values group around a typical or central value.
- The **variation** is the amount of dispersion, or scattering, of values
- The **shape** is the pattern of the distribution of values from the lowest value to the highest value.

Measures of Central Tendency:

The Mean

- The arithmetic mean (often just called “mean”) is the most common measure of central tendency

Pronounced x-bar

- For a sample of size n:

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

The i^{th} value

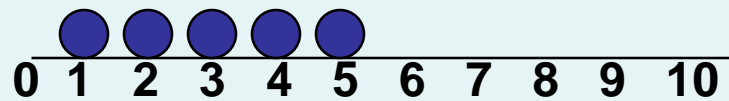
Sample size

Observed values

Measures of Central Tendency: The Mean

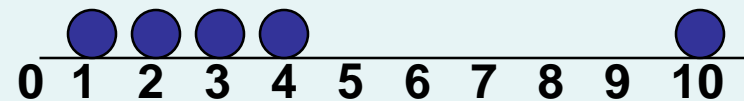
(continued)

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



Mean = 3

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

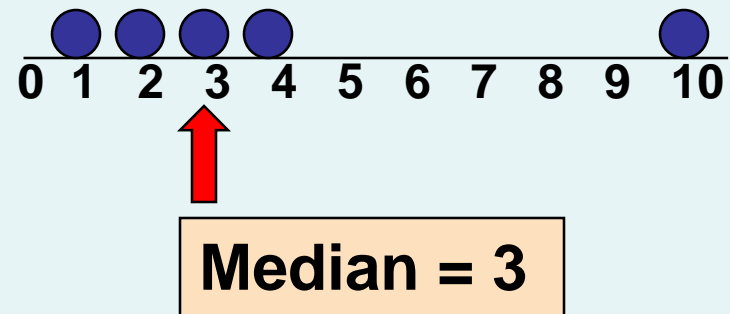
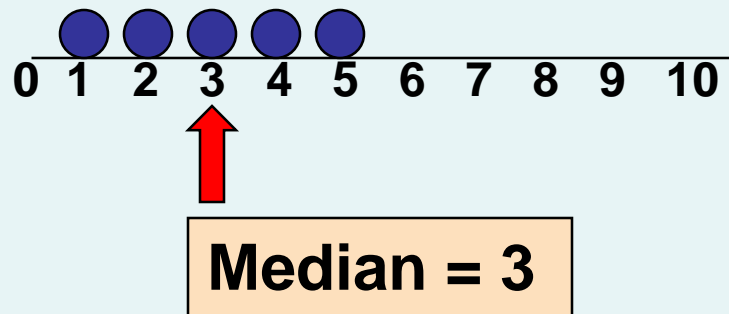


Mean = 4

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

Measures of Central Tendency: The Median

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



- Not affected by extreme values



Measures of Central Tendency: Locating the Median

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

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

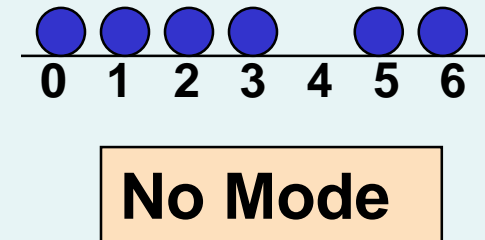
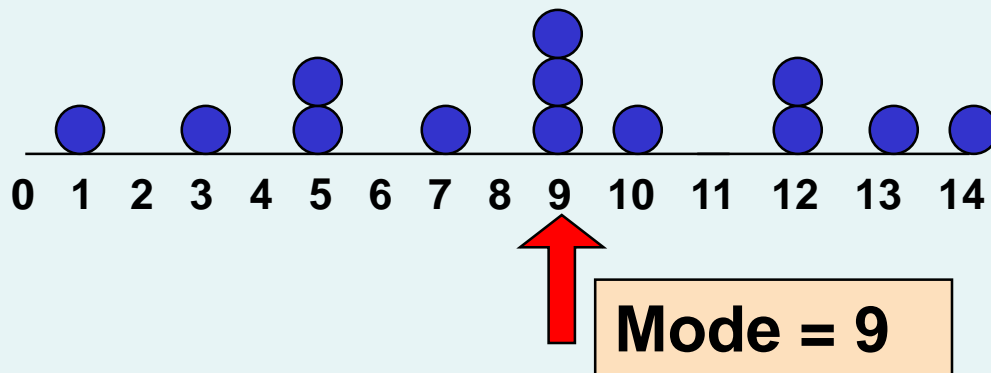
- If the number of values is odd, the median is the middle number
- If the number of values is even, the median is the average of the two middle numbers

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

Measures of Central Tendency:

The Mode

- 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



Measures of Central Tendency: Review Example

House Prices:

\$2,000,000

\$500,000

\$300,000

\$100,000

\$100,000

Sum \$3,000,000

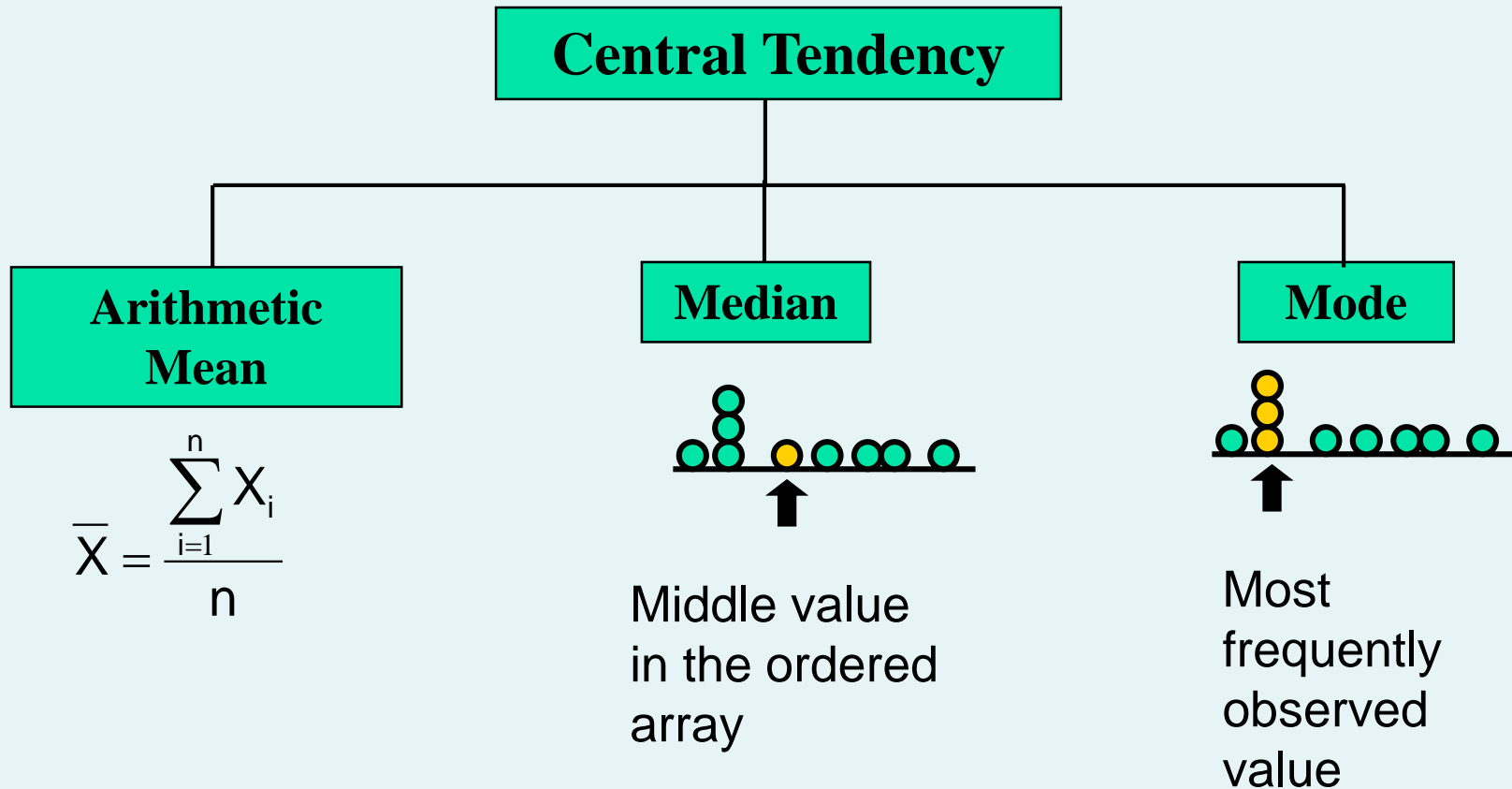
- **Mean:** $(\$3,000,000/5)$
= **\$600,000**
- **Median:** middle value of ranked data
= **\$300,000**
- **Mode:** most frequent value
= **\$100,000**

Measures of Central Tendency: Which Measure to Choose?

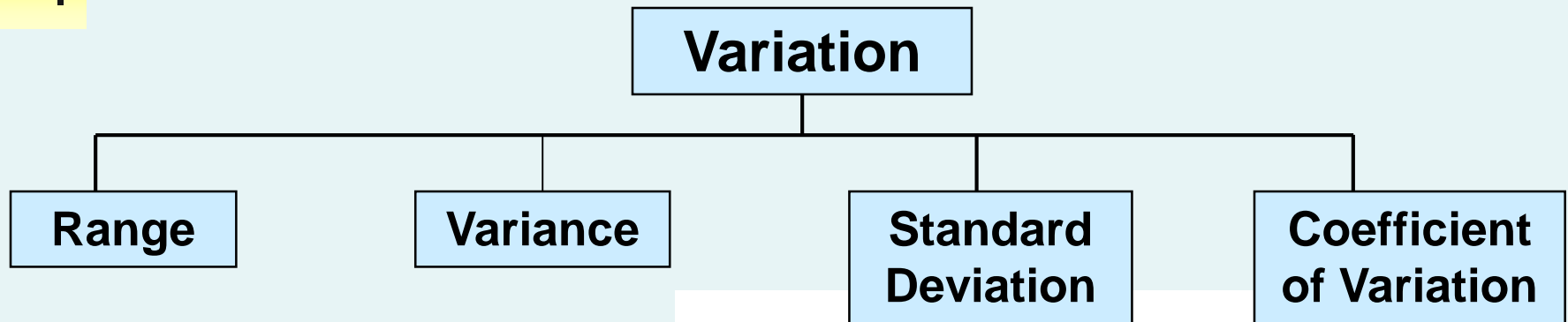


- The **mean** is generally used, unless extreme values (outliers) exist.
- The **median** is often used, since the median is not sensitive to extreme values. For example, median home prices may be reported for a region; it is less sensitive to outliers.
- In some situations it makes sense to report both the **mean** and the **median**.

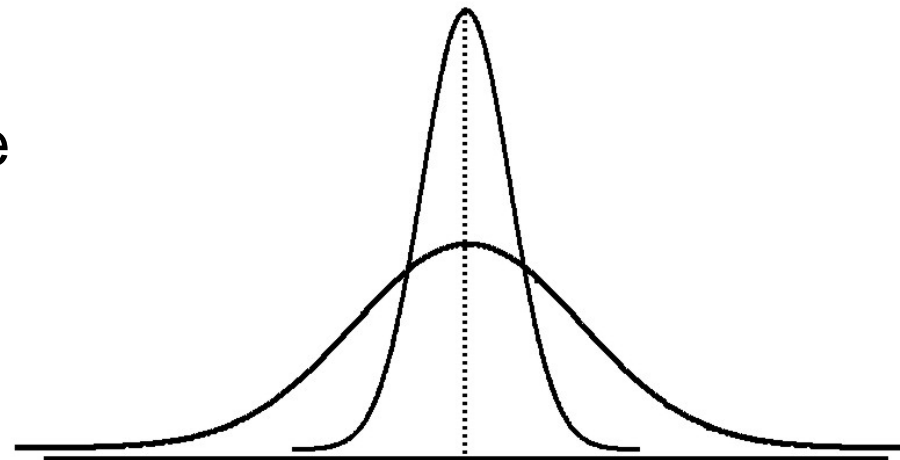
Measures of Central Tendency: Summary



Measures of Variation



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



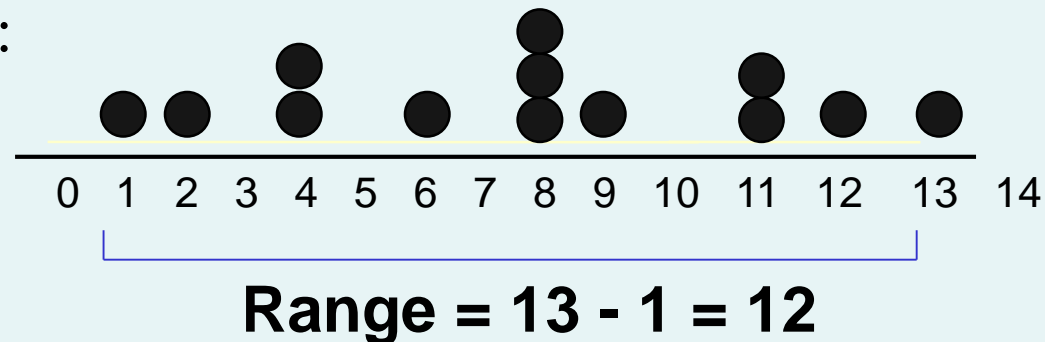
Same center,
different variation

Measures of Variation: The Range

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

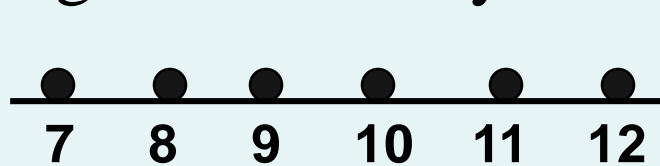
$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

Example:

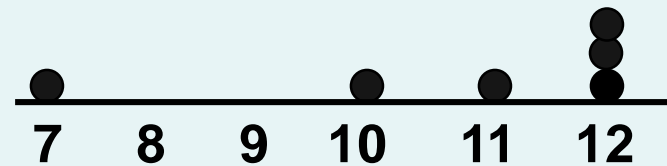


Measures of Variation: Why The Range Can Be Misleading

- Ignores the way in which data are distributed



$$\text{Range} = 12 - 7 = 5$$



$$\text{Range} = 12 - 7 = 5$$

- Sensitive to outliers

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

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

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

$$\text{Range} = 120 - 1 = 119$$

Measures of Variation: The 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

Measures of Variation: The Standard Deviation

- Most commonly used measure of variation
- Shows variation about the mean
- Is the square root of the variance
- Has the **same units as the original data**

- Sample standard deviation:

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



Measures of Variation: The Standard Deviation

Steps for Computing Standard Deviation

1. Compute the difference between each value and the mean.
2. Square each difference.
3. Add the squared differences.
4. Divide this total by $n-1$ to get the sample variance.
5. Take the square root of the sample variance to get the sample standard deviation.

Measures of Variation: Sample Standard Deviation: Calculation Example

Sample

Data (X_i) :

10 12 14 15 17 18 18 24

$n = 8$

Mean = $\bar{X} = 16$

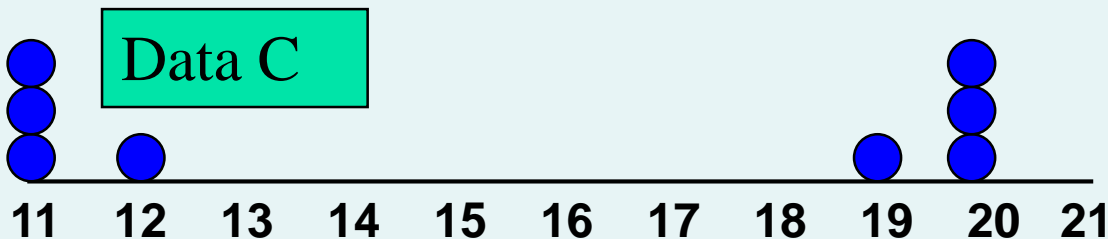
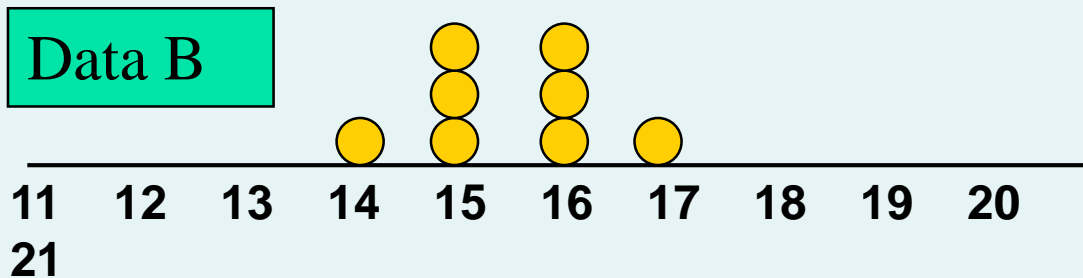
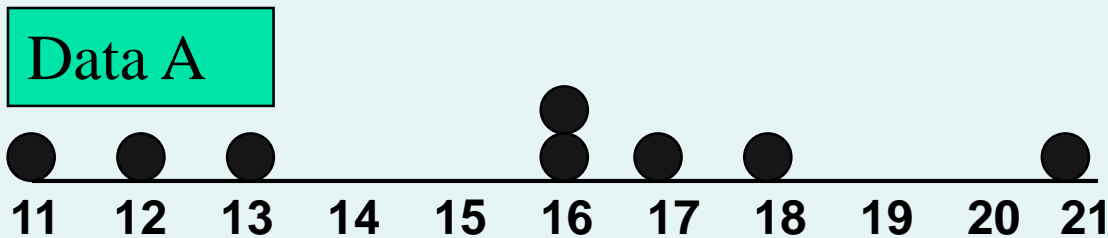
$$S = \sqrt{\frac{(10 - \bar{X})^2 + (12 - \bar{X})^2 + (14 - \bar{X})^2 + \dots + (24 - \bar{X})^2}{n - 1}}$$

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

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

A measure of the “average”
scatter around the mean

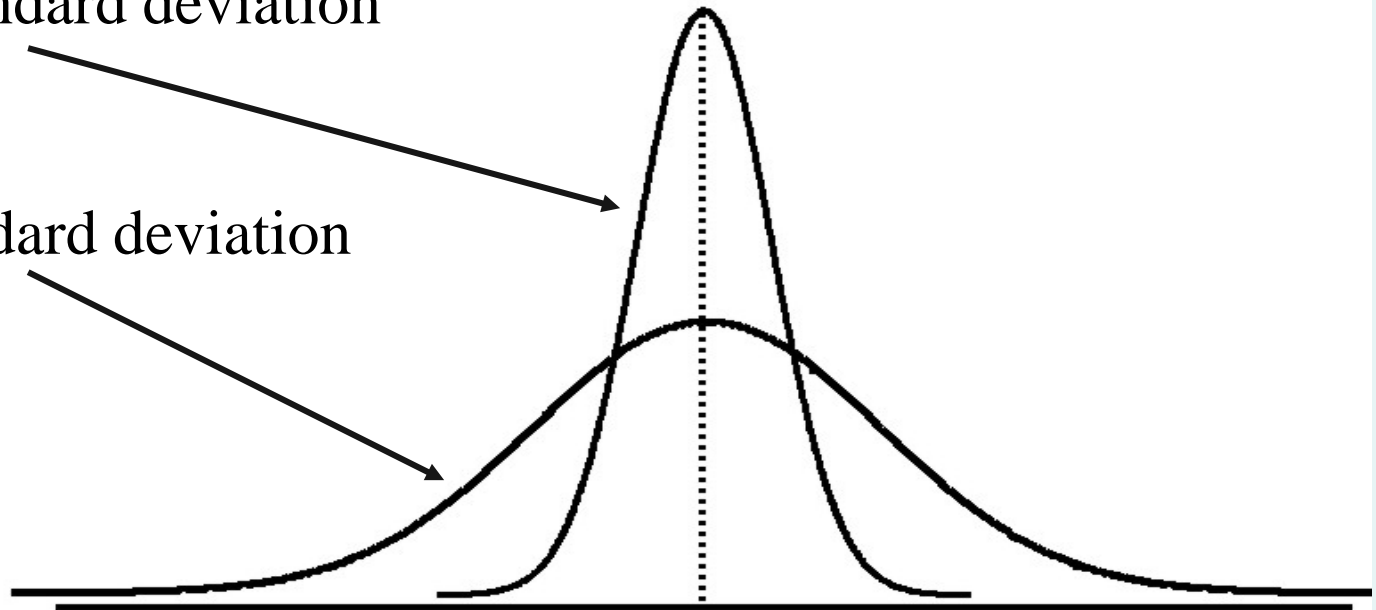
Measures of Variation: Comparing Standard Deviations



Measures of Variation: Comparing Standard Deviations

Smaller standard deviation

Larger standard deviation





Measures of Variation: Summary Characteristics

- The more the data are spread out, the greater the range, variance, and standard deviation.
- The more the data are concentrated, the smaller the range, variance, and standard deviation.
- If the values are all the same (no variation), all these measures will be zero.
- None of these measures are ever negative.



Measures of Variation: The Coefficient of Variation

- Measures **relative variation**
- Always in percentage (%)
- Shows **variation relative to mean**
- Can be used to compare the variability of two or more sets of data measured in different units

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

Measures of Variation: Comparing Coefficients of Variation

■ 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

Locating Extreme Outliers: Z-Score



- To compute the **Z-score** of a data value, subtract the mean and divide by the standard deviation.
- The Z-score is the number of standard deviations a data value is from the mean.
- A data value is considered an extreme outlier if its Z-score is less than -3.0 or greater than $+3.0$.
- The larger the absolute value of the Z-score, the farther the data value is from the mean.

Locating Extreme Outliers: Z-Score



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

where X represents the data value

\bar{X} is the sample mean

S is the sample standard deviation

Locating Extreme Outliers: Z-Score

- Suppose the mean math SAT score is 490, with a standard deviation of 100.
- Compute the Z-score for a test score of 620.

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

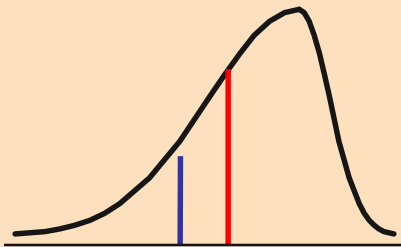
A score of 620 is 1.3 standard deviations above the mean and would not be considered an outlier.

Shape of a Distribution

- Describes how data are distributed
- Measures of shape
 - Symmetric or skewed

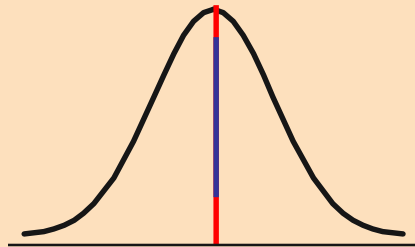
Left-Skewed

Mean < Median



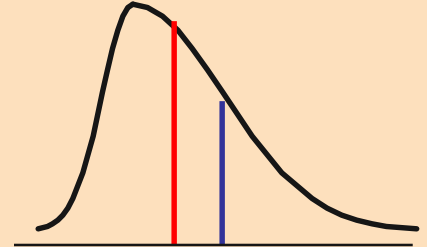
Symmetric

Mean = Median

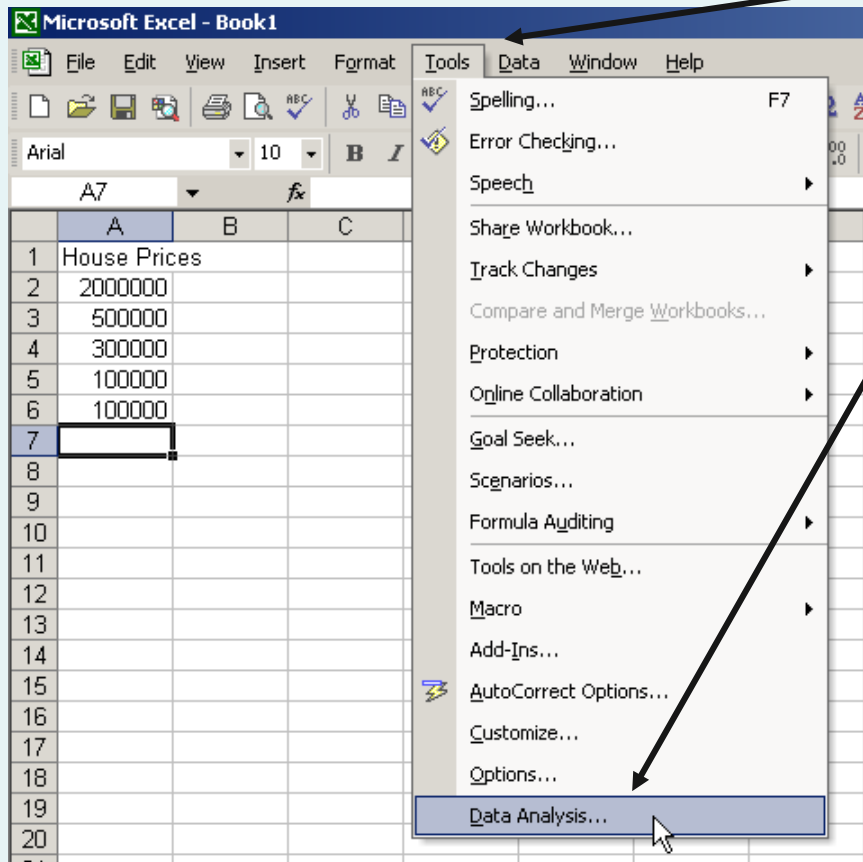


Right-Skewed

Median < Mean



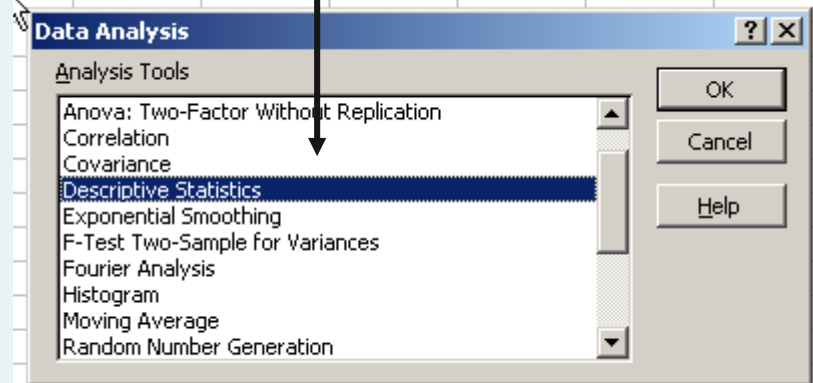
General Descriptive Stats Using Microsoft Excel



1. Select Tools.

2. Select Data Analysis.

3. Select Descriptive Statistics and click OK.



General Descriptive Stats Using Microsoft Excel

4. Enter the cell range.

5. Check the Summary Statistics box.

6. Click OK

The screenshot shows the 'Descriptive Statistics' dialog box in Microsoft Excel. The dialog box is open over a worksheet with data in column A. The 'Input Range' is set to '\$A\$1:\$A\$6'. The 'Summary statistics' checkbox is checked. The 'Confidence Level for Mean' is set to 95%. The 'Output options' section shows 'New Worksheet Ply' selected. The 'OK' button is highlighted.

| | A | B | C | D | E | F | G | H |
|----|--------------|---|---|---|---|---|---|---|
| 1 | House Prices | | | | | | | |
| 2 | 2000000 | | | | | | | |
| 3 | 500000 | | | | | | | |
| 4 | 300000 | | | | | | | |
| 5 | 100000 | | | | | | | |
| 6 | 100000 | | | | | | | |
| 7 | | | | | | | | |
| 8 | | | | | | | | |
| 9 | | | | | | | | |
| 10 | | | | | | | | |
| 11 | | | | | | | | |
| 12 | | | | | | | | |
| 13 | | | | | | | | |
| 14 | | | | | | | | |
| 15 | | | | | | | | |
| 16 | | | | | | | | |
| 17 | | | | | | | | |
| 18 | | | | | | | | |
| 19 | | | | | | | | |
| 20 | | | | | | | | |
| 21 | | | | | | | | |
| 22 | | | | | | | | |

Excel output

Microsoft Excel
descriptive statistics output,
using the house price data:

House Prices:

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

| | A | B |
|----|---------------------|-------------|
| 1 | <i>House Prices</i> | |
| 2 | | |
| 3 | Mean | 600000 |
| 4 | Standard Error | 357770.8764 |
| 5 | Median | 300000 |
| 6 | Mode | 100000 |
| 7 | Standard Deviation | 800000 |
| 8 | Sample Variance | 6.4E+11 |
| 9 | Kurtosis | 4.130126953 |
| 10 | Skewness | 2.006835938 |
| 11 | Range | 1900000 |
| 12 | Minimum | 100000 |
| 13 | Maximum | 2000000 |
| 14 | Sum | 3000000 |
| 15 | Count | 5 |
| 16 | | |
| 17 | | |



Minitab Output

Descriptive Statistics: House Price

| Total | | | | | | | | |
|-------------|-------|--------|---------|--------|-------------|---------|---------|--|
| Variable | Count | Mean | SE Mean | StDev | Variance | Sum | Minimum | |
| House Price | 5 | 600000 | 357771 | 800000 | 6.40000E+11 | 3000000 | 100000 | |

| N for | | | | | | |
|-------------|--------|---------|---------|--------|----------|----------|
| Variable | Median | Maximum | Range | Mode | Skewness | Kurtosis |
| House Price | 300000 | 2000000 | 1900000 | 100000 | 2.01 | 4.13 |



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

- The **population mean** is the sum of the values in the population divided by the population size, N

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

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

- Average of squared deviations of values from the mean

- Population variance:

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

Where μ = population mean

N = population size

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

Numerical Descriptive Measures For A Population: The Standard Deviation σ

- Most commonly used measure of variation
- Shows variation about the mean
- Is the square root of the population variance
- Has the **same units as the original data**

- Population standard deviation:

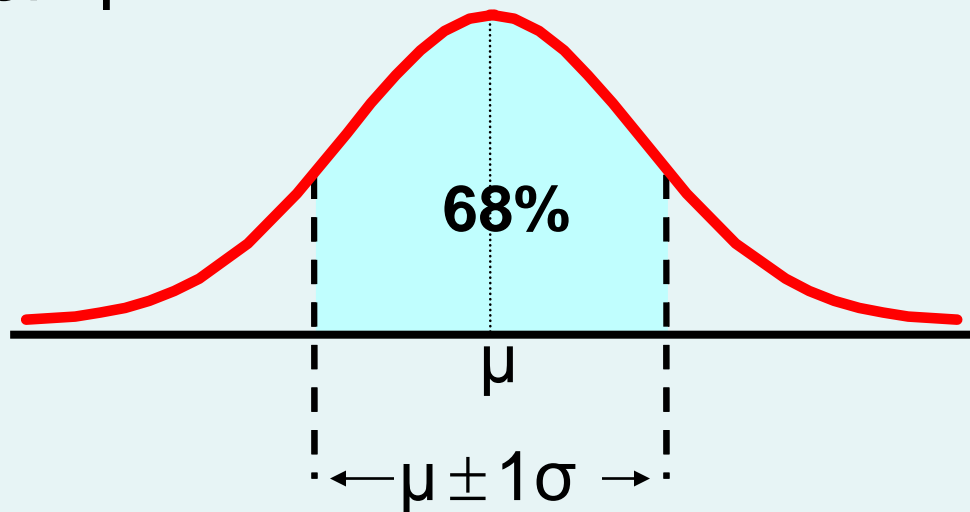
$$\sigma = \sqrt{\frac{\sum_{i=1}^N (X_i - \mu)^2}{N}}$$

Sample statistics versus population parameters

| Measure | Population Parameter | Sample Statistic |
|---------------------------|-----------------------------|-------------------------|
| Mean | μ | \bar{X} |
| Variance | σ^2 | S^2 |
| Standard Deviation | σ | S |

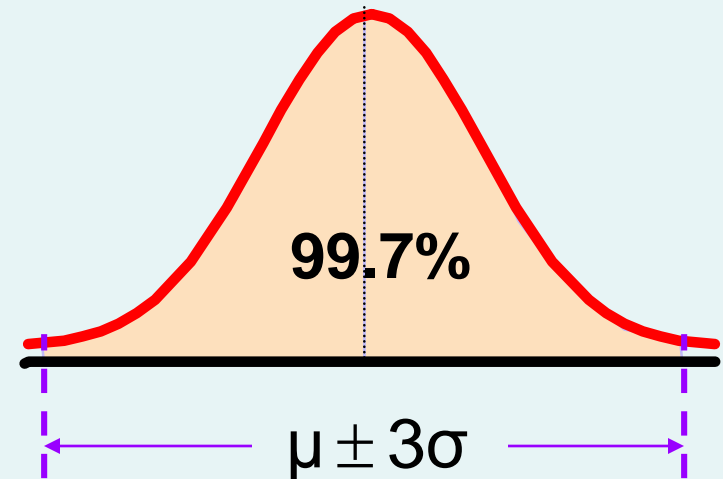
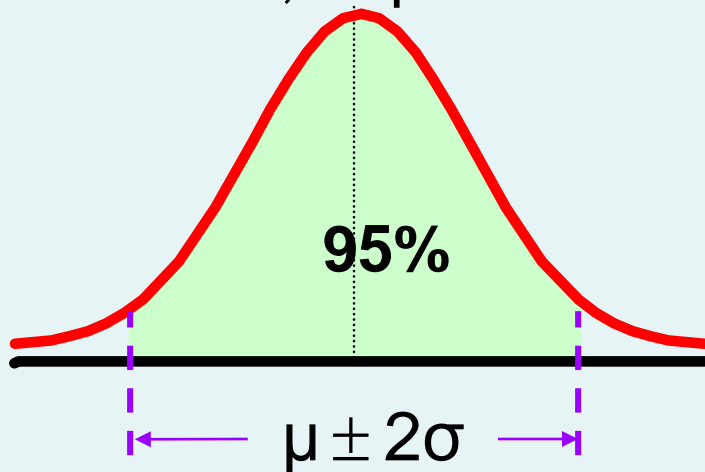
The Empirical Rule

- 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 or $\mu \pm 1\sigma$



The Empirical Rule

- 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).



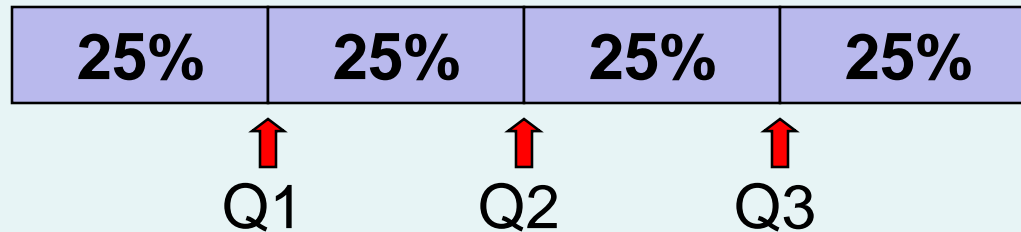
Chebyshev Rule

- Regardless of how the data are distributed, at least $(1 - 1/k^2) \times 100\%$ of the values will fall within k standard deviations of the mean (for $k > 1$)
 - Examples:

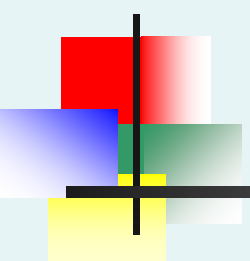
| At least | within |
|-----------------------------------|-------------------------------|
| $(1 - 1/2^2) \times 100\% = 75\%$ | $k=2 \quad (\mu \pm 2\sigma)$ |
| $(1 - 1/3^2) \times 100\% = 89\%$ | $k=3 \quad (\mu \pm 3\sigma)$ |

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
- Q_2 is the same as the median (50% of the observations are smaller and 50% are larger)
- Only 25% of the observations are greater than the third quartile



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

- 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.

Quartile Measures: Locating Quartiles

Sample 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 use the value half way between the 2nd and 3rd values,

so $Q_1 = 12.5$

Q_1 and Q_3 are measures of non-central location
 $Q_2 =$ median, is a measure of central tendency

Quartile Measures

Calculating The Quartiles: Example

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

Q_1 and Q_3 are measures of non-central location
 $Q_2 = \text{median}$, is a measure of central tendency

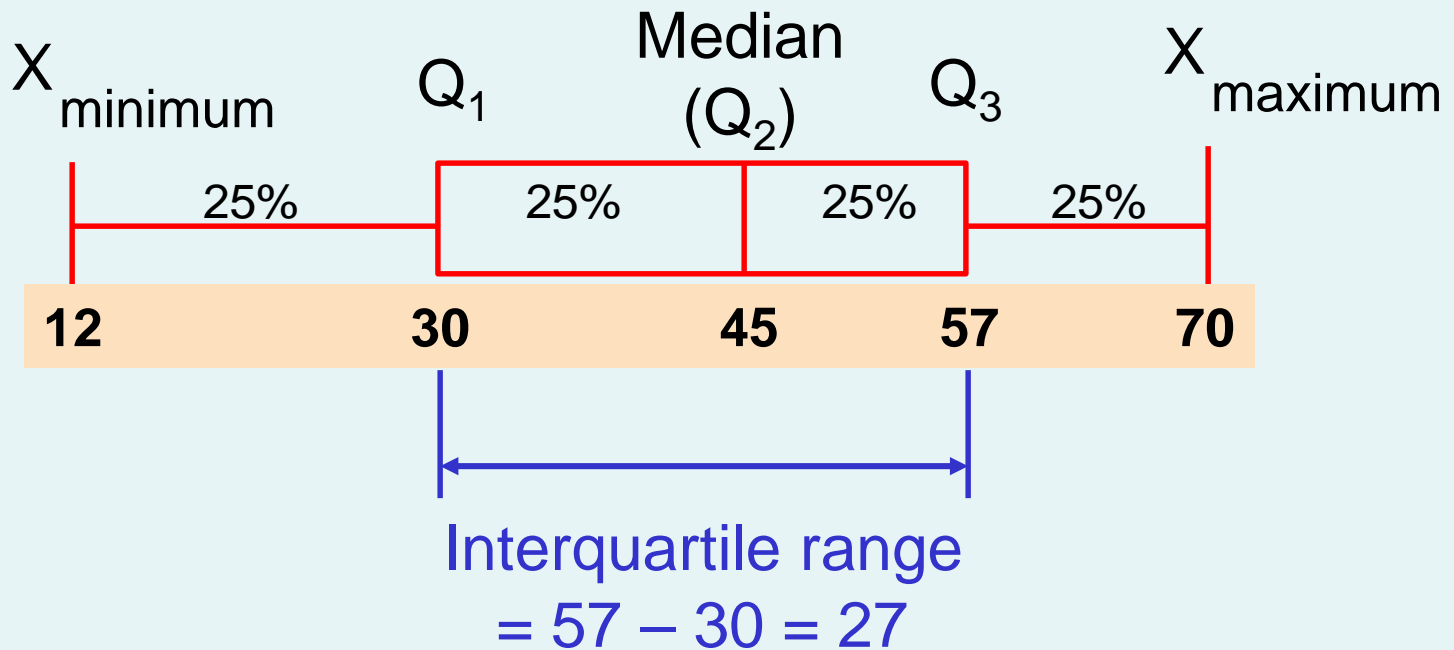


Quartile Measures: The Interquartile Range (IQR)

- The IQR is $Q_3 - Q_1$ and measures the spread in the middle 50% of the data
- The IQR is also called the midspread 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
- Measures like Q_1 , Q_3 , and IQR that are not influenced by outliers are called resistant measures

Calculating The Interquartile Range

Example:





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 and distribution shape

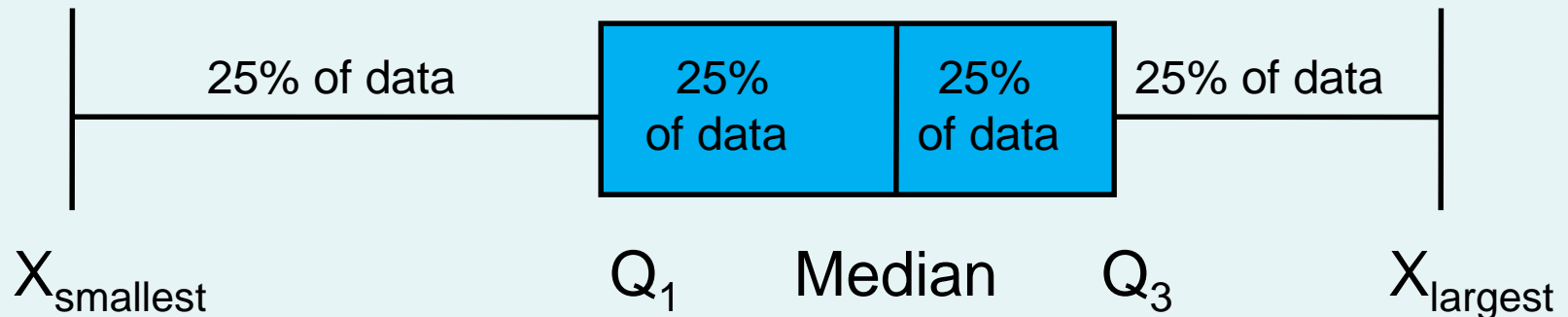
| Left-Skewed | Symmetric | Right-Skewed |
|--------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|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

Five Number Summary and The Boxplot

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

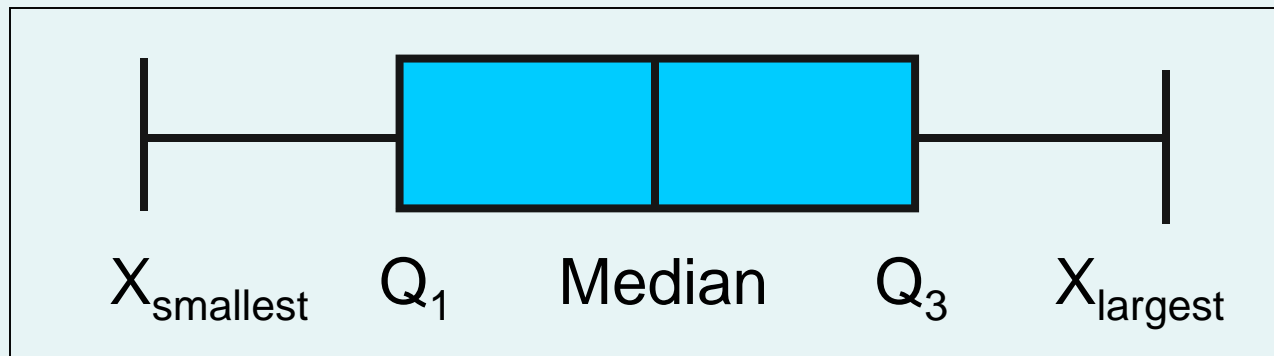
X_{smallest} -- Q_1 -- Median -- Q_3 -- X_{largest}

Example:



Five Number Summary: Shape of Boxplots

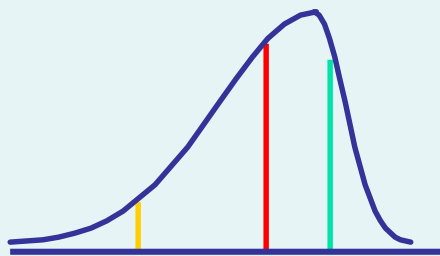
- If data are symmetric around the median then the box and central line are centered between the endpoints



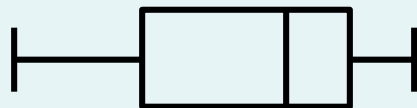
- A Boxplot can be shown in either a vertical or horizontal orientation

Distribution Shape and The Boxplot

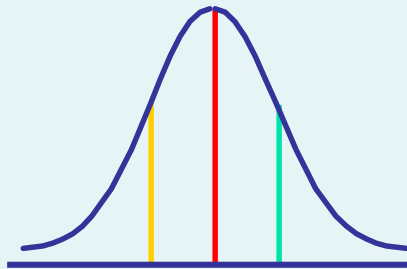
Left-Skewed



Q_1 Q_2 Q_3



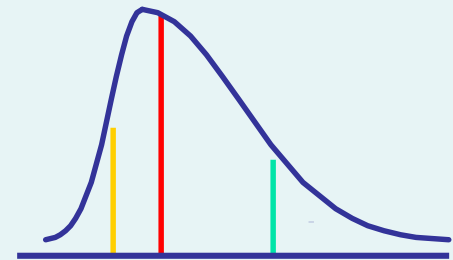
Symmetric



Q_1 Q_2 Q_3



Right-Skewed

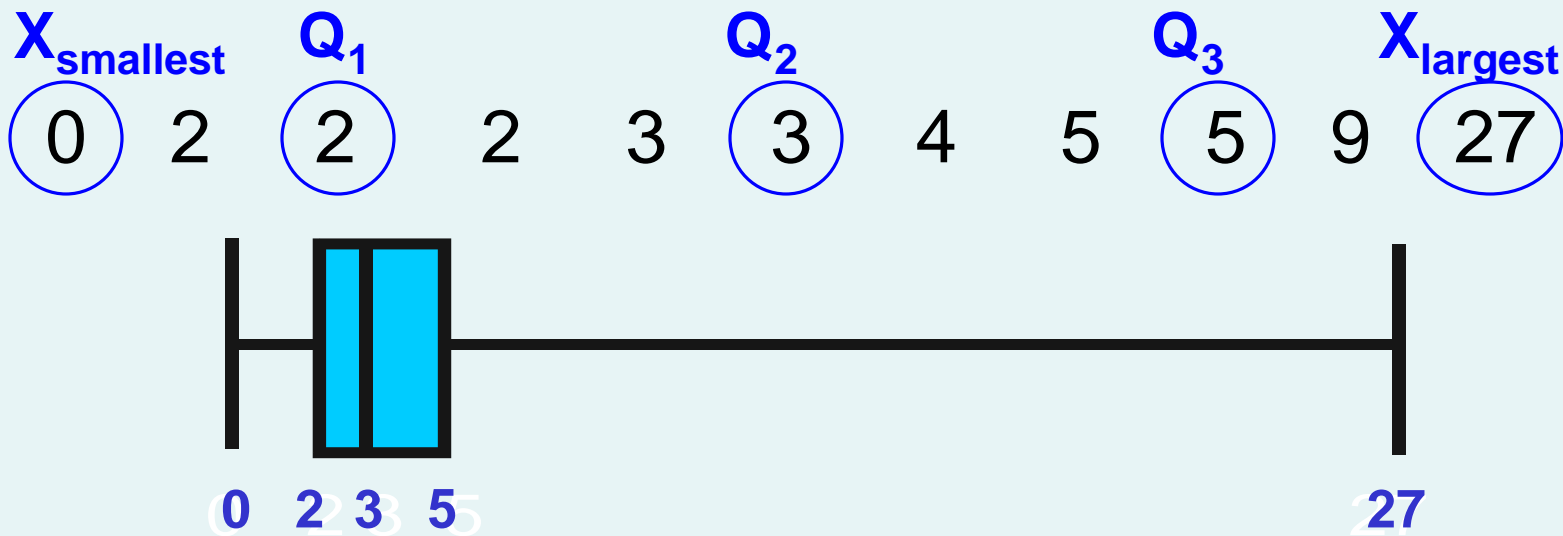


Q_1 Q_2 Q_3



Boxplot Example

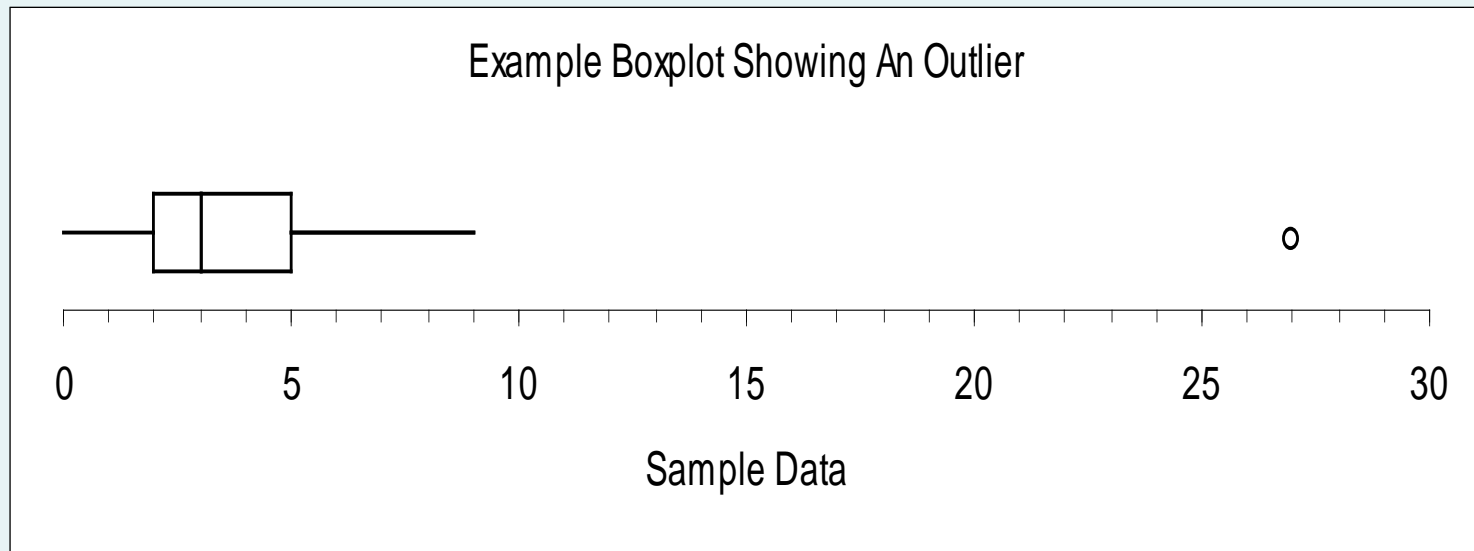
- Below is a Boxplot for the following data:



- The data are right skewed, as the plot depicts

Boxplot example showing an outlier

- The boxplot below of the same data shows the outlier value of 27 plotted separately
- A value is considered an outlier if it is more than 1.5 times the interquartile range below Q_1 or above Q_3





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

- **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

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

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

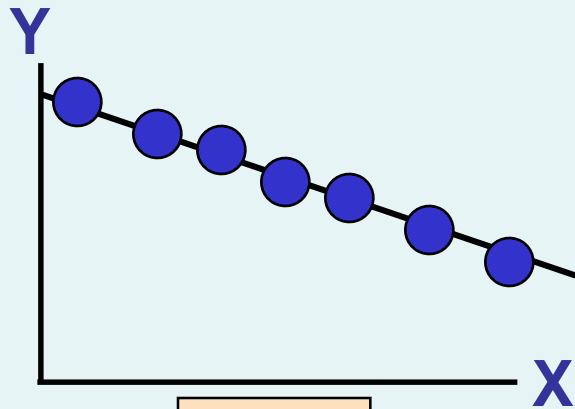
$$S_Y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}}$$



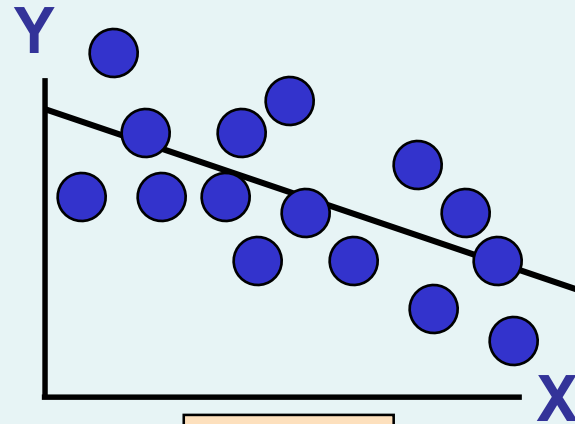
Features of the Coefficient of Correlation

- 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
 - Ranges 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

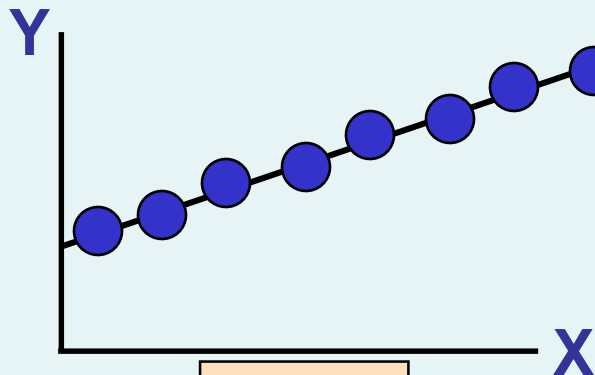
Scatter Plots of Sample Data with Various Coefficients of Correlation



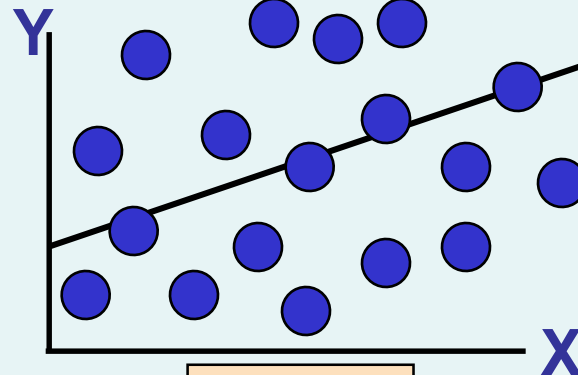
$$r = -1$$



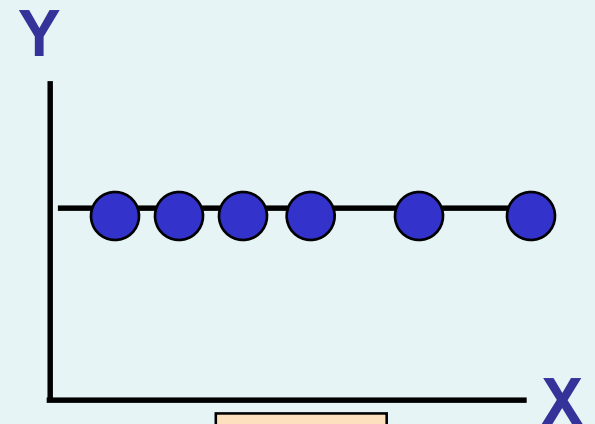
$$r = -.6$$



$$r = +1$$

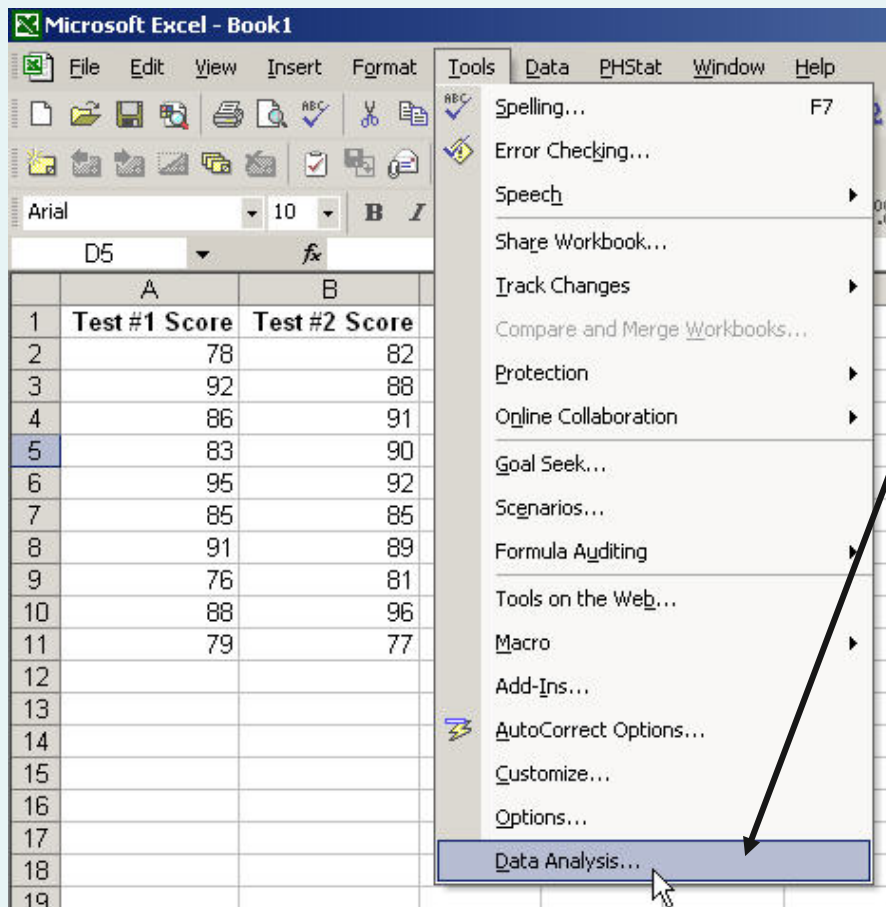


$$r = +.3$$

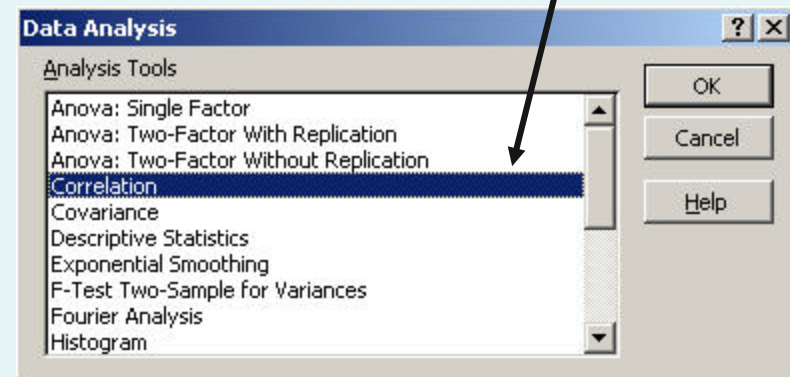


$$r = 0$$

The Coefficient of Correlation Using Microsoft Excel



1. Select Tools/Data Analysis
2. Choose Correlation from the selection menu
3. Click OK . . .



The Coefficient of Correlation Using Microsoft Excel

The screenshot shows an Excel spreadsheet with two columns of test scores. A 'Correlation' dialog box is open, with the 'Input Range' set to '\$A\$1:\$B\$11'. The 'Labels in First Row' checkbox is checked. The 'Grouped By' options are 'Columns' (selected) and 'Rows'. The 'Output options' section shows 'New Worksheet Ply' selected. An arrow points from the dialog box to the data table.

| | A | B | C | D | E | F | G | H | I |
|----|---------------|---------------|---|---|---|---|---|---|---|
| 1 | Test #1 Score | Test #2 Score | | | | | | | |
| 2 | 78 | 82 | | | | | | | |
| 3 | 92 | 88 | | | | | | | |
| 4 | 86 | 91 | | | | | | | |
| 5 | 83 | 90 | | | | | | | |
| 6 | 95 | 92 | | | | | | | |
| 7 | 85 | 85 | | | | | | | |
| 8 | 91 | 89 | | | | | | | |
| 9 | 76 | 81 | | | | | | | |
| 10 | 88 | 96 | | | | | | | |
| 11 | 79 | 77 | | | | | | | |

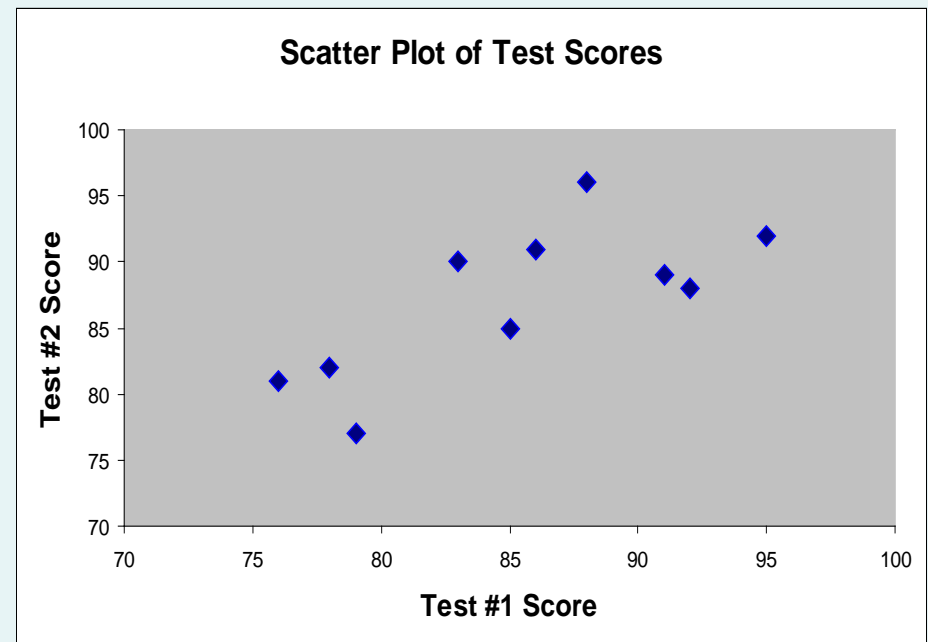
4. Input data range and select appropriate options
5. Click OK to get output

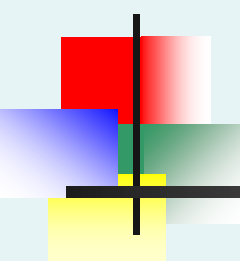
The output table shows the correlation coefficient between the two test scores. The value is 0.733243705.

| | A | B | C |
|---|---------------|---------------|---------------|
| 1 | | Test #1 Score | Test #2 Score |
| 2 | Test #1 Score | 1 | |
| 3 | Test #2 Score | 0.733243705 | 1 |
| 4 | | | |

Interpreting the Coefficient of Correlation Using Microsoft Excel

- $r = .733$
- There is a relatively strong positive linear relationship between test score #1 and test score #2.
- Students who scored high on the first test tended to score high on second test.





Pitfalls in Numerical Descriptive Measures

- Data analysis is objective
 - Should report the summary measures that best describe and communicate the important aspects of the data set
- Data interpretation is subjective
 - Should be done in fair, neutral and clear manner

Ethical Considerations

Numerical descriptive measures:

- Should document both good and bad results
- Should be presented in a fair, objective and neutral manner
- Should not use inappropriate summary measures to distort facts





Chapter Summary

- Described measures of central tendency
 - Mean, median, mode
- Described measures of variation
 - Range, interquartile range, variance and standard deviation, coefficient of variation, Z-scores
- Illustrated shape of distribution
 - Symmetric, skewed
- Described data using the 5-number summary
 - Boxplots

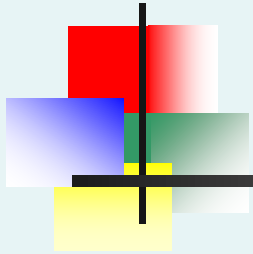


Chapter Summary

(continued)

- Discussed covariance and correlation coefficient
- Addressed pitfalls in numerical descriptive measures and ethical considerations

Business Statistics: A First Course 5th Edition



Chapter 4

Basic Probability



Learning Objectives

In this chapter, you learn:

- Basic probability concepts
- Conditional probability
- To use Bayes' Theorem to revise probabilities



Basic Probability Concepts

- **Probability** – the chance that an uncertain event will occur (always between 0 and 1)
- **Impossible Event** – an event that has no chance of occurring (probability = 0)
- **Certain Event** – an event that is sure to occur (probability = 1)



Assessing Probability

There are three approaches to assessing the probability of an uncertain event:

1. *a priori* -- based on prior knowledge of the process

$$\text{probability of occurrence} = \frac{X}{T} = \frac{\text{number of ways the event can occur}}{\text{total number of elementary outcomes}}$$

Assuming
all
outcomes
are equally
likely

2. empirical probability

$$\text{probability of occurrence} = \frac{\text{number of ways the event can occur}}{\text{total number of elementary outcomes}}$$

3. subjective probability

based on a combination of an individual's past experience, personal opinion, and analysis of a particular situation



Example of a *priori* probability

Find the probability of selecting a face card (Jack, Queen, or King) from a standard deck of 52 cards.

$$\text{Probability of Face Card} = \frac{X}{T} = \frac{\text{number of face cards}}{\text{total number of cards}}$$

$$\frac{X}{T} = \frac{12 \text{ face cards}}{52 \text{ total cards}} = \frac{3}{13}$$



Example of empirical probability

Find the probability of selecting a male taking statistics from the population described in the following table:

| | Taking Stats | Not Taking Stats | Total |
|--------|--------------|------------------|-------|
| Male | 84 | 145 | 229 |
| Female | 76 | 134 | 210 |
| Total | 160 | 279 | 439 |

$$\text{Probability of male taking stats} = \frac{\text{number of males taking stats}}{\text{total number of people}} = \frac{84}{439} = 0.191$$



Events

Each possible outcome of a variable is an **event**.

- **Simple event**

- An event described by a single characteristic
- e.g., A red card from a deck of cards

- **Joint event**

- An event described by two or more characteristics
- e.g., An ace that is also red from a deck of cards

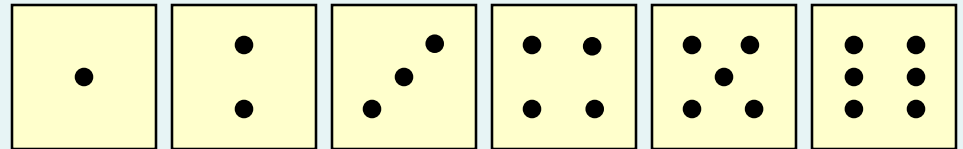
- **Complement of an event A (denoted A')**

- All events that are not part of event A
- e.g., All cards that are not diamonds

Sample Space

The **Sample Space** is the collection of all possible events

e.g. All 6 faces of a die:



e.g. All 52 cards of a bridge deck:



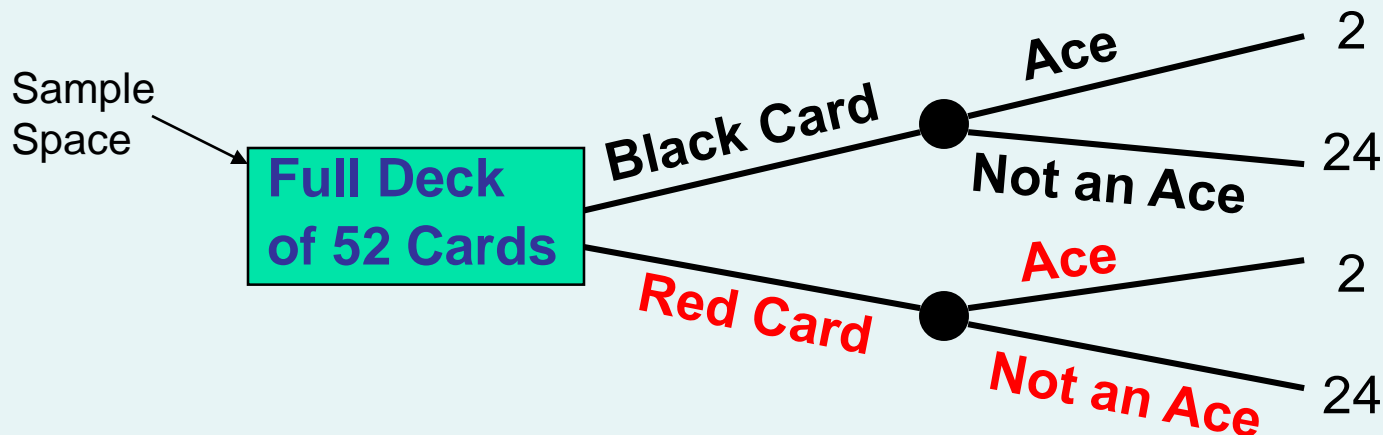
Visualizing Events

■ Contingency Tables

| | Ace | Not Ace | Total |
|-------|-----|---------|-------|
| Black | 2 | 24 | 26 |
| Red | 2 | 24 | 26 |
| Total | 4 | 48 | 52 |

Sample Space

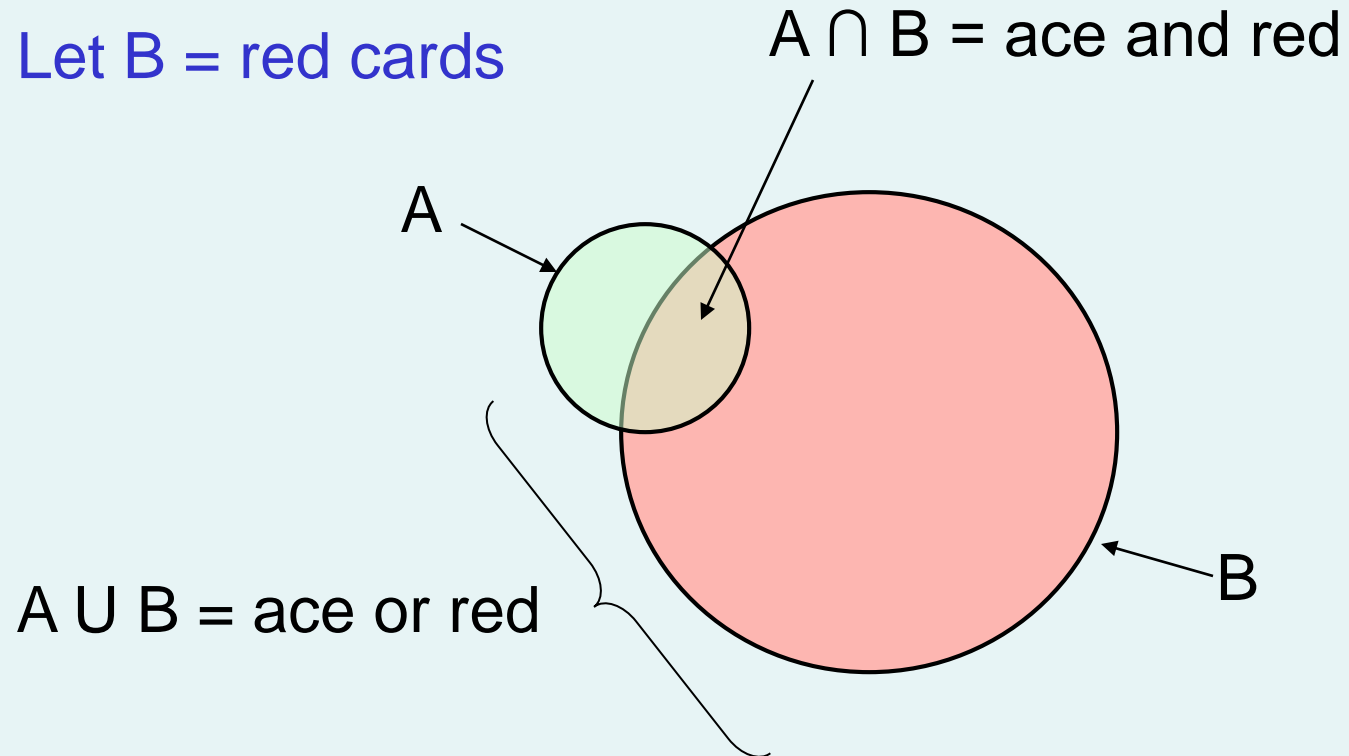
■ Decision Trees



Visualizing Events

- Venn Diagrams

- Let $A = \text{aces}$
- Let $B = \text{red cards}$



Definitions

Simple vs. Joint Probability

- Simple Probability refers to the probability of a simple event.
 - ex. $P(\text{King})$
 - ex. $P(\text{Spade})$
- Joint Probability refers to the probability of an occurrence of two or more events (joint event).
 - ex. $P(\text{King and Spade})$



Mutually Exclusive Events

- Mutually exclusive events
 - Events that cannot occur simultaneously

Example: Drawing one card from a deck of cards

A = queen of diamonds; B = queen of clubs

- Events A and B are mutually exclusive



Collectively Exhaustive Events

- **Collectively exhaustive** events
 - One of the events must occur
 - The set of events covers the entire sample space

example:

A = aces; B = black cards;
C = diamonds; D = hearts

- Events A, B, C and D are collectively exhaustive (but not mutually exclusive – an ace may also be a heart)
- Events B, C and D are collectively exhaustive and also mutually exclusive



Computing Joint and Marginal Probabilities

- The probability of a joint event, A and B:

$$P(A \text{ and } B) = \frac{\text{number of outcomes satisfying A and B}}{\text{total number of elementary outcomes}}$$

- Computing a marginal (or simple) probability:

$$P(A) = P(A \text{ and } B_1) + P(A \text{ and } B_2) + \cdots + P(A \text{ and } B_k)$$

- Where B_1, B_2, \dots, B_k are k mutually exclusive and collectively exhaustive events

Joint Probability Example

P(Red and Ace)

$$= \frac{\text{number of cards that are red and ace}}{\text{total number of cards}} = \frac{2}{52}$$

| Type | Color | | Total |
|---------|-------|-------|-------|
| | Red | Black | |
| Ace | 2 | 2 | 4 |
| Non-Ace | 24 | 24 | 48 |
| Total | 26 | 26 | 52 |




Marginal Probability Example

P(Ace)

$$= P(\text{Ace and Red}) + P(\text{Ace and Black}) = \frac{2}{52} + \frac{2}{52} = \frac{4}{52}$$

| Type | Color | | Total |
|---------|-------|-------|-------|
| | Red | Black | |
| Ace | 2 | 2 | 4 |
| Non-Ace | 24 | 24 | 48 |
| Total | 26 | 26 | 52 |



Marginal & Joint Probabilities In A Contingency Table

| Event | Event | | Total |
|----------------|---------------------------------------|---------------------------------------|--------------------|
| | B ₁ | B ₂ | |
| A ₁ | P(A ₁ and B ₁) | P(A ₁ and B ₂) | P(A ₁) |
| A ₂ | P(A ₂ and B ₁) | P(A ₂ and B ₂) | P(A ₂) |
| Total | P(B ₁) | P(B ₂) | 1 |

Joint Probabilities

Marginal (Simple) Probabilities

Probability Summary So Far

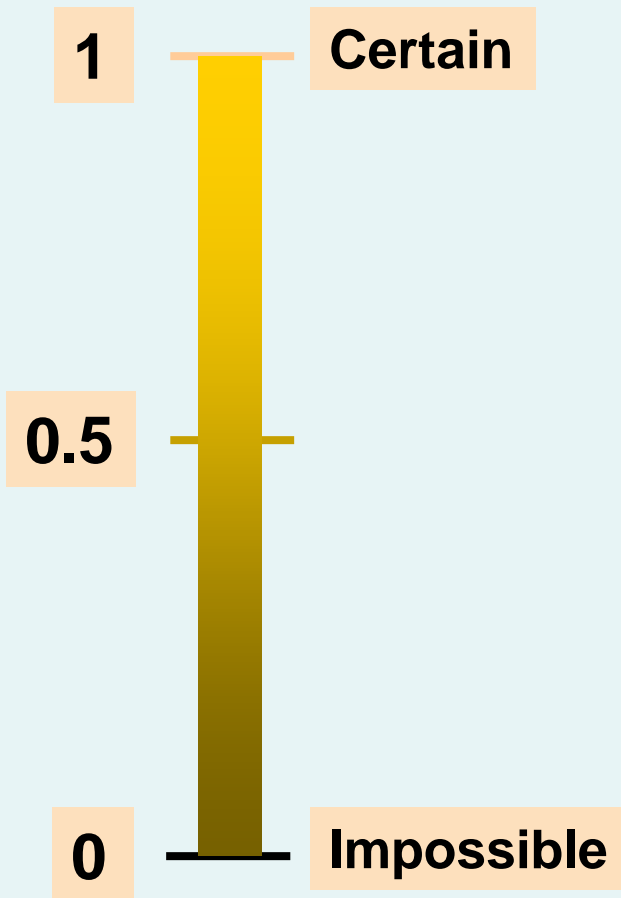
- Probability is the numerical measure of the likelihood that an event will occur
- The probability of any event must be between 0 and 1, inclusively

$$0 \leq P(A) \leq 1 \quad \text{For any event } A$$

- The sum of the probabilities of all mutually exclusive and collectively exhaustive events is 1

$$P(A) + P(B) + P(C) = 1$$

If A, B, and C are mutually exclusive and collectively exhaustive





General Addition Rule

General Addition Rule:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

If A and B are mutually exclusive, then

$P(A \text{ and } B) = 0$, so the rule can be simplified:

$$P(A \text{ or } B) = P(A) + P(B)$$

For mutually exclusive events A and B

General Addition Rule Example

$$P(\text{Red or Ace}) = P(\text{Red}) + P(\text{Ace}) - P(\text{Red and Ace})$$

$$= 26/52 + 4/52 - 2/52 = 28/52$$

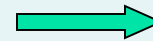
| Type | Color | | Total |
|---------|-------|-------|-------|
| | Red | Black | |
| Ace | 2 | 2 | 4 |
| Non-Ace | 24 | 24 | 48 |
| Total | 26 | 26 | 52 |

Don't count
the two red
aces twice!

Computing Conditional Probabilities

- A **conditional probability** is the probability of one event, given that another event has occurred:

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)}$$



The conditional probability of A given that B has occurred

$$P(B | A) = \frac{P(A \text{ and } B)}{P(A)}$$



The conditional probability of B given that A has occurred

Where $P(A \text{ and } B)$ = joint probability of A and B

$P(A)$ = marginal or simple probability of A

$P(B)$ = marginal or simple probability of B



Conditional Probability Example

- Of the cars on a used car lot, 70% have air conditioning (AC) and 40% have a CD player (CD). 20% of the cars have both.
- What is the probability that a car has a CD player, given that it has AC ?

i.e., we want to find $P(\text{CD} \mid \text{AC})$

Conditional Probability Example

(continued)

- Of the cars on a used car lot, **70%** have air conditioning (AC) and **40%** have a CD player (CD). **20%** of the cars have both.

| | CD | No CD | Total |
|-------|-----|-------|-------|
| AC | 0.2 | 0.5 | 0.7 |
| No AC | 0.2 | 0.1 | 0.3 |
| Total | 0.4 | 0.6 | 1.0 |

$$P(\text{CD} | \text{AC}) = \frac{P(\text{CD and AC})}{P(\text{AC})} = \frac{0.2}{0.7} = 0.2857$$

Conditional Probability Example

(continued)

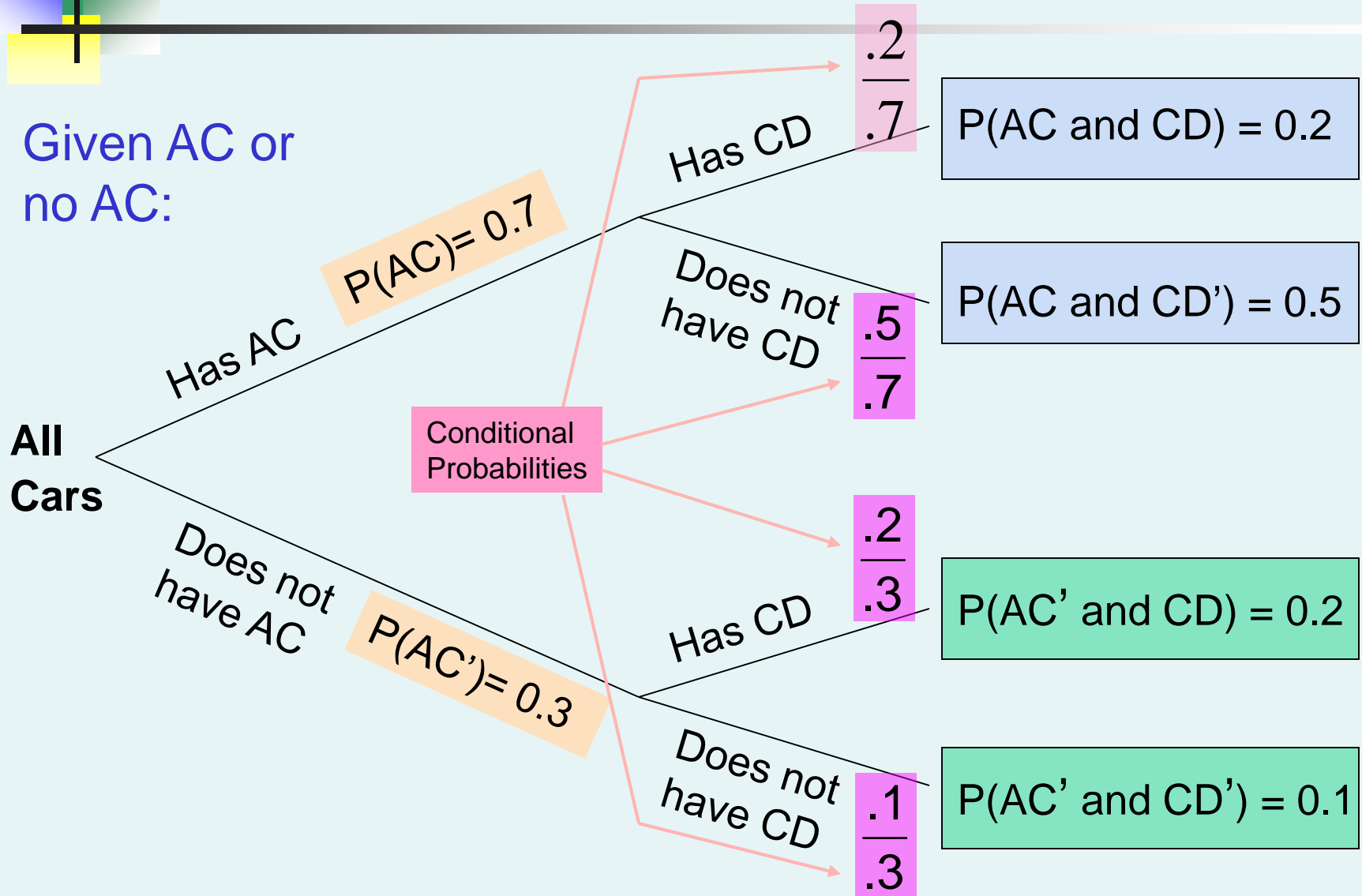
- Given AC, we only consider the top row (70% of the cars). Of these, 20% have a CD player. 20% of 70% is about 28.57%.

| | CD | No CD | Total |
|-------|-----|-------|-------|
| AC | 0.2 | 0.5 | 0.7 |
| No AC | 0.2 | 0.1 | 0.3 |
| Total | 0.4 | 0.6 | 1.0 |

$$P(\text{CD} | \text{AC}) = \frac{P(\text{CD and AC})}{P(\text{AC})} = \frac{0.2}{0.7} = 0.2857$$

Using Decision Trees

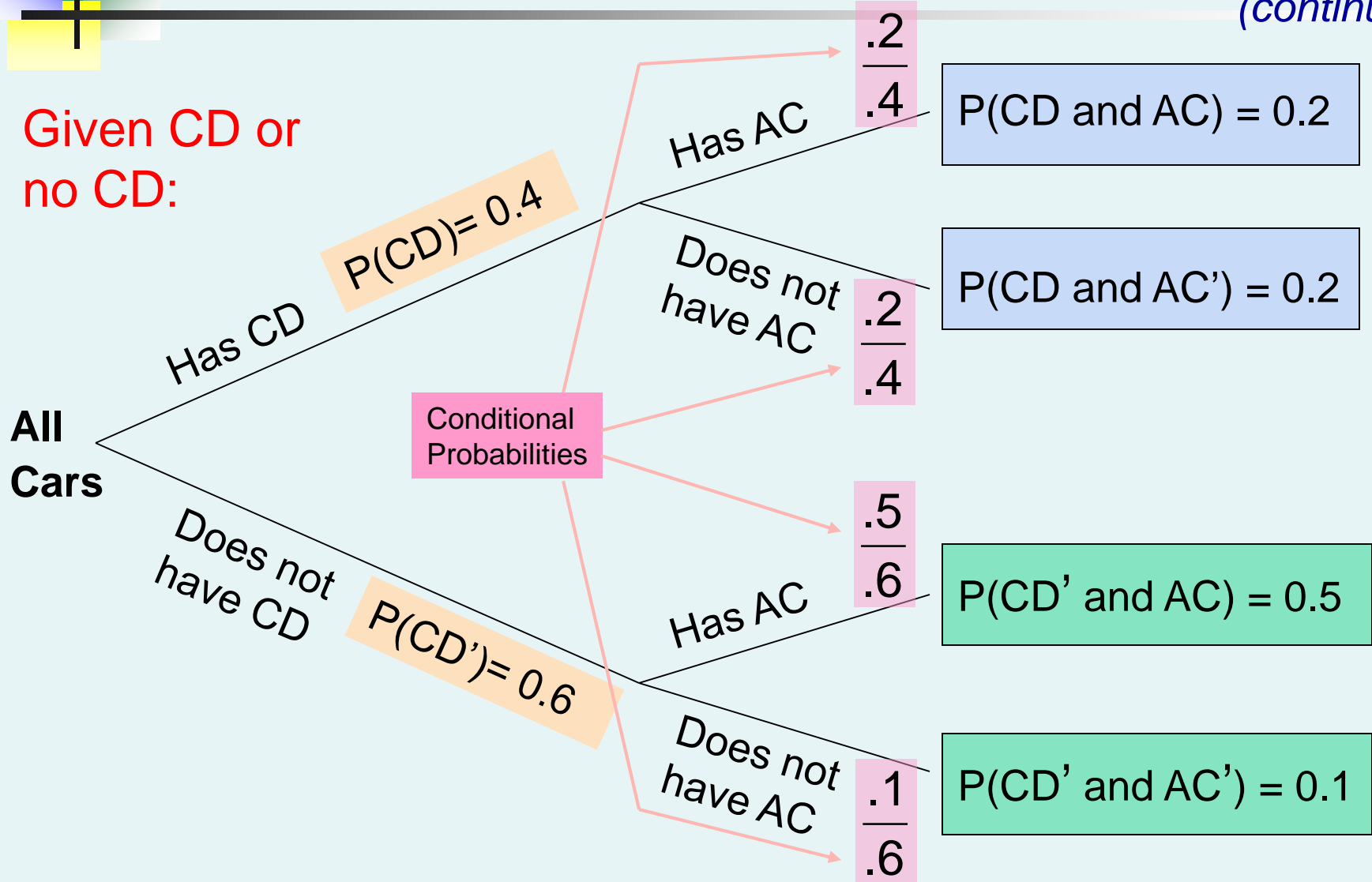
Given AC or
no AC:



Using Decision Trees

(continued)

Given CD or
no CD:





Independence

- Two events are **independent** if and only if:

$$P(A | B) = P(A)$$

- Events A and B are independent when the probability of one event is not affected by the fact that the other event has occurred



Multiplication Rules

- Multiplication rule for two events A and B:

$$P(A \text{ and } B) = P(A | B)P(B)$$

Note: If A and B are independent, then $P(A | B) = P(A)$ and the multiplication rule simplifies to

$$P(A \text{ and } B) = P(A)P(B)$$



Marginal Probability

- Marginal probability for event A:

$$P(A) = P(A | B_1)P(B_1) + P(A | B_2)P(B_2) + \cdots + P(A | B_k)P(B_k)$$

- Where B_1, B_2, \dots, B_k are k mutually exclusive and collectively exhaustive events



Bayes' Theorem

- Bayes' Theorem is used to revise previously calculated probabilities based on new information.
- Developed by Thomas Bayes in the 18th Century.
- It is an extension of conditional probability.



Bayes' Theorem

$$P(B_i | A) = \frac{P(A | B_i)P(B_i)}{P(A | B_1)P(B_1) + P(A | B_2)P(B_2) + \cdots + P(A | B_k)P(B_k)}$$

■ where:

B_i = i^{th} event of k mutually exclusive and collectively exhaustive events

A = new event that might impact $P(B_i)$

Bayes' Theorem Example

- A drilling company has estimated a 40% chance of striking oil for their new well.
- A detailed test has been scheduled for more information. Historically, 60% of successful wells have had detailed tests, and 20% of unsuccessful wells have had detailed tests.
- Given that this well has been scheduled for a detailed test, what is the probability that the well will be successful?



Bayes' Theorem Example

(continued)

- Let S = successful well
 U = unsuccessful well
- $P(S) = 0.4$, $P(U) = 0.6$ (prior probabilities)
- Define the detailed test event as D
- Conditional probabilities:
 $P(D|S) = 0.6$ $P(D|U) = 0.2$
- Goal is to find $P(S|D)$



Bayes' Theorem Example

(continued)

Apply Bayes' Theorem:

$$\begin{aligned} P(S | D) &= \frac{P(D | S)P(S)}{P(D | S)P(S) + P(D | U)P(U)} \\ &= \frac{(0.6)(0.4)}{(0.6)(0.4) + (0.2)(0.6)} \\ &= \frac{0.24}{0.24 + 0.12} = 0.667 \end{aligned}$$



So the revised probability of success, given that this well has been scheduled for a detailed test, is 0.667

Bayes' Theorem Example

(continued)

- Given the detailed test, the revised probability of a successful well has risen to 0.667 from the original estimate of 0.4



| Event | Prior Prob. | Conditional Prob. | Joint Prob. | Revised Prob. |
|------------------|-------------|-------------------|---------------------|---------------------|
| S (successful) | 0.4 | 0.6 | $(0.4)(0.6) = 0.24$ | $0.24/0.36 = 0.667$ |
| U (unsuccessful) | 0.6 | 0.2 | $(0.6)(0.2) = 0.12$ | $0.12/0.36 = 0.333$ |

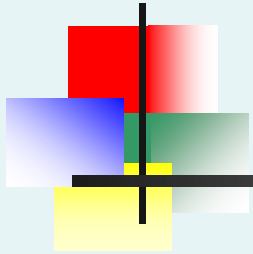
Sum = 0.36



Chapter Summary

- Discussed basic probability concepts
 - Sample spaces and events, contingency tables, Venn diagrams, simple probability, and joint probability
- Examined basic probability rules
 - General addition rule, addition rule for mutually exclusive events, rule for collectively exhaustive events
- Defined conditional probability
 - Statistical independence, marginal probability, decision trees, and the multiplication rule
- Discussed Bayes' theorem

Business Statistics: A First Course 5th Edition



Chapter 5

Some Important Discrete Probability Distributions



Learning Objectives

In this chapter, you learn:

- The properties of a probability distribution
- To calculate the expected value and variance of a probability distribution
- To calculate probabilities from binomial and Poisson distributions
- How to use the binomial and Poisson distributions to solve business problems

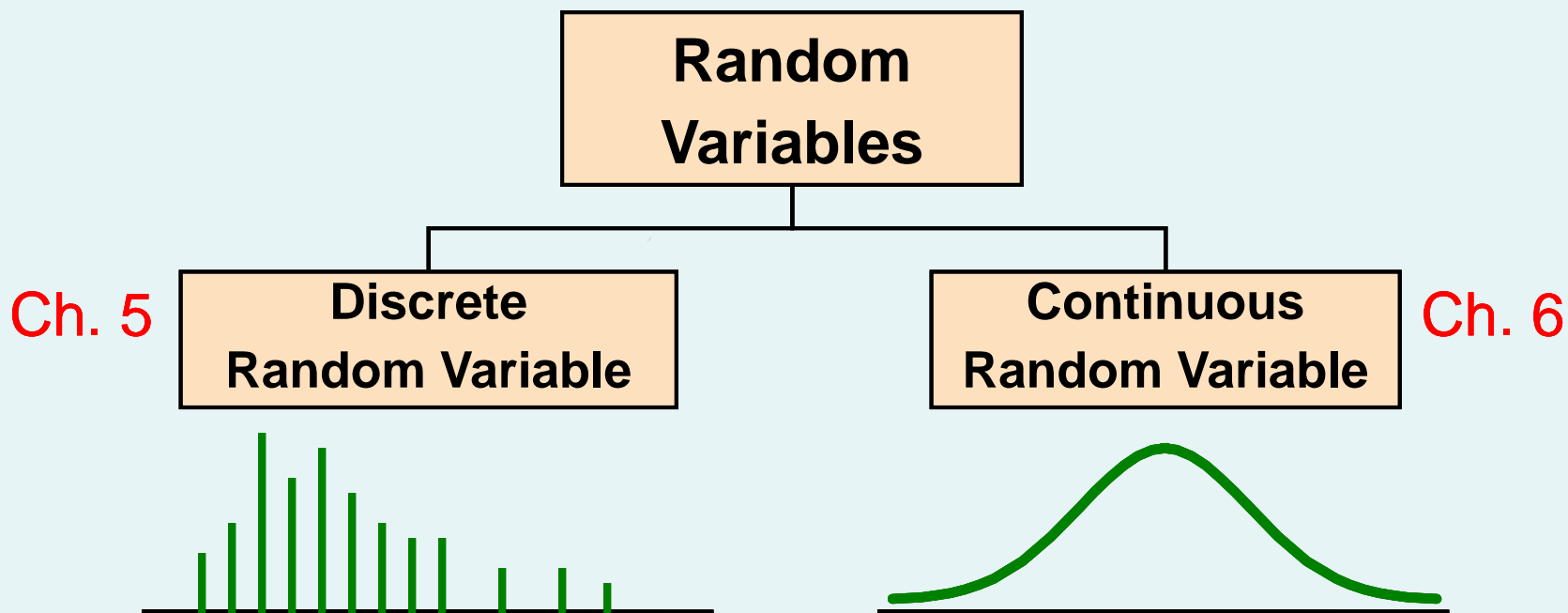


Definitions

Random Variables

- A **random variable** represents a possible numerical value from an uncertain event.
- **Discrete** random variables produce outcomes that come from a counting process (e.g. number of courses you are taking this semester).
- **Continuous** random variables produce outcomes that come from a measurement (e.g. your annual salary, or your weight).

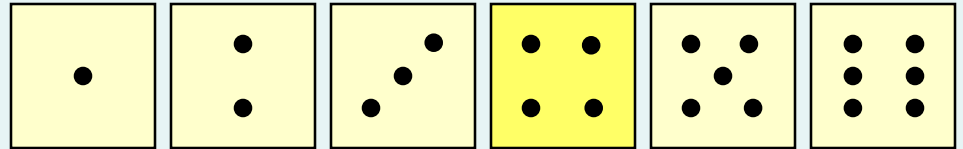
Definitions Random Variables



Discrete Random Variables

- Can only assume a countable number of values

Examples:



- Roll a die twice

Let X be the number of times 4 occurs
(then X could be 0, 1, or 2 times)

- Toss a coin 5 times.

Let X be the number of heads
(then $X = 0, 1, 2, 3, 4, \text{ or } 5$)





Probability Distribution For A Discrete Random Variable

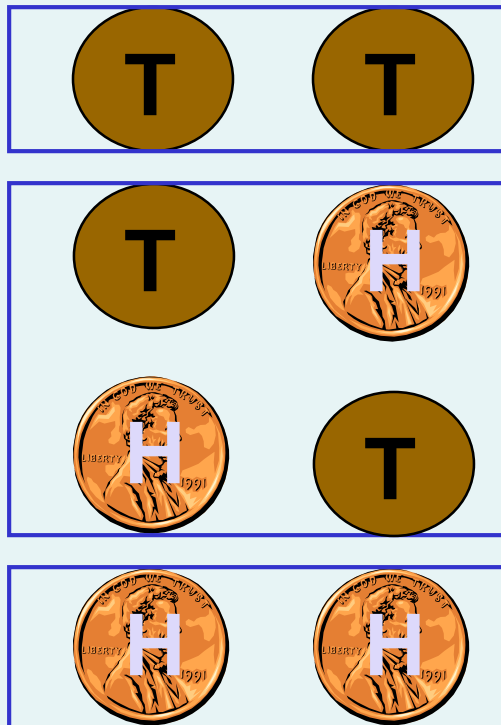
- A **probability distribution for a discrete random variable** is a mutually exclusive listing of all possible numerical outcomes for that variable and a probability of occurrence associated with each outcome.

| Number of Classes Taken | Probability |
|-------------------------|-------------|
| 2 | 0.2 |
| 3 | 0.4 |
| 4 | 0.24 |
| 5 | 0.16 |

Example of a Discrete Random Variable Probability Distribution

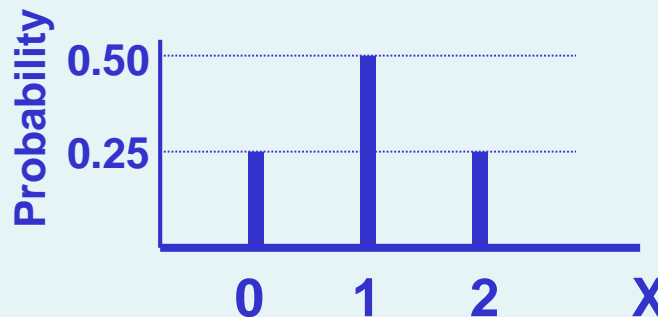
Experiment: Toss 2 Coins. Let $X = \#$ heads.

4 possible outcomes



Probability Distribution

| <u>X Value</u> | <u>Probability</u> |
|----------------|--------------------|
| 0 | $1/4 = 0.25$ |
| 1 | $2/4 = 0.50$ |
| 2 | $1/4 = 0.25$ |



Discrete Random Variables

Expected Value (Measuring Center)

- Expected Value (or mean) of a discrete random variable (Weighted Average)

$$\mu = E(X) = \sum_{i=1}^N X_i P(X_i)$$

- Example:** Toss 2 coins,
 $X = \#$ of heads,
compute expected value of X :

$$E(X) = ((0)(0.25) + (1)(0.50) + (2)(0.25)) \\ = 1.0$$

| X | P(X) |
|---|------|
| 0 | 0.25 |
| 1 | 0.50 |
| 2 | 0.25 |

Discrete Random Variables Measuring Dispersion

- Variance of a discrete random variable

$$\sigma^2 = \sum_{i=1}^N [X_i - E(X)]^2 P(X_i)$$

- Standard Deviation of a discrete random variable

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^N [X_i - E(X)]^2 P(X_i)}$$

where:

$E(X)$ = Expected value of the discrete random variable X

X_i = the i^{th} outcome of X

$P(X_i)$ = Probability of the i^{th} occurrence of X

Discrete Random Variables Measuring Dispersion

(continued)

- **Example:** Toss 2 coins, $X = \#$ heads, compute standard deviation (recall $E(X) = 1$)

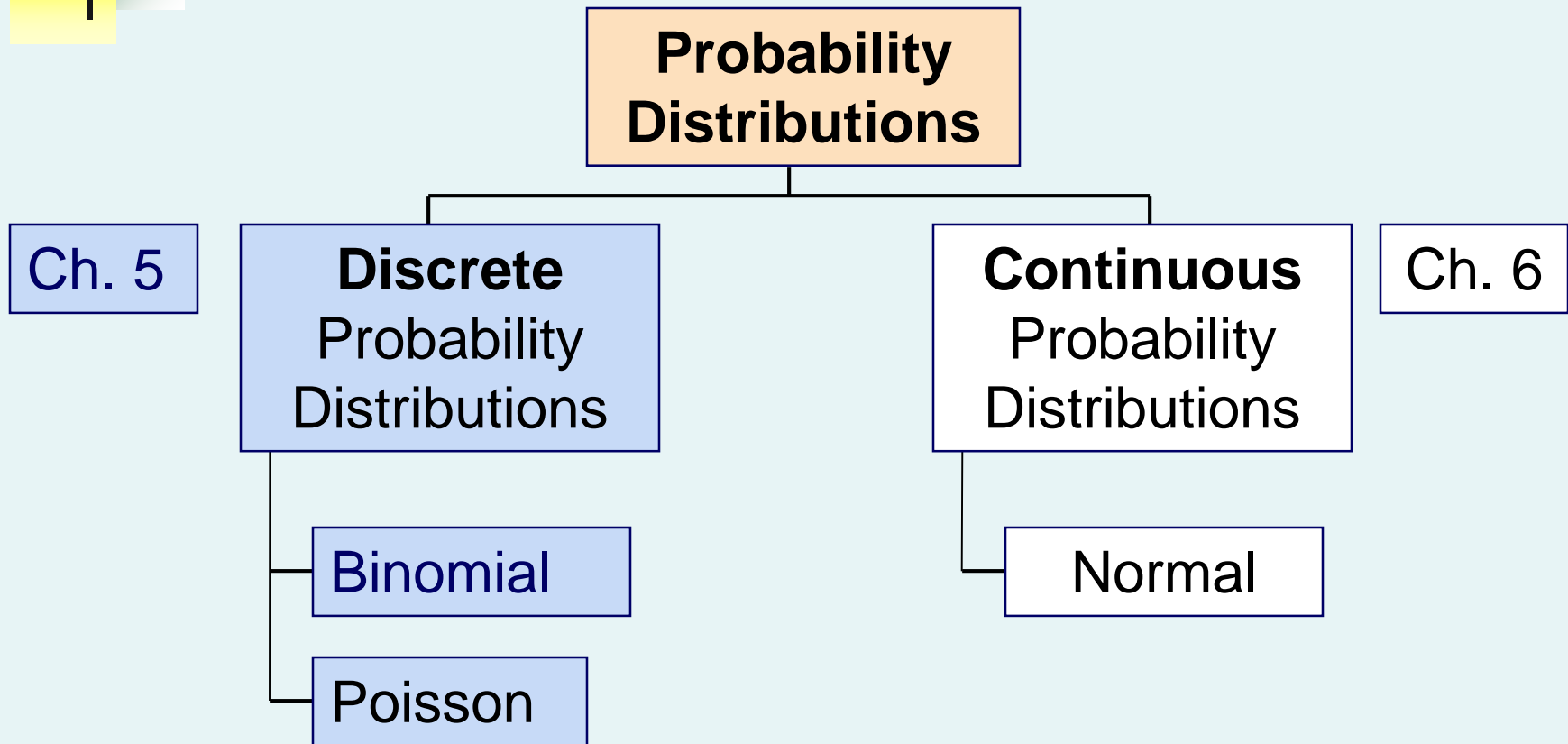
$$\sigma = \sqrt{\sum [X_i - E(X)]^2 P(X_i)}$$

$$\sigma = \sqrt{(0-1)^2(0.25) + (1-1)^2(0.50) + (2-1)^2(0.25)} = \sqrt{0.50} = 0.707$$

Possible number of heads
= 0, 1, or 2



Probability Distributions





Binomial Probability Distribution

- A fixed number of observations, n
 - e.g., 15 tosses of a coin; ten light bulbs taken from a warehouse
- Each observation is categorized as to whether or not the “event of interest” occurred
 - e.g., head or tail in each toss of a coin; defective or not defective light bulb
 - Since these two categories are mutually exclusive and collectively exhaustive
 - When the probability of the event of interest is represented as π , then the probability of the event of interest not occurring is $1 - \pi$
- Constant probability for the event of interest occurring (π) for each observation
 - Probability of getting a tail is the same each time we toss the coin



Binomial Probability Distribution

(continued)

- Observations are independent
 - The outcome of one observation does not affect the outcome of the other
 - Two sampling methods deliver independence
 - Infinite population without replacement
 - Finite population with replacement



Possible Applications for the Binomial Distribution

- A manufacturing plant labels items as either defective or acceptable
- A firm bidding for contracts will either get a contract or not
- A marketing research firm receives survey responses of “yes I will buy” or “no I will not”
- New job applicants either accept the offer or reject it

The Binomial Distribution

Counting Techniques



- Suppose the event of interest is obtaining heads on the toss of a fair coin. You are to toss the coin three times. In how many ways can you get two heads?
- Possible ways: HHT, HTH, THH, so there are three ways you can get two heads.
- This situation is fairly simple. We need to be able to count the number of ways for more complicated situations.

Counting Techniques

Rule of Combinations

- The number of **combinations** of selecting X objects out of n objects is

$${}_n C_x = \frac{n!}{X!(n-X)!}$$

where:

$$n! = (n)(n-1)(n-2) \cdots (2)(1)$$

$$X! = (X)(X-1)(X-2) \cdots (2)(1)$$

$$0! = 1 \quad (\text{by definition})$$

Counting Techniques

Rule of Combinations

- How many possible 3 scoop combinations could you create at an ice cream parlor if you have 31 flavors to select from?
- The total choices is $n = 31$, and we select $X = 3$.

$${}_{31}C_3 = \frac{31!}{3!(31-3)!} = \frac{31!}{3!28!} = \frac{31 \cdot 30 \cdot 29 \cdot 28!}{3 \cdot 2 \cdot 1 \cdot 28!} = 31 \cdot 5 \cdot 29 = 4495$$



Binomial Distribution Formula

$$P(X) = \frac{n!}{X!(n-X)!} \pi^X (1-\pi)^{n-X}$$

$P(X)$ = probability of X events of interest in n trials, with the probability of an “event of interest” being π for each trial

X = number of “events of interest” in sample, ($X = 0, 1, 2, \dots, n$)

n = sample size (number of trials or observations)

π = probability of “event of interest”

Example: Flip a coin four times, let $x = \#$ heads:

$$n = 4$$

$$\pi = 0.5$$

$$1 - \pi = (1 - 0.5) = 0.5$$

$$X = 0, 1, 2, 3, 4$$



Example: Calculating a Binomial Probability

What is the probability of one success in five observations if the probability of an event of interest is .1?

$$X = 1, n = 5, \text{ and } \pi = 0.1$$

$$\begin{aligned} P(X = 1) &= \frac{n!}{X!(n - X)!} \pi^X (1 - \pi)^{n - X} \\ &= \frac{5!}{1!(5 - 1)!} (0.1)^1 (1 - 0.1)^{5 - 1} \\ &= (5)(0.1)(0.9)^4 \\ &= 0.32805 \end{aligned}$$

The Binomial Distribution

Example

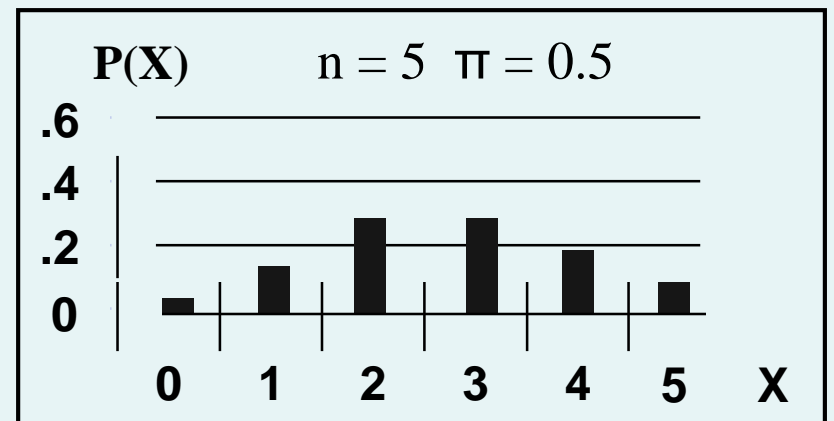
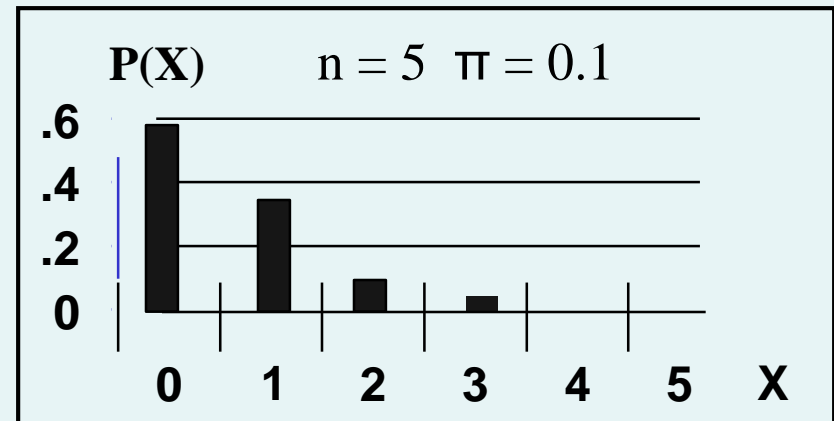
Suppose the probability of purchasing a defective computer is 0.02. What is the probability of purchasing 2 defective computers in a group of 10?

$$X = 2, n = 10, \text{ and } \pi = .02$$

$$\begin{aligned} P(X = 2) &= \frac{n!}{X!(n - X)!} \pi^X (1 - \pi)^{n - X} \\ &= \frac{10!}{2!(10 - 2)!} (.02)^2 (1 - .02)^{10 - 2} \\ &= (45)(.0004)(.8508) \\ &= .01531 \end{aligned}$$

The Binomial Distribution Shape

- The shape of the binomial distribution depends on the values of π and n
- Here, $n = 5$ and $\pi = .1$
- Here, $n = 5$ and $\pi = .5$



The Binomial Distribution Using Binomial Tables

| n = 10 | | | | | | | | | |
|----------|-----|-----------|-----------------------------|-----------|-----------------------------|-----------|-----------|-----------|----------|
| x | ... | $\pi=.20$ | $\pi=.25$ | $\pi=.30$ | $\pi=.35$ | $\pi=.40$ | $\pi=.45$ | $\pi=.50$ | |
| 0 | ... | 0.1074 | 0.0563 | 0.0282 | 0.0135 | 0.0060 | 0.0025 | 0.0010 | 10 |
| 1 | ... | 0.2684 | 0.1877 | 0.1211 | 0.0725 | 0.0403 | 0.0207 | 0.0098 | 9 |
| 2 | ... | 0.3020 | 0.2816 | 0.2335 | 0.1757 | 0.1209 | 0.0763 | 0.0439 | 8 |
| 3 | ... | 0.2013 | 0.2503 | 0.2668 | <u>0.2522</u> | 0.2150 | 0.1665 | 0.1172 | 7 |
| 4 | ... | 0.0881 | 0.1460 | 0.2001 | 0.2377 | 0.2508 | 0.2384 | 0.2051 | 6 |
| 5 | ... | 0.0264 | 0.0584 | 0.1029 | 0.1536 | 0.2007 | 0.2340 | 0.2461 | 5 |
| 6 | ... | 0.0055 | 0.0162 | 0.0368 | 0.0689 | 0.1115 | 0.1596 | 0.2051 | 4 |
| 7 | ... | 0.0008 | 0.0031 | 0.0090 | 0.0212 | 0.0425 | 0.0746 | 0.1172 | 3 |
| 8 | ... | 0.0001 | <u>0.0004</u> | 0.0014 | 0.0043 | 0.0106 | 0.0229 | 0.0439 | 2 |
| 9 | ... | 0.0000 | 0.0000 | 0.0001 | 0.0005 | 0.0016 | 0.0042 | 0.0098 | 1 |
| 10 | ... | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0003 | 0.0010 | 0 |
| | ... | $\pi=.80$ | $\pi=.75$ | $\pi=.70$ | $\pi=.65$ | $\pi=.60$ | $\pi=.55$ | $\pi=.50$ | x |

Examples:

$$n = 10, \pi = .35, x = 3: \quad P(x = 3|n = 10, \pi = .35) = .2522$$

$$n = 10, \pi = .75, x = 2: \quad P(x = 2|n = 10, \pi = .75) = .0004$$

Binomial Distribution Characteristics

- Mean

$$\mu = E(x) = n\pi$$

- Variance and Standard Deviation

$$\sigma^2 = n\pi(1 - \pi)$$

$$\sigma = \sqrt{n\pi(1 - \pi)}$$

Where n = sample size

π = probability of the event of interest for any trial

$(1 - \pi)$ = probability of no event of interest for any trial

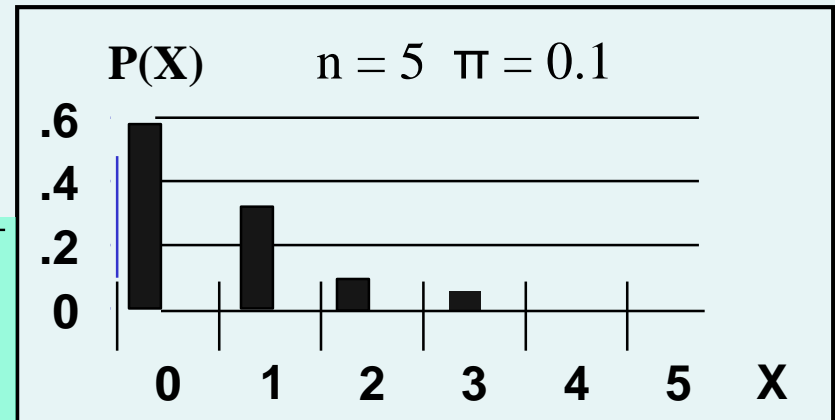
The Binomial Distribution

Characteristics

Examples

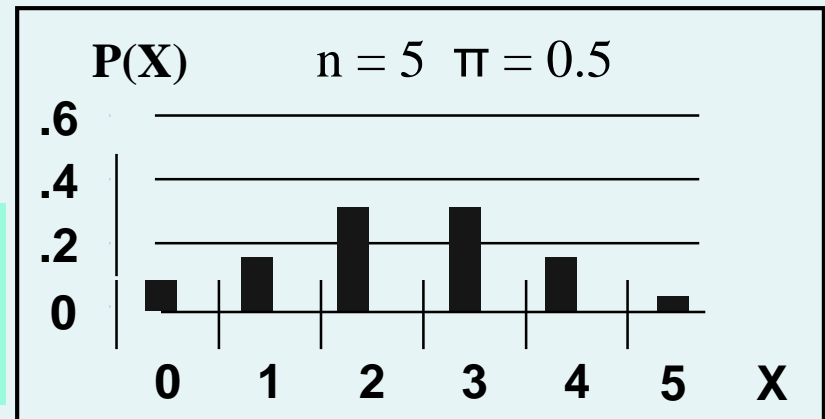
$$\mu = n\pi = (5)(.1) = 0.5$$

$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{(5)(.1)(1-.1)} \\ = 0.6708$$



$$\mu = n\pi = (5)(.5) = 2.5$$

$$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{(5)(.5)(1-.5)} \\ = 1.118$$



Using Excel For The Binomial Distribution

| | A | B |
|----|--------------------------------------------|---------------|
| 1 | Tagged Orders | |
| 2 | | |
| 3 | Data | |
| 4 | Sample size | 4 |
| 5 | Probability of an event of interest | 0.1 |
| 6 | | |
| 7 | Statistics | |
| 8 | Mean | 0.4 |
| 9 | Variance | 0.36 |
| 10 | Standard deviation | 0.6 |
| 11 | | |
| 12 | Binomial Probabilities Table | |
| 13 | X | P(X) |
| 14 | 0 | 0.6561 |
| 15 | 1 | 0.2916 |
| 16 | 2 | 0.0486 |
| 17 | 3 | 0.0036 |
| 18 | 4 | 0.0001 |

=B4 * B5

=B8 * (1 - B5)

=SQRT(B9)

=BINOMDIST(A14, \$B\$4, \$B\$5, FALSE)

=BINOMDIST(A15, \$B\$4, \$B\$5, FALSE)

=BINOMDIST(A16, \$B\$4, \$B\$5, FALSE)

=BINOMDIST(A17, \$B\$4, \$B\$5, FALSE)

=BINOMDIST(A18, \$B\$4, \$B\$5, FALSE)

The Poisson Distribution

Definitions



- You use the **Poisson distribution** when you are interested in the number of times an event occurs in a given **area of opportunity**.
- An **area of opportunity** is a continuous unit or interval of time, volume, or such area in which more than one occurrence of an event can occur.
 - The number of scratches in a car's paint
 - The number of mosquito bites on a person
 - The number of computer crashes in a day



The Poisson Distribution

- Apply the Poisson Distribution when:
 - You wish to count the number of times an event occurs in a given area of opportunity
 - The probability that an event occurs in one area of opportunity is the same for all areas of opportunity
 - The number of events that occur in one area of opportunity is independent of the number of events that occur in the other areas of opportunity
 - The probability that two or more events occur in an area of opportunity approaches zero as the area of opportunity becomes smaller
 - The average number of events per unit is λ (lambda)



Poisson Distribution Formula

$$P(X) = \frac{e^{-\lambda} \lambda^x}{X!}$$

where:

X = number of events in an area of opportunity

λ = expected number of events

e = base of the natural logarithm system (2.71828...)

Poisson Distribution Characteristics

- Mean

$$\mu = \lambda$$

- Variance and Standard Deviation

$$\sigma^2 = \lambda$$

$$\sigma = \sqrt{\lambda}$$

where λ = expected number of events

Using Poisson Tables

| X | λ | | | | | | | | |
|---|-----------|--------|--------|--------|---------------|--------|--------|--------|--------|
| | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 |
| 0 | 0.9048 | 0.8187 | 0.7408 | 0.6703 | 0.6065 | 0.5488 | 0.4966 | 0.4493 | 0.4066 |
| 1 | 0.0905 | 0.1637 | 0.2222 | 0.2681 | 0.3093 | 0.3293 | 0.3476 | 0.3595 | 0.3659 |
| 2 | 0.0045 | 0.0164 | 0.0333 | 0.0536 | 0.0758 | 0.0988 | 0.1217 | 0.1438 | 0.1647 |
| 3 | 0.0002 | 0.0011 | 0.0033 | 0.0072 | 0.0126 | 0.0198 | 0.0284 | 0.0383 | 0.0494 |
| 4 | 0.0000 | 0.0001 | 0.0003 | 0.0007 | 0.0016 | 0.0030 | 0.0050 | 0.0077 | 0.0111 |
| 5 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0002 | 0.0004 | 0.0007 | 0.0012 | 0.0020 |
| 6 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0001 | 0.0002 | 0.0003 |
| 7 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

Example: Find $P(X = 2)$ if $\lambda = 0.50$

$$P(X = 2) = \frac{e^{-\lambda} \lambda^X}{X!} = \frac{e^{-0.50} (0.50)^2}{2!} = 0.0758$$

Using Excel For The Poisson Distribution

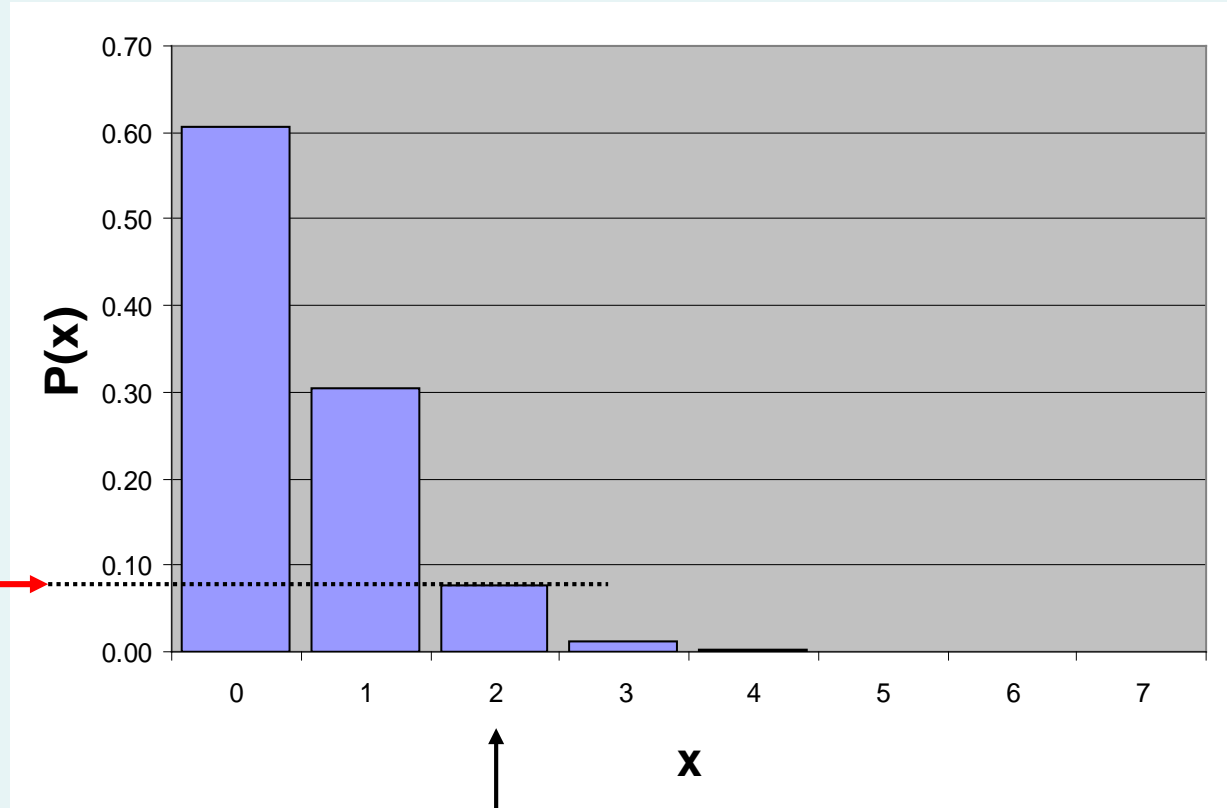
| | A | B | C | D | E |
|----|----------------------------------------------------|-----------------|------------------------------|---|----------|
| 1 | Customer Arrivals Analysis | | | | |
| 2 | | | | | |
| 3 | Data | | | | |
| 4 | Mean/Expected number of events of interest: | | | | 3 |
| 5 | | | | | |
| 6 | Poisson Probabilities Table | | | | |
| 7 | X | P(X) | | | |
| 8 | 0 | 0.049787 | =POISSON(A8, \$E\$4, FALSE) | | |
| 9 | 1 | 0.149361 | =POISSON(A9, \$E\$4, FALSE) | | |
| 10 | 2 | 0.224042 | =POISSON(A10, \$E\$4, FALSE) | | |
| 11 | 3 | 0.224042 | =POISSON(A11, \$E\$4, FALSE) | | |
| 12 | 4 | 0.168031 | =POISSON(A12, \$E\$4, FALSE) | | |
| 13 | 5 | 0.100819 | =POISSON(A13, \$E\$4, FALSE) | | |
| 14 | 6 | 0.050409 | =POISSON(A14, \$E\$4, FALSE) | | |
| 15 | 7 | 0.021604 | =POISSON(A15, \$E\$4, FALSE) | | |
| 16 | 8 | 0.008102 | =POISSON(A16, \$E\$4, FALSE) | | |
| 17 | 9 | 0.002701 | =POISSON(A17, \$E\$4, FALSE) | | |
| 18 | 10 | 0.000810 | =POISSON(A18, \$E\$4, FALSE) | | |
| 19 | 11 | 0.000221 | =POISSON(A19, \$E\$4, FALSE) | | |
| 20 | 12 | 0.000055 | =POISSON(A20, \$E\$4, FALSE) | | |
| 21 | 13 | 0.000013 | =POISSON(A21, \$E\$4, FALSE) | | |
| 22 | 14 | 0.000003 | =POISSON(A22, \$E\$4, FALSE) | | |
| 23 | 15 | 0.000001 | =POISSON(A23, \$E\$4, FALSE) | | |
| 24 | 16 | 0.000000 | =POISSON(A24, \$E\$4, FALSE) | | |
| 25 | 17 | 0.000000 | =POISSON(A25, \$E\$4, FALSE) | | |
| 26 | 18 | 0.000000 | =POISSON(A26, \$E\$4, FALSE) | | |
| 27 | 19 | 0.000000 | =POISSON(A27, \$E\$4, FALSE) | | |
| 28 | 20 | 0.000000 | =POISSON(A28, \$E\$4, FALSE) | | |

Graph of Poisson Probabilities

Graphically:

$\lambda = 0.50$

| X | $\lambda = 0.50$ |
|----------|------------------------------------|
| 0 | 0.6065 |
| 1 | 0.3033 |
| 2 | 0.0758 |
| 3 | 0.0126 |
| 4 | 0.0016 |
| 5 | 0.0002 |
| 6 | 0.0000 |
| 7 | 0.0000 |

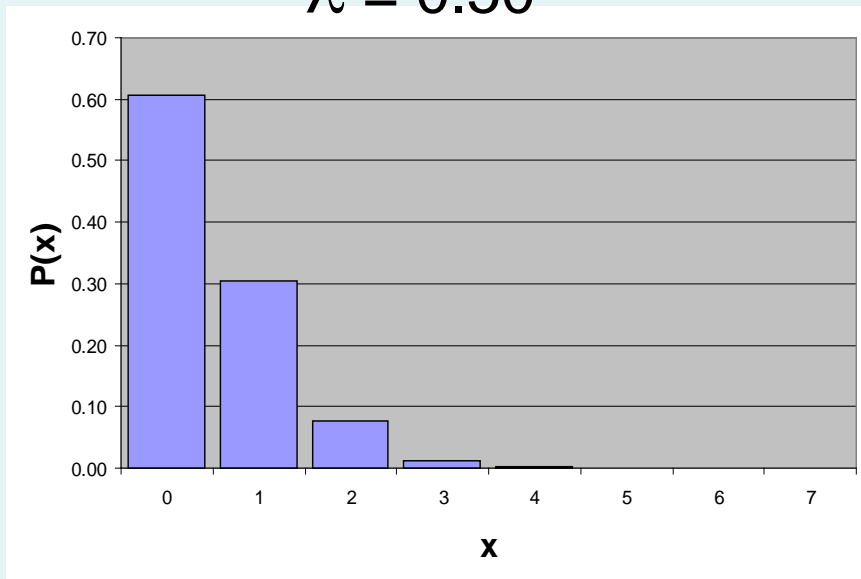


$$P(X = 2) = 0.0758$$

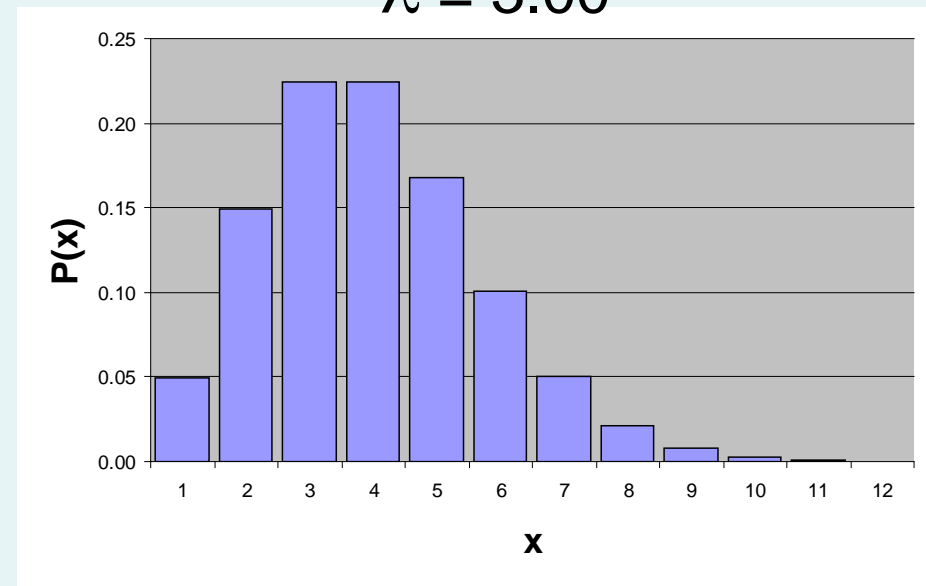
Poisson Distribution Shape

- The shape of the Poisson Distribution depends on the parameter λ :

$\lambda = 0.50$



$\lambda = 3.00$

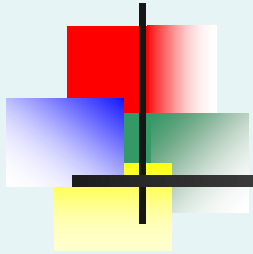




Chapter Summary

- Addressed the probability distribution of a discrete random variable
- Discussed the Binomial distribution
- Discussed the Poisson distribution

Business Statistics: A First Course 5th Edition



Chapter 6

The Normal Distribution



Learning Objectives

In this chapter, you learn:

- To compute probabilities from the normal distribution
- To use the normal probability plot to determine whether a set of data is approximately normally distributed



Continuous Probability Distributions

- A **continuous random variable** is a variable that can assume any value on a continuum (can assume an uncountable number of values)
 - thickness of an item
 - time required to complete a task
 - temperature of a solution
 - height, in inches
- These can potentially take on any value depending only on the ability to precisely and accurately measure

The Normal Distribution

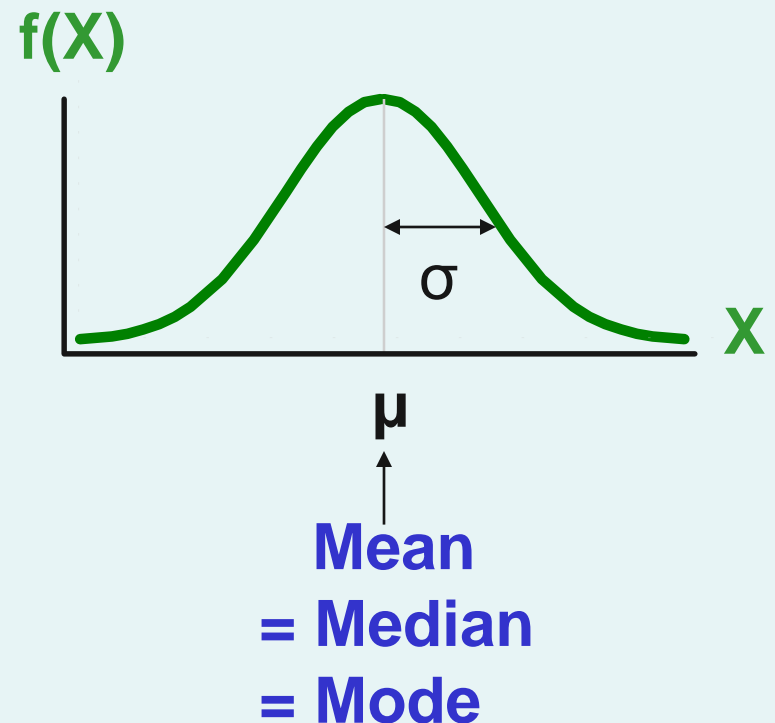
- 'Bell Shaped'
- Symmetrical
- Mean, Median and Mode are Equal

Location is determined by the mean, μ

Spread is determined by the standard deviation, σ

The random variable has an infinite theoretical range:

$+\infty$ to $-\infty$



The Normal Distribution Density Function

- The formula for the normal probability density function is

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

Where e = the mathematical constant approximated by 2.71828

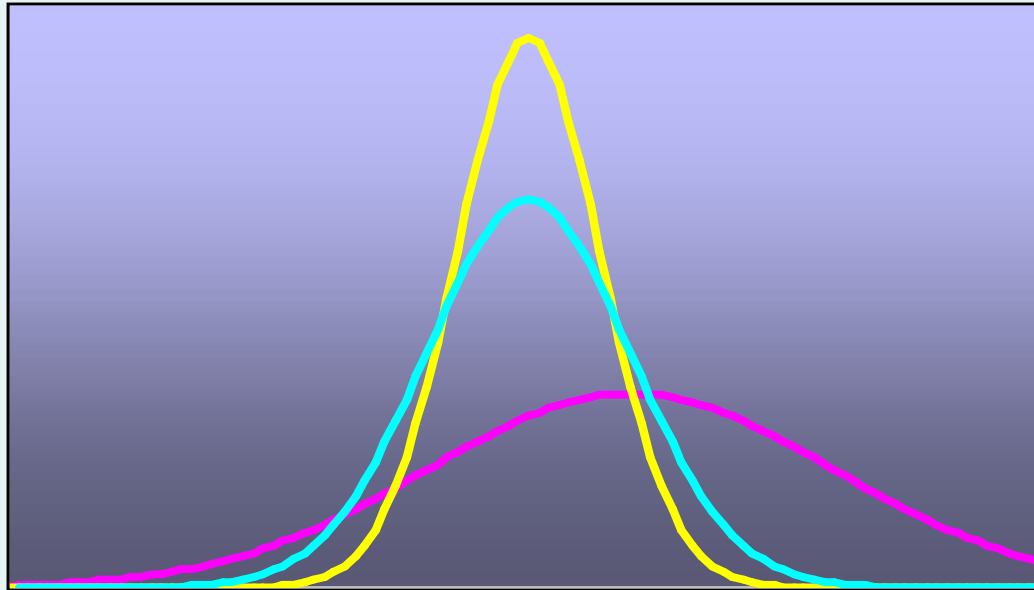
π = the mathematical constant approximated by 3.14159

μ = the population mean

σ = the population standard deviation

X = any value of the continuous variable

Many Normal Distributions

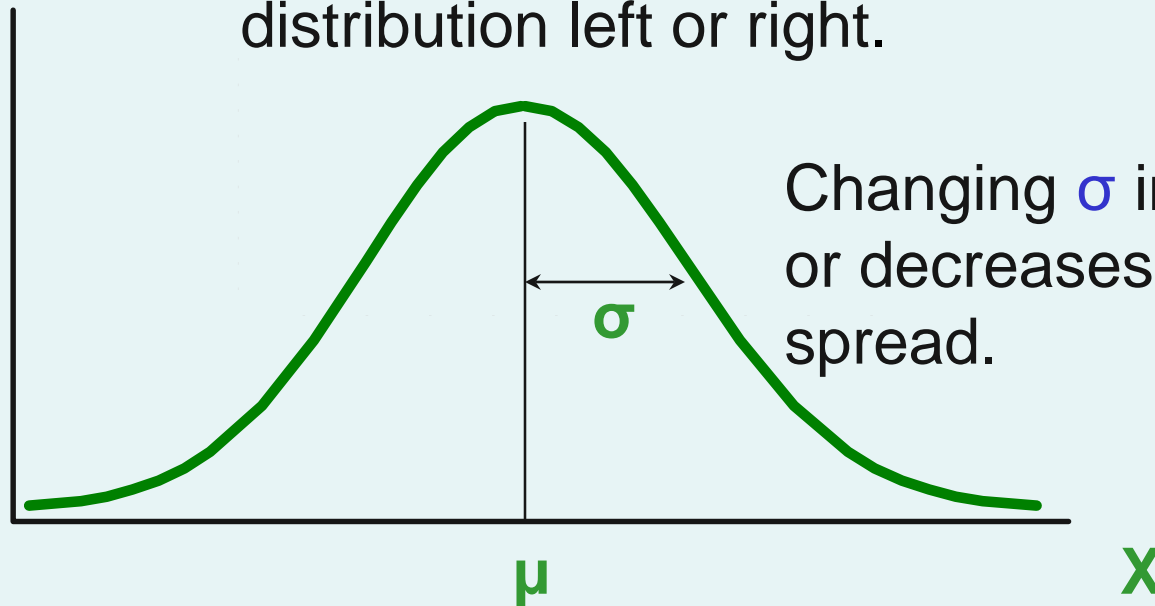


By varying the parameters μ and σ , we obtain different normal distributions

The Normal Distribution Shape

$f(X)$

Changing μ shifts the distribution left or right.





The Standardized Normal

- Any normal distribution (with any mean and standard deviation combination) can be transformed into the standardized normal distribution (Z)
- Need to transform X units into Z units
- The standardized normal distribution (Z) has a mean of 0 and a standard deviation of 1



Translation to the Standardized Normal Distribution

- Translate from X to the standardized normal (the “ Z ” distribution) by **subtracting the mean** of X and **dividing by its standard deviation**:

$$Z = \frac{X - \mu}{\sigma}$$

The Z distribution always has mean = 0 and standard deviation = 1



The Standardized Normal Probability Density Function

- The formula for the standardized normal probability density function is

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)Z^2}$$

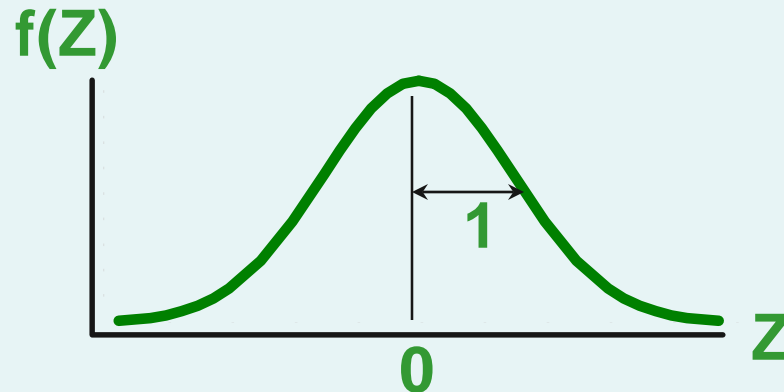
Where e = the mathematical constant approximated by 2.71828

π = the mathematical constant approximated by 3.14159

Z = any value of the standardized normal distribution

The Standardized Normal Distribution

- Also known as the “Z” distribution
- Mean is 0
- Standard Deviation is 1



Values above the mean have **positive** Z-values, values below the mean have **negative** Z-values



Example

- If X is distributed normally with mean of 100 and standard deviation of 50, the Z value for $X = 200$ is

$$Z = \frac{X - \mu}{\sigma} = \frac{200 - 100}{50} = 2.0$$

- This says that $X = 200$ is two standard deviations (2 increments of 50 units) above the mean of 100.

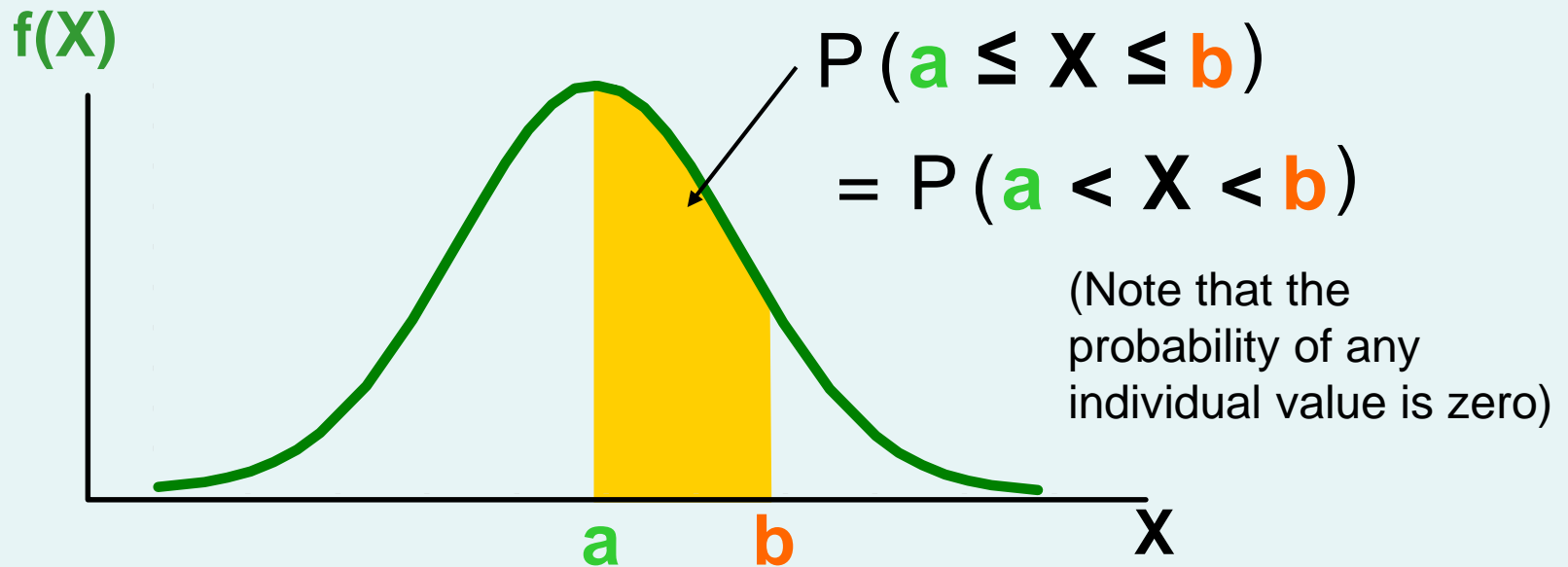
Comparing X and Z units



Note that the shape of the distribution is the same, only the scale has changed. We can express the problem in original units (X) or in standardized units (Z)

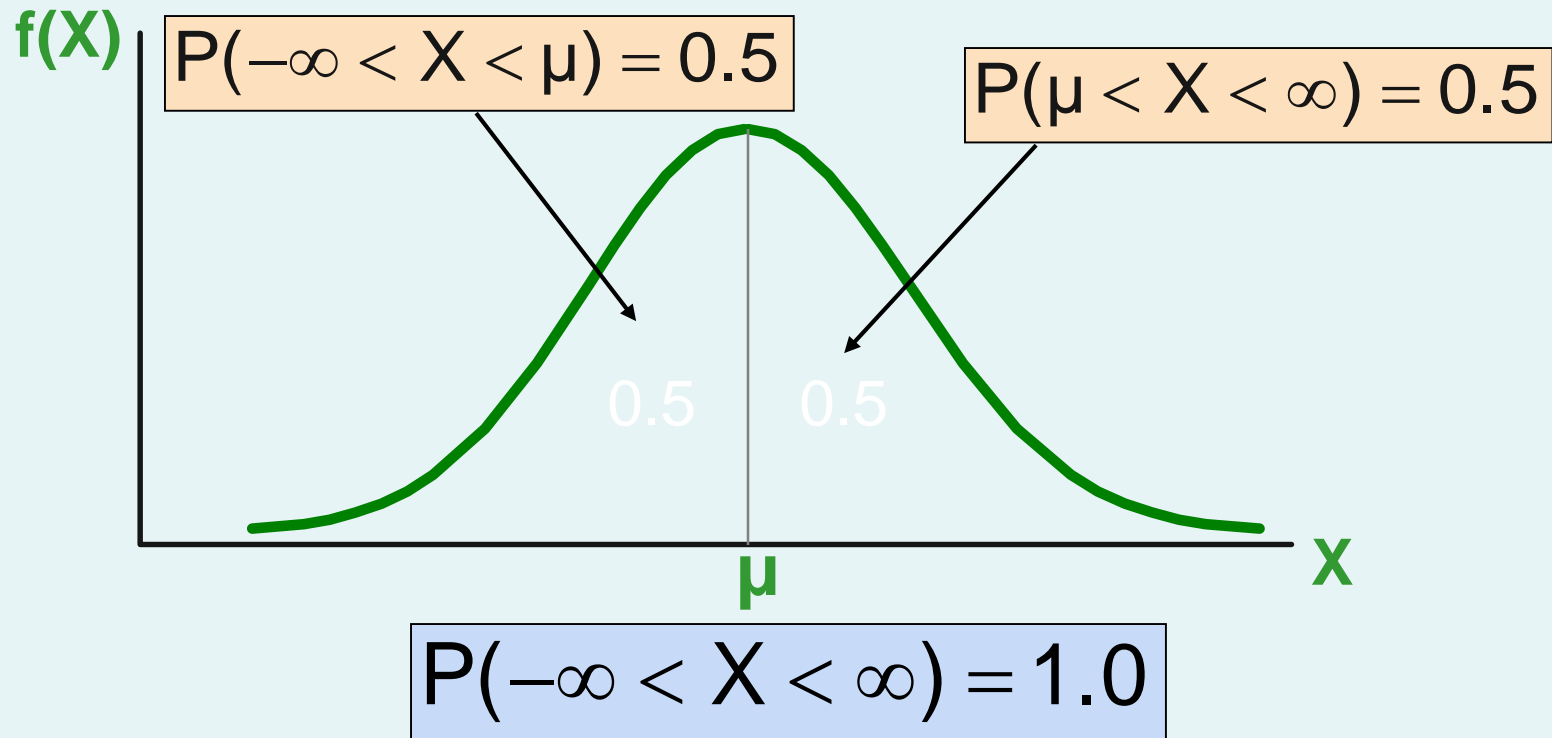
Finding Normal Probabilities

Probability is measured by the area under the curve



Probability as Area Under the Curve

The total area under the curve is 1.0, and the curve is symmetric, so half is above the mean, half is below

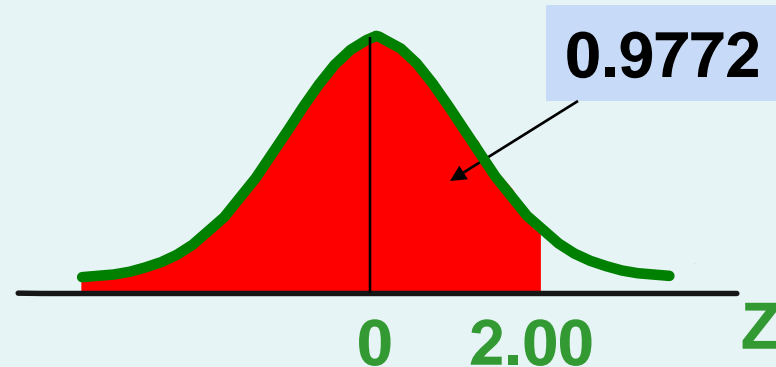


The Standardized Normal Table

- The Cumulative Standardized Normal table in the textbook (Appendix table E.2) gives the probability **less than** a desired value of Z (i.e., from negative infinity to Z)

Example:

$$P(Z < 2.00) = 0.9772$$



The Standardized Normal Table

(continued)

The **column** gives the value of Z to the second decimal point

The **row** shows the value of Z to the first decimal point

| Z | 0.00 | 0.01 | 0.02 ... |
|-----|-------|------|----------|
| 0.0 | | | |
| 0.1 | | | |
| . | | | |
| . | | | |
| 2.0 | .9772 | | |

The value within the table gives the **probability** from $Z = -\infty$ up to the desired Z value

$$P(Z < 2.00) = 0.9772$$



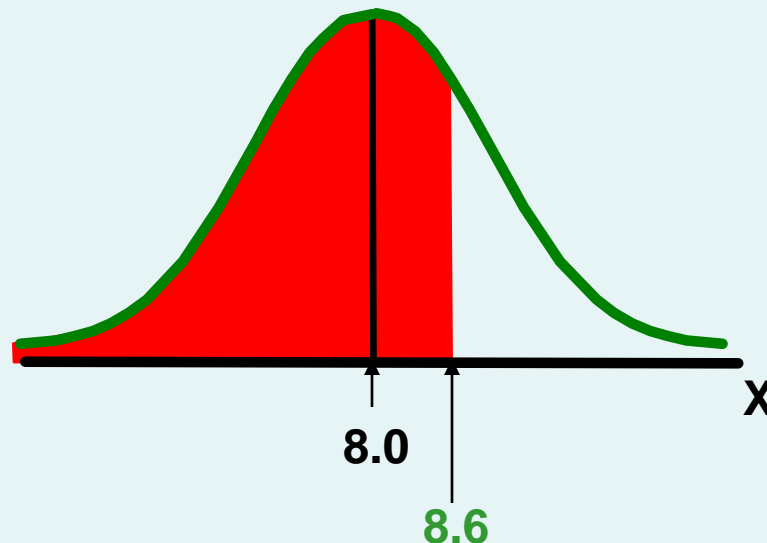
General Procedure for Finding Normal Probabilities

To find $P(a < X < b)$ when X is distributed normally:

- Draw the normal curve for the problem in terms of X
- Translate X -values to Z -values
- Use the Standardized Normal Table

Finding Normal Probabilities

- Let X represent the time it takes to download an image file from the internet.
- Suppose X is normal with mean 8.0 and standard deviation 5.0. Find $P(X < 8.6)$

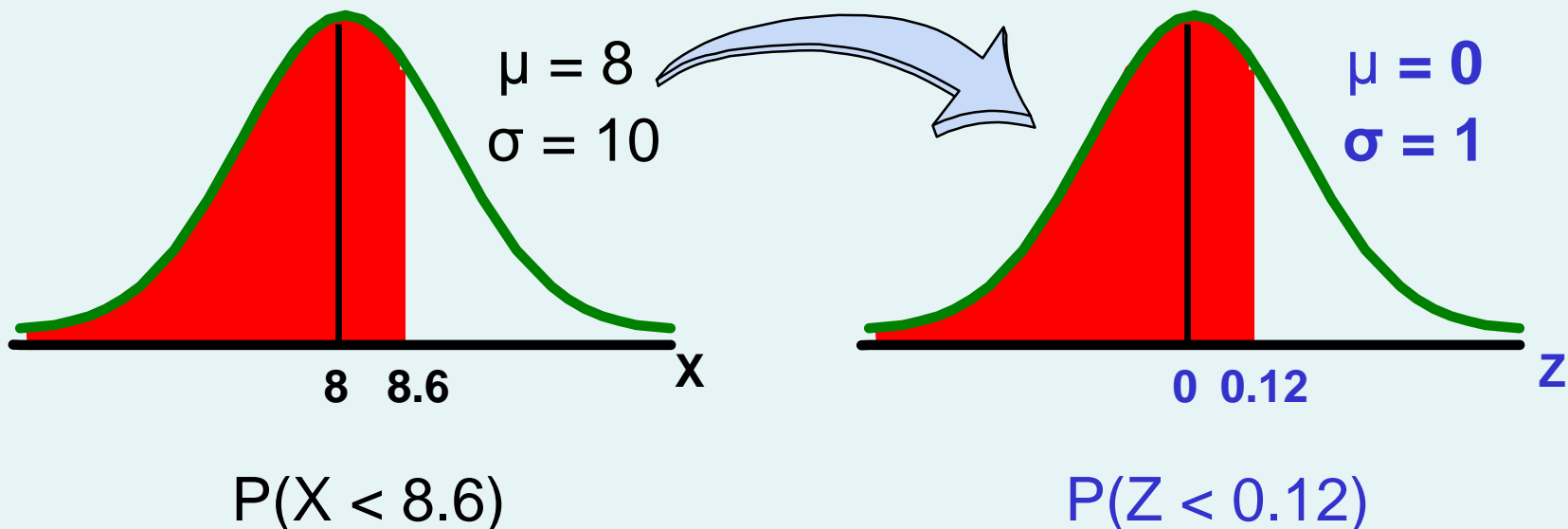


Finding Normal Probabilities

(continued)

- Let X represent the time it takes to download an image file from the internet.
- Suppose X is normal with mean 8.0 and standard deviation 5.0. Find $P(X < 8.6)$

$$Z = \frac{X - \mu}{\sigma} = \frac{8.6 - 8.0}{5.0} = 0.12$$



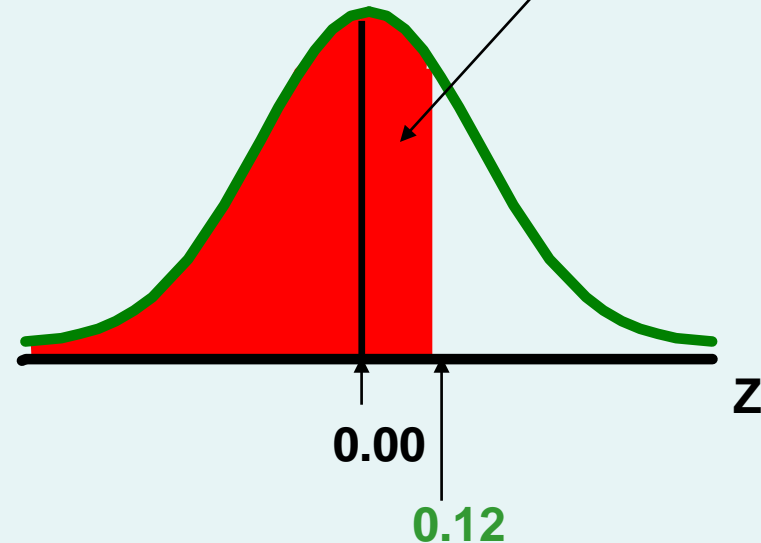
Solution: Finding $P(Z < 0.12)$

Standardized Normal Probability Table (Portion)

| Z | .00 | .01 | .02 |
|------------|-------|-------|--------------|
| 0.0 | .5000 | .5040 | .5080 |
| 0.1 | .5398 | .5438 | .5478 |
| 0.2 | .5793 | .5832 | .5871 |
| 0.3 | .6179 | .6217 | .6255 |

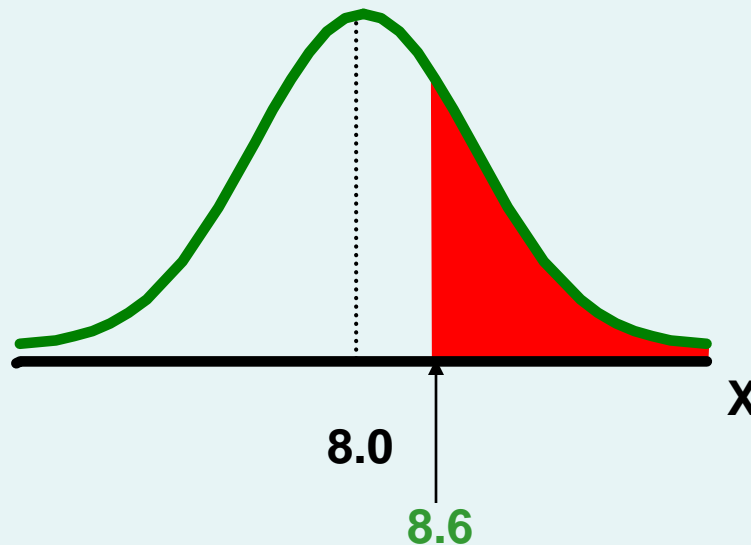
$$P(X < 8.6) = P(Z < 0.12)$$

.5478



Finding Normal Upper Tail Probabilities

- Suppose X is normal with mean 8.0 and standard deviation 5.0.
- Now Find $P(X > 8.6)$

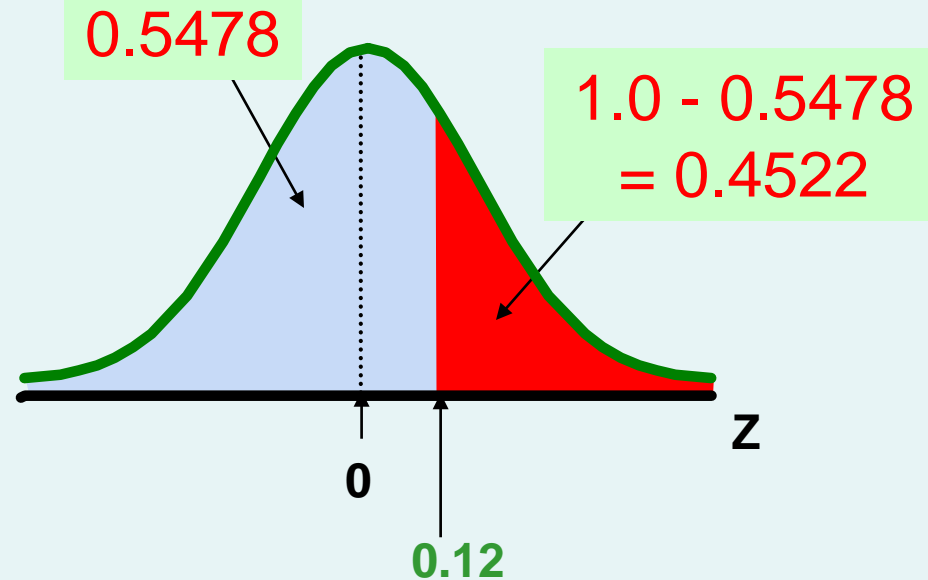
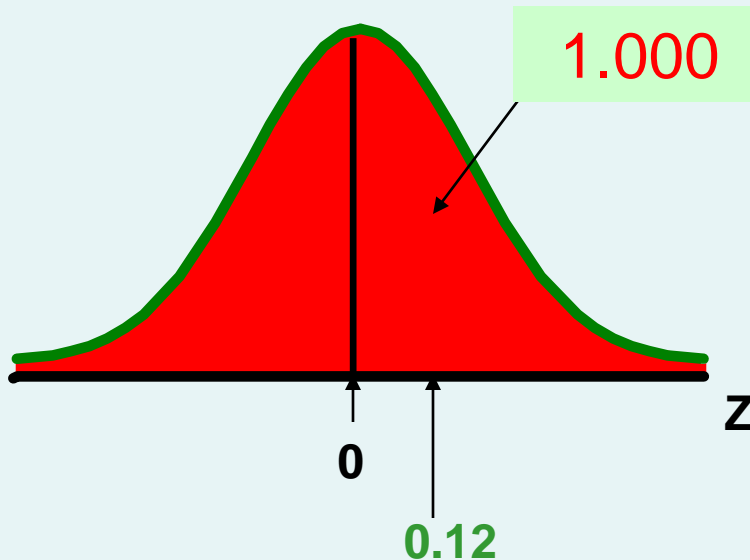


Finding Normal Upper Tail Probabilities

(continued)

- Now Find $P(X > 8.6)$...

$$\begin{aligned} P(X > 8.6) &= P(Z > 0.12) = 1.0 - P(Z \leq 0.12) \\ &= 1.0 - 0.5478 = \mathbf{0.4522} \end{aligned}$$



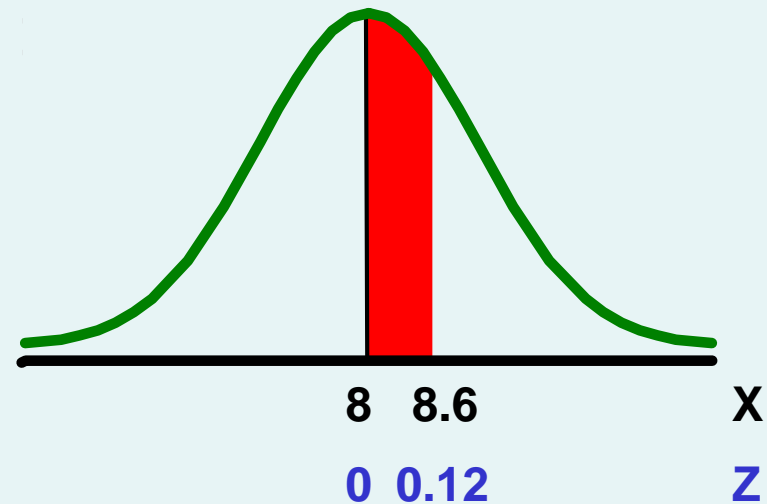
Finding a Normal Probability Between Two Values

- Suppose X is normal with mean 8.0 and standard deviation 5.0. Find $P(8 < X < 8.6)$

Calculate Z-values:

$$Z = \frac{X - \mu}{\sigma} = \frac{8 - 8}{5} = 0$$

$$Z = \frac{X - \mu}{\sigma} = \frac{8.6 - 8}{5} = 0.12$$



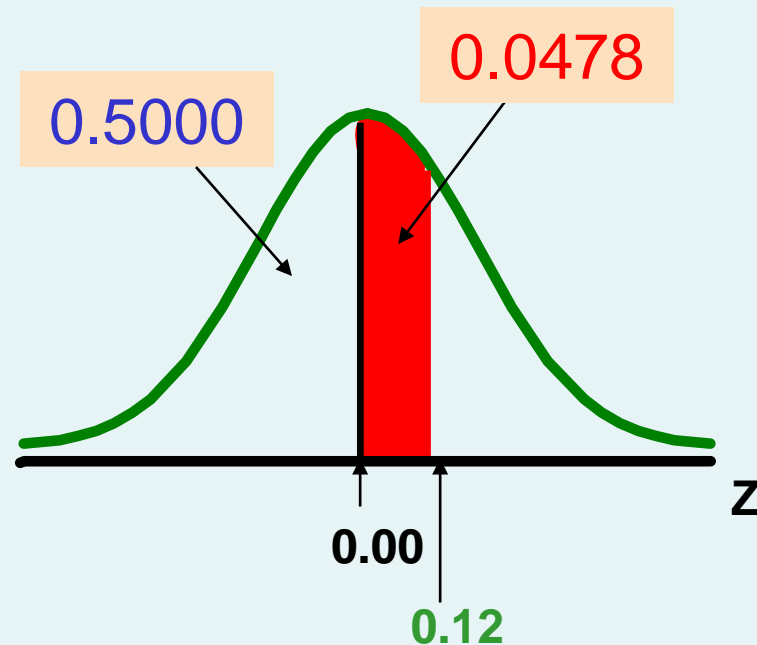
$$\begin{aligned} P(8 < X < 8.6) \\ = P(0 < Z < 0.12) \end{aligned}$$

Solution: Finding $P(0 < Z < 0.12)$

Standardized Normal Probability Table (Portion)

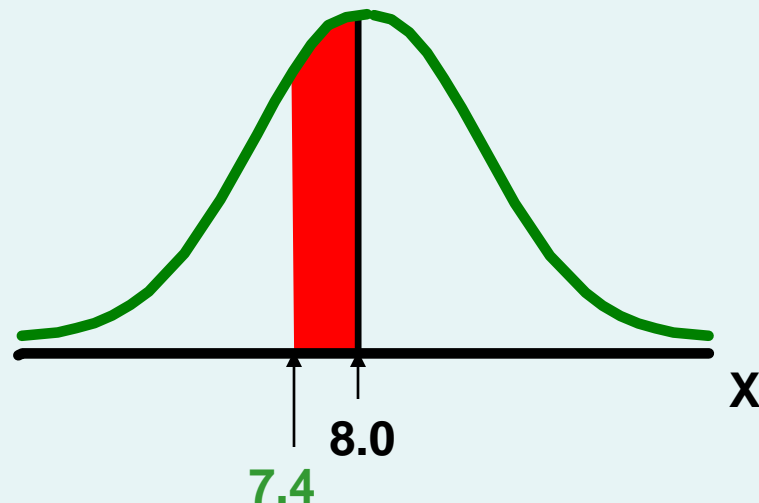
| Z | .00 | .01 | .02 |
|------------|--------------|-------|--------------|
| 0.0 | .5000 | .5040 | .5080 |
| 0.1 | .5398 | .5438 | .5478 |
| 0.2 | .5793 | .5832 | .5871 |
| 0.3 | .6179 | .6217 | .6255 |

$$\begin{aligned} P(8 < X < 8.6) &= P(0 < Z < 0.12) \\ &= P(Z < 0.12) - P(Z \leq 0) \\ &= 0.5478 - .5000 = \mathbf{0.0478} \end{aligned}$$



Probabilities in the Lower Tail

- Suppose X is normal with mean 8.0 and standard deviation 5.0.
- Now Find $P(7.4 < X < 8)$



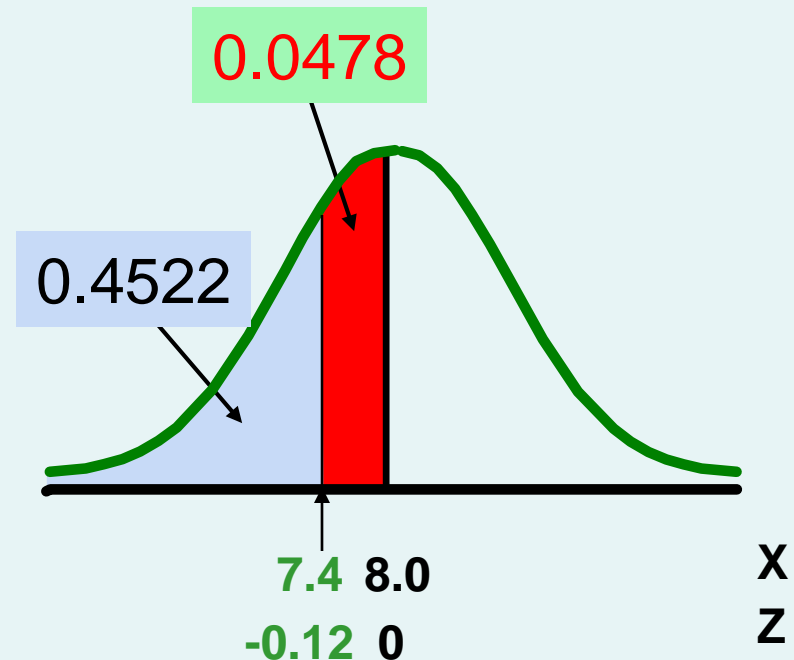
Probabilities in the Lower Tail

(continued)

Now Find $P(7.4 < X < 8)$...

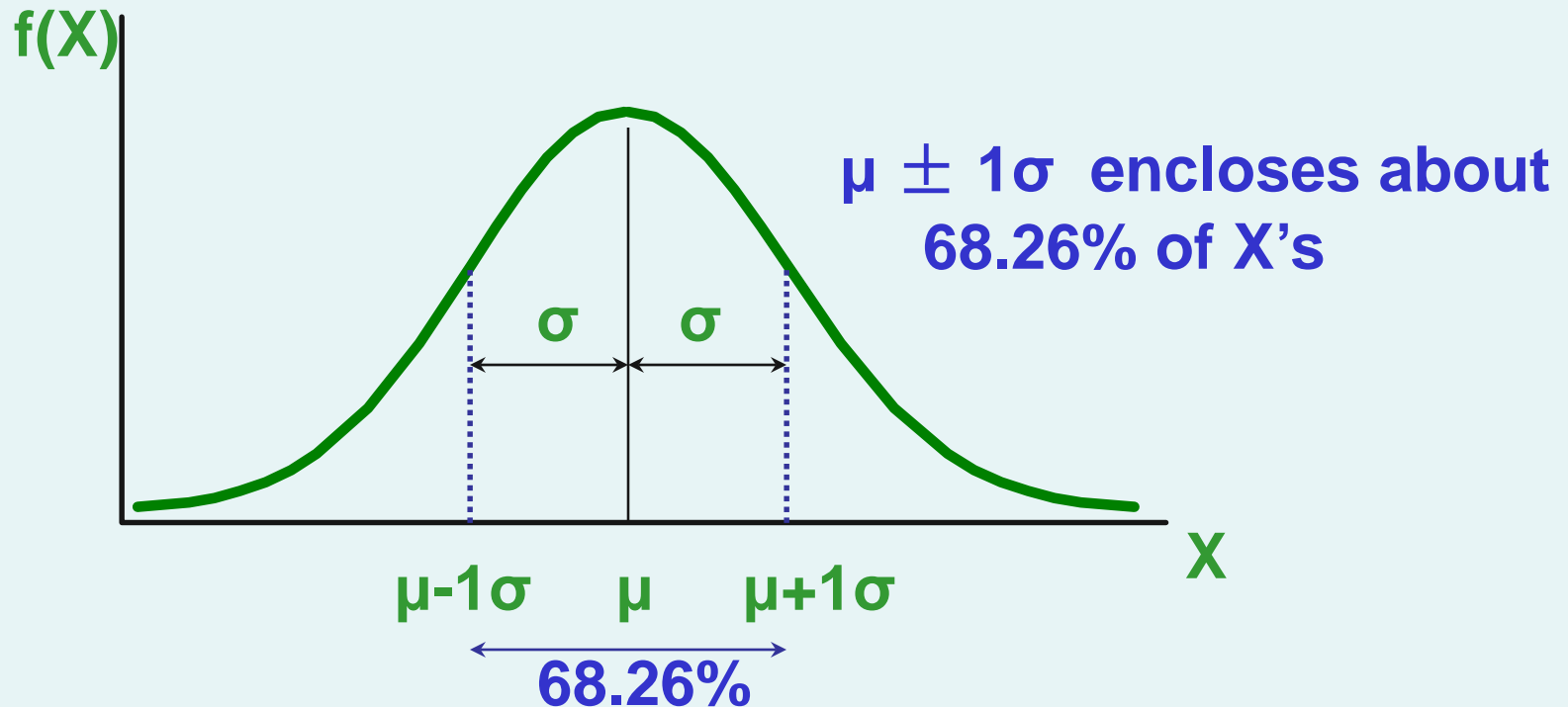
$$\begin{aligned} &P(7.4 < X < 8) \\ &= P(-0.12 < Z < 0) \\ &= P(Z < 0) - P(Z \leq -0.12) \\ &= 0.5000 - 0.4522 = \mathbf{0.0478} \end{aligned}$$

The Normal distribution is symmetric, so this probability is the same as $P(0 < Z < 0.12)$



Empirical Rules

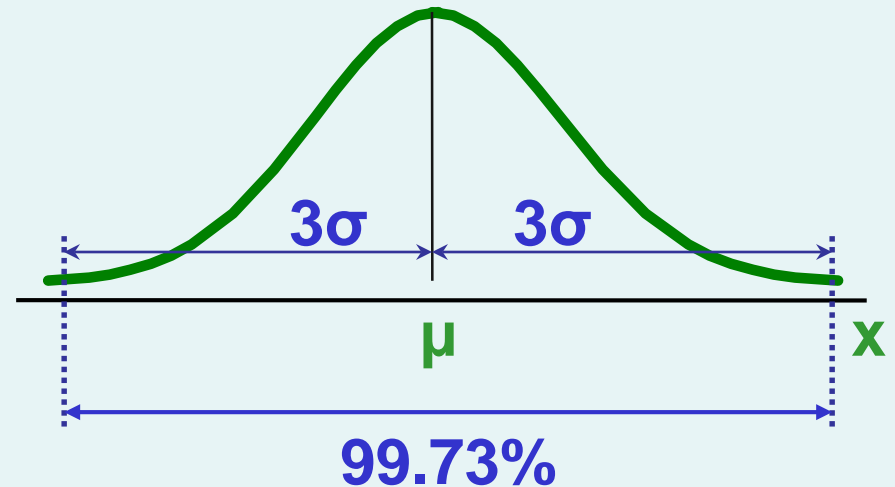
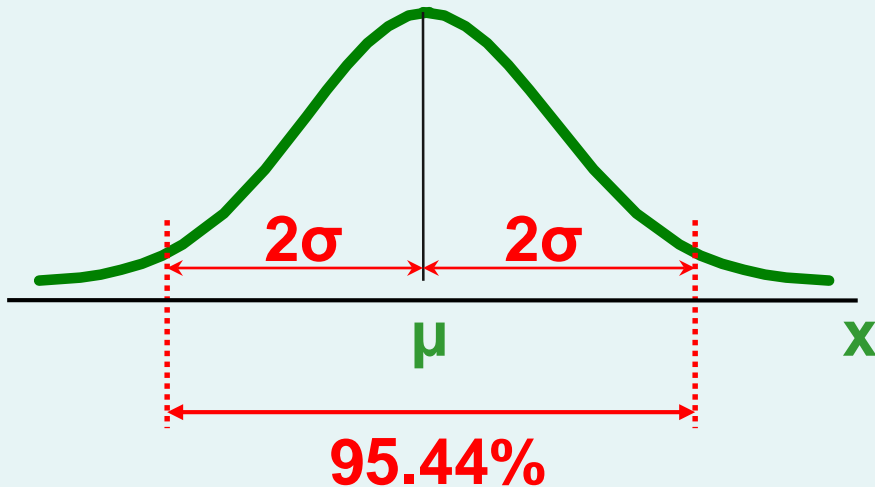
What can we say about the distribution of values around the mean? For any normal distribution:



The Empirical Rule

(continued)

- $\mu \pm 2\sigma$ covers about **95%** of X 's
- $\mu \pm 3\sigma$ covers about **99.7%** of X 's



Given a Normal Probability Find the X Value

- Steps to find the X value for a known probability:
 1. Find the Z value for the known probability
 2. Convert to X units using the formula:

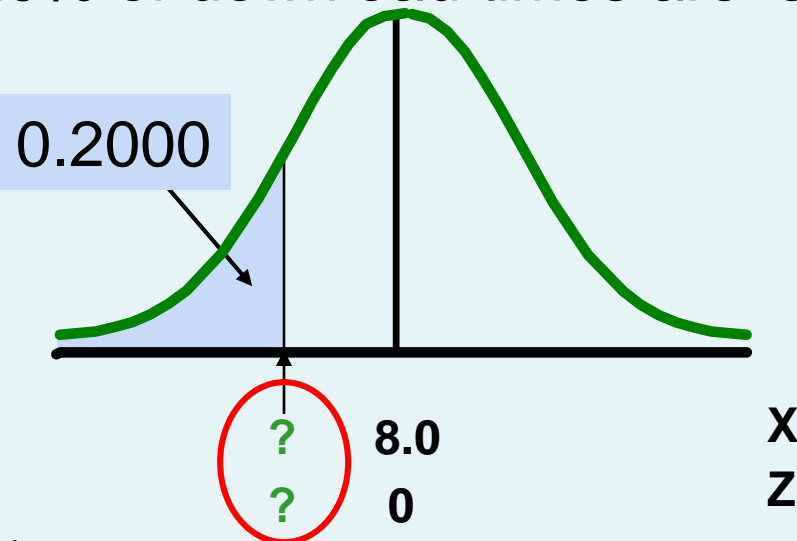
$$X = \mu + Z\sigma$$

Finding the X value for a Known Probability

(continued)

Example:

- Let X represent the time it takes (in seconds) to download an image file from the internet.
- Suppose X is normal with mean 8.0 and standard deviation 5.0
- Find X such that 20% of download times are less than X .



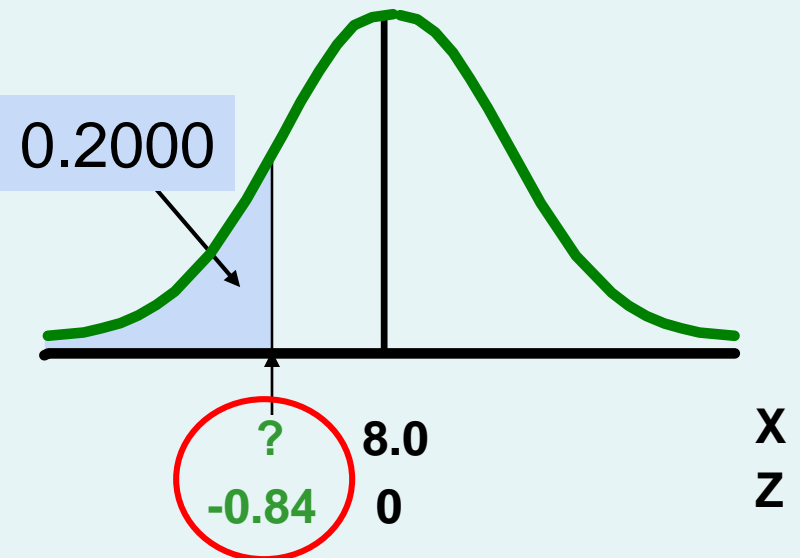
Find the Z value for 20% in the Lower Tail

1. Find the Z value for the known probability

Standardized Normal Probability Table (Portion)

| Z | ... | .03 | .04 | .05 |
|-------------|-----|-------|--------------|-------|
| -0.9 | ... | .1762 | .1736 | .1711 |
| -0.8 | ... | .2033 | .2005 | .1977 |
| -0.7 | ... | .2327 | .2296 | .2266 |

- 20% area in the lower tail is consistent with a Z value of **-0.84**





Finding the X value

2. Convert to X units using the formula:

$$\begin{aligned} X &= \mu + Z\sigma \\ &= 8.0 + (-0.84)5.0 \\ &= 3.80 \end{aligned}$$

So 20% of the values from a distribution with mean 8.0 and standard deviation 5.0 are less than 3.80



Evaluating Normality

- Not all continuous distributions are normal
- It is important to evaluate how well the data set is approximated by a normal distribution.
- Normally distributed data should approximate the theoretical normal distribution:
 - The normal distribution is bell shaped (symmetrical) where the mean is equal to the median.
 - The empirical rule applies to the normal distribution.
 - The interquartile range of a normal distribution is 1.33 standard deviations.



Evaluating Normality

(continued)

Comparing data characteristics to theoretical properties

■ Construct **charts or graphs**

- For small- or moderate-sized data sets, construct a stem-and-leaf display or a boxplot to check for symmetry
- For large data sets, does the histogram or polygon appear bell-shaped?

■ Compute **descriptive summary measures**

- Do the mean, median and mode have similar values?
- Is the interquartile range approximately 1.33σ ?
- Is the range approximately 6σ ?



Evaluating Normality

(continued)

Comparing data characteristics to theoretical properties

- **Observe the distribution** of the data set
 - Do approximately 2/3 of the observations lie within mean ± 1 standard deviation?
 - Do approximately 80% of the observations lie within mean ± 1.28 standard deviations?
 - Do approximately 95% of the observations lie within mean ± 2 standard deviations?
- **Evaluate normal probability plot**
 - Is the normal probability plot approximately linear (i.e. a straight line) with positive slope?

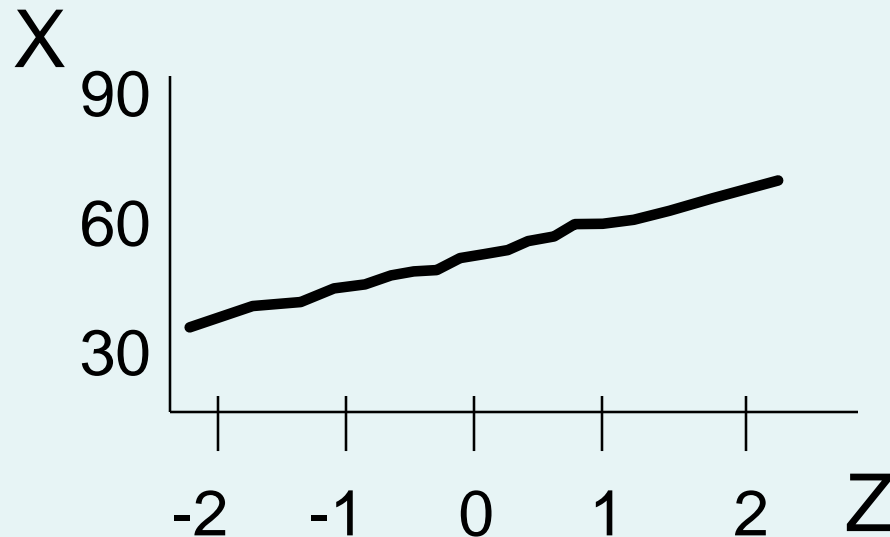


Constructing A Normal Probability Plot

- Normal probability plot
 - Arrange data into ordered array
 - Find corresponding standardized normal quantile values (Z)
 - Plot the pairs of points with observed data values (X) on the vertical axis and the standardized normal quantile values (Z) on the horizontal axis
 - Evaluate the plot for evidence of linearity

The Normal Probability Plot Interpretation

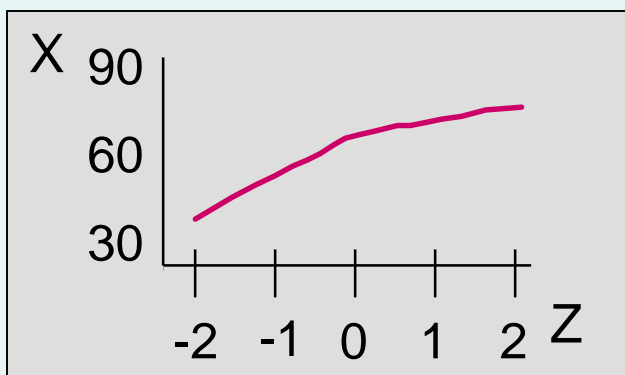
A normal probability plot for data from a normal distribution will be **approximately linear**:



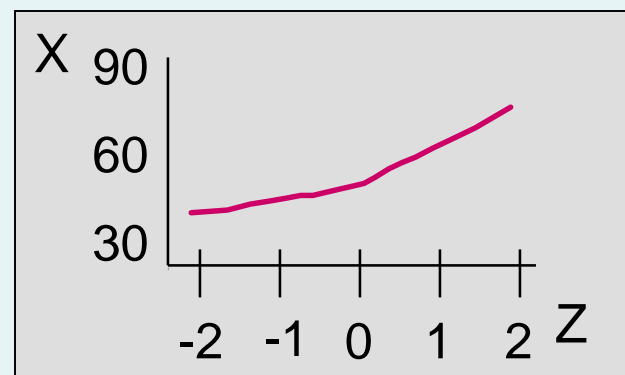
Normal Probability Plot Interpretation

(continued)

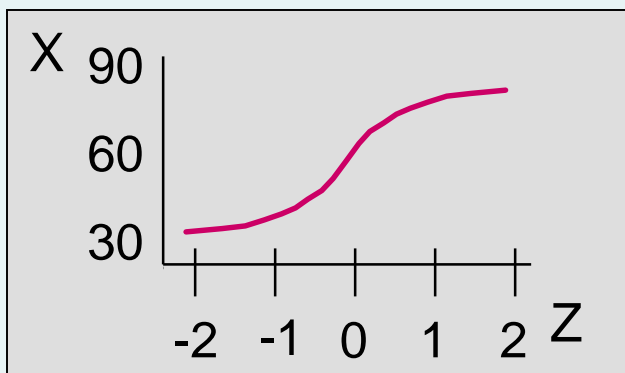
Left-Skewed



Right-Skewed



Rectangular



Nonlinear plots indicate a deviation from normality

Evaluating Normality

An Example: Mutual Funds Returns



The boxplot appears reasonably symmetric, with four lower outliers at -9.0, -8.0, -8.0, -6.5 and one upper outlier at 35.0. (The normal distribution is symmetric.)

Evaluating Normality

An Example: Mutual Funds Returns

(continued)

Descriptive Statistics

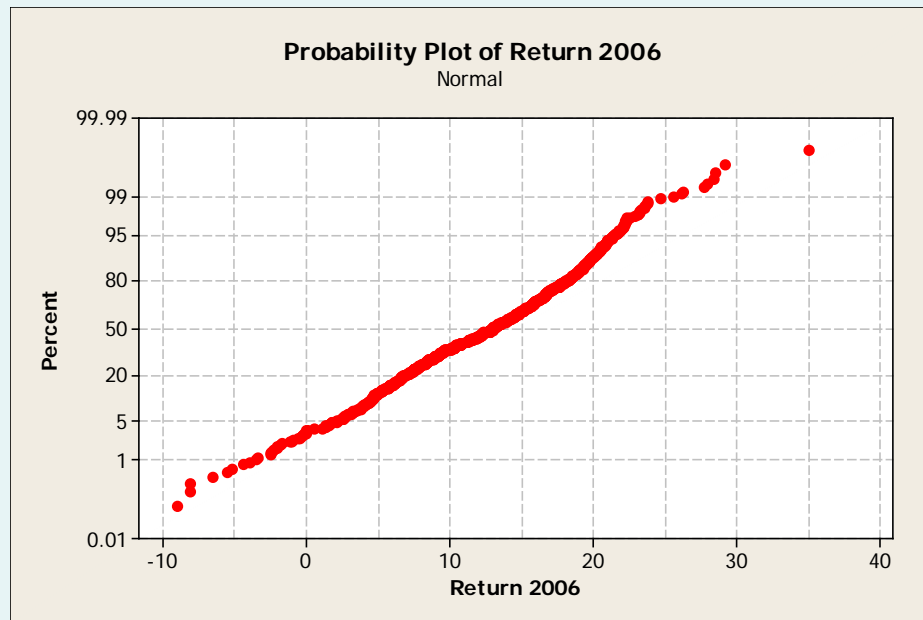
| | A | B |
|----|--------------------|------------|
| 1 | <i>Return 2006</i> | |
| 2 | | |
| 3 | Mean | 12.5142 |
| 4 | Standard Error | 0.2136 |
| 5 | Median | 13.1000 |
| 6 | Mode | 16.6000 |
| 7 | Standard Deviation | 6.2916 |
| 8 | Sample Variance | 39.5840 |
| 9 | Kurtosis | 0.0200 |
| 10 | Skewness | -0.2982 |
| 11 | Range | 44.0000 |
| 12 | Minimum | -9.0000 |
| 13 | Maximum | 35.0000 |
| 14 | Sum | 10862.3000 |
| 15 | Count | 868.0000 |
| 16 | Largest(1) | 35.0000 |
| 17 | Smallest(1) | -9.0000 |

- The mean (12.5142) is slightly less than the median (13.1). (In a normal distribution the mean and median are equal.)
- The interquartile range of 9.2 is approximately 1.46 standard deviations. (In a normal distribution the interquartile range is 1.33 standard deviations.)
- The range of 44 is equal to 6.99 standard deviations. (In a normal distribution the range is 6 standard deviations.)
- 72.2% of the observations are within 1 standard deviation of the mean. (In a normal distribution this percentage is 68.26%.)
- 87% of the observations are within 1.28 standard deviations of the mean. (In a normal distribution percentage is 80%.)

Evaluating Normality

An Example: Mutual Funds Returns

(continued)



Plot is approximately a straight line except for a few outliers at the low end and the high end.



Evaluating Normality

An Example: Mutual Funds Returns

(continued)

■ Conclusions

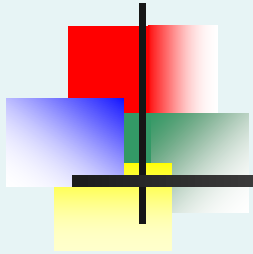
- The returns are slightly left-skewed
- The returns have more values concentrated around the mean than expected
- The range is larger than expected (caused by one outlier at 35.0)
- Normal probability plot is reasonably straight line
- Overall, this data set does not greatly differ from the theoretical properties of the normal distribution



Chapter Summary

- Presented normal distribution
- Found probabilities for the normal distribution
- Applied normal distribution to problems

Business Statistics: A First Course 5th Edition



Chapter 7

Sampling and Sampling Distributions



Learning Objectives

In this chapter, you learn:

- To distinguish between different sampling methods
- The concept of the sampling distribution
- To compute probabilities related to the sample mean and the sample proportion
- The importance of the Central Limit Theorem



Why Sample?

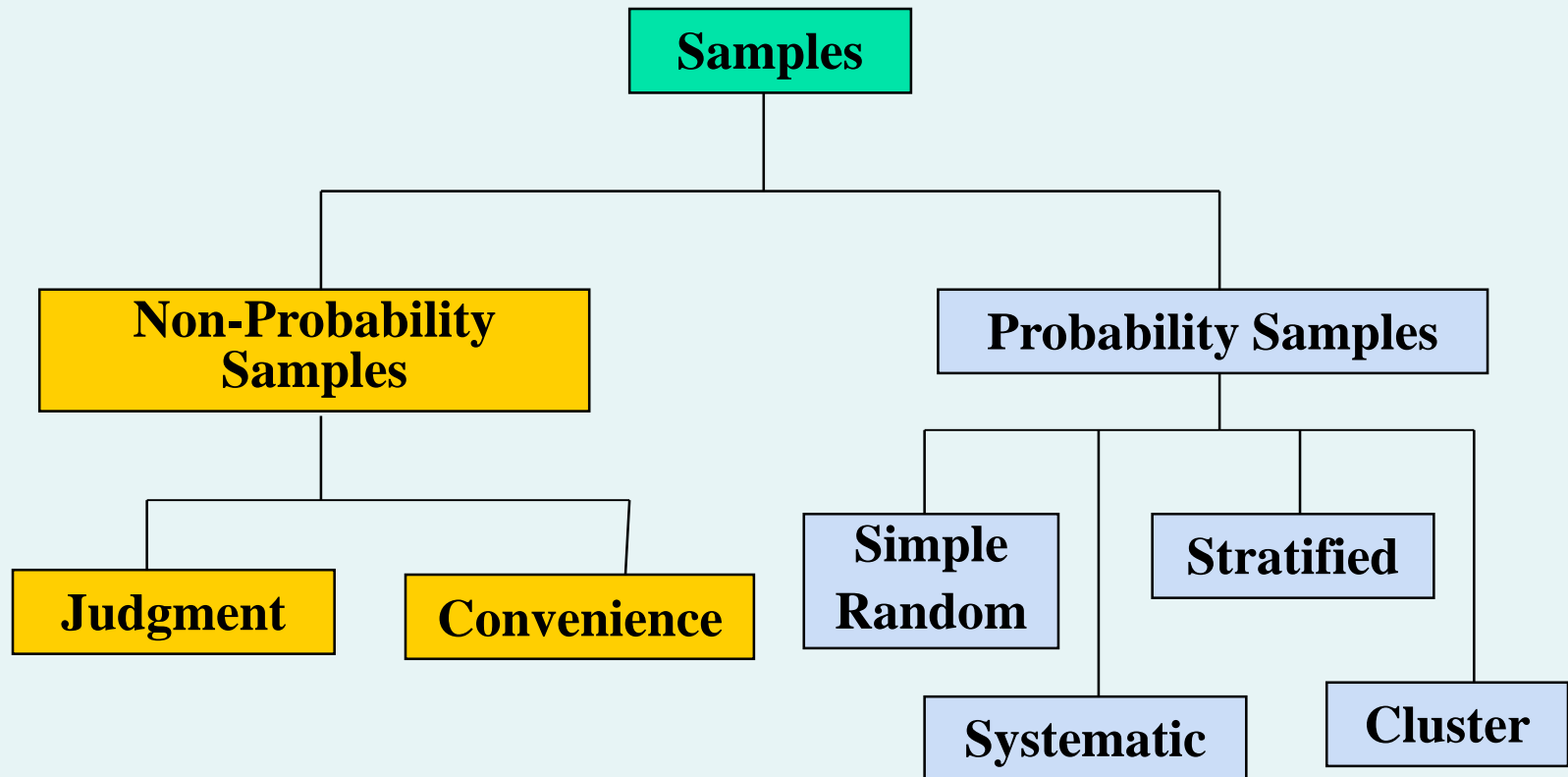
- Selecting a sample is less time-consuming than selecting every item in the population (census).
- Selecting a sample is less costly than selecting every item in the population.
- An analysis of a sample is less cumbersome and more practical than an analysis of the entire population.

A Sampling Process Begins With A Sampling Frame



- The sampling frame is a listing of items that make up the population
- Frames are data sources such as population lists, directories, or maps
- Inaccurate or biased results can result if a frame excludes certain portions of the population
- Using different frames to generate data can lead to dissimilar conclusions

Types of Samples



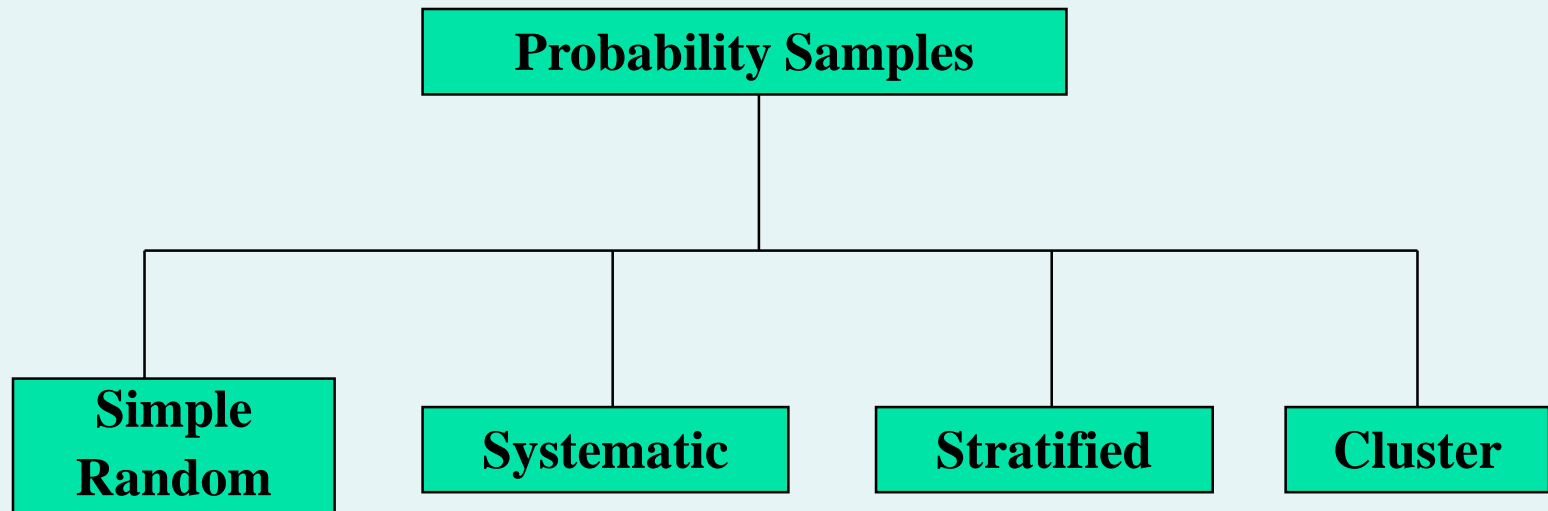


Types of Samples: Nonprobability Sample

- In a nonprobability sample, items included are chosen without regard to their probability of occurrence.
 - In **convenience sampling**, items are selected based only on the fact that they are easy, inexpensive, or convenient to sample.
 - In a **judgment sample**, you get the opinions of pre-selected experts in the subject matter.

Types of Samples: Probability Sample

- In a **probability sample**, items in the sample are chosen on the basis of known probabilities.





Probability Sample: Simple Random Sample

- Every individual or item from the frame has an equal chance of being selected
- Selection may be with replacement (selected individual is returned to frame for possible reselection) or without replacement (selected individual isn't returned to the frame).
- Samples obtained from table of random numbers or computer random number generators.

Selecting a Simple Random Sample Using A Random Number Table

Sampling Frame For Population With 850 Items

| <u>Item Name</u> | <u>Item #</u> |
|------------------|---------------|
| Bev R. | 001 |
| Ulan X. | 002 |
| . | . |
| . | . |
| . | . |
| . | . |
| Joann P. | 849 |
| Paul F. | 850 |

Portion Of A Random Number Table

49280 88924 35779 00283 81163 07275
11100 02340 12860 74697 96644 89439
09893 23997 20048 49420 88872 08401

The First 5 Items in a simple random sample

Item # 492

Item # 808

Item # 892 -- does not exist so ignore

Item # 435

Item # 779

Item # 002

Probability Sample: Systematic Sample

- 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

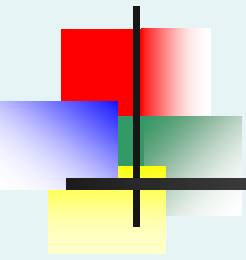
$$N = 40$$

$$n = 4$$

$$k = 10$$

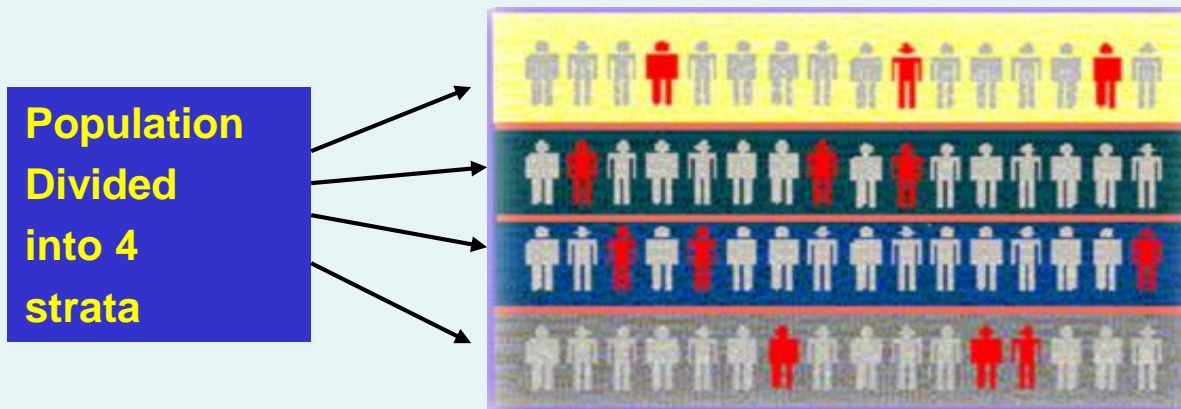
First Group





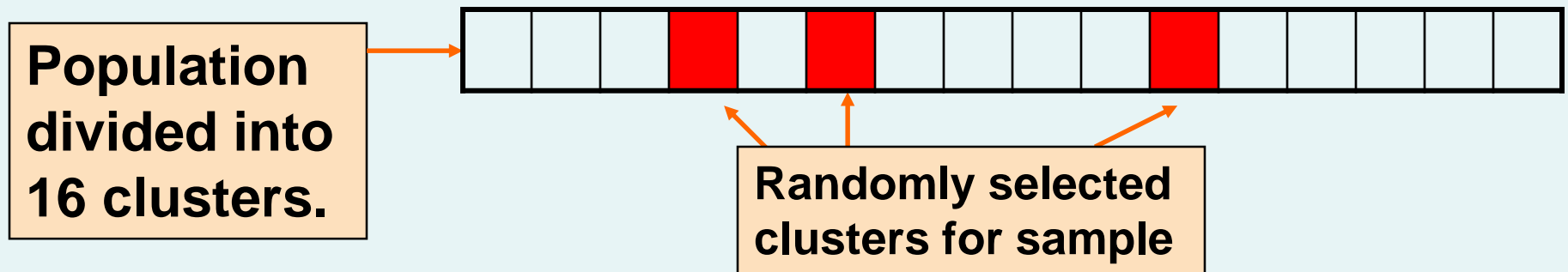
Probability Sample: Stratified Sample

- Divide population into two or more subgroups (called *strata*) according to some common characteristic
- A simple random sample is selected from each subgroup, with sample sizes proportional to strata sizes
- Samples from subgroups are combined into one
- This is a common technique when sampling population of voters, stratifying across racial or socio-economic lines.



Probability Sample Cluster Sample

- Population is divided into several “clusters,” each representative of the population
- A simple random sample of clusters is selected
- All items in the selected clusters can be used, or items can be chosen from a cluster using another probability sampling technique
- A common application of cluster sampling involves election exit polls, where certain election districts are selected and sampled.





Probability Sample: Comparing Sampling Methods

- Simple random sample and Systematic sample
 - Simple to use
 - May not be a good representation of the population's underlying characteristics
- Stratified sample
 - Ensures representation of individuals across the entire population
- Cluster sample
 - More cost effective
 - Less efficient (need larger sample to acquire the same level of precision)



Evaluating Survey Worthiness

- What is the purpose of the survey?
- Is the survey based on a probability sample?
- Coverage error – appropriate frame?
- Nonresponse error – follow up
- Measurement error – good questions elicit good responses
- Sampling error – always exists



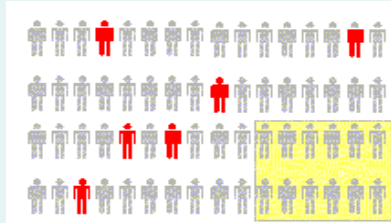
Types of Survey Errors

- Coverage error or selection bias
 - Exists if some groups are excluded from the frame and have no chance of being selected
- Non response error or bias
 - People who do not respond may be different from those who do respond
- Sampling error
 - Variation from sample to sample will always exist
- Measurement error
 - Due to weaknesses in question design, respondent error, and interviewer's effects on the respondent (“Hawthorne effect”)

Types of Survey Errors

(continued)

- Coverage error



Excluded from frame

- Non response error



Follow up on nonresponses

- Sampling error



Random differences from sample to sample

- Measurement error



Bad or leading question



Sampling Distributions

- A sampling distribution is a distribution of all of the possible values of a sample statistic for a given size sample selected from a population.
- For example, suppose you sample 50 students from your college regarding their mean GPA. If you obtained many different samples of 50, you will compute a different mean for each sample. We are interested in the distribution of all potential mean GPA we might calculate for any given sample of 50 students.

Developing a Sampling Distribution

- Assume there is a population ...
- Population size $N=4$
- Random variable, X , is **age** of individuals
- Values of X : 18, 20, 22, 24 (years)



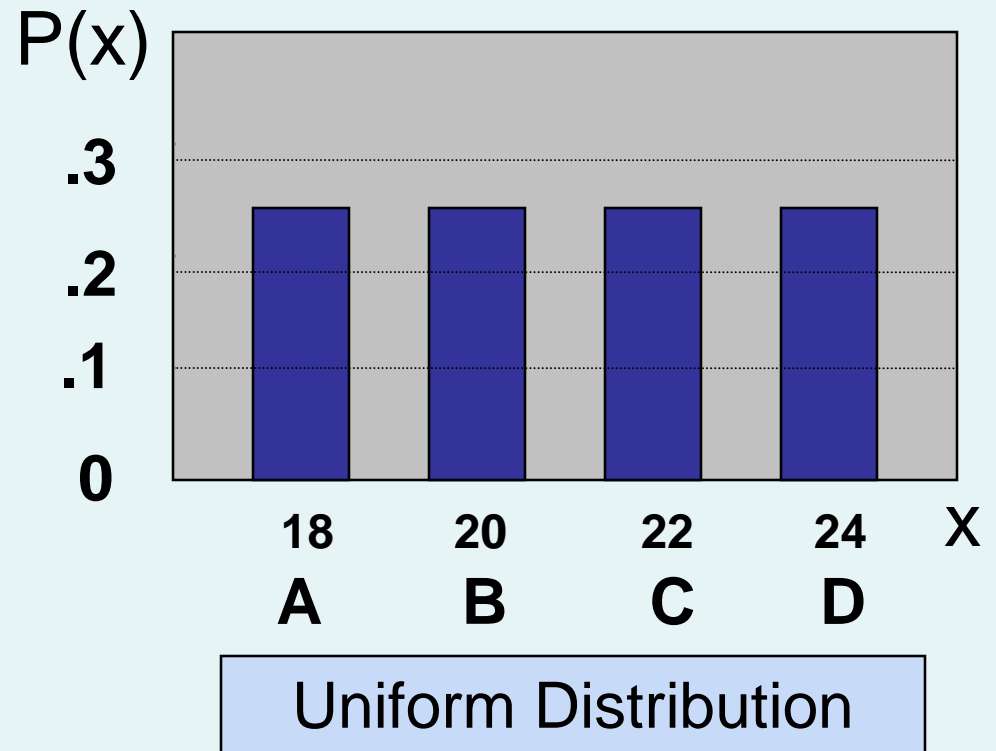
Developing a Sampling Distribution

(continued)

Summary Measures for the Population Distribution:

$$\begin{aligned}\mu &= \frac{\sum X_i}{N} \\ &= \frac{18 + 20 + 22 + 24}{4} = 21\end{aligned}$$

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} = 2.236$$



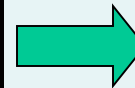
Developing a Sampling Distribution

(continued)

Now consider all possible samples of size $n=2$

| 1 st Obs | 2 nd Observation | | | |
|---------------------|-----------------------------|-------|-------|-------|
| | 18 | 20 | 22 | 24 |
| 18 | 18,18 | 18,20 | 18,22 | 18,24 |
| 20 | 20,18 | 20,20 | 20,22 | 20,24 |
| 22 | 22,18 | 22,20 | 22,22 | 22,24 |
| 24 | 24,18 | 24,20 | 24,22 | 24,24 |

16 possible samples
(sampling with replacement)



| 1 st Obs | 2 nd Observation | | | |
|---------------------|-----------------------------|----|----|----|
| | 18 | 20 | 22 | 24 |
| 18 | 18 | 19 | 20 | 21 |
| 20 | 19 | 20 | 21 | 22 |
| 22 | 20 | 21 | 22 | 23 |
| 24 | 21 | 22 | 23 | 24 |

16 Sample Means

Developing a Sampling Distribution

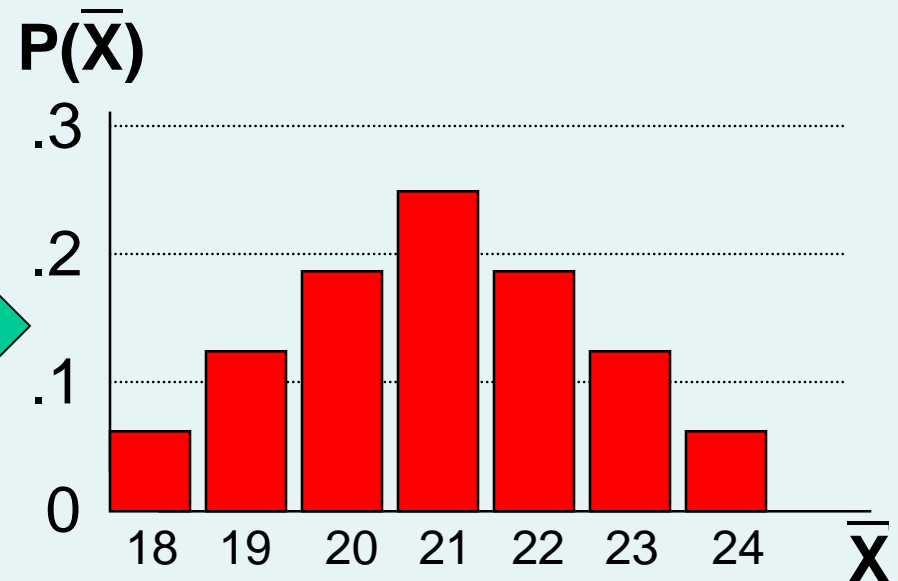
(continued)

Sampling Distribution of All Sample Means

16 Sample Means

| 1st Obs | 2nd Observation | | | |
|------------|-----------------|----|----|----|
| | 18 | 20 | 22 | 24 |
| 18 | 18 | 19 | 20 | 21 |
| 20 | 19 | 20 | 21 | 22 |
| 22 | 20 | 21 | 22 | 23 |
| 24 | 21 | 22 | 23 | 24 |

Sample Means
Distribution



(no longer uniform)



Developing a Sampling Distribution

(continued)

Summary Measures of this Sampling Distribution:

$$\mu_{\bar{X}} = \frac{\sum \bar{X}_i}{N} = \frac{18 + 19 + 19 + \dots + 24}{16} = 21$$

$$\begin{aligned}\sigma_{\bar{X}} &= \sqrt{\frac{\sum (\bar{X}_i - \mu_{\bar{X}})^2}{N}} \\ &= \sqrt{\frac{(18 - 21)^2 + (19 - 21)^2 + \dots + (24 - 21)^2}{16}} = 1.58\end{aligned}$$

Comparing the Population Distribution to the Sample Means Distribution

Population

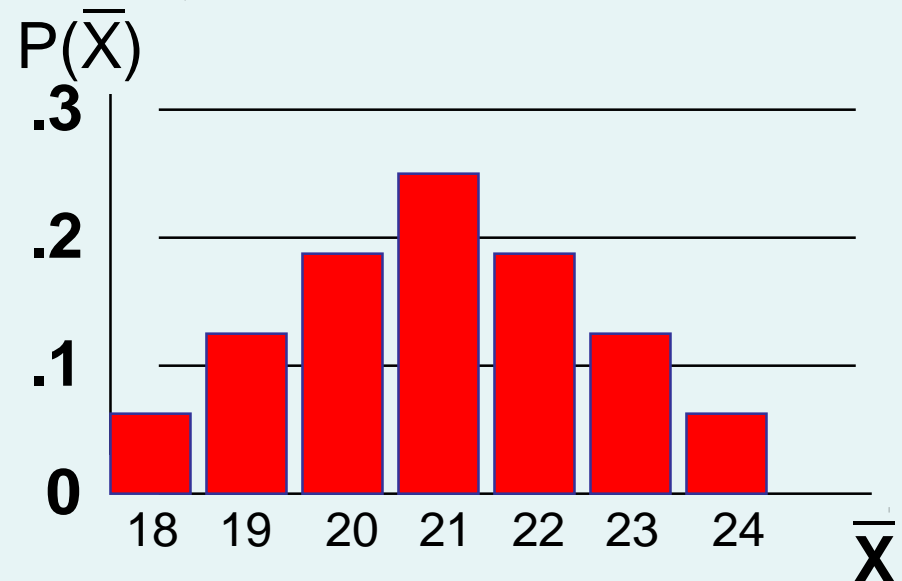
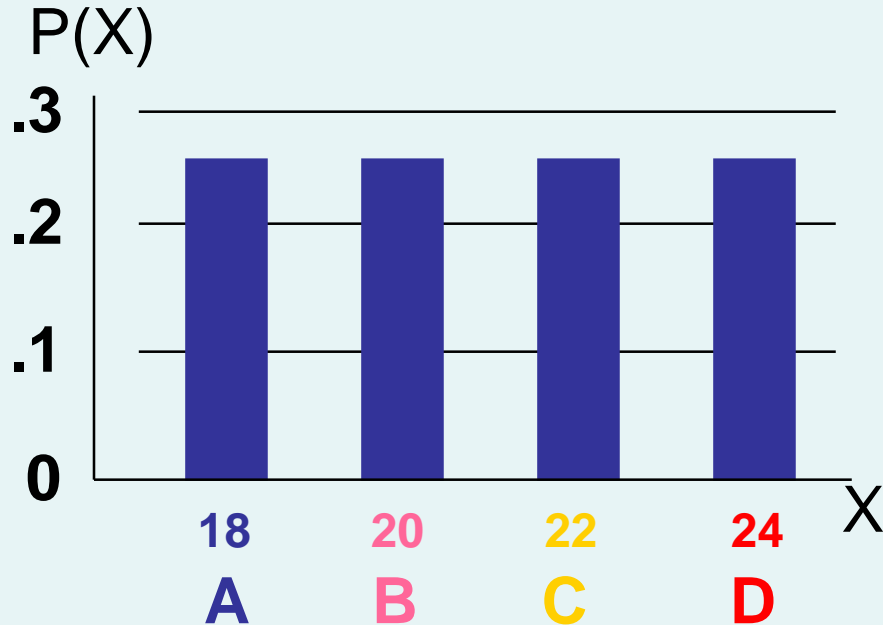
$$N = 4$$

$$\mu = 21 \quad \sigma = 2.236$$

Sample Means Distribution

$$n = 2$$

$$\mu_{\bar{X}} = 21 \quad \sigma_{\bar{X}} = 1.58$$



Sample Mean Sampling Distribution: Standard Error of the Mean

- Different samples of the same size from the same population will yield different sample means
- A measure of the variability in the mean from sample to sample is given by the **Standard Error of the Mean:**

(This assumes that sampling is with replacement or sampling is without replacement from an infinite population)

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- Note that the standard error of the mean decreases as the sample size increases



Sample Mean Sampling Distribution: If the Population is Normal

- If a population is normally distributed with mean μ and standard deviation σ , the sampling distribution of \bar{X} is also normally distributed with

$$\mu_{\bar{X}} = \mu$$

and

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$



Z-value for Sampling Distribution of the Mean

- Z-value for the sampling distribution of \bar{X} :

$$Z = \frac{(\bar{X} - \mu_{\bar{X}})}{\sigma_{\bar{X}}} = \frac{(\bar{X} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

where:

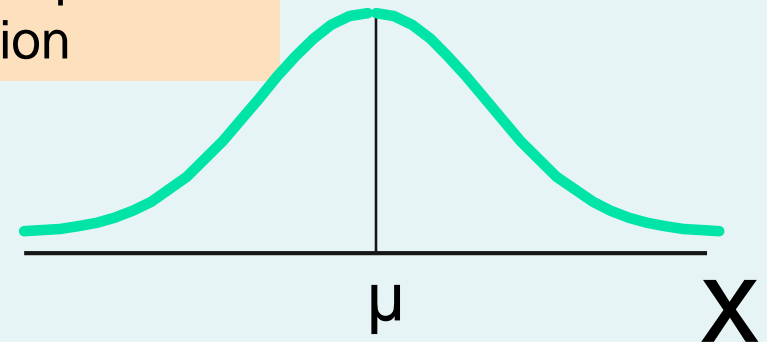
- \bar{X} = sample mean
- μ = population mean
- σ = population standard deviation
- n = sample size

Sampling Distribution Properties

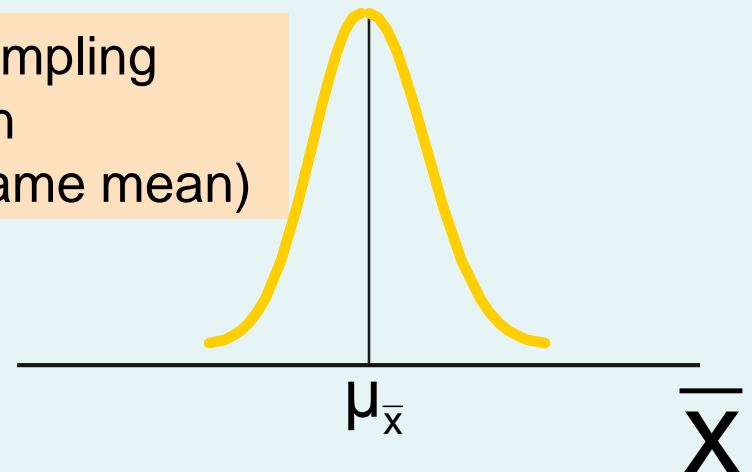
$$\mu_{\bar{X}} = \mu$$

(i.e. \bar{X} is unbiased)

Normal Population Distribution

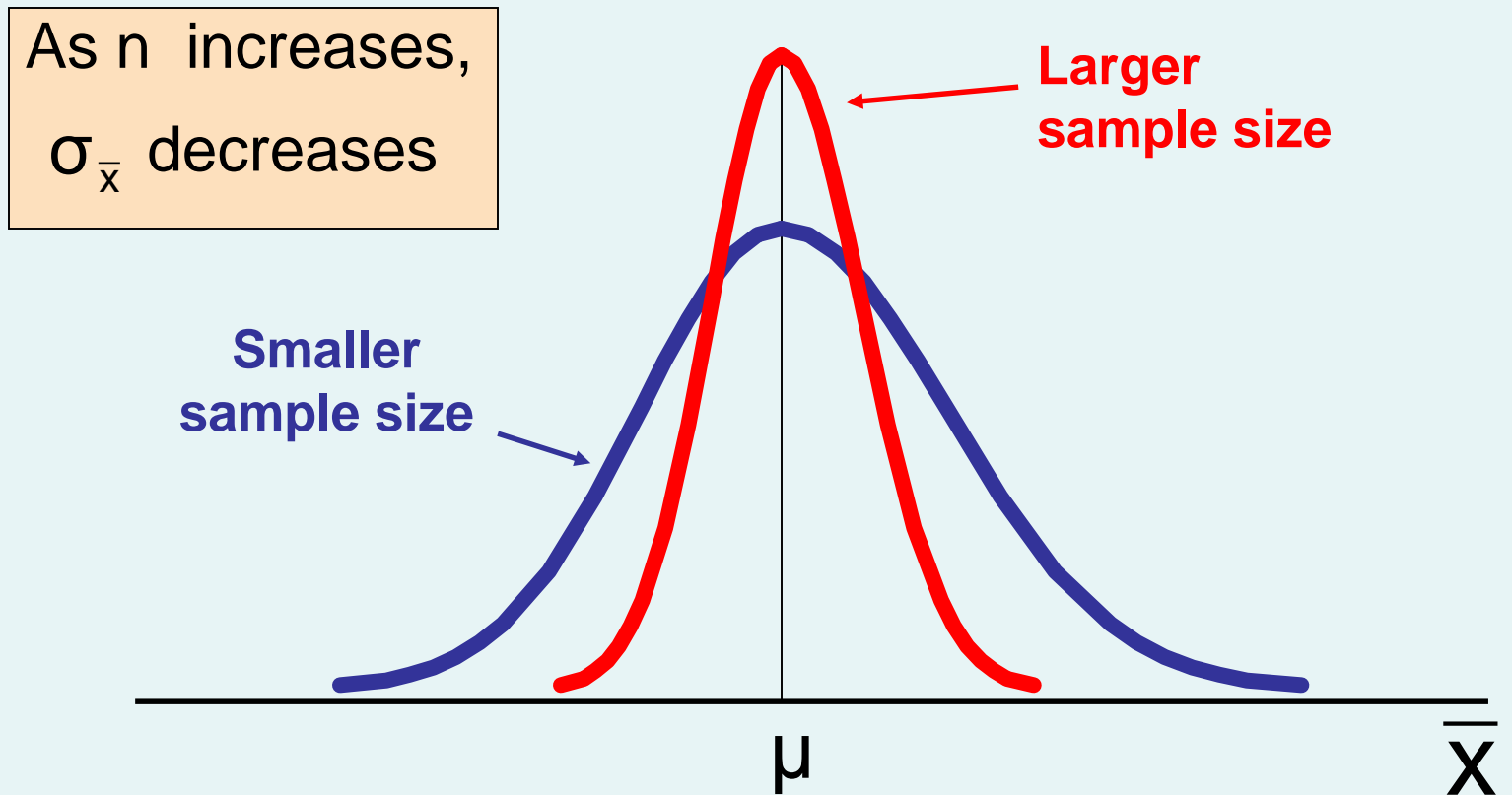


Normal Sampling Distribution
(has the same mean)



Sampling Distribution Properties

(continued)





Determining An Interval Including A Fixed Proportion of the Sample Means

Find a symmetrically distributed interval around μ that will include 95% of the sample means when $\mu = 368$, $\sigma = 15$, and $n = 25$.

- Since the interval contains 95% of the sample means 5% of the sample means will be outside the interval
- Since the interval is symmetric 2.5% will be above the upper limit and 2.5% will be below the lower limit.
- From the standardized normal table, the Z score with 2.5% (0.0250) below it is -1.96 and the Z score with 2.5% (0.0250) above it is 1.96.

Determining An Interval Including A Fixed Proportion of the Sample Means

(continued)

- Calculating the lower limit of the interval

$$\bar{X}_L = \mu + Z \frac{\sigma}{\sqrt{n}} = 368 + (-1.96) \frac{15}{\sqrt{25}} = 362.12$$

- Calculating the upper limit of the interval

$$\bar{X}_U = \mu + Z \frac{\sigma}{\sqrt{n}} = 368 + (1.96) \frac{15}{\sqrt{25}} = 373.88$$

- 95% of all sample means of sample size 25 are between 362.12 and 373.88

Sample Mean Sampling Distribution: If the Population is **not** Normal

- We can apply the **Central Limit Theorem**:
 - Even if the population is **not normal**,
 - ...sample means from the population **will be approximately normal** as long as the sample size is large enough.

Properties of the sampling distribution:

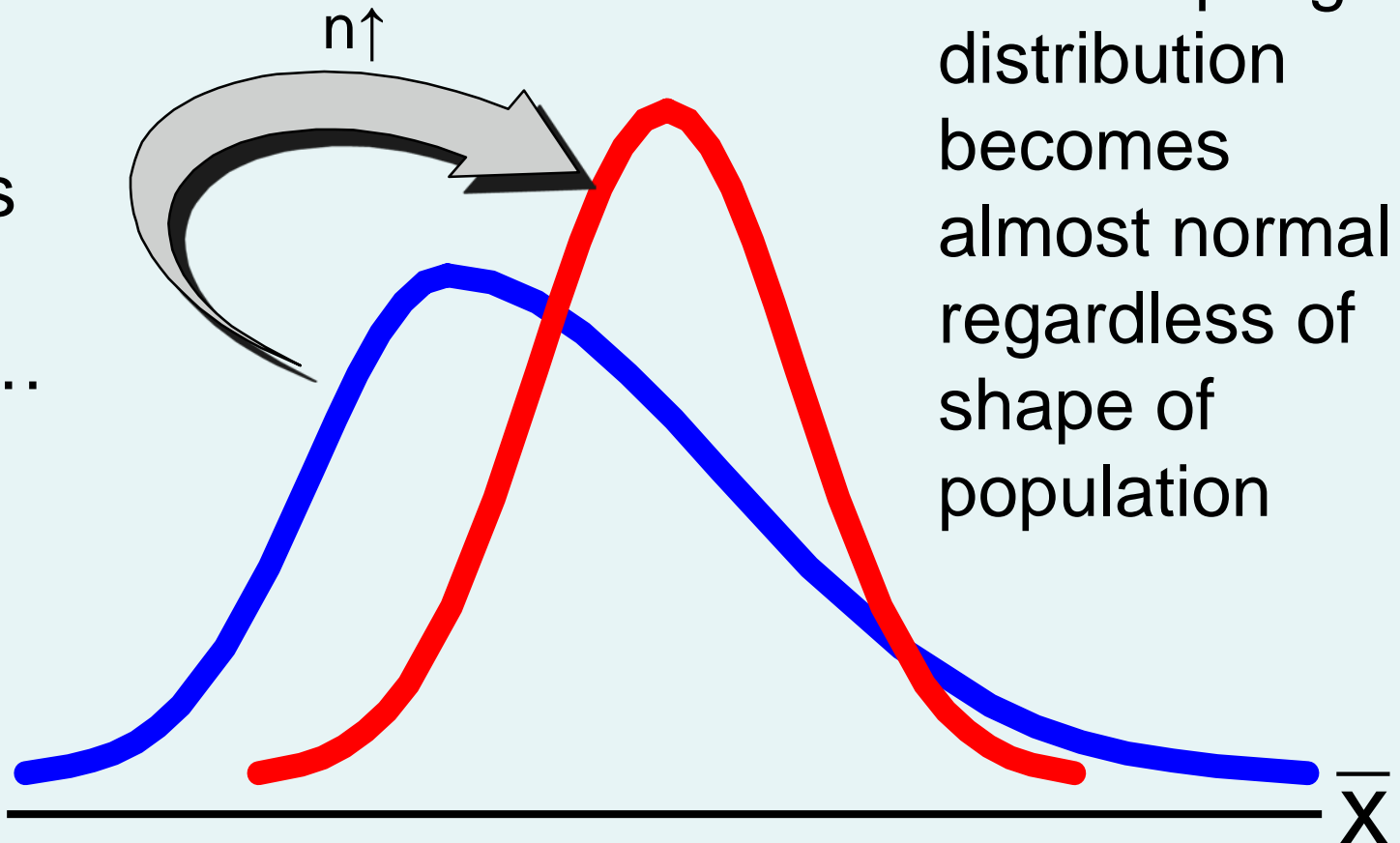
$$\mu_{\bar{x}} = \mu$$

and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem

As the sample size gets large enough...



the sampling distribution becomes almost normal regardless of shape of population

Sample Mean Sampling Distribution: If the Population is **not** Normal

(continued)

Sampling distribution
properties:

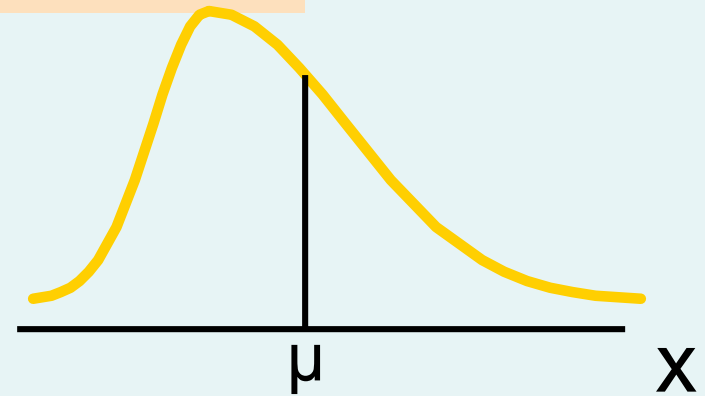
Central Tendency

$$\mu_{\bar{x}} = \mu$$

Variation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

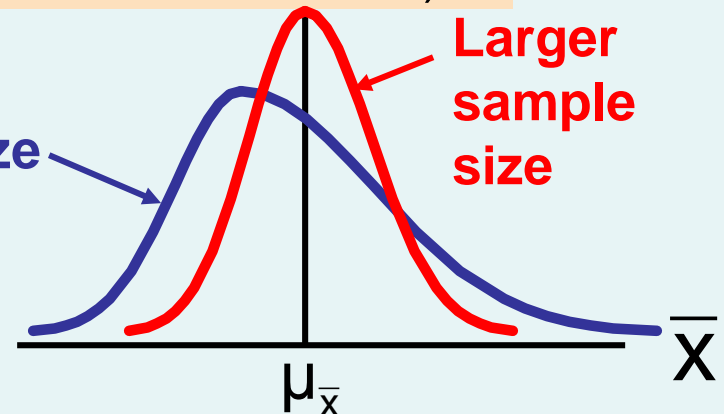
Population Distribution



Sampling Distribution
(becomes normal as n increases)

Smaller
sample size

Larger
sample size





How Large is Large Enough?

- For most distributions, $n > 30$ will give a sampling distribution that is nearly normal
- For fairly symmetric distributions, $n > 15$ will usually give a sampling distribution is almost normal
- For normal population distributions, the sampling distribution of the mean is always normally distributed



Example

- Suppose a population has mean $\mu = 8$ and standard deviation $\sigma = 3$. Suppose a random sample of size $n = 36$ is selected.
- What is the probability that the **sample mean** is between 7.8 and 8.2?



Example

(continued)

Solution:

- Even if the population is not normally distributed, the central limit theorem can be used ($n > 30$)
- ... so the sampling distribution of \bar{X} is approximately normal
- ... with mean $\mu_{\bar{x}} = 8$
- ... and standard deviation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$$

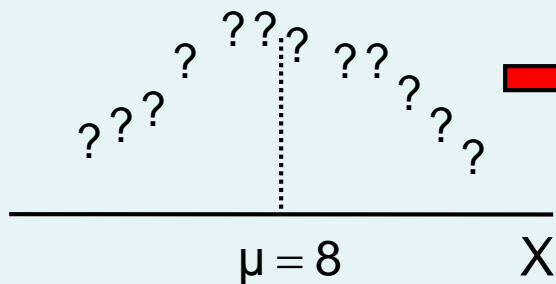
Example

(continued)

Solution (continued):

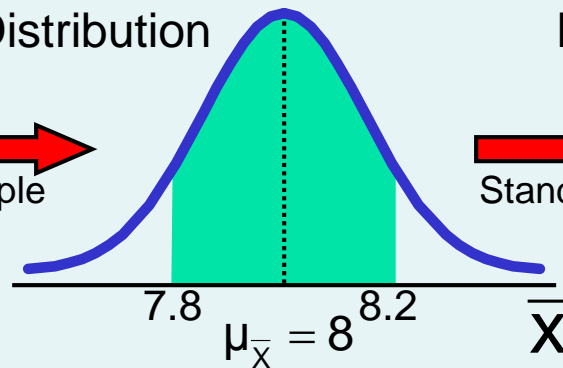
$$\begin{aligned} P(7.8 < \bar{X} < 8.2) &= P\left(\frac{7.8 - 8}{\frac{3}{\sqrt{36}}} < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{8.2 - 8}{\frac{3}{\sqrt{36}}}\right) \\ &= P(-0.4 < Z < 0.4) = \boxed{0.3108} \end{aligned}$$

Population
Distribution



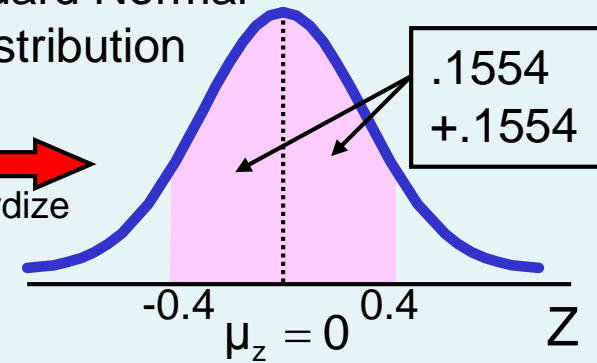
Sampling
Distribution

Sample



Standard Normal
Distribution

Standardize





Population Proportions

π = the proportion of the population having some characteristic

- Sample proportion (p) provides an estimate of π :

$$p = \frac{X}{n} = \frac{\text{number of items in the sample having the characteristic of interest}}{\text{sample size}}$$

- $0 \leq p \leq 1$
- p is approximately distributed as a normal distribution when n is large

(assuming sampling with replacement from a finite population or without replacement from an infinite population)

Sampling Distribution of p

- Approximated by a normal distribution if:

- $n\pi \geq 5$
and
 $n(1 - \pi) \geq 5$

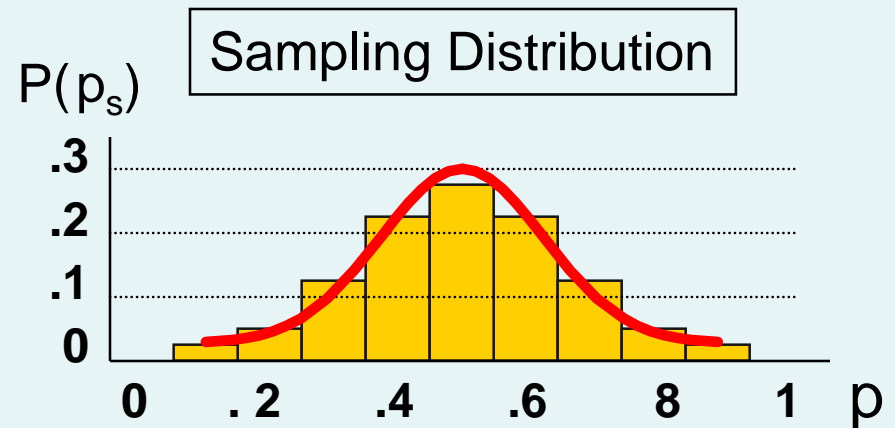
where

$$\mu_p = \pi$$

and

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

(where π = population proportion)





Z-Value for Proportions

Standardize p to a Z value with the formula:

$$Z = \frac{p - \pi}{\sigma_p} = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}}$$



Example

- If the true proportion of voters who support Proposition A is $\pi = 0.4$, what is the probability that a sample of size 200 yields a sample proportion between 0.40 and 0.45?
- i.e.: **if $\pi = 0.4$ and $n = 200$, what is $P(0.40 \leq p \leq 0.45)$?**



Example

(continued)

- if $\pi = 0.4$ and $n = 200$, what is $P(0.40 \leq p \leq 0.45)$?

Find σ_p :

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.4(1-0.4)}{200}} = 0.03464$$

Convert to
standardized
normal:

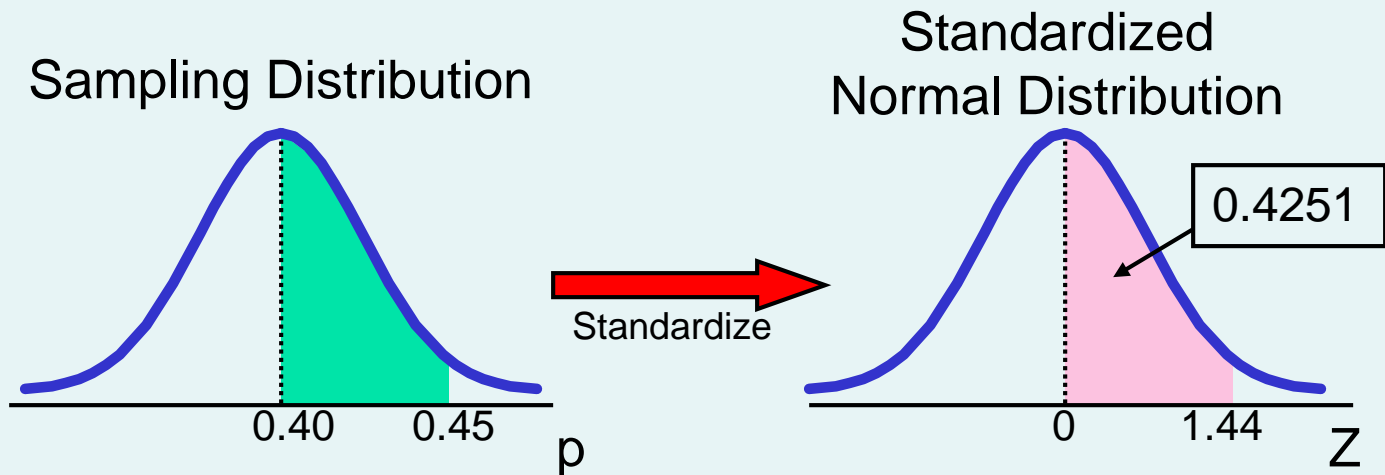
$$\begin{aligned} P(0.40 \leq p \leq 0.45) &= P\left(\frac{0.40 - 0.40}{0.03464} \leq Z \leq \frac{0.45 - 0.40}{0.03464}\right) \\ &= P(0 \leq Z \leq 1.44) \end{aligned}$$

Example

(continued)

- if $\pi = 0.4$ and $n = 200$, what is $P(0.40 \leq p \leq 0.45)$?

Use standardized normal table: $P(0 \leq Z \leq 1.44) = 0.4251$

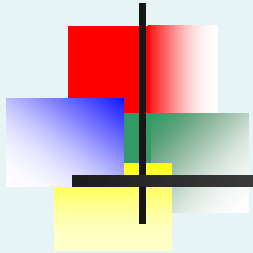




Chapter Summary

- Discussed probability and nonprobability samples
- Described four common probability samples
- Examined survey worthiness and types of survey errors
- Introduced sampling distributions
- Described the sampling distribution of the mean
 - For normal populations
 - Using the Central Limit Theorem
- Described the sampling distribution of a proportion
- Calculated probabilities using sampling distributions

Business Statistics: A First Course 5th Edition



Chapter 8

Confidence Interval Estimation



Learning Objectives

In this chapter, you learn:

- To construct and interpret confidence interval estimates for the mean and the proportion
- How to determine the sample size necessary to develop a confidence interval for the mean or proportion



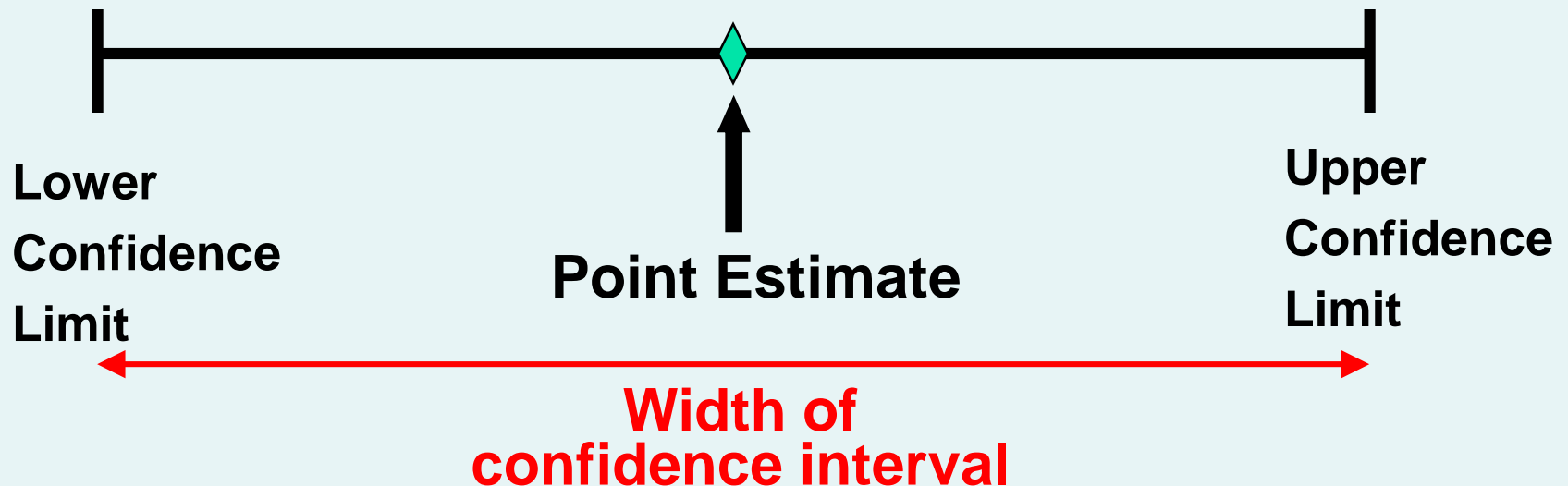
Chapter Outline

Content of this chapter

- Confidence Intervals for the **Population Mean, μ**
 - when Population Standard Deviation σ is **Known**
 - when Population Standard Deviation σ is **Unknown**
- Confidence Intervals for the **Population Proportion, π**
- Determining the **Required Sample Size**

Point and Interval Estimates

- A **point estimate** is a single number
- a **confidence interval** provides additional information about the variability of the estimate





Point Estimates

| We can estimate a Population Parameter ... | | with a Sample Statistic (a Point Estimate) |
|--------------------------------------------|-------|-----------------------------------------------|
| Mean | μ | \bar{X} |
| Proportion | π | p |



Confidence Intervals

- How much uncertainty is associated with a point estimate of a population parameter?
- An **interval estimate** provides more information about a population characteristic than does a **point estimate**
- Such interval estimates are called **confidence intervals**



Confidence Interval Estimate

- An interval gives a **range** of values:
 - Takes into consideration variation in sample statistics from sample to sample
 - Based on observations from 1 sample
 - Gives information about closeness to unknown population parameters
 - Stated in terms of level of confidence
 - e.g. 95% confident, 99% confident
 - Can never be 100% confident



Confidence Interval Example

Cereal fill example

- Population has $\mu = 368$ and $\sigma = 15$.
- If you take a sample of size $n = 25$ you know
 - $368 \pm 1.96 * 15 \sqrt{25} = (362.12, 373.88)$ contains 95% of the sample means
 - When you don't know μ , you use \bar{X} to estimate μ
 - If $\bar{X} = 362.3$ the interval is $362.3 \pm 1.96 * 15 \sqrt{25} = (356.42, 368.18)$
 - Since $356.42 \leq \mu \leq 368.18$, the interval based on this sample makes a correct statement about μ .

But what about the intervals from other possible samples of size 25?



Confidence Interval Example

(continued)

| Sample # | \bar{X} | Lower Limit | Upper Limit | Contain μ ? |
|----------|-----------|-------------|-------------|-----------------|
| 1 | 362.30 | 356.42 | 368.18 | Yes |
| 2 | 369.50 | 363.62 | 375.38 | Yes |
| 3 | 360.00 | 354.12 | 365.88 | No |
| 4 | 362.12 | 356.24 | 368.00 | Yes |
| 5 | 373.88 | 368.00 | 379.76 | Yes |



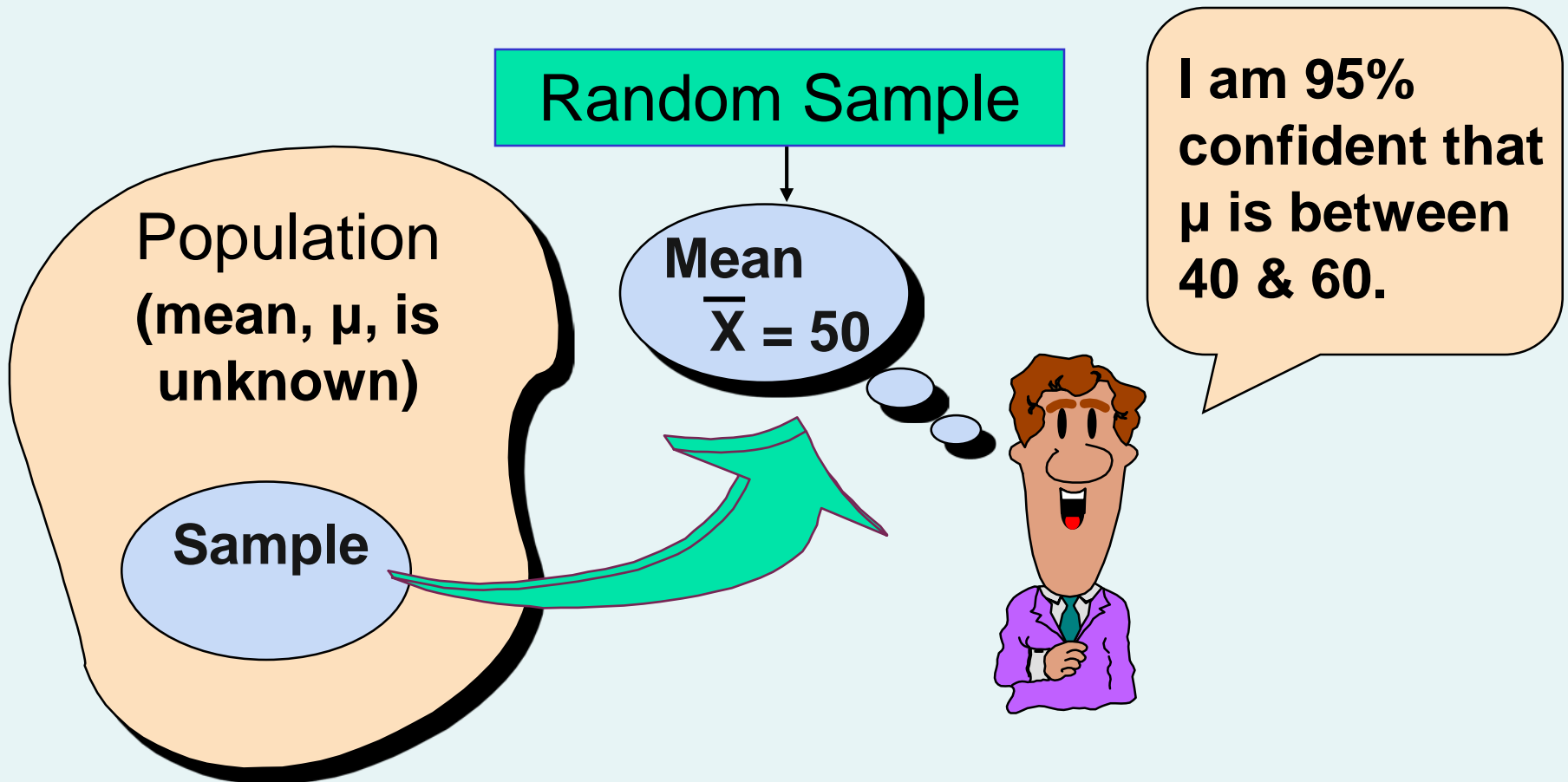
Confidence Interval Example

(continued)

- In practice you only take one sample of size n
- In practice you do not know μ so you do not know if the interval actually contains μ
- However you do know that 95% of the intervals formed in this manner will contain μ
- Thus, based on the one sample, you actually selected you can be 95% confident your interval will contain μ (this is a 95% **confidence interval**)

Note: 95% confidence is based on the fact that we used $Z = 1.96$.

Estimation Process





General Formula

- The general formula for all confidence intervals is:

$$\text{Point Estimate} \pm (\text{Critical Value})(\text{Standard Error})$$

Where:

- **Point Estimate** is the sample statistic estimating the population parameter of interest
- **Critical Value** is a table value based on the sampling distribution of the point estimate and the desired confidence level
- **Standard Error** is the standard deviation of the point estimate



Confidence Level

- Confidence Level
 - The confidence that the interval will contain the unknown population parameter
 - A percentage (less than 100%)

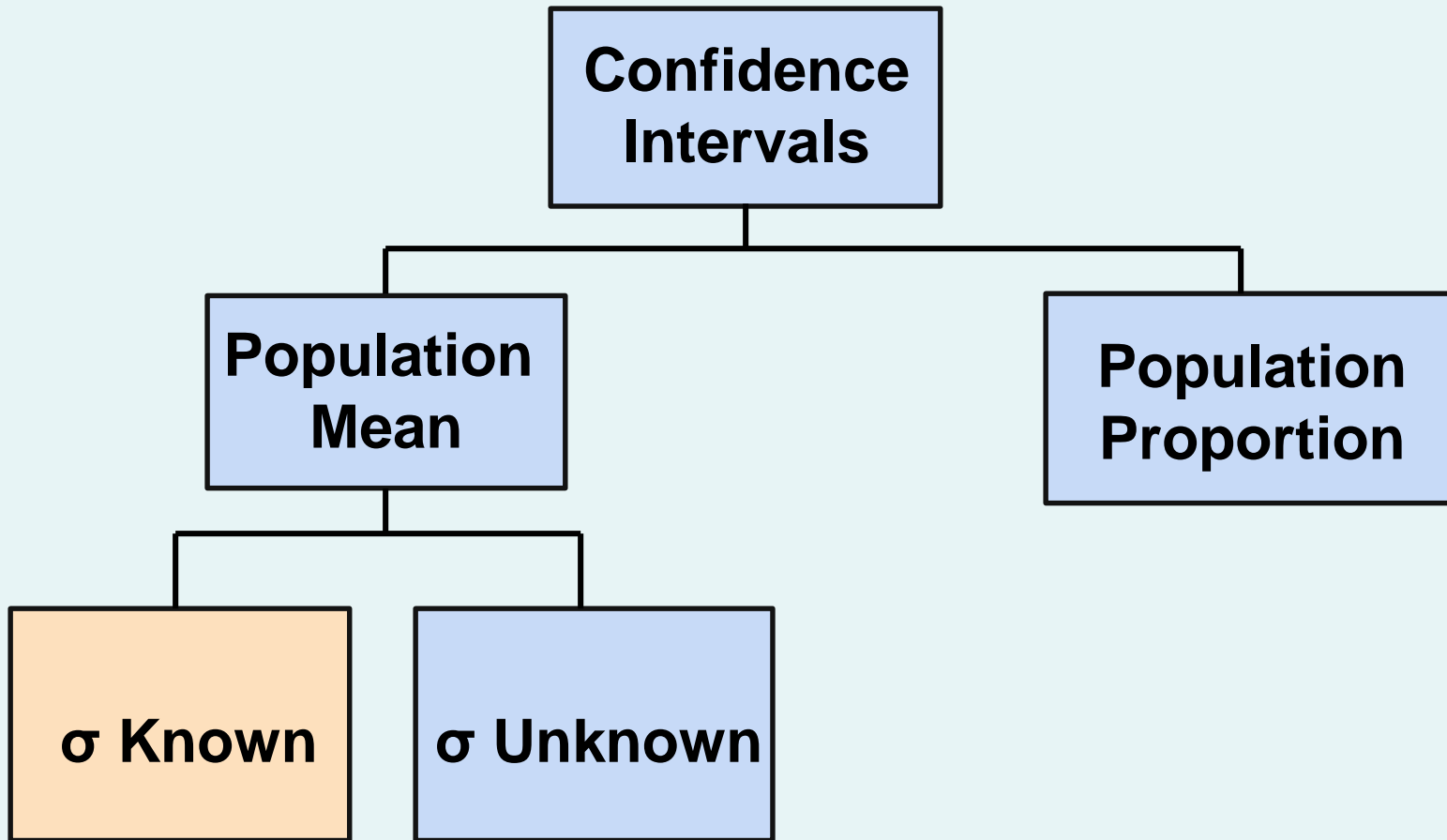


Confidence Level, $(1-\alpha)$

(continued)

- Suppose confidence level = 95%
- Also written $(1 - \alpha) = 0.95$, (so $\alpha = 0.05$)
- A relative frequency interpretation:
 - 95% of all the confidence intervals that can be constructed will contain the unknown true parameter
- A specific interval either will contain or will not contain the true parameter
 - No probability involved in a specific interval

Confidence Intervals



Confidence Interval for μ (σ Known)

- Assumptions
 - Population standard deviation σ is known
 - Population is normally distributed
 - If population is not normal, use large sample
- Confidence interval estimate:

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where \bar{X} is the point estimate

$Z_{\alpha/2}$ is the normal distribution critical value for a probability of $\alpha/2$ in each tail

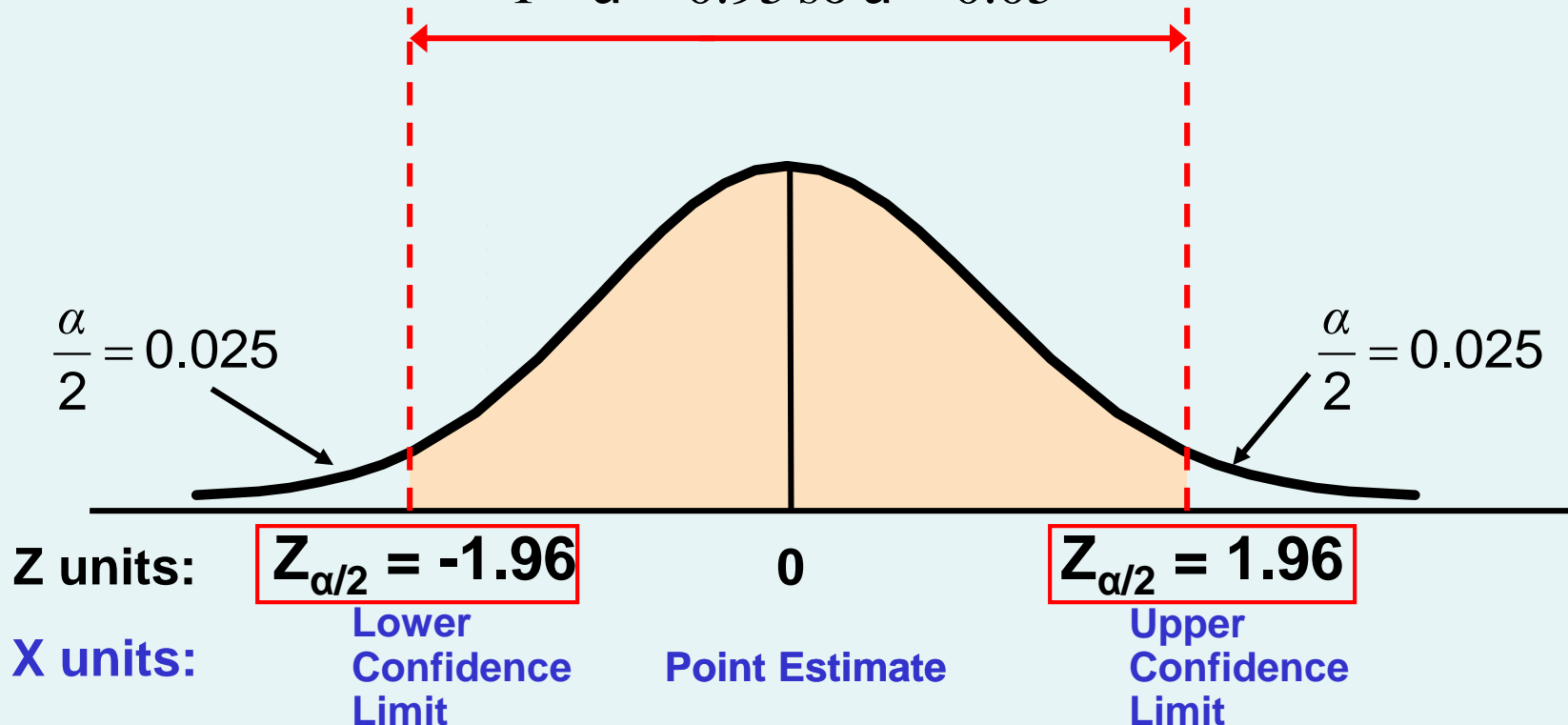
σ/\sqrt{n} is the standard error

Finding the Critical Value, $Z_{\alpha/2}$

$$Z_{\alpha/2} = \pm 1.96$$

- Consider a 95% confidence interval:

$$1 - \alpha = 0.95 \text{ so } \alpha = 0.05$$





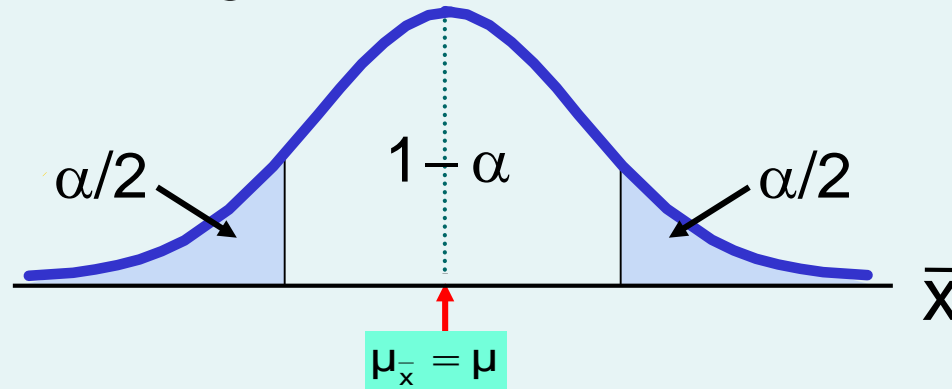
Common Levels of Confidence

- Commonly used confidence levels are 90%, 95%, and 99%

| Confidence Level | Confidence Coefficient, $1 - \alpha$ | $Z_{\alpha/2}$ value |
|-------------------------|------------------------------------------------------------|----------------------------------------|
| 80% | 0.80 | 1.28 |
| 90% | 0.90 | 1.645 |
| 95% | 0.95 | 1.96 |
| 98% | 0.98 | 2.33 |
| 99% | 0.99 | 2.58 |
| 99.8% | 0.998 | 3.08 |
| 99.9% | 0.999 | 3.27 |

Intervals and Level of Confidence

Sampling Distribution of the Mean

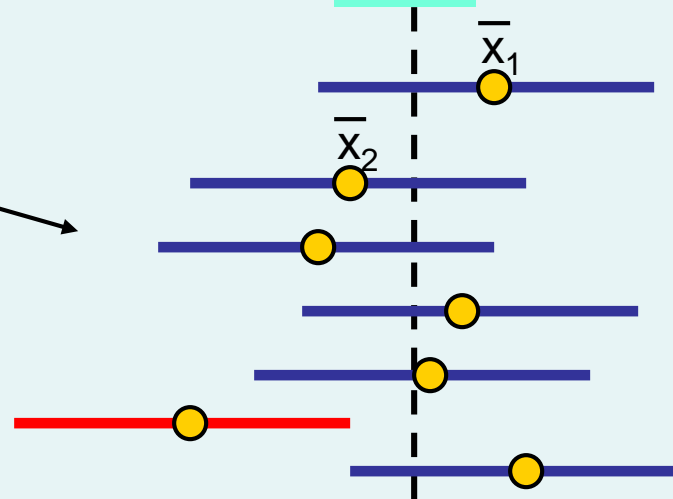


Intervals
extend from

$$\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

to

$$\bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

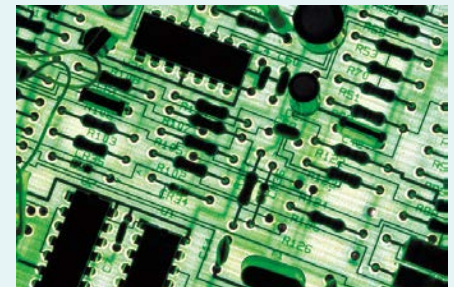


Confidence Intervals

$(1-\alpha) \times 100\%$
of intervals
constructed
contain μ ;
 $(\alpha) \times 100\%$ do
not.

Example

- A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms.
- Determine a 95% confidence interval for the true mean resistance of the population.



Example

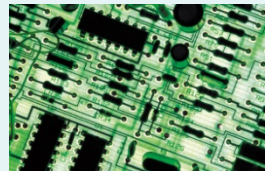
(continued)

- A sample of 11 circuits from a large normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms.

- **Solution:**

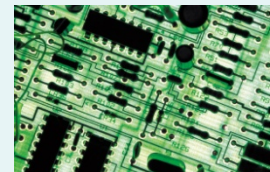
$$\begin{aligned}\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \\ &= 2.20 \pm 1.96 (0.35/\sqrt{11}) \\ &= 2.20 \pm 0.2068\end{aligned}$$

$$1.9932 \leq \mu \leq 2.4068$$

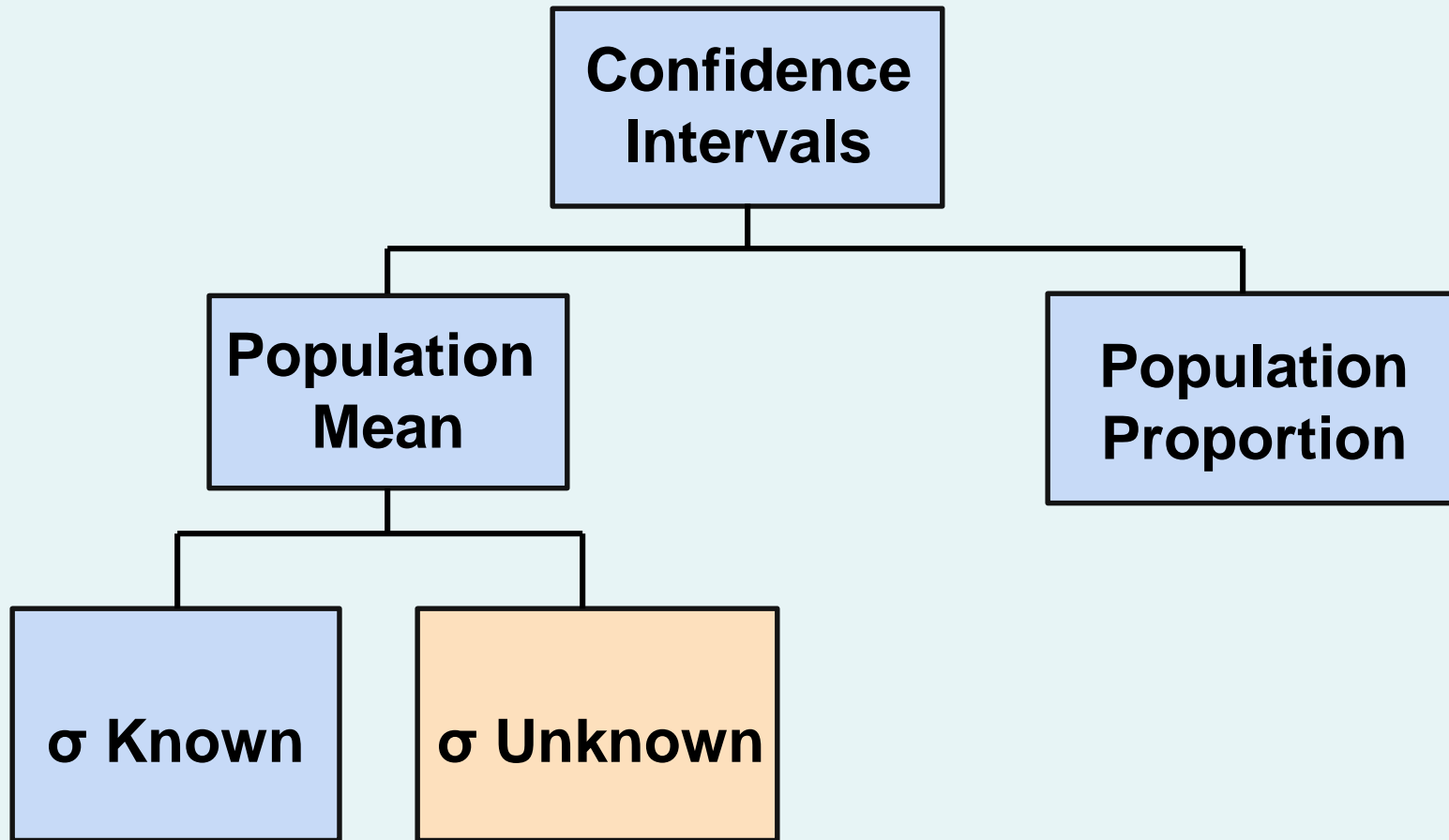


Interpretation

- We are 95% confident that the true mean resistance is between 1.9932 and 2.4068 ohms
- Although the true mean may or may not be in this interval, 95% of intervals formed in this manner will contain the true mean



Confidence Intervals





Do You Ever Truly Know σ ?

- Probably not!
- In virtually all real world business situations, σ is not known.
- If there is a situation where σ is known then μ is also known (since to calculate σ you need to know μ .)
- If you truly know μ there would be no need to gather a sample to estimate it.



Confidence Interval for μ (σ Unknown)

- If the population standard deviation σ is unknown, we can substitute the sample standard deviation, S
- This introduces extra uncertainty, since S is variable from sample to sample
- So we use the t distribution instead of the normal distribution

Confidence Interval for μ (σ Unknown)

(continued)

- Assumptions
 - Population standard deviation is unknown
 - Population is normally distributed
 - If population is not normal, use large sample
- Use Student's t Distribution
- Confidence Interval Estimate:

$$\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}}$$

(where $t_{\alpha/2}$ is the critical value of the t distribution with $n - 1$ degrees of freedom and an area of $\alpha/2$ in each tail)



Student's t Distribution

- The t is a family of distributions
- The $t_{\alpha/2}$ value depends on **degrees of freedom (d.f.)**
 - Number of observations that are free to vary after sample mean has been calculated

$$\text{d.f.} = n - 1$$



Degrees of Freedom (df)

Idea: Number of observations that are free to vary after sample mean has been calculated

Example: Suppose the mean of 3 numbers is 8.0

Let $X_1 = 7$
Let $X_2 = 8$
What is X_3 ?



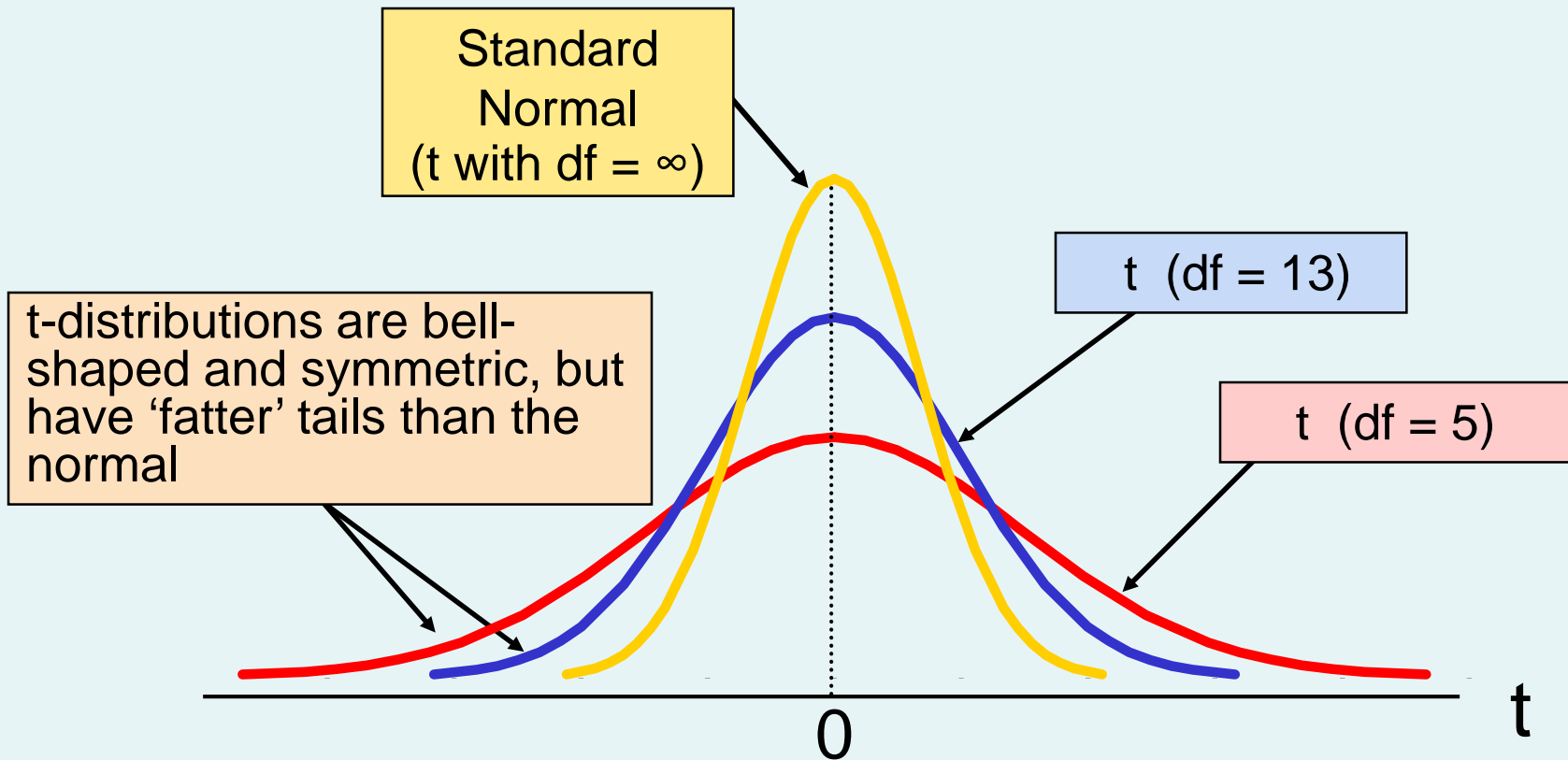
If the mean of these three values is 8.0,
then X_3 **must be 9**
(i.e., X_3 is not free to vary)

Here, $n = 3$, so degrees of freedom = $n - 1 = 3 - 1 = 2$

(2 values can be any numbers, but the third is not free to vary for a given mean)

Student's t Distribution

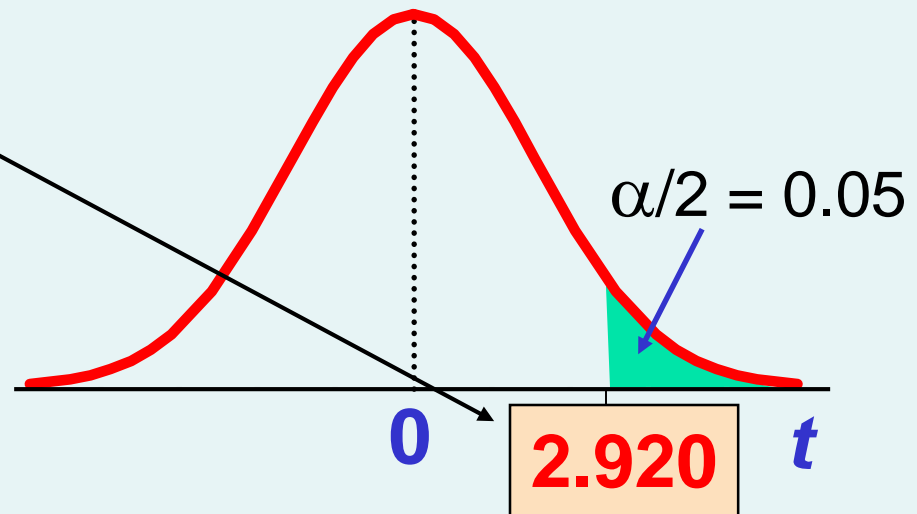
Note: $t \rightarrow Z$ as n increases



Student's t Table

| Upper Tail Area | | | |
|-----------------|-------|-------|--------------|
| df | .25 | .10 | .05 |
| 1 | 1.000 | 3.078 | 6.314 |
| 2 | 0.817 | 1.886 | 2.920 |
| 3 | 0.765 | 1.638 | 2.353 |

Let: $n = 3$
 $df = n - 1 = 2$
 $\alpha = 0.10$
 $\alpha/2 = 0.05$



The body of the table contains t values, not probabilities



Selected t distribution values

With comparison to the Z value

| Confidence Level | t (10 d.f.) | t (20 d.f.) | t (30 d.f.) | Z (∞ d.f.) |
|-------------------------|--------------------|--------------------|--------------------|-------------------------------------|
| 0.80 | 1.372 | 1.325 | 1.310 | 1.28 |
| 0.90 | 1.812 | 1.725 | 1.697 | 1.645 |
| 0.95 | 2.228 | 2.086 | 2.042 | 1.96 |
| 0.99 | 3.169 | 2.845 | 2.750 | 2.58 |

Note: $t \rightarrow Z$ as n increases

Example of t distribution confidence interval

A random sample of $n = 25$ has $\bar{X} = 50$ and $S = 8$. Form a 95% confidence interval for μ

- d.f. = $n - 1 = 24$, so $t_{\alpha/2} = t_{0.025} = 2.0639$

The confidence interval is

$$\bar{X} \pm t_{\alpha/2} \frac{S}{\sqrt{n}} = 50 \pm (2.0639) \frac{8}{\sqrt{25}}$$

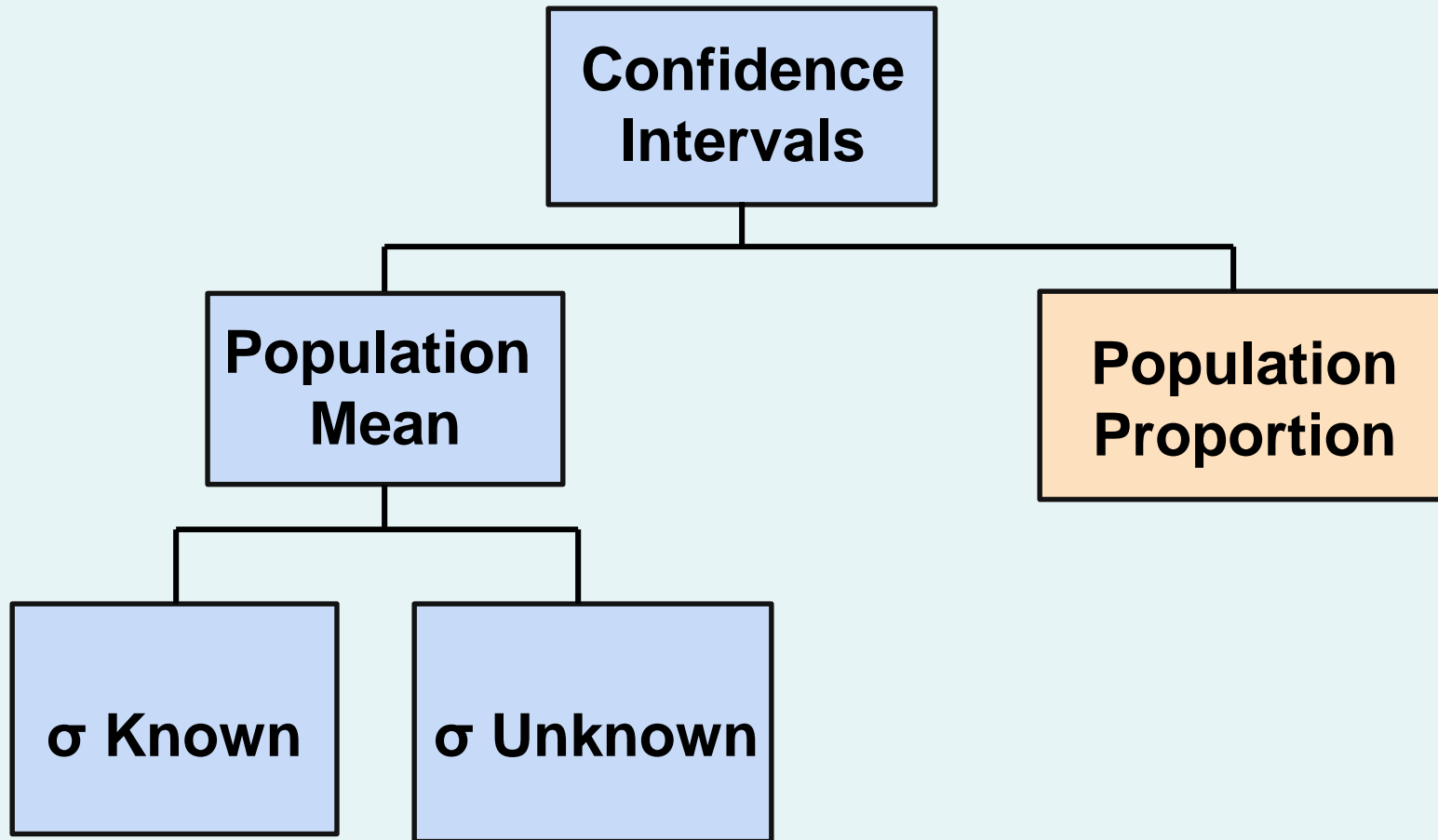
$$46.698 \leq \mu \leq 53.302$$

Example of t distribution confidence interval

(continued)

- Interpreting this interval requires the assumption that the population you are sampling from is approximately a normal distribution (especially since n is only 25).
- This condition can be checked by creating a:
 - Normal probability plot or
 - Boxplot

Confidence Intervals





Confidence Intervals for the Population Proportion, π

- An interval estimate for the population proportion (π) can be calculated by adding an allowance for uncertainty to the sample proportion (p)

Confidence Intervals for the Population Proportion, π

(continued)

- Recall that the distribution of the sample proportion is approximately normal if the sample size is large, with standard deviation

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

- We will estimate this with sample data:

$$\sqrt{\frac{p(1-p)}{n}}$$



Confidence Interval Endpoints

- Upper and lower confidence limits for the population proportion are calculated with the formula

$$p \pm Z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

- where
 - $Z_{\alpha/2}$ is the standard normal value for the level of confidence desired
 - p is the sample proportion
 - n is the sample size
- Note: must have $np > 5$ and $n(1-p) > 5$

Example

- A random sample of 100 people shows that 25 are left-handed.
- Form a 95% confidence interval for the true proportion of left-handers



Example

(continued)

- A random sample of 100 people shows that 25 are left-handed. Form a 95% confidence interval for the true proportion of left-handers.

$$\begin{aligned} p \pm Z_{\alpha/2} \sqrt{p(1-p)/n} \\ = 25/100 \pm 1.96 \sqrt{0.25(0.75)/100} \\ = 0.25 \pm 1.96 (0.0433) \end{aligned}$$

$$0.1651 \leq \pi \leq 0.3349$$



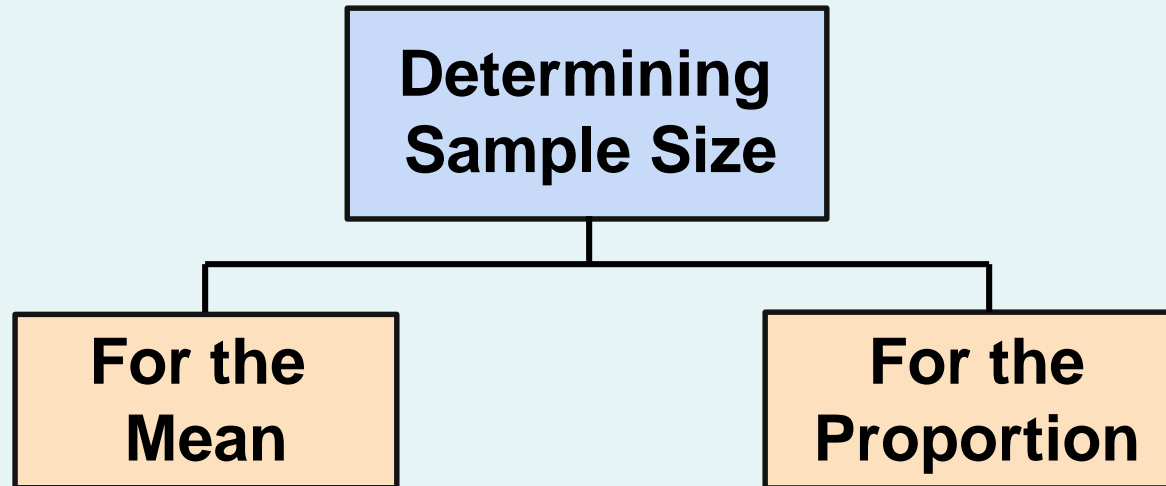
Interpretation

- We are 95% confident that the true percentage of left-handers in the population is between
16.51% and 33.49%.
- Although the interval from 0.1651 to 0.3349 may or may not contain the true proportion, 95% of intervals formed from samples of size 100 in this manner will contain the true proportion.





Determining Sample Size





Sampling Error

- The required sample size can be found to reach a desired **margin of error (e)** with a specified level of confidence $(1 - \alpha)$
- The margin of error is also called **sampling error**
 - the amount of imprecision in the estimate of the population parameter
 - the amount added and subtracted to the point estimate to form the confidence interval

Determining Sample Size

Determining
Sample Size

For the
Mean

Sampling error
(margin of error)

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

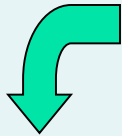
$$e = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Determining Sample Size

(continued)

**Determining
Sample Size**

**For the
Mean**



$$e = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Now solve
for n to get

$$n = \frac{Z_{\alpha/2}^2 \sigma^2}{e^2}$$



Determining Sample Size

(continued)

- To determine the required sample size for the mean, you must know:
 - The desired level of confidence $(1 - \alpha)$, which determines the critical value, $Z_{\alpha/2}$
 - The acceptable sampling error, e
 - The standard deviation, σ



Required Sample Size Example

If $\sigma = 45$, what sample size is needed to estimate the mean within ± 5 with 90% confidence?

$$n = \frac{Z^2 \sigma^2}{e^2} = \frac{(1.645)^2 (45)^2}{5^2} = 219.19$$

So the required sample size is **$n = 220$**

(Always round up)



If σ is unknown

- If unknown, σ can be estimated when using the required sample size formula
 - Use a value for σ that is expected to be at least as large as the true σ
 - Select a pilot sample and estimate σ with the sample standard deviation, S

Determining Sample Size

(continued)

**Determining
Sample Size**

**For the
Proportion**

$$e = Z \sqrt{\frac{\pi(1-\pi)}{n}}$$

Now solve
for n to get

$$n = \frac{Z^2 \pi (1 - \pi)}{e^2}$$



Determining Sample Size

(continued)

- To determine the required sample size for the proportion, you must know:
 - The desired level of confidence ($1 - \alpha$), which determines the critical value, $Z_{\alpha/2}$
 - The acceptable sampling error, e
 - The true proportion of events of interest, π
 - π can be estimated with a pilot sample if necessary (or conservatively use 0.5 as an estimate of π)



Required Sample Size Example

How large a sample would be necessary to estimate the true proportion defective in a large population **within $\pm 3\%$, with 95% confidence?**

(Assume a pilot sample yields $p = 0.12$)

Required Sample Size Example

(continued)

Solution:

For 95% confidence, use $Z_{\alpha/2} = 1.96$

$e = 0.03$

$p = 0.12$, so use this to estimate π

$$n = \frac{Z_{\alpha/2}^2 \pi (1 - \pi)}{e^2} = \frac{(1.96)^2 (0.12)(1 - 0.12)}{(0.03)^2} = 450.74$$

So use $n = 451$



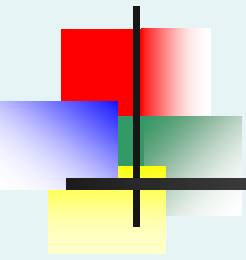
Ethical Issues

- A confidence interval estimate (reflecting sampling error) should always be included when reporting a point estimate
- The level of confidence should always be reported
- The sample size should be reported
- An interpretation of the confidence interval estimate should also be provided



Chapter Summary

- Introduced the concept of confidence intervals
- Discussed point estimates
- Developed confidence interval estimates
- Created confidence interval estimates for the mean (σ known)
- Determined confidence interval estimates for the mean (σ unknown)
- Created confidence interval estimates for the proportion
- Determined required sample size for mean and proportion settings
- Addressed confidence interval estimation and ethical issues



Business Statistics: A First Course 5th Edition

Chapter 9

Fundamentals of Hypothesis Testing: One-Sample Tests



Learning Objectives

In this chapter, you learn:

- The basic principles of hypothesis testing
- How to use hypothesis testing to test a mean or proportion
- The assumptions of each hypothesis-testing procedure, how to evaluate them, and the consequences if they are seriously violated
- How to avoid the pitfalls involved in hypothesis testing
- The ethical issues involved in hypothesis testing

What is a Hypothesis?

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

- population mean

Example: The mean monthly cell phone bill in this city is $\mu = \$42$

- population proportion

Example: The proportion of adults in this city with cell phones is $\pi = 0.68$



The Null Hypothesis, H_0

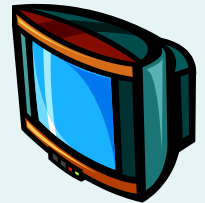
- States the claim or assertion to be tested

Example: The average number of TV sets in U.S. Homes is equal to three ($H_0 : \mu = 3$)

- Is always about a population parameter, not about a sample statistic

$$H_0 : \mu = 3$$

$$\cancel{H_0 : \bar{X} = 3}$$



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 or historical value
- Always contains “=”, “≤” or “≥” sign
- May or may not be rejected



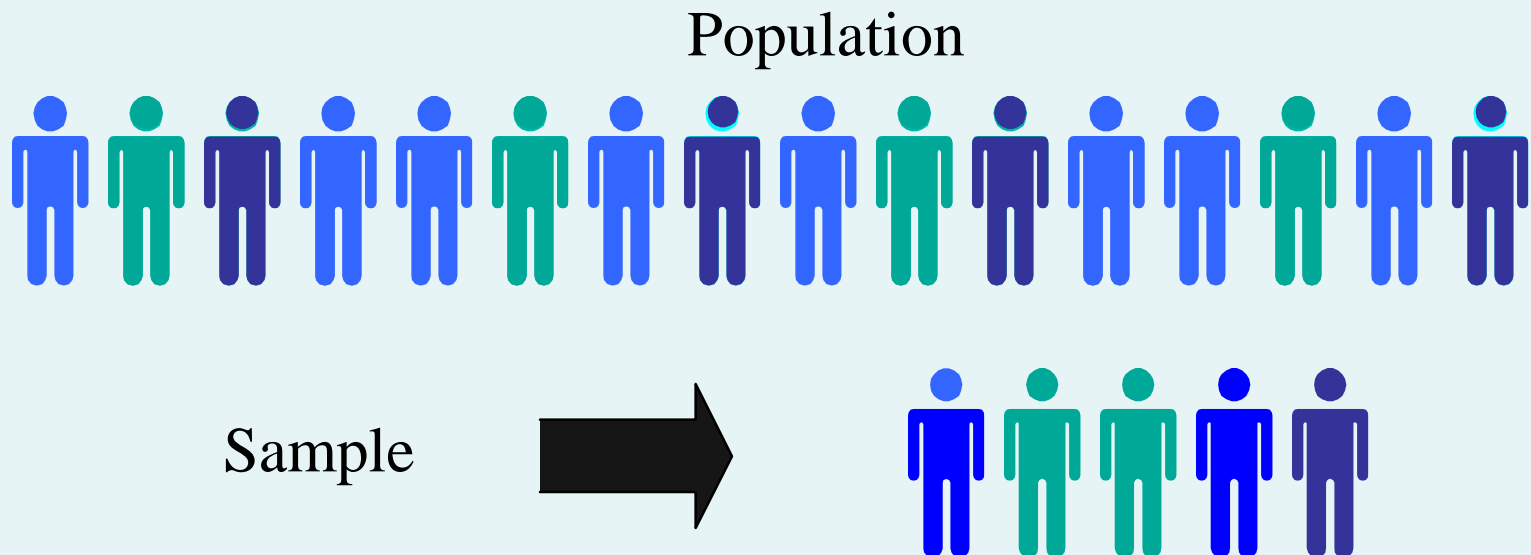


The Alternative Hypothesis, H_1

- Is the opposite of the null hypothesis
 - e.g., The average number of TV sets in U.S. homes is not equal to 3 ($H_1: \mu \neq 3$)
- Challenges the status quo
- Never contains the “=”, “≤” or “≥” sign
- May or may not be proven
- Is generally the hypothesis that the researcher is trying to prove

The Hypothesis Testing Process

- Claim: The population mean age is 50.
 - $H_0: \mu = 50$, $H_1: \mu \neq 50$
- Sample the population and find sample mean.



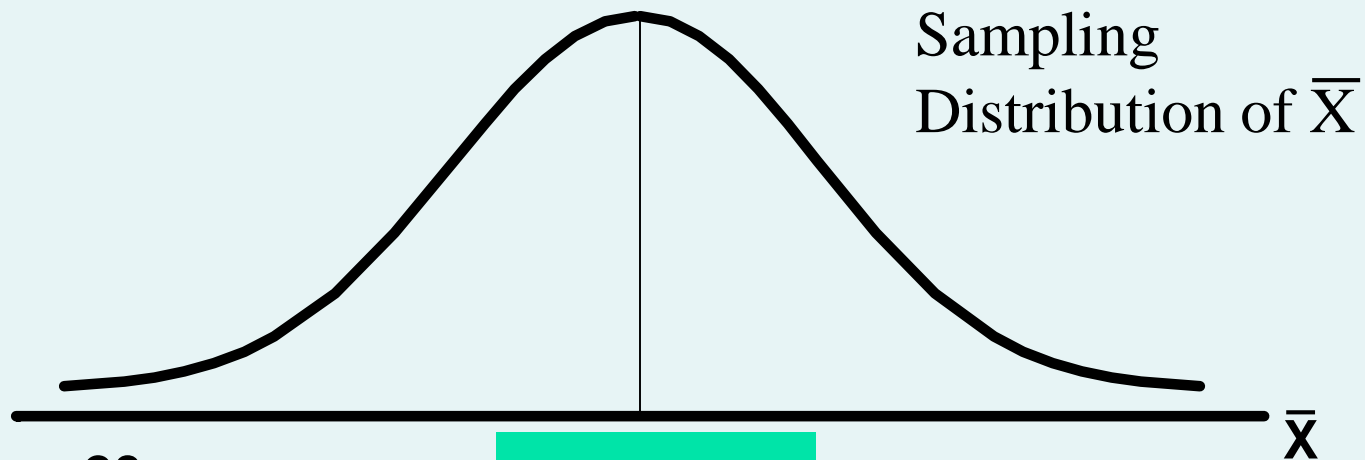
The Hypothesis Testing Process

(continued)

- Suppose the sample mean age was $\bar{X} = 20$.
- This is significantly lower than the claimed mean population age of 50.
- If the null hypothesis were true, the probability of getting such a different sample mean would be very small, so you reject the null hypothesis .
- In other words, getting a sample mean of 20 is so unlikely if the population mean was 50, you conclude that the population mean must not be 50.

The Hypothesis Testing Process

(continued)



If it is unlikely that you would get a sample mean of this value ...

$\mu = 50$
If H_0 is true

... When in fact this were the population mean...

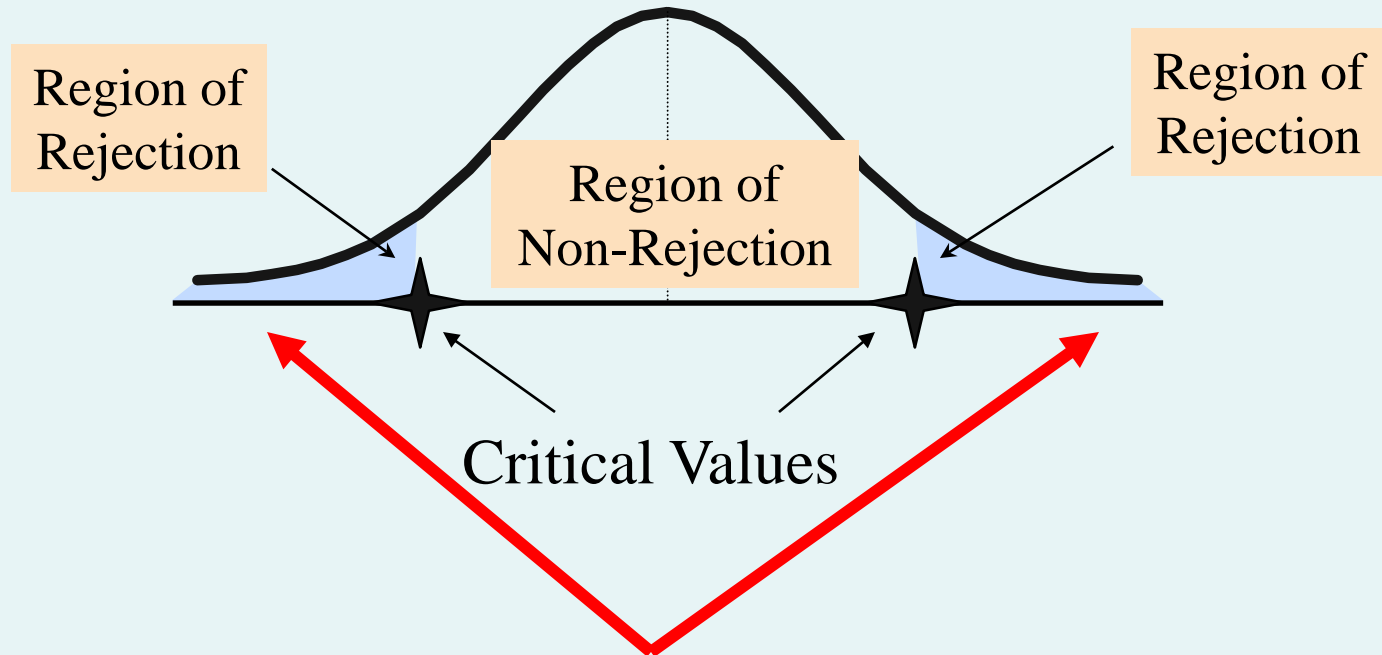
... then you reject the null hypothesis that $\mu = 50$.

The Test Statistic and Critical Values

- If the sample mean is close to the assumed population mean, the null hypothesis is not rejected.
- If the sample mean is far from the assumed population mean, the null hypothesis is rejected.
- How far is “far enough” to reject H_0 ?
- The critical value of a test statistic creates a “line in the sand” for decision making -- it answers the question of how far is far enough.

The Test Statistic and Critical Values

Sampling Distribution of the test statistic



“Too Far Away” From Mean of Sampling Distribution

Possible Errors in Hypothesis Test Decision Making

- **Type I Error**
 - Reject a true null hypothesis
 - Considered a serious type of error
 - The probability of a Type I Error is α
 - Called level of significance of the test
 - Set by researcher in advance
- **Type II Error**
 - Failure to reject false null hypothesis
 - The probability of a Type II Error is β

Possible Errors in Hypothesis Test Decision Making

(continued)

| Possible Hypothesis Test Outcomes | | |
|-----------------------------------|--------------------------------------|--------------------------------------|
| | Actual Situation | |
| Decision | H_0 True | H_0 False |
| Do Not Reject H_0 | No Error Probability $1 - \alpha$ | Type II Error Probability β |
| Reject H_0 | Type I Error Probability α | No Error Probability $1 - \beta$ |

Possible Results in Hypothesis Test Decision Making



(continued)

- The **confidence coefficient** $(1-\alpha)$ is the probability of not rejecting H_0 when it is true.
- The **confidence level** of a hypothesis test is $(1-\alpha)*100\%$.
- The **power of a statistical test** $(1-\beta)$ is the probability of rejecting H_0 when it is false.











Type I & II Error Relationship

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

If Type I error probability (α) , then
Type II error probability (β) 

Factors Affecting Type II Error

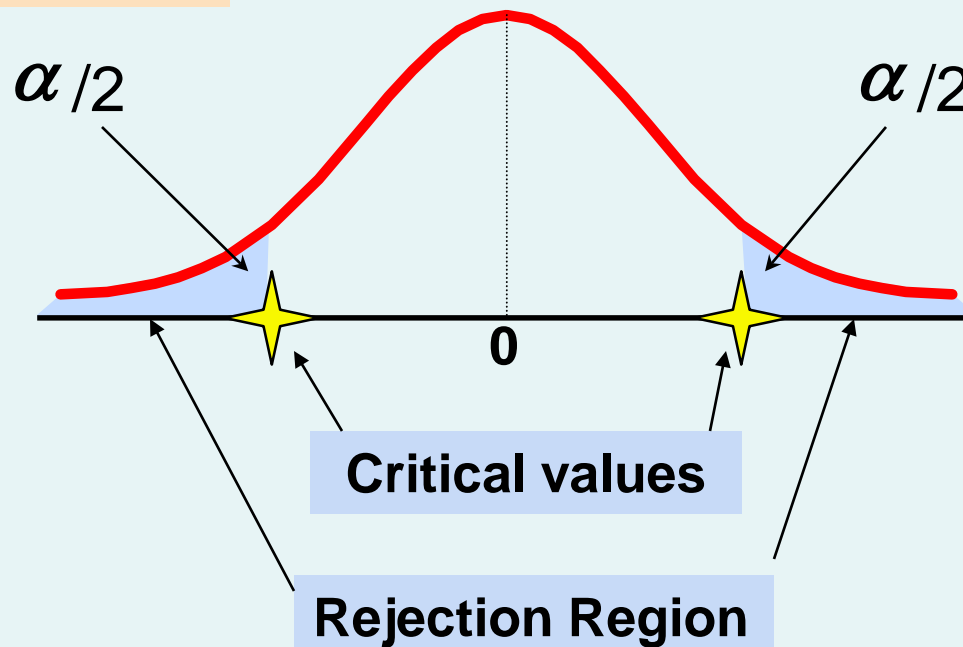
- All else equal,
 - β  when the difference between hypothesized parameter and its true value 
 - β  when α 
 - β  when σ 
 - β  when n 

Level of Significance and the Rejection Region

$$H_0: \mu = 3$$

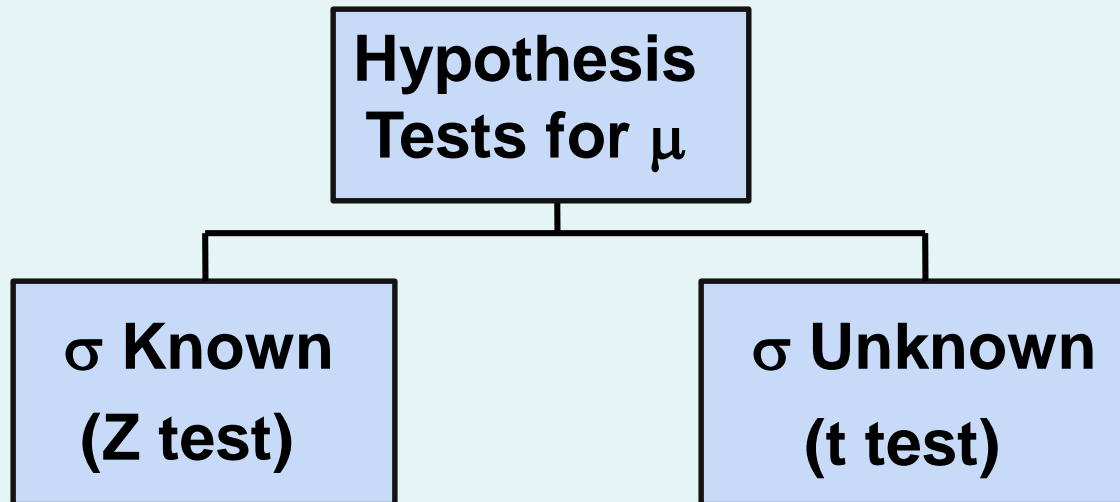
$$H_1: \mu \neq 3$$

Level of significance = α



This is a **two-tail test** because there is a rejection region in both tails

Hypothesis Tests for the Mean



Z Test of Hypothesis for the Mean (σ Known)

- Convert sample statistic (\bar{X}) to a Z_{STAT} test statistic

Hypothesis Tests for μ

σ Known
(Z test)

σ Unknown
(t test)

The test statistic is:

$$Z_{STAT} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

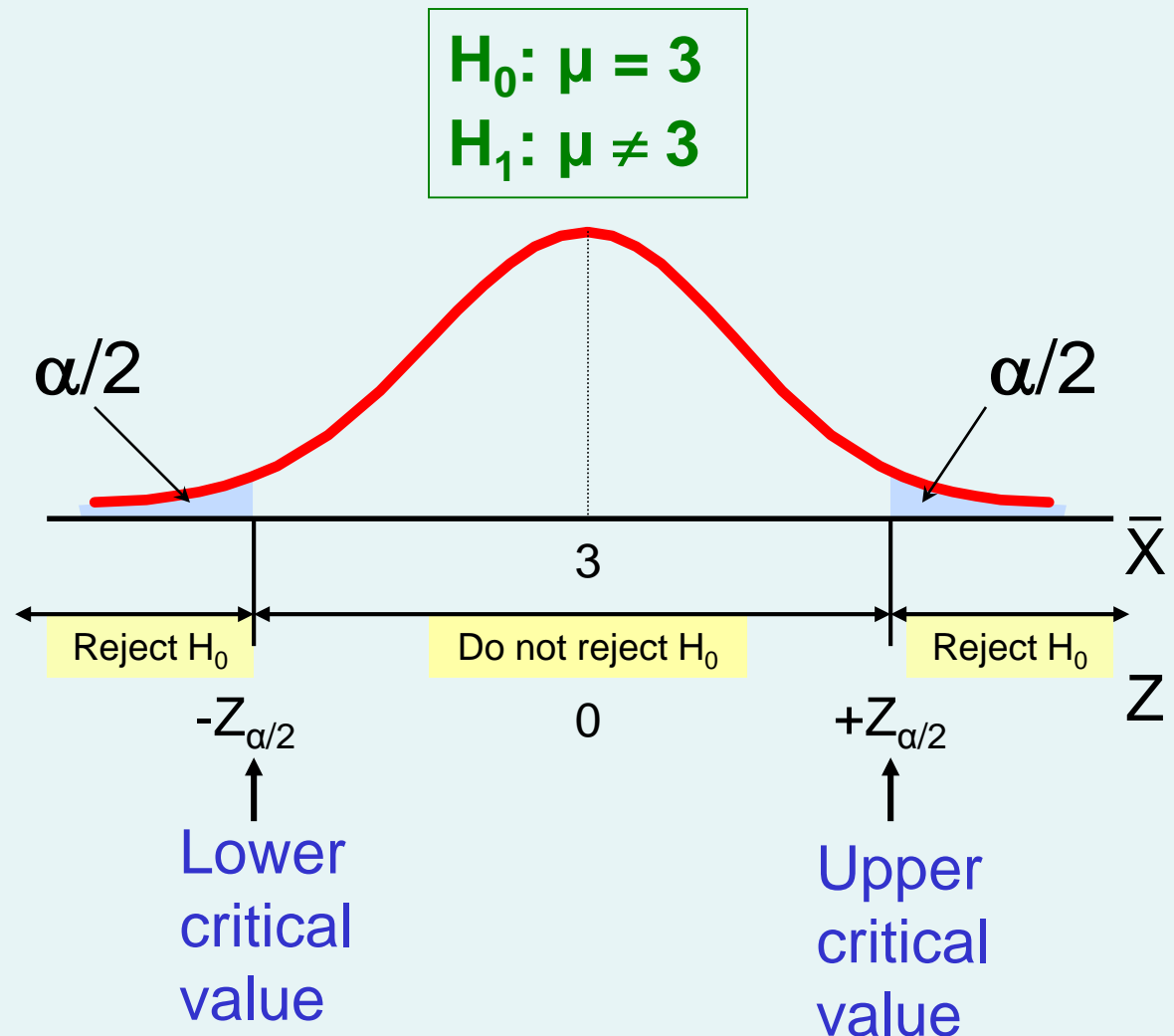


Critical Value Approach to Testing

- For a two-tail test for the mean, σ known:
- Convert sample statistic (\bar{X}) to test statistic (Z_{STAT})
- Determine the critical Z values for a specified level of significance α from a table or computer
- **Decision Rule:** If the test statistic falls in the rejection region, reject H_0 ; otherwise do not reject H_0

Two-Tail Tests

- There are two cutoff values (critical values), defining the regions of rejection





6 Steps in Hypothesis Testing

1. State the null hypothesis, H_0 and the alternative hypothesis, H_1
2. Choose the level of significance, α , and the sample size, n
3. Determine the appropriate test statistic and sampling distribution
4. Determine the critical values that divide the rejection and nonrejection regions



6 Steps in Hypothesis Testing

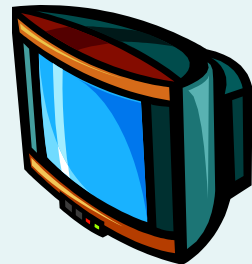
(continued)

5. Collect data and compute the value of the test statistic
6. Make the statistical decision and state the managerial conclusion. If the test statistic falls into the non rejection region, do not reject the null hypothesis H_0 . If the test statistic falls into the rejection region, reject the null hypothesis. Express the managerial conclusion in the context of the problem

Hypothesis Testing Example

**Test the claim that the true mean # of TV sets in US homes is equal to 3.
(Assume $\sigma = 0.8$)**

1. State the appropriate null and alternative hypotheses
 - $H_0: \mu = 3$ $H_1: \mu \neq 3$ (This is a two-tail test)
2. Specify the desired level of significance and the sample size
 - Suppose that $\alpha = 0.05$ and $n = 100$ are chosen for this test



Hypothesis Testing Example

(continued)

3. Determine the appropriate technique
 - σ is assumed known so this is a Z test.
4. Determine the critical values
 - For $\alpha = 0.05$ the critical Z values are ± 1.96
5. Collect the data and compute the test statistic
 - Suppose the sample results are
 $n = 100$, $\bar{X} = 2.84$ ($\sigma = 0.8$ is assumed known)

So the test statistic is:

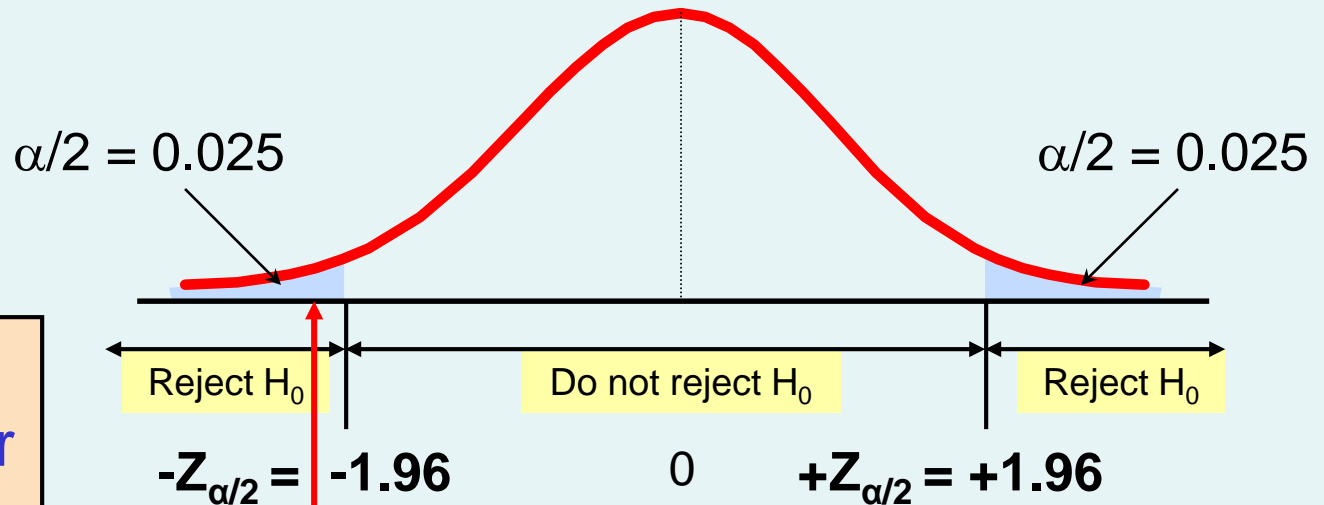
$$Z_{STAT} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2.84 - 3}{\frac{0.8}{\sqrt{100}}} = \frac{-.16}{.08} = -2.0$$



Hypothesis Testing Example

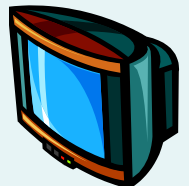
(continued)

- 6. Is the test statistic in the rejection region?



Reject H_0 if
 $Z_{STAT} < -1.96$ or
 $Z_{STAT} > 1.96$;
otherwise do
not reject H_0

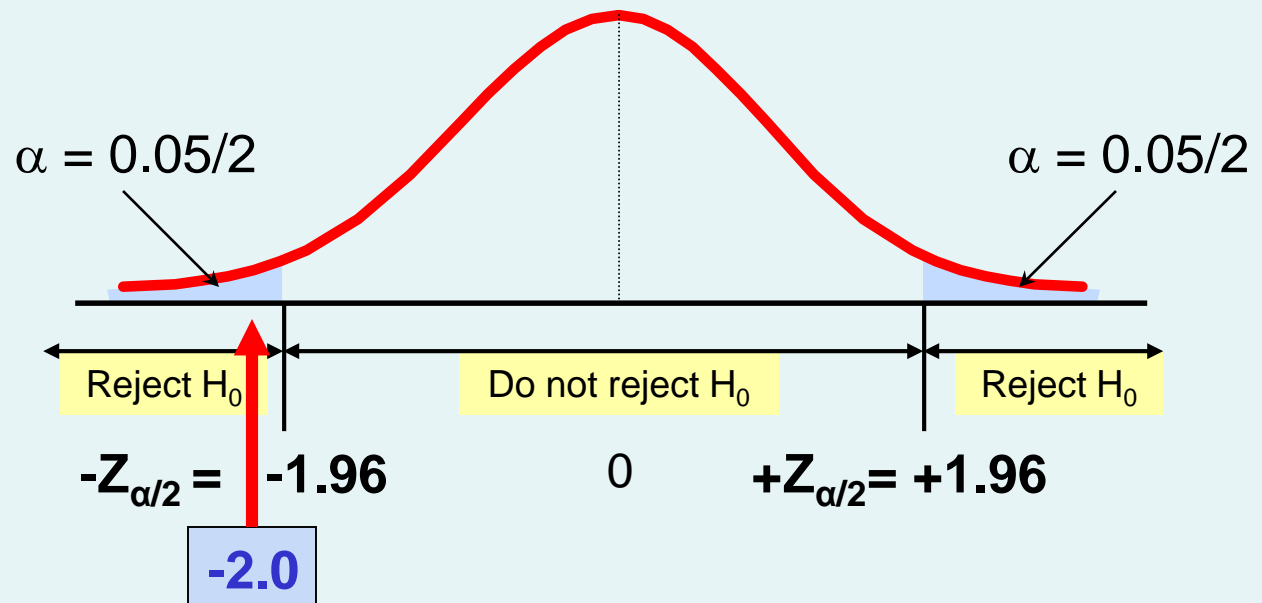
Here, $Z_{STAT} = -2.0 < -1.96$, so the
test statistic is in the rejection
region



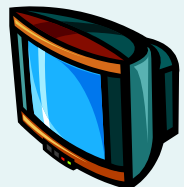
Hypothesis Testing Example

(continued)

6 (continued). Reach a decision and interpret the result



Since $Z_{\text{STAT}} = -2.0 < -1.96$, reject the null hypothesis and conclude there is sufficient evidence that the mean number of TVs in US homes is not equal to 3





p-Value Approach to Testing

- p-value: Probability of obtaining a test statistic equal to or more extreme than the observed sample value **given H_0 is true**
 - The p-value is also called the observed level of significance
 - It is the smallest value of α for which H_0 can be rejected



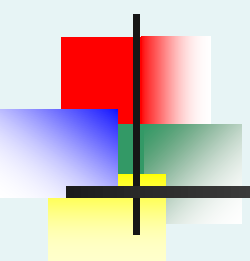
p-Value Approach to Testing: Interpreting the p-value

- Compare the **p-value** with α

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

- Remember

- If the p-value is low then H_0 must go



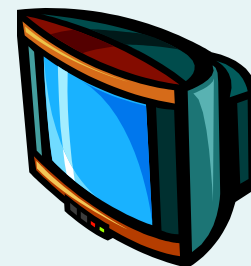
The 5 Step p-value approach to Hypothesis Testing

1. State the null hypothesis, H_0 and the alternative hypothesis, H_1
2. Choose the level of significance, α , and the sample size, n
3. Determine the appropriate test statistic and sampling distribution
4. Collect data and compute the value of the test statistic and the p-value
5. Make the statistical decision and state the managerial conclusion. If the p-value is $< \alpha$ then reject H_0 , otherwise do not reject H_0 . State the managerial conclusion in the context of the problem

p-value Hypothesis Testing Example

**Test the claim that the true mean # of TV sets in US homes is equal to 3.
(Assume $\sigma = 0.8$)**

1. State the appropriate null and alternative hypotheses
 - $H_0: \mu = 3$ $H_1: \mu \neq 3$ (This is a two-tail test)
2. Specify the desired level of significance and the sample size
 - Suppose that $\alpha = 0.05$ and $n = 100$ are chosen for this test



p-value Hypothesis Testing Example

(continued)

- Determine the appropriate technique
 - σ is assumed known so this is a Z test.
- Collect the data, compute the test statistic and the p-value
 - Suppose the sample results are
 $n = 100$, $\bar{X} = 2.84$ ($\sigma = 0.8$ is assumed known)

So the test statistic is:

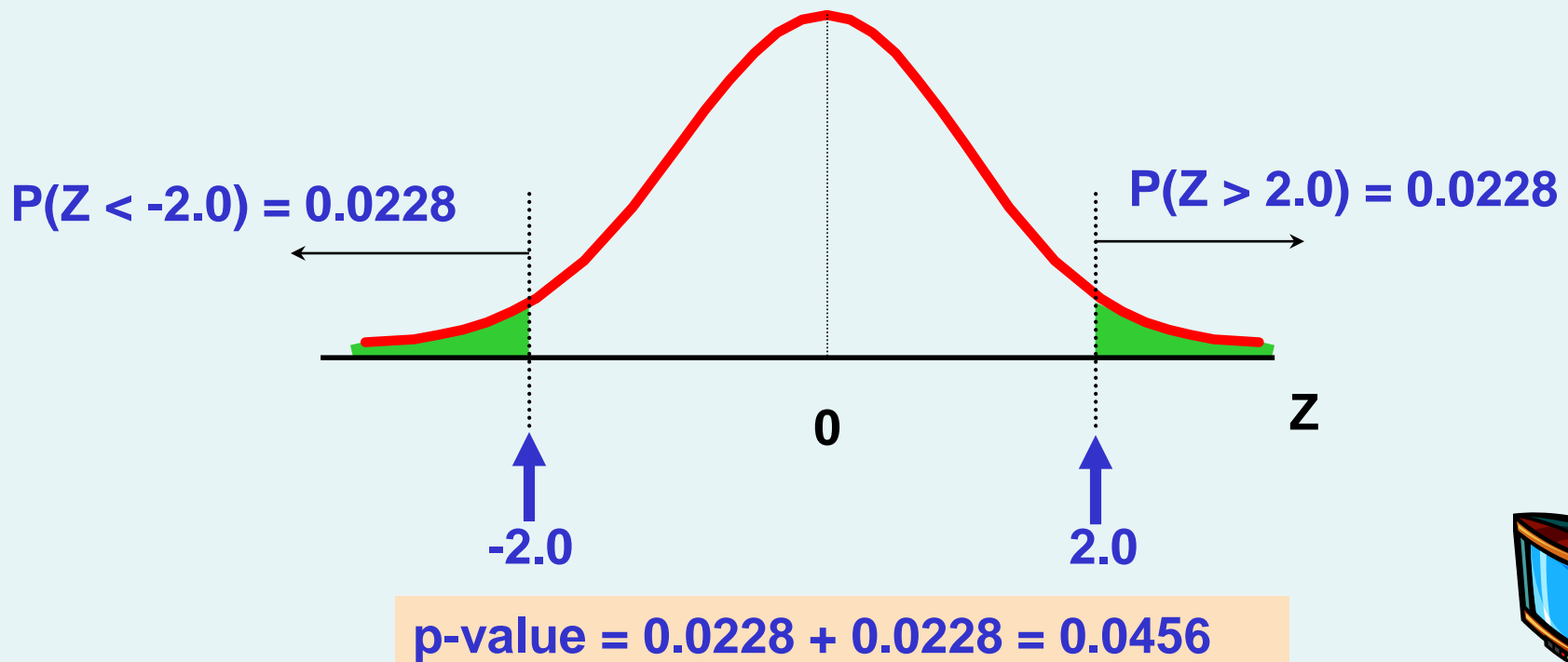
$$Z_{STAT} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2.84 - 3}{\frac{0.8}{\sqrt{100}}} = \frac{-.16}{.08} = -2.0$$



p-Value Hypothesis Testing Example: Calculating the p-value

4. (continued) Calculate the p-value.

- How likely is it to get a Z_{STAT} of -2 (or something further from the mean (0), in either direction) if H_0 is true?



p-value Hypothesis Testing Example

(continued)

- 5. Is the p-value $< \alpha$?
 - Since p-value = 0.0456 $< \alpha = 0.05$ Reject H_0
- 5. (continued) State the managerial conclusion in the context of the situation.
 - There is sufficient evidence to conclude the average number of TVs in US homes is not equal to 3.



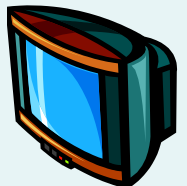
Connection Between Two Tail Tests and Confidence Intervals

- For $\bar{X} = 2.84$, $\sigma = 0.8$ and $n = 100$, the 95% confidence interval is:

$$2.84 - (1.96) \frac{0.8}{\sqrt{100}} \text{ to } 2.84 + (1.96) \frac{0.8}{\sqrt{100}}$$

$$2.6832 \leq \mu \leq 2.9968$$

- Since this interval does not contain the hypothesized mean (3.0), we reject the null hypothesis at $\alpha = 0.05$

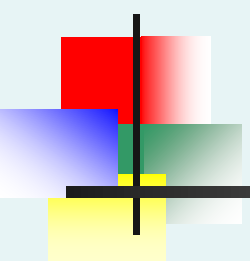




Do You Ever Truly Know σ ?

- Probably not!
- In virtually all real world business situations, σ is not known.
- If there is a situation where σ is known then μ is also known (since to calculate σ you need to know μ .)
- If you truly know μ there would be no need to gather a sample to estimate it.

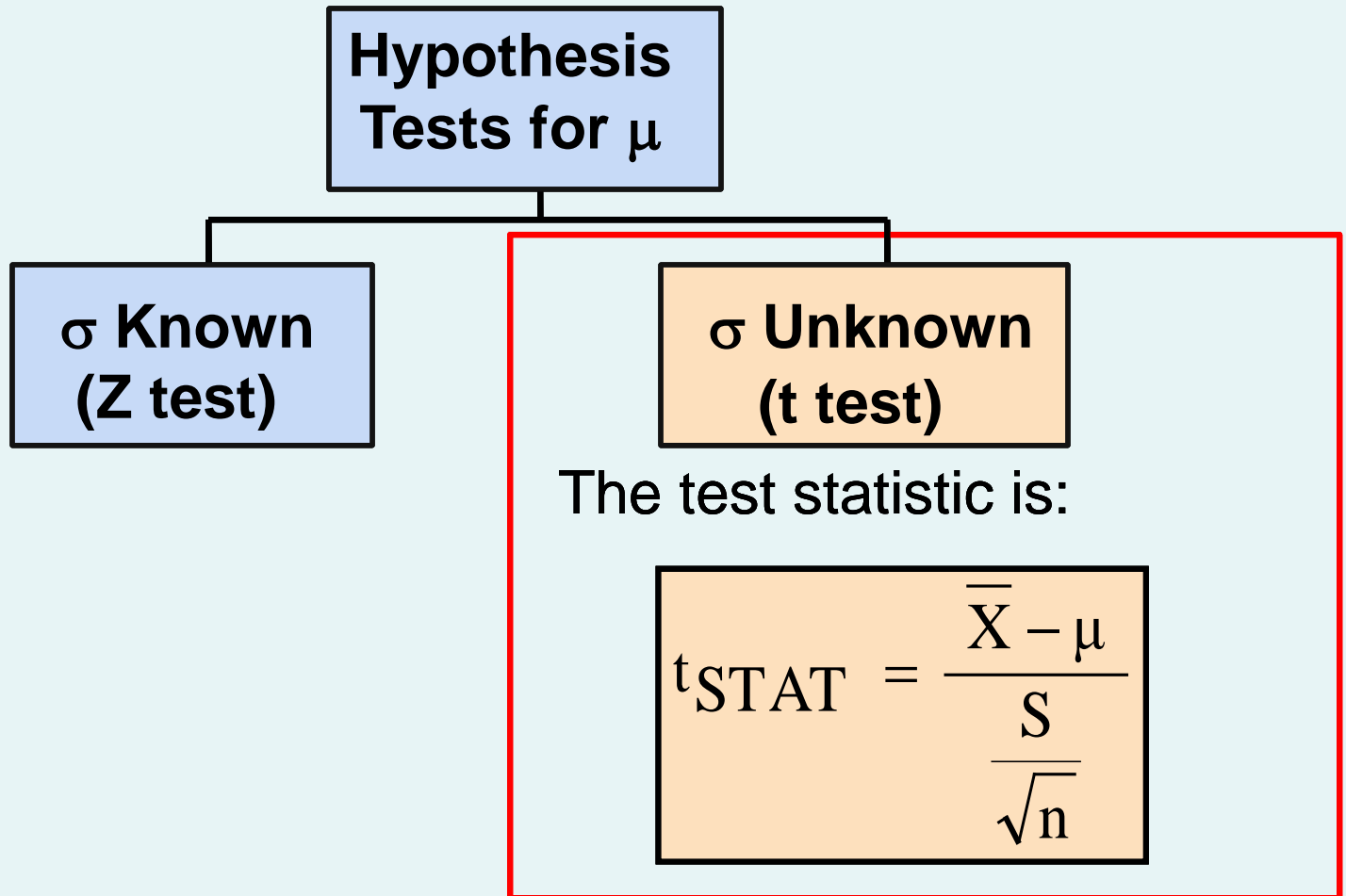
Hypothesis Testing: σ Unknown



- If the population standard deviation is unknown, you instead use the sample standard deviation S .
- Because of this change, you use the t distribution instead of the Z distribution to test the null hypothesis about the mean.
- When using the t distribution you must assume the population you are sampling from follows a normal distribution.
- All other steps, concepts, and conclusions are the same.

t Test of Hypothesis for the Mean (σ Unknown)

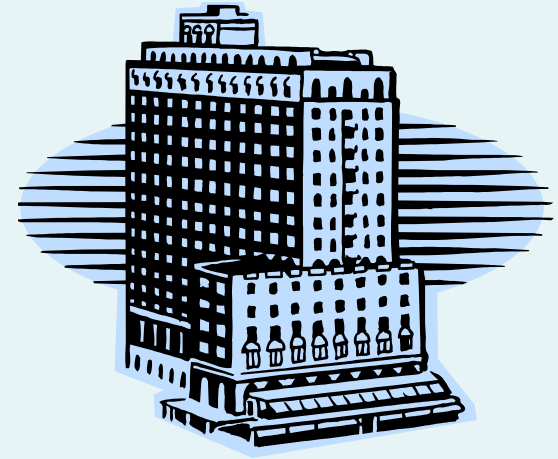
- Convert sample statistic (\bar{X}) to a t_{STAT} test statistic



Example: Two-Tail Test (σ Unknown)

The average cost of a hotel room in New York is said to be \$168 per night. To determine if this is true, a random sample of 25 hotels is taken and resulted in an \bar{X} of \$172.50 and an S of \$15.40. Test the appropriate hypotheses at $\alpha = 0.05$.

(Assume the population distribution is normal)



$$H_0: \mu = 168$$

$$H_1: \mu \neq 168$$

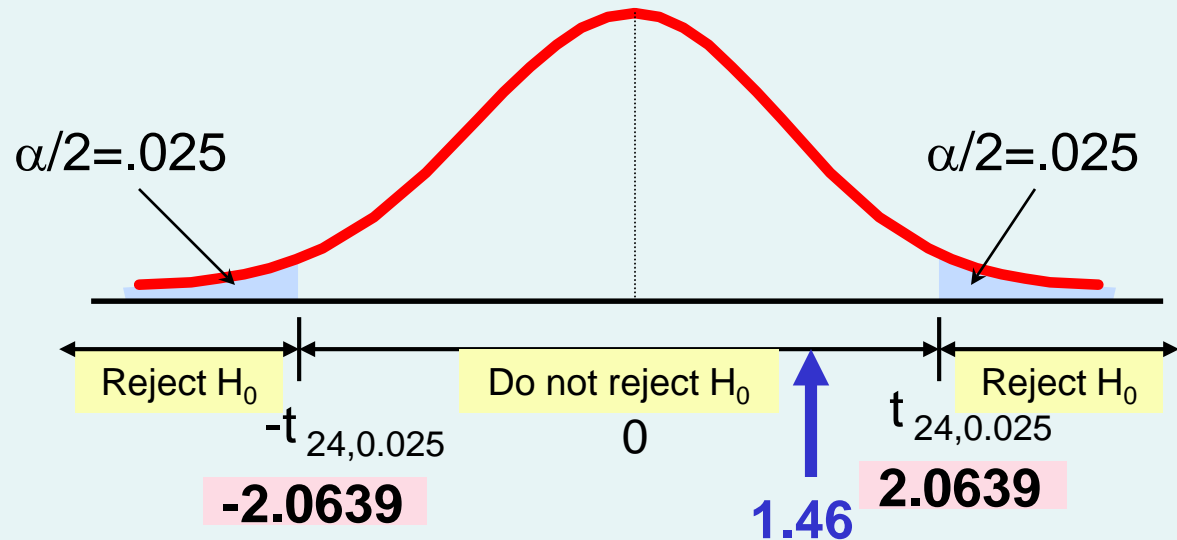
Example Solution: Two-Tail t Test

$$H_0: \mu = 168$$

$$H_1: \mu \neq 168$$

- $\alpha = 0.05$
- $n = 25, df = 25-1=24$
- σ is unknown, so use a **t statistic**
- **Critical Value:**

$$\pm t_{24,0.025} = \pm 2.0639$$



$$t_{STAT} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{172.50 - 168}{\frac{15.40}{\sqrt{25}}} = 1.46$$

Do not reject H_0 : insufficient evidence that true mean cost is different than \$168

Example Two-Tail t Test Using A p-value from Excel

- Since this is a t-test we cannot calculate the p-value without some calculation aid.
- The Excel output below does this:

t Test for the Hypothesis of the Mean

| Data | |
|---------------------------|-----------|
| Null Hypothesis $\mu=$ | \$ 168.00 |
| Level of Significance | 0.05 |
| Sample Size | 25 |
| Sample Mean | \$ 172.50 |
| Sample Standard Deviation | \$ 15.40 |

Intermediate Calculations

| | | | |
|----------------------------|----|-------------|--------------|
| Standard Error of the Mean | \$ | 3.08 | =B8/SQRT(B6) |
| Degrees of Freedom | | 24 | =B6-1 |
| t test statistic | | 1.46 | =(B7-B4)/B11 |

Two-Tail Test

| | | |
|-------------------------------|---------|------------------------------------------------------------------------|
| Lower Critical Value | -2.0639 | =-TINV(B5,B12) |
| Upper Critical Value | 2.0639 | =TINV(B5,B12) |
| p-value | 0.157 | =TDIST(ABS(B13),B12,2) |
| Do Not Reject Null Hypothesis | | =IF(B18<B5, "Reject null hypothesis", "Do not reject null hypothesis") |

p-value > α
So do not reject H_0

Example Two-Tail t Test Using A p-value from Minitab

One-Sample T

Test of $\mu = 168$ vs not = 168

| N | Mean | StDev | SE Mean | 95% CI | T | P |
|----|--------|-------|---------|------------------|------|-------|
| 25 | 172.50 | 15.40 | 3.08 | (166.14, 178.86) | 1.46 | 0.157 |

p-value > α
So do not reject H_0



Connection of Two Tail Tests to Confidence Intervals

- For $\bar{X} = 172.5$, $S = 15.40$ and $n = 25$, the 95% confidence interval for μ is:

$$172.5 - (2.0639) 15.4/\sqrt{25} \quad \text{to} \quad 172.5 + (2.0639) 15.4/\sqrt{25}$$

$$166.14 \leq \mu \leq 178.86$$

- Since this interval contains the Hypothesized mean (**168**), we do not reject the null hypothesis at $\alpha = 0.05$

One-Tail Tests

- In many cases, the alternative hypothesis focuses on a particular direction

$$H_0: \mu \geq 3$$

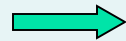
$$H_1: \mu < 3$$



This is a **lower**-tail test since the alternative hypothesis is focused on the lower tail below the mean of 3

$$H_0: \mu \leq 3$$

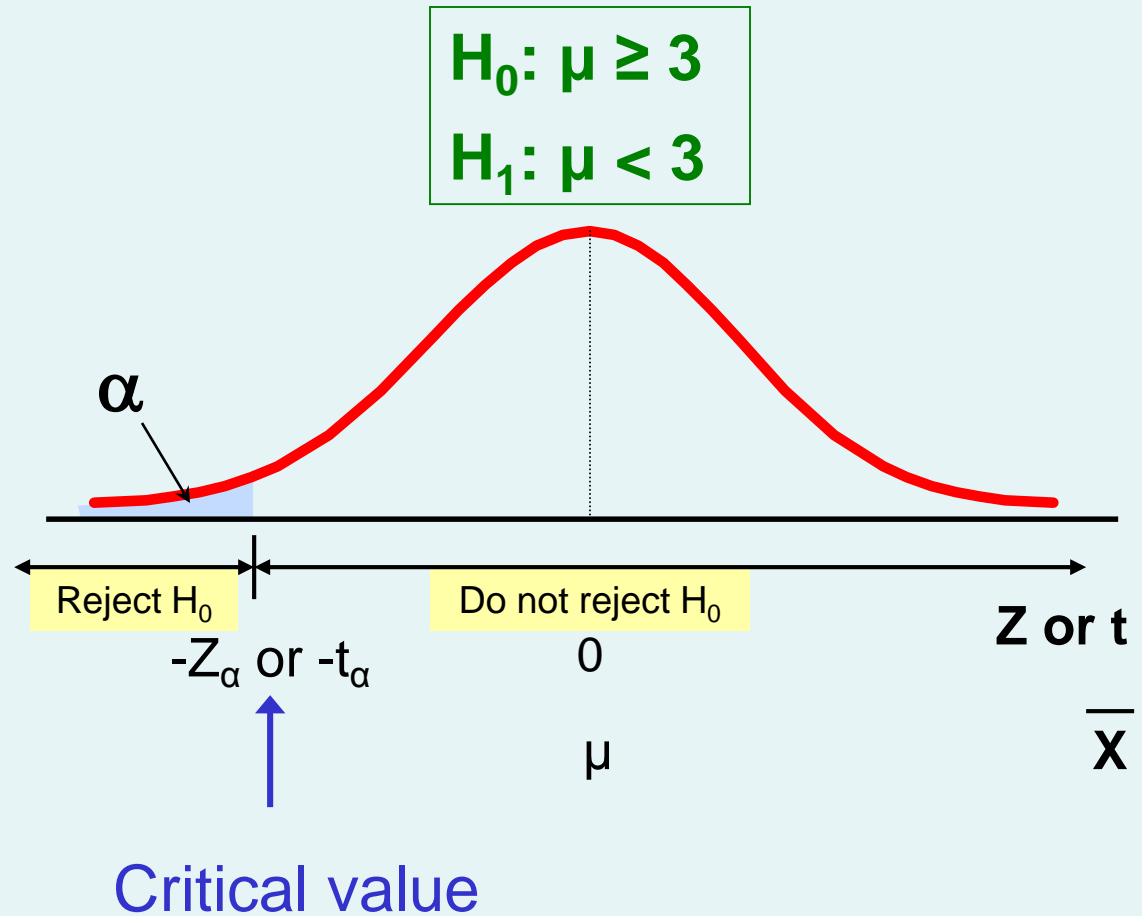
$$H_1: \mu > 3$$



This is an **upper**-tail test since the alternative hypothesis is focused on the upper tail above the mean of 3

Lower-Tail Tests

- There is only one critical value, since the rejection area is in only one tail

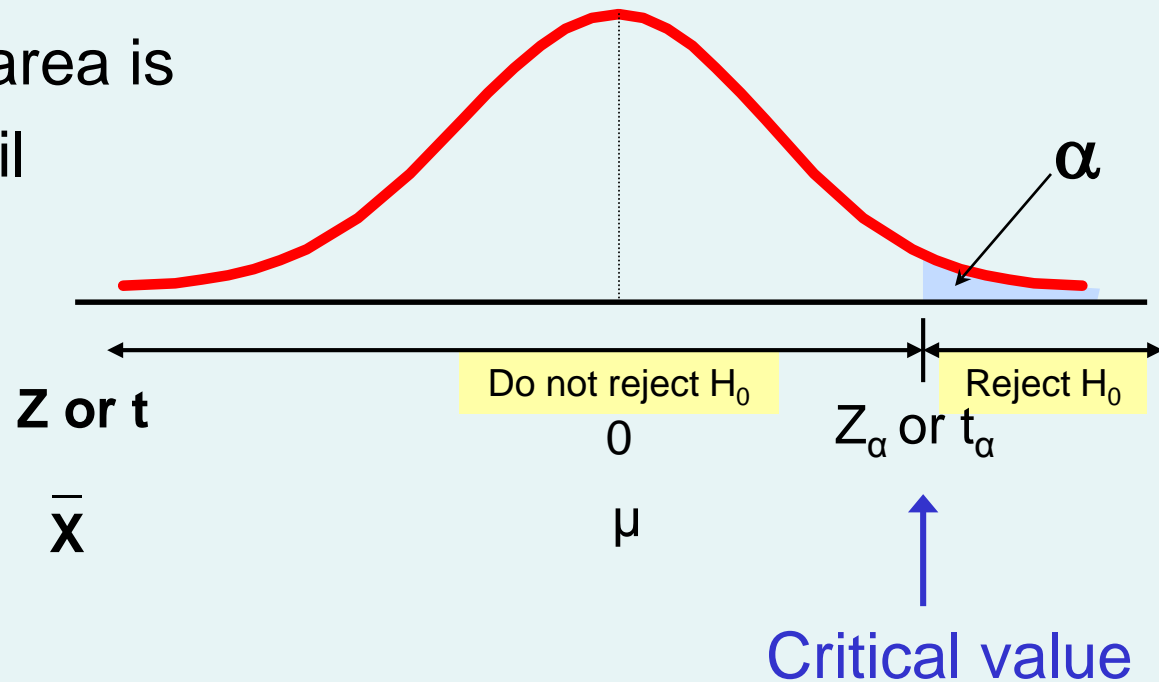


Upper-Tail Tests

- There is only one critical value, since the rejection area is in only one tail

$$H_0: \mu \leq 3$$

$$H_1: \mu > 3$$



Example: Upper-Tail t Test for Mean (σ unknown)

A phone industry manager thinks that customer monthly cell phone bills have increased, and now average over \$52 per month. The company wishes to test this claim. (Assume a normal population)



Form hypothesis test:

$H_0: \mu \leq 52$ the average is not over \$52 per month

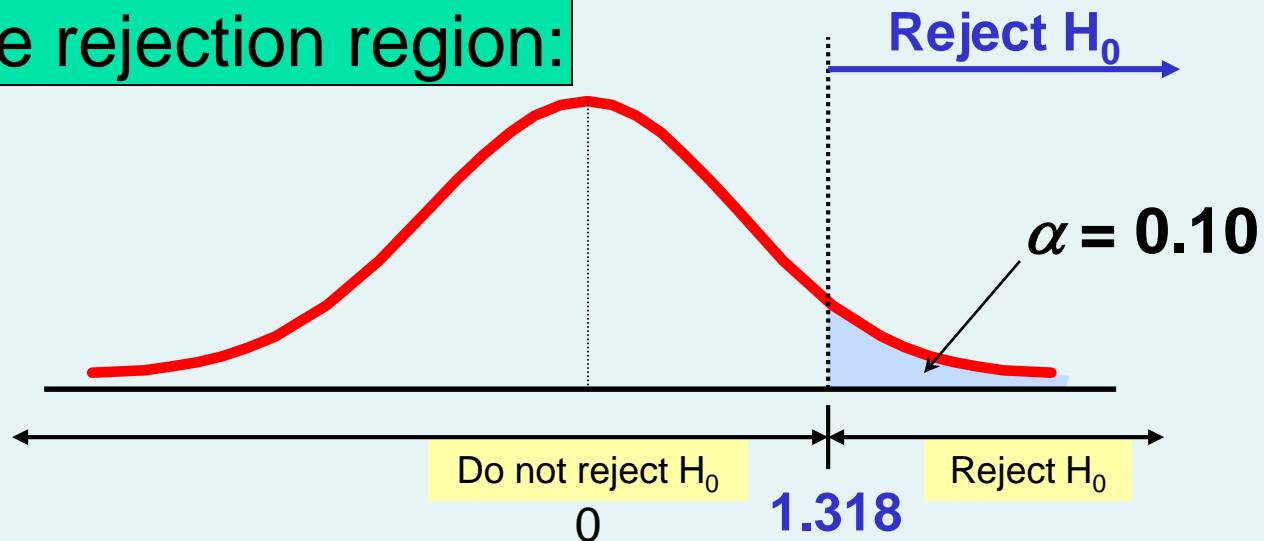
$H_1: \mu > 52$ the average **is** greater than \$52 per month
(i.e., sufficient evidence exists to support the manager's claim)

Example: Find Rejection Region

(continued)

- Suppose that $\alpha = 0.10$ is chosen for this test and $n = 25$.

Find the rejection region:



Reject H_0 if $t_{\text{STAT}} > 1.318$



Example: Test Statistic

(continued)

Obtain sample and compute the test statistic

Suppose a sample is taken with the following results: $n = 25$, $\bar{X} = 53.1$, and $S = 10$

- Then the test statistic is:

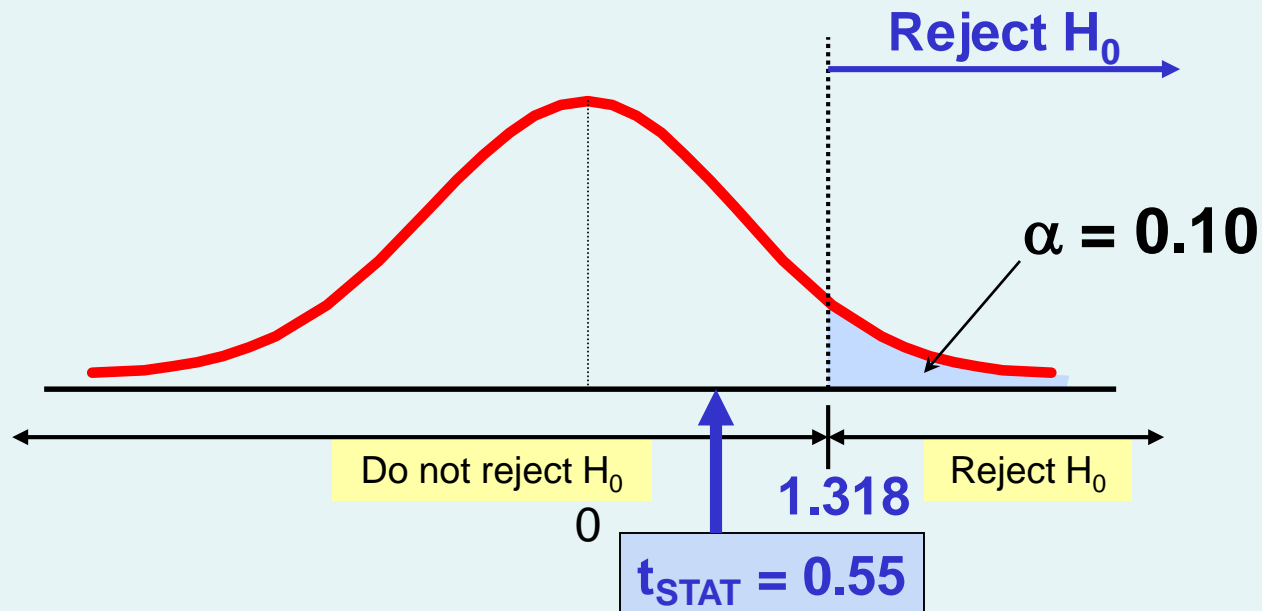
$$t_{\text{STAT}} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{53.1 - 52}{\frac{10}{\sqrt{25}}} = 0.55$$



Example: Decision

(continued)

Reach a decision and interpret the result:



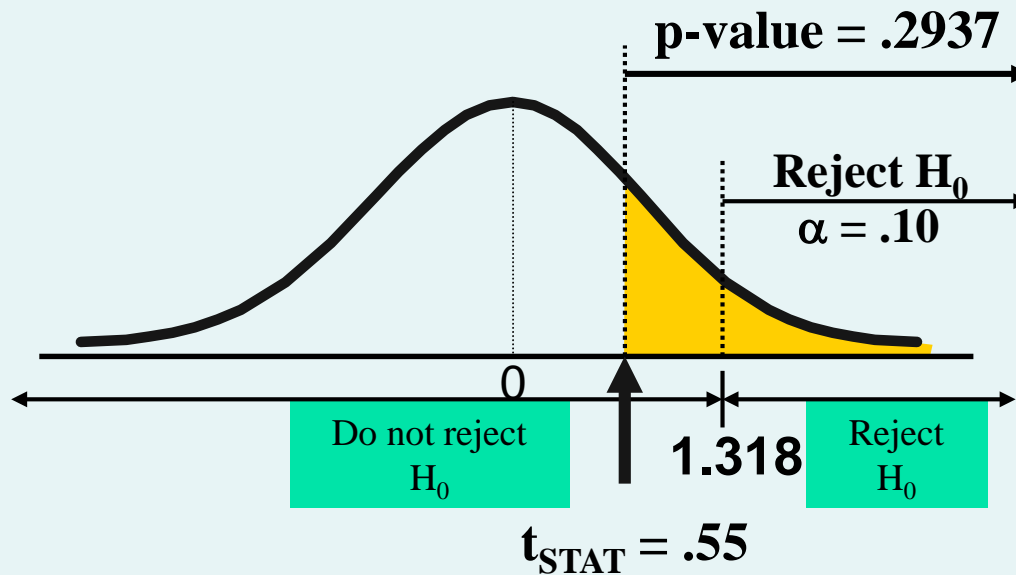
Do not reject H_0 since $t_{STAT} = 0.55 \leq 1.318$

there is not sufficient evidence that the mean bill is over \$52



Example: Utilizing The p-value for The Test

- Calculate the p-value and compare to α (p-value below calculated using excel spreadsheet on next page)



Do not reject H_0 since p-value = .2937 > $\alpha = .10$

Excel Spreadsheet Calculating The p-value for The Upper Tail t Test

t Test for the Hypothesis of the Mean

| Data | |
|---------------------------|-------|
| Null Hypothesis $\mu =$ | 52.00 |
| Level of Significance | 0.1 |
| Sample Size | 25 |
| Sample Mean | 53.10 |
| Sample Standard Deviation | 10.00 |

| Intermediate Calculations | |
|-------------------------------|-------------|
| Standard Error of the Mean | 2.00 |
| Degrees of Freedom | 24 |
| t test statistic | 0.55 |
| Upper Tail Test | |
| Upper Critical Value | 1.318 |
| p-value | 0.2937 |
| Do Not Reject Null Hypothesis | |

=B8/SQRT(B6)

=B6-1

=(B7-B4)/B11

=TINV(2*B5,B12)

=TDIST(ABS(B13),B12,1)

=IF(B18<B5, "Reject null hypothesis",

"Do not reject null hypothesis")



Hypothesis Tests for Proportions

- Involves categorical variables
- Two possible outcomes
 - Possesses characteristic of interest
 - Does not possess characteristic of interest
- Fraction or proportion of the population in the category of interest is denoted by π

Proportions

(continued)

- Sample proportion in the category of interest is denoted by p

- $$p = \frac{X}{n} = \frac{\text{number in category of interest in sample}}{\text{sample size}}$$

- When both $n\pi$ and $n(1-\pi)$ are at least 5, p can be approximated by a normal distribution with mean and standard deviation

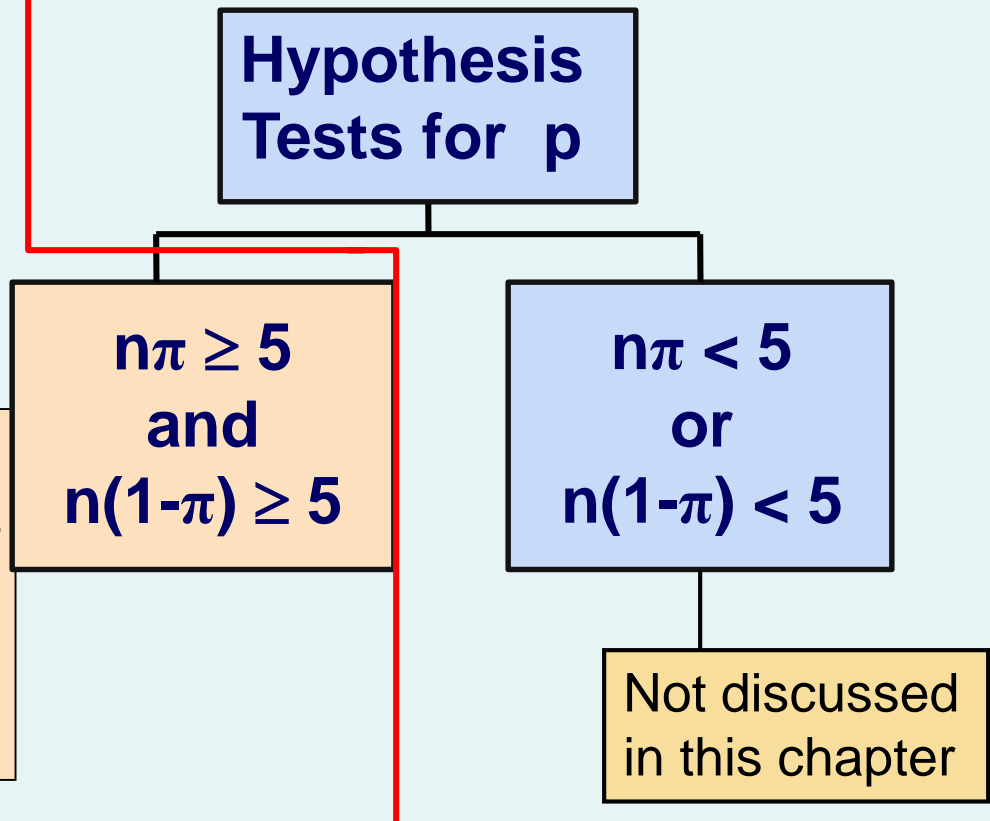
- $$\mu_p = \pi$$

- $$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Hypothesis Tests for Proportions

- The sampling distribution of p is approximately normal, so the test statistic is a Z_{STAT} value:

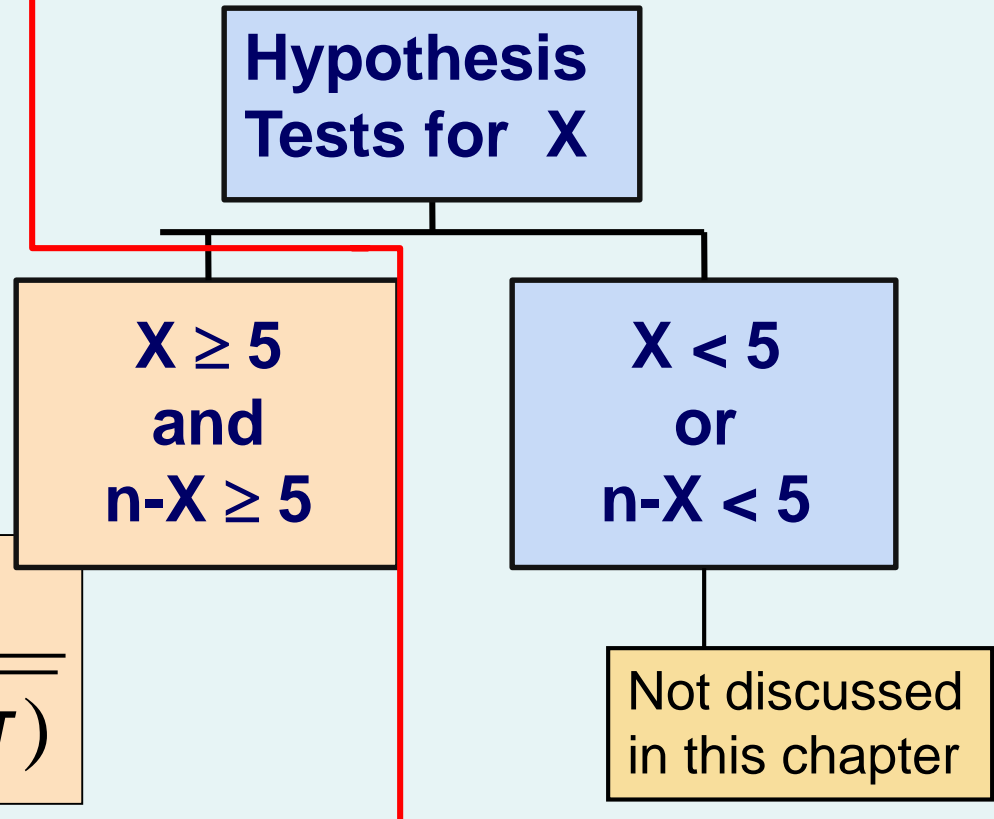
$$Z_{STAT} = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}}$$



Z Test for Proportion in Terms of Number in Category of Interest

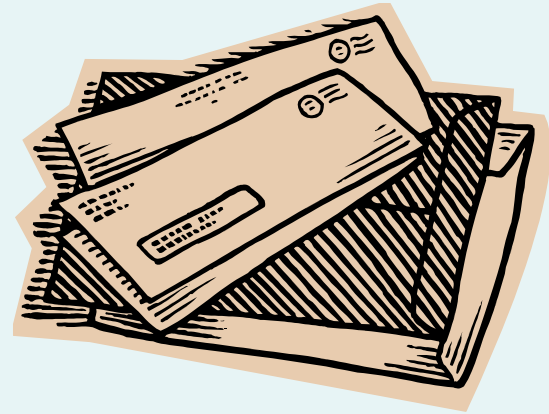
- An equivalent form to the last slide, but in terms of the number in the category of interest, X :

$$Z_{STAT} = \frac{X - n\pi}{\sqrt{n\pi(1 - \pi)}}$$



Example: Z Test for Proportion

A marketing company claims that it receives 8% responses from its mailing. To test this claim, a random sample of 500 were surveyed with 25 responses. Test at the $\alpha = 0.05$ significance level.



Check:

$$n\pi = (500)(.08) = 40$$

$$n(1-\pi) = (500)(.92) = 460$$



Z Test for Proportion: Solution

$$H_0: \pi = 0.08$$

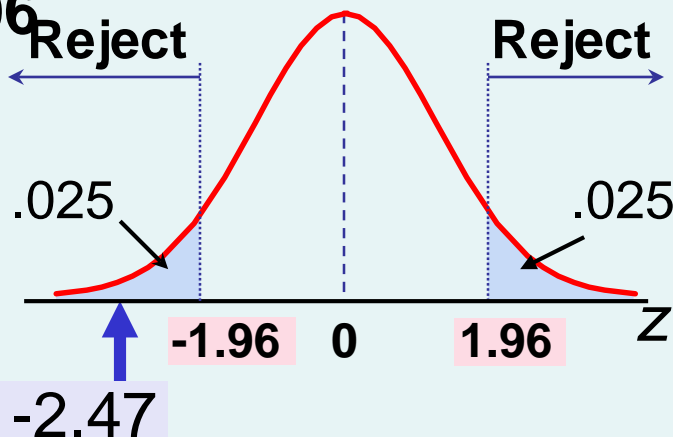
$$H_1: \pi \neq 0.08$$

$$\alpha = 0.05$$

$$n = 500, \quad p = 0.05$$

Critical Values: \pm

1.96



Test Statistic:

$$Z_{\text{STAT}} = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} = \frac{.05 - .08}{\sqrt{\frac{.08(1 - .08)}{500}}} = -2.47$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient evidence to reject the company's claim of 8% response rate.

p-Value Solution

(continued)

Calculate the p-value and compare to α
(For a two-tail test the p-value is always two-tail)



p-value = 0.0136:

$$P(Z \leq -2.47) + P(Z \geq 2.47) \\ = 2(0.0068) = 0.0136$$

Reject H_0 since p-value = 0.0136 < α = 0.05



Potential Pitfalls and Ethical Considerations

- Use randomly collected data to reduce selection biases
- Do not use human subjects without informed consent
- Choose the level of significance, α , and the type of test (one-tail or two-tail) before data collection
- Report all pertinent findings including both statistical significance and practical importance



Chapter Summary

- Addressed hypothesis testing methodology
- Performed Z Test for the mean (σ known)
- Discussed critical value and p-value approaches to hypothesis testing
- Performed one-tail and two-tail tests

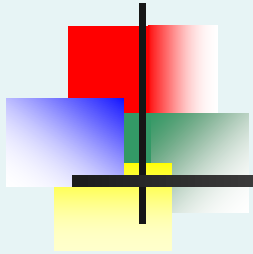


Chapter Summary

(continued)

- Performed t test for the mean (σ unknown)
- Performed Z test for the proportion
- Discussed pitfalls and ethical issues

Business Statistics: A First Course Fifth Edition



Chapter 10 Two-Sample Tests & One-Way ANOVA

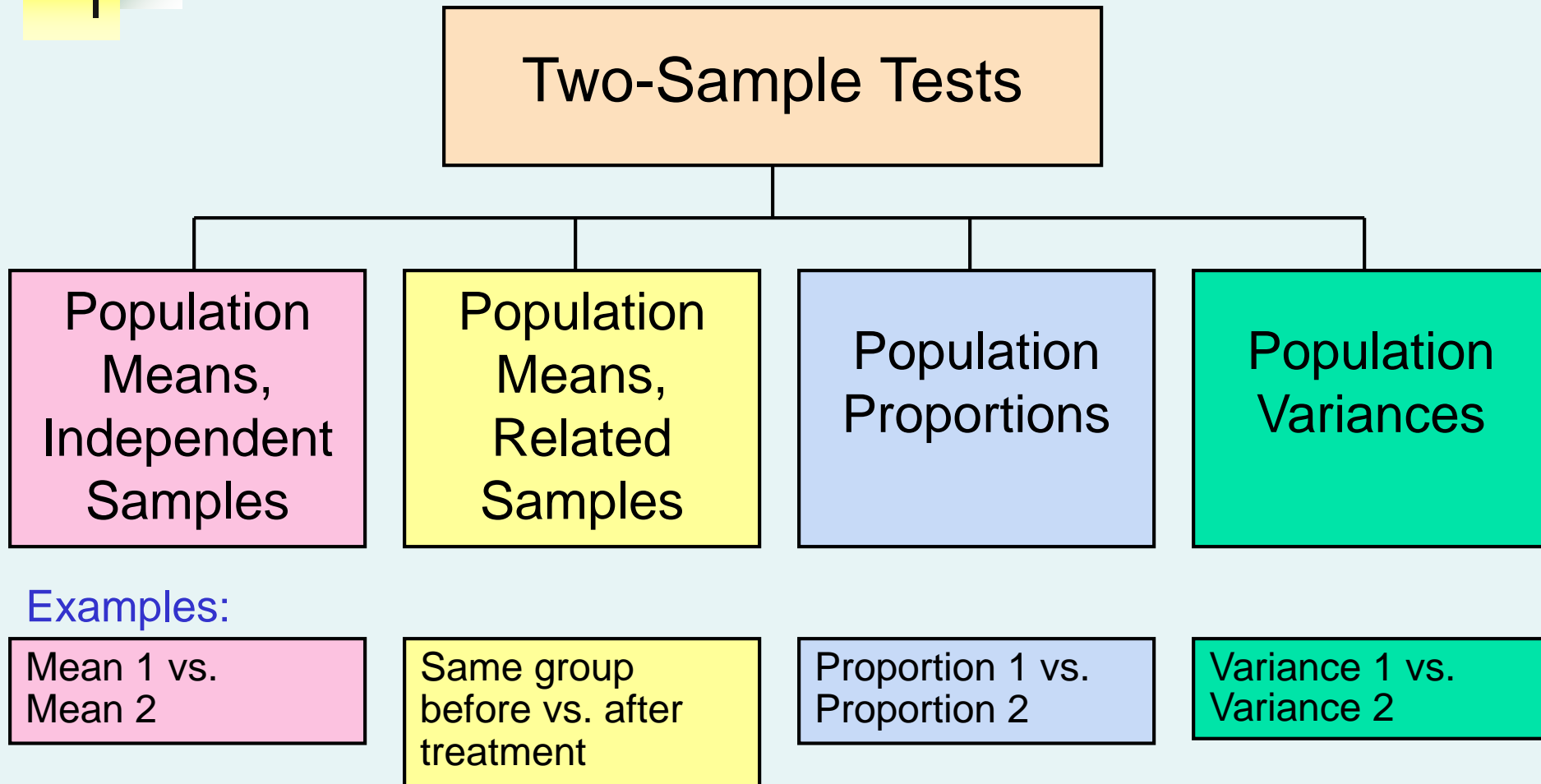


Learning Objectives

In this chapter, you learn:

- How to use hypothesis testing for comparing the difference between
 - The means of two independent populations
 - The means of two related populations
 - The proportions of two independent populations
 - The variances of two independent populations
- How to use one-way analysis of variance (ANOVA) to test for differences among the means of several populations
- How to perform multiple comparisons in a one-way analysis of variance .

Two-Sample Tests



Difference Between Two Means

Population means,
independent
samples

*

σ_1 and σ_2 unknown,
assumed equal

σ_1 and σ_2 unknown,
not assumed equal

Goal: Test hypothesis or form
a confidence interval for the
difference between two
population means, $\mu_1 - \mu_2$

The point estimate for the
difference is

$$\bar{X}_1 - \bar{X}_2$$

Difference Between Two Means: Independent Samples

- Different data sources

- Unrelated
- Independent
 - Sample selected from one population has no effect on the sample selected from the other population

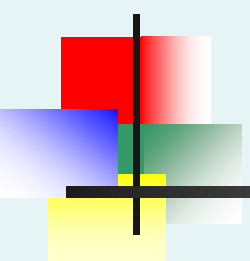
Population means,
independent
samples *

σ_1 and σ_2 unknown,
assumed equal

Use S_p to estimate unknown σ .
Use a **Pooled-Variance t test**.

σ_1 and σ_2 unknown,
not assumed equal

Use S_1 and S_2 to estimate
unknown σ_1 and σ_2 . Use a
Separate-variance t test



Hypothesis Tests for Two Population Means

Two Population Means, Independent Samples

Lower-tail test:

$$H_0: \mu_1 \geq \mu_2$$

$$H_1: \mu_1 < \mu_2$$

i.e.,

$$H_0: \mu_1 - \mu_2 \geq 0$$

$$H_1: \mu_1 - \mu_2 < 0$$

Upper-tail test:

$$H_0: \mu_1 \leq \mu_2$$

$$H_1: \mu_1 > \mu_2$$

i.e.,

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

Two-tail test:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

i.e.,

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

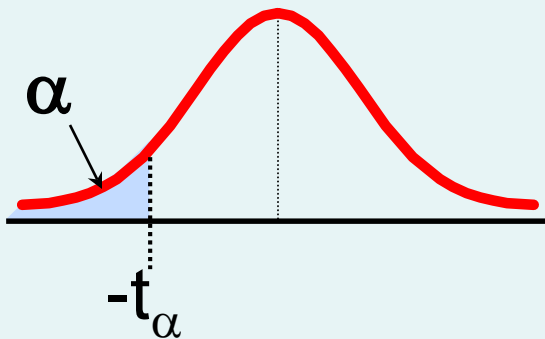
Hypothesis tests for $\mu_1 - \mu_2$

Two Population Means, Independent Samples

Lower-tail test:

$$H_0: \mu_1 - \mu_2 \geq 0$$

$$H_1: \mu_1 - \mu_2 < 0$$

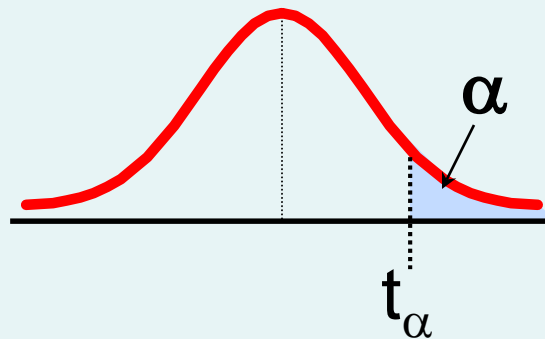


Reject H_0 if $t_{\text{STAT}} < -t_\alpha$

Upper-tail test:

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

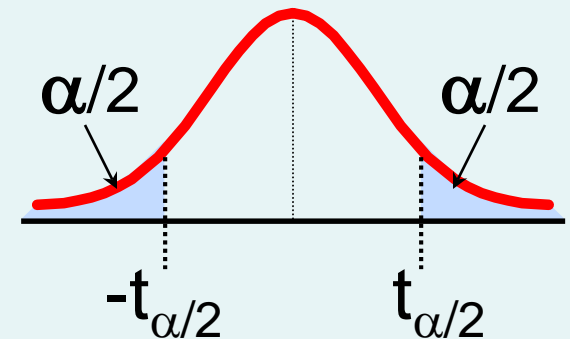


Reject H_0 if $t_{\text{STAT}} > t_\alpha$

Two-tail test:

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$



Reject H_0 if $t_{\text{STAT}} < -t_{\alpha/2}$
or $t_{\text{STAT}} > t_{\alpha/2}$

Hypothesis tests for $\mu_1 - \mu_2$ with σ_1 and σ_2 unknown and assumed equal

Population means,
independent
samples

σ_1 and σ_2 unknown,
assumed equal *

σ_1 and σ_2 unknown,
not assumed equal

Assumptions:

- Samples are randomly and independently drawn
- Populations are normally distributed or both sample sizes are at least 30
- Population variances are unknown but assumed equal

Hypothesis tests for $\mu_1 - \mu_2$ with σ_1 and σ_2 unknown and assumed equal

(continued)

Population means,
independent
samples

σ_1 and σ_2 unknown,
assumed equal

σ_1 and σ_2 unknown,
not assumed equal

- The pooled variance is:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

- The test statistic is:

$$t_{\text{STAT}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

- Where t_{STAT} has d.f. = $(n_1 + n_2 - 2)$

Confidence interval for $\mu_1 - \mu_2$ with σ_1 and σ_2 unknown and assumed equal

Population means,
independent
samples

σ_1 and σ_2 unknown,
assumed equal *

σ_1 and σ_2 unknown,
not assumed equal

The confidence interval for

$\mu_1 - \mu_2$ is:

$$\left(\bar{X}_1 - \bar{X}_2 \right) \pm t_{\alpha/2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

Where $t_{\alpha/2}$ has d.f. = $n_1 + n_2 - 2$

Pooled-Variance t Test Example

You are a financial analyst for a brokerage firm. Is there a difference in dividend yield between stocks listed on the NYSE & NASDAQ? You collect the following data:

| | <u>NYSE</u> | <u>NASDAQ</u> |
|-----------------------|-------------|---------------|
| Number | 21 | 25 |
| Sample mean | 3.27 | 2.53 |
| Sample std dev | 1.30 | 1.16 |

Assuming both populations are approximately normal with equal variances, is there a difference in mean yield ($\alpha = 0.05$)?



Pooled-Variance t Test Example: Calculating the Test Statistic

(continued)

$$H_0: \mu_1 - \mu_2 = 0 \text{ i.e. } (\mu_1 = \mu_2)$$

$$H_1: \mu_1 - \mu_2 \neq 0 \text{ i.e. } (\mu_1 \neq \mu_2)$$

The test statistic is:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{(3.27 - 2.53) - 0}{\sqrt{1.5021 \left(\frac{1}{21} + \frac{1}{25} \right)}} = 2.040$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(21 - 1)1.30^2 + (25 - 1)1.16^2}{(21 - 1) + (25 - 1)} = 1.5021$$

Pooled-Variance t Test Example: Hypothesis Test Solution

$$H_0: \mu_1 - \mu_2 = 0 \text{ i.e. } (\mu_1 = \mu_2)$$

$$H_1: \mu_1 - \mu_2 \neq 0 \text{ i.e. } (\mu_1 \neq \mu_2)$$

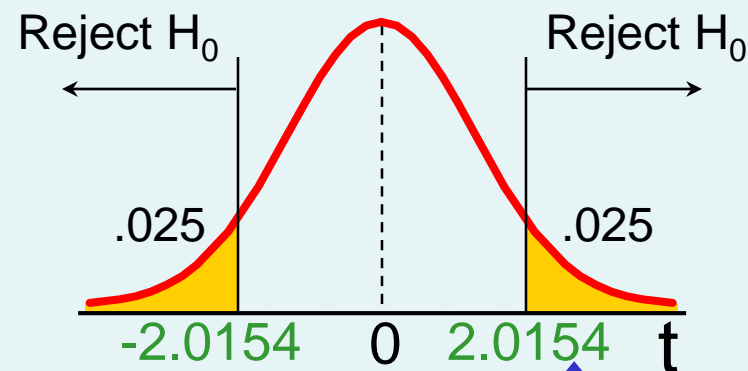
$$\alpha = 0.05$$

$$df = 21 + 25 - 2 = 44$$

$$\text{Critical Values: } t = \pm 2.0154$$

Test Statistic:

$$t = \frac{3.27 - 2.53}{\sqrt{1.5021 \left(\frac{1}{21} + \frac{1}{25} \right)}} = 2.040$$



2.040

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is evidence of a difference in means.



Pooled-Variance t Test Example: Confidence Interval for $\mu_1 - \mu_2$

Since we rejected H_0 can we be 95% confident that $\mu_{\text{NYSE}} > \mu_{\text{NASDAQ}}$?

95% Confidence Interval for $\mu_{\text{NYSE}} - \mu_{\text{NASDAQ}}$

$$\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{\alpha/2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 0.74 \pm 2.0154 \times 0.3628 = (0.09, 1.471)$$

Since 0 is less than the entire interval, we can be 95% confident that $\mu_{\text{NYSE}} > \mu_{\text{NASDAQ}}$

Hypothesis tests for $\mu_1 - \mu_2$ with σ_1 and σ_2 unknown, not assumed equal

Population means,
independent
samples

σ_1 and σ_2 unknown,
assumed equal

σ_1 and σ_2 unknown,
not assumed equal *

Assumptions:

- Samples are randomly and independently drawn
- Populations are normally distributed or both sample sizes are at least 30
- Population variances are unknown and cannot be assumed to be equal

Hypothesis tests for $\mu_1 - \mu_2$ with σ_1 and σ_2 unknown and not assumed equal

(continued)

Population means,
independent
samples

σ_1 and σ_2 unknown,
assumed equal

σ_1 and σ_2 unknown,
not assumed equal *

Excel or Minitab can
be used to perform
the appropriate
calculations

Related Populations

The Paired Difference Test

Related
samples

Tests Means of 2 **Related** Populations

- Paired or matched samples
- Repeated measures (before/after)
- Use **difference** between paired values:

$$D_i = X_{1i} - X_{2i}$$

- Eliminates Variation Among Subjects
- Assumptions:
 - Both Populations Are Normally Distributed
 - Or, if not Normal, use large samples

Related Populations

The Paired Difference Test

(continued)

Related
samples

The i^{th} paired difference is D_i , where

$$D_i = X_{1i} - X_{2i}$$

The point estimate for the
paired difference
population mean μ_D is \bar{D} :

$$\bar{D} = \frac{\sum_{i=1}^n D_i}{n}$$

The sample standard
deviation is S_D

$$S_D = \sqrt{\frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n-1}}$$

n is the number of pairs in the paired sample

The Paired Difference Test: Finding t_{STAT}



Paired
samples

- The test statistic for μ_D is:

$$t_{\text{STAT}} = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{n}}}$$

- Where t_{STAT} has $n - 1$ d.f.

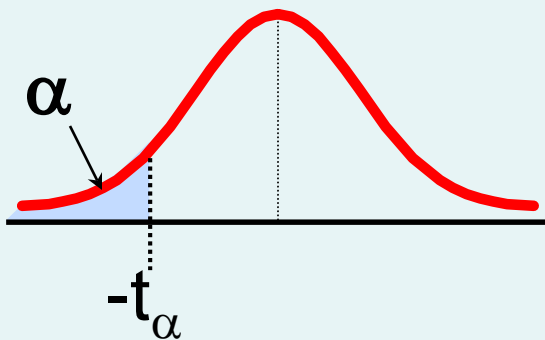
The Paired Difference Test: Possible Hypotheses

Paired Samples

Lower-tail test:

$$H_0: \mu_D \geq 0$$

$$H_1: \mu_D < 0$$

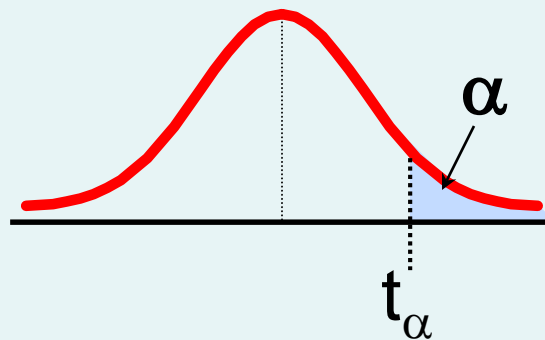


Reject H_0 if $t_{\text{STAT}} < -t_\alpha$

Upper-tail test:

$$H_0: \mu_D \leq 0$$

$$H_1: \mu_D > 0$$

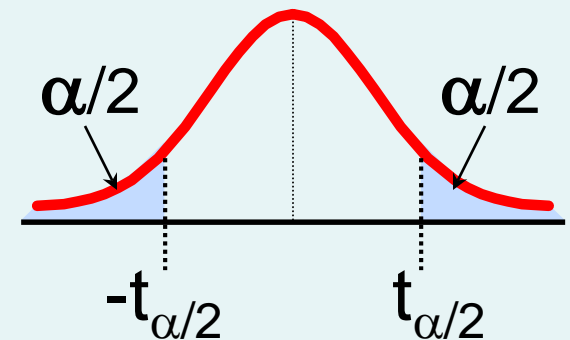


Reject H_0 if $t_{\text{STAT}} > t_\alpha$

Two-tail test:

$$H_0: \mu_D = 0$$

$$H_1: \mu_D \neq 0$$



Reject H_0 if $t_{\text{STAT}} < -t_{\alpha/2}$
or $t_{\text{STAT}} > t_{\alpha/2}$

Where t_{STAT} has $n - 1$ d.f.

The Paired Difference Confidence Interval

Paired
samples

The confidence interval for μ_D is

$$\bar{D} \pm t_{\alpha/2} \frac{S_D}{\sqrt{n}}$$

where $S_D = \sqrt{\frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n-1}}$

Paired Difference Test: Example

- Assume you send your salespeople to a “customer service” training workshop. Has the training made a difference in the number of complaints? You collect the following data:

| Salesperson | Number of Complaints: | | (2) - (1) Difference, D_i |
|-------------|-----------------------|-----------|--------------------------------|
| | Before (1) | After (2) | |
| C.B. | 6 | 4 | - 2 |
| T.F. | 20 | 6 | -14 |
| M.H. | 3 | 2 | - 1 |
| R.K. | 0 | 0 | 0 |
| M.O. | 4 | 0 | - 4 |
| | | | -21 |

$$\bar{D} = \frac{\sum D_i}{n}$$

$$= -4.2$$

$$S_D = \sqrt{\frac{\sum (D_i - \bar{D})^2}{n-1}}$$

$$= 5.67$$

Paired Difference Test: Solution

- Has the training made a difference in the number of complaints (at the 0.01 level)?

$$\begin{aligned} H_0: \mu_D &= 0 \\ H_1: \mu_D &\neq 0 \end{aligned}$$

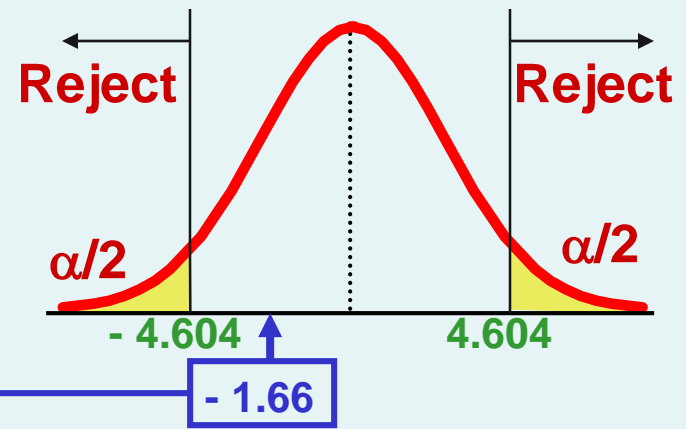
$$\alpha = .01 \quad \bar{D} = -4.2$$

$$t_{0.005} = \pm 4.604$$

d.f. = $n - 1 = 4$

Test Statistic:

$$t_{\text{STAT}} = \frac{\bar{D} - \mu_D}{S_D / \sqrt{n}} = \frac{-4.2 - 0}{5.67 / \sqrt{5}} = -1.66$$



Decision: Do not reject H_0
(t_{stat} is not in the reject region)

Conclusion: There is not a significant change in the number of complaints.



Two Population Proportions

Population proportions

Goal: test a hypothesis or form a confidence interval for the difference between two population proportions,

$$\pi_1 - \pi_2$$

Assumptions:

$$n_1 \pi_1 \geq 5 \quad , \quad n_1(1 - \pi_1) \geq 5$$

$$n_2 \pi_2 \geq 5 \quad , \quad n_2(1 - \pi_2) \geq 5$$

The point estimate for the difference is

$$p_1 - p_2$$



Two Population Proportions

Population proportions

In the null hypothesis we assume the null hypothesis is true, so we assume $\pi_1 = \pi_2$ and pool the two sample estimates

The pooled estimate for the overall proportion is:

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

where X_1 and X_2 are the number of items of interest in samples 1 and 2

Two Population Proportions

(continued)

Population proportions

The test statistic for $\pi_1 - \pi_2$ is a Z statistic:

$$Z_{\text{STAT}} = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\bar{p}(1 - \bar{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where $\bar{p} = \frac{X_1 + X_2}{n_1 + n_2}$, $p_1 = \frac{X_1}{n_1}$, $p_2 = \frac{X_2}{n_2}$



Hypothesis Tests for Two Population Proportions

Population proportions

Lower-tail test:

$$H_0: \pi_1 \geq \pi_2$$

$$H_1: \pi_1 < \pi_2$$

i.e.,

$$H_0: \pi_1 - \pi_2 \geq 0$$

$$H_1: \pi_1 - \pi_2 < 0$$

Upper-tail test:

$$H_0: \pi_1 \leq \pi_2$$

$$H_1: \pi_1 > \pi_2$$

i.e.,

$$H_0: \pi_1 - \pi_2 \leq 0$$

$$H_1: \pi_1 - \pi_2 > 0$$

Two-tail test:

$$H_0: \pi_1 = \pi_2$$

$$H_1: \pi_1 \neq \pi_2$$

i.e.,

$$H_0: \pi_1 - \pi_2 = 0$$

$$H_1: \pi_1 - \pi_2 \neq 0$$

Hypothesis Tests for Two Population Proportions

(continued)

Population proportions

Lower-tail test:

$$H_0: \pi_1 - \pi_2 \geq 0$$

$$H_1: \pi_1 - \pi_2 < 0$$

Upper-tail test:

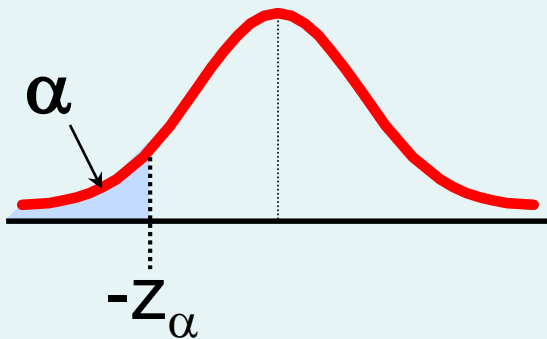
$$H_0: \pi_1 - \pi_2 \leq 0$$

$$H_1: \pi_1 - \pi_2 > 0$$

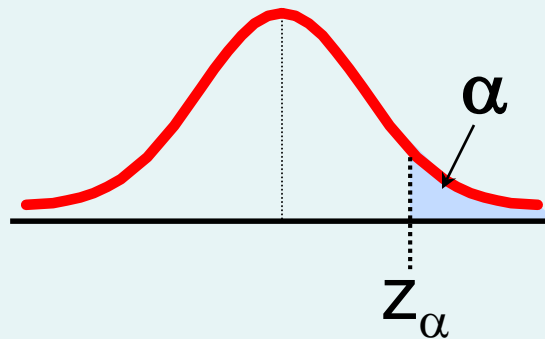
Two-tail test:

$$H_0: \pi_1 - \pi_2 = 0$$

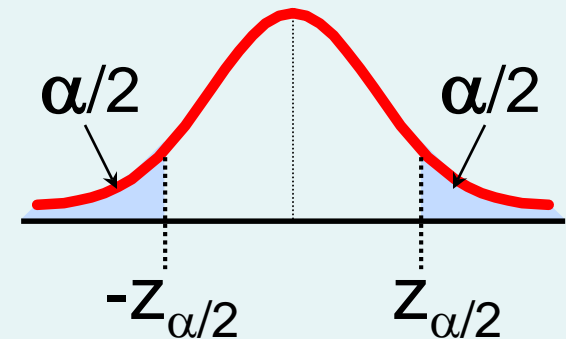
$$H_1: \pi_1 - \pi_2 \neq 0$$



Reject H_0 if $Z_{\text{STAT}} < -Z_\alpha$



Reject H_0 if $Z_{\text{STAT}} > Z_\alpha$



Reject H_0 if $Z_{\text{STAT}} < -Z_{\alpha/2}$
or $Z_{\text{STAT}} > Z_{\alpha/2}$

Hypothesis Test Example: Two population Proportions

Is there a significant difference between the proportion of men and the proportion of women who will vote Yes on Proposition A?

- In a random sample, 36 of 72 men and 31 of 50 women indicated they would vote Yes
- Test at the .05 level of significance



Hypothesis Test Example: Two population Proportions

(continued)

- The hypothesis test is:

$H_0: \pi_1 - \pi_2 = 0$ (the two proportions are equal)

$H_1: \pi_1 - \pi_2 \neq 0$ (there is a significant difference between proportions)

- The sample proportions are:

■ Men: $p_1 = 36/72 = .50$

■ Women: $p_2 = 31/50 = .62$

- The pooled estimate for the overall proportion is:

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{36 + 31}{72 + 50} = \frac{67}{122} = .549$$

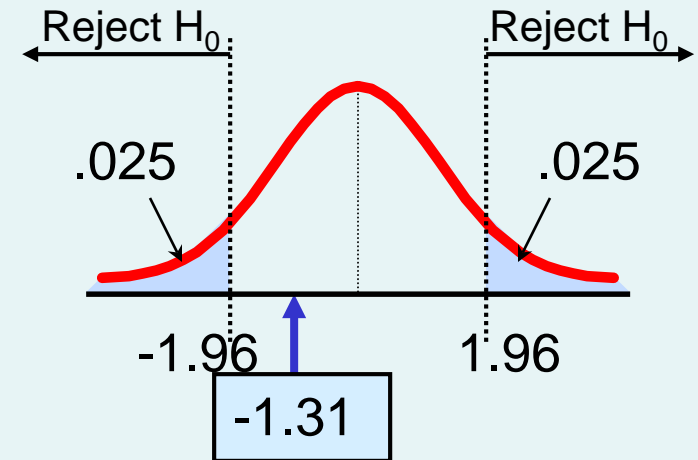
Hypothesis Test Example: Two population Proportions

(continued)

The test statistic for $\pi_1 - \pi_2$ is:

$$Z_{STAT} = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(.50 - .62) - (0)}{\sqrt{.549(1-.549)\left(\frac{1}{72} + \frac{1}{50}\right)}} = -1.31$$

Critical Values =
 ± 1.96
For $\alpha = .05$



Decision: Do not reject H₀

Conclusion: There is not significant evidence of a difference in proportions who will vote yes between men and women.



Confidence Interval for Two Population Proportions

Population proportions

The confidence interval for

$\pi_1 - \pi_2$ is:

$$(p_1 - p_2) \pm Z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Hypothesis Tests for Variances

Tests for Two
Population
Variances

*

F test statistic

Hypotheses

F_{STAT}

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

$$H_0: \sigma_1^2 \leq \sigma_2^2$$

$$H_1: \sigma_1^2 > \sigma_2^2$$

$$S_1^2 / S_2^2$$

Where:

S_1^2 = Variance of sample 1 (the larger sample variance)

n_1 = sample size of sample 1

S_2^2 = Variance of sample 2 (the smaller sample variance)

n_2 = sample size of sample 2

$n_1 - 1$ = numerator degrees of freedom

$n_2 - 1$ = denominator degrees of freedom



The F Distribution

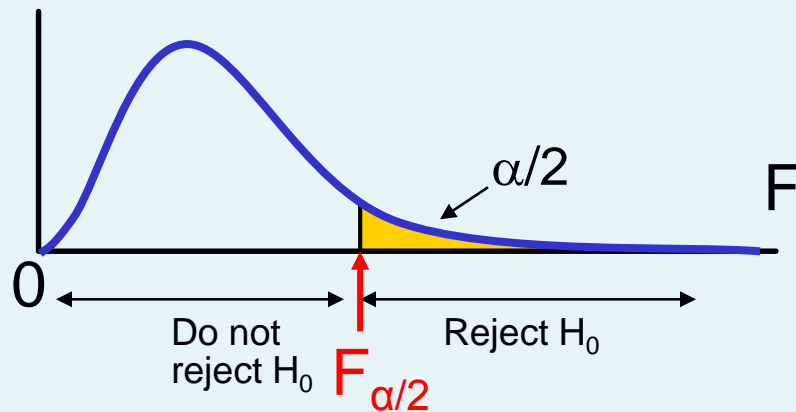
- The F critical value is found from the F table
- There are two degrees of freedom required: numerator and denominator

- When $F_{STAT} = \frac{S_1^2}{S_2^2}$ $df_1 = n_1 - 1$; $df_2 = n_2 - 1$

- In the F table,
 - numerator degrees of freedom determine the column
 - denominator degrees of freedom determine the row

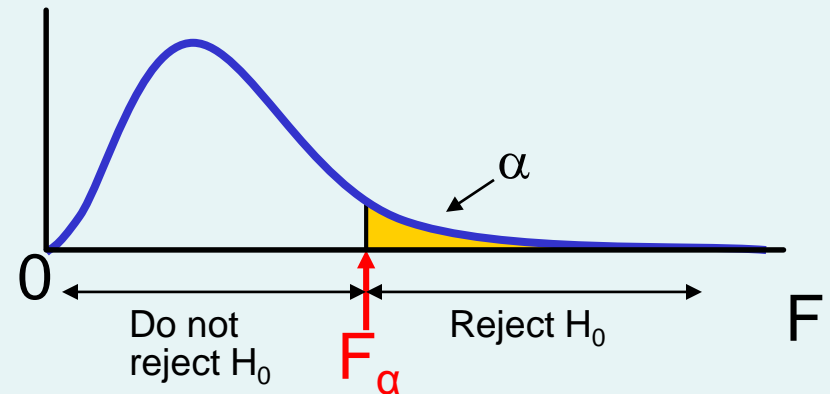
Finding the Rejection Region

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_1: \sigma_1^2 \neq \sigma_2^2$$



Reject H_0 if $F_{\text{STAT}} > F_{\alpha/2}$

$$H_0: \sigma_1^2 \leq \sigma_2^2$$
$$H_1: \sigma_1^2 > \sigma_2^2$$



Reject H_0 if $F_{\text{STAT}} > F_{\alpha}$

F Test: An Example

You are a financial analyst for a brokerage firm. You want to compare dividend yields between stocks listed on the NYSE & NASDAQ. You collect the following data:

| | <u>NYSE</u> | <u>NASDAQ</u> |
|----------------|-------------|---------------|
| Number | 21 | 25 |
| Mean | 3.27 | 2.53 |
| Std dev | 1.30 | 1.16 |

Is there a difference in the variances between the NYSE & NASDAQ at the $\alpha = 0.05$ level?





F Test: Example Solution

- Form the hypothesis test:

$H_0: \sigma^2_1 = \sigma^2_2$ (there is no difference between variances)

$H_1: \sigma^2_1 \neq \sigma^2_2$ (there is a difference between variances)

- Find the F critical value for $\alpha = 0.05$:
- Numerator d.f. = $n_1 - 1 = 21 - 1 = 20$
- Denominator d.f. = $n_2 - 1 = 25 - 1 = 24$

- $F_{\alpha/2} = F_{.025, 20, 24} = 2.33$

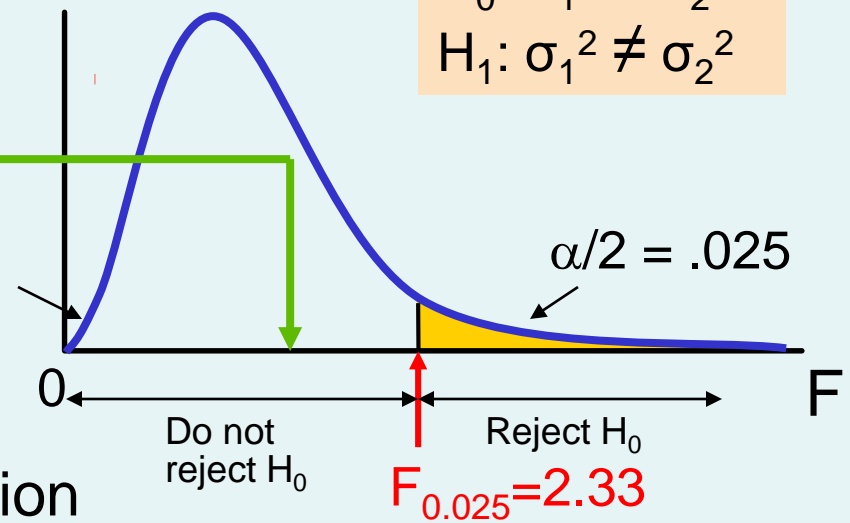
F Test: Example Solution

(continued)

- The test statistic is:

$$F_{STAT} = \frac{S_1^2}{S_2^2} = \frac{1.30^2}{1.16^2} = 1.256$$

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_1: \sigma_1^2 \neq \sigma_2^2$$



- $F_{STAT} = 1.256$ is not in the rejection region, so we **do not reject H_0**
- Conclusion:** There is not sufficient evidence of a difference in variances at $\alpha = .05$



General ANOVA Setting

- Investigator controls one or more factors of interest
 - Each factor contains two or more levels
 - Levels can be numerical or categorical
 - Different levels produce different groups
 - Think of each group as a sample from a different population
- Observe effects on the dependent variable
 - Are the groups the same?
- Experimental design: the plan used to collect the data



Completely Randomized Design

- Experimental units (subjects) are assigned randomly to groups
 - Subjects are assumed homogeneous
- Only one factor or independent variable
 - With two or more levels
- Analyzed by one-factor analysis of variance (ANOVA)



One-Way Analysis of Variance

- Evaluate the difference among the means of three or more groups

Examples: Accident rates for 1st, 2nd, and 3rd shift
Expected mileage for five brands of tires

- **Assumptions**
 - Populations are normally distributed
 - Populations have equal variances
 - Samples are randomly and independently drawn



Hypotheses of One-Way ANOVA

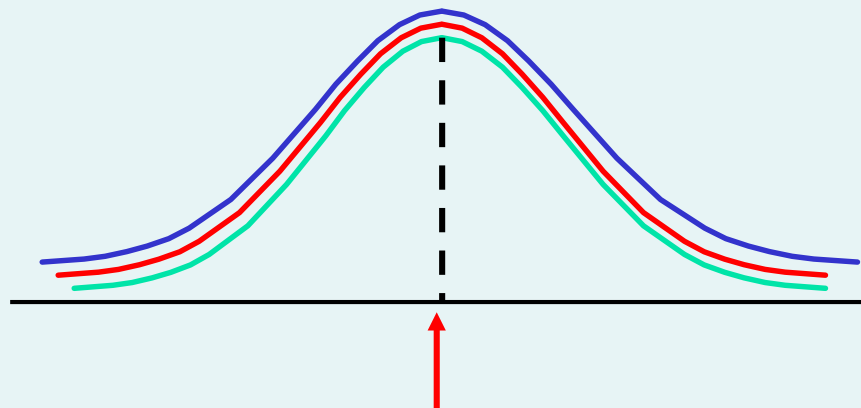
- $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_c$
 - All population means are equal
 - i.e., no factor effect (no variation in means among groups)
- H_1 : Not all of the population means are the same
 - At least one population mean is different
 - i.e., there is a factor effect
 - Does not mean that all population means are different (some pairs may be the same)

One-Way ANOVA

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_c$$

H_1 : Not all μ_j are the same

The Null Hypothesis is True
All Means are the same:
(No Factor Effect)



$$\mu_1 = \mu_2 = \mu_3$$

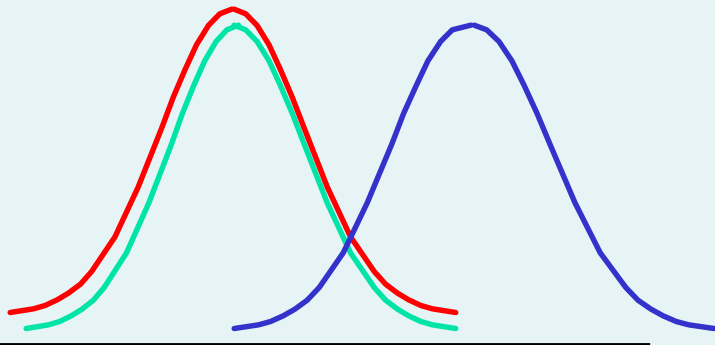
One-Way ANOVA

(continued)

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_c$$

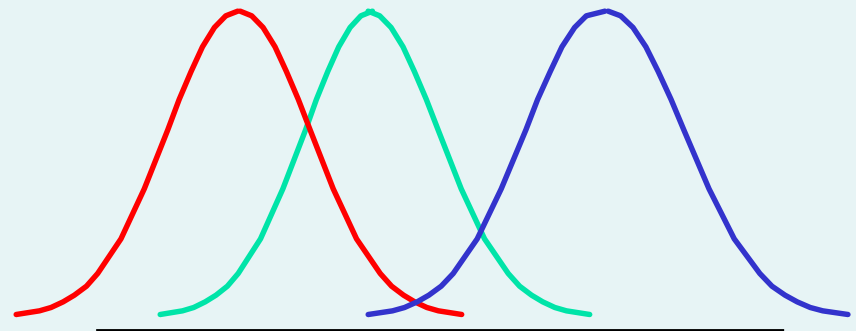
H_1 : Not all μ_j are the same

The Null Hypothesis is NOT true
At least one of the means is different
(Factor Effect is present)



$$\mu_1 = \mu_2 \neq \mu_3$$

or



$$\mu_1 \neq \mu_2 \neq \mu_3$$



Partitioning the Variation

- Total variation can be split into two parts:

$$SST = SSA + SSW$$

SST = Total Sum of Squares
(Total variation)

SSA = Sum of Squares Among Groups
(Among-group variation)

SSW = Sum of Squares Within Groups
(Within-group variation)



Partitioning the Variation

(continued)

$$SST = SSA + SSW$$

Total Variation = the aggregate variation of the individual data values across the various factor levels (SST)

Among-Group Variation = variation among the factor sample means (SSA)

Within-Group Variation = variation that exists among the data values within a particular factor level (SSW)

Partition of Total Variation



Total Variation (SST)

**Variation Due to
Factor (SSA)**

+

**Variation Due to Random
Error (SSW)**

=



Total Sum of Squares

$$\text{SST} = \text{SSA} + \text{SSW}$$

$$\text{SST} = \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2$$

Where:

SST = Total sum of squares

c = number of groups or levels

n_j = number of observations in group j

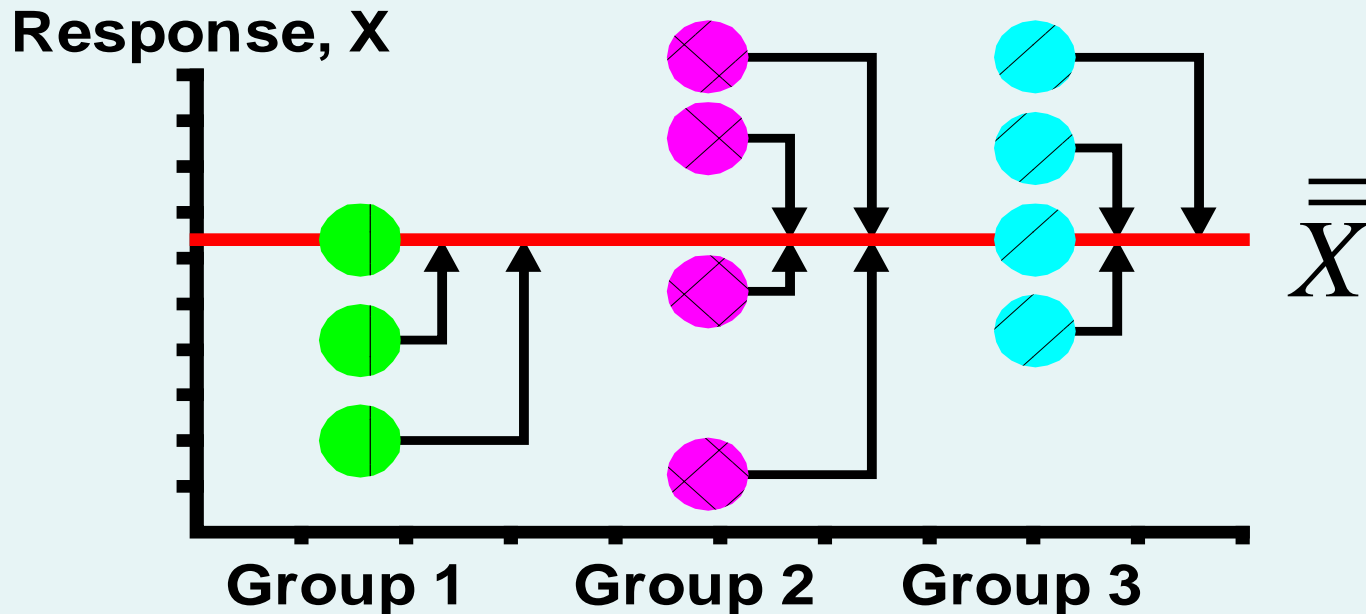
X_{ij} = i^{th} observation from group j

\bar{X} = grand mean (mean of all data values)

Total Variation

(continued)

$$SST = (X_{11} - \bar{X})^2 + (X_{12} - \bar{X})^2 + \dots + (X_{cn_j} - \bar{X})^2$$





Among-Group Variation

$$SST = SSA + SSW$$

$$SSA = \sum_{j=1}^c n_j (\bar{X}_j - \bar{X})^2$$

Where:

SSA = Sum of squares among groups

c = number of groups

n_j = sample size from group j

\bar{X}_j = sample mean from group j

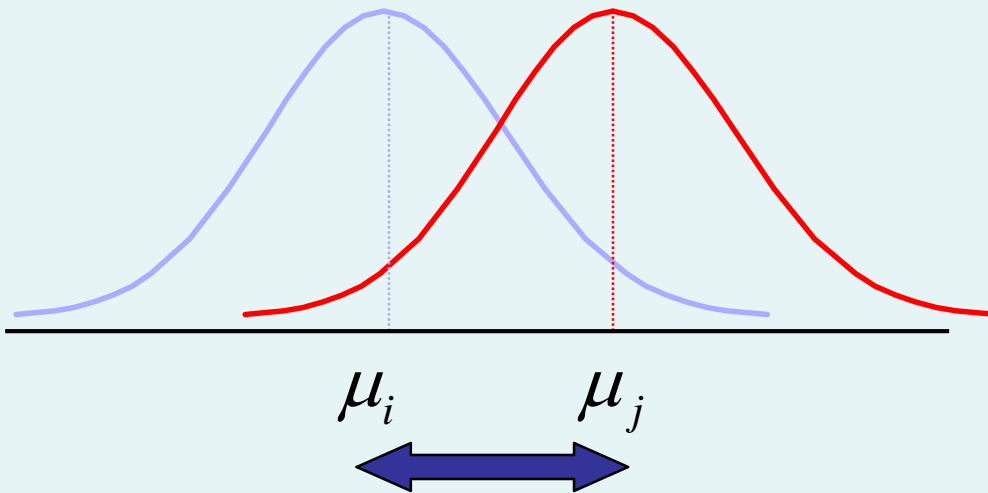
\bar{X} = grand mean (mean of all data values)

Among-Group Variation

(continued)

$$SSA = \sum_{j=1}^c n_j (\bar{X}_j - \bar{X})^2$$

Variation Due to
Differences Among Groups



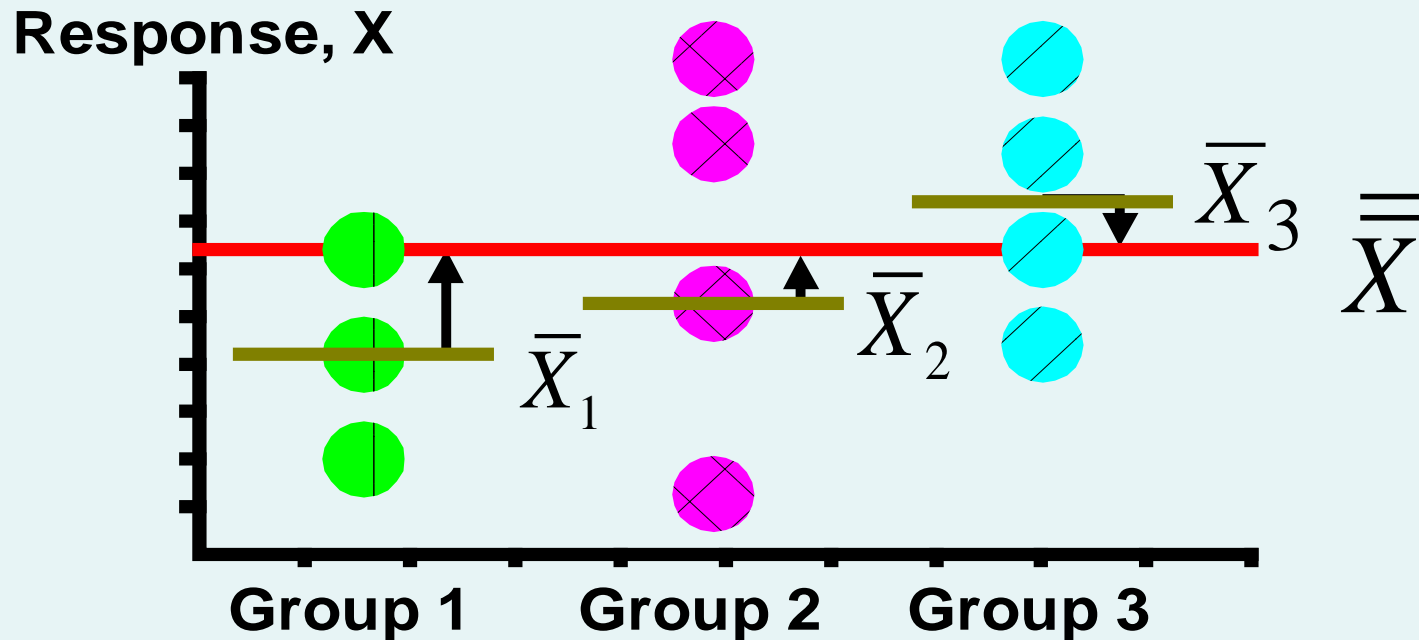
$$MSA = \frac{SSA}{c - 1}$$

Mean Square Among =
SSA/degrees of freedom

Among-Group Variation

(continued)

$$SSA = n_1(\bar{X}_1 - \bar{\bar{X}})^2 + n_2(\bar{X}_2 - \bar{\bar{X}})^2 + \cdots + n_c(\bar{X}_c - \bar{\bar{X}})^2$$





Within-Group Variation

$$SST = SSA + SSW$$

$$SSW = \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$$

Where:

SSW = Sum of squares within groups

c = number of groups

n_j = sample size from group j

\bar{X}_j = sample mean from group j

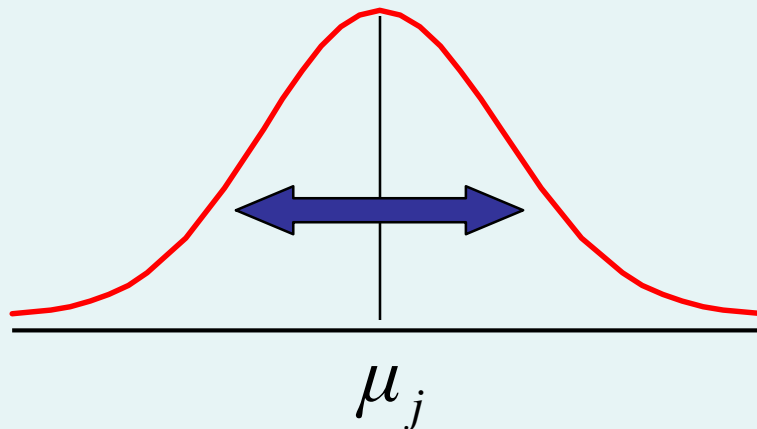
X_{ij} = i^{th} observation in group j

Within-Group Variation

(continued)

$$SSW = \sum_{j=1}^c \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$$

Summing the variation within each group and then adding over all groups



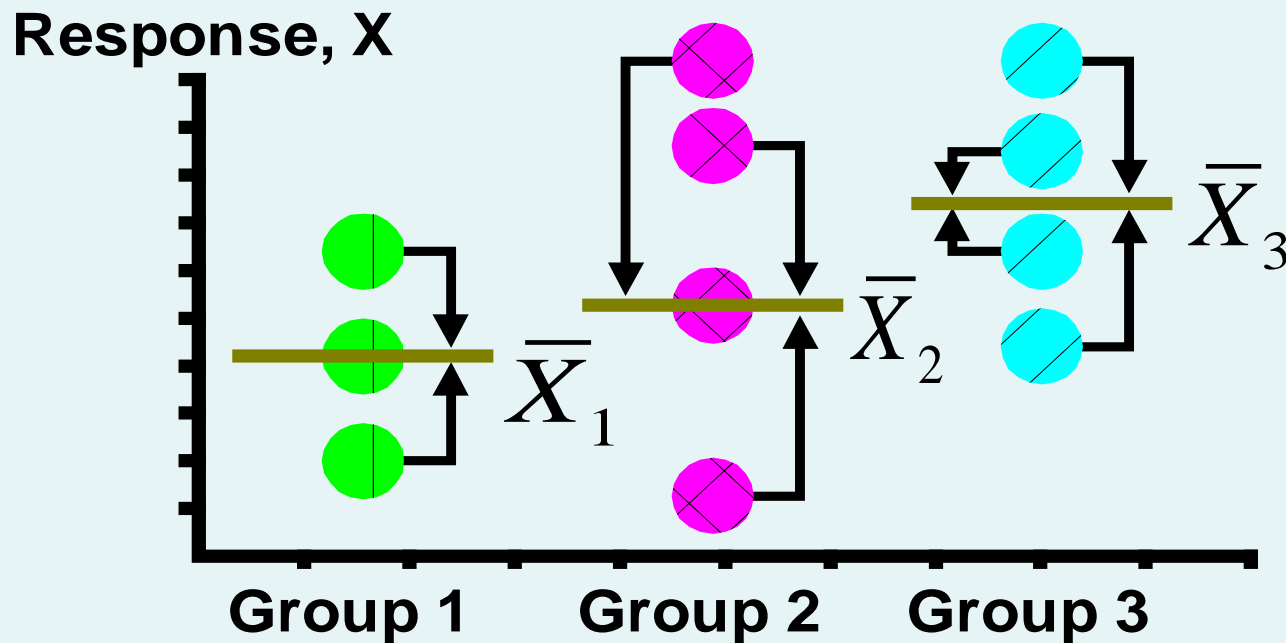
$$MSW = \frac{SSW}{n - c}$$

Mean Square Within =
SSW/degrees of freedom

Within-Group Variation

(continued)

$$SSW = (X_{11} - \bar{X}_1)^2 + (X_{12} - \bar{X}_2)^2 + \cdots + (X_{cn_j} - \bar{X}_c)^2$$





Obtaining the Mean Squares

The Mean Squares are obtained by dividing the various sum of squares by their associated degrees of freedom

$$MSA = \frac{SSA}{c - 1}$$

Mean Square Among
(d.f. = $c-1$)

$$MSW = \frac{SSW}{n - c}$$

Mean Square Within
(d.f. = $n-c$)

$$MST = \frac{SST}{n - 1}$$

Mean Square Total
(d.f. = $n-1$)

One-Way ANOVA Table

| Source of Variation | Degrees of Freedom | Sum Of Squares | Mean Square (Variance) | F |
|---------------------|--------------------|----------------|---------------------------|------------------------------|
| Among Groups | $c - 1$ | SSA | $MSA = \frac{SSA}{c - 1}$ | $F_{STAT} = \frac{MSA}{MSW}$ |
| Within Groups | $n - c$ | SSW | $MSW = \frac{SSW}{n - c}$ | |
| Total | $n - 1$ | SST | | |

c = number of groups

n = sum of the sample sizes from all groups

df = degrees of freedom

One-Way ANOVA

F Test Statistic

$$H_0: \mu_1 = \mu_2 = \dots = \mu_c$$

H_1 : At least two population means are different

- Test statistic

$$F_{STAT} = \frac{MSA}{MSW}$$

MSA is mean squares **among** groups

MSW is mean squares **within** groups

- Degrees of freedom

- $df_1 = c - 1$ (c = number of groups)

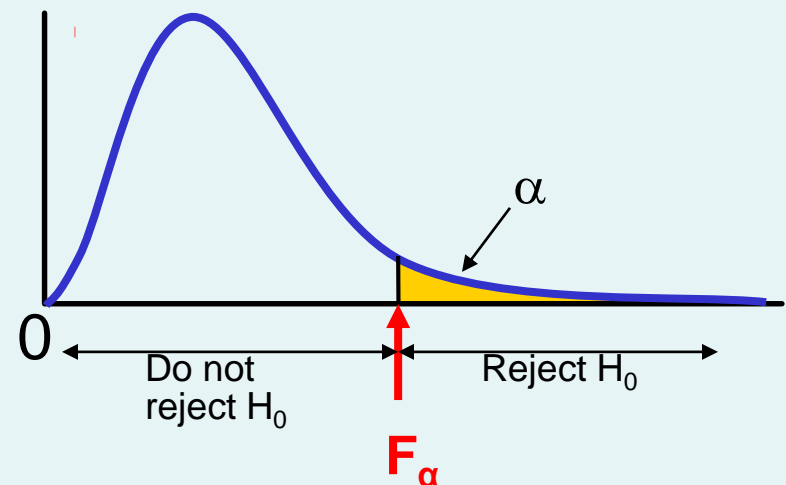
- $df_2 = n - c$ (n = sum of sample sizes from all populations)

Interpreting One-Way ANOVA F Statistic

- The F statistic is the ratio of the **among** estimate of variance and the **within** estimate of variance
 - The ratio must always be positive
 - $df_1 = c - 1$ will typically be small
 - $df_2 = n - c$ will typically be large

Decision Rule:

- Reject H_0 if $F_{STAT} > F_{\alpha}$, otherwise do not reject H_0



One-Way ANOVA F Test Example

You want to see if three different golf clubs yield different distances. You randomly select five measurements from trials on an automated driving machine for each club. At the 0.05 significance level, is there a difference in mean distance?

| Club 1 | Club 2 | Club 3 |
|---------------|---------------|---------------|
| 254 | 234 | 200 |
| 263 | 218 | 222 |
| 241 | 235 | 197 |
| 237 | 227 | 206 |
| 251 | 216 | 204 |



One-Way ANOVA Example: Scatter Plot

| Club 1 | Club 2 | Club 3 |
|--------|--------|--------|
| 254 | 234 | 200 |
| 263 | 218 | 222 |
| 241 | 235 | 197 |
| 237 | 227 | 206 |
| 251 | 216 | 204 |

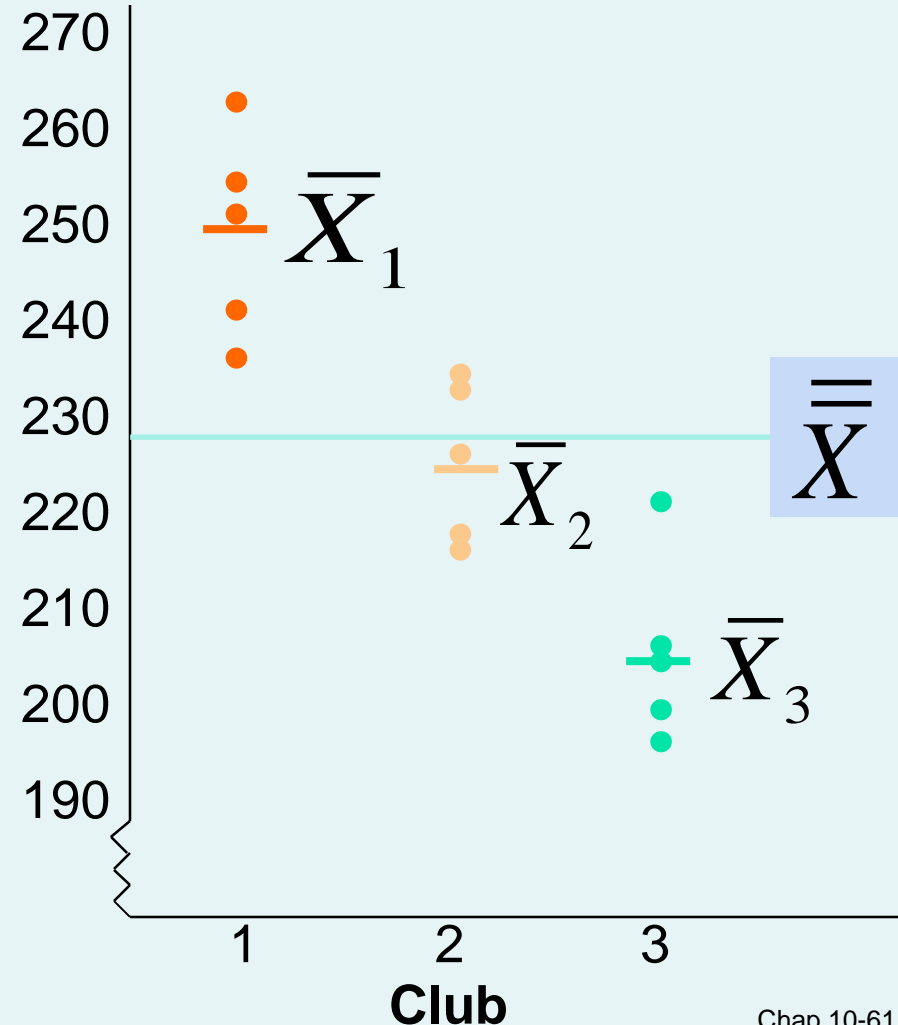


| | | |
|---------------------|---------------------|---------------------|
| $\bar{x}_1 = 249.2$ | $\bar{x}_2 = 226.0$ | $\bar{x}_3 = 205.8$ |
|---------------------|---------------------|---------------------|

| |
|-------------------------|
| $\bar{\bar{x}} = 227.0$ |
|-------------------------|

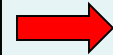


Distance



One-Way ANOVA Example Computations

| Club 1 | Club 2 | Club 3 |
|---------------|---------------|---------------|
| 254 | 234 | 200 |
| 263 | 218 | 222 |
| 241 | 235 | 197 |
| 237 | 227 | 206 |
| 251 | 216 | 204 |



| | |
|-------------------------|-----------|
| $\bar{X}_1 = 249.2$ | $n_1 = 5$ |
| $\bar{X}_2 = 226.0$ | $n_2 = 5$ |
| $\bar{X}_3 = 205.8$ | $n_3 = 5$ |
| $\bar{\bar{X}} = 227.0$ | $n = 15$ |
| | $c = 3$ |



$$SSA = 5 (249.2 - 227)^2 + 5 (226 - 227)^2 + 5 (205.8 - 227)^2 = 4716.4$$

$$SSW = (254 - 249.2)^2 + (263 - 249.2)^2 + \dots + (204 - 205.8)^2 = 1119.6$$

$$MSA = 4716.4 / (3-1) = 2358.2$$

$$MSW = 1119.6 / (15-3) = 93.3$$

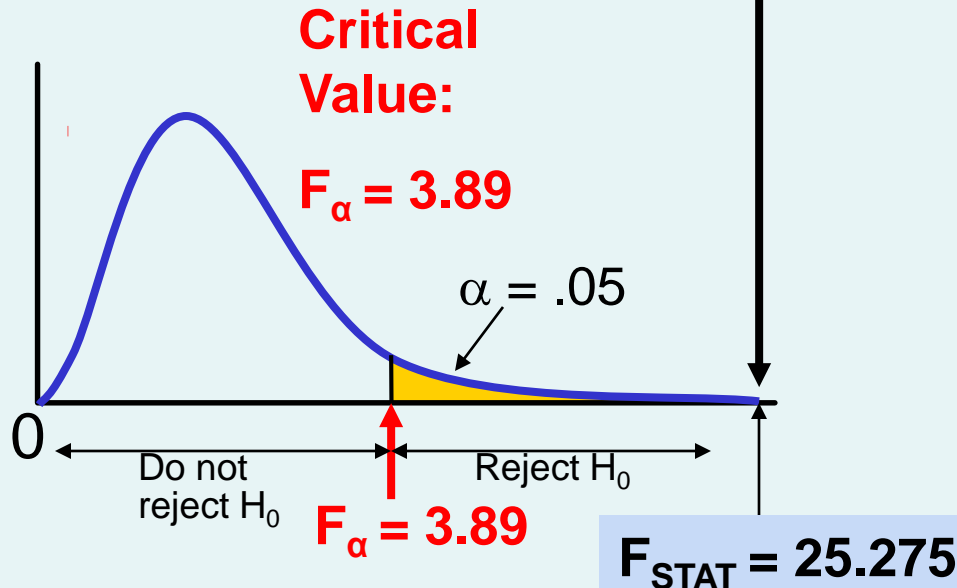
$$F_{STAT} = \frac{2358.2}{93.3} = 25.275$$

One-Way ANOVA Example Solution

$H_0: \mu_1 = \mu_2 = \mu_3$
 $H_1: \mu_j$ not all equal

$\alpha = 0.05$

$df_1 = 2$ $df_2 = 12$



Test Statistic:

$$F_{STAT} = \frac{MSA}{MSW} = \frac{2358.2}{93.3} = 25.275$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is evidence that at least one μ_j differs from the rest

One-Way ANOVA Excel Output

| SUMMARY | | | | | | |
|----------------------------|--------------|------------|----------------|-----------------|----------------|---------------|
| <i>Groups</i> | <i>Count</i> | <i>Sum</i> | <i>Average</i> | <i>Variance</i> | | |
| Club 1 | 5 | 1246 | 249.2 | 108.2 | | |
| Club 2 | 5 | 1130 | 226 | 77.5 | | |
| Club 3 | 5 | 1029 | 205.8 | 94.2 | | |
| ANOVA | | | | | | |
| <i>Source of Variation</i> | <i>SS</i> | <i>df</i> | <i>MS</i> | <i>F</i> | <i>P-value</i> | <i>F crit</i> |
| Between Groups | 4716.4 | 2 | 2358.2 | 25.275 | 4.99E-05 | 3.89 |
| Within Groups | 1119.6 | 12 | 93.3 | | | |
| Total | 5836.0 | 14 | | | | |



One-Way ANOVA

Minitab Output

One-way ANOVA: Distance versus Club

| Source | DF | SS | MS | F | P |
|--------|----|--------|--------|-------|-------|
| Club | 2 | 4716.4 | 2358.2 | 25.28 | 0.000 |
| Error | 12 | 1119.6 | 93.3 | | |
| Total | 14 | 5836.0 | | | |

S = 9.659 R-Sq = 80.82% R-Sq(adj) = 77.62%

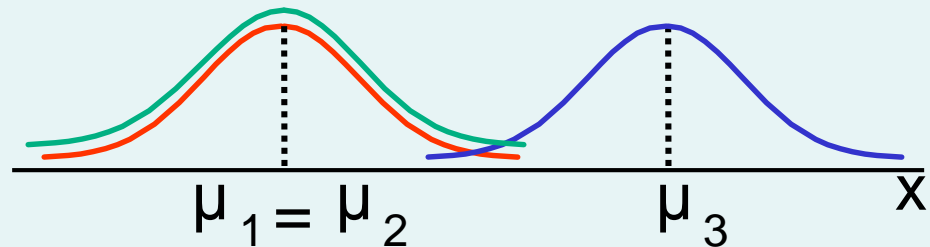
Individual 95% CIs For Mean Based on Pooled StDev

| Level | N | Mean | StDev | -----+-----+-----+-----+--- |
|-------|---|--------|-------|-----------------------------|
| 1 | 5 | 249.20 | 10.40 | (-----*-----) |
| 2 | 5 | 226.00 | 8.80 | (-----*-----) |
| 3 | 5 | 205.80 | 9.71 | (-----*-----) |
| | | | | -----+-----+-----+-----+--- |
| | | | | 208 224 240 256 |

Pooled StDev = 9.66

The Tukey-Kramer Procedure

- Tells **which** population means are significantly different
 - e.g.: $\mu_1 = \mu_2 \neq \mu_3$
 - Done after rejection of equal means in ANOVA
- Allows paired comparisons
 - Compare absolute mean differences with critical range





Tukey-Kramer Critical Range

$$\text{Critical Range} = Q_{\alpha} \sqrt{\frac{\text{MSW}}{2} \left(\frac{1}{n_j} + \frac{1}{n_{j'}} \right)}$$

where:

Q_{α} = Upper Tail Critical Value from Studentized Range Distribution with c and $n - c$ degrees of freedom (see appendix E.8 table)

MSW = Mean Square Within

n_j and $n_{j'}$ = Sample sizes from groups j and j'

The Tukey-Kramer Procedure: Example

| <u>Club 1</u> | <u>Club 2</u> | <u>Club 3</u> |
|---------------|---------------|---------------|
| 254 | 234 | 200 |
| 263 | 218 | 222 |
| 241 | 235 | 197 |
| 237 | 227 | 206 |
| 251 | 216 | 204 |

1. Compute absolute mean differences:

$$|\bar{x}_1 - \bar{x}_2| = |249.2 - 226.0| = 23.2$$

$$|\bar{x}_1 - \bar{x}_3| = |249.2 - 205.8| = 43.4$$

$$|\bar{x}_2 - \bar{x}_3| = |226.0 - 205.8| = 20.2$$

2. Find the Q_α value from the table in appendix E.8 with $c = 3$ and $(n - c) = (15 - 3) = 12$ degrees of freedom:

$$Q_\alpha = 3.77$$



The Tukey-Kramer Procedure: Example

(continued)

3. Compute Critical Range:

$$\text{Critical Range} = Q_{\alpha} \sqrt{\frac{\text{MSW}}{2} \left(\frac{1}{n_j} + \frac{1}{n_{j'}} \right)} = 3.77 \sqrt{\frac{93.3}{2} \left(\frac{1}{5} + \frac{1}{5} \right)} = 16.285$$

4. Compare:

5. All of the absolute mean differences are greater than critical range. Therefore there is a significant difference between each pair of means at 5% level of significance.

Thus, with 95% confidence we can conclude that the mean distance for club 1 is greater than club 2 and 3, and club 2 is greater than club 3.

$$|\bar{x}_1 - \bar{x}_2| = 23.2$$

$$|\bar{x}_1 - \bar{x}_3| = 43.4$$

$$|\bar{x}_2 - \bar{x}_3| = 20.2$$





ANOVA Assumptions

- Randomness and Independence
 - Select random samples from the c groups (or randomly assign the levels)
- Normality
 - The sample values for each group are from a normal population
- Homogeneity of Variance
 - All populations sampled from have the same variance
 - Can be tested with Levene's Test

ANOVA Assumptions

Levene's Test

- Tests the assumption that the variances of each population are equal.
- First, define the null and alternative hypotheses:
 - $H_0: \sigma^2_1 = \sigma^2_2 = \dots = \sigma^2_c$
 - $H_1: \text{Not all } \sigma^2_j \text{ are equal}$
- Second, compute the absolute value of the difference between each value and the median of each group.
- Third, perform a one-way ANOVA on these absolute differences.

Levene Homogeneity Of Variance Test Example

$$H_0: \sigma^2_1 = \sigma^2_2 = \sigma^2_3$$

H1: Not all σ^2_j are equal

Calculate Medians

| Club 1 | Club 2 | Club 3 | |
|--------|--------|--------|---------------|
| 237 | 216 | 197 | |
| 241 | 218 | 200 | |
| 251 | 227 | 204 | Median |
| 254 | 234 | 206 | |
| 263 | 235 | 222 | |

Calculate Absolute Differences

| Club 1 | Club 2 | Club 3 |
|--------|--------|--------|
| 14 | 11 | 7 |
| 10 | 9 | 4 |
| 0 | 0 | 0 |
| 3 | 7 | 2 |
| 12 | 8 | 18 |

Levene Homogeneity Of Variance Test Example

(continued)

Anova: Single Factor

SUMMARY

| <i>Groups</i> | <i>Count</i> | <i>Sum</i> | <i>Average</i> | <i>Variance</i> |
|---------------|--------------|------------|----------------|-----------------|
| Club 1 | 5 | 39 | 7.8 | 36.2 |
| Club 2 | 5 | 35 | 7 | 17.5 |
| Club 3 | 5 | 31 | 6.2 | 50.2 |

| <i>Source of Variation</i> | <i>SS</i> | <i>df</i> | <i>MS</i> | <i>F</i> | <i>P-value</i> | <i>F crit</i> |
|----------------------------|-----------|-----------|-----------|----------|----------------|---------------|
| Between Groups | 6.4 | 2 | 3.2 | 0.092 | 0.912 | 3.885 |
| Within Groups | 415.6 | 12 | 34.6 | | | |
| Total | 422 | 14 | | | | |

Since the p-value is greater than 0.05 we fail to reject H_0 & conclude the variances are equal.



Chapter Summary

- Compared two independent samples
 - Performed pooled-variance t test for the difference in two means
 - Performed separate-variance t test for difference in two means
 - Formed confidence intervals for the difference between two means
- Compared two related samples (paired samples)
 - Performed paired t test for the mean difference
 - Formed confidence intervals for the mean difference

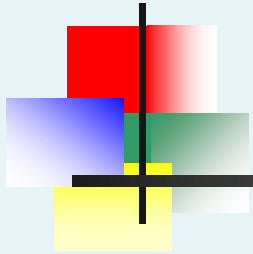


Chapter Summary

(continued)

- Compared two population proportions
 - Formed confidence intervals for the difference between two population proportions
 - Performed Z-test for two population proportions
- Performed F test for the difference between two population variances
- Described one-way analysis of variance
 - The logic of ANOVA
 - ANOVA assumptions
 - F test for difference in c means
 - The Tukey-Kramer procedure for multiple comparisons
 - The Levene test for homogeneity of variance

Business Statistics: A First Course Fifth Edition



Chapter 11

Chi-Square Tests



Learning Objectives

In this chapter, you learn:

- When to use the chi-square test for contingency tables
- How to use the chi-square test for contingency tables



Contingency Tables

Contingency Tables

- Useful in situations involving multiple population proportions
- Used to classify sample observations according to two or more characteristics
- Also called a cross-classification table.



Contingency Table Example

Left-Handed vs. Gender

Dominant Hand: Left vs. Right

Gender: Male vs. Female

- 2 categories for each variable, so called a **2 x 2 table**
- Suppose we examine a sample of 300 children

Contingency Table Example

(continued)

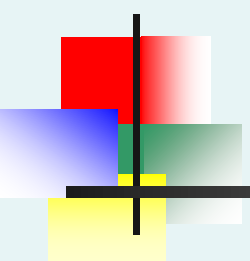
Sample results organized in a contingency table:

sample size = $n = 300$:

120 Females, 12
were left handed
180 Males, 24 were
left handed



| Gender | Hand Preference | | |
|--------|-----------------|-------|-----|
| | Left | Right | |
| Female | 12 | 108 | 120 |
| Male | 24 | 156 | 180 |
| | 36 | 264 | 300 |



χ^2 Test for the Difference Between Two Proportions

$H_0: \pi_1 = \pi_2$ (Proportion of females who are left handed is equal to the proportion of males who are left handed)

$H_1: \pi_1 \neq \pi_2$ (The two proportions are not the same – hand preference is **not** independent of gender)

- If H_0 is true, then the proportion of left-handed females should be the same as the proportion of left-handed males
- The two proportions above should be the same as the proportion of left-handed people overall



The Chi-Square Test Statistic

The Chi-square test statistic is:

$$\chi_{STAT}^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

■ where:

f_o = observed frequency in a particular cell

f_e = expected frequency in a particular cell if H_0 is true

χ_{STAT}^2 **for the 2 x 2 case has 1 degree of freedom**

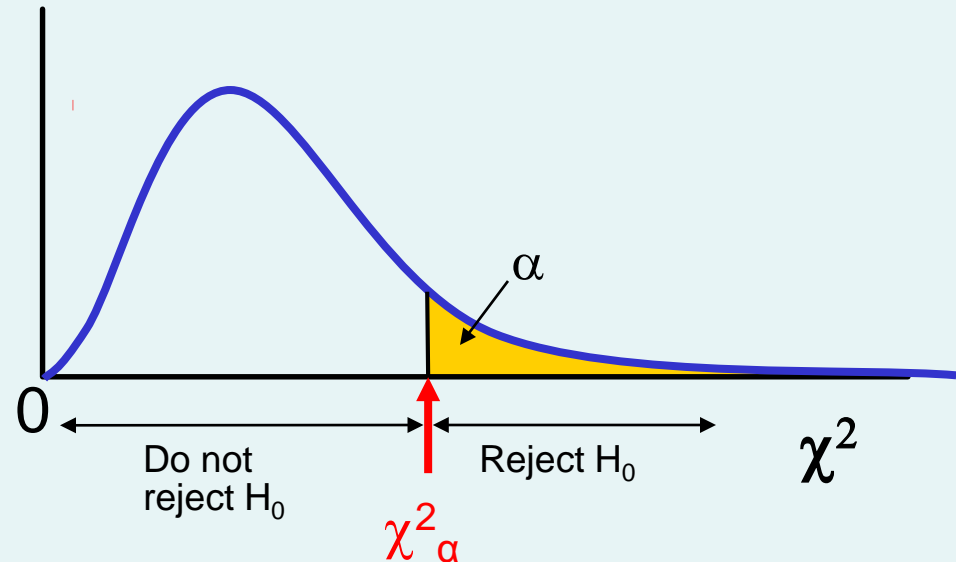
(Assumed: each cell in the contingency table has expected frequency of at least 5)

Decision Rule

The χ^2_{STAT} test statistic approximately follows a chi-squared distribution with one degree of freedom

Decision Rule:

If $\chi^2_{STAT} > \chi^2_{\alpha}$, reject H_0 ,
otherwise, do not reject H_0



Computing the Average Proportion

The average proportion is:

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{X}{n}$$

120 Females, 12 were left handed
180 Males, 24 were left handed

Here:

$$\bar{p} = \frac{12 + 24}{120 + 180} = \frac{36}{300} = 0.12$$

i.e., of all the children the proportion of left handers is 0.12, that is, 12%



Finding Expected Frequencies

- To obtain the expected frequency for left handed females, multiply the average proportion left handed (\bar{p}) by the total number of females
- To obtain the expected frequency for left handed males, multiply the average proportion left handed (\bar{p}) by the total number of males

If the two proportions are equal, then

$$P(\text{Left Handed} \mid \text{Female}) = P(\text{Left Handed} \mid \text{Male}) = .12$$

i.e., we would expect $(.12)(120) = 14.4$ females to be left handed
 $(.12)(180) = 21.6$ males to be left handed

Observed vs. Expected Frequencies

| Gender | Hand Preference | | |
|--------|----------------------------------|------------------------------------|-----|
| | Left | Right | |
| Female | Observed = 12 Expected = 14.4 | Observed = 108 Expected = 105.6 | 120 |
| Male | Observed = 24 Expected = 21.6 | Observed = 156 Expected = 158.4 | 180 |
| | 36 | 264 | 300 |

The Chi-Square Test Statistic

| Gender | Hand Preference | | |
|--------|----------------------------------|------------------------------------|-----|
| | Left | Right | |
| Female | Observed = 12 Expected = 14.4 | Observed = 108 Expected = 105.6 | 120 |
| Male | Observed = 24 Expected = 21.6 | Observed = 156 Expected = 158.4 | 180 |
| | 36 | 264 | 300 |

The test statistic is:

$$\chi^2_{STAT} = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

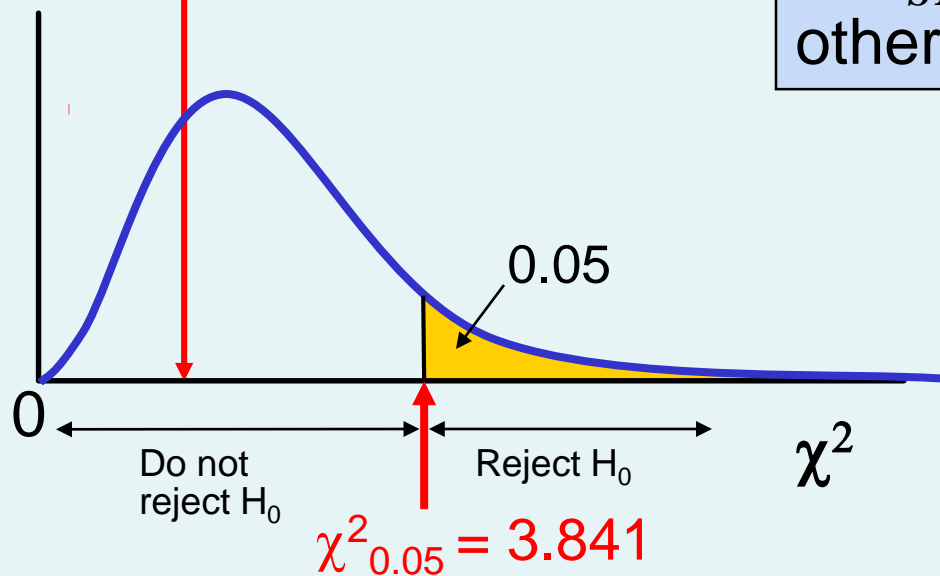
$$= \frac{(12 - 14.4)^2}{14.4} + \frac{(108 - 105.6)^2}{105.6} + \frac{(24 - 21.6)^2}{21.6} + \frac{(156 - 158.4)^2}{158.4} = 0.7576$$

Decision Rule

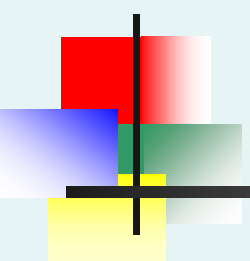
The test statistic is $\chi^2_{STAT} = 0.7576$; $\chi^2_{0.05}$ with 1 d.f. = 3.841

Decision Rule:

If $\chi^2_{STAT} > 3.841$, reject H_0 ,
otherwise, do not reject H_0



Here,
 $\chi^2_{STAT} = 0.7576 < \chi^2_{0.05} = 3.841$,
so we **do not reject H_0** and
conclude that there is not
sufficient evidence that the two
proportions are different at $\alpha =$
0.05



χ^2 Test for Differences Among More Than Two Proportions

- Extend the χ^2 test to the case with more than two independent populations:

$$H_0: \pi_1 = \pi_2 = \cdots = \pi_c$$

$$H_1: \text{Not all of the } \pi_j \text{ are equal } (j = 1, 2, \cdots, c)$$



The Chi-Square Test Statistic

The Chi-square test statistic is:

$$\chi_{STAT}^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

- Where:

f_o = observed frequency in a particular cell of the 2 x c table

f_e = expected frequency in a particular cell if H_0 is true

χ_{STAT}^2 for the 2 x c case has $(2 - 1)(c - 1) = c - 1$ degrees of freedom

(Assumed: each cell in the contingency table has expected frequency of at least 1)

Computing the Overall Proportion

The overall proportion is:

$$\bar{p} = \frac{X_1 + X_2 + \cdots + X_c}{n_1 + n_2 + \cdots + n_c} = \frac{X}{n}$$

- Expected cell frequencies for the c categories are calculated as in the 2×2 case, and the decision rule is the same:

Decision Rule:

If $\chi_{STAT}^2 > \chi_{\alpha}^2$, reject H_0 ,
otherwise, do not reject H_0

Where χ_{α}^2 is from the chi-squared distribution with $c - 1$ degrees of freedom



χ^2 Test of Independence

- Similar to the χ^2 test for equality of more than two proportions, but extends the concept to contingency tables with **r rows** and **c columns**

H_0 : The two categorical variables are independent
(i.e., there is no relationship between them)

H_1 : The two categorical variables are dependent
(i.e., there is a relationship between them)



χ^2 Test of Independence

(continued)

The Chi-square test statistic is:

$$\chi_{STAT}^2 = \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e}$$

■ where:

f_o = observed frequency in a particular cell of the $r \times c$ table

f_e = expected frequency in a particular cell if H_0 is true

χ_{STAT}^2 for the $r \times c$ case has $(r - 1)(c - 1)$ degrees of freedom

(Assumed: each cell in the contingency table has expected frequency of at least 1)



Expected Cell Frequencies

- Expected cell frequencies:

$$f_e = \frac{\text{row total} \times \text{column total}}{n}$$

Where:

row total = sum of all frequencies in the row

column total = sum of all frequencies in the column

n = overall sample size



Decision Rule

- The decision rule is

If $\chi_{STAT}^2 > \chi_{\alpha}^2$, reject H_0 ,
otherwise, do not reject H_0

Where χ_{α}^2 is from the chi-squared distribution
with $(r - 1)(c - 1)$ degrees of freedom



Example

- The meal plan selected by 200 students is shown below:

| Class Standing | Number of meals per week | | | Total |
|----------------|--------------------------|---------|------|-------|
| | 20/week | 10/week | none | |
| Fresh. | 24 | 32 | 14 | 70 |
| Soph. | 22 | 26 | 12 | 60 |
| Junior | 10 | 14 | 6 | 30 |
| Senior | 14 | 16 | 10 | 40 |
| Total | 70 | 88 | 42 | 200 |



Example

(continued)

- The hypothesis to be tested is:

H_0 : Meal plan and class standing are independent
(i.e., there is no relationship between them)

H_1 : Meal plan and class standing are dependent
(i.e., there is a relationship between them)

Example: Expected Cell Frequencies

(continued)

Observed:

| Class Standing | Number of meals per week | | | Total |
|----------------|--------------------------|-------|------|-------|
| | 20/wk | 10/wk | none | |
| Fresh. | 24 | 32 | 14 | 70 |
| Soph. | 22 | 26 | 12 | 60 |
| Junior | 10 | 14 | 6 | 30 |
| Senior | 14 | 16 | 10 | 40 |
| Total | 70 | 88 | 42 | 200 |

Expected cell frequencies if H_0 is true:

| Class Standing | Number of meals per week | | | Total |
|----------------|--------------------------|-------|------|-------|
| | 20/wk | 10/wk | none | |
| Fresh. | 24.5 | 30.8 | 14.7 | 70 |
| Soph. | 21.0 | 26.4 | 12.6 | 60 |
| Junior | 10.5 | 13.2 | 6.3 | 30 |
| Senior | 14.0 | 17.6 | 8.4 | 40 |
| Total | 70 | 88 | 42 | 200 |

Example for one cell:

$$f_e = \frac{\text{row total} \times \text{column total}}{n}$$

$$= \frac{30 \times 70}{200} = 10.5$$



Example: The Test Statistic

(continued)

- The test statistic value is:

$$\begin{aligned}\chi_{STAT}^2 &= \sum_{\text{all cells}} \frac{(f_o - f_e)^2}{f_e} \\ &= \frac{(24 - 24.5)^2}{24.5} + \frac{(32 - 30.8)^2}{30.8} + \dots + \frac{(10 - 8.4)^2}{8.4} = 0.709\end{aligned}$$

$\chi_{0.05}^2 = 12.592$ from the chi-squared distribution
with $(4 - 1)(3 - 1) = 6$ degrees of freedom

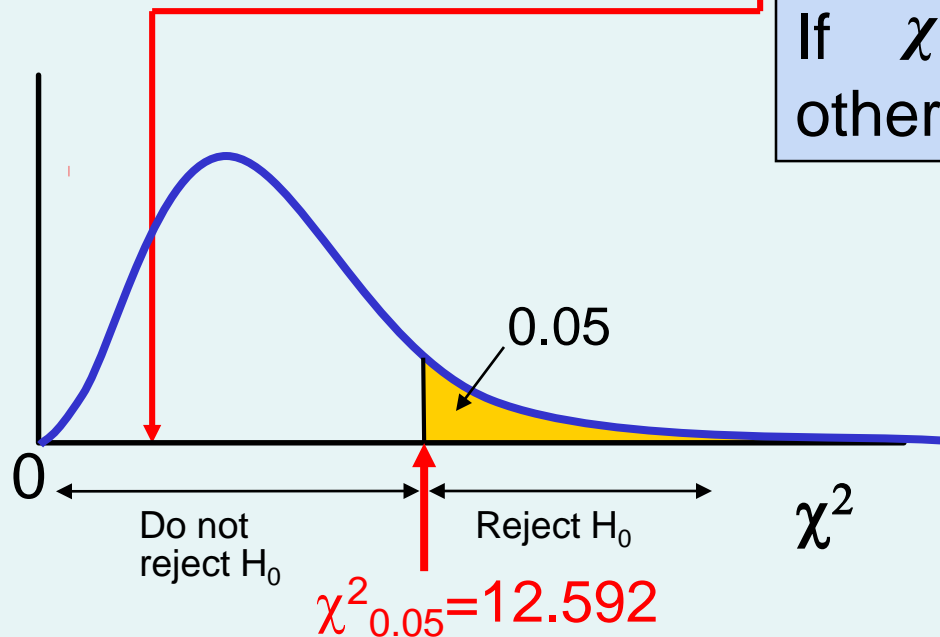
Example: Decision and Interpretation

(continued)

The test statistic is $\chi^2_{STAT} = 0.709$; $\chi^2_{0.05}$ with 6 d.f. = 12.592

Decision Rule:

If $\chi^2_{STAT} > 12.592$, reject H_0 ,
otherwise, do not reject H_0



Here,
 $\chi^2_{STAT} = 0.709 < \chi^2_{0.05} = 12.592$,
so do not reject H_0

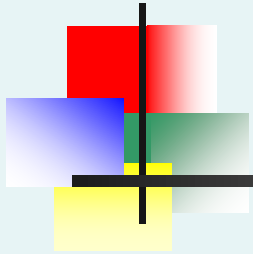
Conclusion: there is not
sufficient evidence that meal
plan and class standing are
related at $\alpha = 0.05$



Chapter Summary

- Developed and applied the χ^2 test for the difference between two proportions
- Developed and applied the χ^2 test for differences in more than two proportions
- Examined the χ^2 test for independence

Business Statistics: A First Course Fifth Edition



Chapter 12

Simple Linear Regression



Learning Objectives

In this chapter, you learn:

- How to use regression analysis to predict the value of a dependent variable based on an independent variable
- The meaning of the regression coefficients b_0 and b_1
- How to evaluate the assumptions of regression analysis and know what to do if the assumptions are violated
- To make inferences about the slope and correlation coefficient
- To estimate mean values and predict individual values



Correlation vs. Regression

- A **scatter plot** can be used to show the relationship between two variables
- **Correlation** analysis is used to measure the strength of the association (linear relationship) between two variables
 - Correlation is only concerned with strength of the relationship
 - No causal effect is implied with correlation
 - Scatter plots were first presented in Ch. 2
 - Correlation was first presented in Ch. 3



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 predict or explain

Independent variable: the variable used to predict or explain the dependent variable

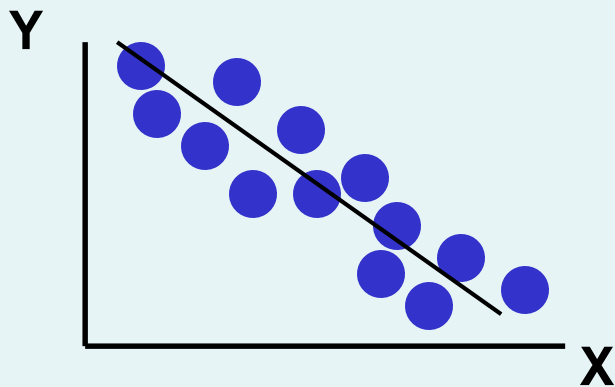
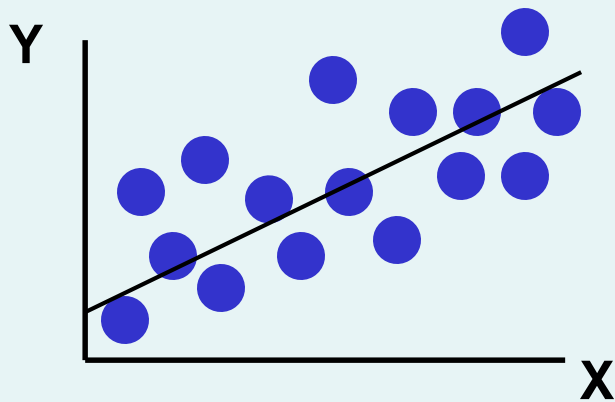


Simple Linear Regression Model

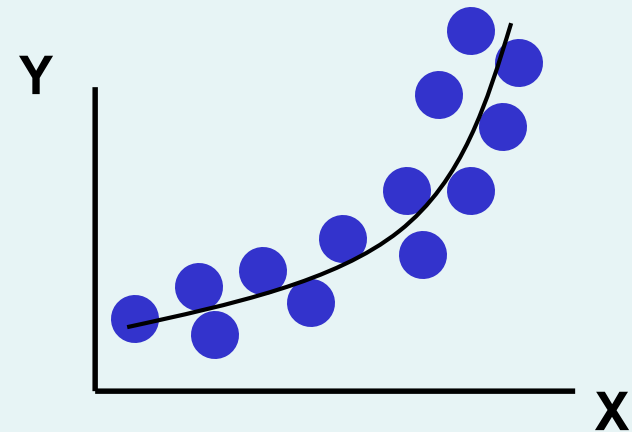
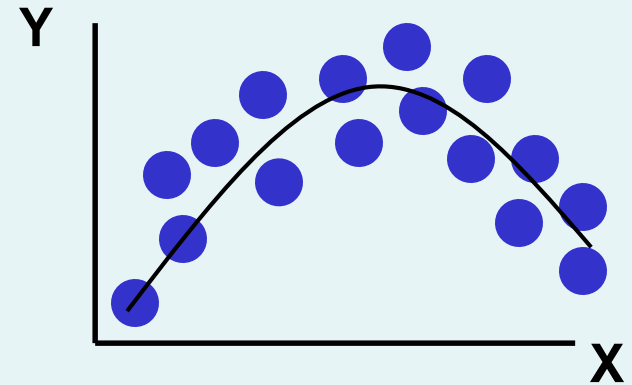
- Only **one** independent variable, X
- Relationship between X and Y is described by a linear function
- Changes in Y are assumed to be related to changes in X

Types of Relationships

Linear relationships



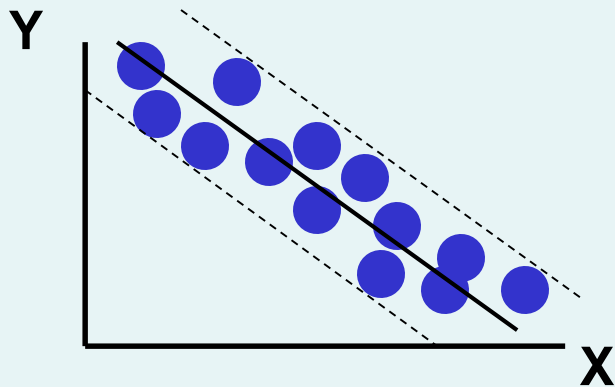
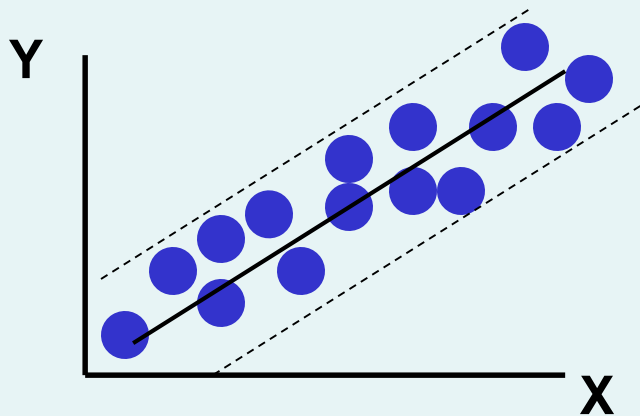
Curvilinear relationships



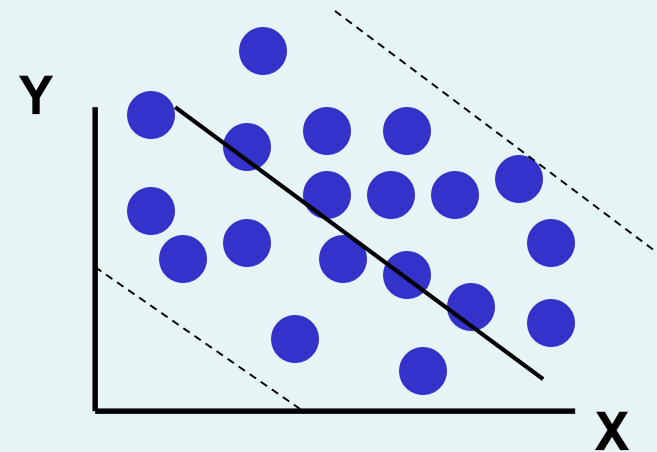
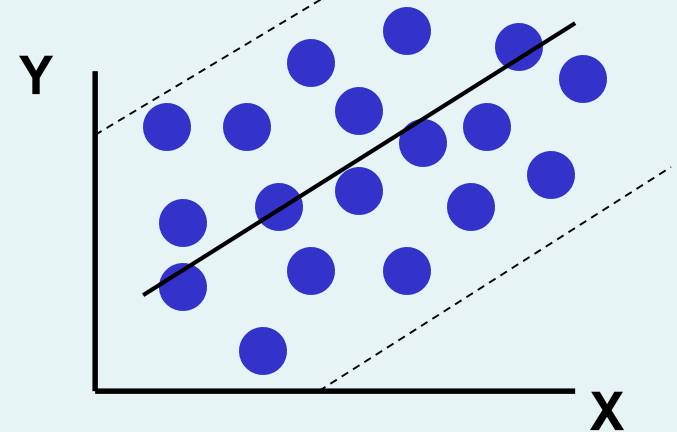
Types of Relationships

(continued)

Strong relationships



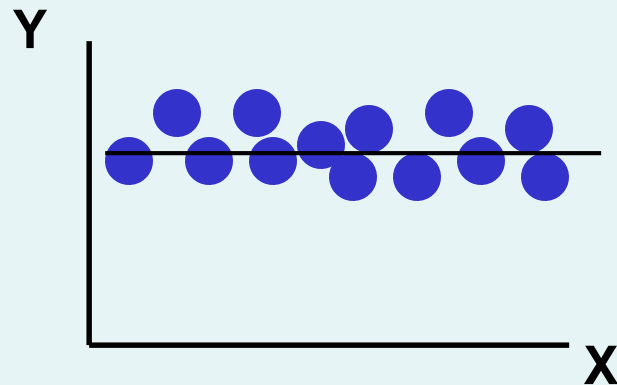
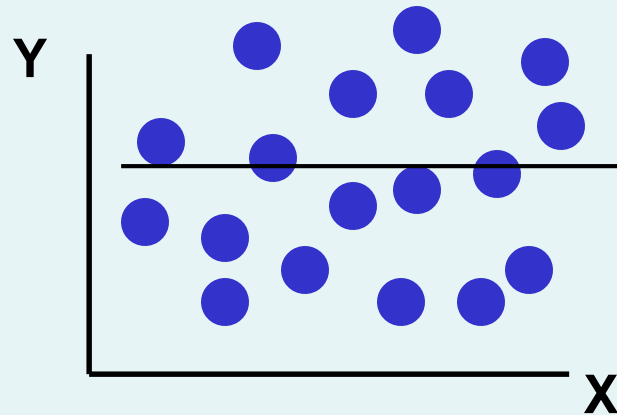
Weak relationships



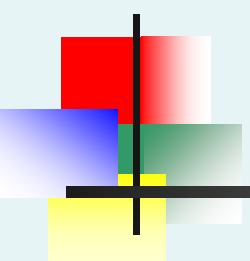
Types of Relationships

(continued)

No relationship



Simple Linear Regression Model



Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

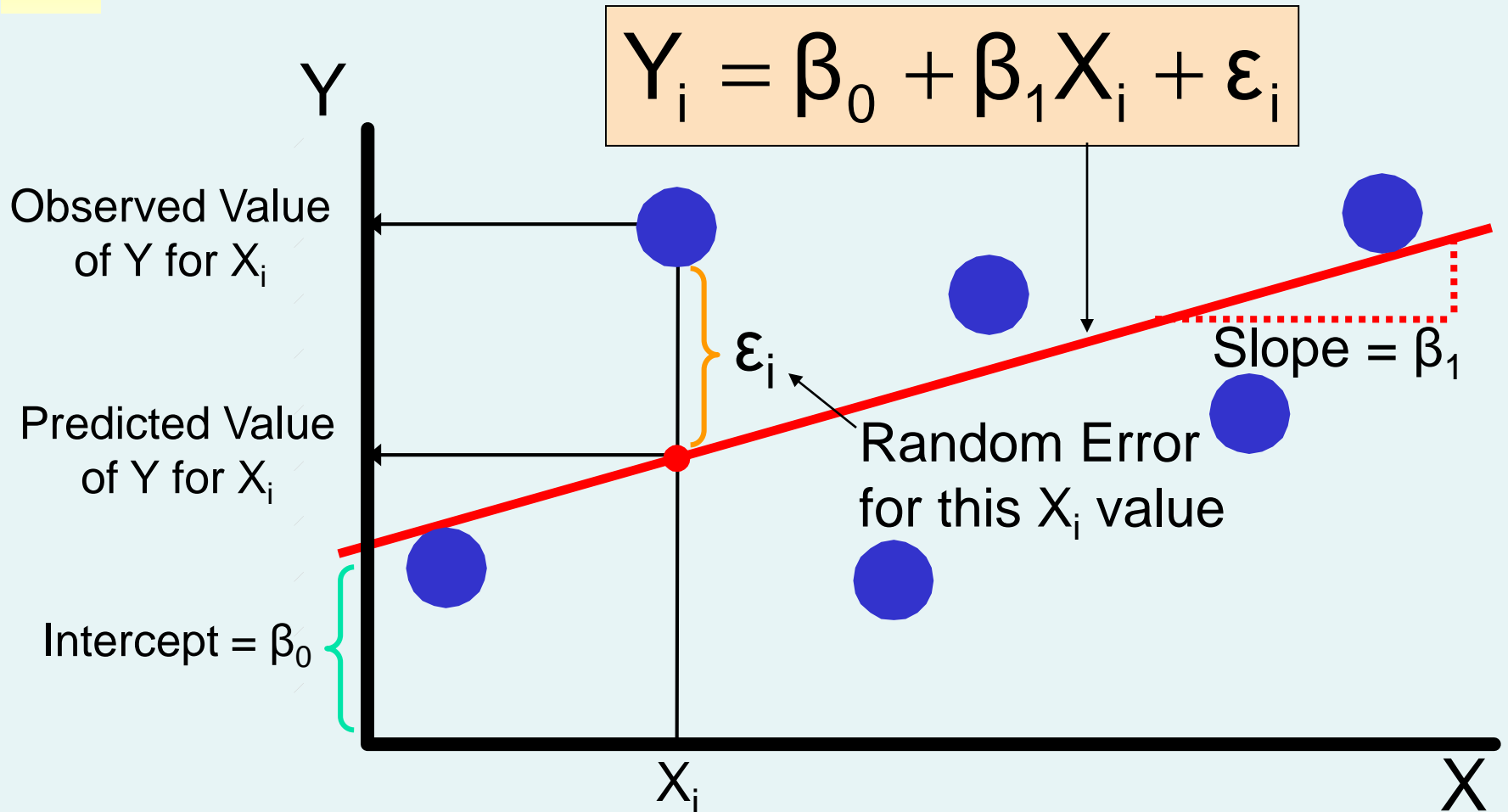
Linear component

Random Error component

The diagram illustrates the Simple Linear Regression Model equation: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. The equation is enclosed in a light orange box. Labels with arrows point to each term: 'Dependent Variable' points to Y_i , 'Population Y intercept' points to β_0 , 'Population Slope Coefficient' points to β_1 , 'Independent Variable' points to X_i , and 'Random Error term' points to ε_i . Below the equation, two blue brackets group the terms: the first bracket under $\beta_0 + \beta_1 X_i$ is labeled 'Linear component', and the second bracket under ε_i is labeled 'Random Error component'.

Simple Linear Regression Model

(continued)



Simple Linear Regression Equation (Prediction Line)

The simple linear regression equation provides an **estimate** of the population regression line

Estimated
(or predicted)
Y value for
observation i

Estimate of
the regression
intercept

Estimate of the
regression slope

Value of X for
observation i

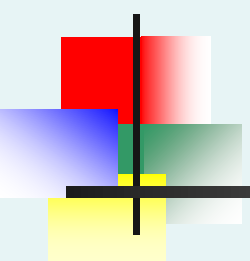
$$\hat{Y}_i = b_0 + b_1 X_i$$



The Least Squares Method

b_0 and b_1 are obtained by finding the values of that minimize the sum of the squared differences between Y and \hat{Y} :

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$



Finding the Least Squares Equation

- The coefficients b_0 and b_1 , and other regression results in this chapter, will be found using Excel or Minitab

Formulas are shown in the text for those who are interested



Interpretation of the Slope and the Intercept

- b_0 is the estimated mean value of Y when the value of X is zero
- b_1 is the estimated change in the mean value of Y as a result of a one-unit change in X

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



Simple Linear Regression

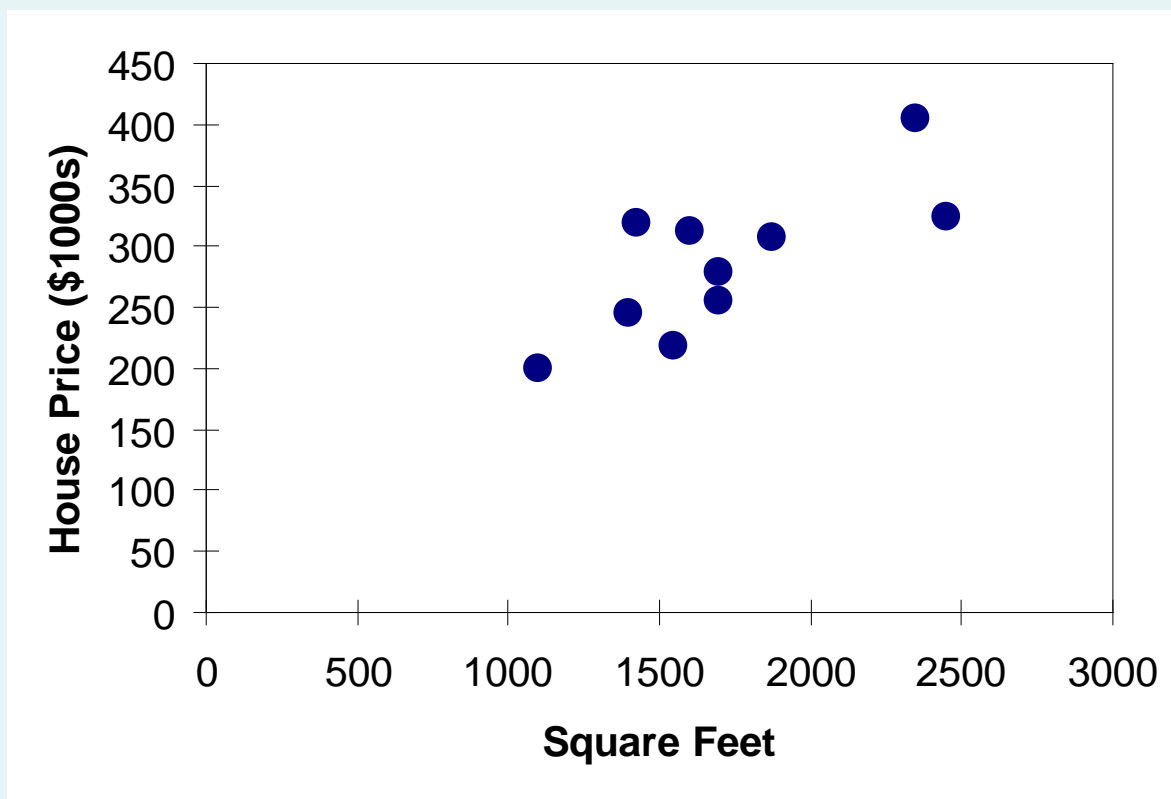
Example: Data

| 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 |



Simple Linear Regression Example: Scatter Plot

House price model: Scatter Plot



Simple Linear Regression Example: Using Excel

The screenshot shows Microsoft Excel with a data table and the Regression dialog box open. The data table has two columns: House Price and Square Feet. The Regression dialog box is configured with 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 data table is as follows:

| | 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 | | |
| 14 | | |
| 15 | | |



Simple Linear Regression Example: 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:

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

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 |



Simple Linear Regression Example: Minitab Output

The regression equation is

$$\text{Price} = 98.2 + 0.110 \text{ Square Feet}$$

| Predictor | Coef | SE Coef | T | P |
|-------------|---------|---------|------|-------|
| Constant | 98.25 | 58.03 | 1.69 | 0.129 |
| Square Feet | 0.10977 | 0.03297 | 3.33 | 0.010 |

S = 41.3303 R-Sq = 58.1% R-Sq(adj) = 52.8%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|-------|-------|-------|-------|
| Regression | 1 | 18935 | 18935 | 11.08 | 0.010 |
| Residual Error | 8 | 13666 | 1708 | | |
| Total | 9 | 32600 | | | |

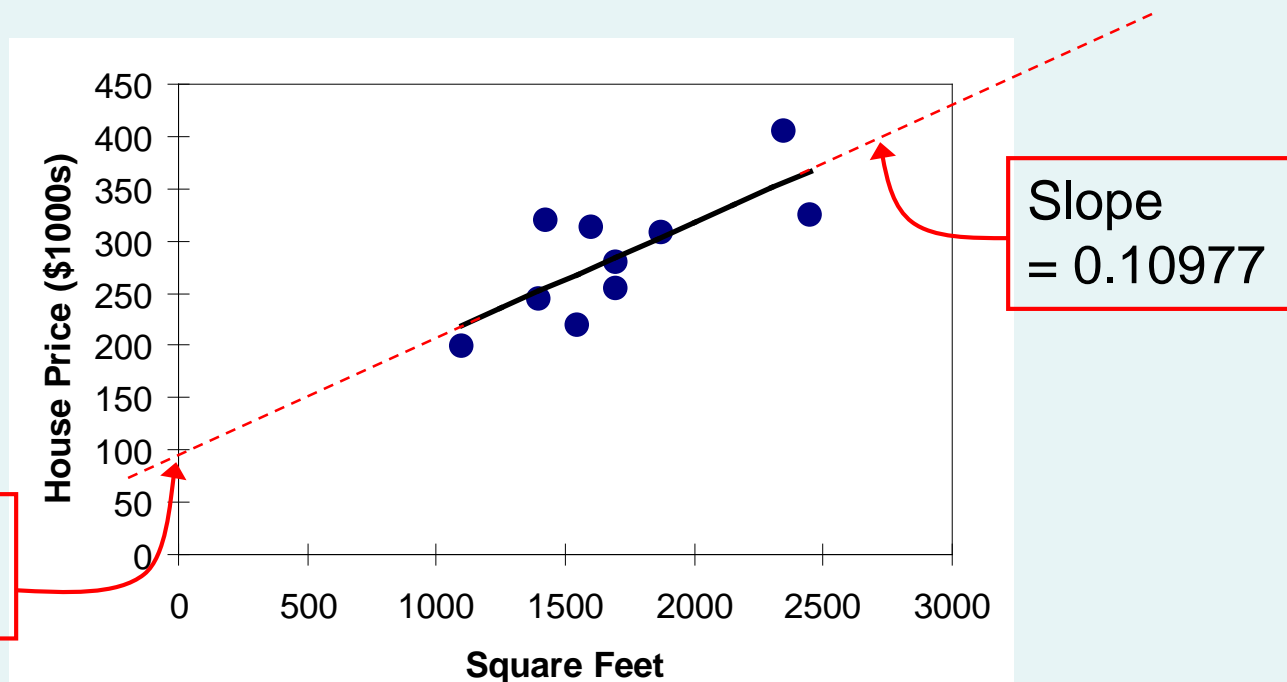
The regression equation is:

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



Simple Linear Regression Example: Graphical Representation

House price model: Scatter Plot and Prediction Line



Intercept
= 98.248

Slope
= 0.10977

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

Simple Linear Regression

Example: Interpretation of b_0

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

- b_0 is the estimated mean value of Y when the value of X is zero (if $X = 0$ is in the range of observed X values)
- Because a house cannot have a square footage of 0, b_0 has no practical application



Simple Linear Regression

Example: Interpreting b_1

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

- b_1 estimates the change in the mean value of Y as a result of a one-unit increase in X
 - Here, $b_1 = 0.10977$ tells us that the mean value of a house increases by $0.10977(\$1000) = \109.77 , on average, for each additional one square foot of size



Simple Linear Regression

Example: Making Predictions

Predict the price for a house with 2000 square feet:

$$\begin{aligned}\widehat{\text{house price}} &= 98.25 + 0.1098 (\text{sq.ft.}) \\ &= 98.25 + 0.1098(2000) \\ &= 317.85\end{aligned}$$

The predicted price for a house with 2000 square feet is $317.85(\$1,000\text{s}) = \$317,850$

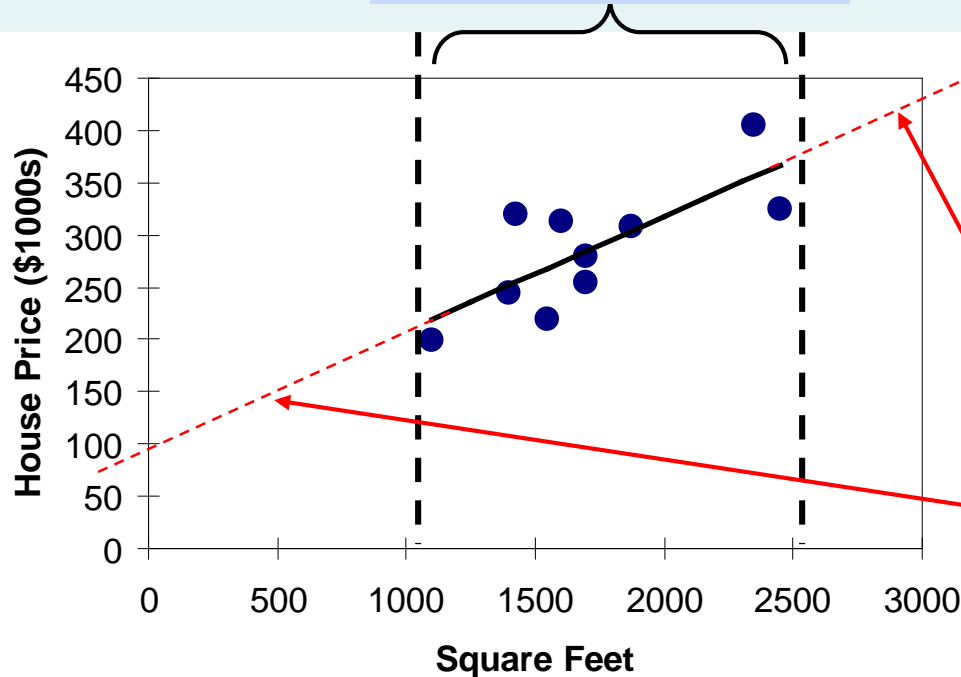


Simple Linear Regression

Example: Making Predictions

- When using a regression model for prediction, only predict within the relevant range of data

Relevant range for interpolation



Do not try to extrapolate beyond the range of observed X's



Measures of Variation

- Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of
Squares

Regression Sum
of Squares

Error Sum of
Squares

$$SST = \sum (Y_i - \bar{Y})^2$$

$$SSR = \sum (\hat{Y}_i - \bar{Y})^2$$

$$SSE = \sum (Y_i - \hat{Y}_i)^2$$

where:

\bar{Y} = Mean value of the dependent variable

Y_i = Observed value of the dependent variable

\hat{Y}_i = Predicted value of Y for the given X_i value



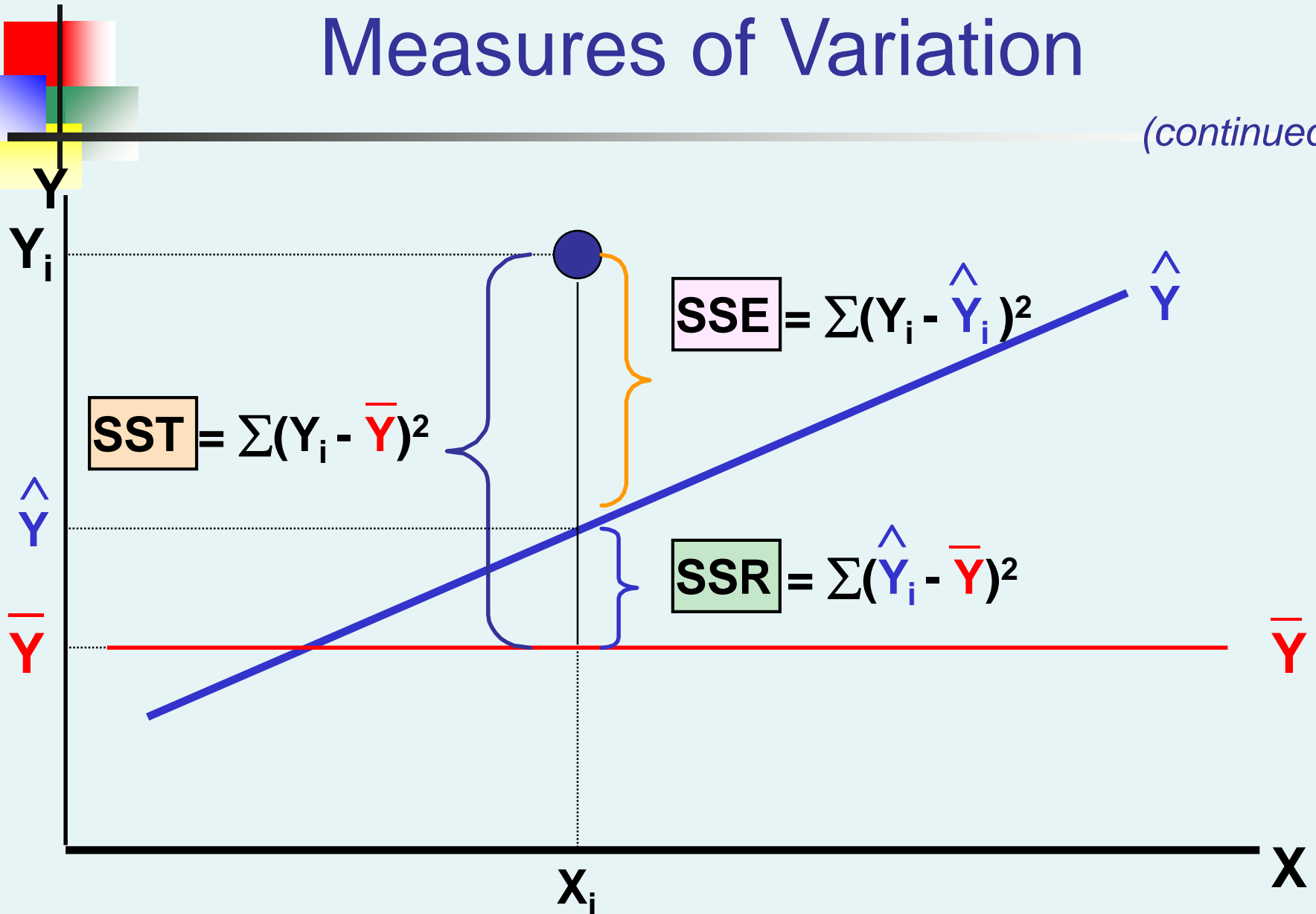
Measures of Variation

(continued)

- SST = total sum of squares (Total Variation)
 - Measures the variation of the Y_i values around their mean \bar{Y}
- SSR = regression sum of squares (Explained Variation)
 - Variation attributable to the relationship between X and Y
- SSE = error sum of squares (Unexplained Variation)
 - Variation in Y attributable to factors other than X

Measures of Variation

(continued)





Coefficient of Determination, r^2

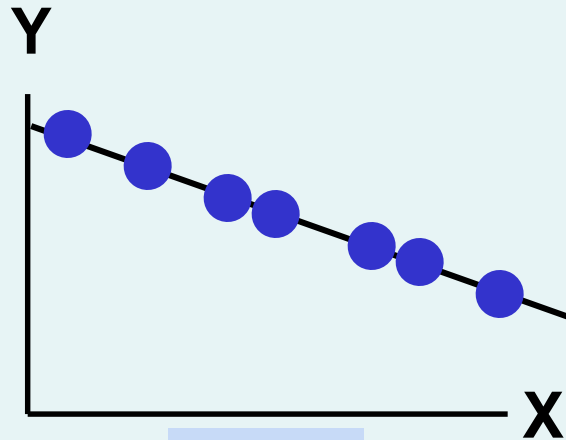
- The **coefficient of determination** is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called **r-squared** and is denoted as r^2

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

note:

$$0 \leq r^2 \leq 1$$

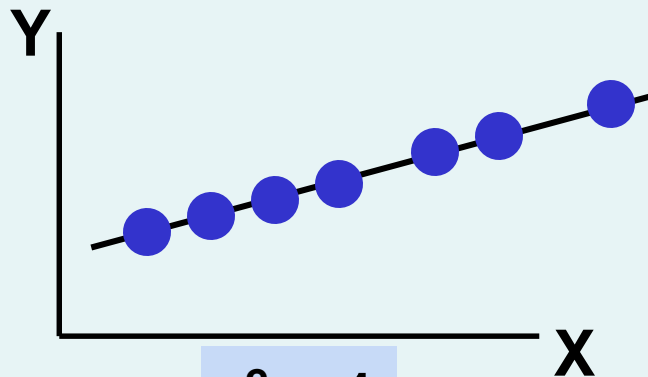
Examples of r^2 Values



$$r^2 = 1$$

$$r^2 = 1$$

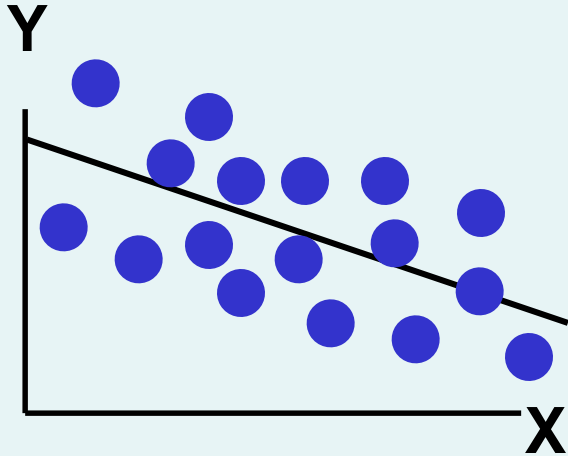
**Perfect linear relationship
between X and Y:**



$$r^2 = 1$$

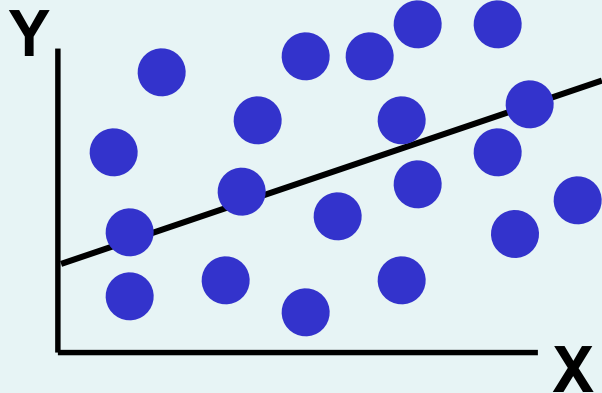
**100% of the variation in Y is
explained by variation in X**

Examples of r^2 Values



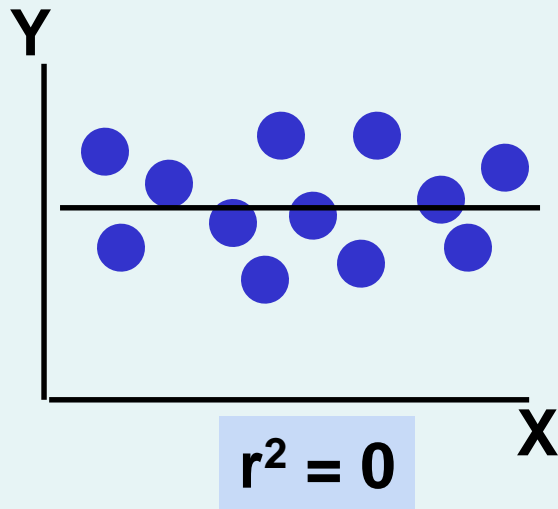
$$0 < r^2 < 1$$

**Weaker linear relationships
between X and Y:**



**Some but not all of the
variation in Y is explained
by variation in X**

Examples of r^2 Values



$$r^2 = 0$$

**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)**

Simple Linear Regression Example: Coefficient of Determination, r^2 in Excel

Regression Statistics

| | |
|-------------------|----------|
| Multiple R | 0.76211 |
| R Square | 0.58082 |
| Adjusted R Square | 0.52842 |
| Standard Error | 41.33032 |
| Observations | 10 |

$$r^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

ANOVA

| | <i>df</i> | SS | <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 |



Simple Linear Regression Example: Coefficient of Determination, r^2 in Minitab



The regression equation is

Price = 98.2 + 0.110 Square Feet

| Predictor | Coef | SE Coef | T | P |
|-------------|---------|---------|------|-------|
| Constant | 98.25 | 58.03 | 1.69 | 0.129 |
| Square Feet | 0.10977 | 0.03297 | 3.33 | 0.010 |

S = 41.3303 R-Sq = 58.1% R-Sq(adj) = 52.8%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|-------|-------|-------|-------|
| Regression | 1 | 18935 | 18935 | 11.08 | 0.010 |
| Residual Error | 8 | 13666 | 1708 | | |
| Total | 9 | 32600 | | | |

$$r^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet



Standard Error of Estimate

- The standard deviation of the variation of observations around the regression line is estimated by

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}}$$

Where

SSE = error sum of squares

n = sample size

Simple Linear Regression Example: Standard Error of Estimate in Excel

Regression Statistics

| | |
|-------------------|----------|
| Multiple R | 0.76211 |
| R Square | 0.58082 |
| Adjusted R Square | 0.52842 |
| Standard Error | 41.33032 |
| Observations | 10 |

$$S_{YX} = 41.33032$$

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 |



Simple Linear Regression Example: Standard Error of Estimate in Minitab

The regression equation is

Price = 98.2 + 0.110 Square Feet

| Predictor | Coef | SE Coef | T | P |
|-------------|---------|---------|------|-------|
| Constant | 98.25 | 58.03 | 1.69 | 0.129 |
| Square Feet | 0.10977 | 0.03297 | 3.33 | 0.010 |

S = 41.3303 R-Sq = 58.1% R-Sq(adj) = 52.8%

Analysis of Variance

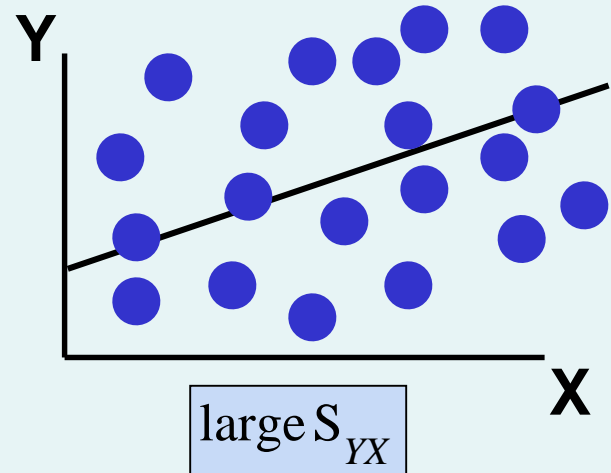
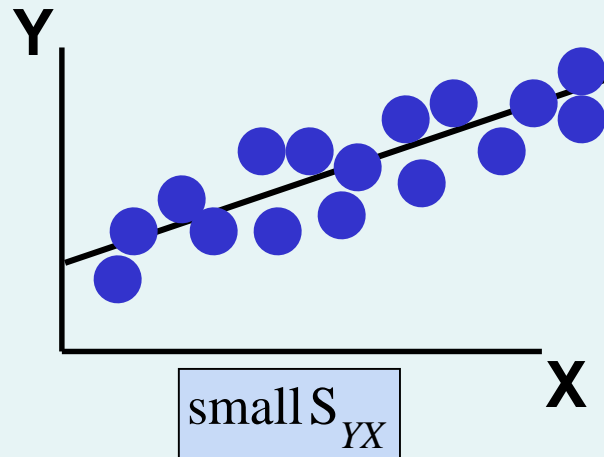
| Source | DF | SS | MS | F | P |
|----------------|----|-------|-------|-------|-------|
| Regression | 1 | 18935 | 18935 | 11.08 | 0.010 |
| Residual Error | 8 | 13666 | 1708 | | |
| Total | 9 | 32600 | | | |

$$S_{YX} = 41.33032$$



Comparing Standard Errors

S_{YX} is a measure of the variation of observed Y values from the regression line



The magnitude of S_{YX} should always be judged relative to the size of the Y values in the sample data

i.e., $S_{YX} = \$41.33K$ is moderately small relative to house prices in the \$200K - \$400K range

Assumptions of Regression

L.I.N.E



- Linearity
 - The relationship between X and Y is linear
- Independence of Errors
 - Error values are statistically independent
- Normality of Error
 - Error values are normally distributed for any given value of X
- Equal Variance (also called homoscedasticity)
 - The probability distribution of the errors has constant variance

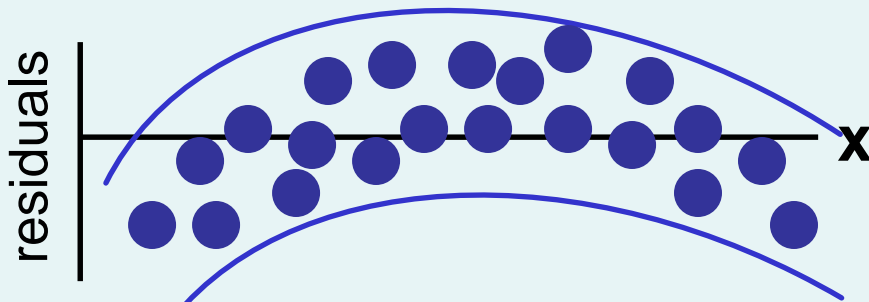
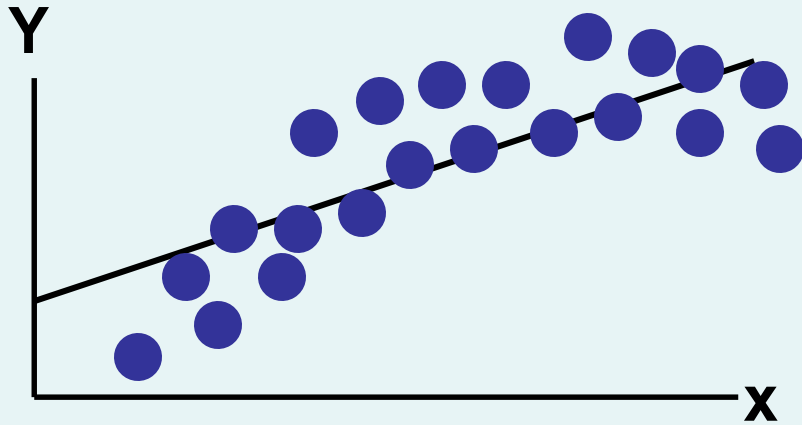


Residual Analysis

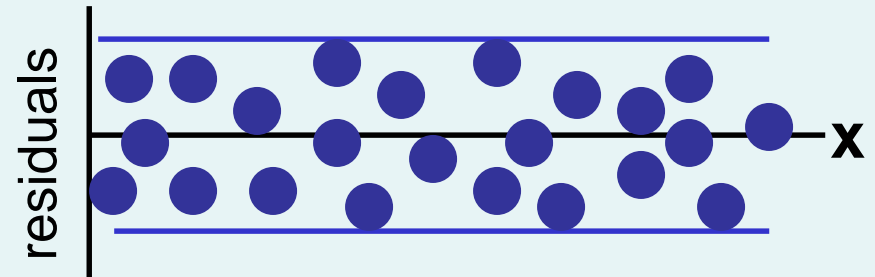
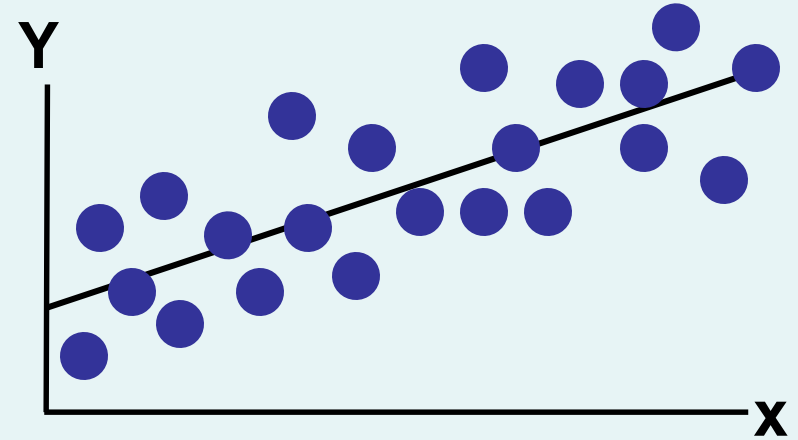
$$e_i = Y_i - \hat{Y}_i$$

- The residual for observation i , e_i , is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
 - Examine for linearity assumption
 - Evaluate independence assumption
 - Evaluate normal distribution assumption
 - Examine for constant variance for all levels of X (homoscedasticity)
- Graphical Analysis of Residuals
 - Can plot residuals vs. X

Residual Analysis for Linearity



Not Linear

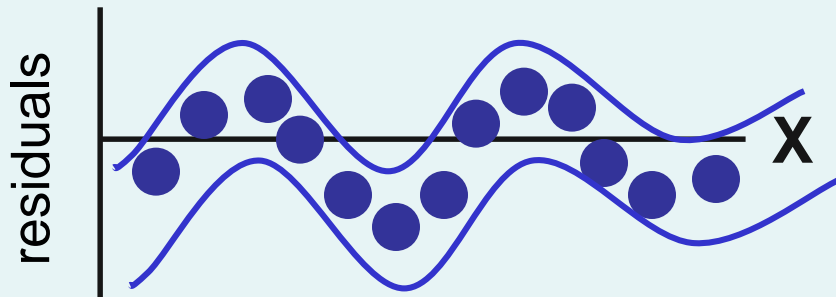
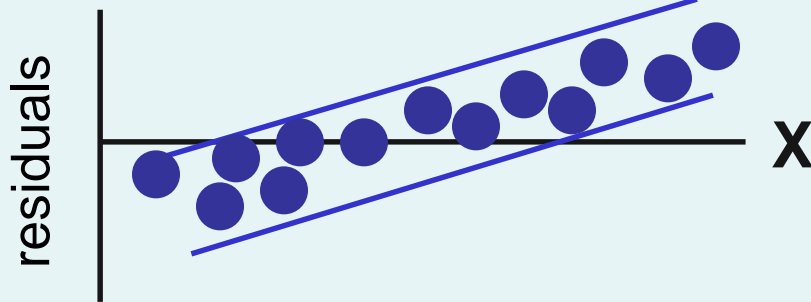


Linear

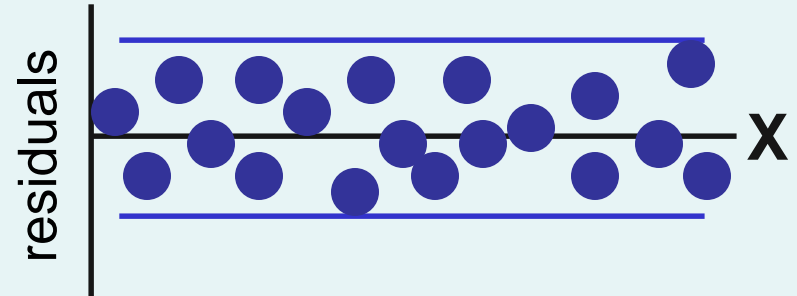
Residual Analysis for Independence



Not Independent



Independent



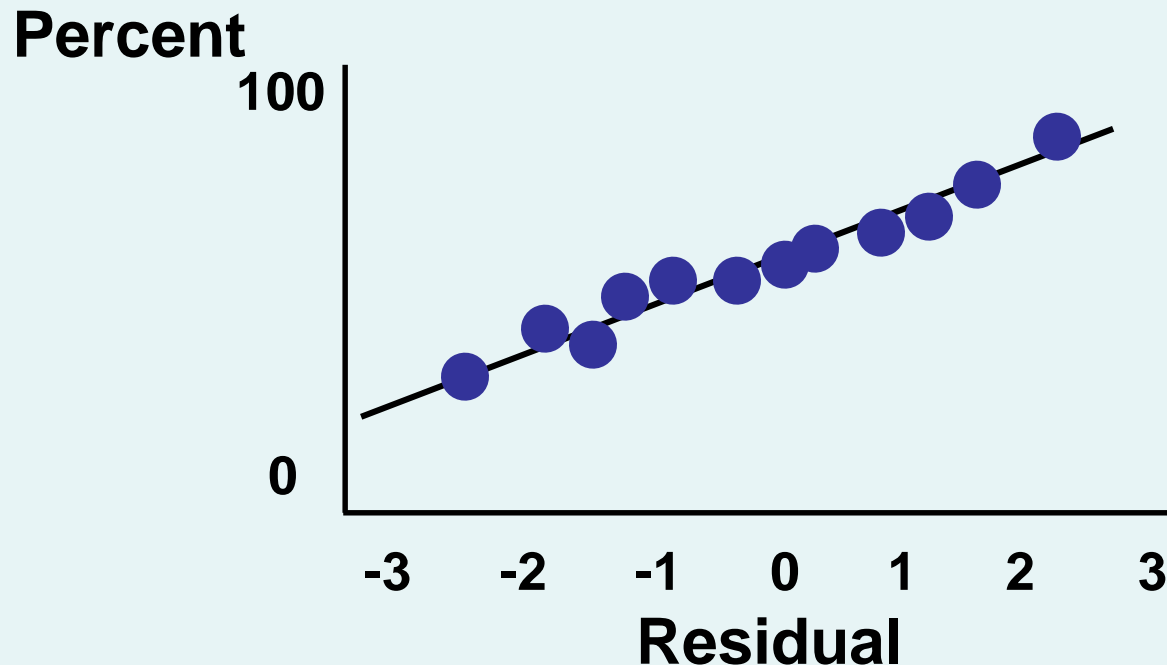


Checking for Normality

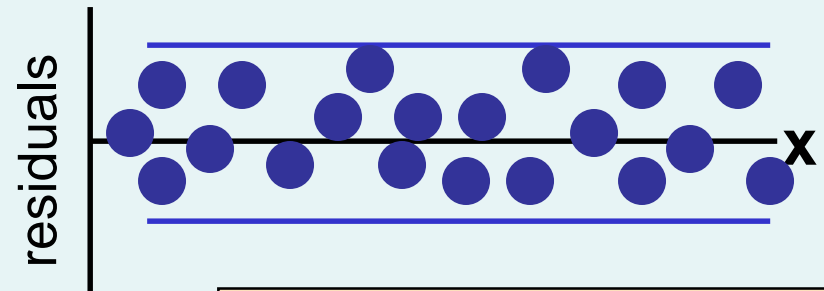
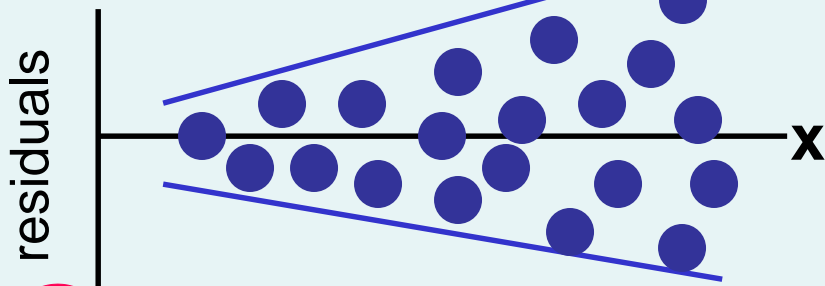
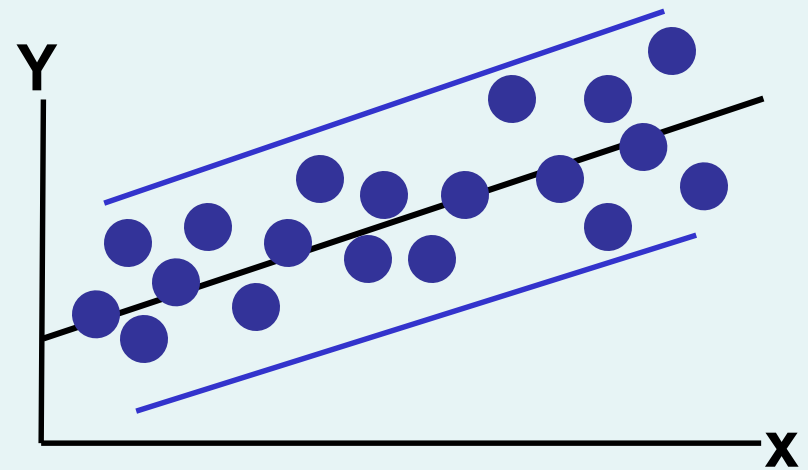
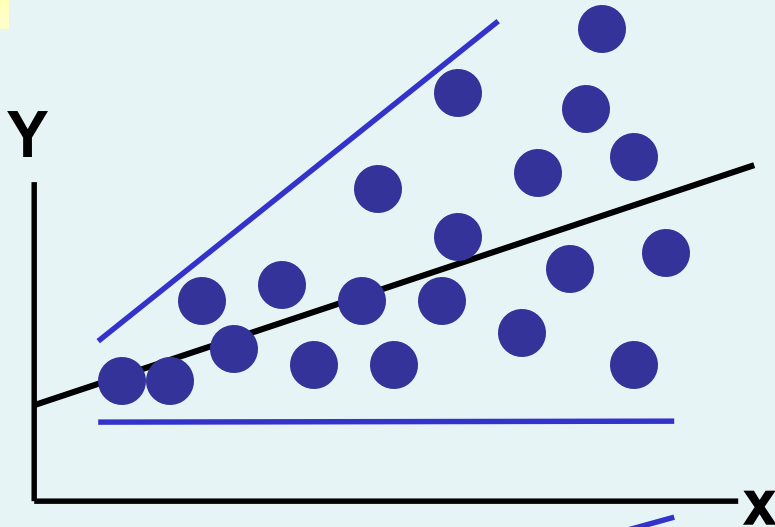
- Examine the Stem-and-Leaf Display of the Residuals
- Examine the Boxplot of the Residuals
- Examine the Histogram of the Residuals
- Construct a Normal Probability Plot of the Residuals

Residual Analysis for Normality

When using a normal probability plot, normal errors will approximately display in a straight line



Residual Analysis for Equal Variance



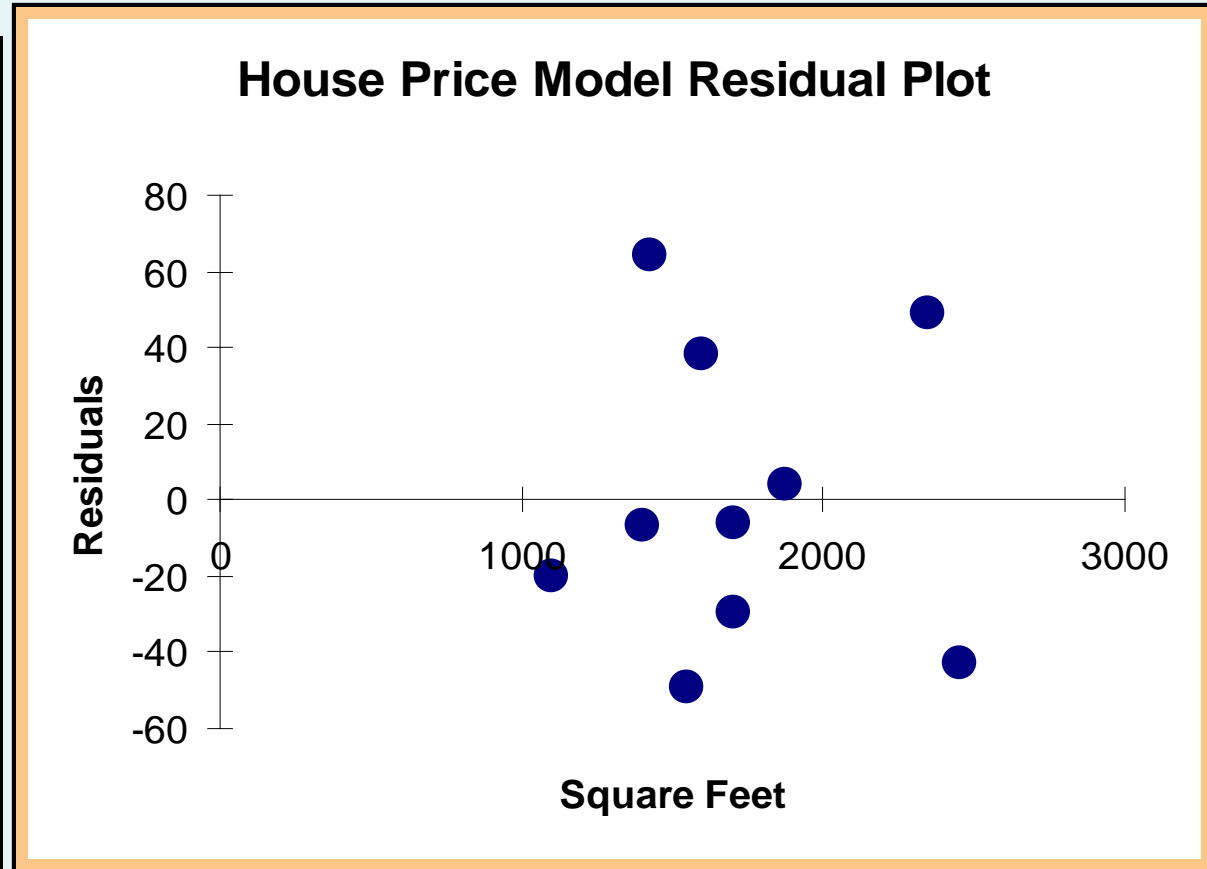
Non-constant variance



Constant variance

Simple Linear Regression Example: Excel Residual Output

| RESIDUAL OUTPUT | | |
|-----------------|------------------------------|------------------|
| | <i>Predicted House Price</i> | <i>Residuals</i> |
| 1 | 251.92316 | -6.923162 |
| 2 | 273.87671 | 38.12329 |
| 3 | 284.85348 | -5.853484 |
| 4 | 304.06284 | 3.937162 |
| 5 | 218.99284 | -19.99284 |
| 6 | 268.38832 | -49.38832 |
| 7 | 356.20251 | 48.79749 |
| 8 | 367.17929 | -43.17929 |
| 9 | 254.6674 | 64.33264 |
| 10 | 284.85348 | -29.85348 |



Does not appear to violate
any regression assumptions



Inferences About the Slope

- The standard error of the regression slope coefficient (b_1) is estimated by

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SSX}} = \frac{S_{YX}}{\sqrt{\sum (X_i - \bar{X})^2}}$$

where:

S_{b_1} = Estimate of the standard error of the slope

$S_{YX} = \sqrt{\frac{SSE}{n-2}}$ = Standard error of the estimate

Inferences About the Slope: t Test

- t test for a population slope
 - Is there a linear relationship between X and Y?
- Null and alternative hypotheses
 - $H_0: \beta_1 = 0$ (no linear relationship)
 - $H_1: \beta_1 \neq 0$ (linear relationship does exist)
- Test statistic

$$t_{\text{STAT}} = \frac{b_1 - \beta_1}{S_{b_1}}$$

$$\text{d.f.} = n - 2$$

where:

b_1 = regression slope
coefficient

β_1 = hypothesized slope

S_{b_1} = standard
error of the slope

Inferences About the Slope: t Test Example

| 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 |

Estimated Regression Equation:

$$\text{house price} = 98.25 + 0.1098 (\text{sq.ft.})$$

The slope of this model is 0.1098

Is there a relationship between the square footage of the house and its sales price?

Inferences About the Slope: t Test Example

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

From Excel output:

| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
|-------------|---------------------|-----------------------|---------------|----------------|
| Intercept | 98.24833 | 58.03348 | 1.69296 | 0.12892 |
| Square Feet | 0.10977 | 0.03297 | 3.32938 | 0.01039 |

From Minitab output:

| Predictor | Coef | SE Coef | T | P |
|-------------|---------|---------|------|-------|
| Constant | 98.25 | 58.03 | 1.69 | 0.129 |
| Square Feet | 0.10977 | 0.03297 | 3.33 | 0.010 |

b_1

S_{b_1}

b_1

S_{b_1}

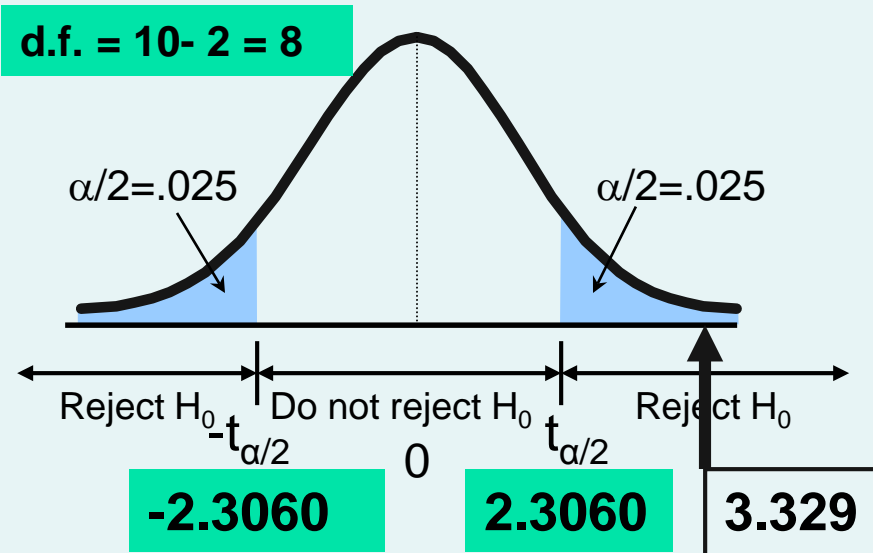
$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{0.10977 - 0}{0.03297} = 3.32938$$

Inferences About the Slope: t Test Example

Test Statistic: $t_{\text{STAT}} = 3.329$

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$



Decision: Reject H_0

There is sufficient evidence that square footage affects house price

Inferences About the Slope: t Test Example

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

From Excel output:

| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> |
|-------------|---------------------|-----------------------|---------------|----------------|
| Intercept | 98.24833 | 58.03348 | 1.69296 | 0.12892 |
| Square Feet | 0.10977 | 0.03297 | 3.32938 | 0.01039 |

From Minitab output:

| Predictor | Coef | SE Coef | T | P |
|-------------|---------|---------|------|-------|
| Constant | 98.25 | 58.03 | 1.69 | 0.129 |
| Square Feet | 0.10977 | 0.03297 | 3.33 | 0.010 |

p-value

Decision: Reject H_0 , since p-value $< \alpha$

There is sufficient evidence that square footage affects house price.



F Test for Significance

- F Test statistic:

$$F_{STAT} = \frac{MSR}{MSE}$$

where

$$MSR = \frac{SSR}{1}$$
$$MSE = \frac{SSE}{n - 2}$$

where F_{STAT} follows an F distribution with 1 numerator and $(n - 2)$ denominator **degrees of freedom**

F-Test for Significance

Excel Output

Regression Statistics

| | |
|-------------------|----------|
| Multiple R | 0.76211 |
| R Square | 0.58082 |
| Adjusted R Square | 0.52842 |
| Standard Error | 41.33032 |
| Observations | 10 |

$$F_{\text{STAT}} = \frac{\text{MSR}}{\text{MSE}} = \frac{18934.9348}{1708.1957} = 11.0848$$

With 1 and 8 degrees of freedom

p-value for the F-Test

| ANOVA | df | SS | MS | F | Significance F |
|------------|----|------------|------------|---------|----------------|
| Regression | 1 | 18934.9348 | 18934.9348 | 11.0848 | 0.01039 |
| Residual | 8 | 13665.5652 | 1708.1957 | | |
| Total | 9 | 32600.5000 | | | |

F-Test for Significance

Minitab Output

| Analysis of Variance | | | | | |
|----------------------|----|-------|-------|-------|-------|
| Source | DF | SS | MS | F | P |
| Regression | 1 | 18935 | 18935 | 11.08 | 0.010 |
| Residual Error | 8 | 13666 | 1708 | | |
| Total | 9 | 32600 | | | |

p-value for
the F-Test

With 1 and 8 degrees
of freedom

$$F_{\text{STAT}} = \frac{\text{MSR}}{\text{MSE}} = \frac{18934.9348}{1708.1957} = 11.0848$$

F Test for Significance

(continued)

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

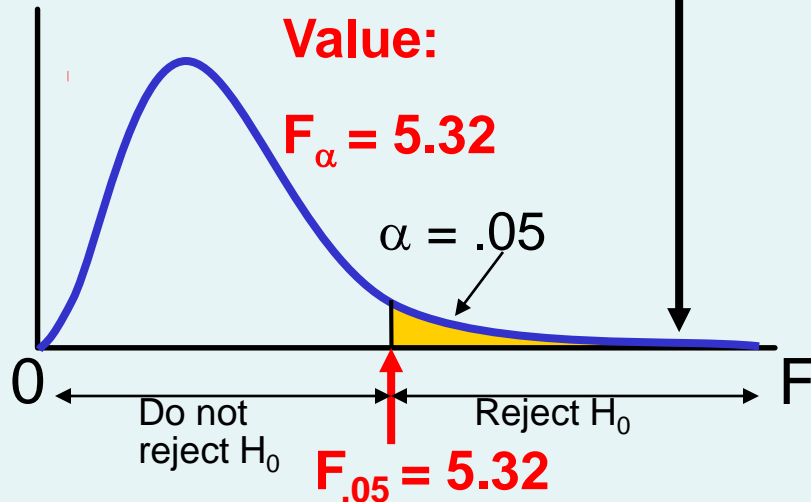
$$\alpha = .05$$

$$df_1 = 1 \quad df_2 = 8$$

Critical Value:

$$F_{\alpha} = 5.32$$

$$\alpha = .05$$



Test Statistic:

$$F_{\text{STAT}} = \frac{MSR}{MSE} = 11.08$$

Decision:

Reject H_0 at $\alpha = 0.05$

Conclusion:

There is sufficient evidence that house size affects selling price

Confidence Interval Estimate for the Slope

Confidence Interval Estimate of the Slope:

$$b_1 \pm t_{\alpha/2} S_{b_1}$$

d.f. = n - 2

Excel Printout for House Prices:

| | <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 |

At 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858)

Confidence Interval Estimate for the Slope

(continued)

| | <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 |

Since the units of the house price variable is \$1000s, we are 95% confident that the average impact on sales price is between \$33.74 and \$185.80 per square foot of house size

This 95% confidence interval **does not include 0**.

Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance



t Test for a Correlation Coefficient

- Hypotheses

$H_0: \rho = 0$ (no correlation between X and Y)

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

- Test statistic

$$t_{\text{STAT}} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

(with $n - 2$ degrees of freedom)

where

$$r = +\sqrt{r^2} \text{ if } b_1 > 0$$

$$r = -\sqrt{r^2} \text{ if } b_1 < 0$$



t-test For A Correlation Coefficient

(continued)

Is there evidence of a linear relationship between square feet and house price at the .05 level of significance?

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

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

$\alpha = .05$, $df = 10 - 2 = 8$

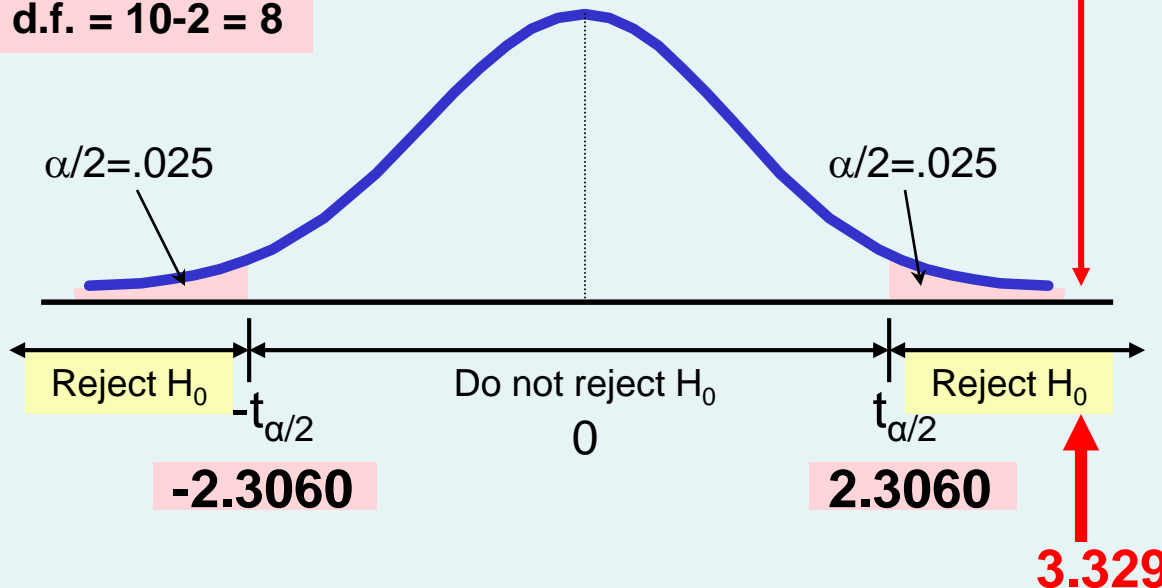
$$t_{\text{STAT}} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{.762 - 0}{\sqrt{\frac{1 - .762^2}{10 - 2}}} = 3.329$$

t-test For A Correlation Coefficient

(continued)

$$t_{\text{STAT}} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{.762 - 0}{\sqrt{\frac{1 - .762^2}{10 - 2}}} = 3.329$$

d.f. = 10 - 2 = 8

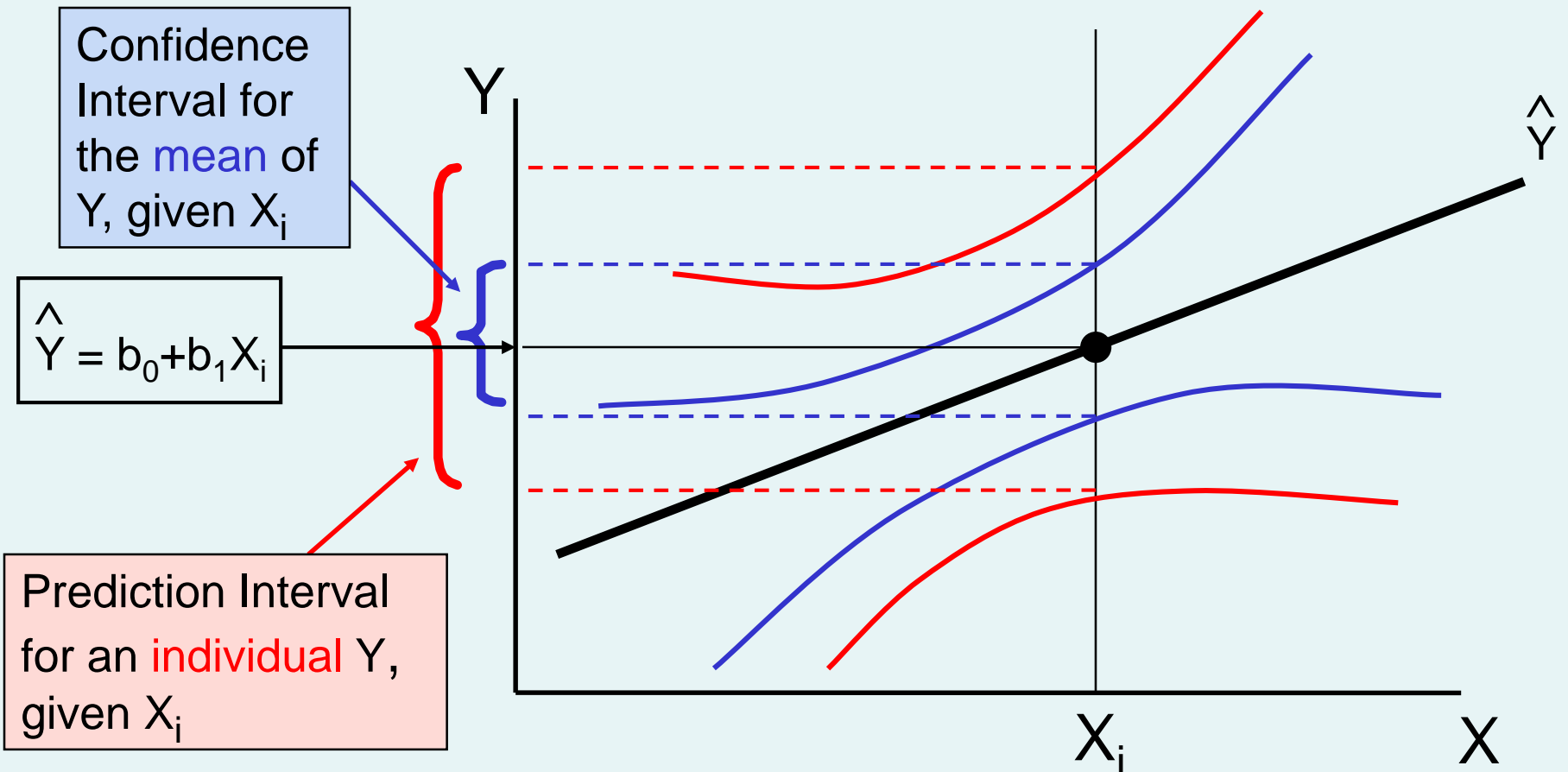


Decision:
Reject H_0

Conclusion:
There is **evidence** of a linear association at the 5% level of significance

Estimating Mean Values and Predicting Individual Values

Goal: Form intervals around Y to express uncertainty about the value of Y for a given X_i



Confidence Interval for the Average Y, Given X

Confidence interval estimate for the **mean value of Y** given a particular X_i

Confidence interval for $\mu_{Y|X=X_i}$:

$$\hat{Y} \pm t_{\alpha/2} S_{YX} \sqrt{h_i}$$

Size of interval varies according to distance away from mean, \bar{X}

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SSX} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}$$

Prediction Interval for an Individual Y, Given X

Prediction interval estimate for an **Individual value of Y** given a particular X_i

Prediction interval for $Y_{X=X_i}$:

$$\hat{Y} \pm t_{\alpha/2} S_{YX} \sqrt{1 + h_i}$$

This extra term adds to the interval width to reflect the added uncertainty for an individual case



Estimation of Mean Values: Example

Confidence Interval Estimate for $\mu_{Y|X=X_i}$

Find the 95% confidence interval for the mean price of 2,000 square-foot houses

Predicted Price $\hat{Y}_i = 317.85$ (\$1,000s)

$$\hat{Y} \pm t_{0.025} S_{YX} \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}} = 317.85 \pm 37.12$$

The confidence interval endpoints are 280.66 and 354.90, or from \$280,660 to \$354,900

Estimation of Individual Values: Example

Prediction Interval Estimate for $Y_{X=X_i}$

Find the 95% prediction interval for an individual house with 2,000 square feet

Predicted Price $\hat{Y}_i = 317.85$ (\$1,000s)

$$\hat{Y} \pm t_{0.025} S_{YX} \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}} = 317.85 \pm 102.28$$

The prediction interval endpoints are 215.50 and 420.07, or from \$215,500 to \$420,070



Finding Confidence and Prediction Intervals in Excel

- From Excel, use
PHStat | regression | simple linear regression ...
- Check the
“confidence and prediction interval for $X=$ ”
box and enter the X -value and confidence level
desired

Finding Confidence and Prediction Intervals in Excel

(continued)

| | A | B |
|----|----------------------------------------|-----------------|
| 1 | Confidence Interval Estimate | |
| 2 | | |
| 3 | Data | |
| 4 | X Value | 2000 |
| 5 | Confidence Level | 95% |
| 6 | | |
| 7 | Intermediate Calculations | |
| 8 | Sample Size | 10 |
| 9 | Degrees of Freedom | 8 |
| 10 | t Value | 2.306006 |
| 11 | Sample Mean | 1715 |
| 12 | Sum of Squared Difference | 1571500 |
| 13 | Standard Error of the Estimate | 41.33032 |
| 14 | h Statistic | 0.151686 |
| 15 | Average Predicted Y (YHat) | 317.7838 |
| 16 | | |
| 17 | For Average Predicted Y (YHat) | |
| 18 | Interval Half Width | 37.11952 |
| 19 | Confidence Interval Lower Limit | 280.6643 |
| 20 | Confidence Interval Upper Limit | 354.9033 |
| 21 | | |
| 22 | For Individual Response Y | |
| 23 | Interval Half Width | 102.2813 |
| 24 | Prediction Interval Lower Limit | 215.5025 |
| 25 | Prediction Interval Upper Limit | 420.0651 |

Input values

\hat{Y}

Confidence Interval Estimate for $\mu_{Y|X=X_i}$

Prediction Interval Estimate for $Y_{X=X_i}$

Finding Confidence and Prediction Intervals in Minitab

Confidence Interval Estimate for $\mu_{Y|X=X_i}$

Predicted Values for New Observations

| New Obs | Fit | SE Fit | 95% CI | 95% PI |
|---------|-------|--------|----------------|----------------|
| 1 | 317.8 | 16.1 | (280.7, 354.9) | (215.5, 420.1) |

\hat{Y}

Values of Predictors for New Observations

| New Obs | Square Feet |
|---------|-------------|
| 1 | 2000 |

Input values

Prediction Interval Estimate for $Y_{X=X_i}$



Pitfalls of Regression Analysis

- Lacking an awareness of the assumptions underlying least-squares regression
- Not knowing how to evaluate the assumptions
- Not knowing the alternatives to least-squares regression if a particular assumption is violated
- Using a regression model without knowledge of the subject matter
- Extrapolating outside the relevant range



Strategies for Avoiding the Pitfalls of Regression

- Start with a scatter plot of X vs. Y to observe possible relationship
- Perform residual analysis to check the assumptions
 - Plot the residuals vs. X to check for violations of assumptions such as homoscedasticity
 - Use a histogram, stem-and-leaf display, boxplot, or normal probability plot of the residuals to uncover possible non-normality



Strategies for Avoiding the Pitfalls of Regression

(continued)

- If there is violation of any assumption, use alternative methods or models
- If there is no evidence of assumption violation, then test for the significance of the regression coefficients and construct confidence intervals and prediction intervals
- Avoid making predictions or forecasts outside the relevant range



Chapter Summary

- Introduced types of regression models
- Reviewed assumptions of regression and correlation
- Discussed determining the simple linear regression equation
- Described measures of variation
- Discussed residual analysis

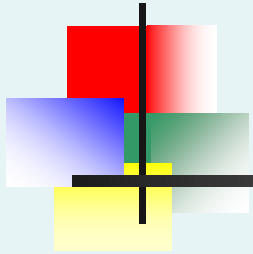


Chapter Summary

(continued)

- Described inference about the slope
- Discussed correlation -- measuring the strength of the association
- Addressed estimation of mean values and prediction of individual values
- Discussed possible pitfalls in regression and recommended strategies to avoid them

Business Statistics: A First Course Fifth Edition



Chapter 13

Multiple Regression



Learning Objectives

In this chapter, you learn:

- How to develop a multiple regression model
- How to interpret the regression coefficients
- How to determine which independent variables to include in the regression model
- How to determine which independent variables are more important in predicting a dependent variable
- How to use categorical variables in a regression model

The Multiple Regression Model

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

Multiple Regression Model with k Independent Variables:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

Diagram illustrating the components of the Multiple Regression Model equation:

- Y-intercept:** β_0
- Population slopes:** $\beta_1, \beta_2, \dots, \beta_k$
- Random Error:** ε_i

Multiple Regression Equation

The coefficients of the multiple regression model are estimated using sample data

Multiple regression equation with k independent variables:

Estimated (or predicted) value of Y

Estimated intercept

Estimated slope coefficients

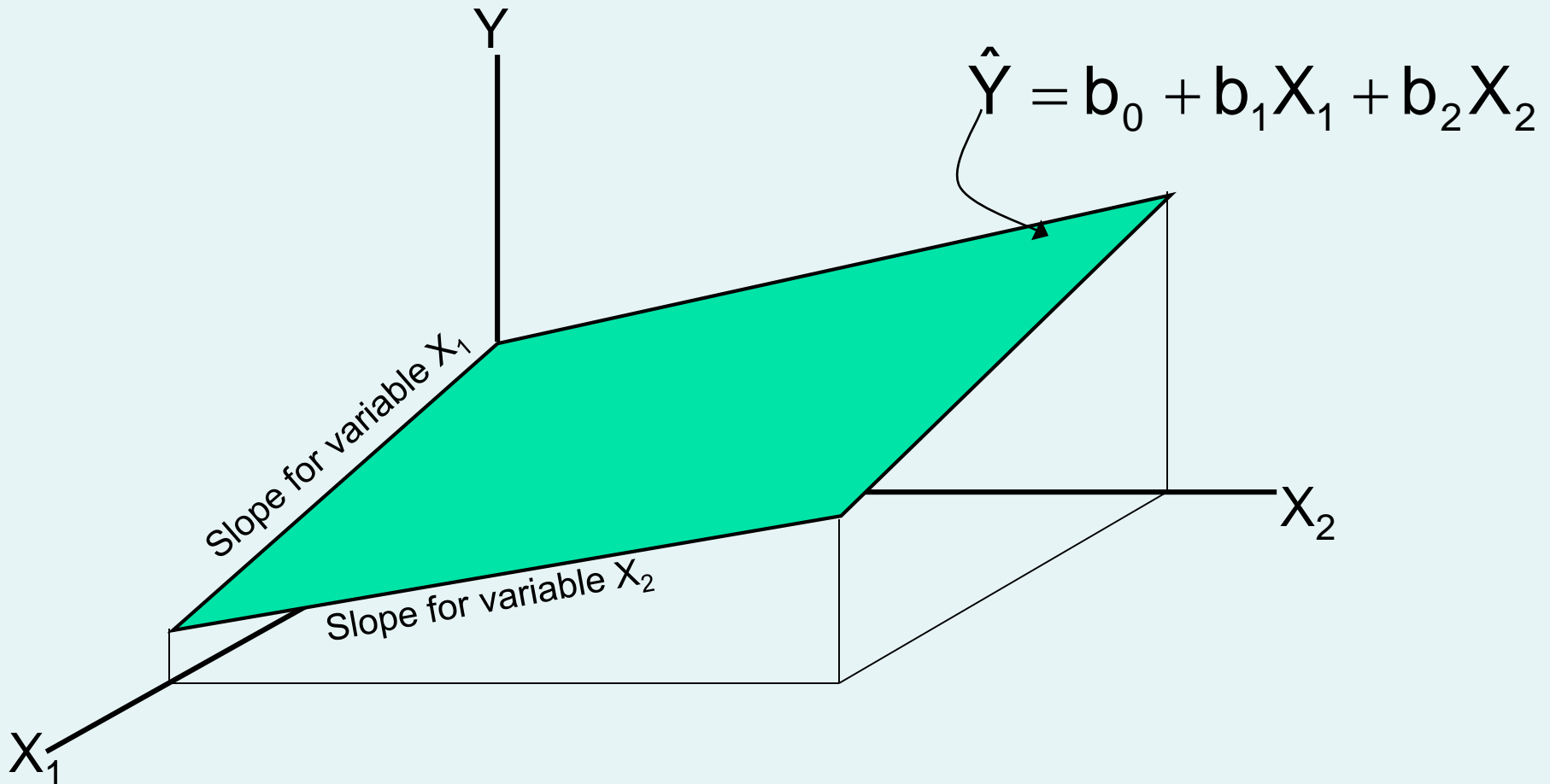
$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \dots + b_k X_{ki}$$

In this chapter we will use Excel or Minitab to obtain the regression slope coefficients and other regression summary measures.

Multiple Regression Equation

(continued)

Two variable model



Example: 2 Independent Variables

- A distributor of frozen dessert pies wants to evaluate factors thought to influence demand
 - Dependent variable: Pie sales (units per week)
 - Independent variables: $\left\{ \begin{array}{l} \text{Price (in \$)} \\ \text{Advertising (\$100's)} \end{array} \right.$
- Data are collected for 15 weeks



Pie Sales Example

| 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 equation:

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



Excel Multiple Regression Output



Regression Statistics

| | |
|-------------------|----------|
| Multiple R | 0.72213 |
| R Square | 0.52148 |
| Adjusted R Square | 0.44172 |
| Standard Error | 47.46341 |
| Observations | 15 |

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

| ANOVA | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
|------------|-----------|-----------|-----------|----------|-----------------------|
| Regression | 2 | 29460.027 | 14730.013 | 6.53861 | 0.01201 |
| Residual | 12 | 27033.306 | 2252.776 | | |
| Total | 14 | 56493.333 | | | |

| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> |
|-------------|---------------------|-----------------------|---------------|----------------|------------------|------------------|
| 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 |

Minitab Multiple Regression Output

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

The regression equation is

Sales = 307 - 25.0 Price + 74.1 Advertising

| Predictor | Coef | SE Coef | T | P |
|-------------|--------|---------|-------|-------|
| Constant | 306.50 | 114.30 | 2.68 | 0.020 |
| Price | -24.98 | 10.83 | -2.31 | 0.040 |
| Advertising | 74.13 | 25.97 | 2.85 | 0.014 |

S = 47.4634 R-Sq = 52.1% R-Sq(adj) = 44.2%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|-------|-------|------|-------|
| Regression | 2 | 29460 | 14730 | 6.54 | 0.012 |
| Residual Error | 12 | 27033 | 2253 | | |
| Total | 14 | 56493 | | | |

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 Equation 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 $X_2 = 3.5$

Predictions in Excel using PHStat

- PHStat | regression | multiple regression ...

| | A | B | C | D |
|----|------|-----------|-------|-------------|
| 1 | Week | Pie Sales | Price | Advertising |
| 2 | 1 | 350 | 5.5 | 3.3 |
| 3 | 2 | 460 | 7.5 | 3.3 |
| 4 | 3 | 350 | 8 | 3 |
| 5 | 4 | 430 | 8 | 4.5 |
| 6 | 5 | 350 | 6.8 | 3 |
| 7 | 6 | 380 | 7.5 | 4 |
| 8 | 7 | 430 | 4.5 | 3 |
| 9 | 8 | 470 | 6.4 | 3.7 |
| 10 | 9 | 450 | 7 | 3.5 |
| 11 | 10 | 490 | 5 | 4 |
| 12 | 11 | 340 | 7.2 | 3.5 |
| 13 | 12 | 300 | 7.9 | 3.2 |
| 14 | 13 | 440 | 5.9 | 4 |
| 15 | 14 | 450 | 5 | 3.5 |
| 16 | 15 | 300 | 7 | 2.7 |

Multiple Regression

Data

Y Variable Cell Range: Sheet1!\$B\$1:\$B\$16

X Variables Cell Range: Sheet1!\$C\$1:\$D\$16

First cells in both ranges contain label

Confidence level for regression coefficients: 95 %

Regression Tool Output Options

Regression Statistics Table

ANOVA and Coefficients Table

Residuals Table

Residual Plots

Output Options

Title: _____

Durbin-Watson Statistic

Coefficients of Partial Determination

Variance Inflationary Factor (VIF)

Confidence and Prediction Interval Estimates

Confidence level for interval estimates: 95 %

Help OK Cancel

Check the
“confidence and
prediction interval
estimates” box

Predictions in PHStat

(continued)

| | A | B |
|----|-----------------------------------------------------|-----------------|
| 1 | Confidence and Prediction Estimate Intervals | |
| 2 | | |
| 3 | Data | |
| 4 | Confidence Level | 95% |
| 5 | | |
| 6 | Price given value | 5.5 |
| 7 | Advertising given value | 3.5 |
| 8 | | |
| 20 | t Statistic | 2.178813 |
| 21 | Predicted Y (YHat) | 428.6216 |
| 22 | | |
| 23 | For Average Predicted Y (Yhat) | |
| 24 | Interval Half Width | 37.50306 |
| 25 | Confidence Interval Lower Limit | 391.1185 |
| 26 | Confidence Interval Upper Limit | 466.1246 |
| 27 | | |
| 28 | For Individual Response Y | |
| 29 | Interval Half Width | 110.0041 |
| 30 | Prediction Interval Lower Limit | 318.6174 |
| 31 | Prediction Interval Upper Limit | 538.6257 |

Input values

Predicted \hat{Y} value

Confidence interval for the mean value of Y, given these X values

Prediction interval for an individual Y value, given these X values

Predictions in Minitab

Predicted Values for New Observations

| New Obs | Fit | SE Fit | 95% CI | 95% PI |
|---------|-------|--------|----------------|----------------|
| 1 | 428.6 | 17.2 | (391.1, 466.1) | (318.6, 538.6) |

Predicted \hat{Y} value

Confidence interval for the mean value of Y, given these X values

Values of Predictors for New Observations

| New Obs | Price | Advertising |
|---------|-------|-------------|
| 1 | 5.50 | 3.50 |

Input values

Prediction interval for an individual Y value, given these X values





Coefficient of Multiple Determination

- Reports the proportion of total variation in Y explained by all X variables taken together

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

Multiple Coefficient of Determination In Excel



Regression Statistics

| | |
|-------------------|----------|
| Multiple R | 0.72213 |
| R Square | 0.52148 |
| Adjusted R Square | 0.44172 |
| Standard Error | 47.46341 |
| Observations | 15 |

$$r^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = .52148$$

52.1% of the variation in pie sales is explained by the variation in price and advertising

| 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 |

Multiple Coefficient of Determination In Minitab



The regression equation is
 Sales = 307 - 25.0 Price + 74.1 Advertising

| Predictor | Coef | SE Coef | T | P |
|-------------|--------|---------|-------|-------|
| Constant | 306.50 | 114.30 | 2.68 | 0.020 |
| Price | -24.98 | 10.83 | -2.31 | 0.040 |
| Advertising | 74.13 | 25.97 | 2.85 | 0.014 |

$$r^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = .52148$$

S = 47.4634 R-Sq = 52.1% R-Sq(adj) = 44.2%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|-------|-------|------|-------|
| Regression | 2 | 29460 | 14730 | 6.54 | 0.012 |
| Residual Error | 12 | 27033 | 2253 | | |
| Total | 14 | 56493 | | | |

52.1% of the variation in pie sales is explained by the variation in price and advertising



Adjusted r^2

- r^2 never decreases when a new X variable is added to the model
 - This can be a disadvantage when comparing models
- What is the net effect of adding a new variable?
 - We lose a degree of freedom when a new X variable is added
 - Did the new X variable add enough explanatory power to offset the loss of one degree of freedom?

Adjusted r^2

(continued)

- Shows the proportion of variation in Y explained by all X variables adjusted for the number of X variables used and sample size

$$r_{adj}^2 = 1 - \left[(1 - r^2) \left(\frac{n - 1}{n - k - 1} \right) \right]$$

(where n = sample size, k = number of independent variables)

- Penalize excessive use of unimportant independent variables
- Smaller than r^2
- Useful in comparing among models

Adjusted r^2 in Excel



| <i>Regression Statistics</i> | |
|------------------------------|----------|
| Multiple R | 0.72213 |
| R Square | 0.52148 |
| Adjusted R Square | 0.44172 |
| Standard Error | 47.46341 |
| Observations | 15 |

$$r_{adj}^2 = .44172$$

44.2% of the variation in pie sales is explained by the variation in price and advertising, taking into account the sample size and number of independent variables

| <i>ANOVA</i> | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
|--------------|-----------|-----------|-----------|----------|-----------------------|
| Regression | 2 | 29460.027 | 14730.013 | 6.53861 | 0.01201 |
| Residual | 12 | 27033.306 | 2252.776 | | |
| Total | 14 | 56493.333 | | | |

| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> |
|-------------|---------------------|-----------------------|---------------|----------------|------------------|------------------|
| 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 |

Adjusted r^2 in Minitab

The regression equation is
Sales = 307 - 25.0 Price + 74.1 Advertising

| Predictor | Coef | SE Coef | T | P |
|-------------|--------|---------|-------|-------|
| Constant | 306.50 | 114.30 | 2.68 | 0.020 |
| Price | -24.98 | 10.83 | -2.31 | 0.040 |
| Advertising | 74.13 | 25.97 | 2.85 | 0.014 |

S = 47.4634 R-Sq = 52.1% R-Sq(adj) = 44.2%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|-------|-------|------|-------|
| Regression | 2 | 29460 | 14730 | 6.54 | 0.012 |
| Residual Error | 12 | 27033 | 2253 | | |
| Total | 14 | 56493 | | | |

$$r_{\text{adj}}^2 = .44172$$

44.2% of the variation in pie sales is explained by the variation in price and advertising, taking into account the sample size and number of independent variables





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)

$H_1: \text{at least one } \beta_i \neq 0$ (at least one independent variable affects Y)



F Test for Overall Significance

- Test statistic:

$$F_{STAT} = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n - k - 1}}$$

where F_{STAT} has numerator d.f. = k and
denominator d.f. = $(n - k - 1)$

F Test for Overall Significance In Excel

(continued)



Regression Statistics

| | |
|-------------------|----------|
| Multiple R | 0.72213 |
| R Square | 0.52148 |
| Adjusted R Square | 0.44172 |
| Standard Error | 47.46341 |
| Observations | 15 |

$$F_{STAT} = \frac{MSR}{MSE} = \frac{14730.0}{2252.8} = 6.5386$$

With 2 and 12 degrees of freedom

P-value for the F Test

| 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 |

F Test for Overall Significance In Minitab



The regression equation is
 Sales = 307 - 25.0 Price + 74.1 Advertising

| Predictor | Coef | SE Coef | T | P |
|-------------|--------|---------|-------|-------|
| Constant | 306.50 | 114.30 | 2.68 | 0.020 |
| Price | -24.98 | 10.83 | -2.31 | 0.040 |
| Advertising | 74.13 | 25.97 | 2.85 | 0.014 |

S = 47.4634 R-Sq = 52.1% R-Sq(adj) = 44.2%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|-------|-------|------|-------|
| Regression | 2 | 29460 | 14730 | 6.54 | 0.012 |
| Residual Error | 12 | 27033 | 2253 | | |
| Total | 14 | 56493 | | | |

$$F_{\text{STAT}} = \frac{\text{MSR}}{\text{MSE}} = \frac{14730.0}{2252.8} = 6.5386$$

With 2 and 12 degrees of freedom

P-value for the F Test

F Test for Overall Significance

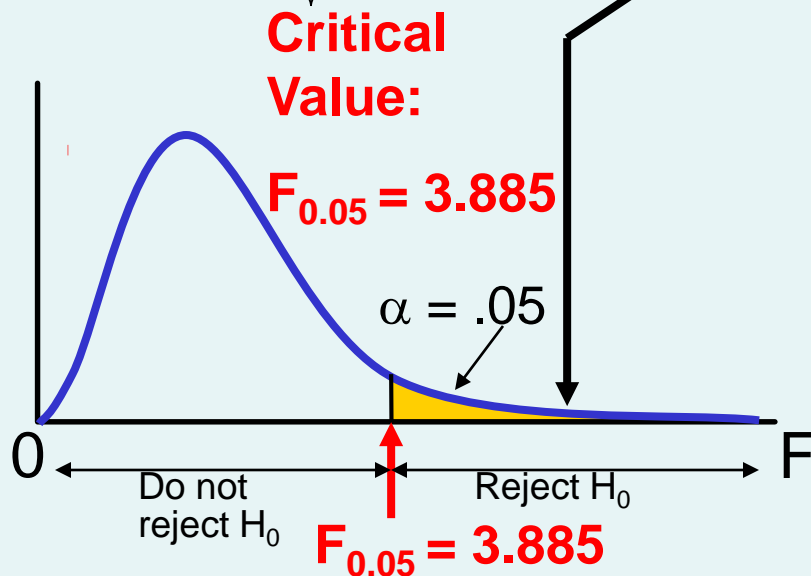
(continued)

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

$$H_1: \beta_1 \text{ and } \beta_2 \text{ not both zero}$$

$$\alpha = .05$$

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



Test Statistic:

$$F_{\text{STAT}} = \frac{MSR}{MSE} = 6.5386$$

Decision:

Since F_{STAT} test statistic is in the rejection region (p-value $< .05$), reject H_0

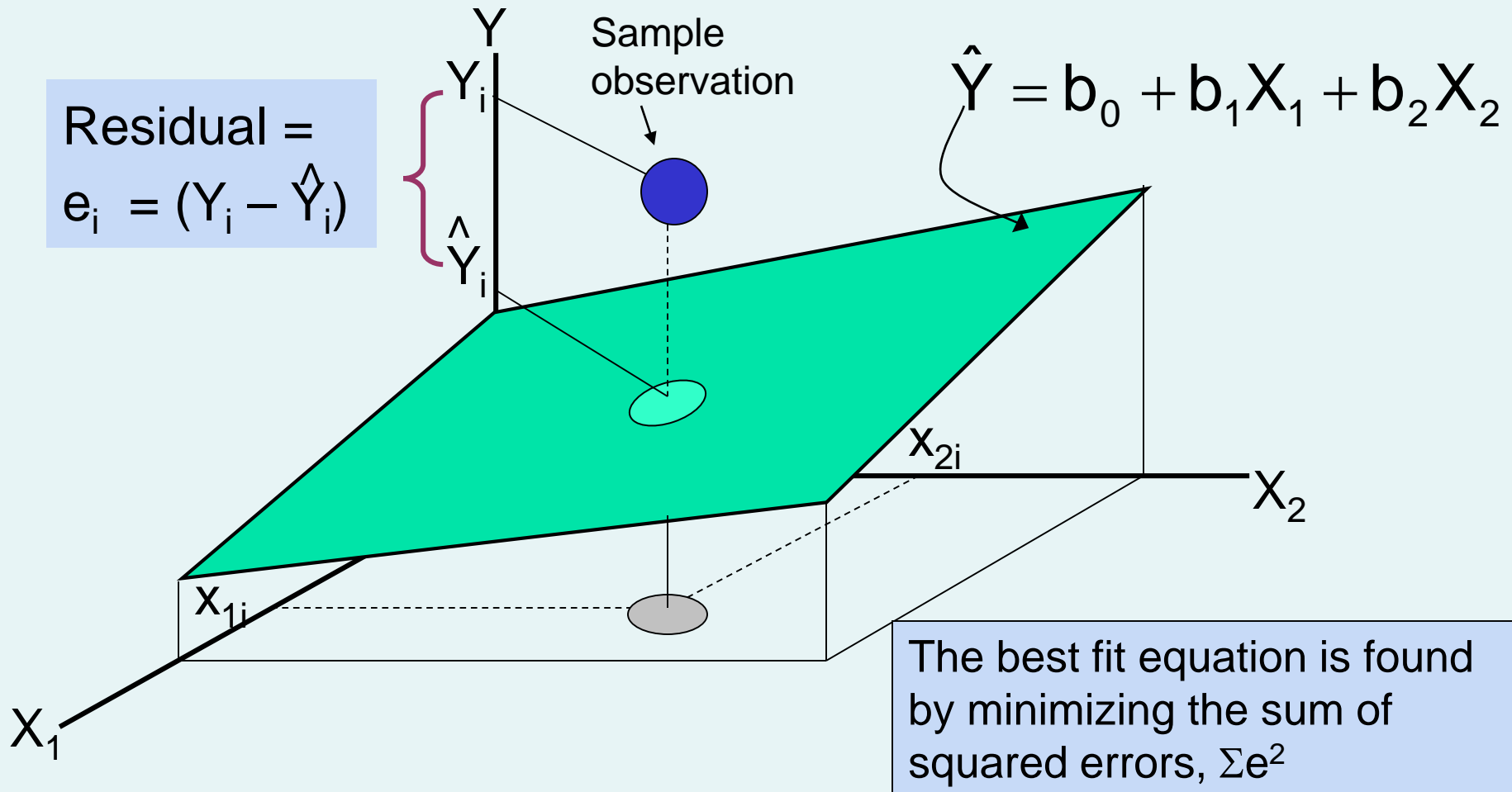
Conclusion:

There is evidence that at least one independent variable affects Y

Residuals in Multiple Regression

Two variable model

Residual =
 $e_i = (Y_i - \hat{Y}_i)$





Multiple Regression Assumptions

Errors (residuals) from the regression model:

$$e_i = (Y_i - \hat{Y}_i)$$

Assumptions:

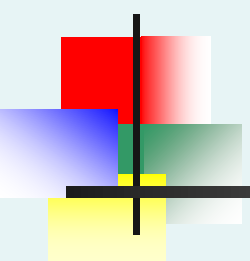
- Independence of errors
 - Error values are statistically independent
- Normality of errors
 - Error values are normally distributed for any given **set of X values**
- Equal Variance (also called Homoscedasticity)
 - The probability distribution of the errors has constant variance



Residual Plots Used in Multiple Regression

- These residual plots are used in multiple regression:
 - Residuals vs. \hat{Y}_i
 - Residuals vs. X_{1i}
 - Residuals vs. X_{2i}
 - Residuals vs. time (if time series data)

Use the residual plots to check for violations of regression assumptions



Are Individual Variables Significant?

- Use t tests of individual variable slopes
- Shows if there is a linear relationship between the variable X_j and Y holding constant the effects of other X variables
- Hypotheses:

- $H_0: \beta_j = 0$ (no linear relationship)
- $H_1: \beta_j \neq 0$ (linear relationship does exist between X_j and Y)

Are Individual Variables Significant?

(continued)

$H_0: \beta_j = 0$ (no linear relationship)

$H_1: \beta_j \neq 0$ (linear relationship does exist between X_j and Y)

Test Statistic:

$$t_{STAT} = \frac{b_j - 0}{S_{b_j}}$$

$$(df = n - k - 1)$$

Are Individual Variables Significant? Excel Output

(continued)



| Regression Statistics | | | | | | |
|-----------------------|---------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------|----------------|------------------|-----------------------|
| Multiple R | 0.72213 | <p>t Stat for Price is $t_{STAT} = -2.306$, with p-value .0398</p> <p>t Stat for Advertising is $t_{STAT} = 2.855$, with p-value .0145</p> | | | | |
| R Square | 0.52148 | | | | | |
| Adjusted R Square | 0.44172 | | | | | |
| Standard Error | 47.46341 | | | | | |
| Observations | 15 | | | | | |
| ANOVA | | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
| Regression | 2 | 29460.027 | 14730.013 | 6.53861 | 0.01201 | |
| Residual | 12 | 27033.306 | 2252.776 | | | |
| Total | 14 | 56493.333 | | | | |
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> |
| 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 |

Are Individual Variables Significant?

Minitab Output



The regression equation is
Sales = 307 - 25.0 Price + 74.1 Advertising

| Predictor | Coef | SE Coef | T | P |
|-------------|--------|---------|-------|-------|
| Constant | 306.50 | 114.30 | 2.68 | 0.020 |
| Price | -24.98 | 10.83 | -2.31 | 0.040 |
| Advertising | 74.13 | 25.97 | 2.85 | 0.014 |

S = 47.4634 R-Sq = 52.1% R-Sq(adj) = 44.2%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|-------|-------|------|-------|
| Regression | 2 | 29460 | 14730 | 6.54 | 0.012 |
| Residual Error | 12 | 27033 | 2253 | | |
| Total | 14 | 56493 | | | |

t Stat for Price is $t_{STAT} = -2.306$, with p-value .0398

t Stat for Advertising is $t_{STAT} = 2.855$, with p-value .0145

Inferences about the Slope: t Test Example

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

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

$$\alpha = .05$$

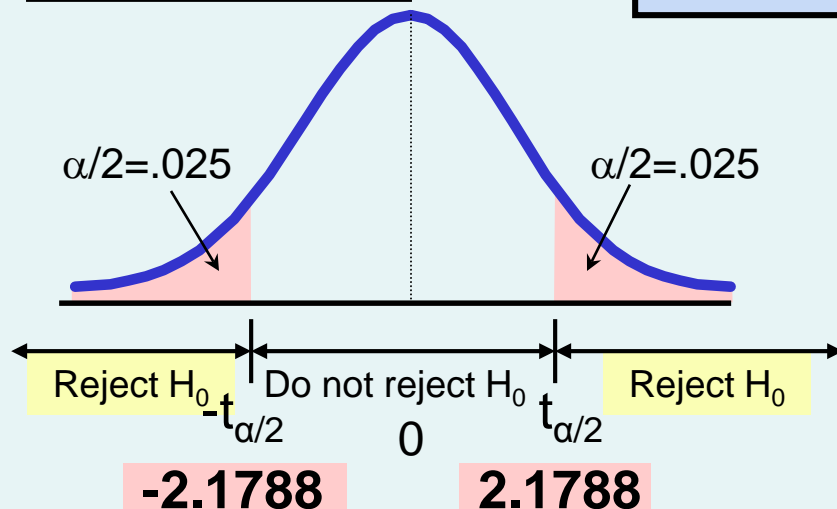
$$t_{\alpha/2} = 2.1788$$

From the Excel and Minitab output:

For Price $t_{\text{STAT}} = -2.306$, with p-value .0398

For Advertising $t_{\text{STAT}} = 2.855$, with p-value .0145

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_j

$$b_j \pm t_{\alpha/2} S_{b_j}$$

where t has
($n - k - 1$) d.f.

| | <i>Coefficients</i> | <i>Standard Error</i> |
|-------------|---------------------|-----------------------|
| Intercept | 306.52619 | 114.25389 |
| Price | -24.97509 | 10.83213 |
| Advertising | 74.13096 | 25.96732 |

Here, t has
($15 - 2 - 1$) = 12 d.f.

Example: Form a 95% confidence interval for the effect of changes in price (X_1) on pie sales:

$$-24.975 \pm (2.1788)(10.832)$$

So the interval is (-48.576 , -1.374)

(This interval does not contain zero, so price has a significant effect on sales)

Confidence Interval Estimate for the Slope

(continued)

Confidence interval for the population slope β_j

| | <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 |

Example: Excel output also reports these interval endpoints:

Weekly sales are estimated to be reduced by between 1.37 to 48.58 pies for each increase of \$1 in the selling price, holding the effect of price constant



Using Dummy Variables

- A dummy variable is a categorical independent variable with two levels:
 - yes or no, on or off, male or female
 - coded as 0 or 1
- Assumes the slopes associated with numerical independent variables do not change with the value for the categorical variable

Dummy-Variable Example (with 2 Levels)

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

Let:

Y = pie sales

X_1 = price

X_2 = holiday ($X_2 = 1$ if a holiday occurred during the week)
($X_2 = 0$ if there was no holiday that week)



Dummy-Variable Example (with 2 Levels)

(continued)

| | |
|-------------------------------------------------------------|-------------------|
| $\hat{Y} = b_0 + b_1 X_1 + b_2 (1) = (b_0 + b_2) + b_1 X_1$ | Holiday |
| $\hat{Y} = b_0 + b_1 X_1 + b_2 (0) = b_0 + b_1 X_1$ | No Holiday |

**Different
intercept**

**Same
slope**

Y (sales)

$b_0 + b_2$

b_0

Holiday ($X_2 = 1$)

No Holiday ($X_2 = 0$)



If $H_0: \beta_2 = 0$ is rejected, then “Holiday” has a significant effect on pie sales

X_1 (Price)

Interpreting the Dummy Variable Coefficient (with 2 Levels)

Example:

$$\text{Sales} = 300 - 30(\text{Price}) + 15(\text{Holiday})$$

Sales: number of pies sold per week

Price: pie price in \$

Holiday: $\begin{cases} 1 & \text{If a holiday occurred during the week} \\ 0 & \text{If no holiday occurred} \end{cases}$

$b_2 = 15$: on average, sales were 15 pies greater in weeks with a holiday than in weeks without a holiday, given the same price



Interaction Between Independent Variables

- Hypothesizes interaction between pairs of X variables
 - Response to one X variable may vary at different levels of another X variable
- Contains two-way cross product terms

- $$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$
$$= b_0 + b_1X_1 + b_2X_2 + b_3(X_1X_2)$$

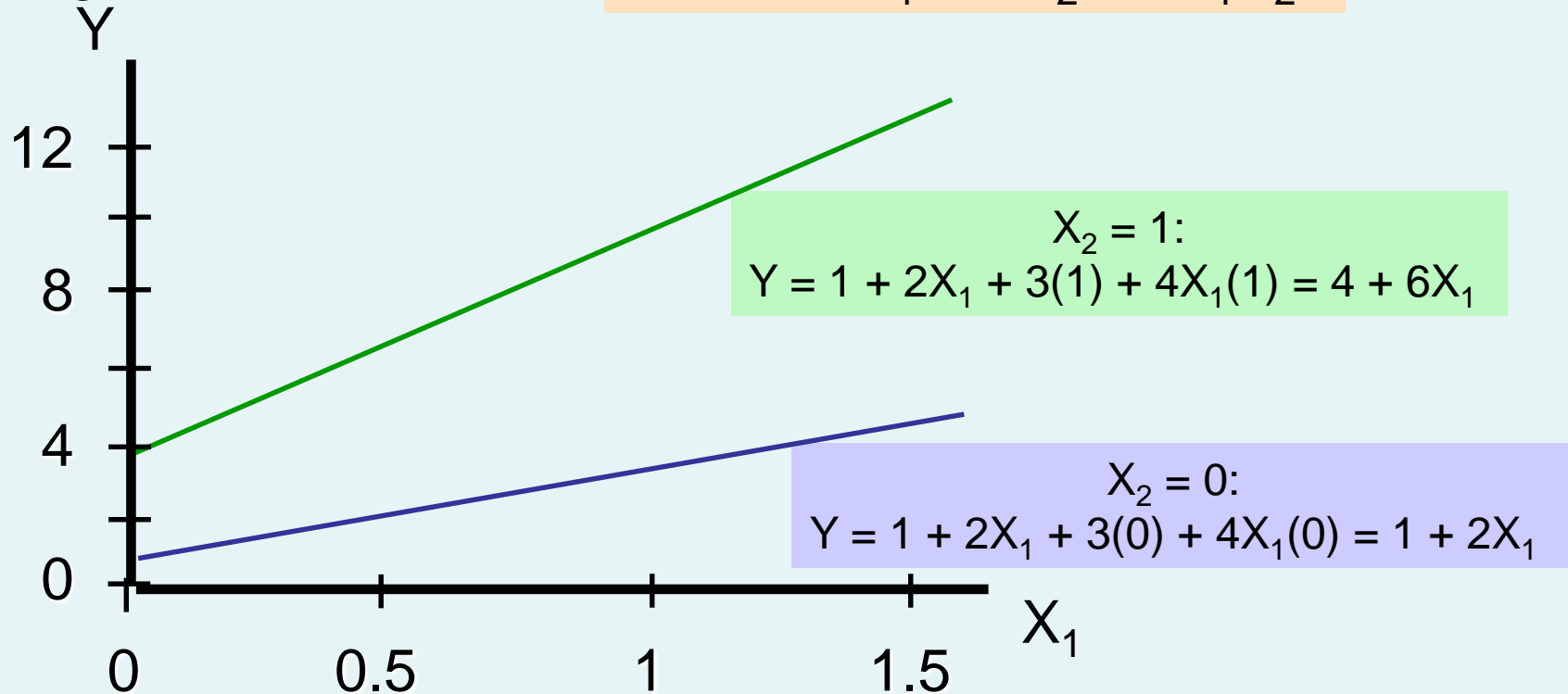


Effect of Interaction

- Given:
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$
- Without interaction term, effect of X_1 on Y is measured by β_1
- With interaction term, effect of X_1 on Y is measured by $\beta_1 + \beta_3 X_2$
- Effect changes as X_2 changes

Interaction Example

Suppose X_2 is a dummy variable and the estimated regression equation is $\hat{Y} = 1 + 2X_1 + 3X_2 + 4X_1X_2$



Slopes are different if the effect of X_1 on Y depends on X_2 value



Significance of Interaction Term

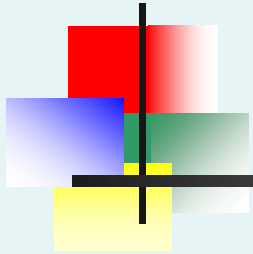
- Tested by utilizing the t-test on the coefficient associated with the interaction term.



Chapter Summary

- Developed the multiple regression model
- Tested the significance of the multiple regression model
- Discussed adjusted r^2
- Discussed using residual plots to check model assumptions
- Tested individual regression coefficients
- Used dummy variables
- Evaluated interaction effects

Business Statistics: A First Course Fifth Edition



Chapter 14

Statistical Applications in Quality Management

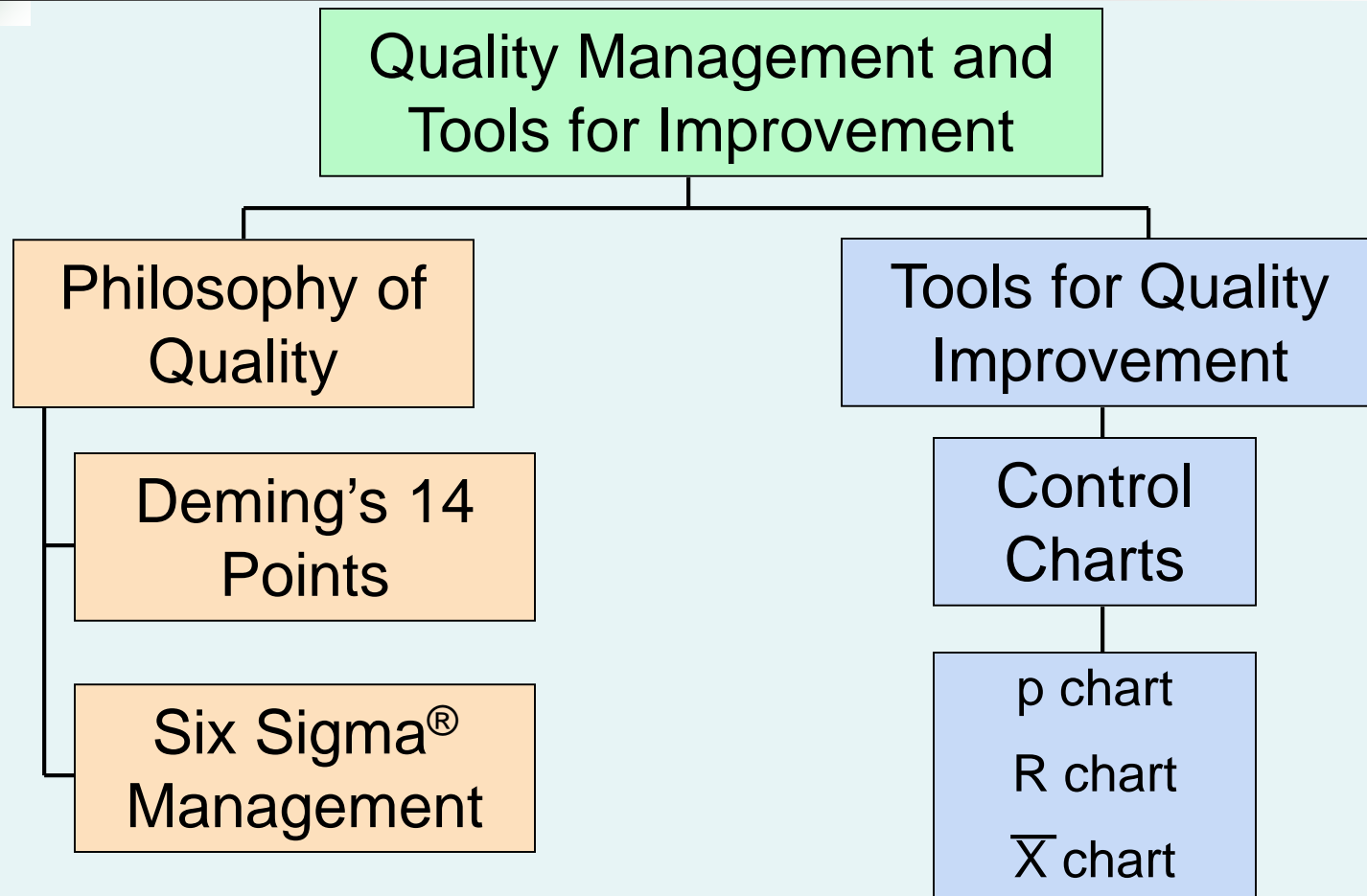


Learning Objectives

In this chapter, you learn:

- How to construct various control charts
- Which control charts to use for a particular type of data
- The basic themes of quality management and Deming's 14 points
- The basic aspects of Six Sigma

Chapter Overview





Theory of Control Charts

- A **process** is the value added transformation of inputs to outputs
- **Control Charts** are used to monitor variation in a measured value from a process
- **Inherent variation** refers to process variation that exists naturally. This variation can be reduced but not eliminated



Theory of Control Charts

(continued)

- Control charts indicate when changes in data are due to:
 - **Special or assignable causes**
 - Fluctuations not inherent to a process
 - Represents problems to be corrected
 - Data outside control limits or trend
 - **Chance or common causes**
 - Inherent random variations
 - Consist of numerous small causes of random variability



Process Variation


Total Process
Variation

=

Common Cause
Variation

+

Special Cause
Variation

- 
- Variation is natural; inherent in the world around us
 - No two products or service experiences are exactly the same
 - With a fine enough gauge, all things can be seen to differ



Total Process Variation

Total Process
Variation

=

Common Cause
Variation

+

Special Cause
Variation

↓

Variation is often due to differences in:

- People
- Machines
- Materials
- Methods
- Measurement
- Environment



Common Cause Variation

Total Process
Variation

=

Common Cause
Variation

+

Special Cause
Variation

↓

Common cause variation

- naturally occurring and expected
- the result of normal variation in materials, tools, machines, operators, and the environment



Special Cause Variation

Total Process
Variation

=

Common Cause
Variation

+

Special Cause
Variation



Special cause variation

- abnormal or unexpected variation
- has an assignable cause
- variation beyond what is considered inherent to the process



Two Kinds Of Errors

- Treating common cause variation as special cause variation
 - Results in over adjusting known as **tampering**
 - Increases process variation
- Treating special cause variation as common cause variation
 - Results in not taking corrective action when it should be taken
- Utilizing control charts greatly reduces the chance of committing either of these errors



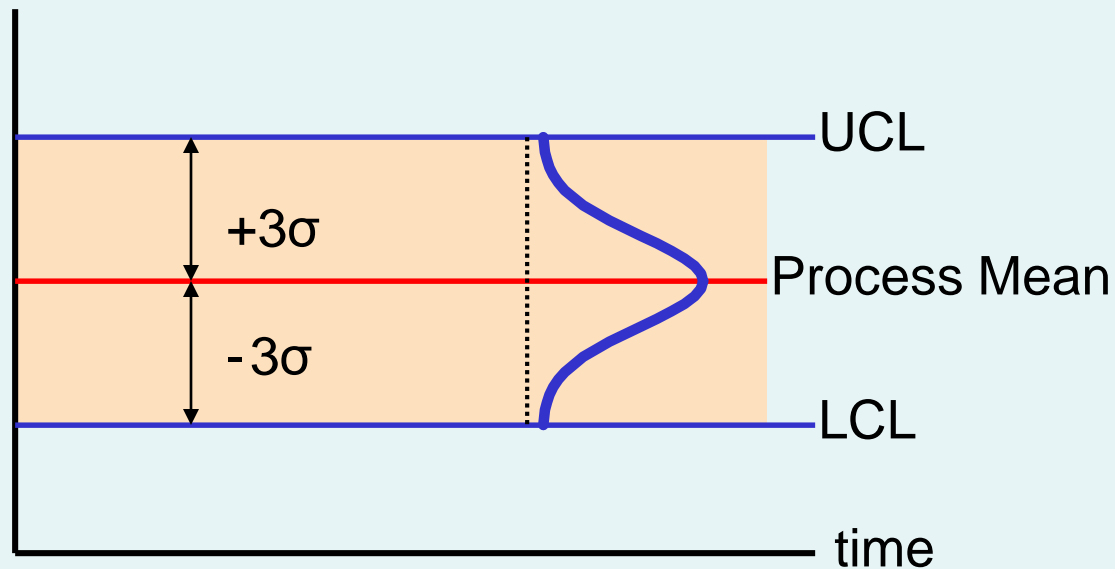
Gathering Data For A Control Chart

- Collect samples from the output of a process over time
 - Each sample is called a **subgroup**
 - Often subgroups are equally spaced over time
- For each subgroup calculate a sample statistic associated with a Critical To Quality (**CTQ**) variable
- Frequently used sample statistics are:
 - For a categorical CTQ -- The proportion of items with “an event of interest” such as non-conforming
 - For a numerical CTQ -- The mean of the sample and the range of the sample

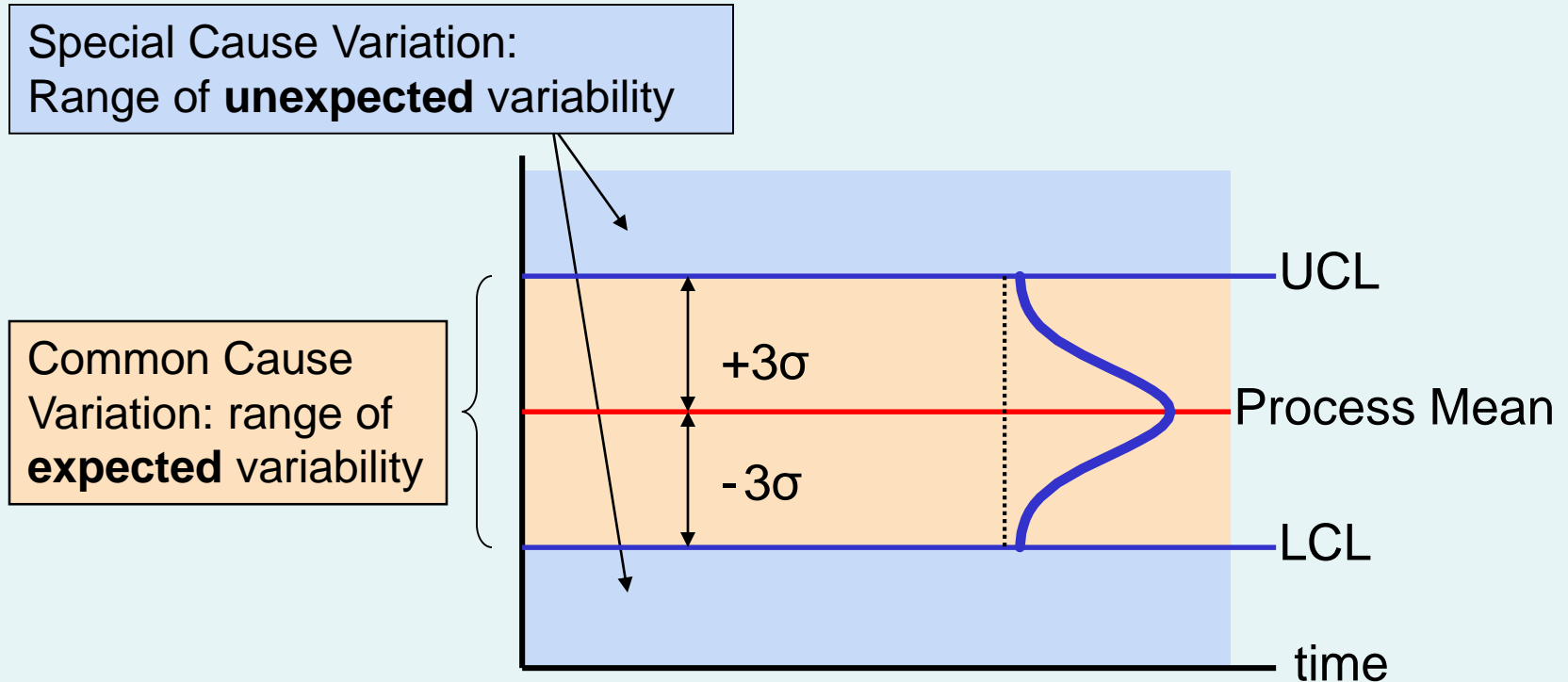
Control Limits

Forming the **Upper control limit (UCL)** and the **Lower control limit (LCL)**:

$$\text{UCL} = \text{Process Mean} + 3 \text{ Standard Deviations}$$
$$\text{LCL} = \text{Process Mean} - 3 \text{ Standard Deviations}$$



Control Chart Basics



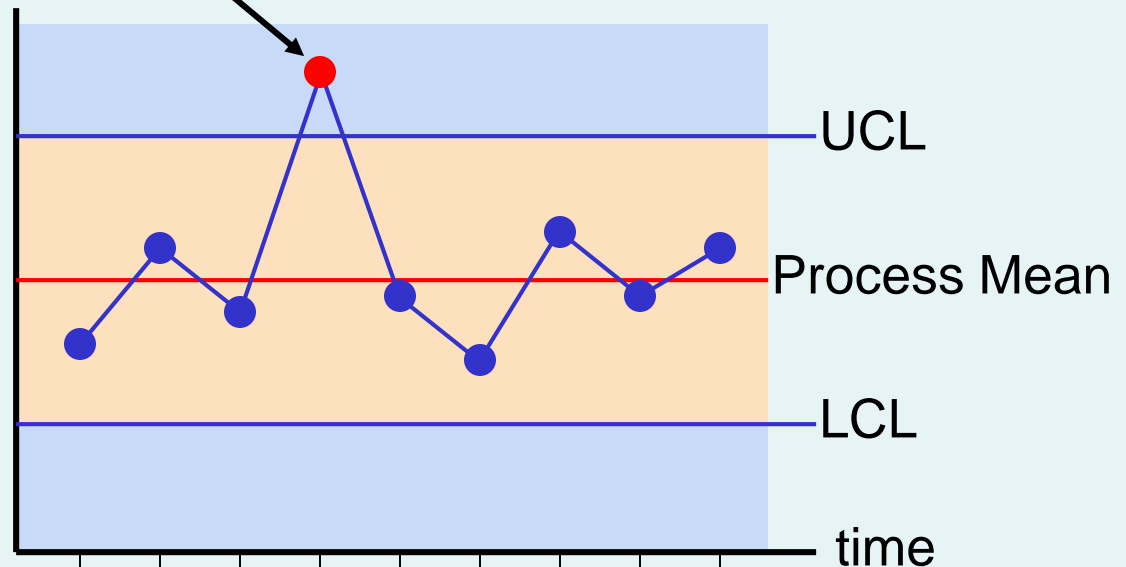
UCL = Process Mean + 3 Standard Deviations

LCL = Process Mean - 3 Standard Deviations

Process Variability

Special Cause of Variation:
A measurement this far from the process mean is very unlikely if only expected variation is present

$\pm 3\sigma \rightarrow 99.7\%$ of process values should be in this range



$UCL = \text{Process Mean} + 3 \text{ Standard Deviations}$
 $LCL = \text{Process Mean} - 3 \text{ Standard Deviations}$



Using Control Charts

- Control Charts are used to check for process control

H_0 : The process is in control

i.e., variation is only due to common causes

H_1 : The process is out of control

i.e., special cause variation exists

- If the process is found to be out of control, steps should be taken to find and eliminate the special causes of variation

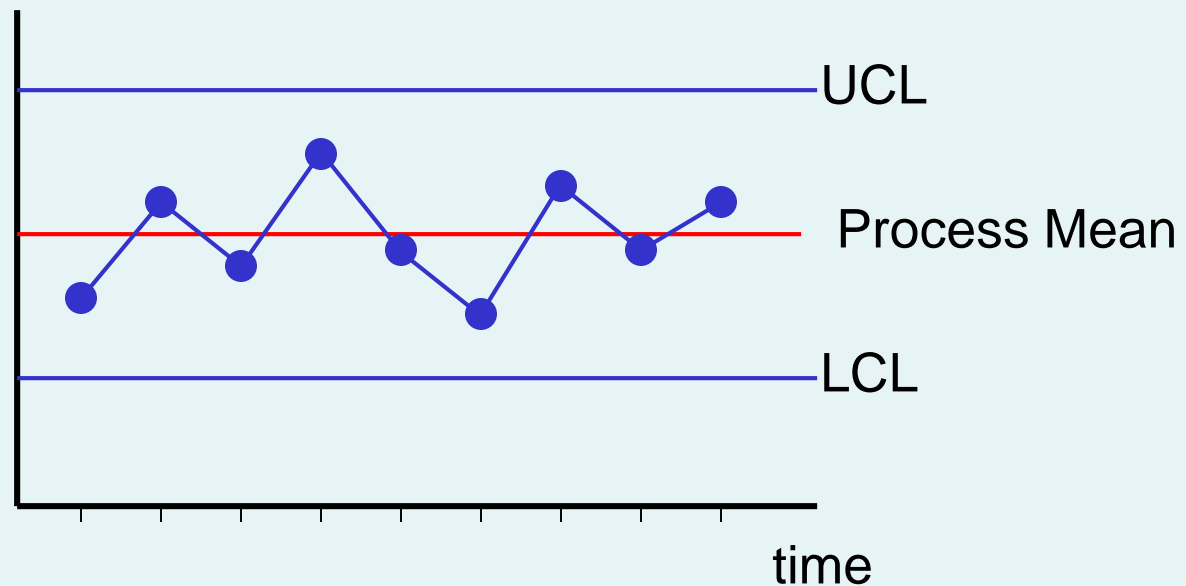


In-control Process

- A process is said to be in control when the control chart does not indicate any out-of-control condition
 - Contains only common causes of variation
 - If the common causes of variation is small, then control chart can be used to monitor the process
 - If the common causes of variation is too large, you need to alter the process

Process In Control

- **Process in control:** points are randomly distributed around the center line and all points are within the control limits





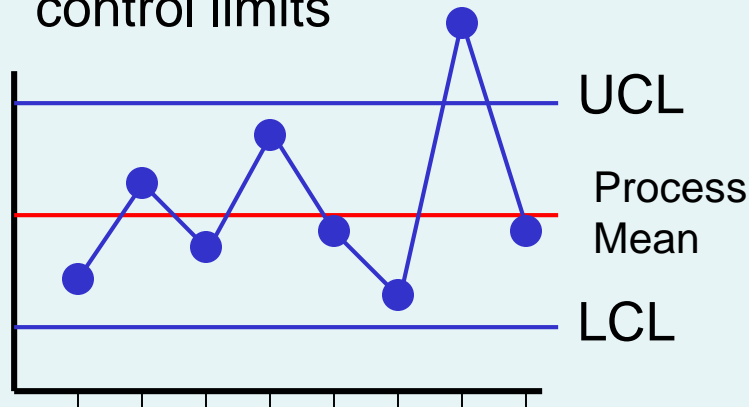
Process Not in Control

Out-of-control conditions:

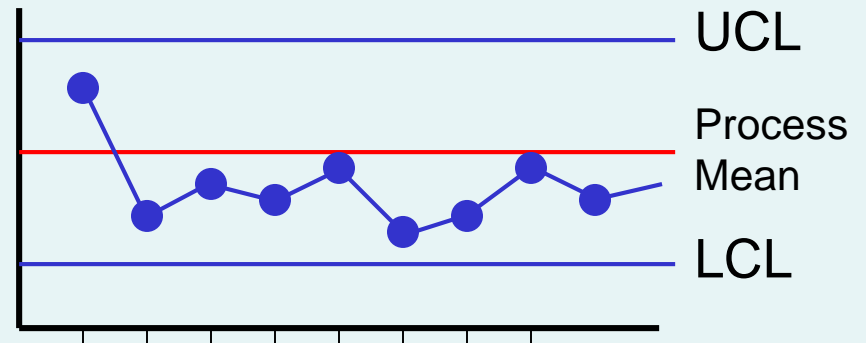
- One or more points **outside control limits**
- 8 or more points in a row **on one side** of the center line

Process Not in Control

- One or more points outside control limits



- Eight or more points in a row on one side of the center line

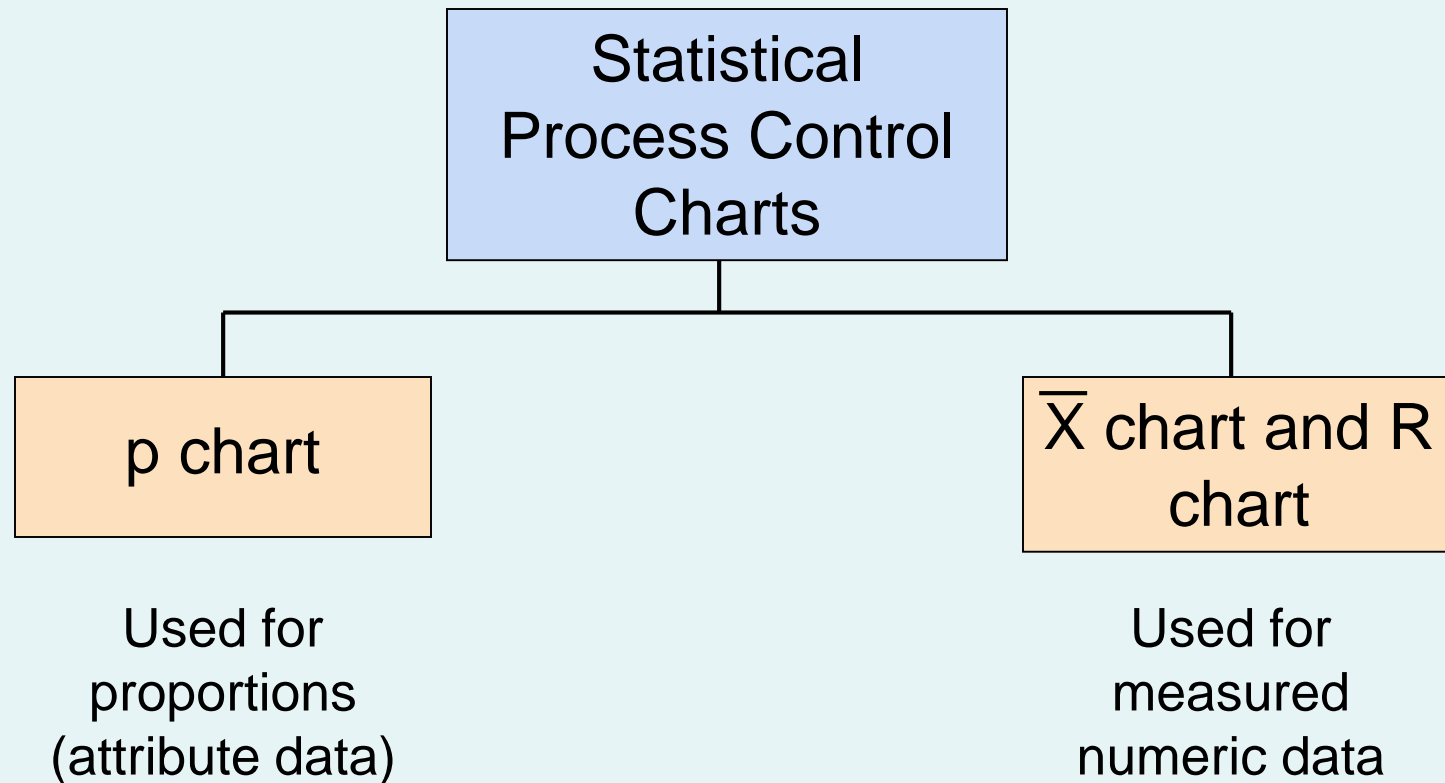




Out-of-control Processes

- When the control chart indicates an out-of-control condition (a point outside the control limits or 8 points in a row on the one side of the centerline, for example)
 - Contains both common causes of variation and special causes of variation
 - The special causes of variation must be identified
 - If detrimental to the quality, special causes of variation must be removed
 - If increases quality, special causes must be incorporated into the process design

Statistical Process Control Charts





p Chart

- Control chart for **proportions**
 - Is an **attribute chart**
- Shows proportion of nonconforming items
 - Example -- Computer chips: Count the number of non conforming chips and divide by total chips inspected
 - Chip is either conforming or not non conforming
 - Finding a non conforming chip can be classified as an “event of interest”



p Chart

(continued)

- Used with equal or unequal sample sizes (subgroups) over time
 - Unequal sample sizes should not differ by more than $\pm 25\%$ from average sample size
 - Easier to develop with equal sample sizes



Creating a p Chart

- Calculate subgroup proportions
- Graph subgroup proportions
- Compute average proportion
- Compute the upper and lower control limits
- Add centerline and control limits to graph



p Chart Example

| Subgroup number | Sample size | Number of events of interest | Sample Proportion, p_s |
|-----------------|-------------|------------------------------|-----------------------------------------|
| 1 | 150 | 15 | .1000 |
| 2 | 150 | 12 | .0800 |
| 3 | 150 | 17 | .1133 |
| ... | | ... | ... |
| | | | Average subgroup proportion = \bar{p} |



Average of Subgroup Proportions

The average of subgroup proportions = \bar{p}

If equal sample sizes:

$$\bar{p} = \frac{\sum_{i=1}^k p_i}{k}$$

where:

p_i = sample proportion
for subgroup i

k = number of subgroups
of size n

If unequal sample sizes:

$$\bar{p} = \frac{\sum_{i=1}^k X_i}{\sum_{i=1}^k n_i}$$

where:

X_i = the number of items with an
event of interest in sample i

$\sum n_i$ = total number of items
sampled in k samples

Computing Control Limits

- The upper and lower control limits for a p chart are

$$\begin{aligned} \text{UCL} &= \text{Average Proportion} + 3 \text{ Standard Deviations} \\ \text{LCL} &= \text{Average Proportion} - 3 \text{ Standard Deviations} \end{aligned}$$

- The standard deviation for the subgroup proportions is

$$\sqrt{\frac{(\bar{p})(1 - \bar{p})}{\bar{n}}}$$

- Where \bar{n} is the average of the subgroup sample sizes or the common n when subgroup sample sizes are all equal.

Computing Control Limits

(continued)

- The upper and lower control limits for the p chart are

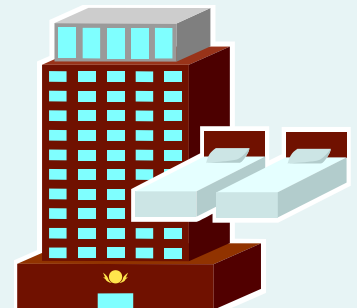
$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}}$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}}$$

Proportions are never negative, so if the calculated lower control limit is negative, set $LCL = 0$

p Chart Example

You are the manager of a 500-room hotel. You want to achieve the highest level of service. For seven days, you collect data on the readiness of 200 rooms. Is the process in control?





p Chart Example: Hotel Data

| Day | # Rooms | # Not Ready | Proportion |
|-----|---------|-------------|------------|
| 1 | 200 | 16 | 0.080 |
| 2 | 200 | 7 | 0.035 |
| 3 | 200 | 21 | 0.105 |
| 4 | 200 | 17 | 0.085 |
| 5 | 200 | 25 | 0.125 |
| 6 | 200 | 19 | 0.095 |
| 7 | 200 | 16 | 0.080 |



p Chart Control Limits Solution

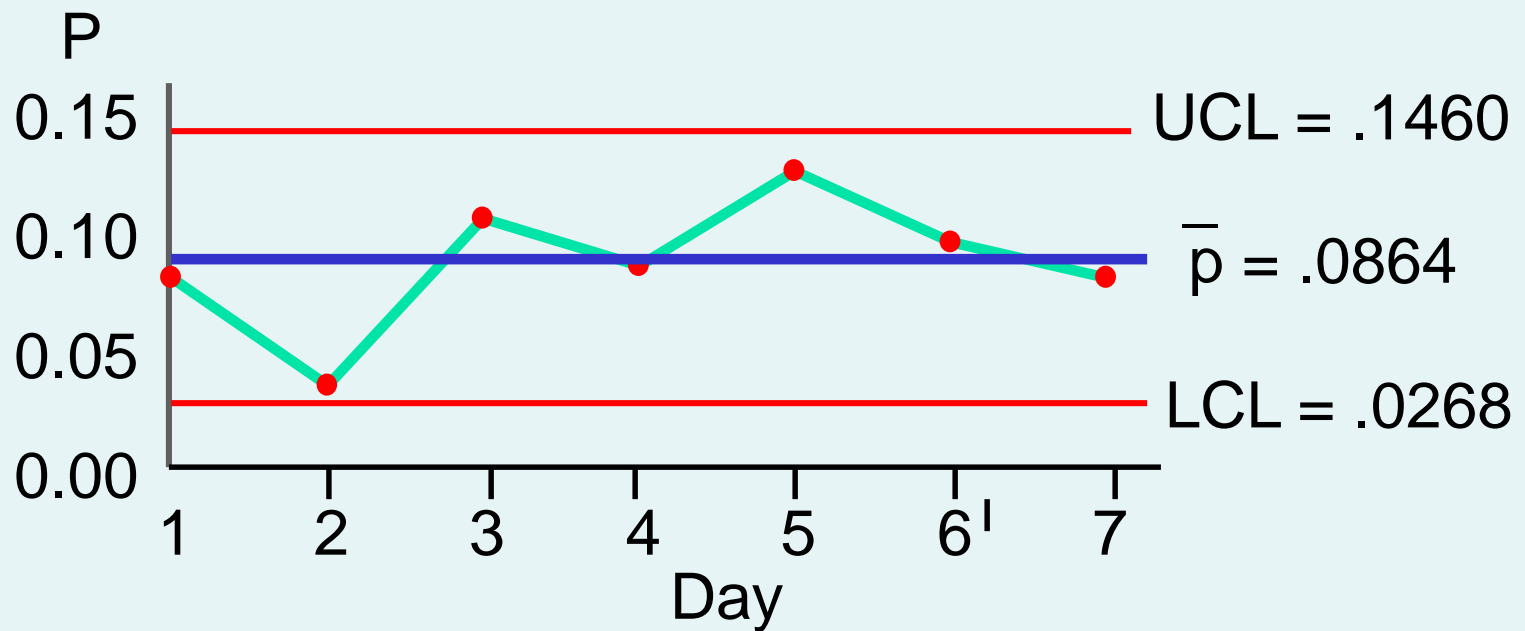
$$\bar{p} = \frac{\sum_{i=1}^k X_i}{\sum_{i=1}^k n_i} = \frac{16 + 7 + \dots + 16}{200 + 200 + \dots + 200} = \frac{121}{1400} = .0864$$

$$\bar{n} = \frac{\sum_{i=1}^k n_i}{k} = \frac{200 + 200 + \dots + 200}{7} = 200$$



$$\begin{aligned} \text{UCL} &= \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}} = .0864 + 3\sqrt{\frac{.0864(1-.0864)}{200}} = .1460 \\ \text{LCL} &= \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{\bar{n}}} = .0864 - 3\sqrt{\frac{.0864(1-.0864)}{200}} = .0268 \end{aligned}$$

p Chart Control Chart Solution



Individual points are distributed around \bar{p} without any pattern. The process is in control. Any improvement in the process must come from reduction of common-cause variation, which is the responsibility of management.



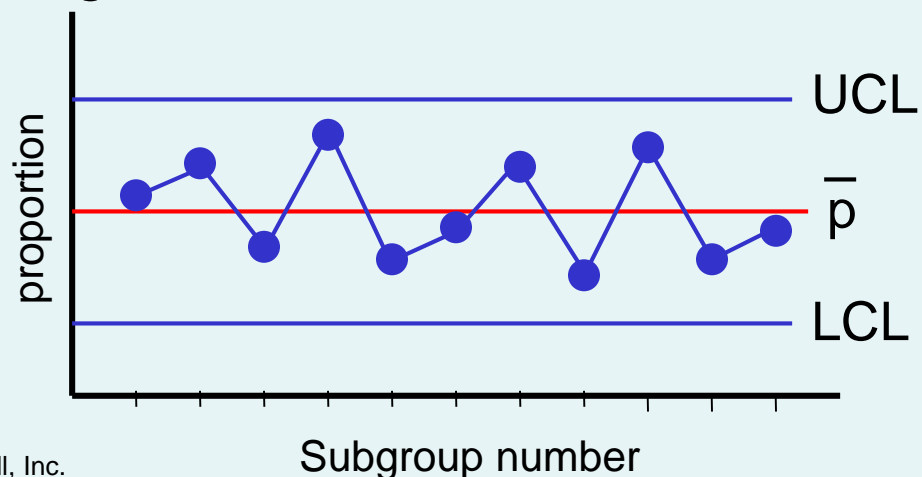
Understanding Process Variability: Red Bead Experiment

The experiment:

- From a box with 20% red beads and 80% white beads, have “workers” scoop out 50 beads
- Tell the workers their job is to get white beads
- 10 red beads out of 50 (20%) is the expected value. Scold workers who get more than 10, praise workers who get less than 10
- Some workers will get better over time, some will get worse

Morals of the Red Bead Experiment

1. Variation is an inherent part of any process.
2. The system is primarily responsible for worker performance.
3. Only management can change the system.
4. Some workers will always be above average, and some will be below.
5. Setting unrealistic goals is detrimental to a firm's well-being.





R chart and \bar{X} chart

- Used for measured numeric data from a process
- Start with at least 20 subgroups of observed values
- Subgroups usually contain 3 to 6 observations each
- For the process to be in control, both the R chart and the \bar{X} -bar chart must be in control

Example: Subgroups

- Process measurements:

| Subgroup number | Individual measurements (subgroup size = 4) | | | | Subgroup measures | |
|-----------------|------------------------------------------------|-----|-----|-----|--------------------------------------|---------------------------------|
| | | | | | Mean, \bar{X} | Range, R |
| 1 | 15 | 17 | 15 | 11 | → 14.5 | 6 |
| 2 | 12 | 16 | 9 | 15 | → 13.0 | 7 |
| 3 | 17 | 21 | 18 | 20 | → 19.0 | 4 |
| ... | ... | ... | ... | ... | ... | ... |
| | | | | | Mean subgroup mean = $\bar{\bar{X}}$ | Mean subgroup range = \bar{R} |



The R Chart

- Monitors **dispersion** (variability) in a process
 - The characteristic of interest is measured on a numerical scale
 - Is a **variables control chart**
- Shows the sample **range** over time
 - Range = difference between smallest and largest values in the subgroup



Steps to create an R chart

- Find the mean of the subgroup ranges (the center line of the R chart)
- Compute the upper and lower control limits for the R chart
- Use lines to show the center and control limits on the R chart
- Plot the successive subgroup ranges as a line chart



Average of Subgroup Ranges

Mean of subgroup ranges:

$$\bar{R} = \frac{\sum R_i}{k}$$

where:

R_i = i^{th} subgroup range

k = number of subgroups



R Chart Control Limits

- The upper and lower control limits for an R chart are

$$\text{LCL} = \bar{R} - 3\bar{R} \frac{d_3}{d_2} = D_3(\bar{R})$$

$$\text{UCL} = \bar{R} + 3\bar{R} \frac{d_3}{d_2} = D_4(\bar{R})$$

where:

d_2 , d_3 , D_3 , and D_4 are found from the table (Appendix Table E.9) for subgroup size = n

R Chart Example

You are the manager of a 500-room hotel. You want to analyze the time it takes to deliver luggage to the room. For 7 days, you collect data on 5 deliveries per day. Is the variation in the process in control?



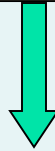
R Chart Example: Subgroup Data

| Day | Subgroup Size | Subgroup Mean | Subgroup Range |
|-----|---------------|---------------|----------------|
| 1 | 5 | 5.32 | 3.85 |
| 2 | 5 | 6.59 | 4.27 |
| 3 | 5 | 4.89 | 3.28 |
| 4 | 5 | 5.70 | 2.99 |
| 5 | 5 | 4.07 | 3.61 |
| 6 | 5 | 7.34 | 5.04 |
| 7 | 5 | 6.79 | 4.22 |



R Chart Center and Control Limits

$$\bar{R} = \frac{\sum R_i}{k} = \frac{3.85 + 4.27 + \dots + 4.22}{7} = 3.894$$



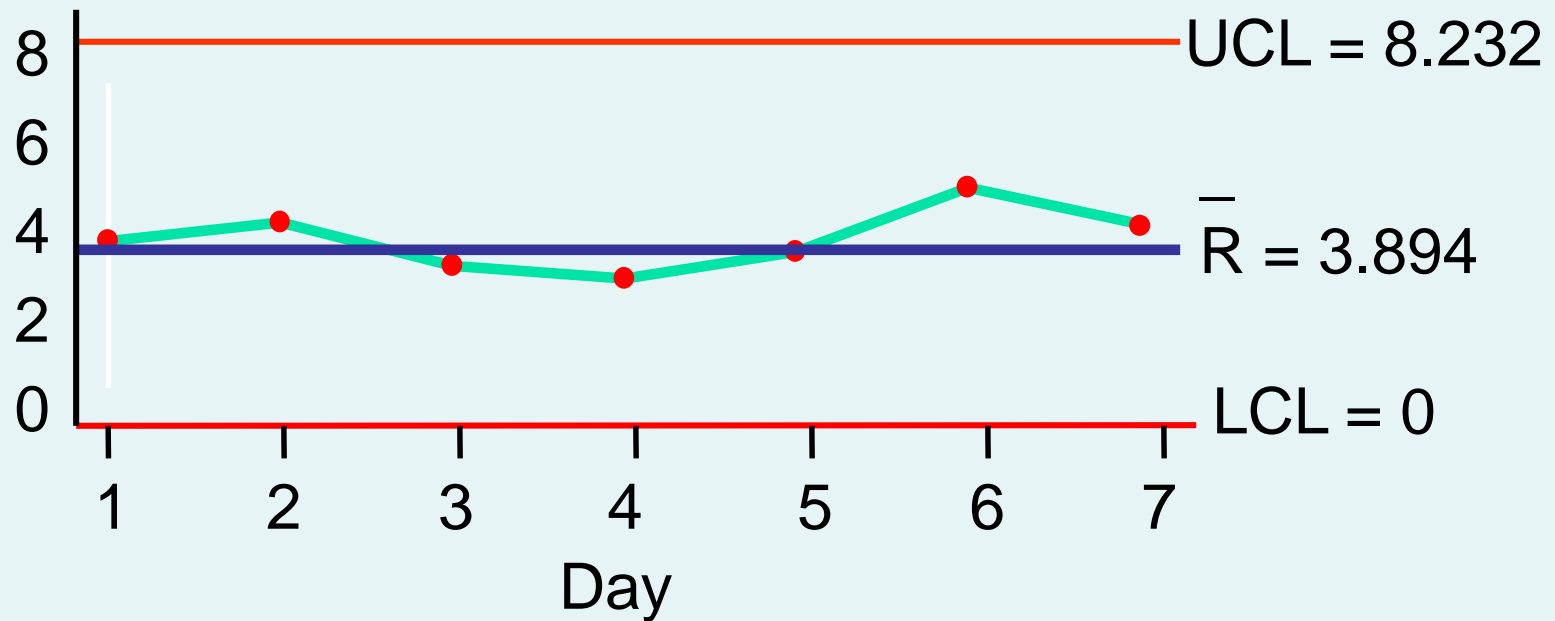
$$UCL = D_4(\bar{R}) = (2.114)(3.894) = 8.232$$

$$LCL = D_3(\bar{R}) = (0)(3.894) = 0$$

D_4 and D_3 are from
Table E.13 ($n = 5$)

R Chart Control Chart Solution

Minutes



Conclusion: Variation is in control



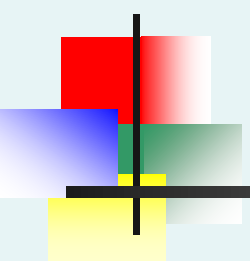
The \bar{X} Chart

- Shows the means of successive subgroups over time
- Monitors process **mean**
- Must be preceded by examination of the R chart to make sure that the variation in the process is in control



Steps to create an \bar{X} chart

- Compute the mean of the subgroup means (the center line of the \bar{X} chart)
- Compute the upper and lower control limits for the \bar{X} chart
- Graph the subgroup means
- Add the center line and control limits to the graph



Mean of Subgroup Means

Mean of subgroup means:

$$\bar{X} = \frac{\sum \bar{X}_i}{k}$$

where:

\bar{X}_i = i^{th} subgroup mean

k = number of subgroups



Computing Control Limits

- The upper and lower control limits for an \bar{X} chart are generally defined as

$$\begin{aligned} \text{UCL} &= \text{Process Mean} + 3 \text{ Standard Deviations} \\ \text{LCL} &= \text{Process Mean} - 3 \text{ Standard Deviations} \end{aligned}$$

- Use $\frac{\bar{R}}{d_2 \sqrt{n}}$ to estimate the standard deviation of the process mean, where d_2 is from appendix Table E.9

Computing Control Limits

(continued)

- The upper and lower control limits for an \bar{X} chart are generally defined as

UCL = Process Mean + 3 Standard Deviations

LCL = Process Mean – 3 Standard Deviations

- so

$$\text{UCL} = \bar{X} + 3 \frac{\bar{R}}{d_2 \sqrt{n}}$$
$$\text{LCL} = \bar{X} - 3 \frac{\bar{R}}{d_2 \sqrt{n}}$$

Computing Control Limits

(continued)

- Simplify the control limit calculations by using

$$UCL = \bar{\bar{X}} + A_2(\bar{R})$$

$$LCL = \bar{\bar{X}} - A_2(\bar{R})$$

where $A_2 = \frac{3}{d_2\sqrt{n}}$

\bar{X} Chart Example

You are the manager of a 500-room hotel. You want to analyze the time it takes to deliver luggage to the room. For seven days, you collect data on five deliveries per day. Is the process mean in control?



\bar{X} Chart Example: Subgroup Data

| Day | Subgroup Size | Subgroup Mean | Subgroup Range |
|-----|---------------|---------------|----------------|
| 1 | 5 | 5.32 | 3.85 |
| 2 | 5 | 6.59 | 4.27 |
| 3 | 5 | 4.89 | 3.28 |
| 4 | 5 | 5.70 | 2.99 |
| 5 | 5 | 4.07 | 3.61 |
| 6 | 5 | 7.34 | 5.04 |
| 7 | 5 | 6.79 | 4.22 |



\bar{X} Chart Control Limits Solution

$$\bar{\bar{X}} = \frac{\sum \bar{X}_i}{k} = \frac{5.32 + 6.59 + \dots + 6.79}{7} = 5.814$$

$$\bar{R} = \frac{\sum R_i}{k} = \frac{3.85 + 4.27 + \dots + 4.22}{7} = 3.894$$



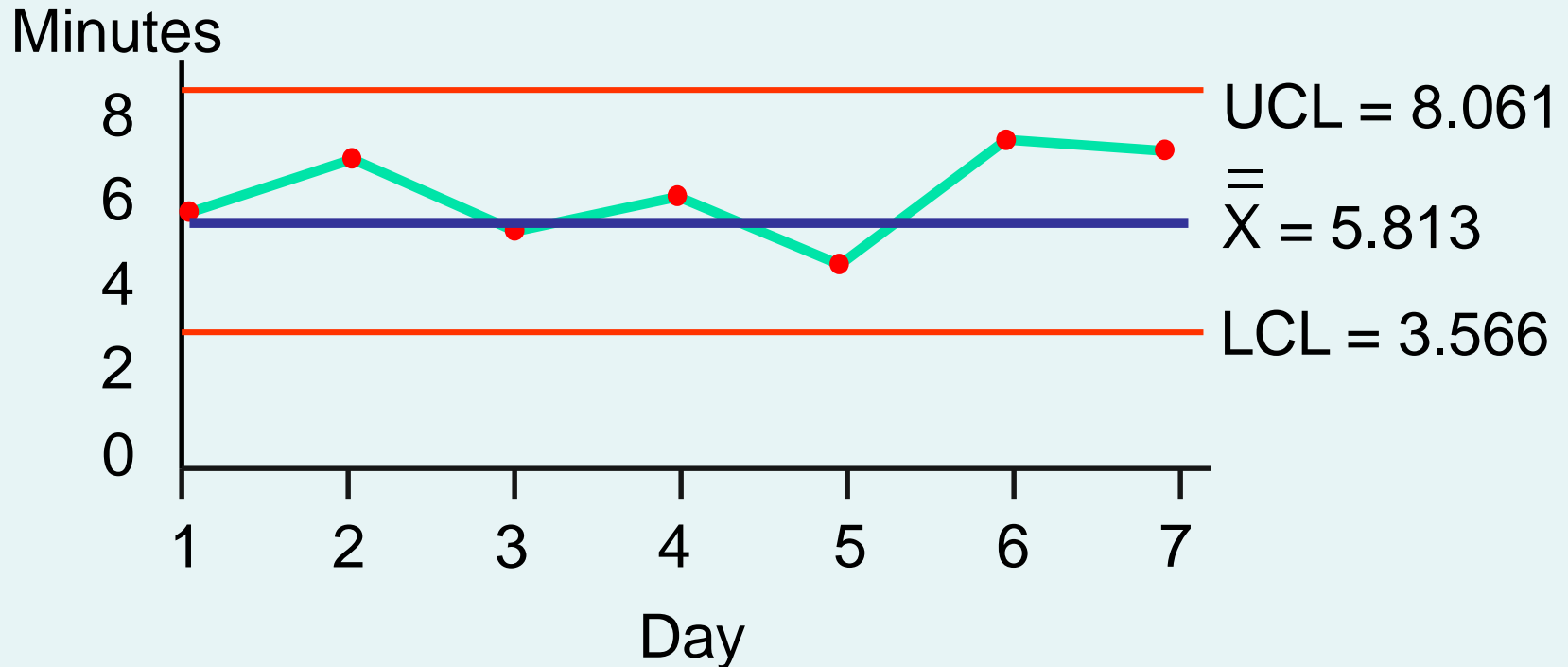
$$UCL = \bar{\bar{X}} + A_2(\bar{R}) = 5.814 + (0.577)(3.894) = 8.061$$

$$LCL = \bar{\bar{X}} - A_2(\bar{R}) = 5.814 - (0.577)(3.894) = 3.566$$



A_2 is from Table
E.9 ($n = 5$)

\bar{X} Chart Control Chart Solution



Conclusion: Process mean is in statistical control



Total Quality Management

- Primary focus is on process improvement
- Most variation in a process is due to the system, not the individual
- Teamwork is integral to quality management
- Customer satisfaction is a primary goal
- Organization transformation is necessary
- Fear must be removed from organizations
- Higher quality costs less, not more



Deming's 14 Points

1. Create a constancy of purpose toward improvement
 - become more competitive, stay in business, and provide jobs
2. Adopt the new philosophy
 - Better to improve now than to react to problems later
3. Stop depending on inspection to achieve quality -- build in quality from the start
 - Inspection to find defects at the end of production is too late
4. Stop awarding contracts on the basis of low bids
 - Better to build long-run purchaser/supplier relationships



Deming's 14 Points

(continued)

5. Improve the system continuously to improve quality and thus constantly reduce costs
6. Institute training on the job
 - Workers and managers must know the difference between common cause and special cause variation
7. Institute leadership
 - Know the difference between leadership and supervision
8. Drive out fear so that everyone may work effectively.
9. Break down barriers between departments so that people can work as a team.

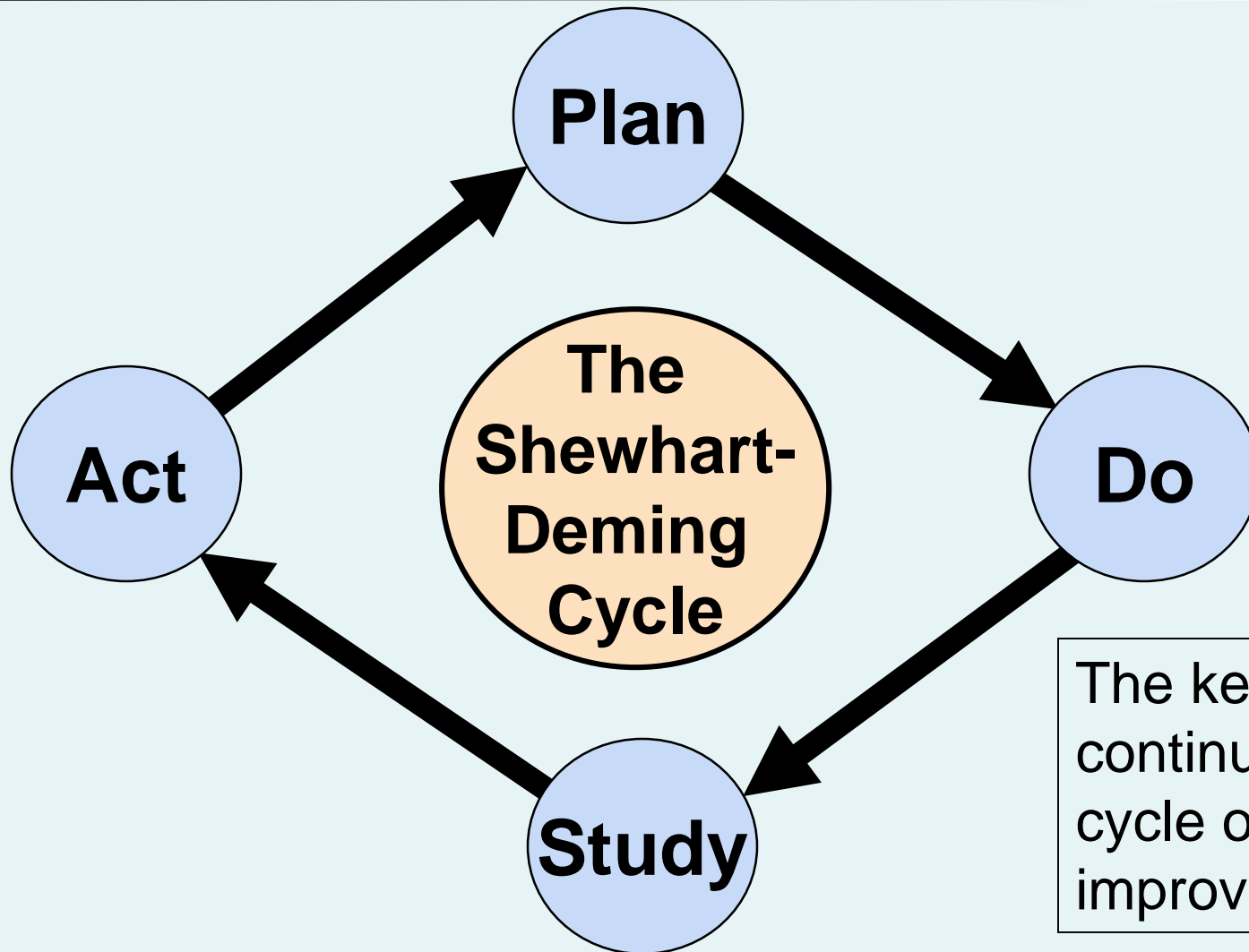


Deming's 14 Points

(continued)

- 10. Eliminate slogans and targets for the workforce
 - They can create adversarial relationships
- 11. Eliminate quotas and management by numerical goals
- 12. Remove barriers to pride of workmanship
- 13. Institute a vigorous program of education and self-improvement
- 14. Make the transformation everyone's job

The Shewhart-Deming Cycle





Six Sigma Management

A method of breaking a process into a series of steps:

- The goal is to reduce defects and produce near perfect results
- The Six Sigma approach allows for a shift of as much as 1.5 standard deviations, so is essentially a ± 4.5 standard deviation goal
- The mean of a normal distribution ± 4.5 standard deviations includes all but 3.4 out of a million items



The Six Sigma DMAIC Model

DMAIC represents

- **Define** -- define the problem to be solved; list costs, benefits, and impact to customer
- **Measure** – need consistent measurements for each Critical-to-Quality characteristic
- **Analyze** – find the root causes of defects
- **Improve** – use experiments to determine importance of each Critical-to-Quality variable
- **Control** – maintain gains that have been made



Roles in a Six Sigma Organization

- **Senior executive** -- clear and committed leadership
- **Executive committee** -- top management of an organization demonstrating commitment
- **Champions** -- strong sponsorship and leadership role in Six Sigma projects.
- **Process owner** -- the manager of the process being studied and improved
- **Master black belt** -- leadership role in the implementation of the Six Sigma process and as an advisor to senior executives



Roles in a Six Sigma Organization

(continued)

- **Black belt** -- works full time on Six Sigma projects
- **Green belt** -- works on Six Sigma projects part-time either as a team member for complex projects or as a project leader for simpler projects



Chapter Summary

- Discussed the theory of control charts
 - Common cause variation vs. special cause variation
- Constructed and interpreted p charts
- Constructed and interpreted \bar{X} and R charts
- Reviewed the philosophy of quality management
 - Deming's 14 points
- Discussed Six Sigma Management
 - Reduce defects to no more than 3.4 per million
 - Using DMAIC model for process improvement
 - Organizational roles