



What is Statistics?



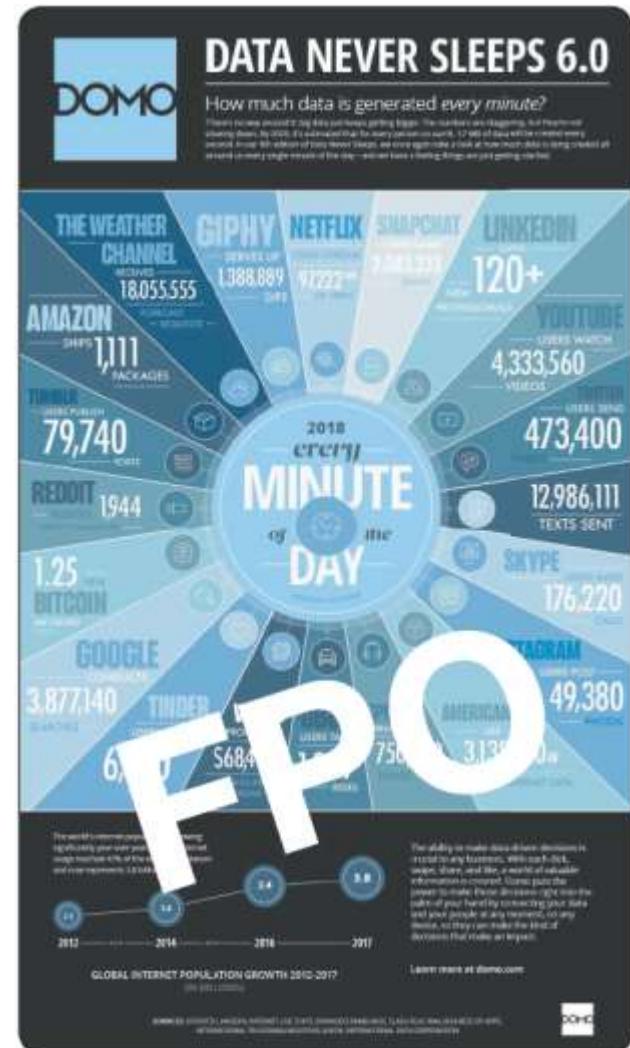
Chapter 1

Learning Objectives

- LOI-1** Explain why knowledge of statistics is important
- LOI-2** Define statistics and provide an example of how statistics is applied
- LOI-3** Differentiate between descriptive and inferential statistics
- LOI-4** Classify variables as qualitative or quantitative, and discrete or continuous
- LOI-5** Distinguish between nominal, ordinal, interval, and ratio levels of measurement
- LOI-6** List the values associated with the practice of statistics

Why Study Statistics

- ▶ Data are collected everywhere and require statistical knowledge to make the information useful
- ▶ Statistics is used to make valid comparisons and to predict the outcomes of decisions
- ▶ Statistical knowledge is useful in any career



What is Meant by Statistics

- ▶ What is statistics?
- ▶ It's more than presenting numerical facts

STATISTICS The science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions.

Example: The inflation rate for the calendar year was 0.7%. By applying statistics we could compare this year's inflation rate to past observations of inflation. Is it higher, lower, or about the same? Is there a trend of increasing or decreasing inflation? Is there a relationship between interest rates and government bonds?

Types of Statistics

- ▶ There are two types of statistics, descriptive and inferential
- ▶ Descriptive statistics can be used to organize data into a meaningful form
- ▶ You can summarize data and provide information that is easy to understand

DESCRIPTIVE STATISTICS Methods of organizing, summarizing, and presenting data in an informative way.

- ▶ Example: There are a total of 46,837 miles of interstate highways in the U.S. The interstate system represents 1% of the nation's roads, but carries more than 20% of the traffic. Texas has the most interstate highways and Alaska doesn't have any.

Types of Statistics (2 of 3)

- ▶ Inferential statistics can be used to estimate properties of a population
- ▶ You can make decisions based on a limited set of data

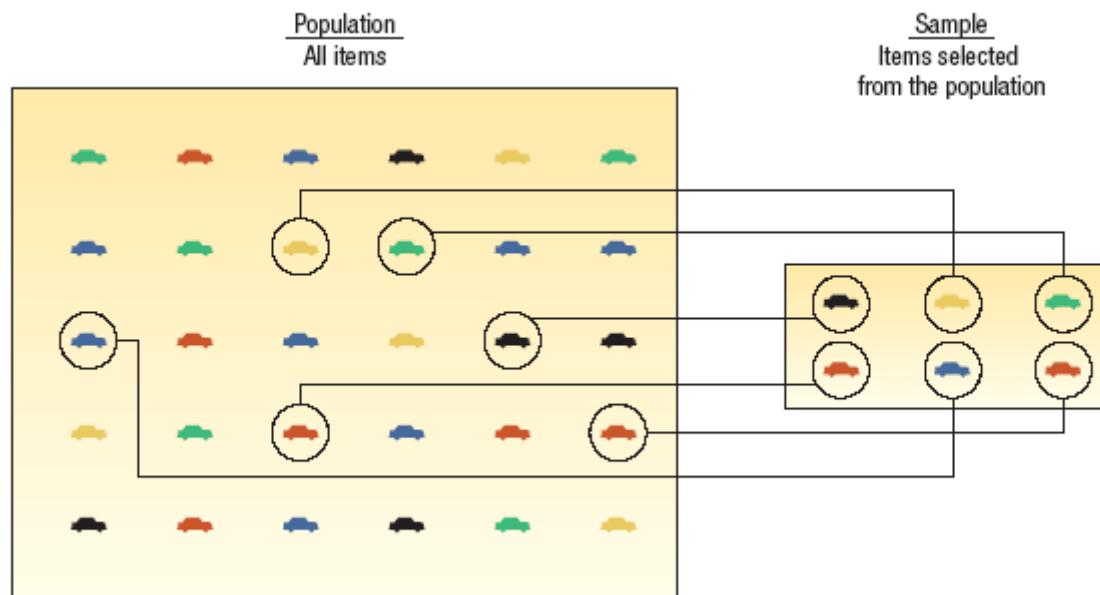
INFERENTIAL STATISTICS The methods used to estimate a property of a population on the basis of a sample.

- ▶ Example: In 2015, a sample of U.S. Internal Revenue Service tax preparation volunteers were tested with three standard tax returns. The sample indicated that tax returns were completed with a 49% accuracy rate. In other words, there were errors on about half of the returns.

Types of Statistics (3 of 3)

POPULATION The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest.

SAMPLE A portion or part of the population of interest.



Types of Variables

- ▶ There are two basic types of variables

QUALITATIVE VARIABLE An object or individual is observed and recorded as a non-numeric characteristic or attribute.

Examples: gender, state of birth, eye color

QUANTITATIVE VARIABLE A variable that is reported numerically.

Examples: balance in your checking account, the life of a car battery, the number of people employed by a company

Types of Variables (2 of 2)

- ▶ Quantitative variables can be discrete or continuous
- ▶ Discrete variables are typically the result of counting
 - ▶ Values have “gaps” between the values
 - ▶ Examples: the number of bedrooms in a house (1, 2, 3, 4, etc.), the number of students in a statistics course (326, 421, etc.)
- ▶ Continuous variables are usually the result of measuring something
 - ▶ Can assume any value within a specific range
 - ▶ Examples: Duration of flights from Orlando to San Diego (5.25 hours), grade point average (3.258)

Types of Variables Summary

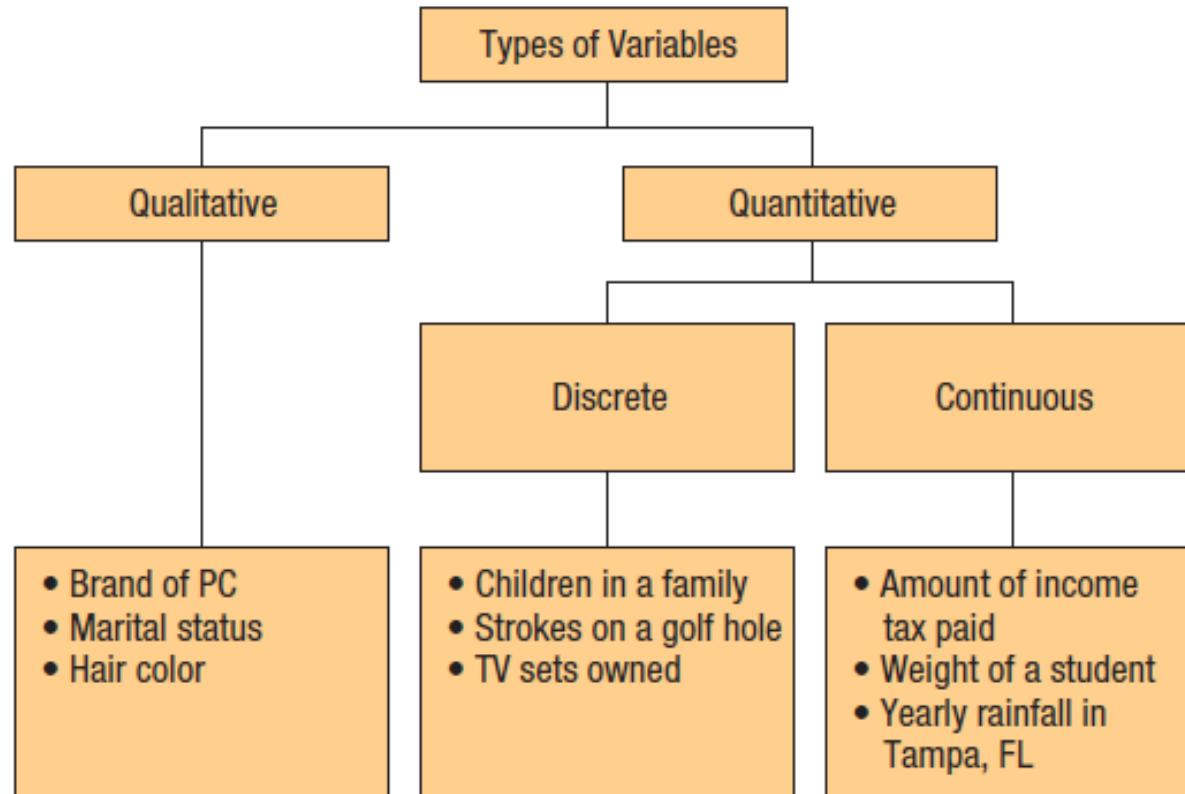


CHART 1–2 Summary of the Types of Variables

Levels of Measurement

- ▶ There are four levels of measurement
 - ▶ Nominal, ordinal, interval, and ratio
- ▶ The level of measurement determines the type of statistical analysis that can be performed
- ▶ Nominal is the lowest level of measurement

NOMINAL LEVEL OF MEASUREMENT Data recorded at the nominal level of measurement is represented as labels or names. They have no order. They can only be classified and counted.

- ▶ Examples: classifying M&M candies by color, identifying students at a football game by gender

Levels of Measurement (2 of 4)

- ▶ The next level of measurement is the ordinal level
- ▶ The rankings are known but not the magnitude of differences between groups

ORDINAL LEVEL OF MEASUREMENT Data recorded at the ordinal level of measurement is based on a relative ranking or rating of items based on a defined attribute or qualitative variable. Variables based on this level of measurement are only ranked and counted.

- ▶ Examples: the list of top ten states for best business climate, student ratings of professors

Levels of Measurement (3 of 4)

- ▶ The next level of measurement is the interval level
- ▶ This data has all the characteristics of ordinal level data, plus the differences between the values are meaningful
- ▶ There is no natural 0 point

INTERVAL LEVEL OF MEASUREMENT For data recorded at the interval level of measurement, the interval or the distance between values is meaningful. The interval level of measurement is based on a scale with a known unit of measurement.

- ▶ Examples: the Fahrenheit temperature scale, dress sizes

Levels of Measurement (4 of 4)

- ▶ The highest level of measurement is the ratio level
- ▶ The data has all the characteristics of the interval scale and ratios between numbers are meaningful
- ▶ The 0 point represents the absence of the characteristic

RATIO LEVEL OF MEASUREMENT Data recorded at the ratio level of measurement are based on a scale with a known unit of measurement and a meaningful interpretation of zero on the scale.

- ▶ Examples: wages, changes in stock prices, and height

Levels of Measurement Summary

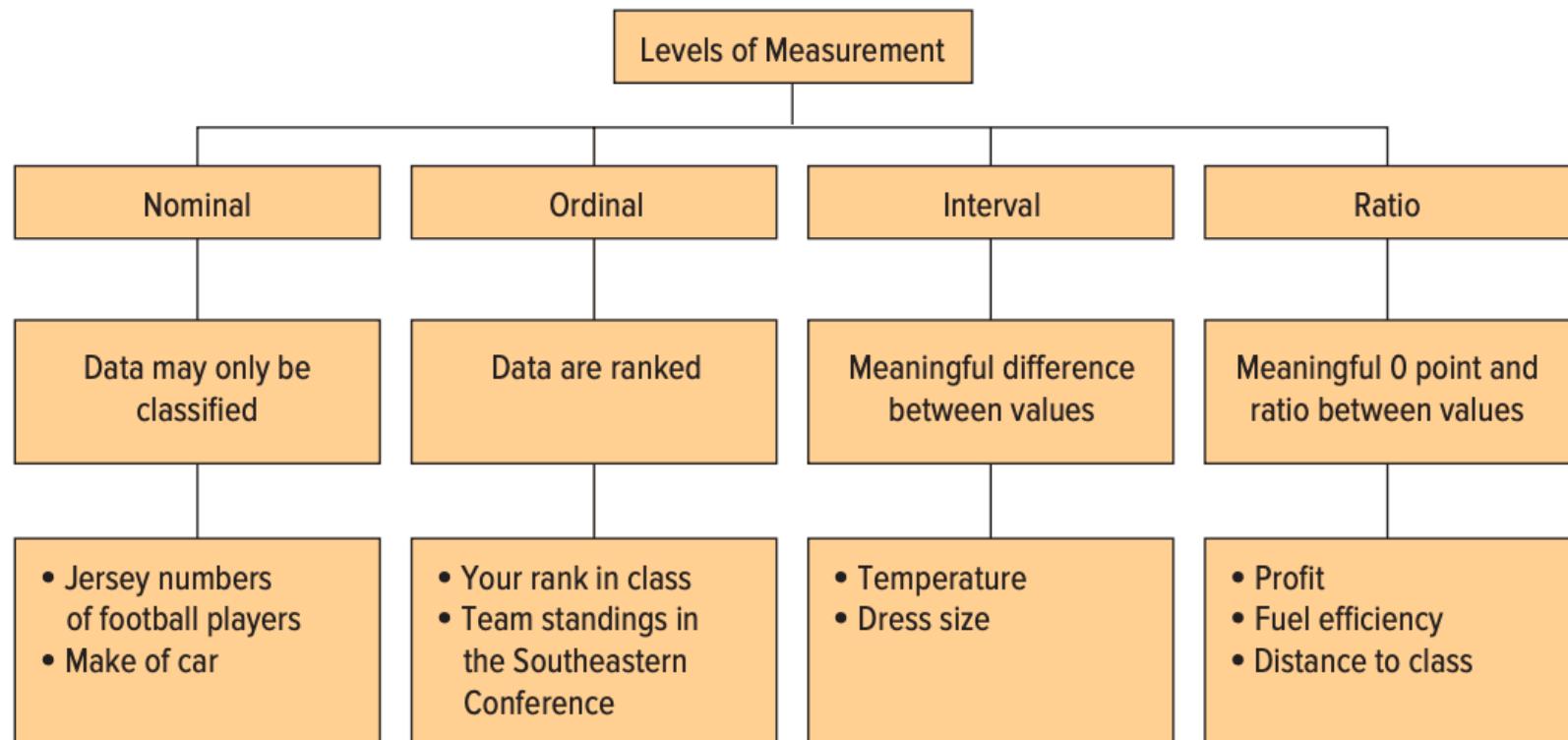


CHART 1-3 Summary and Examples of the Characteristics for Levels of Measurement

Ethics and Statistics

- ▶ Practice statistics with integrity and honesty when collecting, organizing, summarizing, analyzing, and interpreting numerical information
- ▶ Maintain an independent and principled point of view when analyzing and reporting findings and results
- ▶ Question reports that are based on data that
 - ▶ does not fairly represent the population
 - ▶ does not include all relevant statistics
 - ▶ introduces bias in an attempt to mislead or misrepresent

Basic Business Analytics

- ▶ Business Analytics is used to process and analyze data and information to support a story or narrative of a company
- ▶ Using computer software to summarize, organize, analyze, and present the findings of statistical analysis is essential

| | A | B | C | D | E | F | G | H |
|----|-----|---------|-----------|--------------|----------|--------------------|-----------|---|
| 1 | Age | Profit | Location | Vehicle-Type | Previous | | Profit | |
| 2 | 33 | \$1,889 | Olean | SUV | 1 | | | |
| 3 | 47 | \$1,461 | Kane | Sedan | 0 | Mean | 1843.17 | |
| 4 | 44 | \$1,532 | Tionesta | SUV | 3 | Standard Error | 47.97 | |
| 5 | 53 | \$1,220 | Olean | Sedan | 0 | Median | 1882.50 | |
| 6 | 51 | \$1,674 | Sheffield | Sedan | 1 | Mode | 1915.00 | |
| 7 | 41 | \$2,389 | Kane | Truck | 1 | Standard Deviation | 643.63 | |
| 8 | 58 | \$2,058 | Kane | SUV | 1 | Sample Variance | 414256.61 | |
| 9 | 35 | \$1,919 | Tionesta | SUV | 1 | Kurtosis | -0.22 | |
| 10 | 45 | \$1,266 | Olean | Sedan | 0 | Skewness | -0.24 | |
| 11 | 54 | \$2,991 | Tionesta | Sedan | 0 | Range | 2998 | |
| 12 | 56 | \$2,695 | Kane | Sedan | 2 | Minimum | 294 | |
| 13 | 41 | \$2,165 | Tionesta | SUV | 0 | Maximum | 3292 | |
| 14 | 38 | \$1,766 | Sheffield | SUV | 0 | Sum | 331770 | |
| 15 | 48 | \$1,952 | Tionesta | Compact | 1 | Count | 180 | |



Chapter 1 Practice Problems

Question 1

LO1-5

What is the level of measurement for each of the following variables?

- a. Student IQ ratings
- b. Distance students travel to class
- c. The jersey numbers of a sorority soccer team
- d. A student's state of birth
- e. A student's academic class – that is, freshman, sophomore, junior, or senior
- f. Number of hours students study per week

Question 13

LO1-4,5

For each of the following, determine whether the variable is continuous or discrete, quantitative or qualitative, and level of measurement

- a. Salary
- b. Gender
- c. Sales volume of MP3 players
- d. Soft drink preference
- e. Temperature
- f. SAT scores
- g. Student rank in class
- h. Rating of a finance professor
- i. Number of home video screens



Describing Data: Frequency Tables, Frequency Distributions, and Graphic Presentation



Chapter 2

Learning Objectives

- LO2-1** Summarize qualitative variables with frequency and relative frequency tables
- LO2-2** Display a frequency table using a bar or pie chart
- LO2-3** Summarize quantitative variables with frequency and relative frequency distributions
- LO2-4** Display a frequency distribution using a histogram or frequency polygon

Constructing Frequency Tables

FREQUENCY TABLE A grouping of qualitative data into mutually exclusive and collectively exhaustive classes showing the number of observations in each class.

- ▶ Mutually exclusive means the data fit in just one class
- ▶ Collectively exhaustive means there is a class for each value

TABLE 2–1 Frequency Table for Vehicles Sold Last Month at Applewood Auto Group by Location

| Location | Number of Cars |
|-----------|----------------|
| Kane | 52 |
| Olean | 40 |
| Sheffield | 45 |
| Tionesta | 43 |
| Total | 180 |

Constructing Frequency Tables (2 of 3)

- ▶ To construct a frequency table
 - ▶ First sort the data into classes
 - ▶ Count the number in each class and report as the class frequency

TABLE 2–1 Frequency Table for Vehicles Sold Last Month at Applewood Auto Group by Location

| Location | Number of Cars |
|-----------|----------------|
| Kane | 52 |
| Olean | 40 |
| Sheffield | 45 |
| Tionesta | 43 |
| Total | <hr/> 180 |

Constructing Frequency Tables (3 of 3)

- ▶ Convert each frequency to a relative frequency
 - ▶ Each of the class frequencies is divided by the total number of observations
 - ▶ Shows the fraction of the total number observations in each class

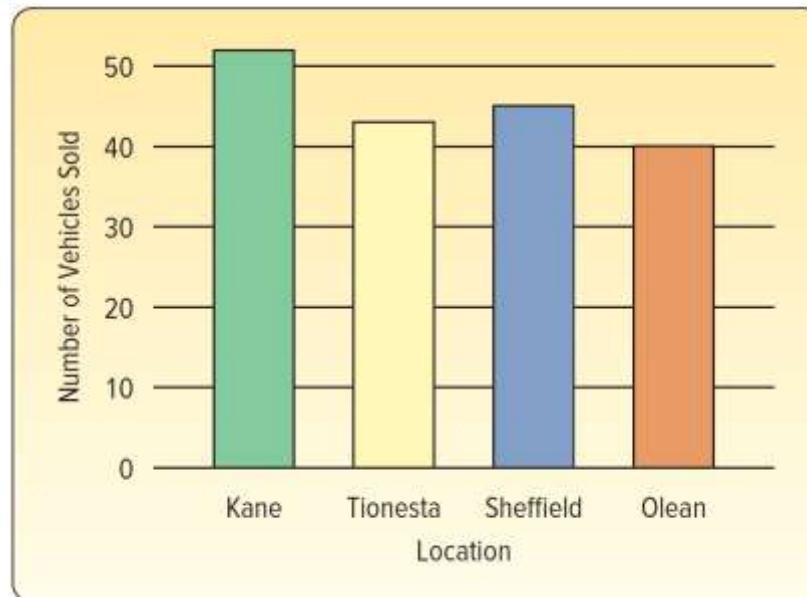
TABLE 2–2 Relative Frequency Table of Vehicles Sold by Location Last Month at Applewood Auto Group

| Location | Number of Cars | Relative Frequency | Found by |
|-----------|----------------|--------------------|----------|
| Kane | 52 | .289 | 52/180 |
| Olean | 40 | .222 | 40/180 |
| Sheffield | 45 | .250 | 45/180 |
| Tionesta | 43 | .239 | 43/180 |
| Total | 180 | 1.000 | |

Graphic Presentation of Qualitative Data

BAR CHART A graph that shows the qualitative classes on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are proportional to the heights of the bars.

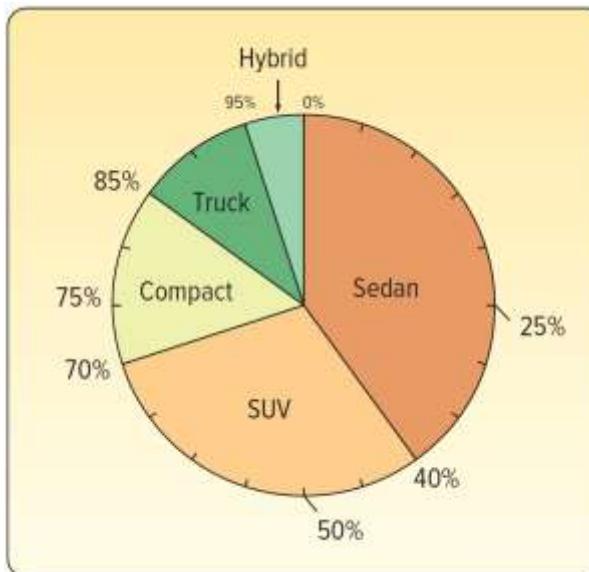
- Use a bar chart when you wish to compare the number of observations for each class of a qualitative variable.



Graphic Presentation of Qualitative Data (2 of 2)

PIE CHART A chart that shows the proportion or percentage that each class represents of the total number of frequencies.

- Use a pie chart when you wish to compare relative differences in the percentage of observations for each class of a qualitative variable.



Constructing Frequency Distributions

FREQUENCY DISTRIBUTION A grouping of quantitative data into mutually exclusive and collectively exhaustive classes showing the number of observations in each class.

- ▶ This is a four-step process
 1. Decide on the number of classes
 2. Determine the class interval
 3. Set the individual class limits
 4. Tally the data into classes and determine the number of the observations in each class

Frequency Distributions

- ▶ Step I Decide on the number of classes
- ▶ Use the $2^k > n$ rule, where $n=180$
 - ▶ k is the number of classes
 - ▶ n is the number of values in the data set
 - ▶ $2^k > 180$, let $k = 8$
 - ▶ So use 8 classes

TABLE 2-4 Profit on Vehicles Sold Last Month by the Applewood Auto Group

| \$1,387 | \$2,148 | \$2,201 | \$ 963 | \$ 820 | \$2,230 | \$3,043 | \$2,584 | \$2,370 |
|---------|---------|---------|--------|--------|---------|---------|---------|---------|
| 1,754 | 2,207 | 996 | 1,298 | 1,266 | 2,341 | 1,059 | 2,666 | 2,637 |
| 1,817 | 2,252 | 2,813 | 1,410 | 1,741 | 3,292 | 1,674 | 2,991 | 1,426 |
| 1,040 | 1,428 | 323 | 1,553 | 1,772 | 1,108 | 1,807 | 934 | 2,944 |
| 1,273 | 1,889 | 352 | 1,648 | 1,932 | 1,295 | 2,056 | 2,063 | 2,147 |
| 1,529 | 1,166 | 482 | 2,071 | 2,350 | 1,344 | 2,236 | 2,083 | 1,973 |
| 3,082 | 1,320 | 1,144 | 2,116 | 2,422 | 1,906 | 2,928 | 2,856 | 2,502 |
| 1,951 | 2,265 | 1,485 | 1,500 | 2,446 | 1,952 | 1,269 | 2,989 | 783 |
| 2,692 | 1,323 | 1,509 | 1,549 | 369 | 2,070 | 1,717 | 910 | 1,538 |
| 1,206 | 1,760 | 1,638 | 2,348 | 978 | 2,454 | 1,797 | 1,536 | 2,339 |
| 1,342 | 1,919 | 1,961 | 2,498 | 1,238 | 1,606 | 1,955 | 1,957 | 2,700 |
| 443 | 2,357 | 2,127 | 294 | 1,818 | 1,680 | 2,199 | 2,240 | 2,222 |
| 754 | 2,866 | 2,430 | 1,115 | 1,824 | 1,827 | 2,482 | 2,695 | 2,597 |
| 1,621 | 732 | 1,704 | 1,124 | 1,907 | 1,915 | 2,701 | 1,325 | 2,742 |
| 870 | 1,464 | 1,876 | 1,532 | 1,938 | 2,084 | 3,210 | 2,250 | 1,837 |
| 1,174 | 1,626 | 2,010 | 1,688 | 1,940 | 2,639 | 377 | 2,279 | 2,842 |
| 1,412 | 1,762 | 2,165 | 1,822 | 2,197 | 842 | 1,220 | 2,626 | 2,434 |
| 1,809 | 1,915 | 2,231 | 1,897 | 2,646 | 1,963 | 1,401 | 1,501 | 1,640 |
| 2,415 | 2,119 | 2,389 | 2,445 | 1,461 | 2,059 | 2,175 | 1,752 | 1,821 |
| 1,546 | 1,766 | 335 | 2,886 | 1,731 | 2,338 | 1,118 | 2,058 | 2,487 |

Frequency Distributions (2 of 4)

- ▶ Step 2 Determine the class interval, i

$$i \geq \frac{\text{Maximum Value} - \text{Minimum Value}}{k}$$

- ▶ Round up to some convenient number

$$i \geq \frac{\$3,292 - \$294}{8} = \$374.75$$

- ▶ So decide to use an interval of \$400
- ▶ The interval is also referred to as the class width

Frequency Distributions (3 of 4)

- ▶ Step 3 Set the individual class limits
- ▶ Lower limits should be rounded to an easy-to-read number when possible

| Classes | |
|--------------|--------|
| \$ 200 up to | \$ 600 |
| 600 up to | 1,000 |
| 1,000 up to | 1,400 |
| 1,400 up to | 1,800 |
| 1,800 up to | 2,200 |
| 2,200 up to | 2,600 |
| 2,600 up to | 3,000 |
| 3,000 up to | 3,400 |

Frequency Distributions (4 of 4)

- ▶ Step 4 Tally the individual data into the classes and determine the number of observations in each class
- ▶ The number of observations is the class frequency

| Profit | Frequency |
|---------------------|-----------|
| \$ 200 up to \$ 600 | III |
| 600 up to 1,000 | |
| 1,000 up to 1,400 | III |
| 1,400 up to 1,800 | III |
| 1,800 up to 2,200 | |
| 2,200 up to 2,600 | II |
| 2,600 up to 3,000 | IIII |
| 3,000 up to 3,400 | |

| Profit | Frequency |
|---------------------|-----------|
| \$ 200 up to \$ 600 | 8 |
| 600 up to 1,000 | 11 |
| 1,000 up to 1,400 | 23 |
| 1,400 up to 1,800 | 38 |
| 1,800 up to 2,200 | 45 |
| 2,200 up to 2,600 | 32 |
| 2,600 up to 3,000 | 19 |
| 3,000 up to 3,400 | 4 |
| Total | 180 |

Relative Frequency Distributions

- To find the relative frequencies, simply take the class frequency and divide by the total number of observations

| Profit | Frequency | Relative Frequency | Found by |
|---------------------|-----------|--------------------|----------|
| \$ 200 up to \$ 600 | 8 | .044 | 8/180 |
| 600 up to 1,000 | 11 | .061 | 11/180 |
| 1,000 up to 1,400 | 23 | .128 | 23/180 |
| 1,400 up to 1,800 | 38 | .211 | 38/180 |
| 1,800 up to 2,200 | 45 | .250 | 45/180 |
| 2,200 up to 2,600 | 32 | .178 | 32/180 |
| 2,600 up to 3,000 | 19 | .106 | 19/180 |
| 3,000 up to 3,400 | 4 | .022 | 4/180 |
| Total | 180 | 1.000 | |

Graphic Presentation of a Frequency Distribution

HISTOGRAM A graph in which the classes are marked on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are represented by the heights of the bars, and the bars are drawn adjacent to each other.

- ▶ A histogram shows the shape of a distribution
- ▶ Each class is depicted as a rectangle, with the height of the bar representing the number in each class

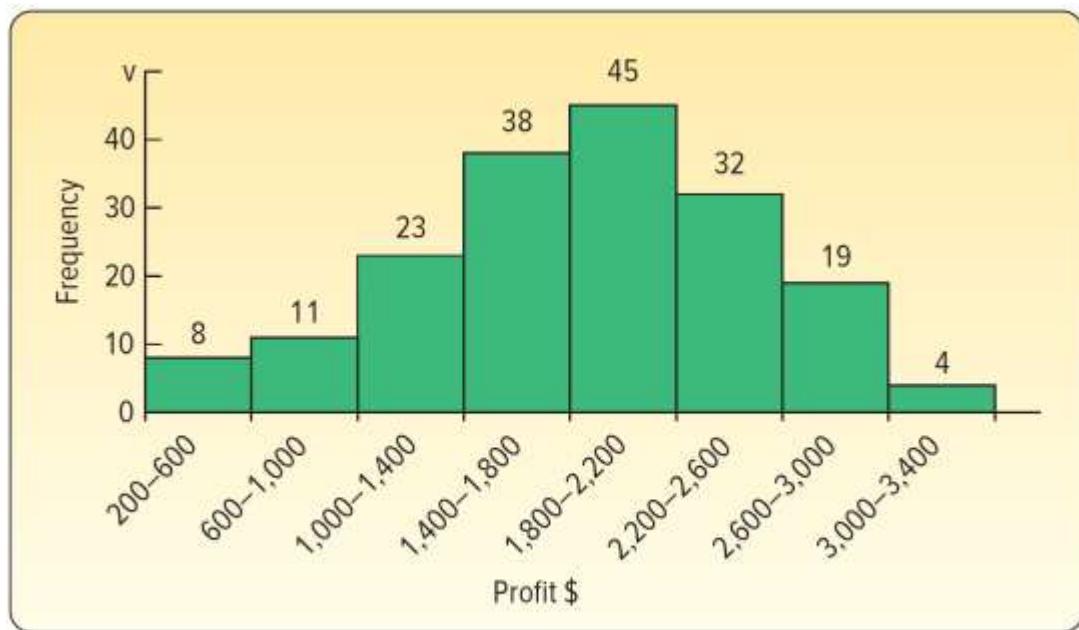


CHART 2-4 Histogram of the Profit on 180 Vehicles Sold at the Applewood Auto Group

Graphical Presentation of a Frequency Distribution

- ▶ A frequency polygon, similar to a histogram, also shows the shape of a distribution
- ▶ These are good to use when comparing two or more distributions

| Profit | Midpoint | Frequency |
|---------------------|----------|-----------|
| \$ 200 up to \$ 600 | \$ 400 | 8 |
| 600 up to 1,000 | 800 | 11 |
| 1,000 up to 1,400 | 1,200 | 23 |
| 1,400 up to 1,800 | 1,600 | 38 |
| 1,800 up to 2,200 | 2,000 | 45 |
| 2,200 up to 2,600 | 2,400 | 32 |
| 2,600 up to 3,000 | 2,800 | 19 |
| 3,000 up to 3,400 | 3,200 | 4 |
| Total | | 180 |

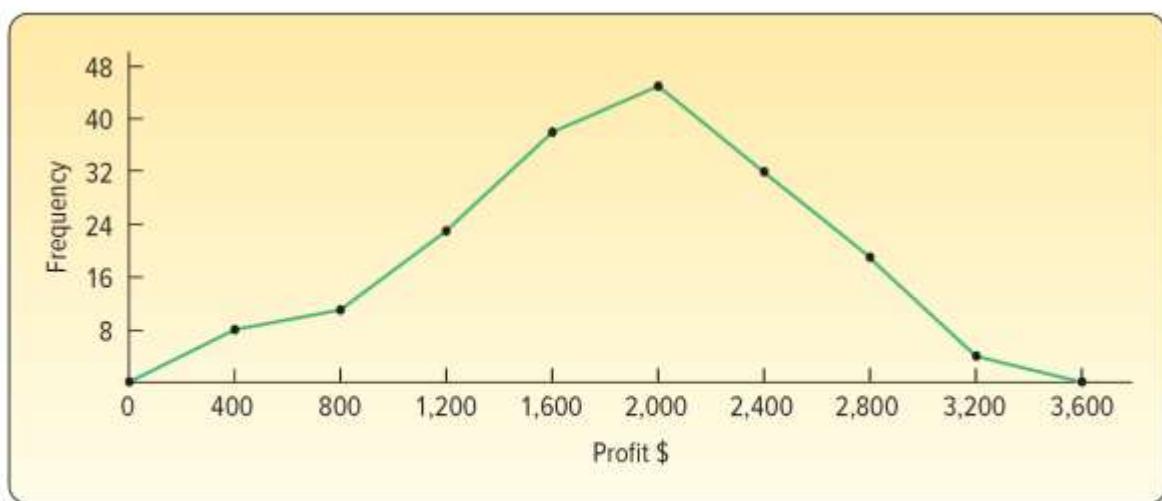


CHART 2–5 Frequency Polygon of Profit on 180 Vehicles Sold at Applewood Auto Group

Cumulative Frequency Distributions

- ▶ To construct a cumulative frequency distribution, add each frequency to the frequencies before it
- ▶ This shows how many values have accumulated as you move from one class down to the next class

TABLE 2–8 Cumulative Frequency Distribution for Profit on Vehicles Sold Last Month at Applewood Auto Group

| Profit | Cumulative Frequency | Found by |
|------------------|----------------------|-------------------------------------|
| Less than \$ 600 | 8 | 8 |
| Less than 1,000 | 19 | 8 + 11 |
| Less than 1,400 | 42 | 8 + 11 + 23 |
| Less than 1,800 | 80 | 8 + 11 + 23 + 38 |
| Less than 2,200 | 125 | 8 + 11 + 23 + 38 + 45 |
| Less than 2,600 | 157 | 8 + 11 + 23 + 38 + 45 + 32 |
| Less than 3,000 | 176 | 8 + 11 + 23 + 38 + 45 + 32 + 19 |
| Less than 3,400 | 180 | 8 + 11 + 23 + 38 + 45 + 32 + 19 + 4 |

Cumulative Relative Frequency Distribution

- ▶ To construct a cumulative relative frequency distribution, we divide the cumulative frequencies by the total number of observations

TABLE 2–9 Cumulative Relative Frequency Distribution for Profit on Vehicles Sold Last Month at Applewood Auto Group

| Profit | Cumulative Frequency | Cumulative Relative Frequency |
|------------------|----------------------|-------------------------------|
| Less than \$ 600 | 8 | $8/180 = 0.044 = 4.4\%$ |
| Less than 1,000 | 19 | $19/180 = 0.106 = 10.6\%$ |
| Less than 1,400 | 42 | $42/180 = 0.233 = 23.3\%$ |
| Less than 1,800 | 80 | $80/180 = 0.444 = 44.4\%$ |
| Less than 2,200 | 125 | $125/180 = 0.694 = 69.4\%$ |
| Less than 2,600 | 157 | $157/180 = 0.872 = 87.2\%$ |
| Less than 3,000 | 176 | $176/180 = 0.978 = 97.8\%$ |
| Less than 3,400 | 180 | $180/180 = 1.000 = 100\%$ |

Cumulative Frequency Polygon

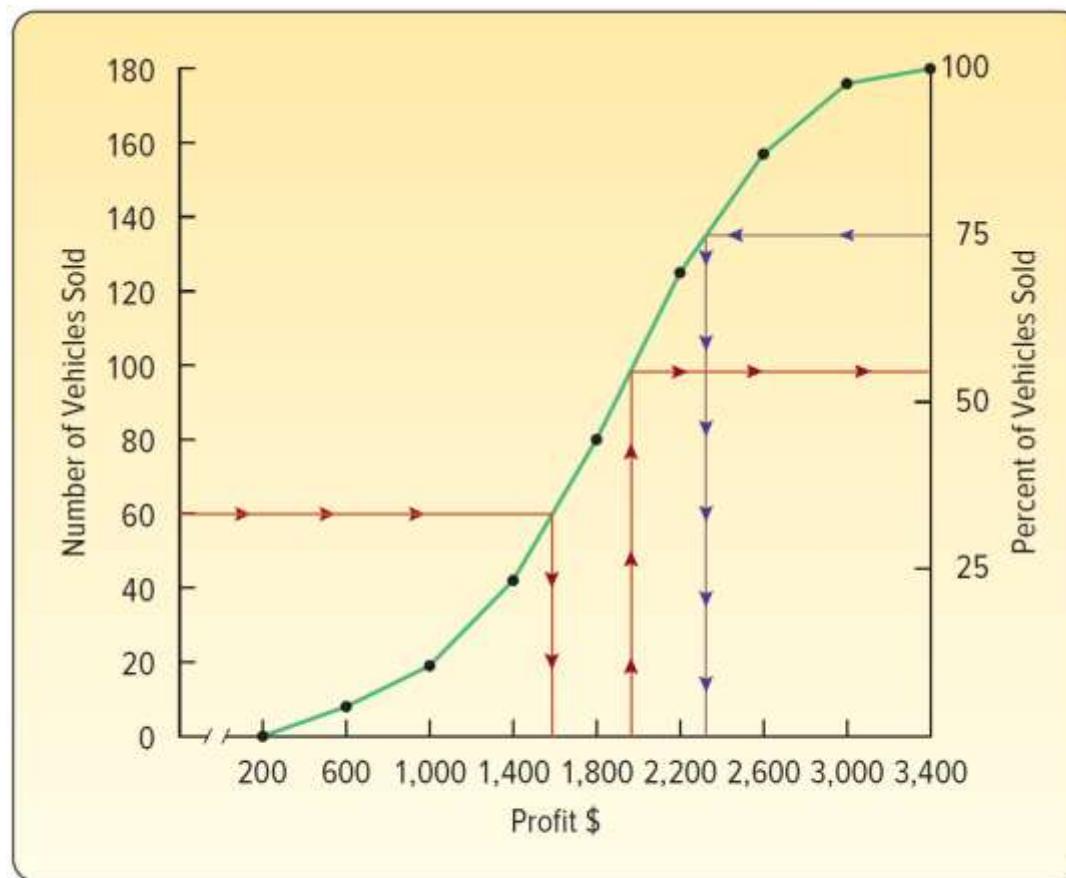


CHART 2–7 Cumulative Frequency Polygon for Profit on Vehicles Sold Last Month at Applewood Auto Group



Chapter 2 Practice Problems

Question 5

LO2-2

Wellstone Inc. produces and markets replacement covers for cell phones in five different colors: bright white, metallic black, magnetic lime, tangerine orange, and fusion red. To estimate the demand for each color, the company set up a kiosk for several hours in the Mall of America and asked randomly selected people which cover color was their favorite.

- a. What is the table called?
- b. Draw a bar chart for the table.
- c. Draw a pie chart.
- d. If Wellstone Inc. plans to produce one million cell phone covers, how many of each color should it produce?

| | |
|------------------|-----|
| Bright white | 130 |
| Metallic black | 104 |
| Magnetic lime | 325 |
| Tangerine orange | 455 |
| Fusion red | 286 |

Question 11

LO2-3

Wachesaw Manufacturing Inc. produced the following number of units in the last 16 days.

| | | | | | | | |
|----|----|----|----|----|----|----|----|
| 27 | 27 | 27 | 28 | 27 | 25 | 25 | 28 |
| 26 | 28 | 26 | 28 | 31 | 30 | 26 | 26 |

The information is to be organized into a frequency distribution.

- a. How many classes would you recommend?
- b. What class interval would you suggest?
- c. What lower limit would you recommend for the first class?
- d. Organize the information into a frequency distribution and determine the relative frequency distribution.
- e. Comment on the shape of the distribution.

Question 17

LO2-4

The following frequency distribution reports the number of frequent flier miles, reported in thousands, for employees of Brumley Statistical Consulting Inc. during the most recent quarter.

| Frequent Flier Miles (000) | Number of Employees |
|-------------------------------|------------------------|
| 0 up to 3 | 5 |
| 3 up to 6 | 12 |
| 6 up to 9 | 23 |
| 9 up to 12 | 8 |
| 12 up to 15 | 2 |
| Total | 50 |

- a. How many employees were studied?
- b. What is the midpoint of the first class? What lower limit would you recommend for the first class?
- c. Construct a histogram.
- d. A frequency polygon is to be drawn. What are the coordinates of the plot for the first class?
- e. Construct a frequency polygon.
- f. Interpret the frequent flier miles accumulated using the two charts.



Describing Data: Numerical Measures



Chapter 3

Learning Objectives

- LO3-1** Compute and interpret the mean, the median, and the mode
- LO3-2** Compute a weighted mean
- LO3-3** Compute and interpret the geometric mean
- LO3-4** Compute and interpret the range, variance, and standard deviation
- LO3-5** Explain and apply Chebyshev's theorem and the Empirical Rule
- LO3-6** Compute the mean and standard deviation of grouped data

Measures of Location

- ▶ A measure of location is a value used to describe the central tendency of a set of data
- ▶ Common measures of location
 - ▶ Mean
 - ▶ Median
 - ▶ Mode
- ▶ The arithmetic mean is the most widely reported measure of location

Population Mean

POPULATION MEAN

$$\mu = \frac{\Sigma x}{N} \quad (3-1)$$

where:

- μ represents the population mean. It is the Greek lowercase letter “mu.”
- N is the number of values in the population.
- x represents any particular value.
- Σ is the Greek capital letter “sigma” and indicates the operation of adding.
- Σx is the sum of the x values in the population.

- ▶ A measurable characteristic of a population is a parameter

PARAMETER A characteristic of a population.

Example: Population Mean

There are 42 exits on I-75 through the state of Kentucky. Listed below are the distances between exits (in miles).

| | | | | | | | | | | | | | |
|----|---|----|---|---|---|---|----|---|----|---|----|---|----|
| 11 | 4 | 10 | 4 | 9 | 3 | 8 | 10 | 3 | 14 | 1 | 10 | 3 | 5 |
| 2 | 2 | 5 | 6 | 1 | 2 | 2 | 3 | 7 | 1 | 3 | 7 | 8 | 10 |
| 1 | 4 | 7 | 5 | 2 | 2 | 5 | 1 | 1 | 3 | 3 | 1 | 2 | 1 |

1. Why is this information a population?
2. What is the mean number of miles between exits?

Example: Population Mean Continued

There are 42 exits on I-75 through the state of Kentucky. Listed below are the distances between exits (in miles).

| | | | | | | | | | | | | | |
|----|---|----|---|---|---|---|----|---|----|---|----|---|----|
| 11 | 4 | 10 | 4 | 9 | 3 | 8 | 10 | 3 | 14 | 1 | 10 | 3 | 5 |
| 2 | 2 | 5 | 6 | 1 | 2 | 2 | 3 | 7 | 1 | 3 | 7 | 8 | 10 |
| 1 | 4 | 7 | 5 | 2 | 2 | 5 | 1 | 1 | 3 | 3 | 1 | 2 | 1 |

- I. Why is this information a population?

This is a population because we are considering all of the exits in Kentucky.

2. What is the mean number of miles between exits?

$$\mu = \frac{\sum x}{N} = \frac{11 + 4 + 10 + \dots + 1}{42} = \frac{192}{42} = 4.57$$

Sample Mean

SAMPLE MEAN

$$\bar{x} = \frac{\Sigma x}{n}$$

(3–2)

where:

- \bar{x} represents the sample mean. It is read “x bar.”
- n is the number of values in the sample.
- x represents any particular value.
- Σ is the Greek capital letter “sigma” and indicates the operation of adding.
- Σx is the sum of the x values in the sample.

- ▶ A measurable characteristic of a sample is a statistic

STATISTIC A characteristic of a sample.

Example: Sample Mean

Verizon is studying the number of hours per day that people use their mobile phones. A random sample of 12 customers showed the following daily usage in hours.

| | | | | | |
|-----|-----|-----|-----|-----|-----|
| 4.1 | 3.7 | 4.3 | 4.2 | 5.5 | 5.1 |
| 4.2 | 5.1 | 4.2 | 4.6 | 5.2 | 3.8 |

What is the arithmetic mean number of hours used last month?

$$\text{Sample mean} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{4.1 + 3.7 + \dots + 3.8}{12} = \frac{54.0}{12} = 4.5$$

The arithmetic mean number of hours per day that people use their mobile phones is 4.5 hours.

Properties of the Arithmetic Mean

- ▶ Interval or ratio scale of measurement is required
- ▶ All the data values are used in the calculation
- ▶ The mean is unique
- ▶ The sum of the deviations from the mean equals zero

A weakness of the mean is that it is affected by extreme values.

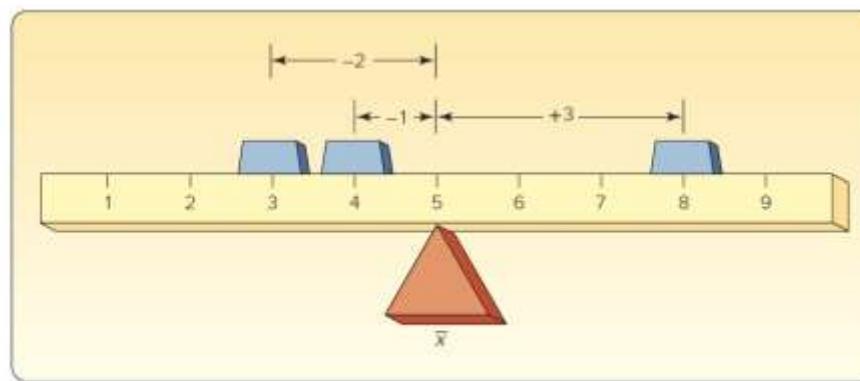


CHART 3-1 Mean as a Balance Point.

The Median

MEDIAN The midpoint of the values after they have been ordered from the minimum to the maximum values.

| Prices Ordered from Minimum to Maximum | Prices Ordered from Maximum to Minimum |
|----------------------------------------|----------------------------------------|
| \$ 60,000 | \$275,000 |
| 65,000 | 80,000 |
| 70,000 | 70,000 |
| 80,000 | 65,000 |
| 275,000 | 60,000 |

Characteristics of the Median

- ▶ The median is the value in the middle of a set of ordered data
- ▶ At least the ordinal scale of measurement is required
- ▶ It is not influenced by extreme values
- ▶ Fifty percent of the observations are larger than the median
- ▶ It is unique to a set of data

Finding the Median

- ▶ To find the median for an even numbered data set, sort the observations and calculate the average of the two middle values

The number of hours a sample of 10 adults used Facebook last month:

3 5 7 5 9 | 3 9 17 10

Arranging the data in ascending order gives:

| 3 3 5 5 7 9 9 10 17

Thus, the median is 6.

The Mode

MODE The value of the observation that occurs most frequently.

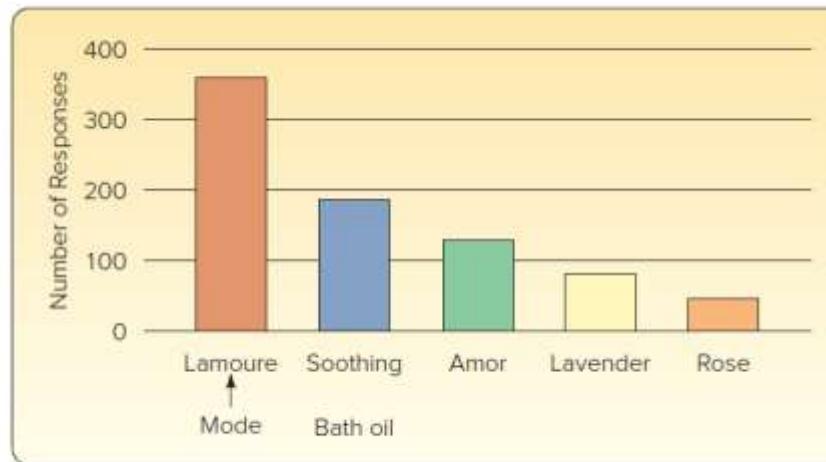


CHART 3-2 Number of Respondents Favoring Various Bath Oils

- ▶ The mode can be found for nominal level data
- ▶ A set of data can have more than one mode
- ▶ A set of data could have no mode

Relative Positions of Mean, Median, and Mode

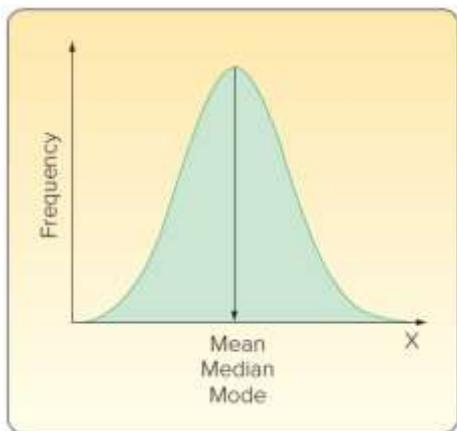


CHART 3-3 A Symmetric Distribution

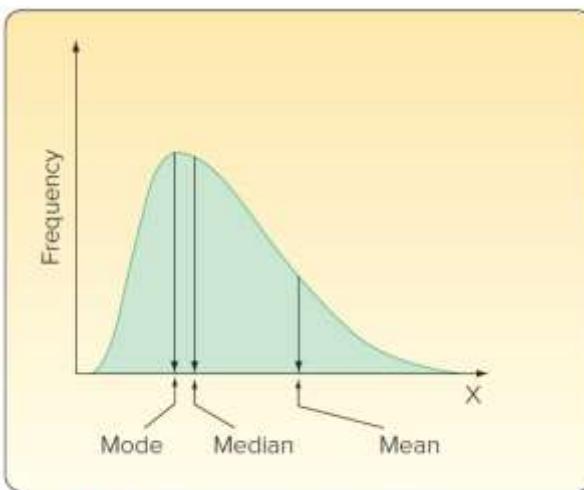


CHART 3-4 A Positively Skewed Distribution

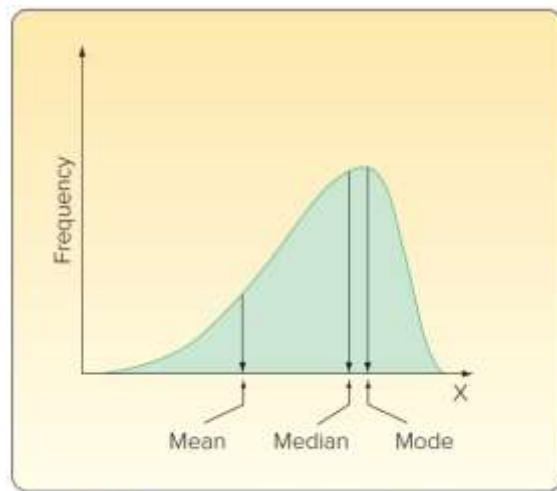


CHART 3-5 A Negatively Skewed Distribution

The Weighted Mean

- ▶ The weighted mean is found by multiplying each observation, x , by its corresponding weight, w

WEIGHTED MEAN

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + w_3x_3 + \dots + w_nx_n}{w_1 + w_2 + w_3 + \dots + w_n} \quad (3-3)$$

- ▶ The Carter Construction Company pays its hourly employees \$16.50, \$19.00, or \$25.00 per hour. There are 26 hourly employees: 14 are paid at the \$16.50 rate, 10 at the \$19.00 rate, and 2 at the \$25.00 rate
- ▶ What is the mean hourly rate paid for the 26 employees?

$$\bar{x}_w = \frac{14(\$16.50) + 10(\$19.00) + 2(\$25.00)}{14 + 10 + 2} = \frac{\$471.00}{26} = \$18.1154$$

The Geometric Mean

- ▶ The geometric mean is the nth root of the product of n positive values
- ▶ The formula for geometric mean is:

GEOMETRIC MEAN

$$GM = \sqrt[n]{(x_1)(x_2) \cdots (x_n)}$$

(3–4)

- ▶ The geometric mean is also used to find the rate of change from one period to another

RATE OF INCREASE OVER TIME

$$GM = \sqrt[n]{\frac{\text{Value at end of period}}{\text{Value at start of period}}} - 1 \quad (3-5)$$

Why Study Dispersion?

- ▶ The dispersion is the variation or spread in a set of data
- ▶ Measures of dispersion include:
 - ▶ Range
 - ▶ Variance
 - ▶ Standard Deviation

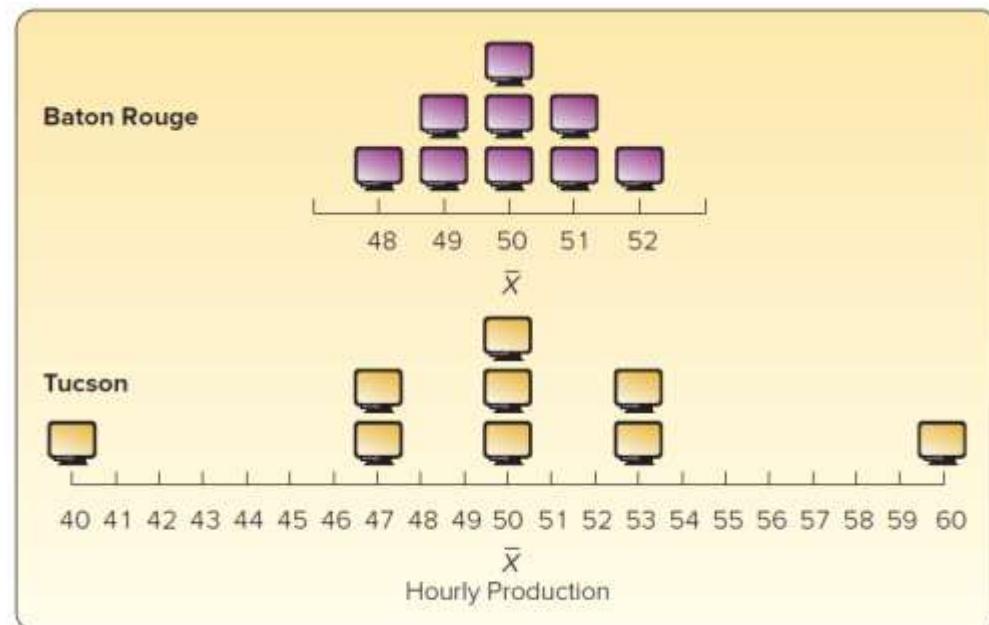


CHART 3-6 Hourly Production of Computer Monitors at the Baton Rouge and Tucson Plants

Range

- ▶ The range is the difference between the maximum and minimum values in a set of data

RANGE

Range = Maximum value – Minimum value

(3–6)

- ▶ The major characteristics of the range are
 - ▶ Only two values are used in its calculation
 - ▶ It is influenced by extreme values
 - ▶ It is easy to compute and to understand

Population Variance

POPULATION VARIANCE

$$\sigma^2 = \frac{\sum(x - \mu)^2}{N} \quad (3-7)$$

where:

- σ² is the population variance (σ is the lowercase Greek letter sigma). It is read as “sigma squared.”
- x is the value of a particular observation in the population.
- μ is the arithmetic mean of the population.
- N is the number of observations in the population.

- ▶ Major characteristics of the variance are:
 - ▶ All observations are used in the calculation
 - ▶ The units are somewhat difficult to work with; they are the original units squared

Example: Population Variance

The number of traffic citations issued last year by month in Beaufort County, South Carolina is reported below.

| Citations by Month | | | | | | | | | | | | |
|--------------------|----------|-------|-------|-----|------|------|--------|-----------|---------|----------|----------|--|
| January | February | March | April | May | June | July | August | September | October | November | December | |
| 19 | 17 | 22 | 18 | 28 | 34 | 45 | 39 | 38 | 44 | 34 | 10 | |

Determine the population variance.

| Month | Citations (x) | $x - \mu$ | $(x - \mu)^2$ |
|-----------|------------------|-----------|---------------|
| January | 19 | -10 | 100 |
| February | 17 | -12 | 144 |
| March | 22 | -7 | 49 |
| April | 18 | -11 | 121 |
| May | 28 | -1 | 1 |
| June | 34 | 5 | 25 |
| July | 45 | 16 | 256 |
| August | 39 | 10 | 100 |
| September | 38 | 9 | 81 |
| October | 44 | 15 | 225 |
| November | 34 | 5 | 25 |
| December | 10 | -19 | 361 |
| Total | 348 | 0 | 1,488 |

$$\mu = \frac{\sum x}{N} = \frac{19 + 17 + \dots + 10}{12} = \frac{348}{12} = 29$$

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N} = \frac{1,488}{12} = 124$$

So, the population variance for the number of citations is 124.

Standard Deviation

- ▶ The major characteristics of the standard deviation are:
 - ▶ It is in the same units as the original data
 - ▶ It is the square root of the average squared distance from the mean
 - ▶ It cannot be negative
 - ▶ It is the most widely used measure of dispersion

POPULATION STANDARD DEVIATION

$$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}} \quad (3-8)$$

Sample Variance and Standard Deviation

SAMPLE VARIANCE

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} \quad (3-9)$$

where:

s^2 is the sample variance.

x is the value of each observation in the sample.

\bar{x} is the mean of the sample.

n is the number of observations in the sample.

SAMPLE STANDARD DEVIATION

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} \quad (3-10)$$

Interpretations and Uses of the Standard Deviation

CHEBYSHEV'S THEOREM For any set of observations (sample or population), the proportion of the values that lie within k standard deviations of the mean is at least $1 - 1/k^2$, where k is any value greater than 1.

Dupree Paint Company employees contribute a mean of \$51.54 to the company's profit-sharing plan every two weeks. The standard deviation of biweekly contributions is \$7.51. At least what percent of the contributions lie within plus 3.5 standard deviations and minus 3.5 standard deviations of the mean, that is, between \$25.26 and \$77.83?

About 92%, found by:

$$1 - \frac{1}{k^2} = 1 - \frac{1}{(3.5)^2} = 1 - \frac{1}{12.25} = 0.92$$

Interpretations and Uses of the Standard Deviation (2 of 2)

THE EMPIRICAL RULE For a symmetrical, bell-shaped frequency distribution, approximately 68% of the observations will lie within plus and minus one standard deviation of the mean, about 95% of the observations will lie within plus or minus 2 standard deviations of the mean, and practically all (99.7%) will lie within 3 standard deviations of the mean.

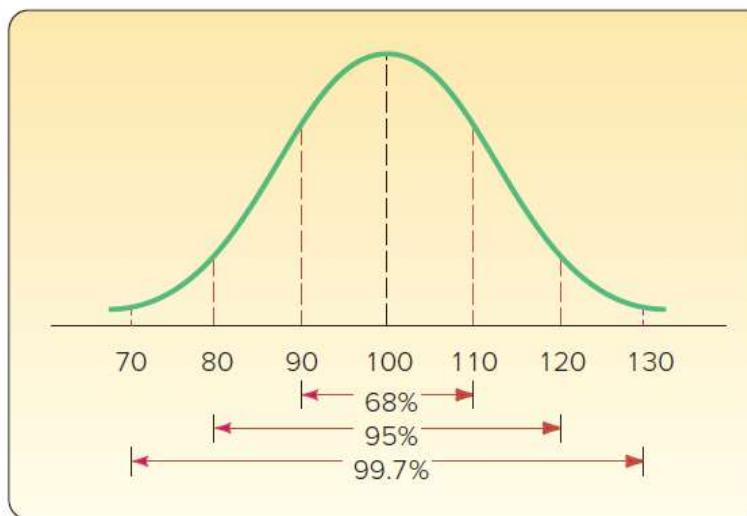


CHART 3-7 A Symmetrical, Bell-Shaped Curve Showing the Relationships between the Standard Deviation and the Percentage of Observations

Sample Mean of Grouped Data

ARITHMETIC MEAN OF GROUPED DATA

$$\bar{x} = \frac{\sum fM}{n} \quad (3-11)$$

where:

\bar{x} is the sample mean.

M is the midpoint of each class.

f is the frequency in each class.

fM is the frequency in each class times the midpoint of the class.

$\sum fm$ is the sum of these products.

n is the total number of frequencies.

Standard Deviation of Grouped Data

STANDARD DEVIATION, GROUPED DATA

$$s = \sqrt{\frac{\sum f(M - \bar{x})^2}{n - 1}} \quad (3-12)$$

where:

- s is the sample standard deviation.
- M is the midpoint of the class.
- f is the class frequency.
- n is the number of observations in the sample.
- \bar{x} is the sample mean.

Calculating the Standard Deviation of Grouped Data

- ▶ Applewood Auto Group Frequency Distribution
Compute the standard deviation of the vehicle profits.

| Profit | Frequency (<i>f</i>) | Midpoint (<i>M</i>) | <i>fM</i> | (<i>M</i> – \bar{x}) | (<i>M</i> – \bar{x}) ² | <i>f(M</i> – \bar{x}) ² |
|---------------------|------------------------|-----------------------|-----------|--------------------------|---------------------------------------|---------------------------------------|
| \$ 200 up to \$ 600 | 8 | 400 | 3,200 | -1,451 | 2,105,401 | 16,843,208 |
| 600 up to 1,000 | 11 | 800 | 8,800 | -1,051 | 1,104,601 | 12,150,611 |
| 1,000 up to 1,400 | 23 | 1,200 | 27,600 | -651 | 423,801 | 9,747,423 |
| 1,400 up to 1,800 | 38 | 1,600 | 60,800 | -251 | 63,001 | 2,394,038 |
| 1,800 up to 2,200 | 45 | 2,000 | 90,000 | 149 | 22,201 | 999,045 |
| 2,200 up to 2,600 | 32 | 2,400 | 76,800 | 549 | 301,401 | 9,644,832 |
| 2,600 up to 3,000 | 19 | 2,800 | 53,200 | 949 | 900,601 | 17,111,419 |
| 3,000 up to 3,400 | 4 | 3,200 | 12,800 | 1,349 | 1,819,801 | 7,279,204 |
| Total | 180 | | 333,200 | | | 76,169,780 |

$$s = \sqrt{\frac{\sum f(M - \bar{x})^2}{n - 1}} = \sqrt{\frac{76,169,780}{180 - 1}} = 652.33$$

Ethics and Reporting Results

- ▶ Useful to know the advantages and disadvantages of mean, median, and mode as we report statistics and as we use statistics to make decisions
- ▶ Important to maintain an independent and principled point of view
- ▶ Statistical reporting requires objective and honest communication of any results



Chapter 3 Practice Problems

Question 19

LO3-I

The accounting firm of Rowatti and Koppel specializes in income tax returns for self-employed professionals, such as physicians, dentists, architects, and lawyers. The firm employs 11 accountants who prepare the returns. For last year, the number of returns prepared by each accountant was:

58 75 31 58 46 65 60 71 45 58 80

Find the mean, median, and mode for the number of returns prepared by each accountant. If you could report only one, which measure of location would you recommend reporting?

Question 25

LO3-2

The Loris Healthcare System employs 200 persons on the nursing staff. Fifty are nurse's aides, 50 are practical nurses, and 100 are registered nurses. Nurse's aides receive \$12 an hour, practical nurses \$20 an hour, and registered nurses \$29 an hour. What is the weighted mean hourly wage?

Question 27

LO3-3

Compute the geometric mean of the following monthly percent increases: 8, 12, 14, 26, and 5.

Question 33

LO3-3

In 2011 there were 232.2 million cell phone subscribers in the United States. By 2017 the number of subscribers increased to 265.9 million.

- a. What is the geometric mean annual percent increase for the period?
- b. Further, the number of subscribers is forecast to increase to 276.7 million by 2020.
- c. What is the rate of increase from 2017 to 2020?
- d. Is the rate of increase expected to slow?

Question 37

LO3-1, 4

Dave's Automatic Door installs automatic garage door openers. The following list indicates the number of minutes needed to install 10 door openers: 28, 32, 24, 46, 44, 40, 54, 38, 32, and 42.

Calculate the following:

- a. Range
- b. Mean
- c. Variance



Question 45

LO3-1, 4

Plywood Inc. reported these returns on stockholder equity for the past 5 years: 4.3, 4.9, 7.2, 6.7, and 11.6. Consider these as population values.

Compute the following:

- a. Range
- b. Arithmetic mean
- c. Variance
- d. Standard deviation



Question 49

LO3-4

Dave's Automatic Door installs automatic garage door openers. Based on a sample, following are the times, in minutes, required to install 10 door openers: 28, 32, 24, 46, 44, 40, 54, 38, 32, and 42.

- a. Compute the sample variance.
- b. Determine the sample standard deviation.

Question 53

LO3-5

According to Chebyshev's theorem, at least what percent of any set of observations will be within 1.8 standard deviations of the mean?



Question 55

LO3-5

The distribution of the weights of a sample of 1,400 cargo containers is symmetric and bell-shaped. According to the Empirical Rule, what percent of the weights will lie:

- ▶ Between $\bar{x} - 2s$ and $\bar{x} + 2s$?
- ▶ Between \bar{x} and $\bar{x} + 2s$? Above $\bar{x} + 2s$?

Question 59

LO3-6

Estimate the mean and the standard deviation of the following frequency distribution showing the ages of the first 60 people in line on Black Friday at a retail store.

| Class | Frequency |
|-------------|-----------|
| 20 up to 30 | 7 |
| 30 up to 40 | 12 |
| 40 up to 50 | 21 |
| 50 up to 60 | 18 |
| 60 up to 70 | 12 |



Describing Data: Displaying and Exploring Data



Chapter 4

Measures of Position

- The standard deviation is the most widely used measure of dispersion.
- Another way to describe the spread of data is using the position of values that divide a set of observations into equal parts.
- These measures of position include **quartiles**, **deciles**, and **percentiles**.

Measures of Position

- **Quartiles** divide a set of observations into four equal parts.
- **Deciles** divide a set of observations into ten equal parts.
- **Percentile** divide a set of observations into hundred equal parts.

Percentile Computation

- To compute a percentile, let L_p refer to the location of a desired percentile. So if we wanted to find the 33rd percentile we would use L_{33} and if we wanted the median, the 50th percentile, then L_{50} .

LOCATION OF A PERCENTILE

$$L_p = (n + 1) \frac{P}{100} \quad [4-1]$$

- The number of observations is n , so if we want to locate the median, its position is at $(n + 1)/2$, or we could write this as $(n + 1)(P/100)$, where P is the desired percentile.

Percentiles - Example

Listed below are the commissions earned last month by a sample of 15 brokers at Salomon Smith Barney's Oakland, California, office.

| | | | |
|---------|---------|---------|---------|
| \$2,038 | \$1,758 | \$1,721 | \$1,637 |
| \$2,097 | \$2,047 | \$2,205 | \$1,787 |
| \$2,287 | \$1,940 | \$2,311 | \$2,054 |
| \$2,406 | \$1,471 | \$1,460 | |

Locate the median, the first quartile, and the third quartile for the commissions earned.

Percentiles – Example

Step 1: Organize the data from lowest to largest value.

| | | | | | | | |
|---------|---------|---------|---------|---------|---------|---------|---------|
| \$1,460 | \$1,471 | \$1,637 | \$1,721 | \$1,758 | \$1,787 | \$1,940 | \$2,038 |
| 2,047 | 2,054 | 2,097 | 2,205 | 2,287 | 2,311 | 2,406 | |

Percentiles – Example

Step 2: Compute the first and third quartiles. Locate L_{25} and L_{75}

LOCATION OF A PERCENTILE

$$L_p = (n + 1) \frac{P}{100} \quad [4-1]$$

$$L_{25} = (15 + 1) \frac{25}{100} = 4 \quad L_{75} = (15 + 1) \frac{75}{100} = 12$$

Therefore, the first and third quartiles are located at the 4th and 12th positions, respectively : $L_{25} = \$1,721$; $L_{75} = \$2,205$

The median corresponds to the one in the middle, or $L_{50} = 2038$

| | | | |
|---------|---------|---------|----------------|
| \$1,460 | \$1,471 | \$1,637 | \$1,721 |
| \$1,758 | \$1,787 | \$1,940 | \$2,038 |
| \$2,047 | \$2,054 | \$2,097 | \$2,205 |
| \$2,287 | \$2,311 | \$2,406 | |

Percentiles – Example

In the previous example the location formula yielded a whole number. What if there were 6 observations in the sample with the following ordered observations: 43, 61, 75, 91, 101, and 104 , that is $n=6$, and we wanted to locate the first quartile?

$$L_{25} = (6 + 1) \frac{25}{100} = 1.75$$

Locate the first value in the ordered array and then move .75 of the distance between the first and second values and report that as the first quartile. Like the median, the quartile does not need to be one of the actual values in the data set.

The 1st and 2nd values are 43 and 61. Moving 0.75 of the distance between these numbers, the 25th percentile is **56.5**, obtained as **43 + 0.75*(61- 43)**.

Skewness

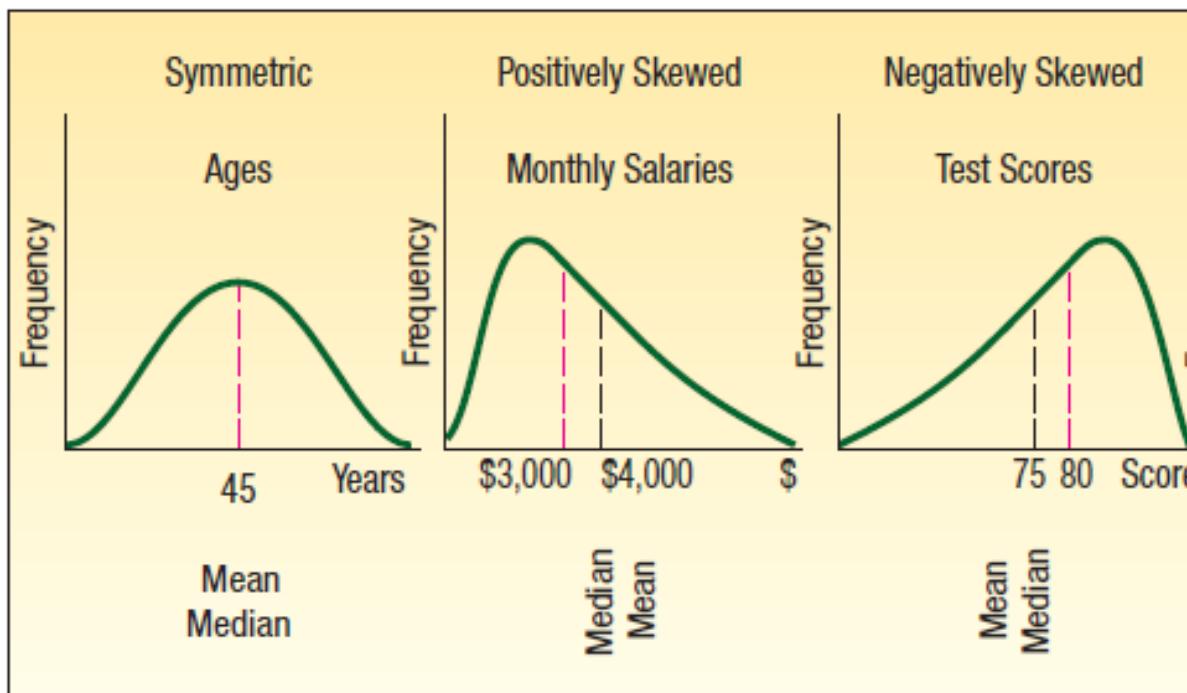
- Chapter 3 introduced measures of central location (the mean, median, and mode) and measures of dispersion (the range and standard deviation) for a distribution of data.
- **Shape** is another characteristic of a distribution.
- There are four shapes commonly observed:
 1. symmetric,
 2. positively skewed,
 3. negatively skewed, and

Computing the Coefficient of Skewness

Skewness can be calculated using Pearson's Coefficient of Skewness formula:

$$\text{Skewness} = \frac{n}{(n - 1)(n - 2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Commonly Observed Shapes



Excel Application

The screenshot shows a Microsoft Excel interface with the following details:

- File Tab:** Home, Insert, Page Layout, Formulas, Data (selected), Review, View, Bloomberg, Tell me what you want to do.
- Data Tab Submenu:** Get External Data, New Query, Refresh All, Connections, Sort & Filter, Data Tools, Forecast, Outline, Analysis.
- Table:** A data table with columns: Age, Profit, Location, Vehicle-Type, Previous. Rows 1 through 15 are visible.
- Dialog Box:** Data Analysis dialog box is open, showing the following options:
 - Anova - Single Factor
 - Anova - Two Factor With Replication
 - Anova - Two Factor Without Replication
 - Correlation
 - Covariance
 - Descriptive Statistics** (highlighted)
 - Exponential Smoothing
 - F-Test Two-Sample for Variances
 - Fourier Analysis
 - Histogram
- Taskbar:** Shows the Windows Start button, taskbar icons for File Explorer, Edge, Google Chrome, File, and Excel, system tray icons, and the date/time (12:01, 26.8.2016).

Red numbered arrows indicate the steps:

- 1: Points to the Data tab in the ribbon.
- 2: Points to the Data Analysis button in the ribbon.
- 3: Points to the Descriptive Statistics option in the Data Analysis dialog box.

4. Choose the profit column for the Input Range
5. Click «Summary Statistics». You'll have

Applewood_Dataset - Excel

File Home Insert Page Layout Formulas Data Review View Bloomberg Tell me what you want to do Sign in Share

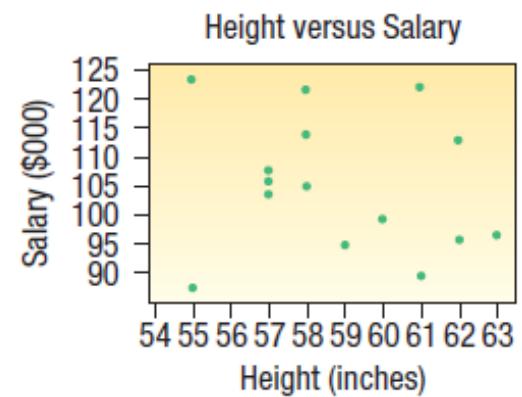
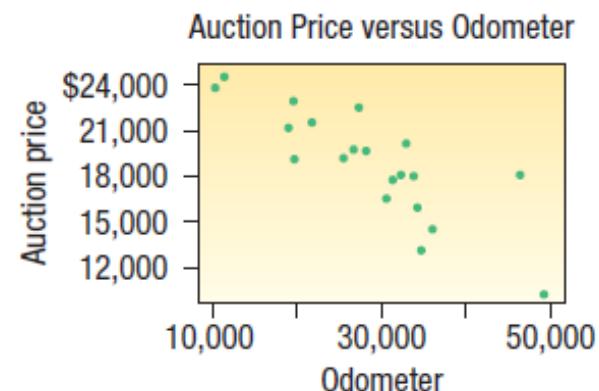
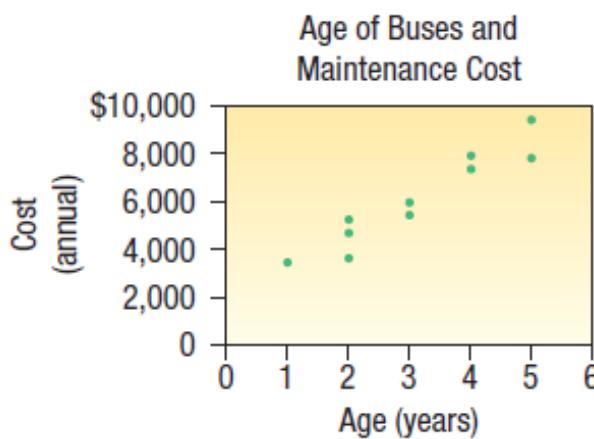
C23

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|----|--------------------------|-----------|---|---|---|---|---|---|---|---|---|---|---|
| 1 | <i>Profit Statistics</i> | | | | | | | | | | | | |
| 3 | Mean | 1843.17 | | | | | | | | | | | |
| 4 | Standard Error | 47.97 | | | | | | | | | | | |
| 5 | Median | 1882.50 | | | | | | | | | | | |
| 6 | Mode | 1761.00 | | | | | | | | | | | |
| 7 | Standard Deviation | 643.63 | | | | | | | | | | | |
| 8 | Sample Variance | 414256.60 | | | | | | | | | | | |
| 9 | Kurtosis | -0.22 | | | | | | | | | | | |
| 10 | Skewness | -0.24 | | | | | | | | | | | |
| 11 | Range | 2998.00 | | | | | | | | | | | |
| 12 | Minimum | 294.00 | | | | | | | | | | | |
| 13 | Maximum | 3292.00 | | | | | | | | | | | |
| 14 | Sum | 331770.00 | | | | | | | | | | | |
| 15 | Count | 180.00 | | | | | | | | | | | |
| 16 | | | | | | | | | | | | | |
| 17 | | | | | | | | | | | | | |
| 18 | | | | | | | | | | | | | |

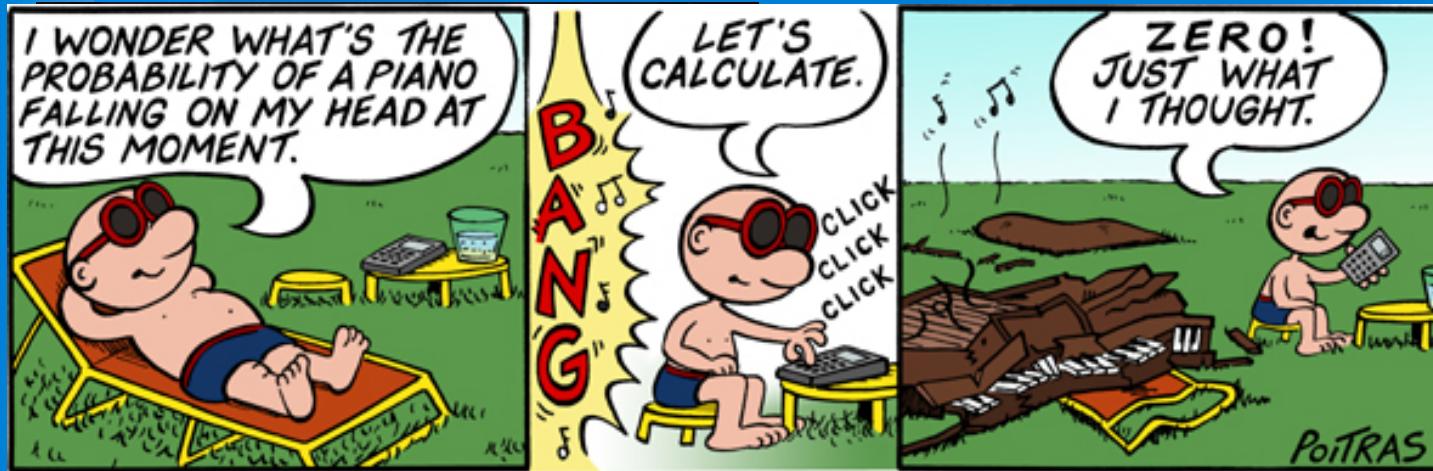
Describing the Relationship between Two Variables

- When we study the relationship between two variables we refer to the data as **bivariate**.
- One graphical technique we use to show the relationship between two variables is called a **scatter diagram**.
- To draw a scatter diagram, we scale one variable along the horizontal axis (X-axis) of a graph and the other variable along the vertical axis (Y-axis).

Scatter Diagram Examples



Chapter 5: A Survey of Probability Concepts

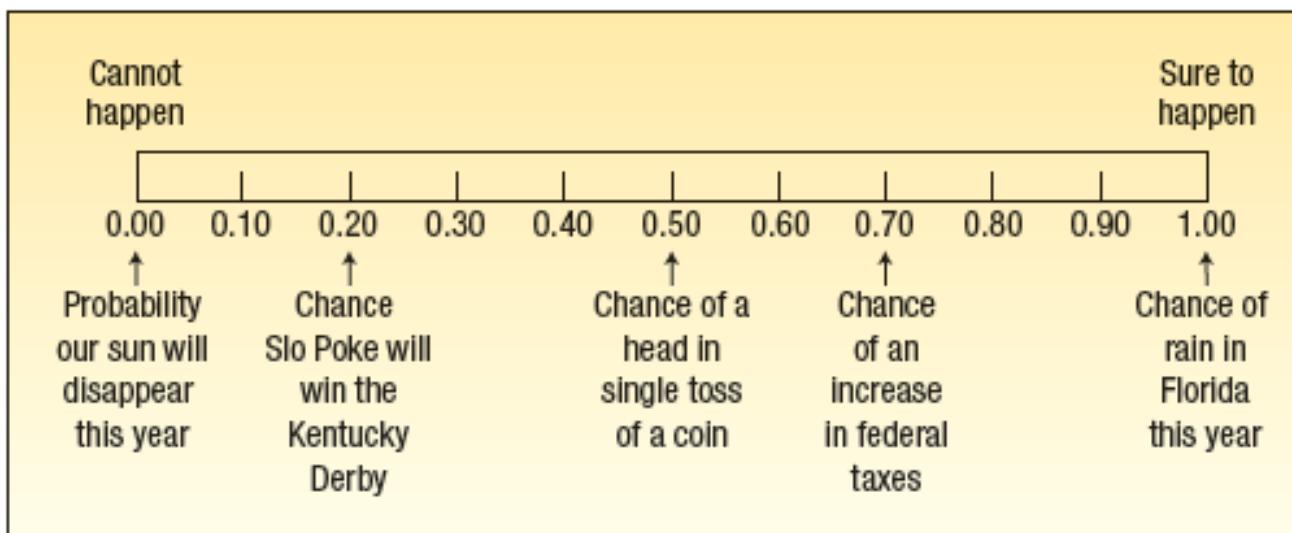


Learning Objectives

- **LO5-1** Define the terms *probability*, *experiment*, *event*, and *outcome*.
- **LO5-2** Assign probabilities using a classical, empirical, or subjective approach.
- **LO5-3** Calculate probabilities using the rules of addition.
- **LO5-4** Calculate probabilities using the rules of multiplication.
- **LO5-5** Compute probabilities using a contingency table.
- **LO5-7** Determine the number of outcomes using principles of counting.

Probability

PROBABILITY A value between zero and one, inclusive, describing the relative possibility (chance or likelihood) an event will occur.



Experiment, Outcome, and Event

- An **experiment** is a process that leads to the occurrence of one and only one of several possible results.
- An **outcome** is the particular result of an experiment.
- An **event** is the collection of one or more outcomes of an experiment.

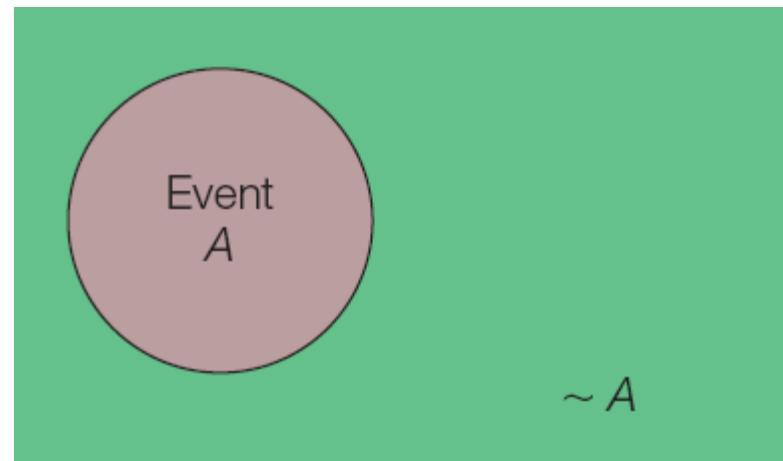
| |  |  |
|-----------------------|-----------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|
| Experiment | Roll a die | Count the number of members of the board of directors for Fortune 500 companies who are over 60 years of age |
| All possible outcomes | Observe a 1 Observe a 2 Observe a 3 Observe a 4 Observe a 5 Observe a 6 | None is over 60 One is over 60 Two are over 60 ... 29 are over 60 48 are over 60 ... |
| Some possible events | Observe an even number Observe a number greater than 4 Observe a number 3 or less | More than 13 are over 60 Fewer than 20 are over 60 |

Basic Probability: The Complement Rule

The **complement rule** is used to determine the probability of an event occurring by subtracting the probability of the event *not* occurring from 1.

$$P(A) + P(\sim A) = 1$$

or $P(A) = 1 - P(\sim A)$.



The Complement Rule - Example

An experiment has two mutually exclusive outcomes. Based on the rules of probability, the sum of the probabilities must be one. If the probability of the first outcome is .61, then logically, AND by the complement rule, the probability of the other outcome is $(1.0 - .61) = .39$.

$$\begin{aligned}P(B) &= 1 - P(\sim B) \\&= 1 - .61 \\&= .39\end{aligned}$$

Ways to Assign Probabilities

There are three ways to assign probabilities:

1. CLASSICAL PROBABILITY

Based on the assumption that the outcomes of an experiment are *equally likely*.

2. EMPIRICAL PROBABILITY

The probability of an event happening is the fraction of the time similar events happened in the past.

3. SUBJECTIVE PROBABILITY

The likelihood (probability) of a particular event happening that is assigned by an individual based on whatever information is available.

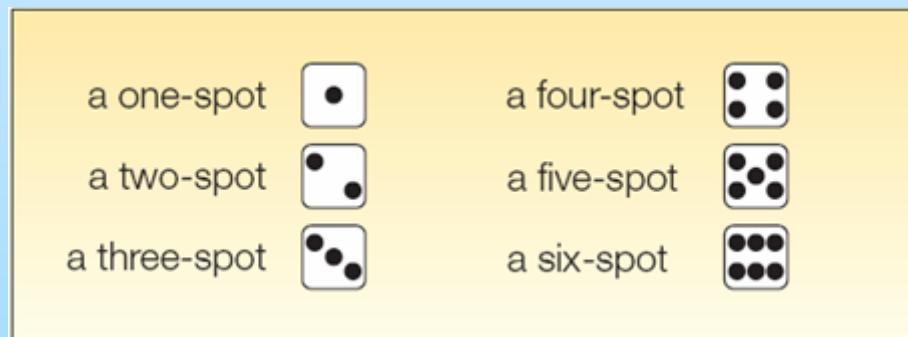
1. Classical Probability

CLASSICAL PROBABILITY

$$\text{Probability of an event} = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}} \quad [5-1]$$

Consider an experiment of rolling a six-sided die. What is the probability of the event: “an **even number** of spots appear face up”?

The possible outcomes are:



There are three “favorable” outcomes (a two, a four, and a six) in the collection of six equally likely possible outcomes.

2. Empirical Probability

EMPIRICAL PROBABILITY The probability of an event happening is the fraction of the time similar events happened in the past.

Empirical approach to probability is based on what is called the Law of Large Numbers.

LAW OF LARGE NUMBERS Over a large number of trials, the empirical probability of an event will approach its true probability.

The key to establishing probabilities empirically: a larger number of observations provides a more accurate estimate of the probability.

2. Empirical Probability - Example

On February 1, 2003, the Space Shuttle Columbia exploded. This was the second disaster in 123 space missions for NASA. On the basis of this information, what is the probability that a future mission is successfully completed?

$$\begin{aligned} \text{Probability of a successful flight} &= \frac{\text{Number of successful flights}}{\text{Total number of flights}} \\ &= \frac{121}{123} = 0.98 \end{aligned}$$

Subjective Probability - Example

SUBJECTIVE CONCEPT OF PROBABILITY The likelihood (probability) of a particular event happening that is assigned by an individual based on whatever information is available.

- If there is little or no data or information to calculate a probability, it may be arrived at subjectively.
- Illustrations of subjective probability are:
 1. Estimating the likelihood the New England Patriots will play in the Super Bowl next year.
 2. Estimating the likelihood a person will be married before the age of 30.
 3. Estimating the likelihood the U.S. budget deficit will be reduced by half in the next 10 years.

Rules of Addition for Computing Probabilities

- **Special Rule of Addition** - If two events A and B are mutually exclusive, the probability of one or the other event occurring equals the sum of their probabilities.

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

- Events are mutually exclusive if the occurrence of any one event means that none of the others can occur at the same time.

Special Rule of Addition- Example

Mutually Exclusive Events

A machine fills plastic bags with a mixture of beans, broccoli, and other vegetables. Most of the bags contain the correct weight, but because of the variation in the size of the beans and other vegetables, a package might be underweight or overweight. A check of 4,000 packages filled in the past month revealed:

| Weight | Event | Number of Packages | Probability of Occurrence |
|--------------|-------|--------------------|---------------------------|
| Underweight | A | 100 | .025 |
| Satisfactory | B | 3,600 | .900 |
| Overweight | C | 300 | .075 |
| | | 4,000 | 1.000 |

$\leftarrow \frac{100}{4,000}$

What is the probability that a particular package will be either underweight or overweight?

Being overweight/underweight cannot happen at the same time → mutually exclusive

$$P(A \text{ or } C) = P(A) + P(C) = .025 + .075 = .10$$

Complement Rule- Example Mutually Exclusive Events

The complement rule can also be used:

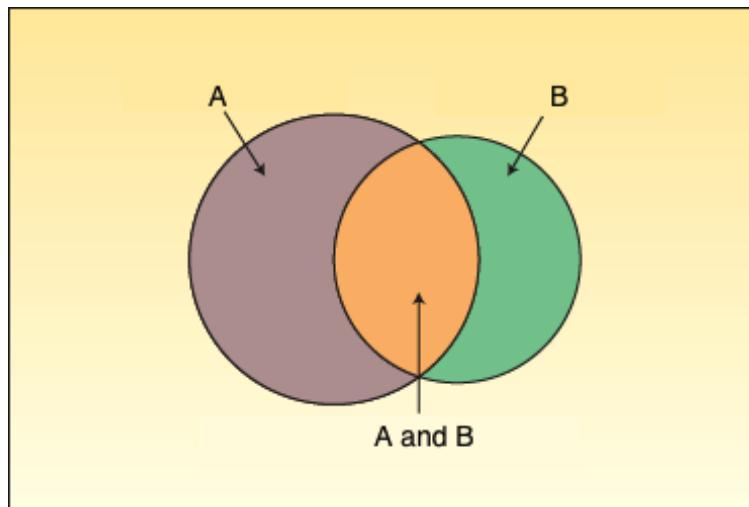
Note that $P(A \text{ or } C) = P(\sim B)$, so
 $P(\sim B) = 1 - P(B) = 1 - .900 = .10$

| Weight | Event | Number of Packages | Probability of Occurrence | |
|--------------|-------|--------------------|---------------------------|-----------------------|
| Underweight | A | 100 | .025 | ← $\frac{100}{4,000}$ |
| Satisfactory | B | 3,600 | .900 | |
| Overweight | C | 300 | .075 | |
| | | 4,000 | 1.000 | |

Rules of Addition for Computing Probabilities

The General Rule of Addition - If A and B are two events that are not mutually exclusive, then $P(A \text{ or } B)$ is given by the following formula:

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

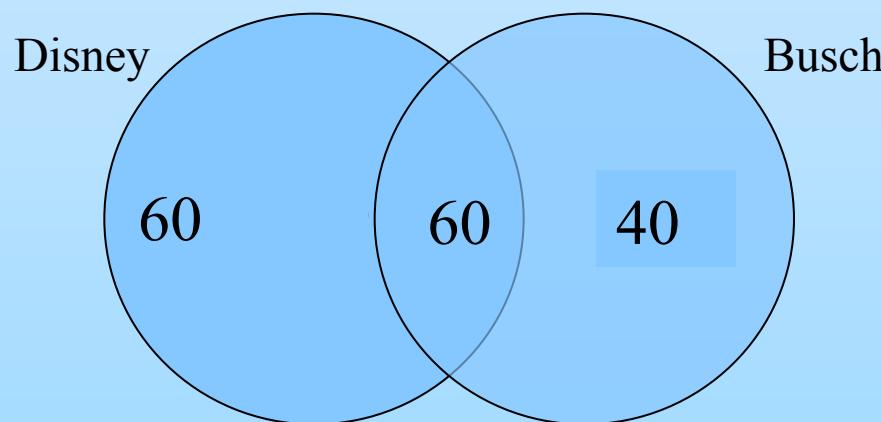


$P(A \text{ and } B)$ is called a **joint probability**.

The General Rule of Addition

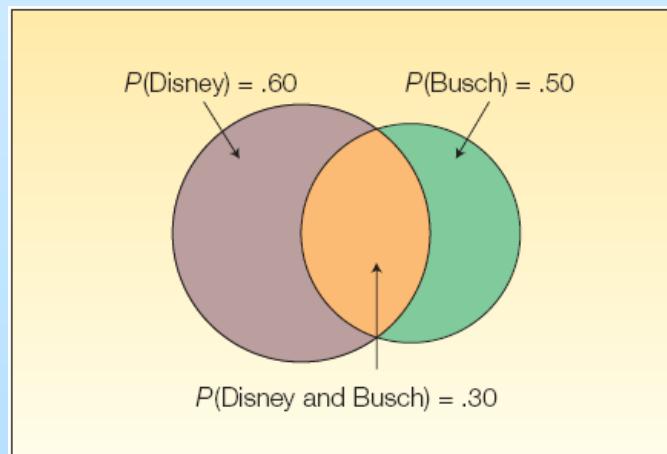
The Venn Diagram shows the results of a survey of 200 tourists who visited Florida during the year. The results revealed that 120 went to Disney World, 100 went to Busch Gardens, and 60 visited both.

What is the probability a selected person visited either Disney World or Busch Gardens?



The General Rule of Addition

What is the probability a selected person visited either Disney World or Busch Gardens?



$$\begin{aligned}P(\text{Disney or Busch}) &= P(\text{Disney}) + P(\text{Busch}) - P(\text{both Disney and Busch}) \\&= 120/200 + 100/200 - 60/200 \\&= .60 + .50 - .30 = 0.80\end{aligned}$$

Special Rule of Multiplication

- The **special rule of multiplication** requires that the two events A and B are *independent*.
- Two events A and B are independent if the occurrence of one has no effect on the probability of the occurrence of the other.
- This rule is written: $P(A \text{ and } B) = P(A)P(B)$

Special Rule of Multiplication-Example

A survey by the American Automobile Association (AAA) revealed 60 percent of its members made airline reservations last year. Two members are selected at random. What is the probability *both* made airline reservations last year?

(Since the number of AAA members is very large, we can assume that R_1 and R_2 are *independent*.)

Solution:

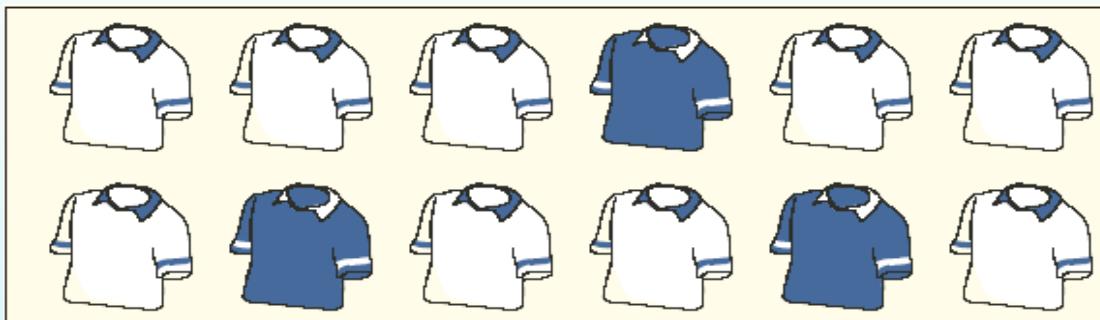
The probability the first member made an airline reservation last year is .60, written as $P(R_1) = .60$

The probability that the second member selected made a reservation is also .60, so $P(R_2) = .60$.

$$P(R_1 \text{ and } R_2) = P(R_1)P(R_2) = (.60)(.60) = .36$$

General Rule of Multiplication - Example

A golfer has 12 golf shirts in his closet. Suppose 9 of these shirts are white and the others blue. He gets dressed in the dark, so he just grabs a shirt and puts it on.



Conditional Probability

- A **conditional probability** is the probability of a particular event occurring, given that another event has occurred.
- The probability of the event A given that the event B has occurred is written $P(A|B)$.

General Rule of Multiplication

GENERAL RULE OF MULTIPLICATION

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

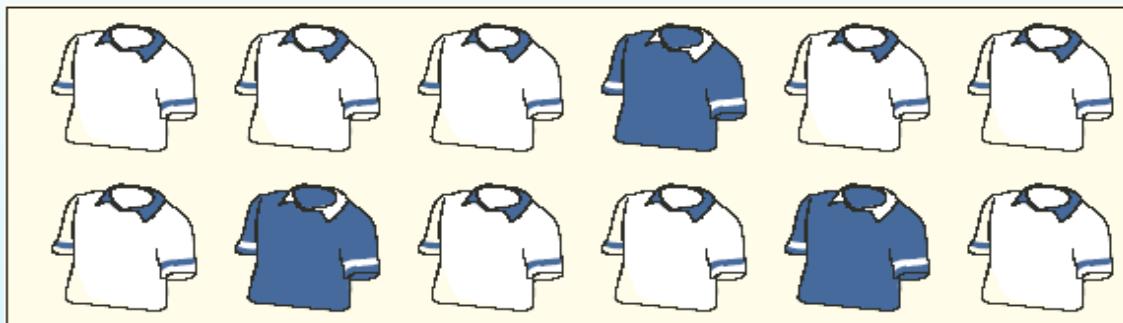
[5-6]

The **general rule of multiplication** is used to find the joint probability that two events will occur when they are not independent.

It states that for two events, A and B , the joint probability that both events will happen is found by multiplying the probability that event A will happen by the **conditional probability** of event B occurring given that A has occurred.

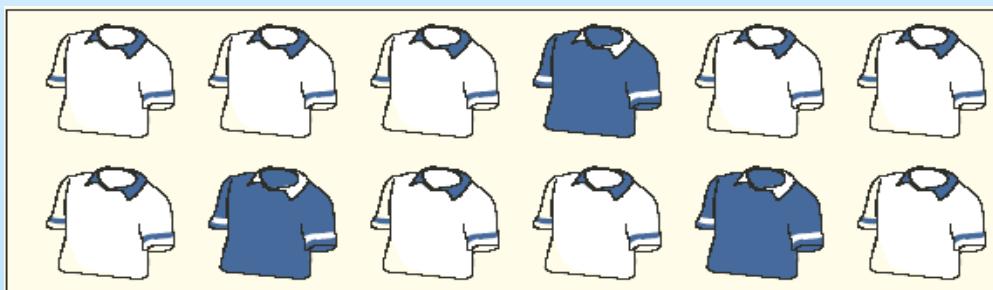
General Rule of Multiplication - Example

A golfer has 12 golf shirts in his closet. Suppose 9 of these shirts are white and the others blue. He gets dressed in the dark, so he just grabs a shirt and puts it on. He plays golf two days in a row and does return it to the closet.



What is the likelihood both shirts selected are white?

General Rule of Multiplication - Example



- The probability that the first shirt selected is white is $P(W_1) = 9/12$.
- The conditional probability is the probability the second shirt selected is white, given that first shirt selected is also white: $P(W_2 | W_1) = 8/11$.
- Apply the General Multiplication Rule: $P(A \text{ and } B) = P(A) P(B|A)$
- The joint probability of selecting 2 white shirts is:
$$P(W_1 \text{ and } W_2) = P(W_1)P(W_2 | W_1) = (9/12)(8/11) = 0.55$$

Contingency Tables

A **contingency table** is used to classify sample observations according to two or more identifiable characteristics measured.

For example, 150 adults are surveyed about their attendance of movies during the last 12 months. Each respondent is classified according to two criteria—the **number of movies attended** and **gender**.

| Movies Attended | Gender | | Total |
|-----------------|--------|-------|-------|
| | Men | Women | |
| 0 | 20 | 40 | 60 |
| 1 | 40 | 30 | 70 |
| 2 or more | 10 | 10 | 20 |
| Total | 70 | 80 | 150 |

Contingency Table - Example

Based on the survey, what is the probability that a person attended zero movies?

| Movies Attended | Gender | | Total |
|-----------------|--------|-------|-------|
| | Men | Women | |
| 0 | 20 | 40 | 60 |
| 1 | 40 | 30 | 70 |
| 2 or more | 10 | 10 | 20 |
| Total | 70 | 80 | 150 |

Based on the empirical information,
 $P(\text{zero movies}) = 60/150 = 0.4$

Contingency Table - Example

Based on the survey, what is the probability that a person attended zero movies or is male?

| Movies Attended | Gender | | Total |
|-----------------|--------|-------|-------|
| | Men | Women | |
| 0 | 20 | 40 | 60 |
| 1 | 40 | 30 | 70 |
| 2 or more | 10 | 10 | 20 |
| Total | 70 | 80 | 150 |

Applying the General Rule of Addition:

$$P(\text{zero movies or male}) = P(\text{zero movies}) + P(\text{male}) - P(\text{zero movies and male})$$

$$P(\text{zero movies or male}) = 60/150 + 70/150 - 20/150 = 0.733$$

Contingency Table - Example

Based on the survey, what is the probability that a person attended zero movies if a person is male?

| Movies Attended | Gender | | Total |
|-----------------|--------|-------|-------|
| | Men | Women | |
| 0 | 20 | 40 | 60 |
| 1 | 40 | 30 | 70 |
| 2 or more | 10 | 10 | 20 |
| Total | 70 | 80 | 150 |

Applying the concept of conditional probability:
 $P(\text{zero movies} | \text{male}) = 20/70 = 0.286$

Contingency Table - Example

Based on the survey, what is the probability that a person is male and attended zero movies?

| Movies Attended | Gender | | Total |
|-----------------|--------|-------|-------|
| | Men | Women | |
| 0 | 20 | 40 | 60 |
| 1 | 40 | 30 | 70 |
| 2 or more | 10 | 10 | 20 |
| Total | 70 | 80 | 150 |

Applying the General Rule of Multiplication
 $P(\text{male and zero movies}) = (20/150) = 0.133$

Counting Rules – Multiplication

The **multiplication formula** indicates that if there are m ways of doing one thing and n ways of doing another thing, there are $m \times n$ ways of doing both.

MULTIPLICATION FORMULA

Total number of arrangements = $(m)(n)$ [5-8]

Example: Dr. Delong has 10 shirts and 8 ties.

How many shirt and tie outfits does he have?

$$(10)(8) = 80$$

Counting Rules - Permutation

A **permutation** is any arrangement of r objects selected from n possible objects. The order of arrangement is important in permutations.

PERMUTATION FORMULA

$${}_n P_r = \frac{n!}{(n - r)!}$$

[5-9]

where:

n is the total number of objects.

r is the number of objects selected.

Example

- There are 3 electronics part (A,B,C) that are to be assembled in any order, in how many different ways can they be assembled?

$${}_3P_3 = \frac{3!}{(3-3)!} = 6$$

ABC, BAC, CAB, ACB, BCA, CBA



Screenshot of Microsoft Excel showing the PERMUT function dialog box.

The formula bar shows the formula `=PERMUT(3;3)`. A large red arrow points from the formula bar down to the "OK" button in the dialog box.

Function Arguments

PERMUT

| | | |
|---------------|---|-----|
| Number | 3 | = 3 |
| Number_chosen | 3 | = 3 |
| = 6 | | |

Returns the number of permutations for a given number of objects that can be selected from the total objects.

Number_chosen is the number of objects in each permutation.

Formula result = 6

[Help on this function](#)

OK Cancel

Sheet1

File Home Insert Page Layout Formulas Data Review View Tell me what you want to do

Cut Copy Format Painter

Clipboard Font Alignment Number Styles Cells Editing

PERMUT : X ✓ fx =PERMUT(3;3)

A B C D E F G H I J K L M N O P Q R

1 1UT(3;3) 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

OK Cancel

AutoSum Fill Clear Sort & Filter Select

Counting Rules - Combination

A **combination** is the number of ways to choose r objects from a group of n objects without regard to order.

COMBINATION FORMULA

$${}_nC_r = \frac{n!}{r!(n-r)!}$$

[5-10]

where:

n is the total number of objects.

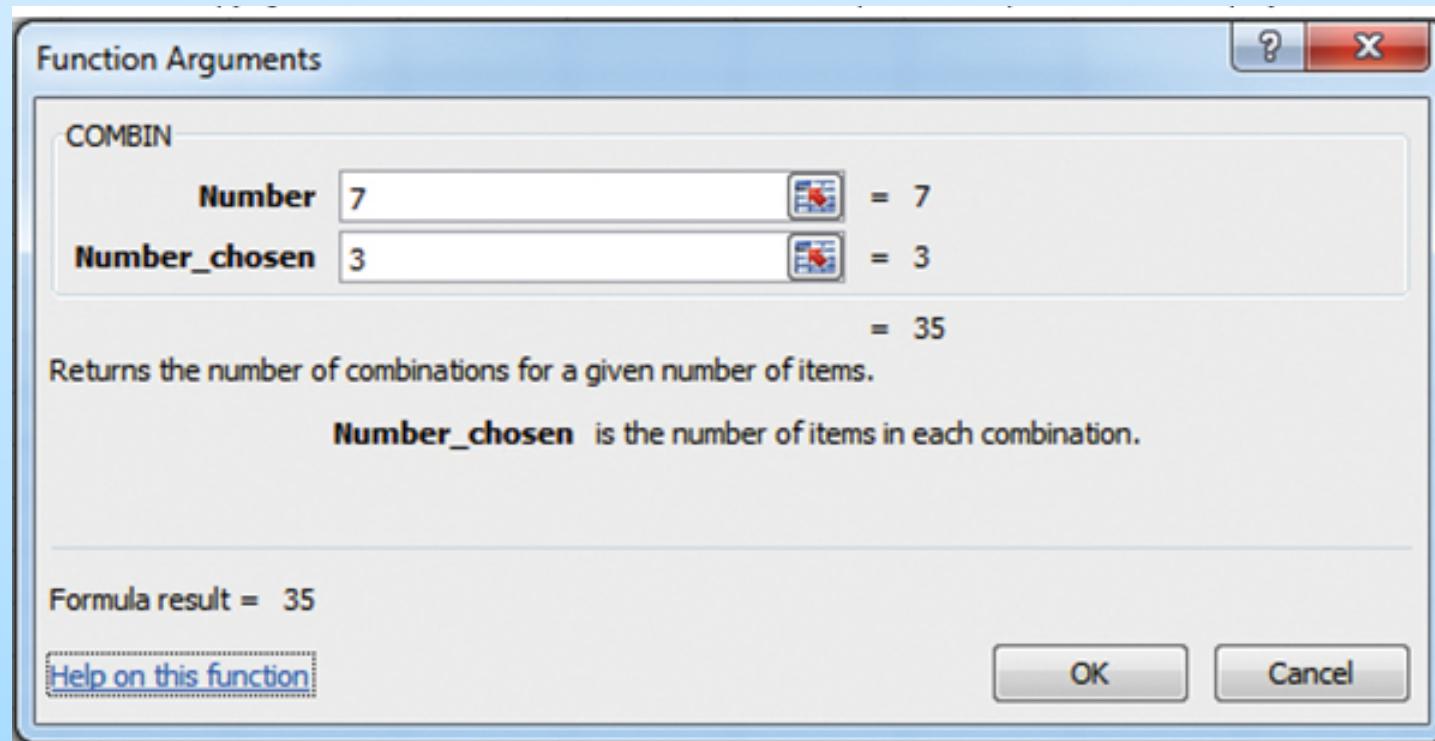
r is the number of objects selected.

Example

- The Grand 16 movie theatre uses teams of three employees to work concession stand each evening. There are seven employees available to work each evening. How many different teams can be scheduled to staff the concession stand?

$${}_7C_3 = \frac{7!}{3!(7-3)!} = 35$$

Excel Application



Discrete Probability Distributions

Chapter 6



Learning Objectives

- **LO6-1** Identify the characteristics of a probability distribution.
- **LO6-2** Distinguish between discrete and continuous random variables.
- **LO6-3** Compute the mean, variance, and standard deviation of a discrete probability distribution.
- **LO6-4** Explain the assumptions of the binomial distribution and apply it to calculate probabilities.
- **LO6-5** Explain the assumptions of the hypergeometric distribution and apply it to calculate probabilities.
- **LO6-6** Explain the assumptions of the Poisson distribution and apply it to calculate probabilities.

What is a Probability Distribution?

PROBABILITY DISTRIBUTION A listing of all the outcomes of an experiment and the probability associated with each outcome.

Characteristics of a Probability Distribution

- The probability of a particular outcome is between 0 and 1 inclusive.
- The outcomes are mutually exclusive events.
- The list is exhaustive. So the sum of the probabilities of the various events is equal to 1.

Probability Distribution - Example

Experiment: Toss a coin three times. Observe the number of heads.

The possible experimental outcomes are: zero heads, one head, two heads, and three heads.

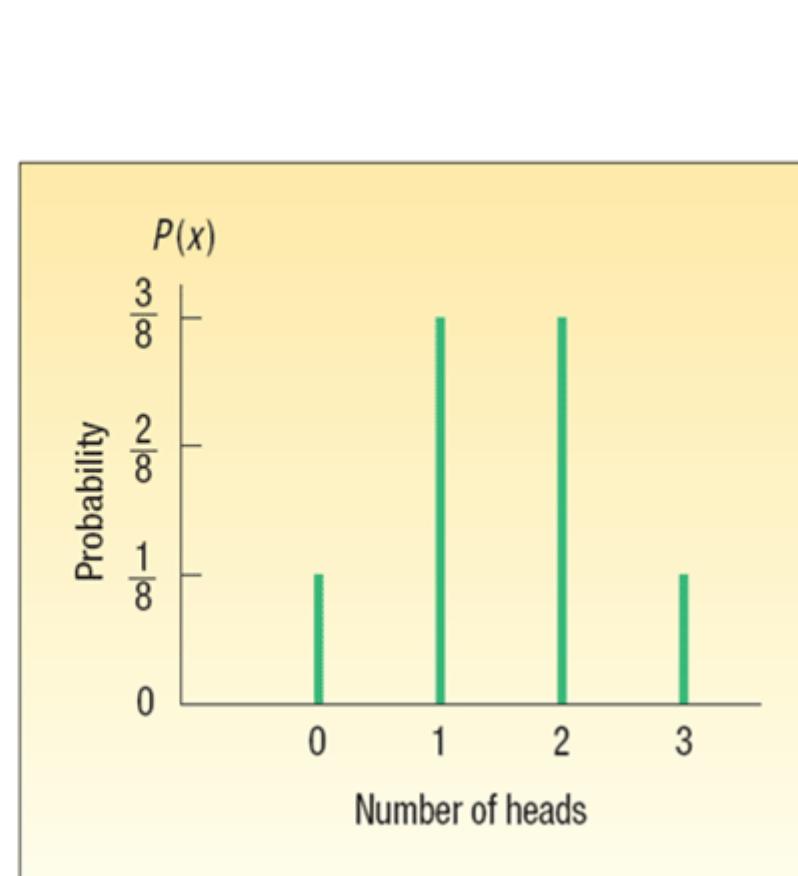
| Possible Result | Coin Toss | | | Number of Heads |
|-----------------|-----------|--------|-------|-----------------|
| | First | Second | Third | |
| 1 | T | T | T | 0 |
| 2 | T | T | H | 1 |
| 3 | T | H | T | 1 |
| 4 | T | H | H | 2 |
| 5 | H | T | T | 1 |
| 6 | H | T | H | 2 |
| 7 | H | H | T | 2 |
| 8 | H | H | H | 3 |

$\frac{1}{8}$

What is the probability distribution for the number of heads?

Probability Distribution: Number of Heads in 3 Tosses of a Coin

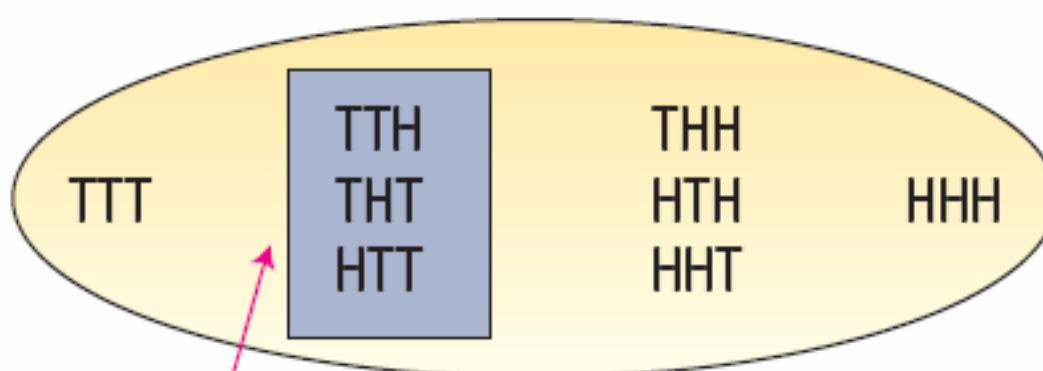
| Number of Heads, x | Probability of Outcome, $P(x)$ |
|-------------------------|-----------------------------------|
| 0 | $\frac{1}{8} = .125$ |
| 1 | $\frac{3}{8} = .375$ |
| 2 | $\frac{3}{8} = .375$ |
| 3 | $\frac{1}{8} = .125$ |
| Total | $\frac{8}{8} = 1.000$ |



Random Variables

RANDOM VARIABLE A quantity resulting from an experiment that, by chance, can assume different values.

Possible *outcomes* for three coin tosses



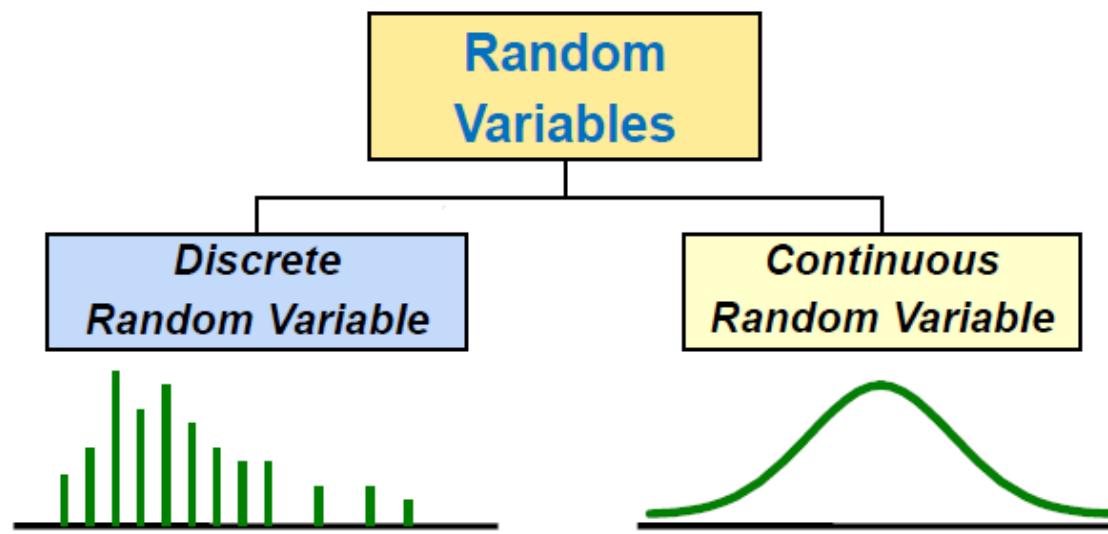
The *event* {one head} occurs and the *random variable* $x = 1$.

Types of Random Variables

DISCRETE RANDOM VARIABLE A random variable that can assume only certain clearly separated values. It is usually the result of counting something.

CONTINUOUS RANDOM VARIABLE A random variable that can assume an infinite number of values within a given range. It is usually the result of some type of measurement.

Types of Random Variables



Discrete Random Variable

DISCRETE RANDOM VARIABLE A random variable that can assume only certain clearly separated values. It is usually the result of counting something.

EXAMPLES:

- The number of students in a class
- The number of children in a family
- The number of cars entering a carwash in a hour
- The number of home mortgages approved by Coastal Federal Bank last week

Continuous Random Variable

CONTINUOUS RANDOM VARIABLE A random variable that can assume an infinite number of values within a given range. It is usually the result of some type of measurement

EXAMPLES:

- The length of each song on the latest Tim McGraw CD
- The weight of each student in this class
- The amount of money earned by each player in the National Football League

The Mean of a Discrete Probability Distribution

- The mean is a typical value used to represent the central location of a probability distribution.
- The mean of a probability distribution is also referred to as its **expected value**.

MEAN OF A PROBABILITY DISTRIBUTION

$$\mu = \Sigma [xP(x)]$$

[6-1]

The Mean of a Discrete Probability Distribution - Example



John has developed the following probability distribution for the number of cars he expects to sell on a particular Saturday.

| Number of Cars Sold, x | Probability, $P(x)$ |
|--------------------------|---------------------|
| 0 | .1 |
| 1 | .2 |
| 2 | .3 |
| 3 | .3 |
| 4 | .1 |
| Total | 1.0 |

1. What type of distribution is this?
2. On a typical Saturday, how many cars does John expect to sell?

The Mean of a Discrete Probability Distribution - Example

$$\begin{aligned}\mu &= \Sigma [xP(x)] \\ &= 0(.1) + 1(.2) + 2(.3) + 3(.3) + 4(.1) \\ &= 2.1\end{aligned}$$

These calculations are summarized in the following table.

| Number of Cars Sold, <i>x</i> | Probability, <i>P(x)</i> | <i>x</i> • <i>P(x)</i> |
|----------------------------------|-----------------------------|------------------------|
| 0 | .1 | 0.0 |
| 1 | .2 | 0.2 |
| 2 | .3 | 0.6 |
| 3 | .3 | 0.9 |
| 4 | .1 | 0.4 |
| Total | 1.0 | $\mu = \overline{2.1}$ |

The Variance and Standard Deviation of a Discrete Probability Distribution

Measures the amount of spread in a distribution.

VARIANCE OF A PROBABILITY DISTRIBUTION $\sigma^2 = \Sigma [(x - \mu)^2 P(x)]$ [6-2]

The computational steps are:

1. Subtract the mean from each value, and square this difference.
2. Multiply each squared difference by its probability.
3. Sum the resulting products to arrive at the variance.

The Variance and Standard Deviation of a Discrete Probability Distribution - Example

$$\sigma^2 = \sum[(x - \mu)^2 P(x)]$$

| Number of Cars Sold, <i>x</i> | Probability, <i>P(x)</i> | $(x - \mu)$ | $(x - \mu)^2$ | $(x - \mu)^2 P(x)$ |
|----------------------------------|-----------------------------|-------------|---------------|--------------------|
| 0 | .10 | 0 - 2.1 | 4.41 | 0.441 |
| 1 | .20 | 1 - 2.1 | 1.21 | 0.242 |
| 2 | .30 | 2 - 2.1 | 0.01 | 0.003 |
| 3 | .30 | 3 - 2.1 | 0.81 | 0.243 |
| 4 | .10 | 4 - 2.1 | 3.61 | 0.361 |
| | | | | $\sigma^2 = 1.290$ |

$$\sigma = \sqrt{\sigma^2} = \sqrt{1.290} = 1.136$$

Binomial Probability Distribution

- A widely occurring discrete probability distribution

Characteristics of a Binomial Probability Experiment

- The outcome of each trial is classified into one of two mutually exclusive categories—a **success** or a **failure**.
- The random variable, x , is the number of successes in a **fixed number of trials**.
- The **probability of success and failure stay the same** for each trial.
- The **trials are independent**, meaning that the outcome of one trial does not affect the outcome of any other trial.

Binomial Probability Formula

BINOMIAL PROBABILITY FORMULA

$$P(x) = {}_n C_x \pi^x (1 - \pi)^{n-x}$$

[6-3]

where:

C denotes a combination.

n is the number of trials.

x is the random variable defined as the number of successes.

π is the probability of a success on each trial.

Binomial Probability - Example

There are **five** flights daily from Pittsburgh via US Airways into the Bradford Regional Airport. Suppose the probability that any flight arrives late is **0.20**.

What is the probability that **none** of the flights are late today?

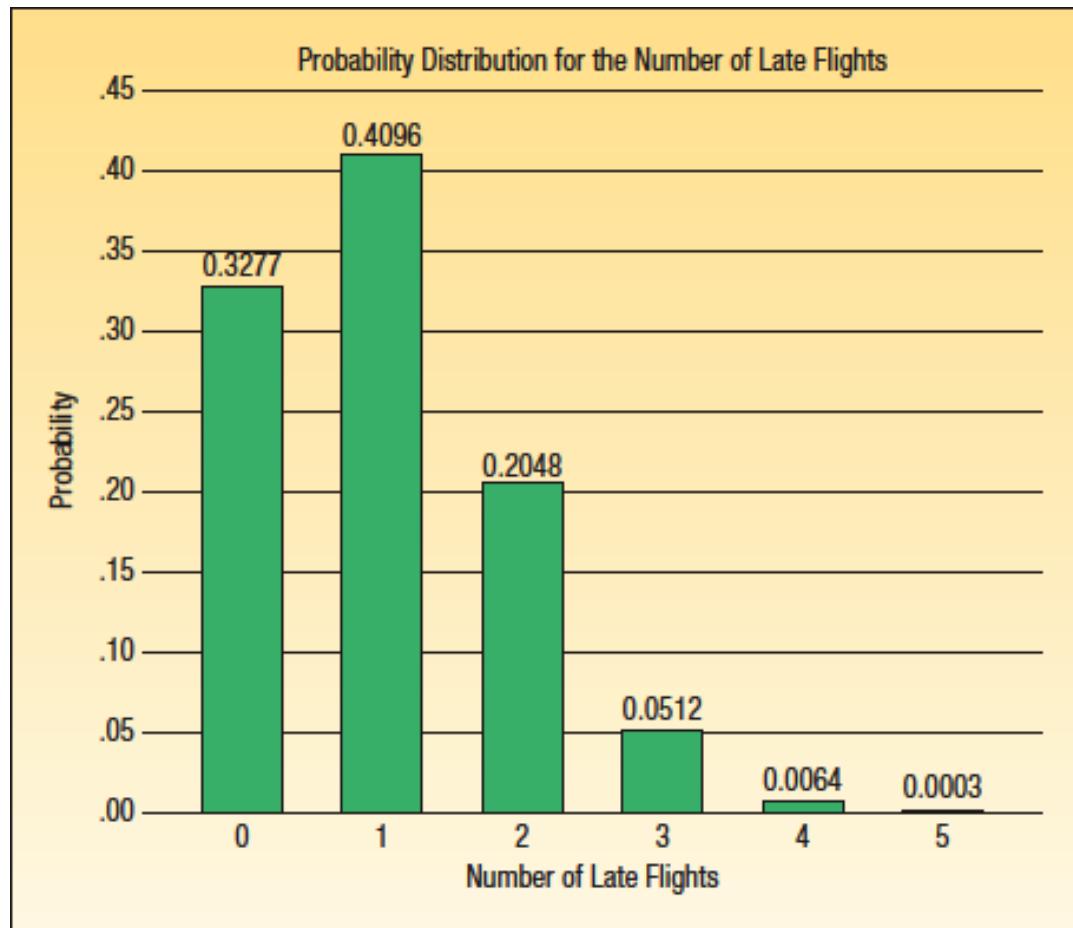
$$\begin{aligned}P(x=0) &= {}_nC_x \pi^x (1-\pi)^{n-x} \\&= {}_5C_0 (.20)^0 (1-.20)^{5-0} \\&= (1)(1)(.3277) \\&= 0.3277\end{aligned}$$

$${}_nC_r = \frac{n!}{r!(n-r)!}$$

Recall: $0! = 1$, and, any variable with a 0 exponent is equal to one.

Binomial Distribution Probability

The probabilities for each value of the random variable, number of late flights (0 through 5), can be calculated to create the entire binomial probability distribution.



Mean and Variance of a Binomial Distribution

Knowing the number of trials, n , and the probability of a success, π , for a binomial distribution, we can compute the mean and variance of the distribution.

MEAN OF A BINOMIAL DISTRIBUTION

$$\mu = n\pi$$

[6-4]

VARIANCE OF A BINOMIAL DISTRIBUTION

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

[6-5]

Mean and Variance of a Binomial Distribution - Example

For the example regarding the number of late flights, recall that $\pi = .20$ and $n = 5$.

What is the average number of late flights?

What is the variance of the number of late flights?

$$\begin{aligned}\mu &= n\pi \\ &= (5)(0.20) = 1.0\end{aligned}$$

$$\begin{aligned}\sigma^2 &= n\pi(1 - \pi) \\ &= (5)(0.20)(1 - 0.20) \\ &= (5)(0.20)(0.80) \\ &= 0.80\end{aligned}$$

Mean and Variance of a Binomial Distribution – Example

Using the general formulas for discrete probability distributions:

| Number of Late Flights, | | x | $P(x)$ | $xP(x)$ | $x - \mu$ | $(x - \mu)^2$ | $(x - \mu)^2 P(x)$ |
|-------------------------|--------|-----|--------|----------------|-----------|---------------|---------------------|
| 0 | 0.3277 | | 0.0000 | | -1 | 1 | 0.3277 |
| 1 | 0.4096 | | 0.4096 | | 0 | 0 | 0 |
| 2 | 0.2048 | | 0.4096 | | 1 | 1 | 0.2048 |
| 3 | 0.0512 | | 0.1536 | | 2 | 4 | 0.2048 |
| 4 | 0.0064 | | 0.0256 | | 3 | 9 | 0.0576 |
| 5 | 0.0003 | | 0.0015 | | 4 | 16 | 0.0048 |
| | | | | $\mu = 1.0000$ | | | $\sigma^2 = 0.7997$ |

$$\mu = \sum [xP(x)]$$

$$\sigma^2 = \sum [(x - \mu)^2 P(x)]$$

Binomial Probability Distributions Tables

Binomial probability distributions can be listed in tables. The calculations have already been done. In the table below, the binomial distributions for $n=6$ trials, and the different values of the probability of success are listed.

TABLE 6-2 Binomial Probabilities for $n = 6$ and Selected Values of π

| | | $n = 6$ Probability | | | | | | | | | |
|-------------------|------|------------------------|------|------|------|------|------|------|------|------|------|
| $x \setminus \pi$ | .05 | .1 | .2 | .3 | .4 | .5 | .6 | .7 | .8 | .9 | .95 |
| 0 | .735 | .531 | .262 | .118 | .047 | .016 | .004 | .001 | .000 | .000 | .000 |
| 1 | .232 | .354 | .393 | .303 | .187 | .094 | .037 | .010 | .002 | .000 | .000 |
| 2 | .031 | .098 | .246 | .324 | .311 | .234 | .138 | .060 | .015 | .001 | .000 |
| 3 | .002 | .015 | .082 | .185 | .276 | .313 | .276 | .185 | .082 | .015 | .002 |
| 4 | .000 | .001 | .015 | .060 | .138 | .234 | .311 | .324 | .246 | .098 | .031 |
| 5 | .000 | .000 | .002 | .010 | .037 | .094 | .187 | .303 | .393 | .354 | .232 |
| 6 | .000 | .000 | .000 | .001 | .004 | .016 | .047 | .118 | .262 | .531 | .735 |

Binomial Probability Distribution Tables – Example

Five percent of the worm gears produced by an automatic, high-speed Carter-Bell milling machine are defective.

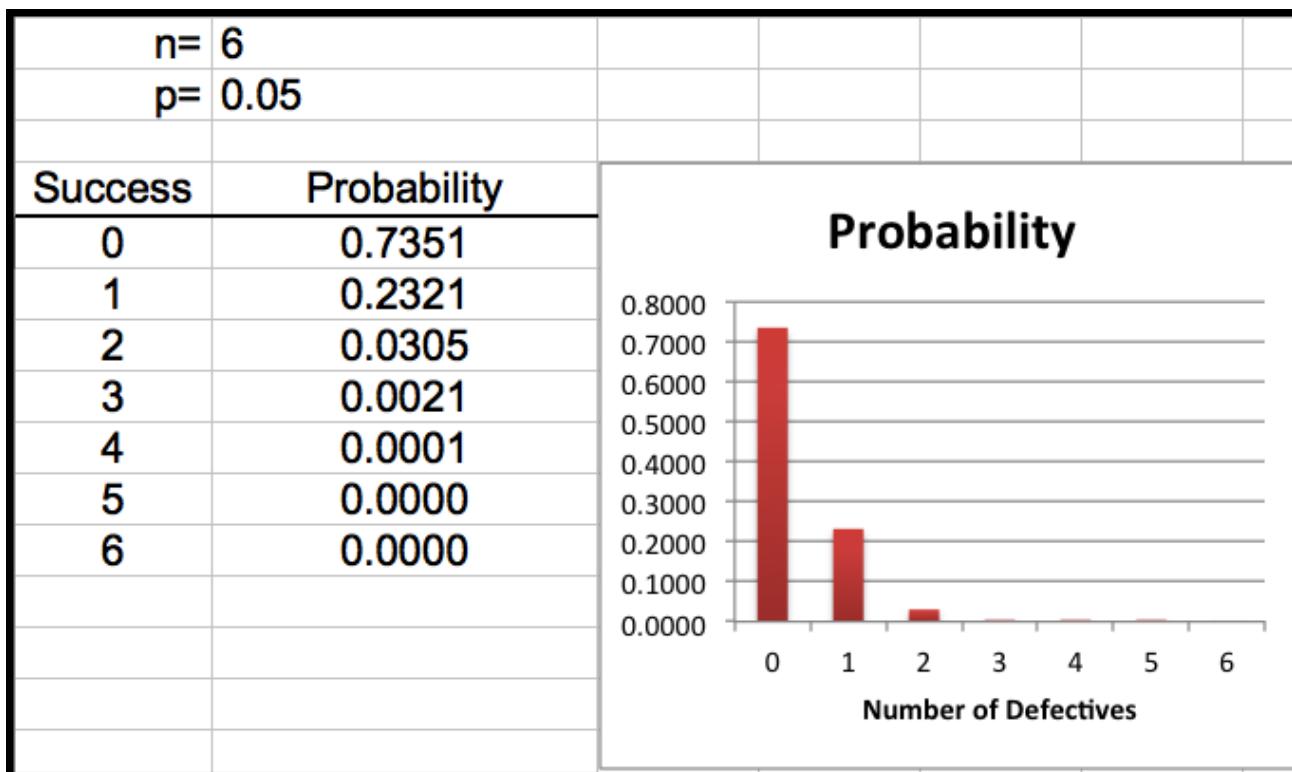
What is the probability that out of six gears selected at random none will be defective? Exactly one? Exactly two? Exactly three? Exactly four? Exactly five? Exactly six out of six?

TABLE 6–2 Binomial Probabilities for $n = 6$ and Selected Values of π

| | | $n = 6$ Probability | | | | | | | | | |
|-------------------|------|------------------------|------|------|------|------|------|------|------|------|------|
| $x \setminus \pi$ | .05 | .1 | .2 | .3 | .4 | .5 | .6 | .7 | .8 | .9 | .95 |
| 0 | .735 | .531 | .262 | .118 | .047 | .016 | .004 | .001 | .000 | .000 | .000 |
| 1 | .232 | .354 | .393 | .303 | .187 | .094 | .037 | .010 | .002 | .000 | .000 |
| 2 | .031 | .098 | .246 | .324 | .311 | .234 | .138 | .060 | .015 | .001 | .000 |
| 3 | .002 | .015 | .082 | .185 | .276 | .313 | .276 | .185 | .082 | .015 | .002 |
| 4 | .000 | .001 | .015 | .060 | .138 | .234 | .311 | .324 | .246 | .098 | .031 |
| 5 | .000 | .000 | .002 | .010 | .037 | .094 | .187 | .303 | .393 | .354 | .232 |
| 6 | .000 | .000 | .000 | .001 | .004 | .016 | .047 | .118 | .262 | .531 | .735 |

Binomial Probability Distribution- Excel Example

Using the Excel Function: Binom.dist(x,n,π,false)

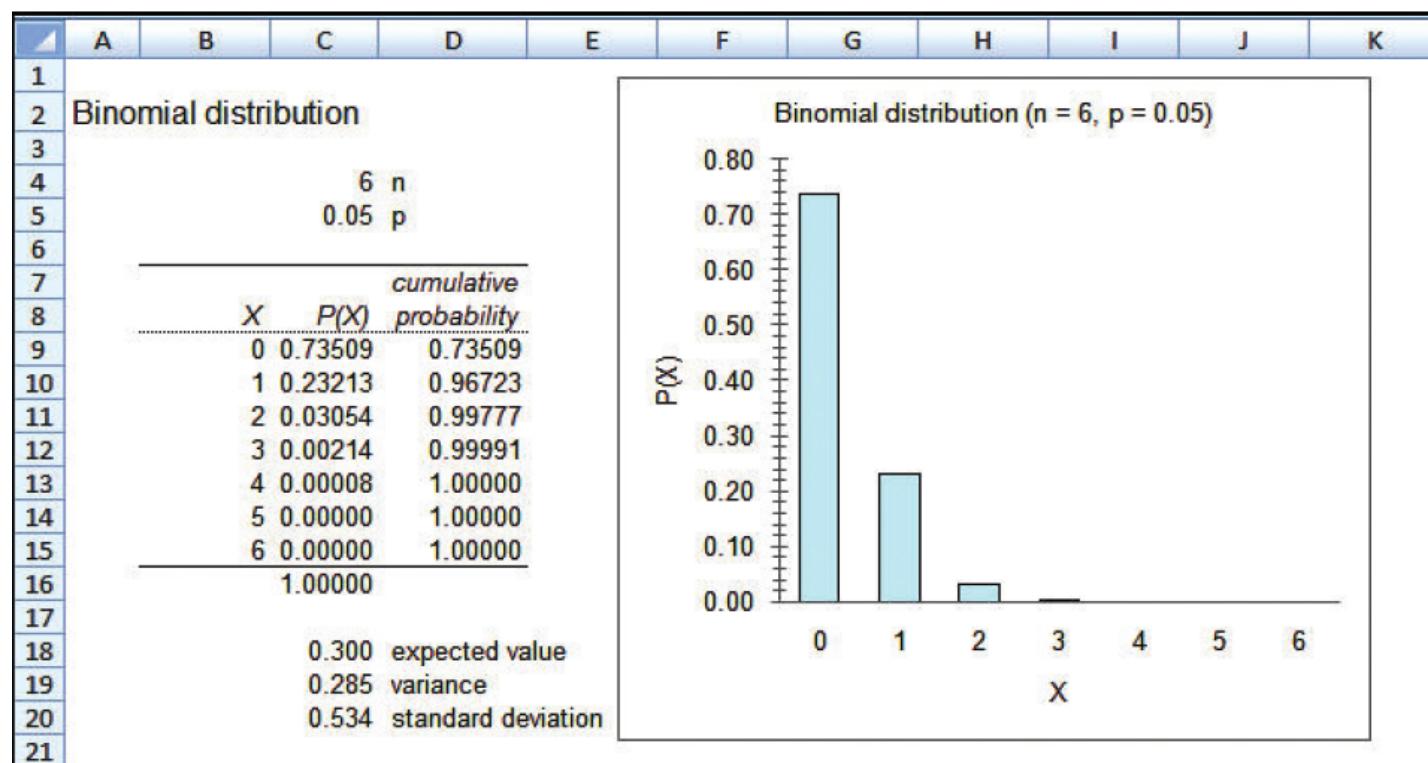


Binomial Probability Distribution – MegaStat Example

Statistical software, such as Megastat, can also create the values and graph for any binomial probability distribution.

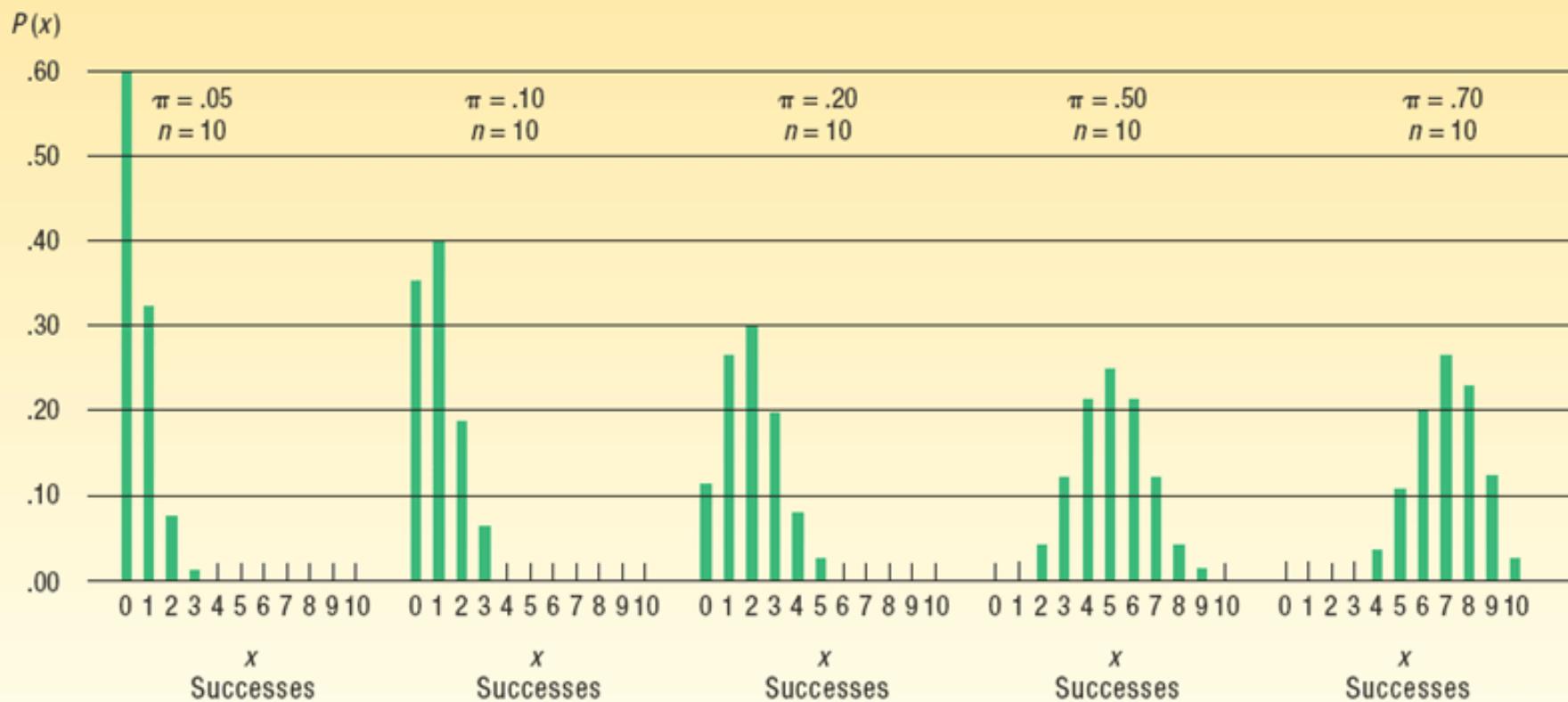
Five percent of the worm gears produced by an automatic, high-speed Carter-Bell milling machine are defective.

What is the binomial probability distribution of the number defective when six gears are selected?



Binomial – Shapes or Skewness for Varying π and $n=10$

The shape of a binomial distribution changes as n and π change.



Binomial Probability Distributions – Excel Example

A study by the Illinois Department of Transportation showed that 76.2 percent of front seat occupants used seat belts. If a sample of 12 cars traveling on a highway are selected, the binomial probability distribution of cars with front seat occupants using seat belts can be calculated as shown.

| =BINOM.DIST(A9,\$B\$2,\$B\$3,FALSE) | | |
|-------------------------------------|----------|-------------|
| A | B | C |
| 1 | | |
| 2 | n= 12 | |
| 3 | p= 0.762 | |
| 5 | Success | Probability |
| 6 | 0 | 0.0000 |
| 7 | 1 | 0.0000 |
| 8 | 2 | 0.0000 |
| 9 | 3 | 0.0002 |
| 10 | 4 | 0.0017 |
| 11 | 5 | 0.0088 |
| 12 | 6 | 0.0329 |
| 13 | 7 | 0.0902 |
| 14 | 8 | 0.1805 |
| 15 | 9 | 0.2569 |
| 16 | 10 | 0.2467 |
| 17 | 11 | 0.1436 |
| 18 | 12 | 0.0383 |

Binomial Probability Distributions – Excel Example

What is the probability the front seat occupants in **exactly 7** of the 12 vehicles are wearing seat belts?

| Success | Probability |
|---------|-------------|
| 0 | 0.0000 |
| 1 | 0.0000 |
| 2 | 0.0000 |
| 3 | 0.0002 |
| 4 | 0.0017 |
| 5 | 0.0088 |
| 6 | 0.0329 |
| 7 | 0.0902 |
| 8 | 0.1805 |
| 9 | 0.2569 |
| 10 | 0.2467 |
| 11 | 0.1436 |
| 12 | 0.0383 |

$$\begin{aligned}P(x = 7 | n = 12 \text{ and } \pi = .762) \\&= {}_{12}C_7(.762)^7(1 - .762)^{12-7} \\&= 792(.149171)(.000764) = .0902\end{aligned}$$

Cumulative Binomial Probability Distributions – Excel Example

What is the probability the front seat occupants in **at least** 7 of the 12 vehicles are wearing seat belts?

| Success | Probability |
|---------|-------------|
| 0 | 0.0000 |
| 1 | 0.0000 |
| 2 | 0.0000 |
| 3 | 0.0002 |
| 4 | 0.0017 |
| 5 | 0.0088 |
| 6 | 0.0329 |
| 7 | 0.0902 |
| 8 | 0.1805 |
| 9 | 0.2569 |
| 10 | 0.2467 |
| 11 | 0.1436 |
| 12 | 0.0383 |
| | 0.9562 |

Sum of
Probabilities
for 7 or
more
successes

$$P(x \geq 7 | n = 12 \text{ and } \pi = .762)$$

$$= P(x = 7) + P(x = 8) + P(x = 9) + P(x = 10) + P(x = 11) + P(x = 12)$$

$$= .0902 + .1805 + .2569 + .2467 + .1436 + .0383$$

$$= .9562$$

Poisson Probability Distribution

The **Poisson probability distribution** describes the number of times some **event occurs** during a **specified interval**. The interval may be time, distance, area, or volume.

Assumptions of the Poisson Distribution:

- The **probability is proportional** to the length of the interval.
- The intervals are **independent**.

Poisson Probability Distribution

The Poisson probability distribution is characterized by the number of times an event happens during some interval or continuum.

Examples:

- The number of financial crisis per decade
- The number of calls per hour received by Dyson Vacuum Cleaner Company
- The number of vehicles sold per day at Hyatt Buick GMC in Durham, North Carolina
- The number of goals scored in a college soccer game

Poisson Probability Distribution

The **Poisson** distribution can be described mathematically by the formula:

POISSON DISTRIBUTION

$$P(x) = \frac{\mu^x e^{-\mu}}{x!} \quad [6-7]$$

where:

μ (mu) is the mean number of occurrences (successes) in a particular interval.

e is the constant 2.71828 (base of the Napierian logarithmic system).

x is the number of occurrences (successes).

$P(x)$ is the probability for a specified value of x.

Poisson Probability Distribution

- The mean number of successes, μ , can be determined in Poisson situations by $n\pi$, where n is the number of trials and π the probability of a success.

MEAN OF A POISSON DISTRIBUTION

$$\mu = n\pi$$

[6-8]

- The variance of the Poisson distribution is also equal to $n \pi$.

Poisson Probability Distribution – Example

Assume baggage is rarely lost by Northwest Airlines. Suppose a random sample of 1,000 flights shows a total of 300 bags were lost. Thus, the arithmetic mean number of lost bags per flight is 0.3 ($300/1,000$). If the number of lost bags per flight follows a Poisson distribution with $\mu = 0.3$, find the probability of not losing any bags.

$$P(0) = \frac{\mu^x e^{-\mu}}{x!} = \frac{0.3^0 e^{-0.3}}{0!} = .7408$$

Poisson Probability Distribution Table – Example

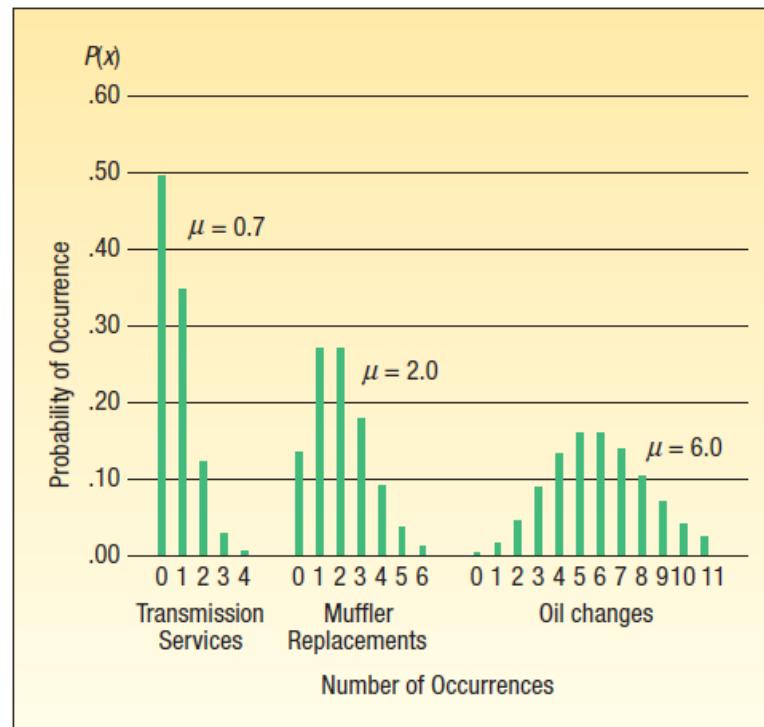
Recall from the previous illustration that the number of lost bags follows a Poisson distribution with a mean of 0.3. A table can be used to find the probability that no bags will be lost on a particular flight. What is the probability **no bag** will be lost on a particular flight?

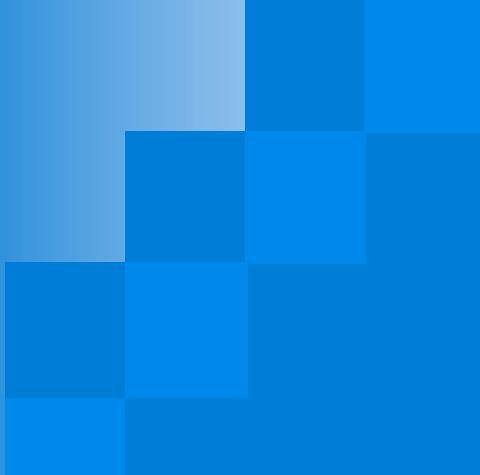
TABLE 6-6 Poisson Table for Various Values of μ (from Appendix B.2)

| x | μ | | | | | | | | | |
|-----|--------|--------|--------|--------|--------|--------|--------|--------|--------|--|
| | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | |
| 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 | 0.3033 | 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 | |

More About the Poisson Probability Distribution

- The Poisson probability distribution is always positively skewed and the random variable has no specific upper limit.
- The Poisson distribution for the lost bags illustration, where $\mu=0.3$, is highly skewed.
- As μ becomes larger, the Poisson distribution becomes more symmetrical.





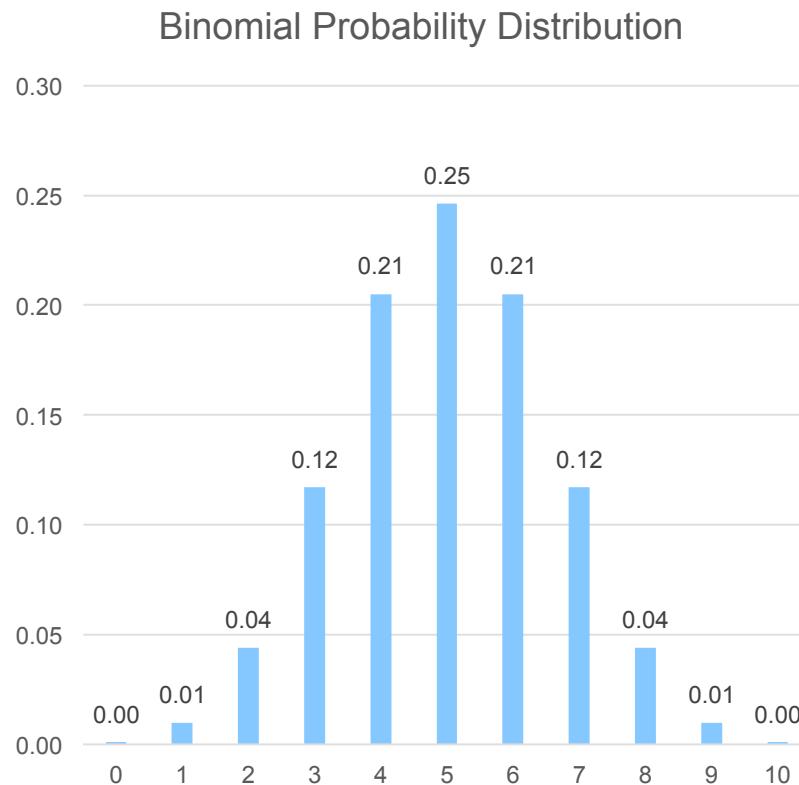
Continuous Probability Distributions



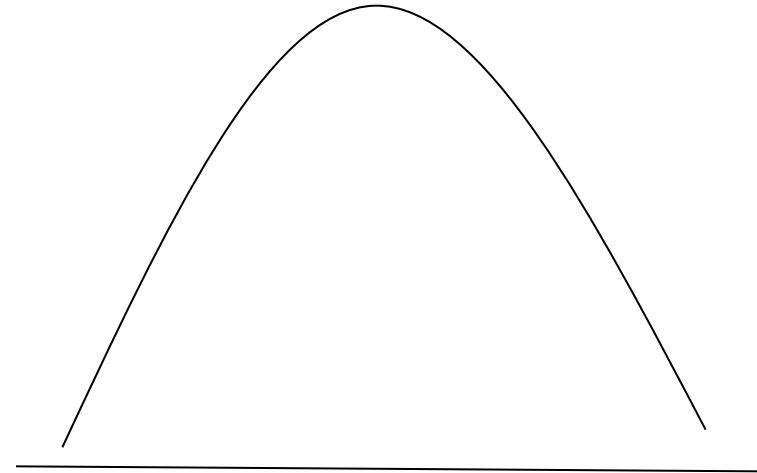
Chapter 7

Continuous Probability Distributions

Discrete Prob. Dist.



Continuous Prob. Dist.



The Uniform Distribution

The uniform probability distribution is perhaps the **simplest distribution for a continuous random variable**.

This distribution is **rectangular in shape** and is defined by minimum (a) and maximum (b) values.

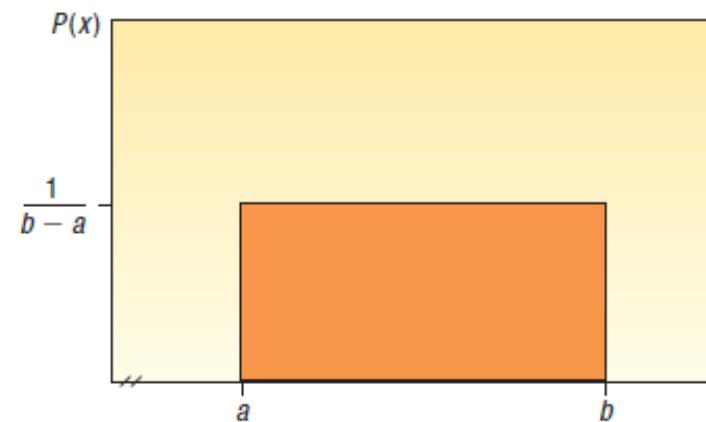


CHART 7-1 A Continuous Uniform Distribution

The Uniform Distribution – Mean and Standard Deviation

Knowing the minimum and maximum values of a uniform distribution, we can define the probability function, and calculate the mean, variance, and standard deviation of the distribution.

UNIFORM DISTRIBUTION $P(x) = \frac{1}{b-a}$ if $a \leq x \leq b$ and 0 elsewhere [7-3]

MEAN OF THE UNIFORM DISTRIBUTION $\mu = \frac{a+b}{2}$ [7-1]

STANDARD DEVIATION OF THE UNIFORM DISTRIBUTION $\sigma = \sqrt{\frac{(b-a)^2}{12}}$ [7-2]

The Uniform Distribution – Example

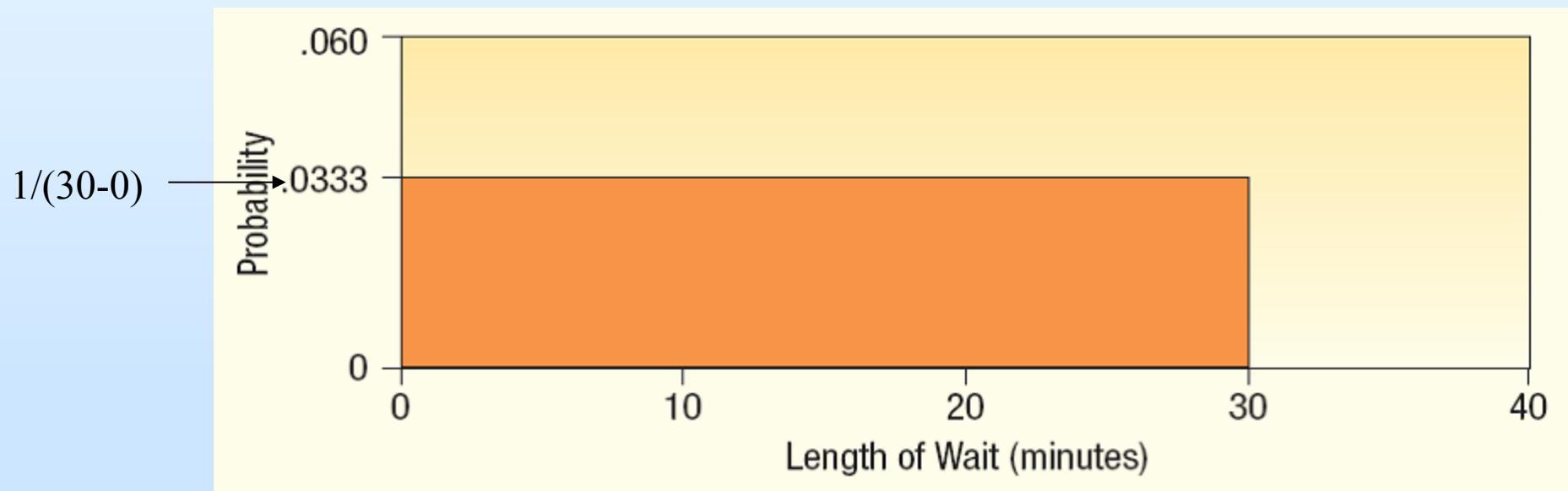
Southwest Arizona State University provides bus service to students. On weekdays, a bus arrives at the North Main Street and College Drive stop every 30 minutes between 6 a.m. and 11 p.m.

Students arrive at the bus stop at random times. The time that a student waits is uniformly distributed from 0 to 30 minutes.

1. Draw a graph of this distribution.
2. Show that the area of this uniform distribution is 1.00.
3. How long will a student “typically” have to wait for a bus? In other words what is the mean waiting time?
4. What is the probability a student will wait more than 25 minutes?
5. What is the probability a student will wait between 10 and 20 minutes?

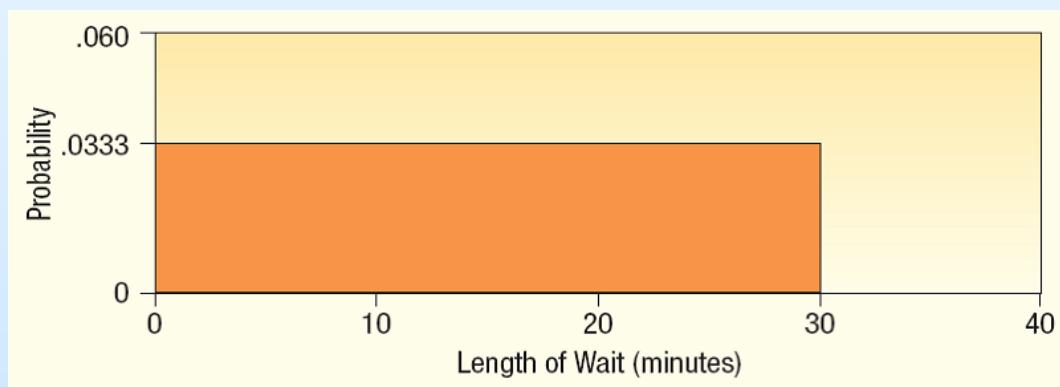
The Uniform Distribution - Example

1. Graph of uniformly distributed waiting times between 0 and 30:



The Uniform Distribution – Example

2. Show that the area of this distribution is 1.00.

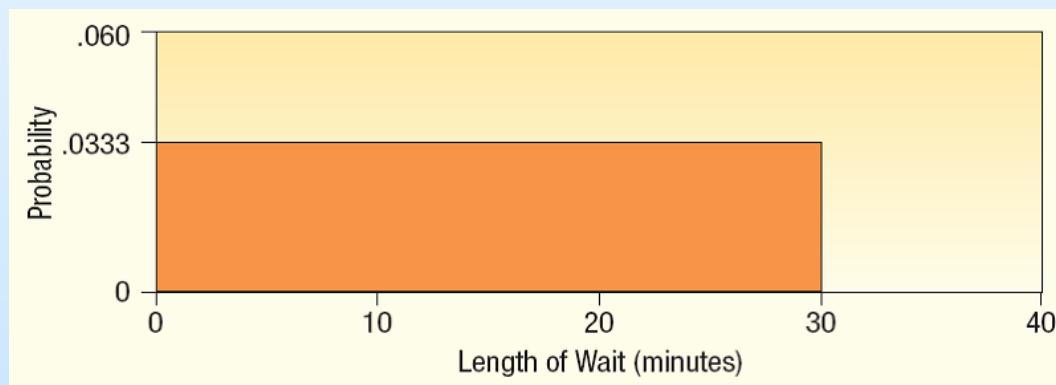


The times students must wait for the bus is uniform over the interval from 0 minutes to 30 minutes, so in this case a is 0 and b is 30.

$$\text{Area} = (\text{height})(\text{base}) = \frac{1}{(30 - 0)} (30 - 0) = 1.00$$

The Uniform Distribution – Example

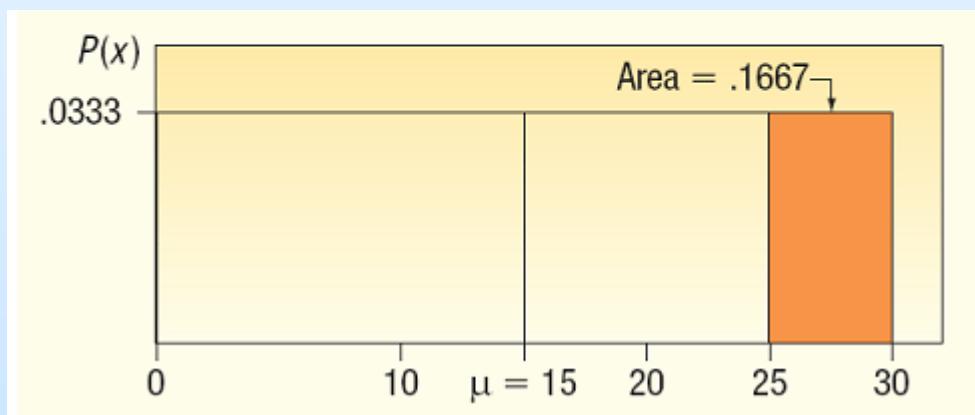
3. How long will a student “typically” have to wait for a bus? In other words what is the **mean waiting time**?



$$\mu = \frac{a + b}{2} = \frac{0 + 30}{2} = 15$$

The Uniform Distribution – Example

4. What is the probability a student will wait **more than 25 minutes?**

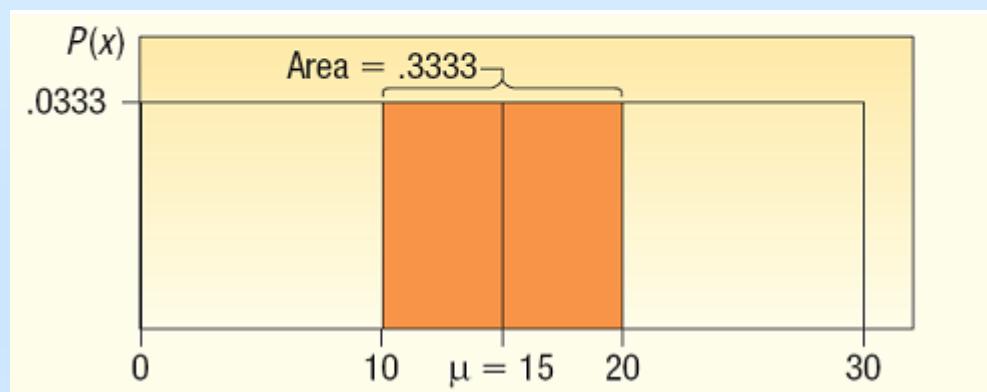


$$P(25 < \text{wait time} < 30) = (\text{height})(\text{base}) = \frac{1}{(30 - 0)} (5) = .1667$$

The Uniform Distribution – Example

5. What is the probability a student will wait **between 10 and 20 minutes?**

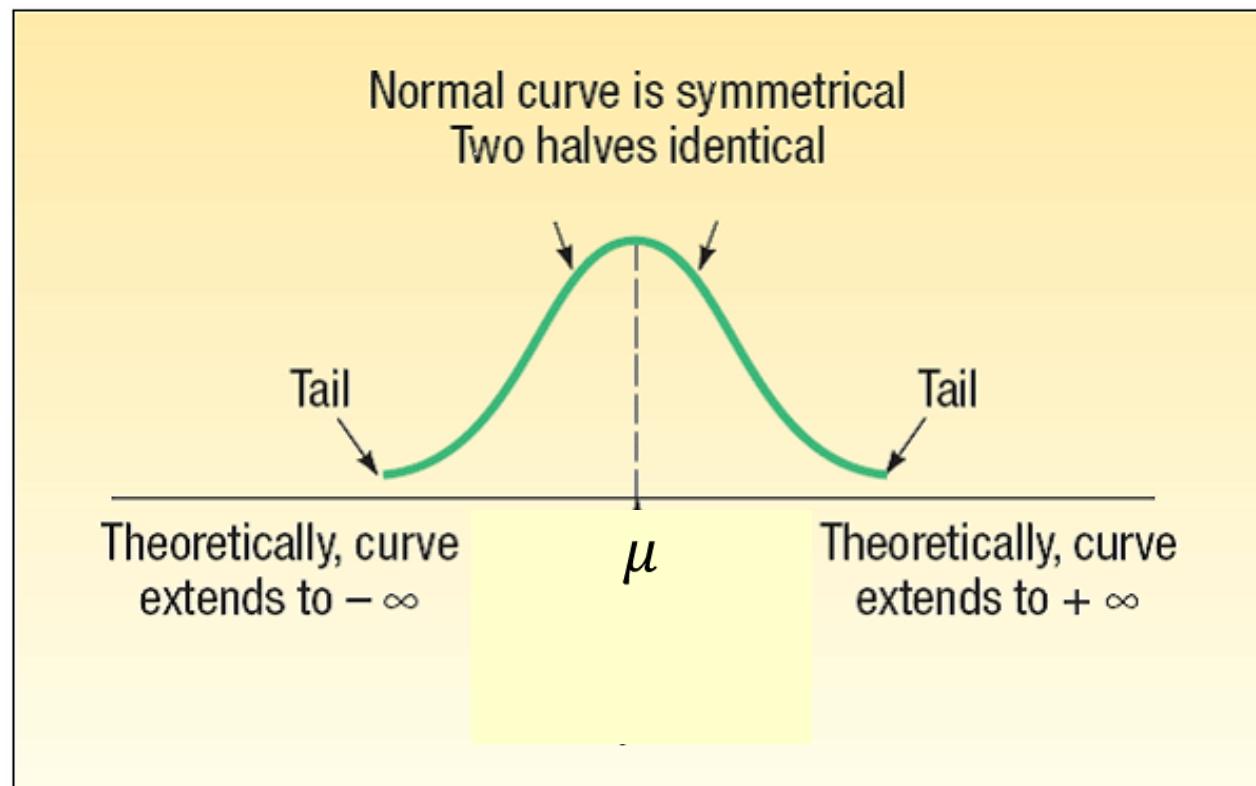
$$P(10 < \text{wait time} < 20) = (\text{height})(\text{base}) = \frac{1}{(30 - 0)} (10) = .3333$$



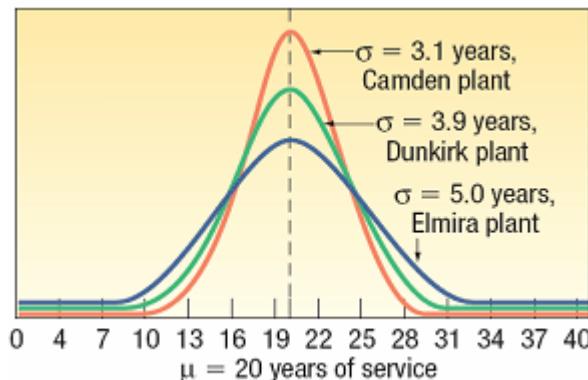
Characteristics of a Normal Probability Distribution

- It is **bell-shaped** and has a single peak at the center of the distribution.
- It is **symmetrical** about the mean.
- It is **asymptotic**: The curve gets closer and closer to the X-axis but never actually touches it. To put it another way, the tails of the curve extend indefinitely in both directions.
- The location of a normal distribution is determined by the mean, μ . The dispersion or spread of the distribution is determined by the standard deviation, σ .
- The arithmetic **mean, median, and mode are equal**.
- As a probability distribution, the total **area under the curve is defined to be 1.00**.
- Because the distribution is symmetrical about the mean, half the area under the normal curve is to the right of the mean, and the other half to the left of it.

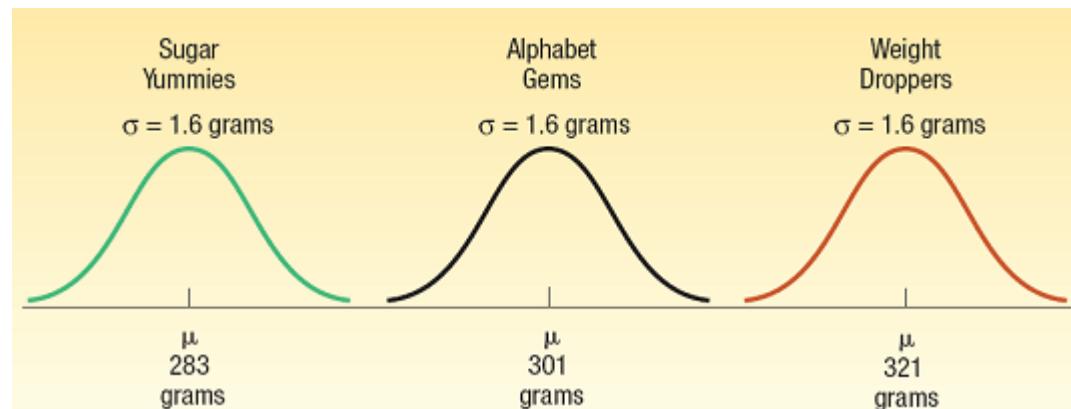
The Normal Distribution – Graphically



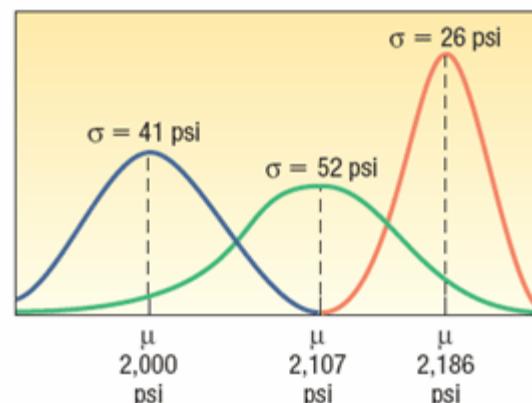
The Family of Normal Distributions



Equal Means and Different Standard Deviations



Different Means and Equal Standard Deviations



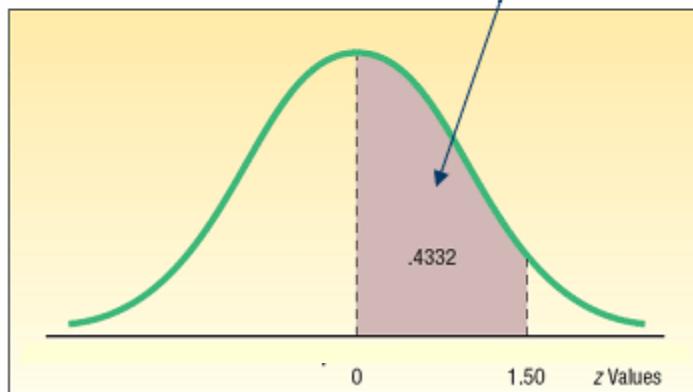
Different Means and Standard Deviations

The Standard Normal Probability Distribution

- The standard normal distribution is a normal distribution with a **mean of 0** and a **standard deviation of 1**.
- It is also called the **z distribution**.

Areas Under the Normal Curve Using a Standard Normal Table

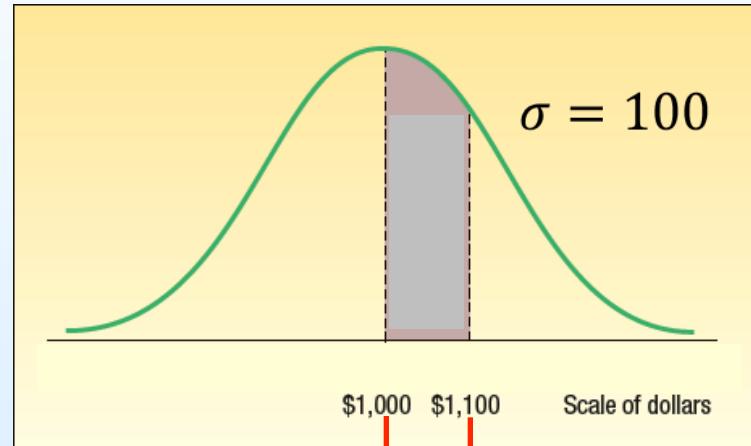
| <i>z</i> | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | ... |
|-----------------|-------------|-------------|-------------|-------------|-------------|-------------|------------|
| 1.3 | 0.4032 | 0.4049 | 0.4066 | 0.4082 | 0.4099 | 0.4115 | |
| 1.4 | 0.4192 | 0.4207 | 0.4222 | 0.4236 | 0.4251 | 0.4265 | |
| 1.5 | 0.4332 | 0.4345 | 0.4357 | 0.4370 | 0.4382 | 0.4394 | |
| 1.6 | 0.4452 | 0.4463 | 0.4474 | 0.4484 | 0.4495 | 0.4505 | |
| 1.7 | 0.4554 | 0.4564 | 0.4573 | 0.4582 | 0.4591 | 0.4599 | |
| 1.8 | 0.4641 | 0.4649 | 0.4656 | 0.4664 | 0.4671 | 0.4678 | |
| 1.9 | 0.4713 | 0.4719 | 0.4726 | 0.4732 | 0.4738 | 0.4744 | |
| . | | | | | | | |
| . | | | | | | | |



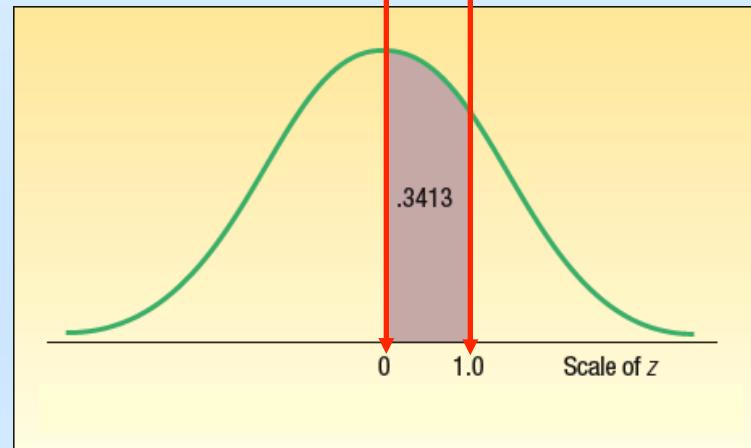
The Standard Normal Probability Distribution

- A **z-value** is the signed distance between a selected value, designated x , and the population mean, μ , divided by the population standard deviation, σ .
- The formula is:

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



Normal Distribution



Standard Normal Distribution

The Normal Distribution – Example

The weekly incomes of shift foremen in the glass industry follow the normal probability distribution with a **mean of \$1,000** and a **standard deviation of \$100**.

What is the **z value** for the income, let's call it x , of a foreman who earns **\$1,100** per week? For a foreman who earns **\$900** per week?

For $x = \$1,100$:

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{\$1,100 - \$1,000}{\$100} \\ &= 1.00 \end{aligned}$$

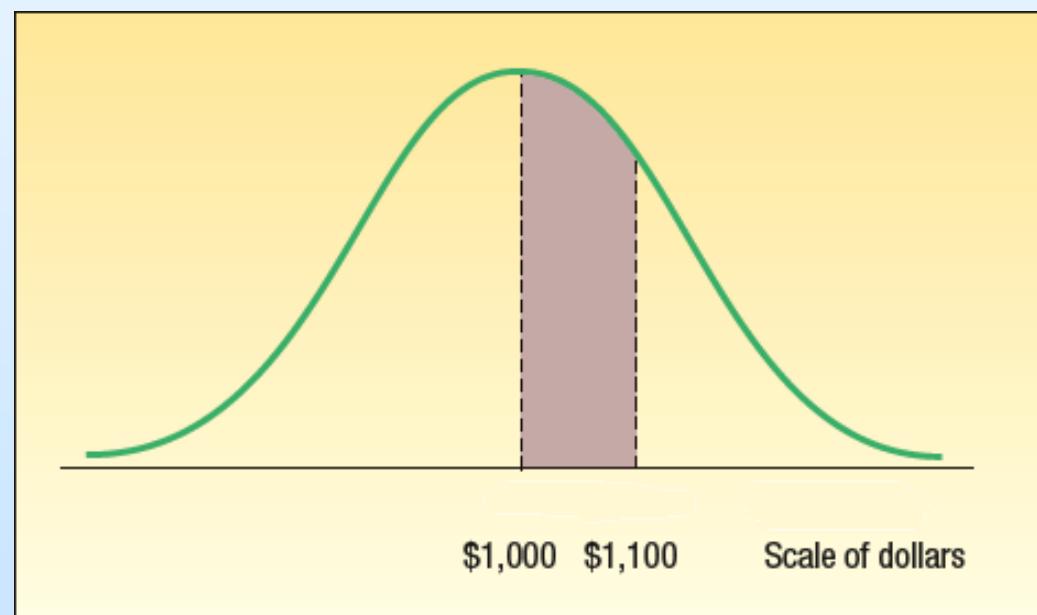
For $x = \$900$:

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{\$900 - \$1,000}{\$100} \\ &= -1.00 \end{aligned}$$

Normal Distribution – Finding Probabilities (Example 1)

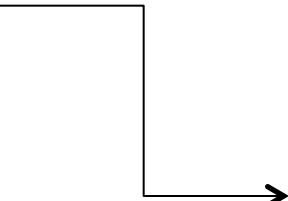
In an earlier example we reported that the mean weekly income of a shift foreman in the glass industry is normally distributed with a mean of \$1,000 and a standard deviation of \$100.

What is the likelihood of selecting a foreman whose weekly income is between \$1,000 and \$1,100?



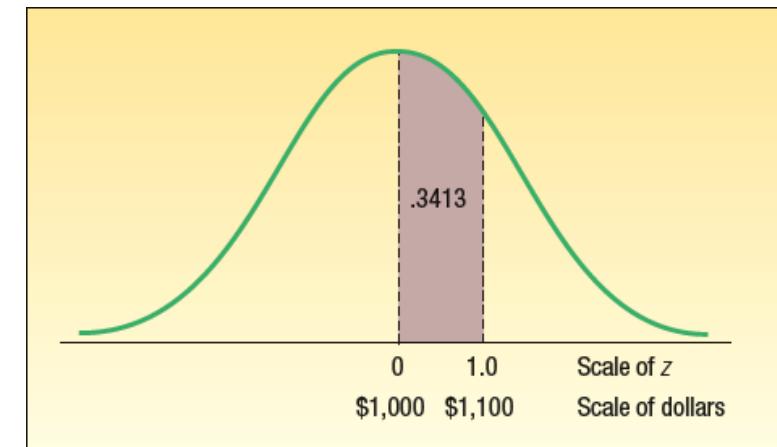
Normal Distribution – Finding Probabilities (Example 1)

$$z = \frac{x - \mu}{\sigma} = \frac{\$1,100 - \$1,000}{\$100} = 1.00$$



A z-table showing cumulative probabilities for standard normal distribution values. The table has columns for z-values (0.00, 0.01, 0.02) and rows for corresponding cumulative probabilities.

| <i>z</i> | 0.00 | 0.01 | 0.02 |
|----------|-------|-------|-------|
| : | : | : | : |
| 0.7 | .2580 | .2611 | .2642 |
| 0.8 | .2881 | .2910 | .2939 |
| 0.9 | .3159 | .3186 | .3212 |
| 1.0 | .3413 | .3438 | .3461 |
| 1.1 | .3643 | .3665 | .3686 |
| : | : | : | : |



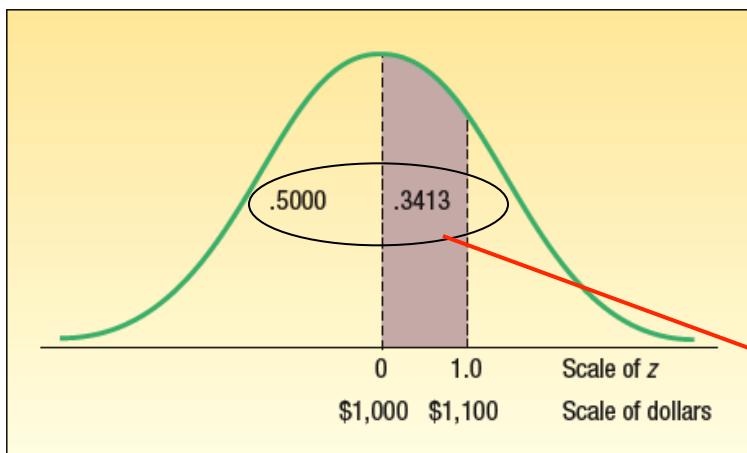
Finding Areas for z Using Excel

The Excel function:

=NORM.DIST(x,Mean,Standard_dev,Cumu)

=NORM.DIST(1100,1000,100,true)

calculates the probability (area) for $z=1$.



Screenshot of Microsoft Excel showing the Home tab selected. The formula bar displays the formula =NORM.DIST(1100,1000,100,TRUE). The cell A1 contains the same formula. The rest of the cells A2 through A5 are empty.

| A | B | C | D | E | F | G | H |
|---|--------------------------------|---|---|---|---|---|---|
| 1 | =NORM.DIST(1100,1000,100,TRUE) | | | | | | |
| 2 | | | | | | | |
| 3 | | | | | | | |
| 4 | | | | | | | |
| 5 | | | | | | | |

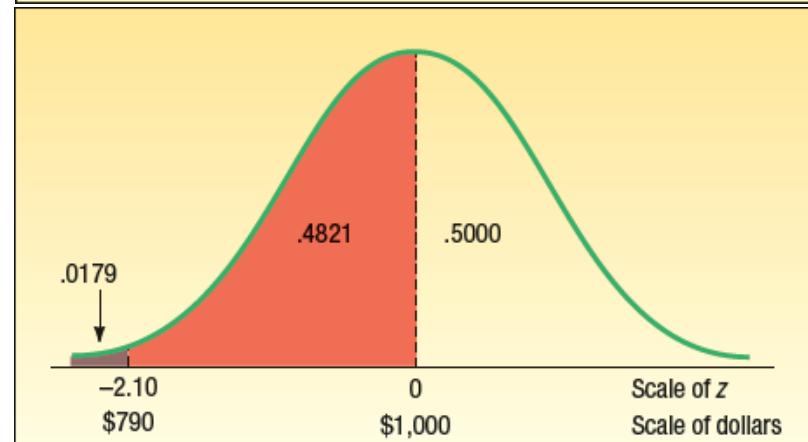
Normal Distribution – Finding Probabilities (Example 2)

Refer to the information regarding the weekly income of shift foremen in the glass industry. The distribution of weekly incomes follows the normal probability distribution with a mean of \$1,000 and a standard deviation of \$100.

What is the probability of selecting a shift foreman in the glass industry whose income is **between \$790 and \$1,000?**

$$z = \frac{x - \mu}{s} = \frac{\$790 - \$1,000}{\$100} = -2.10$$

| <i>z</i> | 0.00 | 0.01 | 0.02 |
|----------|--------------|-------|-------|
| : | : | : | : |
| 2.0 | .4772 | .4778 | .4783 |
| 2.1 | .4821 | .4826 | .4830 |
| 2.2 | .4861 | .4864 | .4868 |
| 2.3 | .4893 | .4896 | .4898 |
| : | : | : | : |

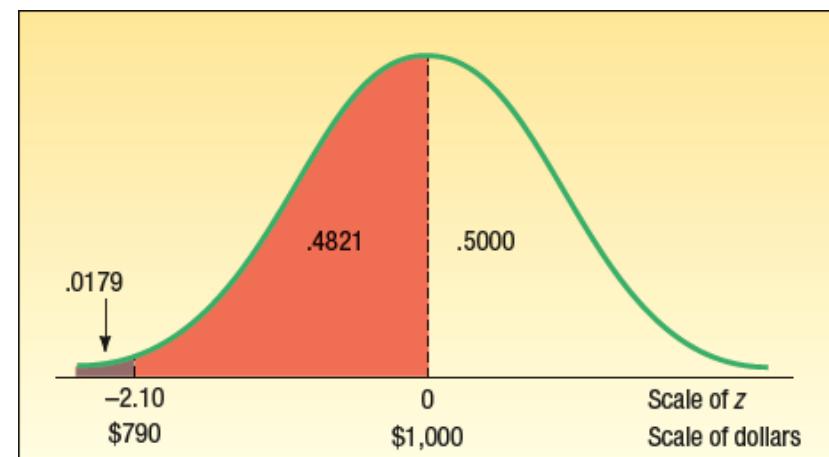


Normal Distribution – Finding Probabilities (Example 3)

Refer to the information regarding the weekly income of shift foremen in the glass industry. The distribution of weekly incomes follows the normal probability distribution with a mean of \$1,000 and a standard deviation of \$100.

What is the probability of selecting a shift foreman in the glass industry whose income is **less than \$790?**

$$z = \frac{x - \mu}{s} = \frac{\$790 - \$1,000}{\$100} = -2.10$$



The probability of selecting a shift foreman with income less than \$790 is $0.5 - .4821 = .0179$.

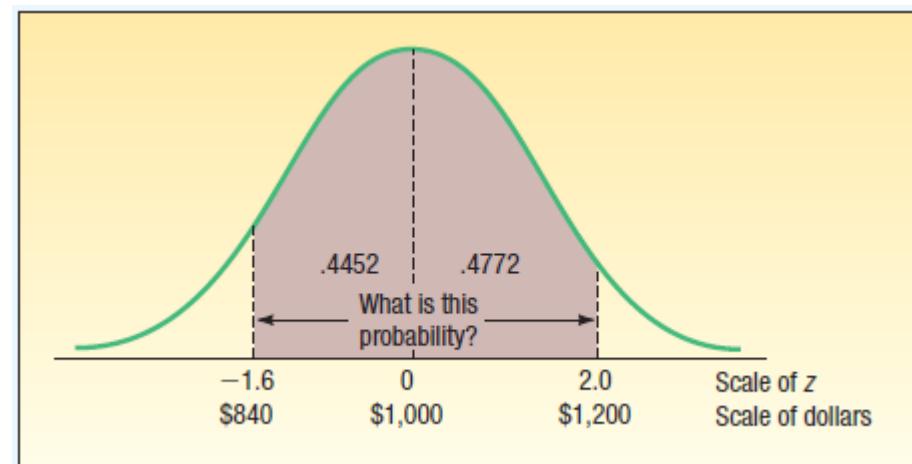
Normal Distribution – Finding Probabilities (Example 4)

Refer to the information regarding the weekly income of shift foremen in the glass industry. The distribution of weekly incomes follows the normal probability distribution with a mean of \$1,000 and a standard deviation of \$100.

What is the probability of selecting a shift foreman in the glass industry whose income is **between \$840 and \$1,200**?

$$z = \frac{\$840 - \$1,000}{\$100} = \frac{-\$160}{\$100} = -1.60$$

$$z = \frac{\$1,200 - \$1,000}{\$100} = \frac{\$200}{\$100} = 2.00$$



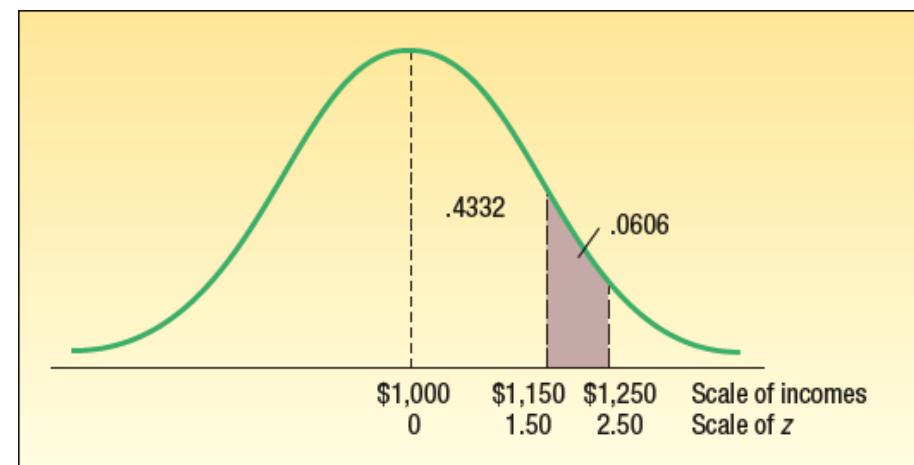
Normal Distribution – Finding Probabilities (Example 5)

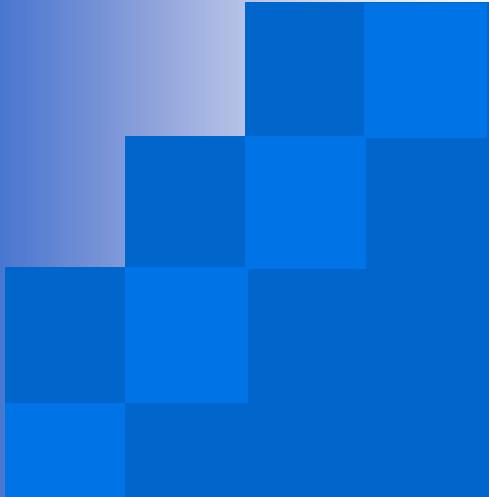
Refer to the information regarding the weekly income of shift foremen in the glass industry. The distribution of weekly incomes follows the normal probability distribution with a mean of \$1,000 and a standard deviation of \$100.

What is the probability of selecting a shift foreman in the glass industry whose income is **between \$1,150 and \$1,250**?

$$z = \frac{\$1,250 - \$1,000}{\$100} = 2.50$$

$$z = \frac{\$1,150 - \$1,000}{\$100} = 1.50$$





Sampling Methods and the Central Limit Theorem



Chapter 8

Why Sample a Population?

- Contacting the whole population
 - Would be **time consuming**.
 - Costly
 - May sometimes be **physically impossible**

Sampling Error

| | Sample | Population | Sampling Error |
|----------|-----------|------------|------------------|
| Mean | \bar{X} | μ | $\bar{X} - \mu$ |
| St. Dev | s | σ | $s - \sigma$ |
| Variance | s^2 | σ^2 | $s^2 - \sigma^2$ |



Developing a Sampling Distribution

- Assume there is a population with size $N=4$
- Random variable, X , is age of individuals



Amy=18, Valerie=20, Celine=22, Megan=24

Sampling

- We can form samples of size 1,2,3,4.
- Let's do it for $n=2$.



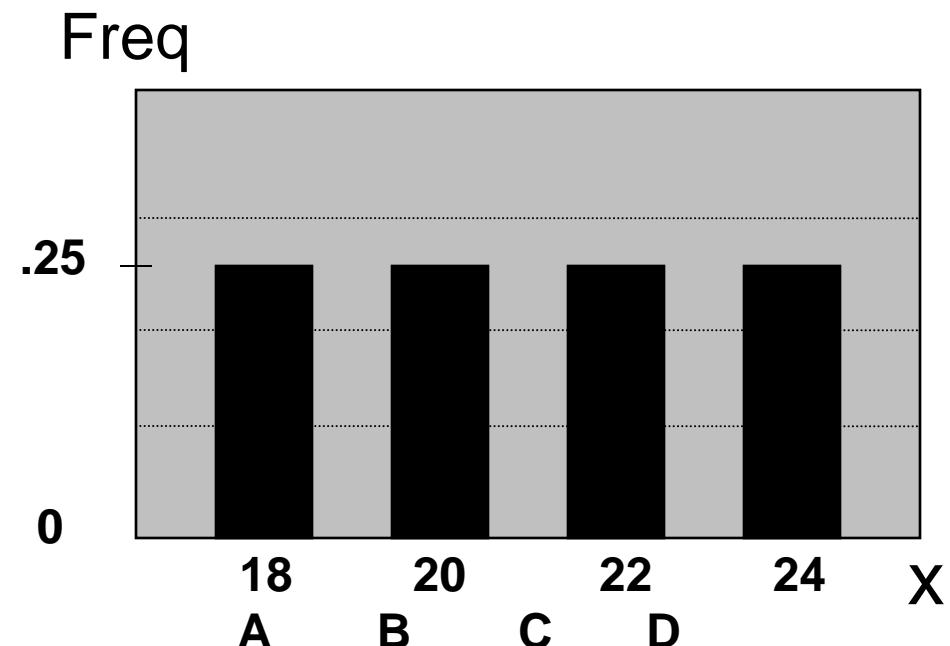
Developing a Sampling Distribution

(continued)

Summary Measures for the Population Distribution:

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

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



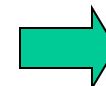
Uniform Distribution

Developing a Sampling Distribution

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

| 1 st | 2 nd Observation | | | |
|-----------------|-----------------------------|----------|---------|---------|
| Obs | Amy(18) | Val (20) | Cel(22) | Meg(24) |
| Amy | - | 18,20 | 18,22 | 18,24 |
| Val | - | - | 20,22 | 20,24 |
| Cel | - | - | - | 22,24 |
| Meg | - | - | - | - |

6 Sample Means



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

Developing a Sampling Distribution

(continued)

Sampling Distribution of All Sample Means

6 Sample Means

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



Sample Means
Distribution

| Means | Frequency |
|-------|------------|
| 19 | $1/6=0.16$ |
| 20 | $1/6=0.16$ |
| 21 | $2/6=0.32$ |
| 22 | $1/6=0.16$ |
| 23 | $1/6=0.16$ |

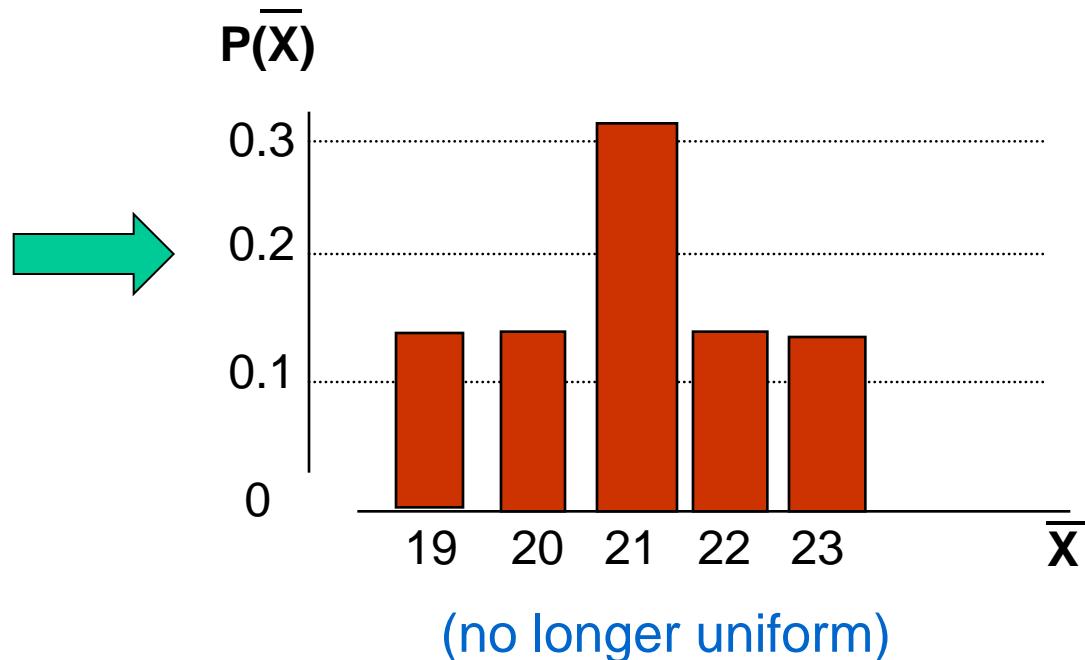
Developing a Sampling Distribution

Sampling Distribution of All Sample Means

6 Sample Means

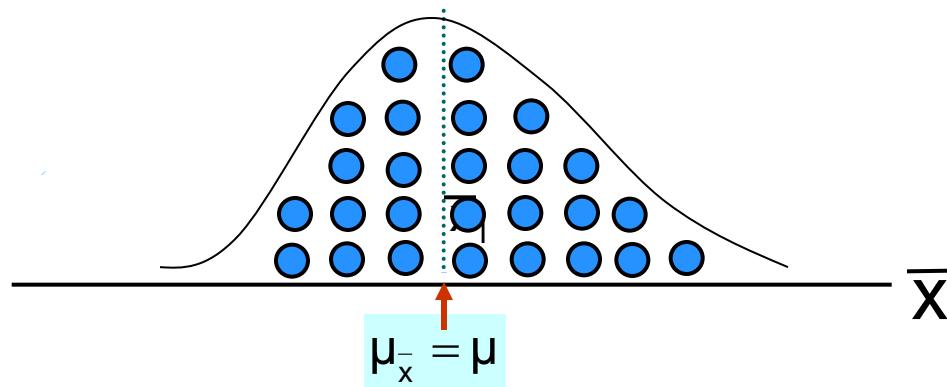
| 1st Obs | 2nd Observation | | | |
|------------|-----------------|----|----|----|
| | 18 | 20 | 22 | 24 |
| 18 | - | 19 | 20 | 21 |
| 20 | - | - | 21 | 22 |
| 22 | - | - | - | 23 |
| 24 | - | - | - | - |

Sample Means
Distribution



Sampling Distribution of the Sample Mean

The **sampling distribution of the sample mean** is a probability distribution consisting of all possible sample means of a given sample size selected from a population.



Sampling Distribution of the Sample Mean

- The Standard Error of the Mean:

STANDARD ERROR OF THE MEAN

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad [8-1]$$

The dispersion of the sampling distribution of the sample mean is narrower than the population distribution.(n)

Sampling Distribution of the Sample Mean – Example

Tartus Industries has seven production employees (considered the population). The hourly earnings of each employee are given in the table below.

| Employee | Hourly Earnings | Employee | Hourly Earnings |
|----------|-----------------|----------|-----------------|
| Joe | \$7 | Jan | \$7 |
| Sam | 7 | Art | 8 |
| Sue | 8 | Ted | 9 |
| Bob | 8 | | |

- What is the population mean?
- What is the population standard deviation?
- What is the sampling distribution of the sample mean for samples of size 2?
- What is the mean of the sampling distribution?
- What is the standard deviation of the sampling distribution?

Sampling Distribution of the Sample Mean – Example

1. The population mean is \$7.71, found by:

$$\mu = \frac{\Sigma X}{N} = \frac{\$7 + \$7 + \$8 + \$8 + \$7 + \$8 + \$9}{7} = \$7.71$$

The population standard deviation $s = \sqrt{\frac{(x - \bar{m})^2}{N}} = 0.70$

Sampling Distribution of the Sample Mean – Example

- What is the sampling distribution of the sample mean for samples of size 2?

| Sample | Employees | Hourly Earnings | Sum | Mean | Sample | Employees | Hourly Earnings | Sum | Mean |
|--------|-----------|-----------------|------|--------|--------|-----------|-----------------|------|--------|
| 1 | Joe, Sam | \$7, \$7 | \$14 | \$7.00 | 12 | Sue, Bob | \$8, \$8 | \$16 | \$8.00 |
| 2 | Joe, Sue | 7, 8 | 15 | 7.50 | 13 | Sue, Jan | 8, 7 | 15 | 7.50 |
| 3 | Joe, Bob | 7, 8 | 15 | 7.50 | 14 | Sue, Art | 8, 8 | 16 | 8.00 |
| 4 | Joe, Jan | 7, 7 | 14 | 7.00 | 15 | Sue, Ted | 8, 9 | 17 | 8.50 |
| 5 | Joe, Art | 7, 8 | 15 | 7.50 | 16 | Bob, Jan | 8, 7 | 15 | 7.50 |
| 6 | Joe, Ted | 7, 9 | 16 | 8.00 | 17 | Bob, Art | 8, 8 | 16 | 8.00 |
| 7 | Sam, Sue | 7, 8 | 15 | 7.50 | 18 | Bob, Ted | 8, 9 | 17 | 8.50 |
| 8 | Sam, Bob | 7, 8 | 15 | 7.50 | 19 | Jan, Art | 7, 8 | 15 | 7.50 |
| 9 | Sam, Jan | 7, 7 | 14 | 7.00 | 20 | Jan, Ted | 7, 9 | 16 | 8.00 |
| 10 | Sam, Art | 7, 8 | 15 | 7.50 | 21 | Art, Ted | 8, 9 | 17 | 8.50 |
| 11 | Sam, Ted | 7, 9 | 16 | 8.00 | | | | | |

Sampling Distribution of the Sample Mean – Example

| Sample Mean | Number of Means | Probability |
|-------------|-----------------|-------------|
| \$7.00 | 3 | .1429 |
| 7.50 | 9 | .4285 |
| 8.00 | 6 | .2857 |
| 8.50 | 3 | .1429 |
| | 21 | 1.0000 |

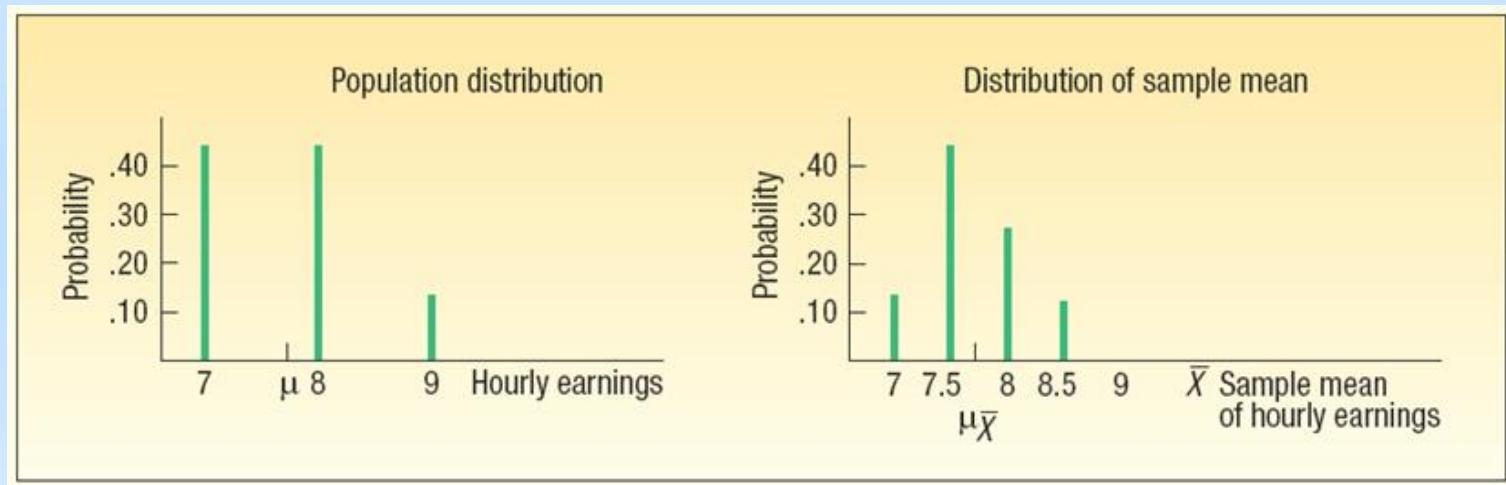
Example

- What is the mean of the sampling distribution?

$$\begin{aligned}\mu_{\bar{X}} &= \frac{\text{Sum of all sample means}}{\text{Total number of samples}} = \frac{\$7.00 + \$7.50 + \dots + \$8.50}{21} \\ &= \frac{\$162}{21} = \$7.71\end{aligned}$$

- What is the standard deviation of the sampling distribution?

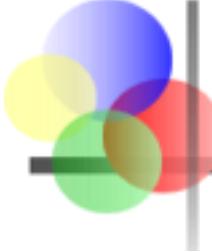
$$S_{\bar{X}} = \frac{S}{\sqrt{n}} = \frac{0.70}{\sqrt{2}} = 0.49$$



Central Limit Theorem

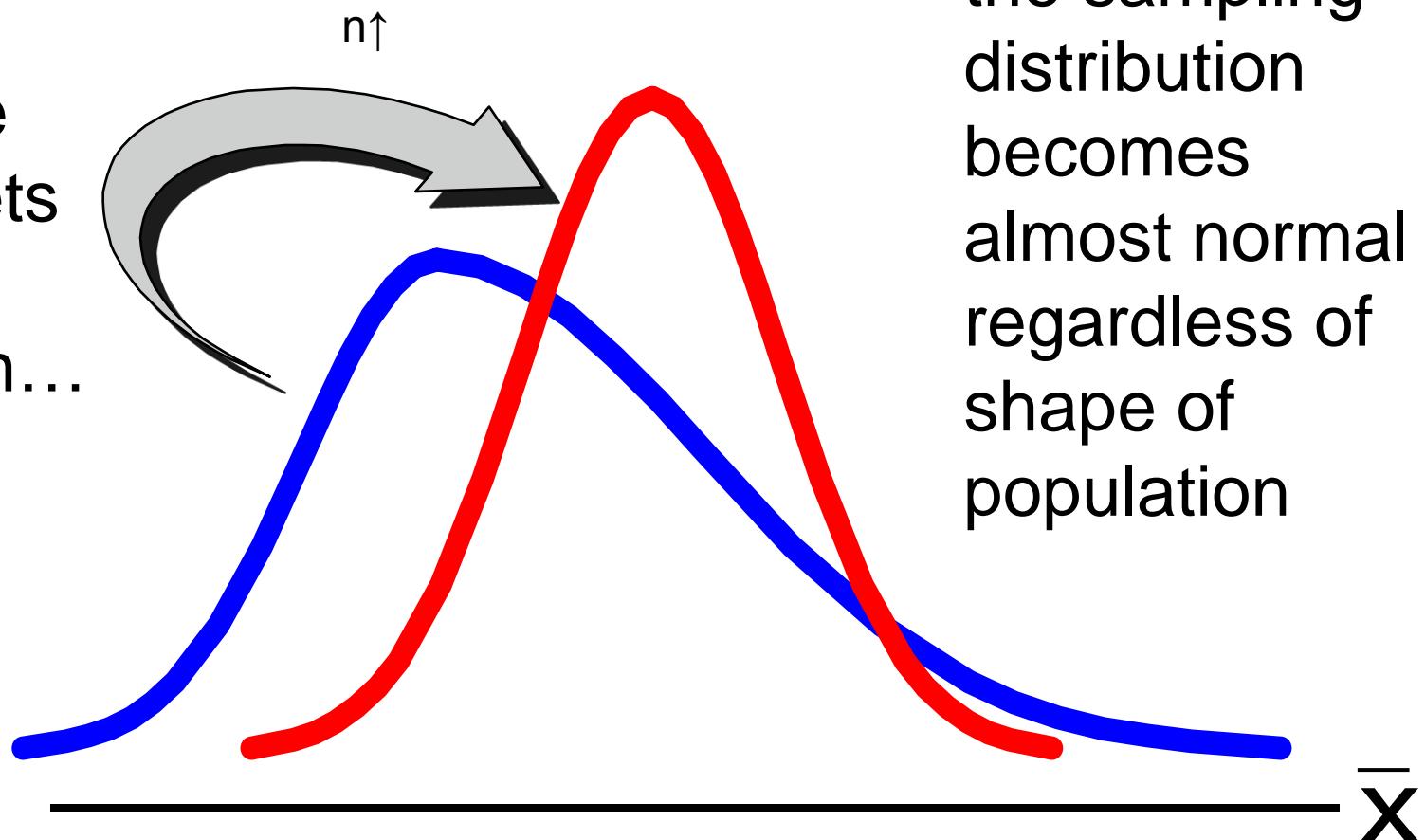
CENTRAL LIMIT THEOREM If all samples of a particular size are selected from any population, the sampling distribution of the sample mean is approximately a normal distribution. This approximation improves with larger samples.

- If the population follows a normal probability distribution, then for any sample size the sampling distribution of the sample mean will also be normal.
- If the population does not follow a normal p.d, or we do not have an idea about its shape, a sample of 30 or higher guarantees the normal shape of the distribution of the sample mean.
- The mean of the sampling distribution is equal to μ . The variance is equal to σ^2/n and the standard deviation is equal to s/\sqrt{n}



Central Limit Theorem

As the sample size gets large enough...



Using the Sampling Distribution of the Sample Mean

- When the population standard deviation is known, a z-statistic for the sampling distribution of the sample mean is calculated as:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Example

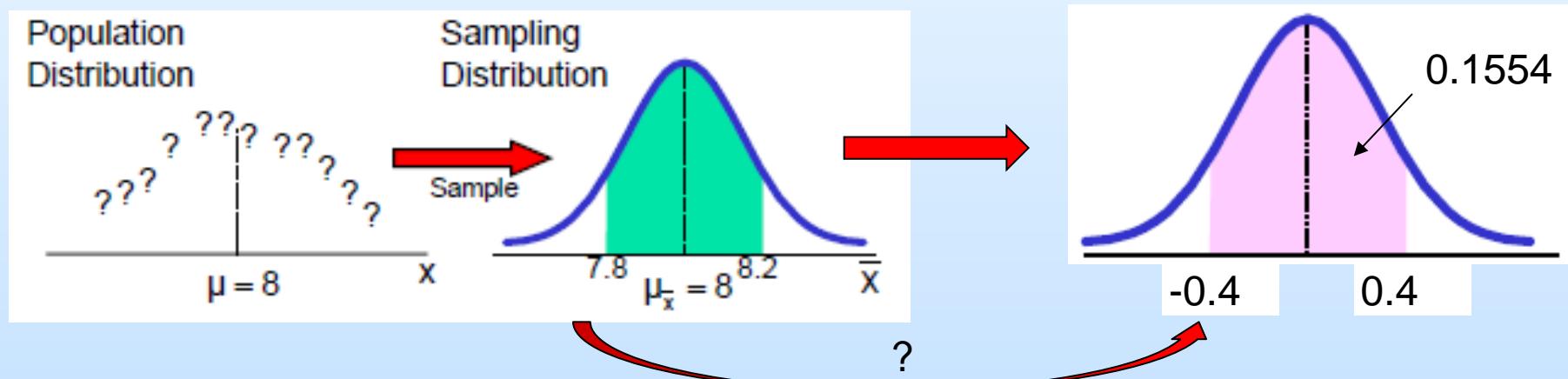
- Suppose a population has mean $\mu = 8$ and standard deviation $\sigma = 3$.
- Suppose a random sample of size $n = 36$ is selected.
- What is the probability that the **sample mean** is between 7.8 and 8.2?

$$P(7.8 < \bar{x} < 8.2) = ?$$

Solution:

- Even if the population is not normally distributed, the central limit theorem can be used ($n > 30$)
- The sampling distribution of \bar{x} is approx. normal
- ... with mean $\mu_{\bar{x}} = \mu = 8$
- ...and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{7.8 - 8}{3/\sqrt{36}} = -0.4$$

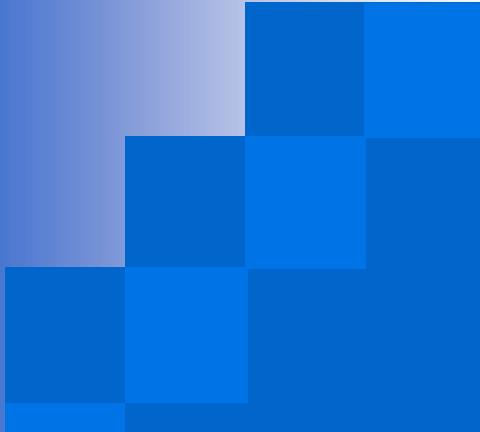


$$\begin{aligned} P(7.8 < \bar{x} < 8.2) &= P\left(\frac{7.8 - 8}{3/\sqrt{36}} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < \frac{8.2 - 8}{3/\sqrt{36}}\right) \\ &= P(-0.4 < z < 0.4) = 0.3108 \end{aligned}$$

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?

$$P(7.8 < \bar{x} < 8.2) = ?$$



Estimation and Confidence Intervals

Chapter 9



Point Estimates

- A **point estimate** is a single value (point) derived from a sample and used to estimate a population value.

$$\bar{X} \rightarrow \mu$$

$$S \rightarrow \sigma$$

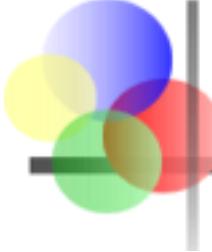
$$S^2 \rightarrow \sigma^2$$

$$p \rightarrow \pi$$

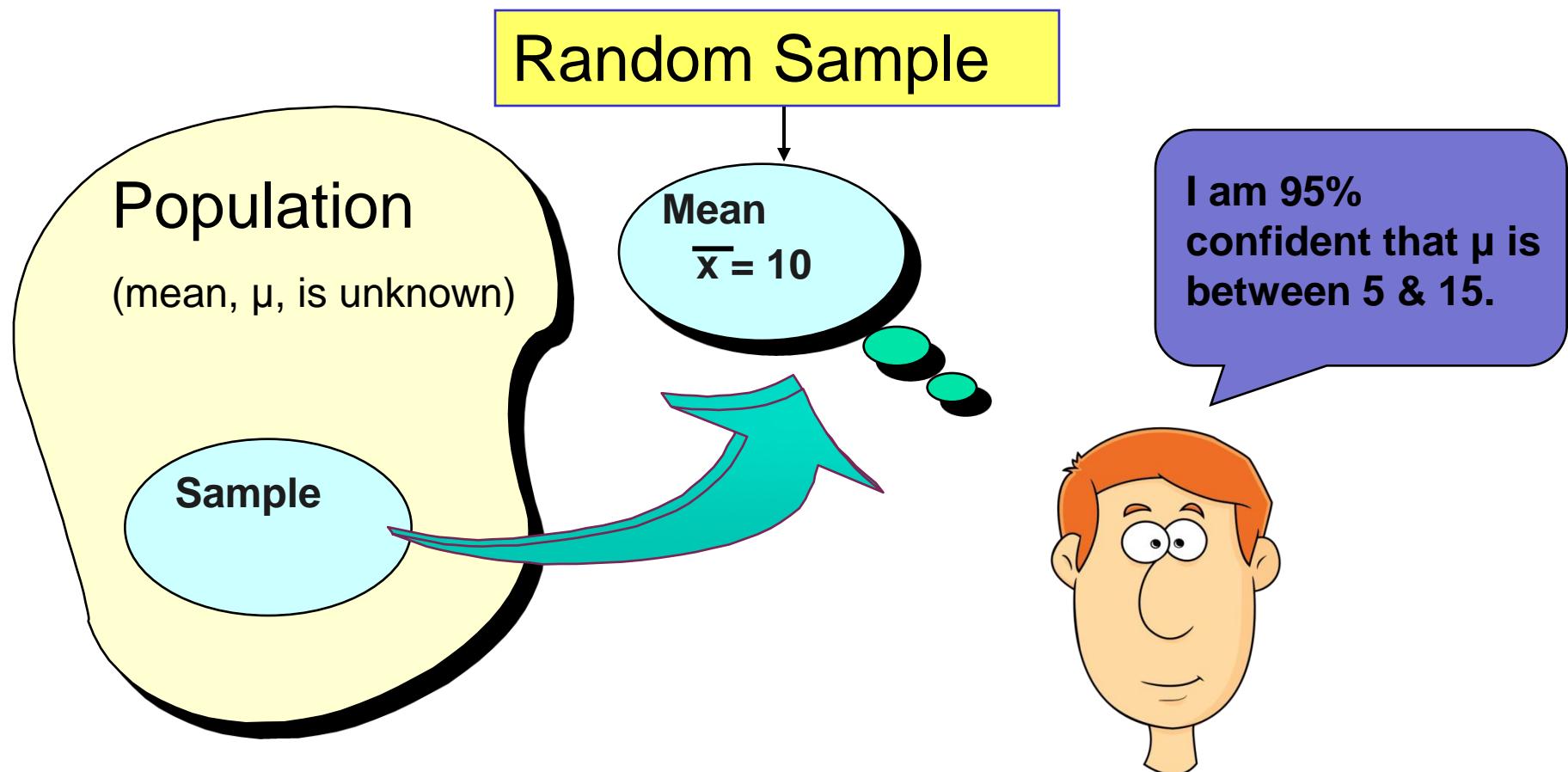
Confidence Interval Estimates

- A **confidence interval estimate** is a range of values constructed from sample data so that the population parameter is likely to occur within that range at a specified probability.

C.I. = point estimate \pm margin of error



Estimation Process



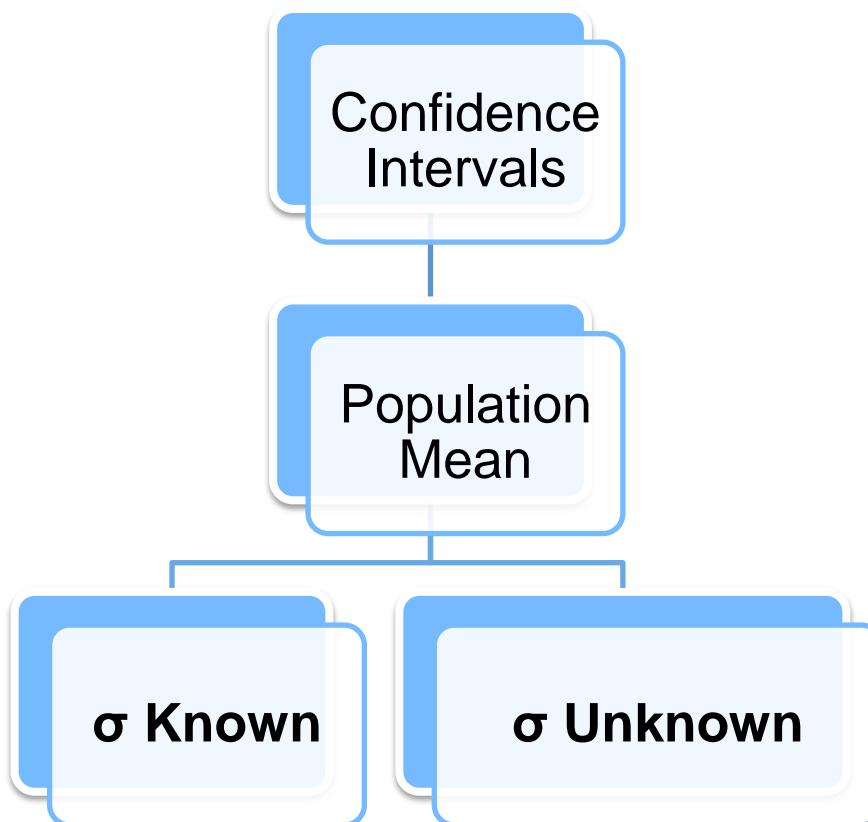
Confidence Level

- Confidence Level
 - Confidence in which the interval will contain the unknown population parameter
- A percentage (less than 100%)

Confidence Level, $(1-\alpha)$

- Suppose confidence level = 95%
- Also written $(1 - \alpha) = .95$
- interpretation:
 - In the long run, 95% of all the confidence intervals that can be constructed will contain the unknown true parameter

Confidence Intervals



Confidence Intervals for a Mean, σ Known

**CONFIDENCE INTERVAL FOR A POPULATION
MEAN WITH σ KNOWN**

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}} \quad [9-1]$$

\bar{x} – sample mean

z – z - value for a particular confidence level

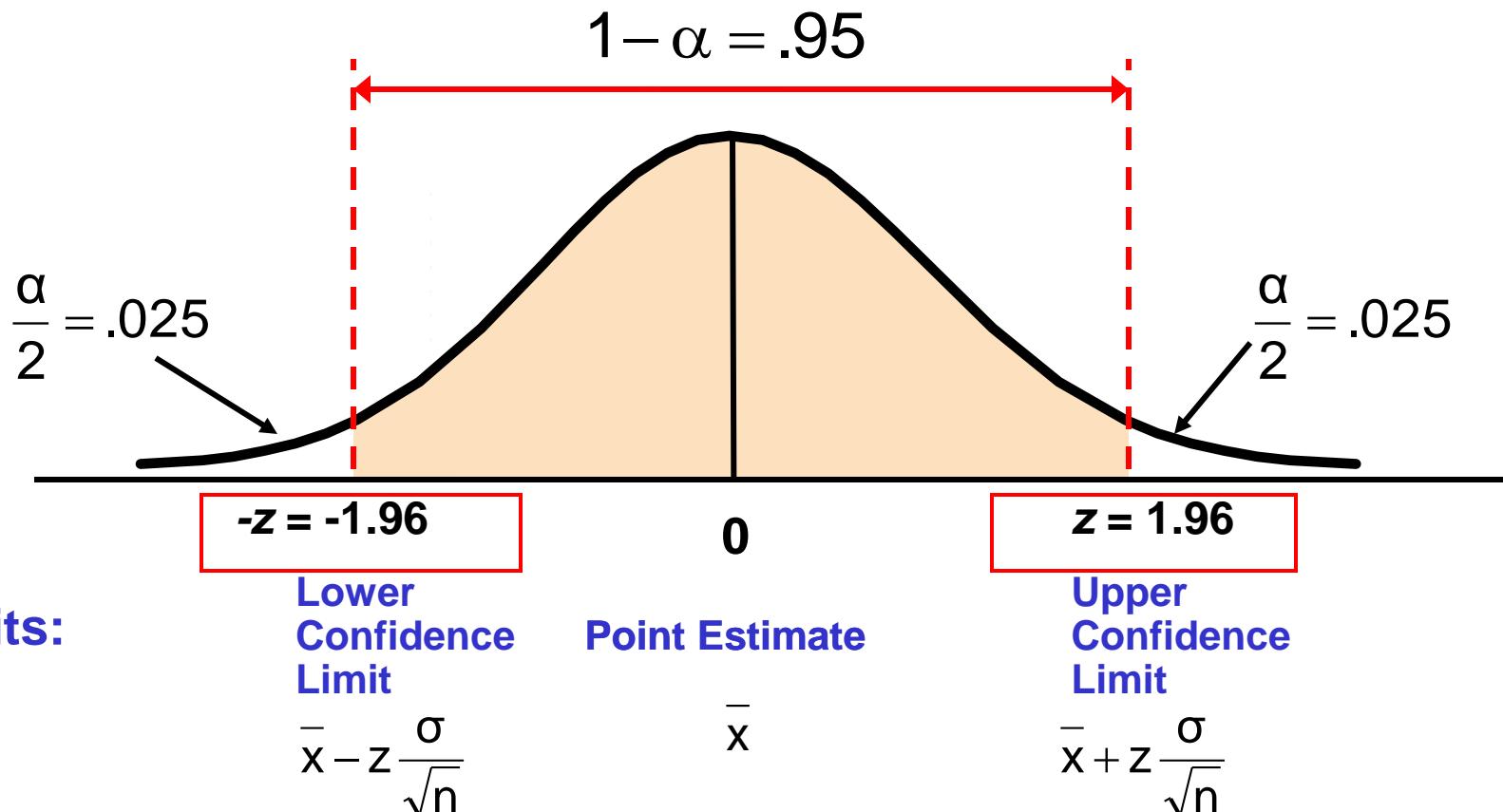
σ – the population standard deviation

n – the number of observations in the sample

Finding the Reliability Factor, $z_{\alpha/2}$

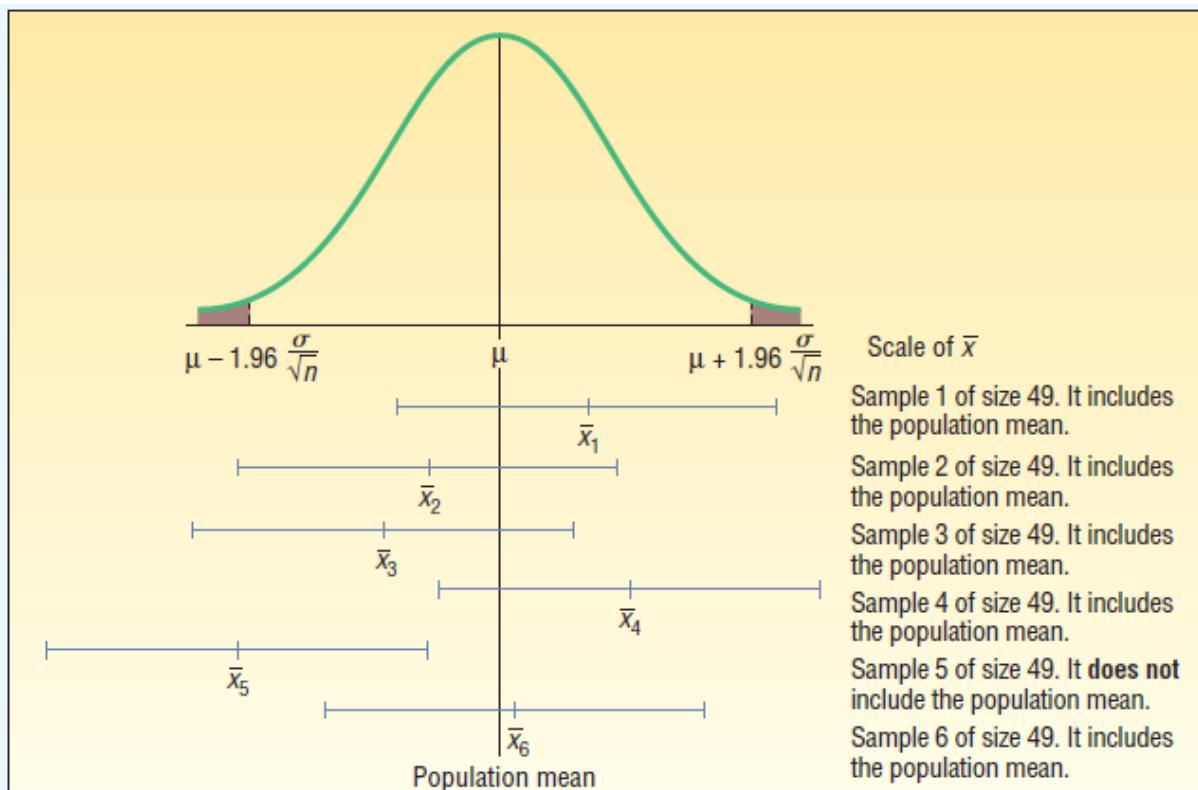
- Consider a 95% confidence interval:

$$z = \pm 1.96$$



Interval Estimates - Interpretation

For a 95% confidence interval, about 95% of similarly constructed intervals will contain the parameter being estimated. Also 95% of the sample means for a specified sample size will lie within 1.96 standard deviations of the hypothesized population.



Confidence Interval for a Mean, σ Known - Example

The American Management Association surveys middle managers in the retail industry and wants to estimate their mean annual income. A random sample of 49 managers reveals a sample mean of \$45,420. The standard deviation of this population is \$2,050.

- What is a reasonable range of values for the population mean?
- What do these results mean?

Confidence Interval for a Mean, σ Known - Example

The American Management Association surveys middle managers in the retail industry and wants to estimate their mean annual income. A random sample of 49 managers reveals a sample mean of \$45,420. The standard deviation of this population is \$2,050.

- What is a reasonable range of values for the population mean?

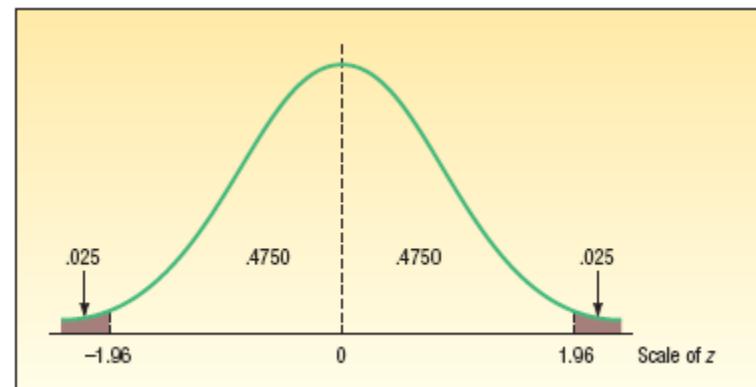
CONFIDENCE INTERVAL FOR A POPULATION
MEAN WITH σ KNOWN

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}} \quad [9-1]$$

- Suppose the association decides to use the 95 percent level of confidence.

How to Obtain a z-value for a Given Confidence Level

The *95 percent confidence* refers to the middle 95 percent of the observations. Therefore, the remaining 5 percent are equally divided between the two tails.



Following is a portion of Appendix B.3.

| <i>z</i> | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| 1.5 | 0.4332 | 0.4345 | 0.4357 | 0.4370 | 0.4382 | 0.4394 | 0.4406 | 0.4418 | 0.4429 | 0.4441 |
| 1.6 | 0.4452 | 0.4463 | 0.4474 | 0.4484 | 0.4495 | 0.4505 | 0.4515 | 0.4525 | 0.4535 | 0.4545 |
| 1.7 | 0.4554 | 0.4564 | 0.4573 | 0.4582 | 0.4591 | 0.4599 | 0.4608 | 0.4616 | 0.4625 | 0.4633 |
| 1.8 | 0.4641 | 0.4649 | 0.4656 | 0.4664 | 0.4671 | 0.4678 | 0.4686 | 0.4693 | 0.4699 | 0.4706 |
| 1.9 | 0.4713 | 0.4719 | 0.4726 | 0.4732 | 0.4738 | 0.4744 | 0.4750 | 0.4756 | 0.4761 | 0.4767 |
| 2.0 | 0.4772 | 0.4778 | 0.4783 | 0.4788 | 0.4793 | 0.4798 | 0.4803 | 0.4808 | 0.4812 | 0.4817 |
| 2.1 | 0.4821 | 0.4826 | 0.4830 | 0.4834 | 0.4838 | 0.4842 | 0.4846 | 0.4850 | 0.4854 | 0.4857 |
| 2.2 | 0.4861 | 0.4864 | 0.4868 | 0.4871 | 0.4875 | 0.4878 | 0.4881 | 0.4884 | 0.4887 | 0.4890 |
| 2.3 | 0.4893 | 0.4896 | 0.4898 | 0.4901 | 0.4904 | 0.4906 | 0.4909 | 0.4911 | 0.4913 | 0.4916 |
| 2.4 | 0.4918 | 0.4920 | 0.4922 | 0.4925 | 0.4927 | 0.4929 | 0.4931 | 0.4932 | 0.4934 | 0.4936 |

Confidence Interval for a Mean, σ Known – Example

The 95 percent confidence interval estimate is:

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}} = \$45,420 \pm 1.96 \frac{\$2,050}{\sqrt{49}} = \$45,420 \pm \$574$$

The confidence interval limits are \$44,846 and \$45,994. The degree or level of confidence is 95% and the confidence interval is from \$44,846 to \$45,994. The $\pm \$574$ is called the margin of error.

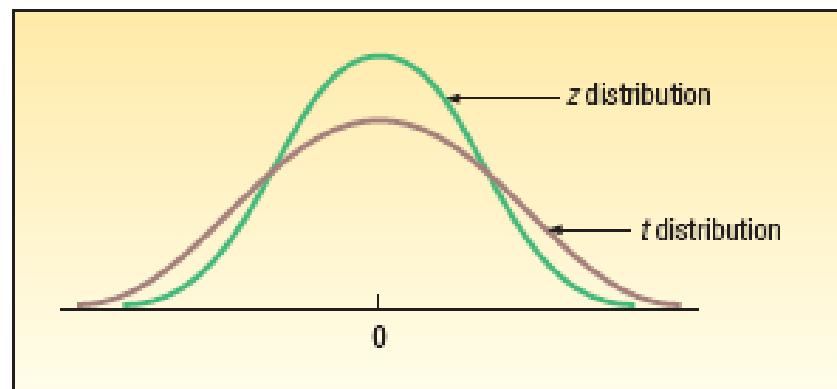
Confidence Intervals for a Mean, σ Unknown

In most sampling situations the population standard deviation (σ) is not known. Below are some examples where it is unlikely the population standard deviations would be known.

- The Dean of Students wants to estimate the distance the typical commuter student travels to class. She selects a sample of 40 commuter students, contacts each, and determines the one-way distance from each student's home to the center of campus.

Using the *t*-Distribution: Confidence Intervals for a Mean, σ Unknown

- It is, like the *z* distribution, **bell-shaped and symmetrical**.
- There is **not one *t* distribution**, but rather a family of *t* distributions. All *t* distributions have a mean of 0, but their standard deviations differ according to the sample size, n .
- The *t* distribution is more spread out and flatter at the center than the standard normal distribution. As the sample size increases, however, the *t* distribution approaches the standard normal distribution.



If σ Unknown

- In order to use t dist,
 - either
 - You should know that the population is normally distributed
 - or
 - Sample size is larger than 30

Confidence Interval when σ Unknown

$$\bar{x} - t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1, \alpha/2} \frac{s}{\sqrt{n}}$$

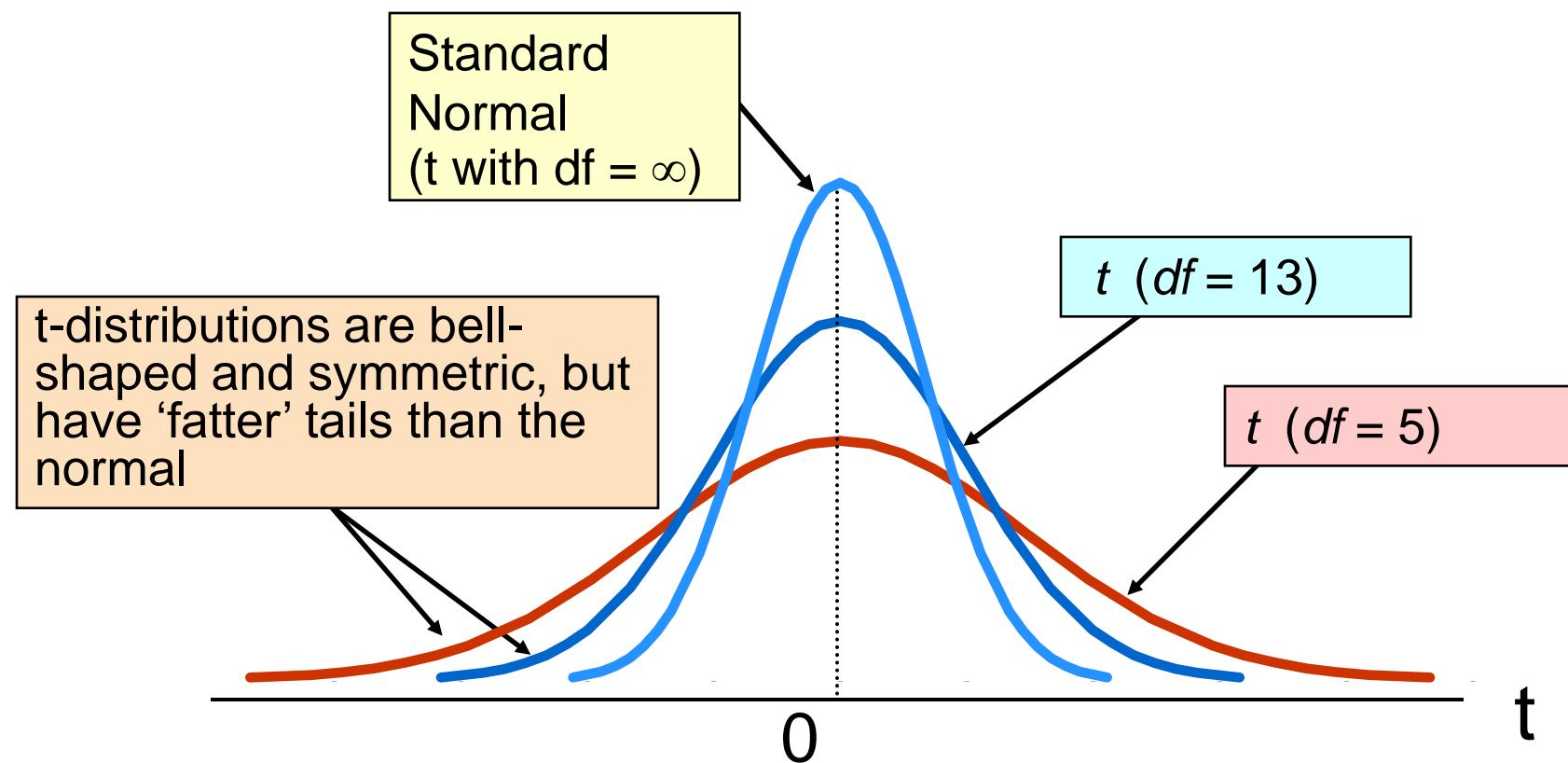
Student's t Distribution

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

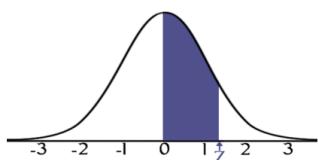
$$d.f. = n - 1$$

Student's t Distribution

Note: $t \rightarrow z$ as n increases



Z-Table



STANDARD NORMAL TABLE (Z)

Entries in the table give the area under the curve between the mean and z standard deviations above the mean. For example, for $z = 1.25$ the area under the curve between the mean (0) and z is 0.3944.

| <i>z</i> | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|-----------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| 0.0 | 0.0000 | 0.0040 | 0.0080 | 0.0120 | 0.0160 | 0.0190 | 0.0239 | 0.0279 | 0.0319 | 0.0359 |
| 0.1 | 0.0398 | 0.0438 | 0.0478 | 0.0517 | 0.0557 | 0.0596 | 0.0636 | 0.0675 | 0.0714 | 0.0753 |
| 0.2 | 0.0793 | 0.0832 | 0.0871 | 0.0910 | 0.0948 | 0.0987 | 0.1026 | 0.1064 | 0.1103 | 0.1141 |
| 0.3 | 0.1179 | 0.1217 | 0.1255 | 0.1293 | 0.1331 | 0.1368 | 0.1406 | 0.1443 | 0.1480 | 0.1517 |
| 0.4 | 0.1554 | 0.1591 | 0.1628 | 0.1664 | 0.1700 | 0.1736 | 0.1772 | 0.1808 | 0.1844 | 0.1879 |
| 0.5 | 0.1915 | 0.1950 | 0.1985 | 0.2019 | 0.2054 | 0.2088 | 0.2123 | 0.2157 | 0.2190 | 0.2224 |
| 0.6 | 0.2257 | 0.2291 | 0.2324 | 0.2357 | 0.2389 | 0.2422 | 0.2454 | 0.2486 | 0.2517 | 0.2549 |
| 0.7 | 0.2580 | 0.2611 | 0.2642 | 0.2673 | 0.2704 | 0.2734 | 0.2764 | 0.2794 | 0.2823 | 0.2852 |
| 0.8 | 0.2881 | 0.2910 | 0.2939 | 0.2969 | 0.2995 | 0.3023 | 0.3051 | 0.3078 | 0.3106 | 0.3133 |
| 0.9 | 0.3159 | 0.3186 | 0.3212 | 0.3238 | 0.3264 | 0.3289 | 0.3315 | 0.3340 | 0.3365 | 0.3389 |
| 1.0 | 0.3413 | 0.3438 | 0.3461 | 0.3485 | 0.3508 | 0.3513 | 0.3554 | 0.3577 | 0.3529 | 0.3621 |
| 1.1 | 0.3643 | 0.3665 | 0.3686 | 0.3708 | 0.3729 | 0.3749 | 0.3770 | 0.3790 | 0.3810 | 0.3830 |
| 1.2 | 0.3849 | 0.3869 | 0.3888 | 0.3907 | 0.3925 | 0.3944 | 0.3962 | 0.3980 | 0.3997 | 0.4015 |
| 1.3 | 0.4032 | 0.4049 | 0.4066 | 0.4082 | 0.4099 | 0.4115 | 0.4131 | 0.4147 | 0.4162 | 0.4177 |

T-table

| <i>df</i> | Confidence Intervals | | | | |
|------------------|--------------------------------------------------|--------------|---------------|---------------|---------------|
| | 80% | 90% | 95% | 98% | 99% |
| <i>df</i> | Level of Significance for One-Tailed Test | | | | |
| | 0.10 | 0.05 | 0.025 | 0.010 | 0.005 |
| <i>df</i> | Level of Significance for Two-Tailed Test | | | | |
| | 0.20 | 0.10 | 0.05 | 0.02 | 0.01 |
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |

Using the *t*-Distribution; Confidence Intervals for a Mean, σ Unknown

A tire manufacturer wishes to investigate the tread life of its tires. A **sample of 10** tires driven 50,000 miles revealed a **sample mean of 0.32 inch** of tread remaining with a **standard deviation of 0.09 inch**.

Construct a 95 percent confidence interval for the population mean.

Would it be reasonable for the manufacturer to conclude that after 50,000 miles the population mean amount of tread remaining is 0.30 inches?

Given in the problem:

$$n = 10$$

$$\bar{x} = 0.32$$

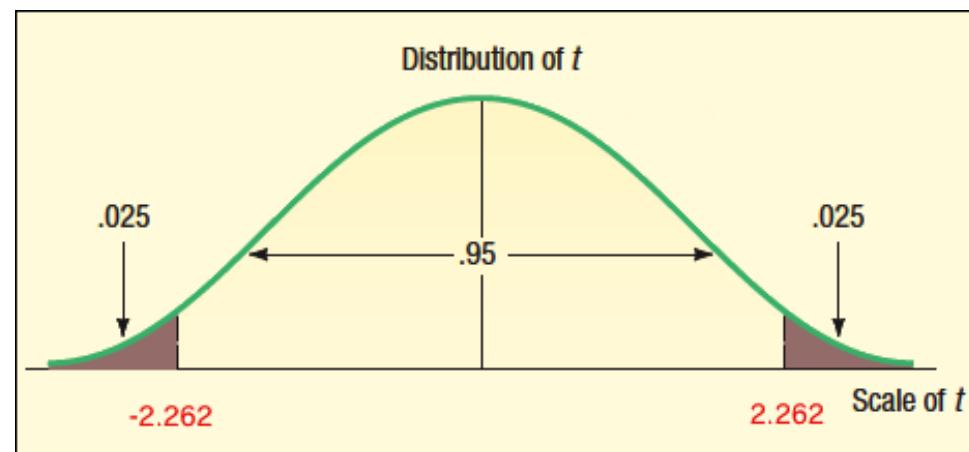
$$s = 0.09$$

Compute the confidence interval using the *t*-distribution (since s is unknown).

$$\bar{x} \pm t \frac{s}{\sqrt{n}}$$

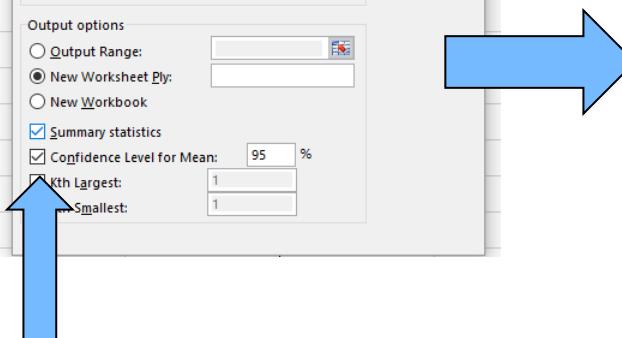
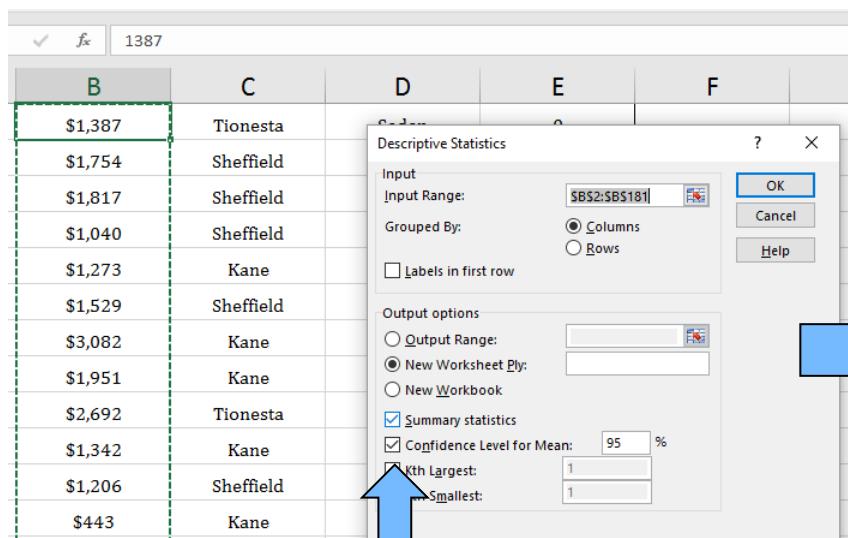
Using the Student's t -Distribution Table

| Confidence Intervals | | | | | |
|-------------------------------------------|-------------------------------------------|-------|--------|--------|--------|
| | 80% | 90% | 95% | 98% | 99% |
| <i>df</i> | Level of Significance for One-Tailed Test | | | | |
| | 0.10 | 0.05 | 0.025 | 0.010 | 0.005 |
| Level of Significance for Two-Tailed Test | | | | | |
| | 0.20 | 0.10 | 0.05 | 0.02 | 0.01 |
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 |



$$\bar{x} \pm t \frac{s}{\sqrt{n}} = 0.32 \pm 2.262 \frac{0.09}{\sqrt{10}} = 0.32 \pm .064$$

Confidence Interval Estimates for the Mean – Using Excel



The screenshot shows the results of the Descriptive Statistics analysis in a new worksheet. The results are listed in columns A and B. The 'Mean' is 1843.166667, the 'Standard Error' is 47.97317525, the 'Median' is 1882.5, the 'Mode' is 1761, the 'Standard Deviation' is 643.6276857, the 'Sample Variance' is 414256.5978, the 'Kurtosis' is -0.21659402, the 'Skewness' is -0.24294466, the 'Range' is 2998, the 'Minimum' is 294, the 'Maximum' is 3292, the 'Sum' is 331770, the 'Count' is 180, and the 'Confidence Level(95.0%)' is 94.66572739.

| A | B |
|----|-------------------------------------|
| 1 | Column1 |
| 2 | |
| 3 | Mean 1843.166667 |
| 4 | Standard Error 47.97317525 |
| 5 | Median 1882.5 |
| 6 | Mode 1761 |
| 7 | Standard Deviation 643.6276857 |
| 8 | Sample Variance 414256.5978 |
| 9 | Kurtosis -0.21659402 |
| 10 | Skewness -0.24294466 |
| 11 | Range 2998 |
| 12 | Minimum 294 |
| 13 | Maximum 3292 |
| 14 | Sum 331770 |
| 15 | Count 180 |
| 16 | Confidence Level(95.0%) 94.66572739 |
| 17 | |

$$1843 \pm 94.66$$

1748.50.....1937.83

When to Use the z or t Distribution for Confidence Interval Computation

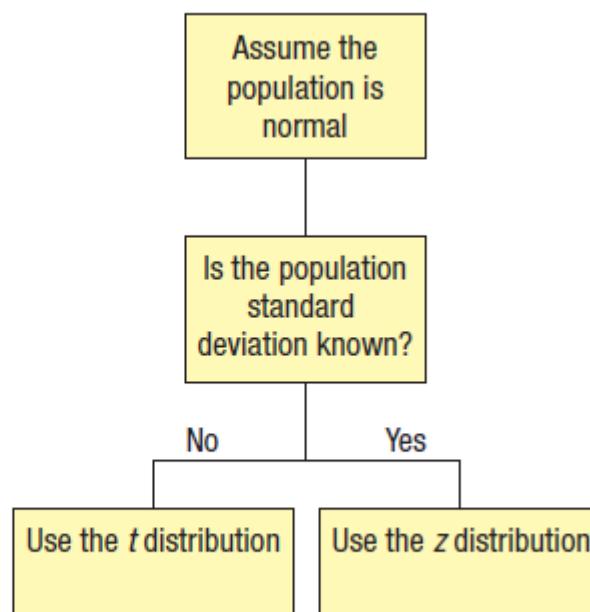


CHART 9–3 Determining When to Use the z Distribution or the t Distribution

When to Use the z or t Distribution for Confidence Interval Computation

Use Z -distribution,

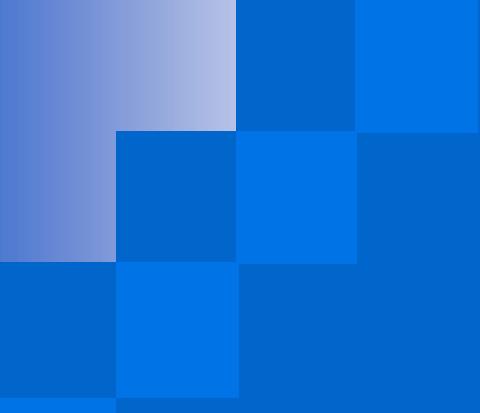
If the population standard deviation is known.

$$\bar{x} \pm z \frac{\sigma}{\sqrt{n}}$$

Use t -distribution,

If the population standard deviation is unknown.

$$\bar{x} \pm t \frac{s}{\sqrt{n}}$$



One-Sample Tests of Hypothesis



Chapter 10

Hypothesis

HYPOTHESIS A statement about a population parameter subject to verification.

Examples:

- The mean speed of automobiles on the West Virginia Turnpike is 68 miles per hour.
- The mean monthly cell phone bill of this city is \$60.

Hypothesis Testing

HYPOTHESIS TESTING A procedure based on sample evidence to determine whether the hypothesis is a reasonable statement.

Null and the Alternate Hypothesis

NULL HYPOTHESIS A statement about the value of a population parameter developed for the purpose of testing numerical evidence. It is represented by H_0 .

Always contains “=” , “ \leq ” or “ \geq ” sign

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

Example:

The mean monthly cell phone bill of this city is \$60.

$H_0: \mu = \$60$

Null and the Alternate Hypothesis

- *Begin with the assumption that the null hypothesis is true

- *Similar to the notion of innocent until proven guilty



Null and the Alternate Hypothesis

ALTERNATE HYPOTHESIS It is the opposite of the null hypothesis. It is represented by H_1 .

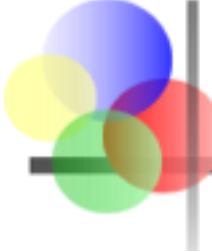
Never contains the “=” , “≤” or “≥” sign

EXAMPLE: The mean monthly cell phone bill of this city is NOT \$60.

$H_1: \mu \neq \$60$

Example

- The average annual income of people in Dallas is claimed to be \$75,000 per year. An industry analyst would like to test this claim.
- What is the appropriate hypothesis test?

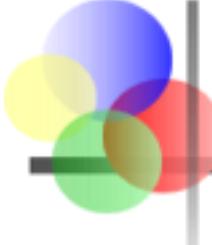


Formulating Hypotheses

- **Example:** The average annual income of people in Dallas is claimed to be \$75,000 per year. An industry analyst would like to test this claim.
 - What is the appropriate test?
-

$H_0: \mu = 75,000$ (income is as claimed) **status quo**

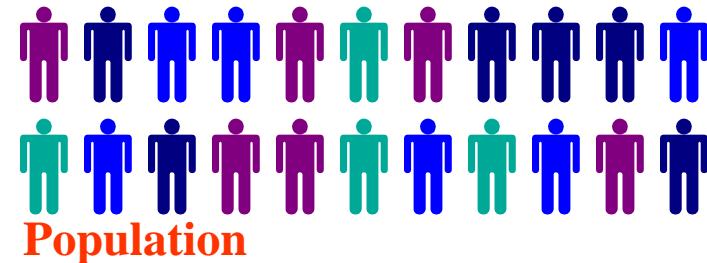
$H_A: \mu \neq 75,000$ (income is different than claimed)



Hypothesis Testing Process

Claim: the population mean phone bill \$60.

Null Hypothesis: $H_0: \mu = \$60$



Now select a random sample:



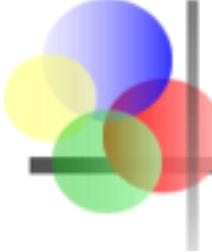
Suppose the sample mean phone bill is \$40:
 $\bar{x} = 40$



Is $\bar{x} = 40$ likely if $\mu = 60$?

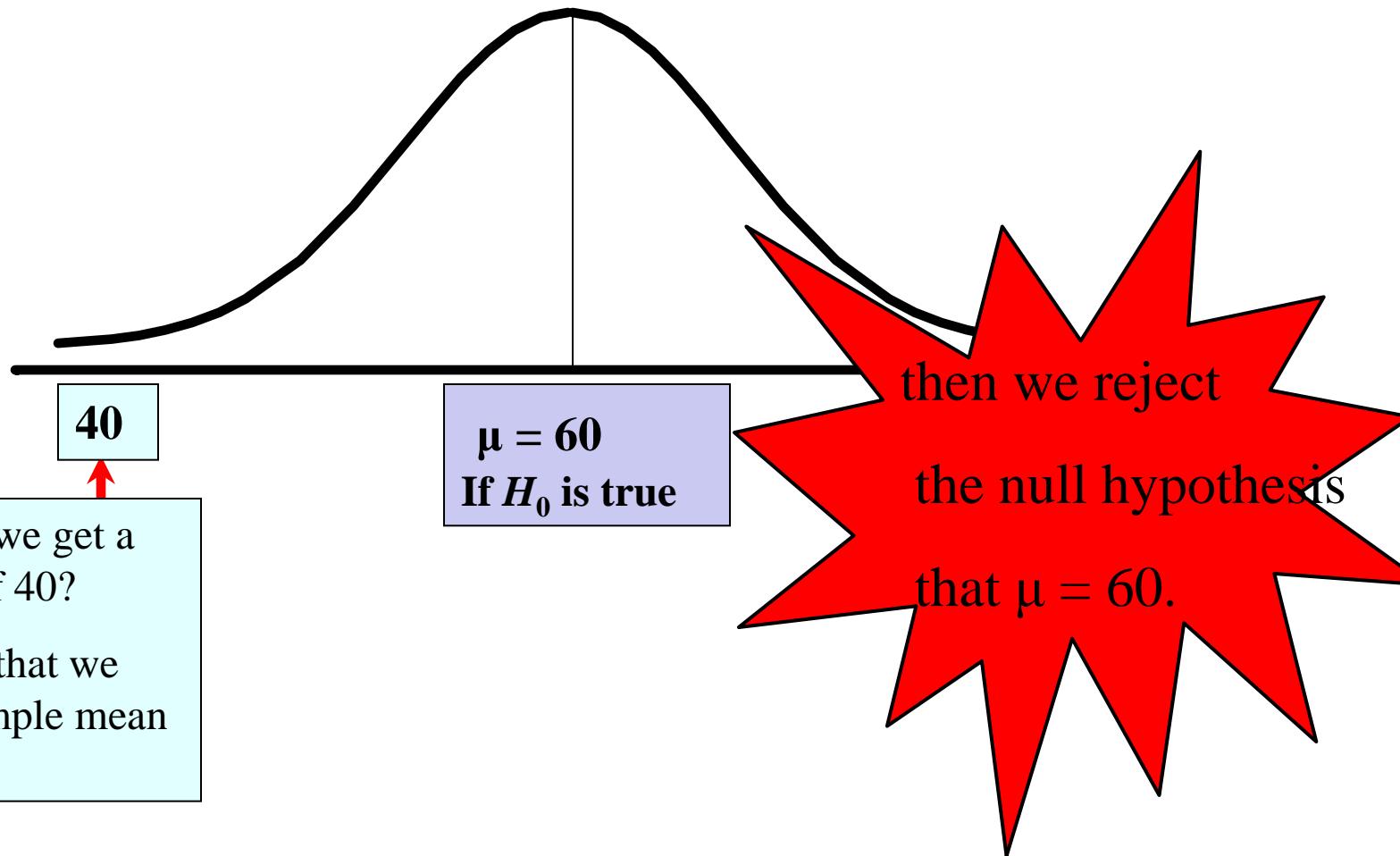


If not likely,
REJECT
Null Hypothesis

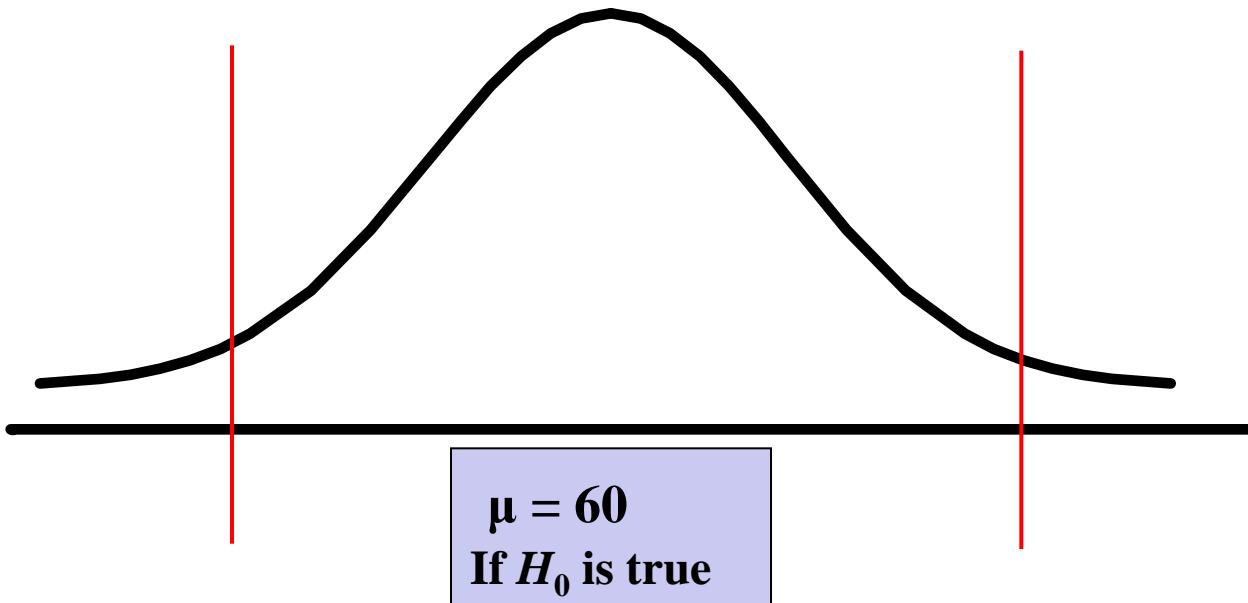


Reason for Rejecting H_0

Sampling Distribution of \bar{x}



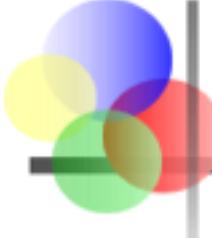
- So we have to determine the likeliness boundaries which is called the significance level.



Level of Significance: Errors in Hypothesis Testing

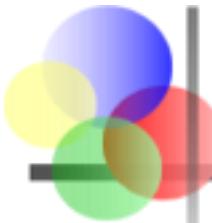
The ***significance level*** of a test:

- Defined as the probability of rejecting the null hypothesis when it is actually true.
- This is denoted by the Greek letter “ α ”.
- Also known as **Type I Error**.
- A typical value of “ α ” is 0.05.



Steps

1. State the Null and the Alternate Hypothesis
2. Specify the desired significance level, α
3. Identify test statistic
4. Formulate the decision rule
5. Take a random sample and determine whether or not the sample result is in the rejection region
6. Reach a decision and draw a conclusion

- 
- There can be three kinds of hypothesis

$$H_0: \mu \geq 60$$

$$H_A: \mu < 60$$

$$H_0: \mu = 60$$

$$H_A: \mu \neq 60$$

$$H_0: \mu \leq 60$$

$$H_A: \mu > 60$$

Level of Significance and the Rejection Region

Level of significance = $\alpha=5\%$

Lower tail test

Example:

$$H_0: \mu \geq 60$$

$$H_A: \mu < 60$$

Two tailed test

Example:

$$H_0: \mu = 60$$

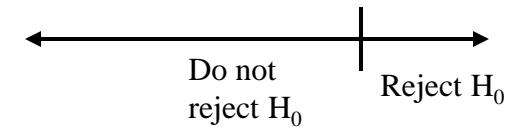
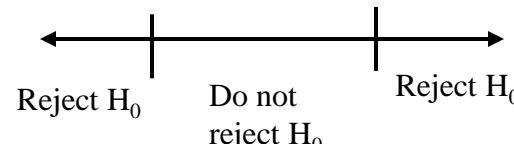
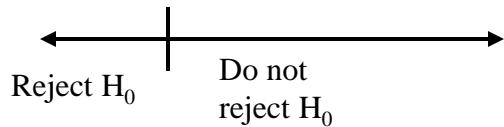
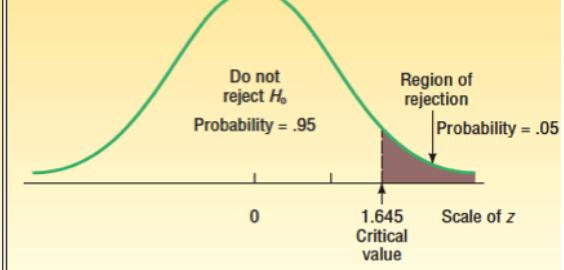
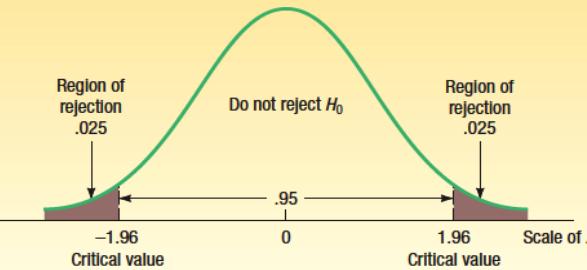
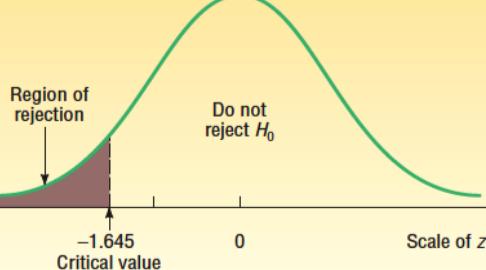
$$H_A: \mu \neq 60$$

Upper tail test

Example:

$$H_0: \mu \leq 60$$

$$H_A: \mu > 60$$



If the test statistic is in the rejection region, then reject the null hypothesis.

Hypothesis Setups for Testing a Mean (μ)

$$H_0: \mu = \text{value}$$

$$H_1: \mu \neq \text{value}$$

Reject H_0 if:

$$|Z| > Z_{\alpha/2}$$

$$|t| > t_{\alpha/2, n-1}$$

$$H_0: \mu \geq \text{value}$$

$$H_1: \mu < \text{value}$$

Reject H_0 if:

$$Z < -Z_\alpha$$

$$t < -t_{\alpha, n-1}$$

$$H_0: \mu \leq \text{value}$$

$$H_1: \mu > \text{value}$$

Reject H_0 if:

$$Z > Z_\alpha$$

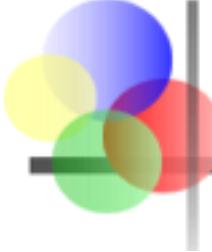
$$t > t_{\alpha, n-1}$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

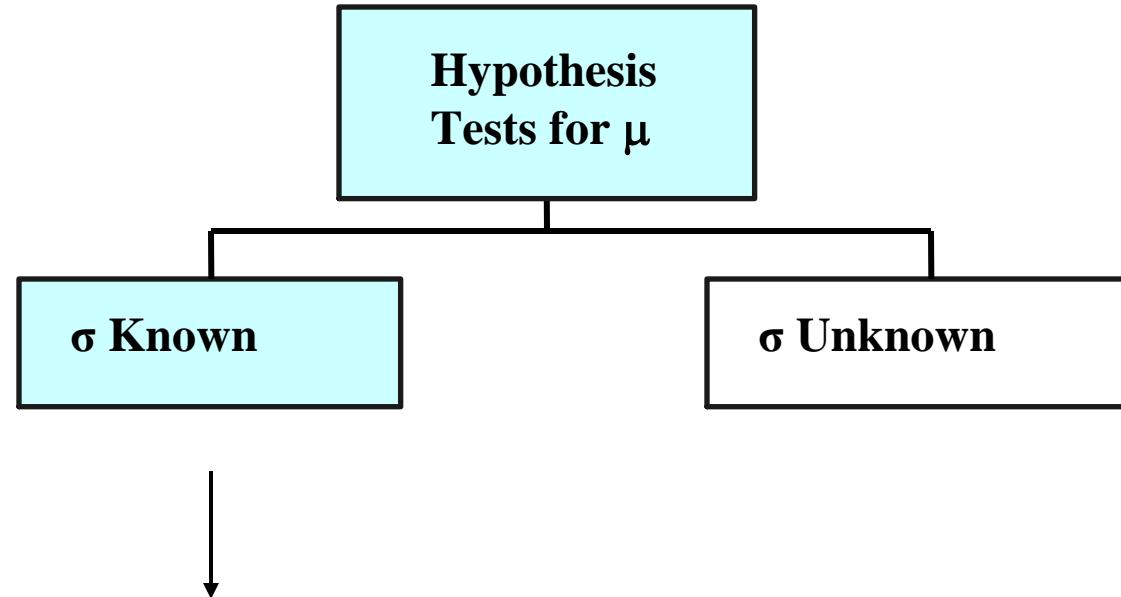
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Important Things to Remember about H_0 and H_1

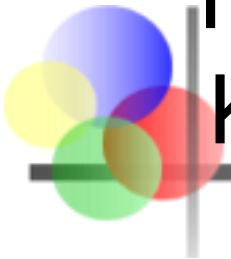
- H_0 is the null hypothesis; H_1 is the alternate hypothesis.
- H_0 and H_1 are mutually exclusive and collectively exhaustive.
- H_0 is always presumed to be true.
- H_1 has the burden of proof.
- A random sample (n) is used to “reject H_0 .”
- If we conclude “do not reject H_0 ,” this does not necessarily mean that the null hypothesis is true, it only suggests that there is not sufficient evidence to reject H_0 ; rejecting the null hypothesis, suggests that the alternative hypothesis may be true given the probability of Type I error.
- Equality is always part of H_0 (e.g. “ $=$ ”, “ \geq ”, “ \leq ”).
- Inequality is always part of H_1 (e.g. “ \neq ”, “ $<$ ”, “ $>$ ”).



Hypothesis Tests for the Mean

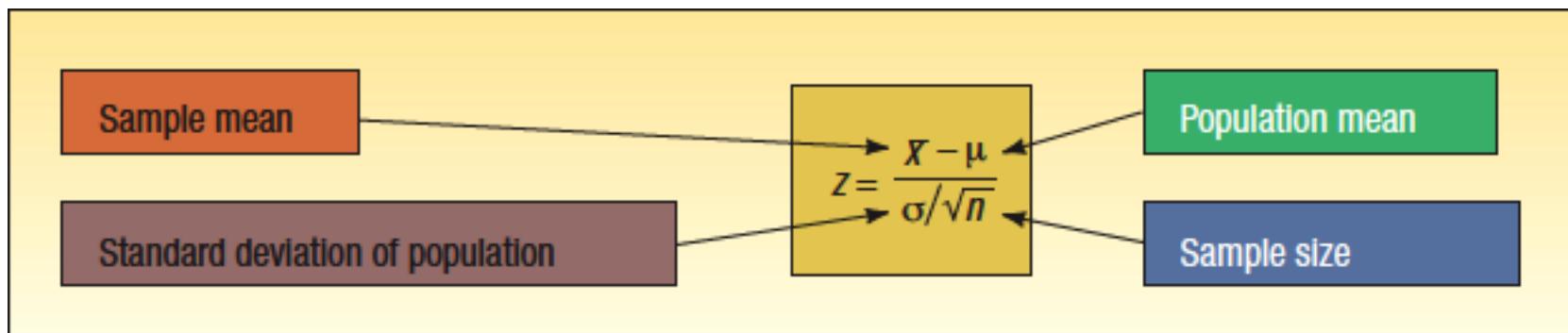


- Assume first that the population standard deviation σ is known



Hypothesis Test of a Population Mean, Known Population Standard Deviation

Use z-distribution since σ is known.



Example

Jamestown Steel Company manufactures office equipment. Test the claim that the true mean of weekly production is 200.

Suppose a sample is taken with the following results:

($n=50$ week $\bar{x}=203.5$ $\sigma = 16$, $\alpha=.01$)



Example

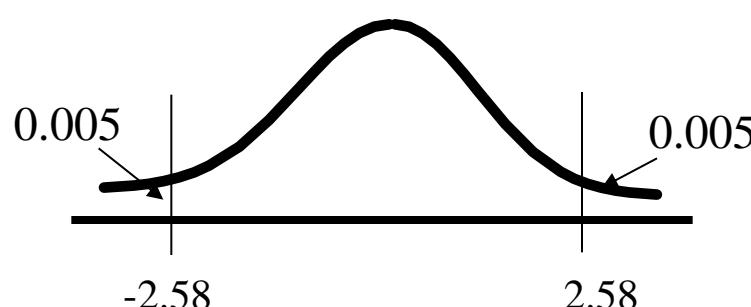
Step 1: State the null and alternate hypotheses.

$$H_0: \mu = 200$$

$$H_1: \mu \neq 200$$

Step 2: Select the level of significance.

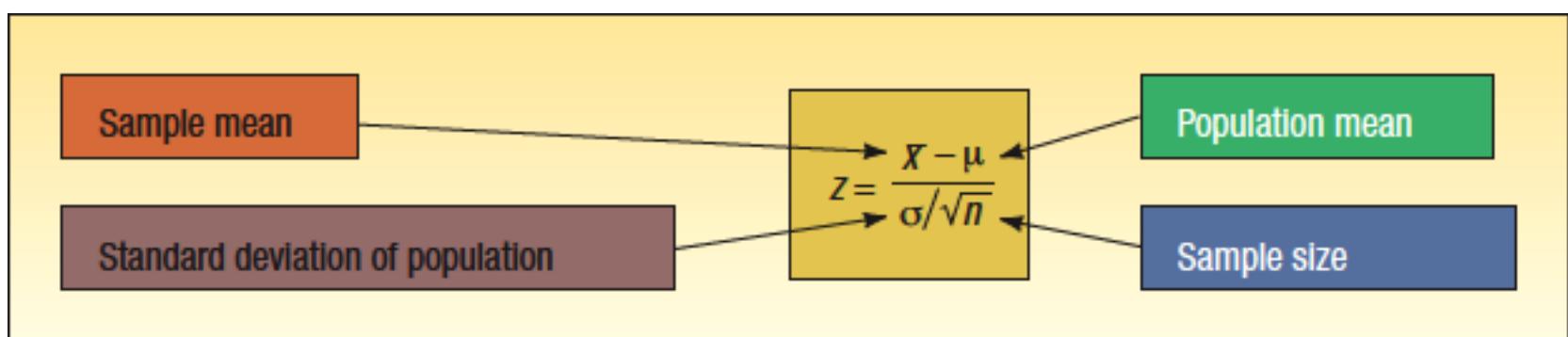
$\alpha = 0.01$ as stated in the problem.



Example

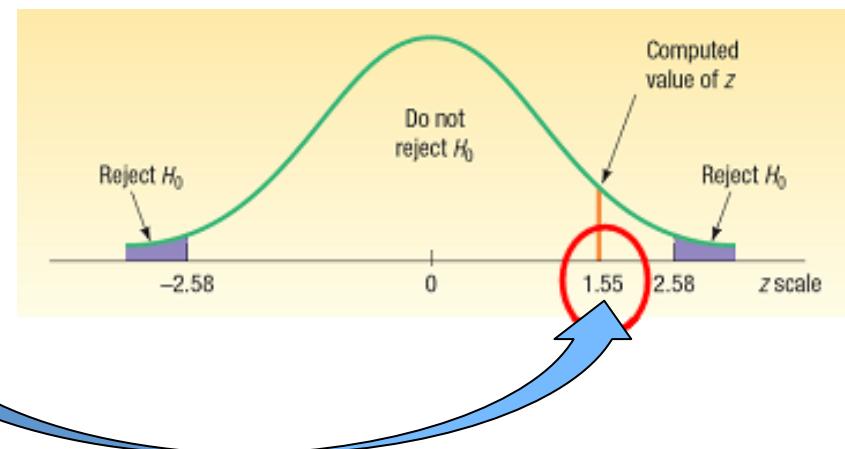
Step 3: Select the test statistic.

Use z-distribution since σ is known.



■ Step 4: Formulate the decision rule.

$$\blacksquare \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{203.5 - 200}{16 / \sqrt{50}} = 1.55$$



- **Step 5: Make a decision and interpret the result.**
- H_0 is not rejected because 1.55 does not fall in the rejection region.

- **Step 6: Interpret the result.**
- We conclude that the population mean is not different from 200.

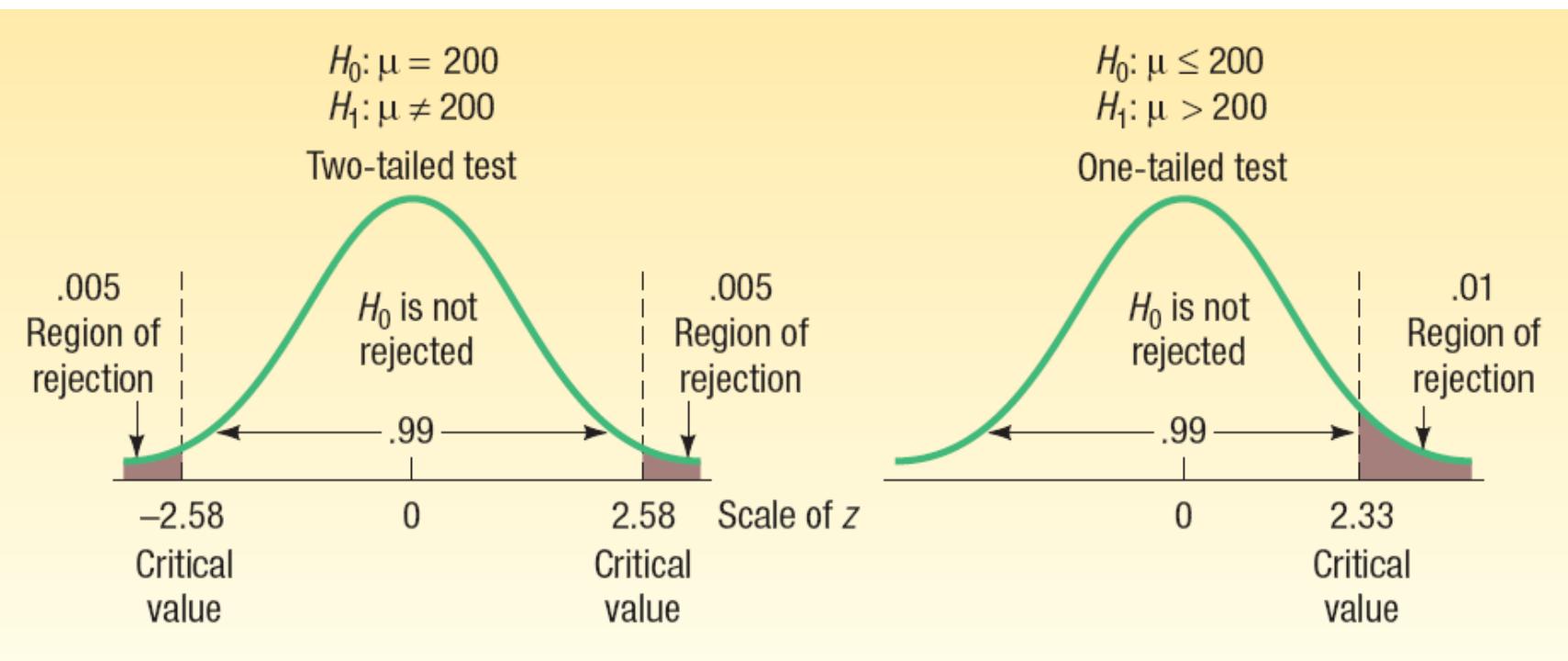
Hypothesis Test of a Population Mean, Known Population Standard Deviation – One-tail Example

Suppose in the previous problem the vice president wants to know whether the true mean is higher than 200? To put it another way, can we conclude, because of the improved production methods, that the mean number of desks assembled in the last 50 weeks was **more than 200?**

Recall: $\sigma=16$, $n=50$, $\alpha=.01$



One-Tailed Test versus Two-Tailed Test



Rejection Regions for Two-Tailed and One-Tailed Tests, $\alpha = .01$

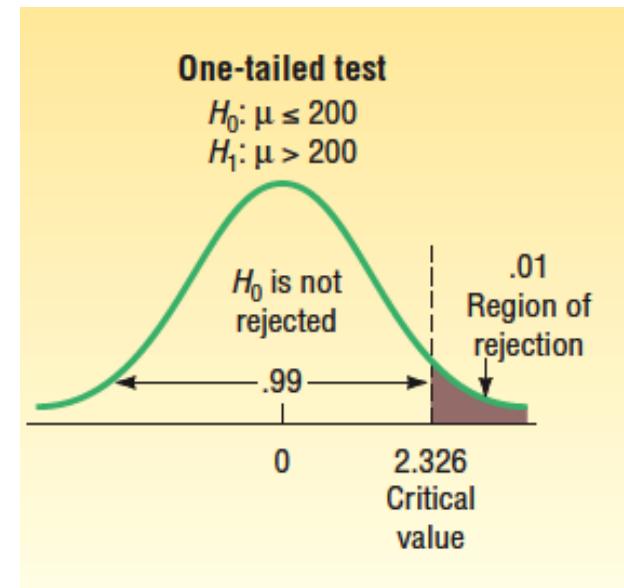
Testing for a Population Mean, Known Population Standard Deviation – One-Tail Example

Step 4: Formulate the decision rule.

Reject H_0 if $z > z_\alpha$.

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$
$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{203.5 - 200}{16/\sqrt{50}} = 1.55$$
$$z_{.01} = 2.326$$

z is not greater than $z_{.01}$.



Step 5: Make a decision. Because 1.55 does not fall in the rejection region, H_0 is not rejected.

Step 6. Interpret the result. Based on the evidence, we cannot conclude that the average number of desks assembled increased in the last 50 weeks.

p-Value in Hypothesis Testing

- A ***p*-value** is the probability of observing a sample value as extreme as, or more extreme than, the value observed (the test statistic), given that the null hypothesis is true.
- In testing a hypothesis, we can **also compare the *p*-value to the significance level (α)**.
- Decision rule using the *p*-value:

Reject H_0 if p -value < significance level.

Testing for a Population Mean, Known Population Standard Deviation – One-Tail Example

Step 1: State the null hypothesis and the alternate hypothesis.

$$H_0: \mu \leq 200$$

$$H_1: \mu > 200$$

(Note: The keyword in the problem “an *increase*.”)

Step 2: Select the level of significance.

$\alpha = 0.01$ as stated in the problem.

Step 3: Select the test statistic.

Use z -distribution since σ is known.

p-Value in Hypothesis Testing – Example

a P value is the probability of observing a 203.5 or greater weekly production in repeated experiments if the null hypothesis were true

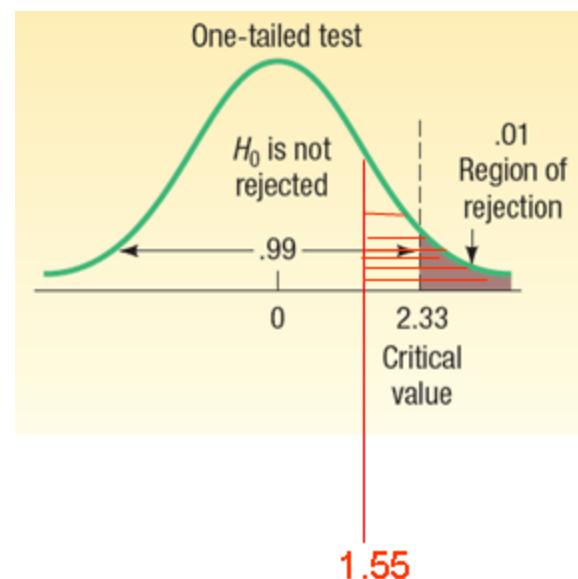
$$H_0: \mu \leq 200$$

$$H_1: \mu > 200$$

Reject H_0 if $z > z_\alpha$,
where $z = 1.55$ and $z_\alpha = 2.33$.

Reject H_0 if p -value $< \alpha$:

0.0606 is not < 0.01 .



$$\begin{aligned} P(Z > 1.55) &= .5000 - .4394 \\ P\text{-value} &= .0606 \end{aligned}$$

Conclude: Fail to reject H_0 .

Testing for the Population Mean: Population Standard Deviation Unknown

- When the population standard deviation (σ) is unknown, the sample standard deviation (s) is used in its place
- The *t*-distribution is used as the test statistic, which is computed using the formula:

TESTING A MEAN, σ UNKNOWN

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad [10-2]$$

with $n - 1$ degrees of freedom, where:

\bar{x} is the sample mean.

μ is the hypothesized population mean.

s is the sample standard deviation.

n is the number of observations in the sample.

Testing for the Population Mean: Population Standard Deviation Unknown – Example



The McFarland Insurance Company Claims Department reports the **mean** cost to process a claim is **\$60**. An industry comparison showed this amount to be larger than most other insurance companies, so the company instituted cost-cutting measures. To evaluate the effect of the cost-cutting measures, the Supervisor of the Claims Department selected a random sample of **26** claims processed last month. The sample information is reported below.

At the **.01** significance level, is it reasonable to conclude that a claim is **now less than \$60?**

| | | | | | |
|------|------|------|------|------|------|
| \$45 | \$49 | \$62 | \$40 | \$43 | \$61 |
| 48 | 53 | 67 | 63 | 78 | 64 |
| 48 | 54 | 51 | 56 | 63 | 69 |
| 58 | 51 | 58 | 59 | 56 | 57 |
| 38 | 76 | | | | |

Testing for a Population Mean: Population Standard Deviation Unknown – Example

Step 1: State the null hypothesis and the alternate hypothesis.

$$H_0: \mu \geq \$60$$

$$H_1: \mu < \$60$$

(Note: The keyword in the problem is “now **less** than.”)

Step 2: Select the level of significance.

$\alpha = 0.01$ as stated in the problem.

Step 3: Select the test statistic.

Since σ is unknown, use a t -distribution with $n-1$ ($26 - 1 = 25$) degrees of freedom.

t-Distribution Table (Portion)

TABLE 10-1 A Portion of the *t* Distribution Table

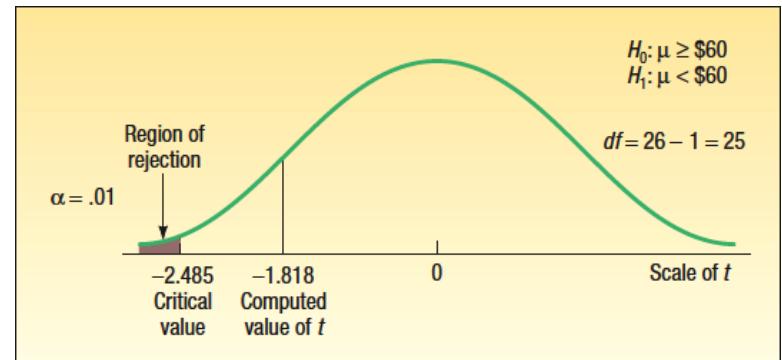
| | | Confidence Intervals | | | | | |
|-----------|-----------------------------------------------------|-----------------------------------------------------|-------|-------|-------|--------|--|
| | | Level of Significance for One-Tailed Test, α | | | | | |
| <i>df</i> | 80% | 90% | 95% | 98% | 99% | 99.9% | |
| | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.0005 | |
| | Level of Significance for Two-Tailed Test, α | | | | | | |
| | 0.20 | 0.10 | 0.05 | 0.02 | 0.01 | 0.001 | |
| | : | : | : | : | : | : | |
| 21 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.819 | |
| 22 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.792 | |
| 23 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.768 | |
| 24 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.745 | |
| 25 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.725 | |
| 26 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.707 | |
| 27 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.690 | |
| 28 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.674 | |
| 29 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.659 | |
| 30 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.646 | |

Testing for a Population Mean: Population Standard Deviation Unknown – Example

Step 4: Formulate the decision rule.

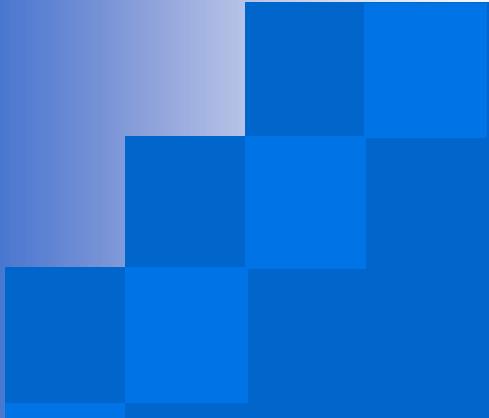
Reject H_0 if $t < -t_{\alpha, n-1}$.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{50.64 - 60}{10.02/\sqrt{26}} = -1.818$$



Step 5: Make a decision. Because -1.818 does not fall in the rejection region, H_0 is not rejected at the .01 significance level.

Step 6: Interpret the result. We have not demonstrated that the cost-cutting measures reduced the mean cost per claim to less than \$60. The difference of \$3.58 (\$56.42 - \$60) between the sample mean and the population mean could be due to sampling error.



Two-Sample Tests of Hypothesis



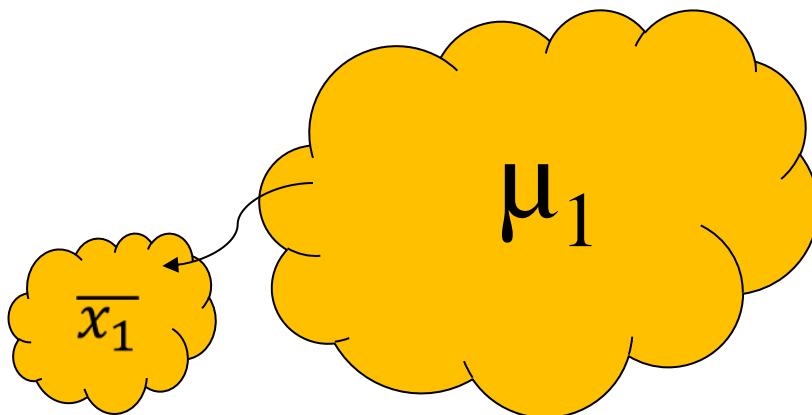
Chapter 11

Comparing Two Populations – Examples

- Is there a difference in the mean value of residential real estate sold by male agents and female agents in south Florida?
- Is there an increase in the production rate if music is piped into the production area?

Idea

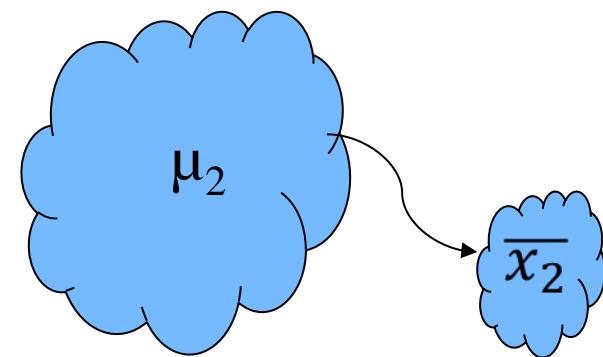
Population 1: TAMUC Univ



$$H_0: \mu_1 = \mu_2$$

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

Population 2: University X



Calculate a statistics which is based on the sample mean differences: $\bar{x}_1 - \bar{x}_2$

Hypothesis Tests for Two Population Means

Two Population Means, Independent Samples

Lower tail test:

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

$$H_A: \mu_1 < \mu_2$$

i.e.,

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

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

Two-tailed test:

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 \neq \mu_2$$

i.e.,

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

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

Upper tail test:

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

$$H_A: \mu_1 > \mu_2$$

i.e.,

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

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

Comparing Two Population Means: **Unequal** Known Population Variances

- The two populations follow normal distributions.
- The samples are from independent populations.
- The formula for computing the value of z is:

If σ_1 and σ_2 are known:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Comparing Two Population Means: Equal, Known Population Variances – Example

The Fast Lane procedure was recently installed at the local food market. The store manager would like to know if the mean checkout time using the standard checkout method is longer than using the Fast Lane procedure. She gathered the following sample information. The time is measured from when the customer enters the line until their bags are in the cart. Hence, the time includes both waiting in line and checking out.

| Customer Type | Sample Mean | Population Standard | |
|---------------|--------------|---------------------|-------------|
| | | Deviation | Sample Size |
| Standard | 5.50 minutes | 0.40 minute | 50 |
| Fast Lane | 5.30 minutes | 0.30 minute | 100 |

Interested in $\mu_S > \mu_F$



Comparing Two Population Means: Equal, Known Population Variances – Example

Applying the six-step hypothesis testing procedure:

Step 1: State the null and alternate hypotheses.

(keyword: “longer than”)

$$H_0: \mu_S \leq \mu_F$$

$$H_1: \mu_S > \mu_F$$

Step 2: Select the level of significance.

The .01 significance level is requested in the problem.

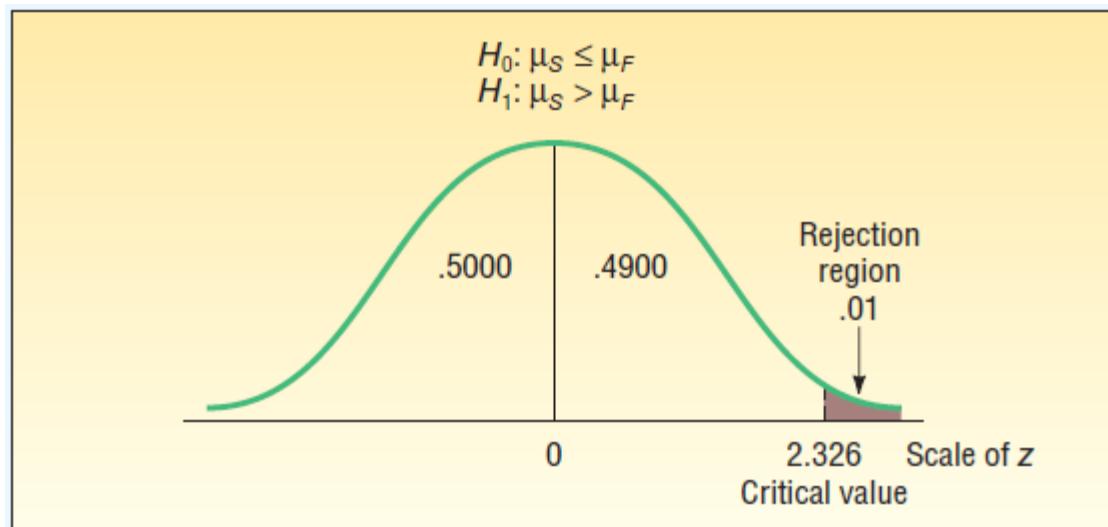
Step 3: Determine the appropriate test statistic.

Because both population standard deviations are known, we can use the *z-distribution* as the test statistic.

Comparing Two Population Means: Equal, Known Population Variances – Example

Step 4: Formulate a decision rule.

Reject H_0 if $z > z_\alpha$
 > 2.326



Comparing Two Population Means: Equal, Known Population Variances – Example

Step 5: Take a sample and make a decision.

$$\begin{aligned}
 z &= \frac{\bar{x}_s - \bar{x}_f}{\sqrt{\frac{\sigma_s^2}{n_s} + \frac{\sigma_f^2}{n_f}}} \\
 &= \frac{5.5 - 5.3}{\sqrt{\frac{0.40^2}{50} + \frac{0.30^2}{100}}} \\
 &= \frac{0.2}{0.064031} = 3.123
 \end{aligned}$$

| Customer Type | Sample Mean | Population Standard Deviation | Sample Size |
|---------------|--------------|-------------------------------|-------------|
| Standard | 5.50 minutes | 0.40 minute | 50 |
| Fast Lane | 5.30 minutes | 0.30 minute | 100 |

The computed value of 3.123 is larger than the critical value of 2.326.

Our decision is to reject the null hypothesis.

Step 6: Interpret the result. The difference of .20 minutes between the mean checkout time using the standard method is too large to have occurred by chance. We conclude the Fast Lane method is faster.

Comparing Population Means: Equal, Unknown Population Standard Deviations (The Pooled t -test)

The t distribution is used as the test statistic if one or more of the samples have less than 30 observations. The required assumptions are:

- Both populations must follow the normal distribution.
- The populations must have equal standard deviations.
- The samples are from independent populations.

Comparing Population Means: Equal, Unknown Population Standard Deviations (The Pooled *t*-test)

Finding the value of the test statistic requires two steps:

1. Pool the sample standard deviations.
2. Use the pooled standard deviation to compute the *t*-statistic.

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Comparing Population Means: Equal, Unknown Population Standard Deviations (The Pooled *t*-test) – Example

Owens Lawn Care, Inc., manufactures and assembles lawnmowers that are shipped to dealers throughout the United States and Canada. Two different procedures have been proposed for mounting the engine on the frame of the lawnmower. The question is: **Is there a difference in the mean time to mount the engines on the frames of the lawnmowers?**

The first procedure was developed by longtime Owens employee Herb Welles (designated “W”), and the other procedure was developed by Owens Vice President of Engineering William Atkins (designated “A”). To evaluate the two methods, it was decided to conduct a time and motion study.

A sample of five employees was timed using the Welles method and six using the Atkins method. The results, in minutes, are shown on the right.

Is there a difference in the mean mounting times? Use the .10 significance level.

| Welles (minutes) | Atkins (minutes) |
|---------------------|---------------------|
| 2 | 3 |
| 4 | 7 |
| 9 | 5 |
| 3 | 8 |
| 2 | 4 |
| | 3 |

Comparing Population Means: Equal, Unknown Population Standard Deviations (The Pooled *t*-test) – Example

Step 1: State the null and alternate hypotheses.

(Keyword: “Is there a *difference*”)

$$H_0: \mu_W = \mu_A$$

$$H_1: \mu_W \neq \mu_A$$

Step 2: State the level of significance.

The 0.10 significance level is stated in the problem.

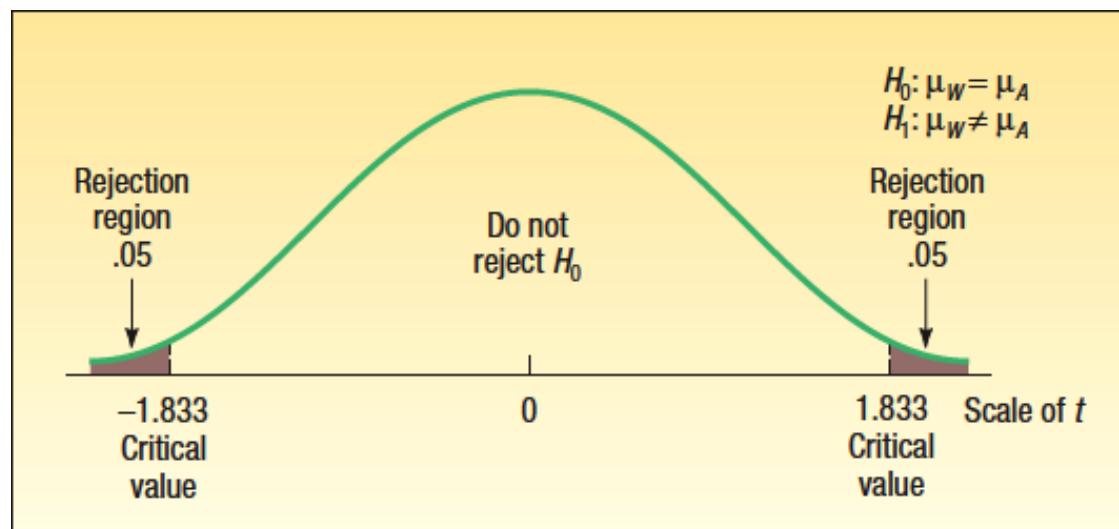
Step 3: Select the appropriate test statistic.

Because the population standard deviations are not known but are assumed to be equal, we use the pooled *t*-test.

Comparing Population Means: Equal, Unknown Population Standard Deviations (The Pooled *t*-test) – Example

Step 4: State the decision rule.

Reject H_0 if $t > t_{\alpha/2, n_W+n_A-2}$ or $t < -t_{\alpha/2, n_W+n_A-2}$
 $t > t_{.05, 9}$ or $t < -t_{.05, 9}$
 $t > 1.833$ or $t < -1.833$



Comparing Population Means: Equal, Unknown Population Standard Deviations (The Pooled *t*-test) – Example

Step 5: Compute the value of *t* and make a decision.

(a) Calculate the sample standard deviations.

| Welles Method | | Atkins Method | |
|---------------|-----------------------|---------------|-----------------------|
| x_W | $(x_W - \bar{x}_W)^2$ | x_A | $(x_A - \bar{x}_A)^2$ |
| 2 | $(2 - 4)^2 = 4$ | 3 | $(3 - 5)^2 = 4$ |
| 4 | $(4 - 4)^2 = 0$ | 7 | $(7 - 5)^2 = 4$ |
| 9 | $(9 - 4)^2 = 25$ | 5 | $(5 - 5)^2 = 0$ |
| 3 | $(3 - 4)^2 = 1$ | 8 | $(8 - 5)^2 = 9$ |
| 2 | $(2 - 4)^2 = 4$ | 4 | $(4 - 5)^2 = 1$ |
| 20 | $\overline{34}$ | 30 | $\overline{22}$ |

$$\bar{x}_W = \frac{\sum x_W}{n_W} = \frac{20}{5} = 4$$

$$\bar{x}_A = \frac{\sum x_A}{n_A} = \frac{30}{6} = 5$$

$$s_W = \sqrt{\frac{\sum (x_W - \bar{x}_W)^2}{n_W - 1}} = \sqrt{\frac{34}{5 - 1}} = 2.9155 \quad s_A = \sqrt{\frac{\sum (x_A - \bar{x}_A)^2}{n_A - 1}} = \sqrt{\frac{22}{6 - 1}} = 2.0976$$

(b) Calculate the **pooled** sample standard deviation.

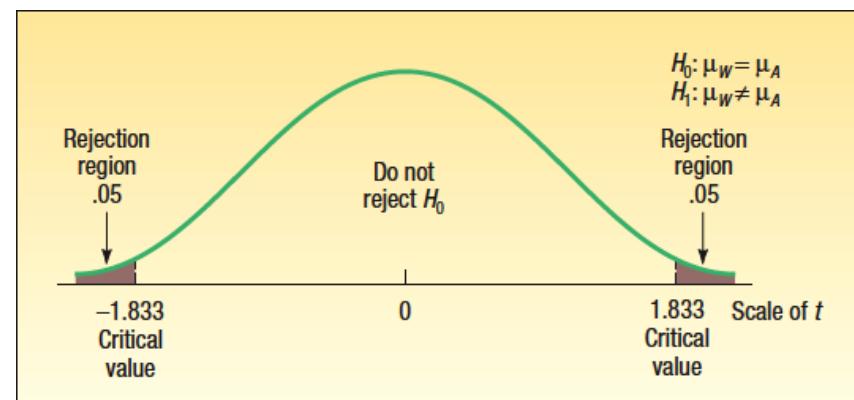
$$s_p^2 = \frac{(n_W - 1)s_W^2 + (n_A - 1)s_A^2}{n_W + n_A - 2} = \frac{(5 - 1)(2.9155)^2 + (6 - 1)(2.0976)^2}{5 + 6 - 2} = 6.2222$$

Comparing Population Means: Equal, Unknown Population Standard Deviations (The Pooled *t*-test) – Example

Step 5 (continued): Take a sample and make a decision.

(c) Determine the value of *t*.

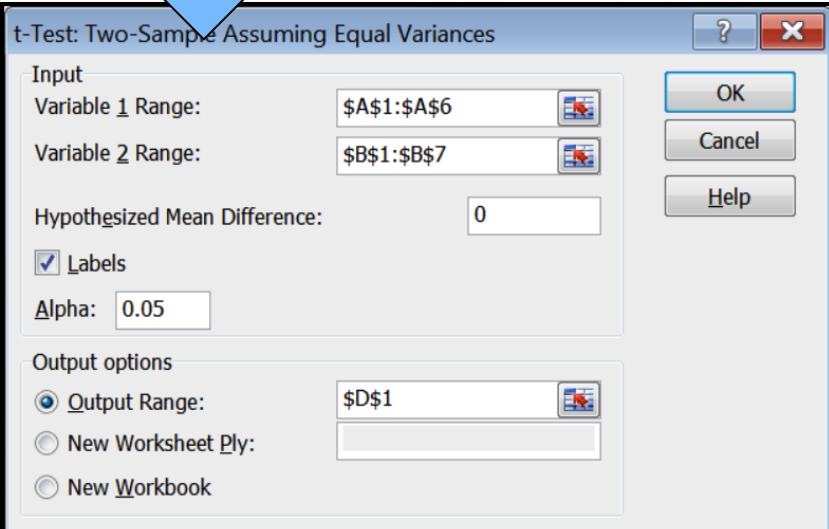
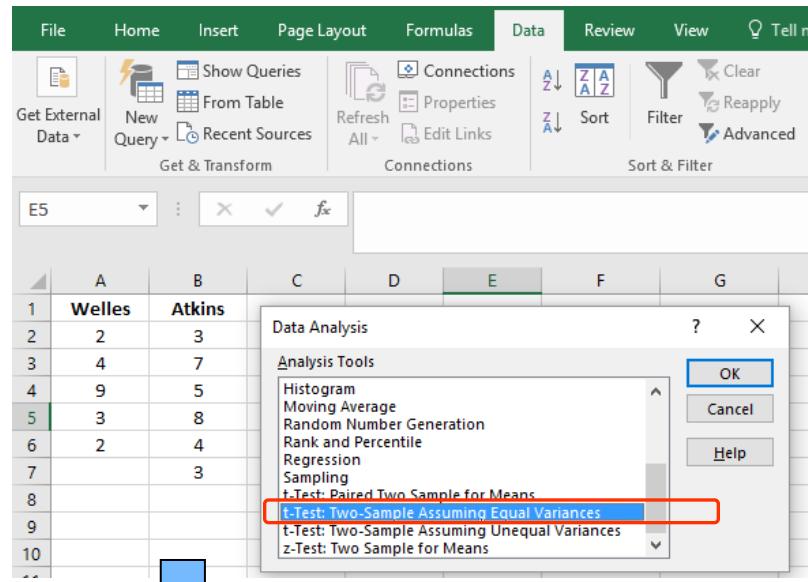
$$t = \frac{\bar{x}_W - \bar{x}_A}{\sqrt{s_p^2 \left(\frac{1}{n_W} + \frac{1}{n_A} \right)}} = \frac{4.00 - 5.00}{\sqrt{6.2222 \left(\frac{1}{5} + \frac{1}{6} \right)}} = -0.662$$



The decision is not to reject the null hypothesis because – 0.662 falls in the region between -1.833 and 1.833.

Step 6: Interpret the Result. The data show no evidence that there is a difference in the mean times to mount the engine on the frame between the Welles and Atkins methods.

Excel Example (Equal Variance)



| | A | B | C | D | E | F |
|----|--------|--------|---|---------------------------------------------|------------------------------|-------------|
| 1 | Welles | Atkins | | t-Test: Two-Sample Assuming Equal Variances | | |
| 2 | 2 | 3 | | | Welles | Atkins |
| 3 | 4 | 7 | | | Mean | 4.000 5.000 |
| 4 | 9 | 5 | | | Variance | 8.500 4.400 |
| 5 | 3 | 8 | | | Observations | 5.000 6.000 |
| 6 | 2 | 4 | | | Pooled Variance | 6.222 |
| 7 | | 3 | | | Hypothesized Mean Difference | 0.000 |
| 8 | | | | | df | 9.000 |
| 9 | | | | | t Stat | -0.662 |
| 10 | | | | | P(T<=t) one-tail | 0.262 |
| 11 | | | | | t Critical one-tail | 1.833 |
| 12 | | | | | P(T<=t) two-tail | 0.525 |
| 13 | | | | | t Critical two-tail | 2.262 |
| 14 | | | | | | |

Comparing Population Means with Unknown AND Unequal Population Standard Deviations

Use the formula for the t -statistic shown if it is not reasonable to assume the population standard deviations are equal.

TEST STATISTIC FOR NO DIFFERENCE
IN MEANS, UNEQUAL VARIANCES

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad [11-5]$$

The degrees of freedom are adjusted downward by a rather complex approximation formula. The effect is to reduce the number of degrees of freedom in the test, which will require a larger value of the test statistic to reject the null hypothesis.

DEGREES OF FREEDOM FOR
UNEQUAL VARIANCE TEST

$$df = \frac{[(s_1^2/n_1) + (s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \quad [11-6]$$

Excel Example (Unequal Variance)

Screenshot of Microsoft Excel showing a t-test analysis for unequal variances.

The Data tab is selected in the ribbon. A screenshot of the Data Analysis dialog box shows the "t-Test: Two-Sample Assuming Unequal Variances" option highlighted.

A blue arrow points from the "t-Test: Two-Sample Assuming Unequal Variances" option in the Data Analysis dialog to the "t-Test: Two-Sample Assuming Unequal Variances" dialog box.

The "t-Test: Two-Sample Assuming Unequal Variances" dialog box displays the following input parameters:

- Input
- Variable 1 Range: \$A\$1:\$A\$6
- Variable 2 Range: \$B\$1:\$B\$7
- Hypothesized Mean Difference: 0
- Labels (checkbox checked)
- Alpha: 0.05
- Output options:
 - Output Range: \$D\$1 (radio button selected)
 - New Worksheet Ply: (radio button unselected)
 - New Workbook: (radio button unselected)

A large blue arrow points from the "t-Test: Two-Sample Assuming Unequal Variances" dialog box to the resulting output table in the worksheet.

The output table in the worksheet includes the following results:

| | Variable 1 | Variable 2 |
|------------------------------|--------------|------------|
| Mean | 4 | 5 |
| Variance | 8.5 | 4.4 |
| Observations | 5 | 6 |
| Hypothesized Mean Difference | 0 | |
| df | 7 | |
| t Stat | -0.641060765 | |
| P(T<=t) one-tail | 0.270947096 | |
| t Critical one-tail | 1.894578605 | |
| P(T<=t) two-tail | 0.541894193 | |
| t Critical two-tail | 2.364624252 | |

Comparing Population Means: Hypothesis Testing with Dependent Samples

Dependent samples are samples that are paired or related in some fashion.

For example:

- If you wished to buy a car, you would look at the *same* car at two (or more) *different* dealerships and compare the prices.
- If you wished to measure the effectiveness of a new diet you would weigh the dieters at the start and at the finish of the program.



Comparing Population Means: Hypothesis Testing with Dependent Samples

Use the following test when the samples are **dependent**:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

Where

\bar{d} is the mean of the differences

s_d is the standard deviation of the differences

n is the number of pairs (differences)

Comparing Population Means: Hypothesis Testing with Dependent Samples – Example



Nickel Savings and Loan wishes to compare the two companies, Schadek and Bowyer, it uses to appraise the value of residential homes. Nickel Savings selected a sample of 10 residential properties and scheduled both firms for an appraisal. The results, reported in \$000, are shown in the table (right).

| Home | Schadek | Bowyer |
|------|---------|--------|
| 1 | 235 | 228 |
| 2 | 210 | 205 |
| 3 | 231 | 219 |
| 4 | 242 | 240 |
| 5 | 205 | 198 |
| 6 | 230 | 223 |
| 7 | 231 | 227 |
| 8 | 210 | 215 |
| 9 | 225 | 222 |
| 10 | 249 | 245 |

At the .05 significance level, can we conclude there is a difference in the mean appraised values of the homes?

Comparing Population Means: Hypothesis Testing with Dependent Samples – Example

Step 1: State the null and alternate hypotheses.

$$H_0: \mu_d = 0$$

$$H_1: \mu_d \neq 0$$

Step 2: State the level of significance.

The .05 significance level is stated in the problem.

Step 3: Select the appropriate test statistic.

To test the difference between two population means with dependent samples, we use the *t*-statistic.

Comparing Population Means: Hypothesis Testing with Dependent Samples – Example

Step 4: State the decision rule.

Reject H_0 if

$$t > t_{\alpha/2, n-1} \text{ or } t < -t_{\alpha/2, n-1}$$

$$t > t_{.025, 9} \text{ or } t < -t_{.025, 9}$$

$$t > 2.262 \text{ or } t < -2.262$$

| df | Confidence Intervals | | | | | <i>p</i> -value between 0.01 and 0.001 | |
|-------------------------------------------|-------------------------------------------|-------|--------|--------|--------|-------------------------------------------------|--|
| | Level of Significance for One-Tailed Test | | | | | | |
| | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | | |
| Level of Significance for Two-Tailed Test | | | | | | | |
| | 0.20 | 0.10 | 0.05 | 0.02 | 0.01 | 0.001 | |
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 636.619 | |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 31.599 | |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 12.924 | |
| Critical <i>t</i> -statistic for 0.05 | | 2.132 | 2.776 | 3.747 | 4.604 | 8.610 | |
| | | 2.015 | 2.571 | 3.365 | 4.032 | 6.869 | |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.959 | |
| 7 | 1.415 | 1.995 | 2.365 | 2.998 | 3.499 | 5.408 | |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 5.041 | |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.781 | |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.587 | |

Comparing Population Means: Hypothesis Testing with Dependent Samples – Example

Step 5: Take a sample and make a decision.

| Home | Schadek | Bowyer | Difference, d | $(d - \bar{d})$ | $(d - \bar{d})^2$ |
|------|---------|--------|-----------------|-----------------|-------------------|
| 1 | 235 | 228 | 7 | 2.4 | 5.76 |
| 2 | 210 | 205 | 5 | 0.4 | 0.16 |
| 3 | 231 | 219 | 12 | 7.4 | 54.76 |
| 4 | 242 | 240 | 2 | -2.6 | 6.76 |
| 5 | 205 | 198 | 7 | 2.4 | 5.76 |
| 6 | 230 | 223 | 7 | 2.4 | 5.76 |
| 7 | 231 | 227 | 4 | -0.6 | 0.36 |
| 8 | 210 | 215 | -5 | -9.6 | 92.16 |
| 9 | 225 | 222 | 3 | -1.6 | 2.56 |
| 10 | 249 | 245 | 4 | -0.6 | 0.36 |
| | | | 46 | 0 | 174.40 |

The computed value of t , 3.305, is greater than the higher critical value, 2.262, so our decision is to reject the null hypothesis.

$$\bar{d} = \frac{\sum d}{n} = \frac{46}{10} = 4.60$$

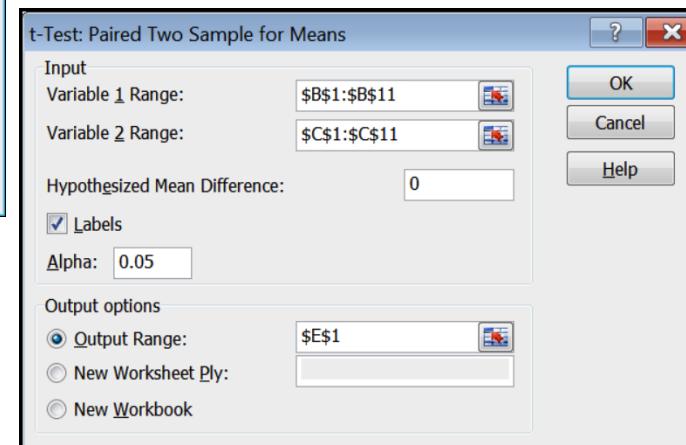
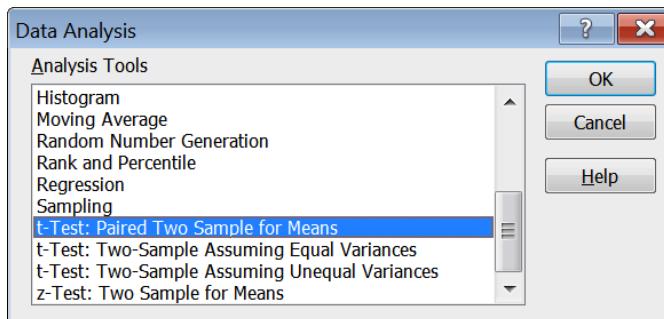
$$s_d = \sqrt{\frac{\sum (d - \bar{d})^2}{n - 1}} = \sqrt{\frac{174.4}{10 - 1}} = 4.402$$

Using formula (11–7), the value of the test statistic is 3.305, found by

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} = \frac{4.6}{4.402 / \sqrt{10}} = \frac{4.6}{1.3920} = 3.305$$

Step 6: Interpret the result. The data indicate that there is a significant statistical difference in the property appraisals from the two firms. We would hope that appraisals of a property would be similar.

Comparing Population Means: Hypothesis Testing with Dependent Samples – Excel Example



| t-Test: Paired Two Sample for Means | | |
|-------------------------------------|-------------|-------------|
| | Variable 1 | Variable 2 |
| Mean | 226.8 | 222.2 |
| Variance | 208.8444444 | 204.1777778 |
| Observations | 10 | 10 |
| Pearson Correlation | 0.953143808 | |
| Hypothesized Mean Difference | 0 | |
| df | 9 | |
| t Stat | 3.304500684 | |
| P(T<=t) one-tail | 0.00458195 | |
| t Critical one-tail | 1.833112933 | |
| P(T<=t) two-tail | 0.0091639 | |
| t Critical two-tail | 2.262157163 | |

Analysis of Variance

Chapter 12



Learning Objectives

- LO1** List the characteristics of the F distribution and locate values in an F table.
- LO2** Perform a test of hypothesis to determine whether the variances of two populations are equal.
- LO3** Describe the ANOVA approach for testing difference in sample means.
- LO4** Organize data into ANOVA tables for analysis.
- LO5** Conduct a test of hypothesis among three or more treatment means and describe the results.
- LO6** Develop confidence intervals for the difference in treatment means and interpret the results.
- LO7** Carry out a test of hypothesis among treatment means using a blocking variable and understand the results.
- LO8** Perform a two-way ANOVA with interaction and describe the results.

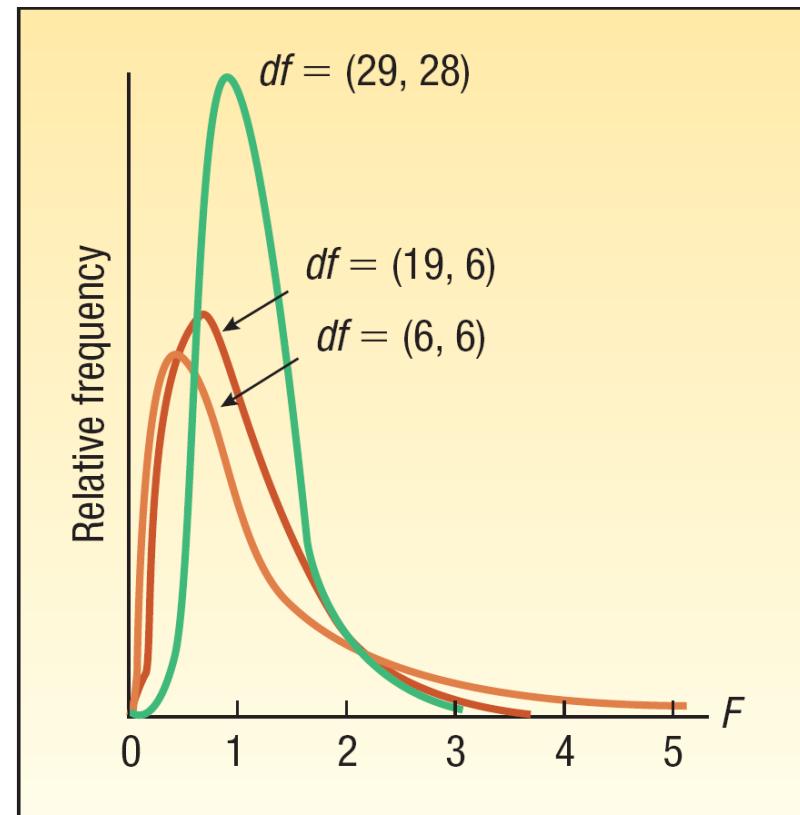
LO1 List the characteristics of the *F* distribution and locate values in an F table.

The *F* Distribution

- It was named to honor Sir Ronald Fisher, one of the founders of modern-day statistics.
- It is
 - used to test whether two samples are from populations having equal variances
 - applied when we want to compare several population means simultaneously. The simultaneous comparison of several population means is called analysis of variance(ANOVA).
 - In both of these situations, the populations must follow a normal distribution, and the data must be at least interval-scale.

Characteristics of F -Distribution

1. There is a “family” of F Distributions. A particular member of the family is determined by two parameters: the degrees of freedom in the numerator and the degrees of freedom in the denominator.
2. The F distribution is continuous
3. F value cannot be negative.
4. The F distribution is positively skewed.
5. It is asymptotic. As $F \rightarrow \infty$ the curve approaches the X -axis but never touches it.



LO2 Perform a test of hypothesis to determine whether the variances of two populations are equal.

Comparing Two Population Variances

The *F* distribution is used to test the hypothesis that the variance of one normal population equals the variance of another normal population.

Examples:

- Two Barth shearing machines are set to produce steel bars of the same length. The bars, therefore, should have the same mean length. We want to ensure that in addition to having the same mean length they also have similar variation.
- The mean rate of return on two types of common stock may be the same, but there may be more variation in the rate of return in one than the other. A sample of 10 technology and 10 utility stocks shows the same mean rate of return, but there is likely more variation in the Internet stocks.
- A study by the marketing department for a large newspaper found that men and women spent about the same amount of time per day reading the paper. However, the same report indicated there was nearly twice as much variation in time spent per day among the men than the women.

Test for Equal Variances

$$\begin{aligned} H_0: \sigma_1^2 &= \sigma_2^2 \\ H_1: \sigma_1^2 &\neq \sigma_2^2 \end{aligned}$$

To conduct the test, we select a random sample of n_1 observations from one population, and a random sample of n_2 observations from the second population. The test statistic is defined as follows.

**TEST STATISTIC FOR COMPARING
TWO VARIANCES**

$$F = \frac{s_1^2}{s_2^2}$$

[12-1]

Test for Equal Variances - Example



Lammers Limos offers limousine service from the city hall in Toledo, Ohio, to Metro Airport in Detroit. The president of the company, is considering two routes. One is via U.S. 25 and the other via I-75. He wants to study the time it takes to drive to the airport using each route and then compare the results. He collected the following sample data, which is reported in minutes.

Using the .10 significance level, **is there a difference in the variation** in the driving times for the two routes?

| U.S. Route 25 | Interstate 75 |
|---------------|---------------|
| 52 | 59 |
| 67 | 60 |
| 56 | 61 |
| 45 | 51 |
| 70 | 56 |
| 54 | 63 |
| 64 | 57 |
| | 65 |

Test for Equal Variances - Example

Step 1: The hypotheses are:

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

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

Step 2: The significance level is .10.

Step 3: The test statistic is the *F* distribution.

Test for Equal Variances - Example

Step 4: State the decision rule.

Reject H_0 if

$$F > F_{\alpha/2, v_1, v_2}$$

$$F > F_{.10/2, 7-1, 8-1}$$

$$F > F_{.05, 6, 7}$$

TABLE 12–1 Critical Values of the F Distribution, $\alpha = .05$

| Degrees of Freedom for Denominator | Degrees of Freedom for Numerator | | | |
|------------------------------------|----------------------------------|------|------|------|
| | 5 | 6 | 7 | 8 |
| 1 | 230 | 234 | 237 | 239 |
| 2 | 19.3 | 19.3 | 19.4 | 19.4 |
| 3 | 9.01 | 8.94 | 8.89 | 8.85 |
| 4 | 6.26 | 6.16 | 6.09 | 6.04 |
| 5 | 5.05 | 4.95 | 4.88 | 4.82 |
| 6 | 4.39 | 4.28 | 4.21 | 4.15 |
| 7 | 3.97 | 3.87 | 3.79 | 3.73 |
| 8 | 3.69 | 3.58 | 3.50 | 3.44 |
| 9 | 3.48 | 3.37 | 3.29 | 3.23 |
| 10 | 3.33 | 3.22 | 3.14 | 3.07 |

Test for Equal Variances - Example

Step 5: Compute the value of F and make a decision

U.S. Route 25

$$\bar{X} = \frac{\Sigma X}{n} = \frac{408}{7} = 58.29 \quad s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}} = \sqrt{\frac{485.43}{7 - 1}} = 8.9947$$

Interstate 75

$$\bar{X} = \frac{\Sigma X}{n} = \frac{472}{8} = 59.00 \quad s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}} = \sqrt{\frac{134}{8 - 1}} = 4.3753$$

$$F = \frac{s_1^2}{s_2^2} = \frac{(8.9947)^2}{(4.3753)^2} = 4.23$$

The decision is to **reject the null hypothesis**, because the computed F value (4.23) is larger than the critical value (3.87).

We conclude that **there is a difference in the variation** of the travel times along the two routes.

Test for Equal Variances – Excel Example

The screenshot shows a Microsoft Excel interface. At the top, there is a blue ribbon tab labeled "Data Analysis". Below the ribbon, a "Data Analysis" dialog box is open, showing a list of analysis tools. The tool "F-Test Two-Sample for Variances" is highlighted with a blue selection bar. To the right of the dialog box are buttons for "OK", "Cancel", and "Help".

The main Excel window displays a worksheet titled "num 1 variance test". The worksheet contains two sets of data in columns A and B, labeled "U. S. 25" and "Interstate 75". To the right of these data sets, a summary table for the "F-Test Two-Sample for Variances" is shown. The summary table includes the following data:

| | A | B | D | E | F | G |
|----|----------|---------------|---|---------------------------------|----------|---------------|
| 1 | U. S. 25 | Interstate 75 | | F-Test Two-Sample for Variances | | |
| 2 | 52 | 59 | | | U. S. 25 | Interstate 75 |
| 3 | 67 | 60 | | Mean | 58.29 | 59.00 |
| 4 | 56 | 61 | | Variance | 80.90 | 19.14 |
| 5 | 45 | 51 | | Observations | 7.00 | 8.00 |
| 6 | 70 | 56 | | df | 6.00 | 7.00 |
| 7 | 54 | 63 | | F | 4.23 | |
| 8 | 64 | 57 | | P(F<=f) one-tail | 0.04 | |
| 9 | | 65 | | F Critical one-tail | 3.87 | |
| 10 | | | | | | |
| 11 | | | | | | |

LO3 Describe the ANOVA approach for testing difference in sample means.

Comparing Means of Two or More Populations

The F distribution is also used for testing whether two or more sample means came from the same or equal populations.

Assumptions:

- The sampled populations follow the normal distribution.
- The populations have equal standard deviations.
- The samples are randomly selected and are independent.

Comparing Means of Two or More Populations

- The **Null Hypothesis** is that the population means are all the same. The **Alternative Hypothesis** is that at least one of the means is different.
- The **Test Statistic** is the F distribution.
- The **Decision rule** is to reject the null hypothesis if F (computed) is greater than F (table) with numerator and denominator degrees of freedom.
- Hypothesis Setup and Decision Rule:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

H_1 : The means are not all equal

Reject H_0 if $F > F_{\alpha, k-1, n-k}$

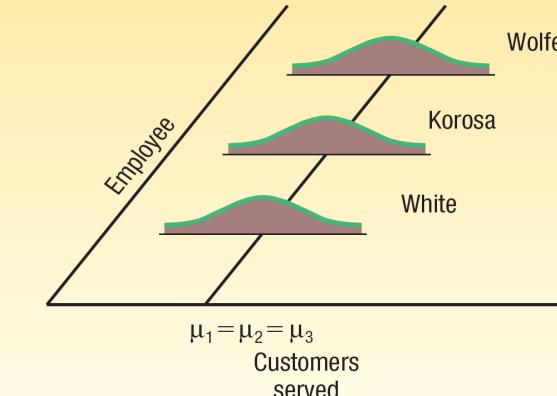
Analysis of Variance – F statistic

- If there are k populations being sampled, the numerator degrees of freedom is $k - 1$.
- If there are a total of n observations the denominator degrees of freedom is $n - k$.
- The test statistic is computed by:

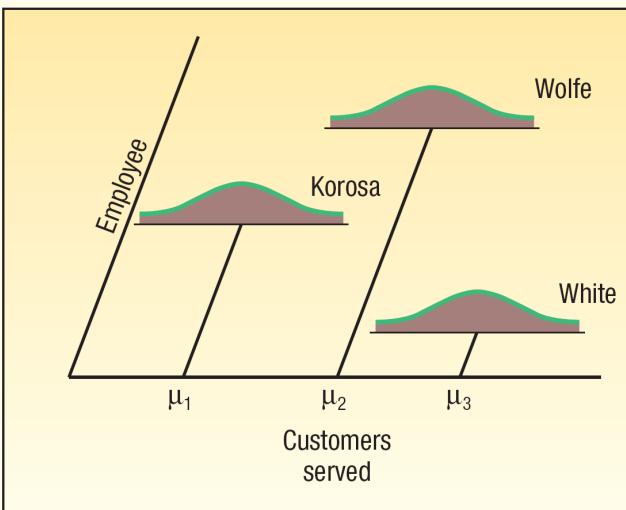
$$F = \frac{SST/(k-1)}{SSE/(n-k)}$$

Comparing Means of Two or More Populations – Illustrative Example

Joyce Kuhlman manages a regional financial center. She wishes to compare the productivity, as measured by the number of customers served, among three employees. Four days are randomly selected and the number of customers served by each employee is recorded.



Case Where Treatment Means Are the Same



Case Where Treatment Means Are Different

Comparing Means of Two or More Populations – Example

Recently a group of four major carriers joined in hiring Brunner Marketing Research, Inc., to survey recent passengers regarding their level of satisfaction with a recent flight. The survey included questions on ticketing, boarding, in-flight service, baggage handling, pilot communication, and so forth.

Twenty-five questions offered a range of possible answers: excellent, good, fair, or poor. A response of excellent was given a score of 4, good a 3, fair a 2, and poor a 1. These responses were then totaled, so the total score was an indication of the satisfaction with the flight. Brunner Marketing Research, Inc., randomly selected and surveyed passengers from the four airlines.

| American | Delta | United | US Airways |
|----------|-------|--------|------------|
| 94 | 75 | 70 | 68 |
| 90 | 68 | 73 | 70 |
| 85 | 77 | 76 | 72 |
| 80 | 83 | 78 | 65 |
| | 88 | 80 | 74 |
| | | 68 | 65 |
| | | | 65 |

Is there a **difference** in the **mean** satisfaction level among the four airlines? Use the .01 significance level.

Comparing Means of Two or More Populations – Example

Step 1: State the null and alternate hypotheses.

$$H_0: \mu_A = \mu_D = \mu_U = \mu_{US}$$

$H_1:$ The means are not all equal

Reject H_0 if $F > F_{\alpha, k-1, n-k}$

Step 2: State the level of significance.

The .01 significance level is stated in the problem.

Step 3: Find the appropriate test statistic.

Because we are comparing means of more than two groups, use the F statistic

Comparing Means of Two or More Populations – Example

Step 4: State the decision rule.

Reject H_0 if $F > F_{\alpha, k-1, n-k}$

$$F > F_{.01, 4-1, 22-4}$$

$$F > F_{.01, 3, 18}$$

$$F > 5.09$$

Comparing Means of Two or More Populations – Example

Step 5: Compute the value of F and make a decision

| ANOVA Table | | | | |
|---------------------|----------------|--------------------|---------------------|---------|
| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | F |
| Treatments | SST | $k - 1$ | $SST/(k - 1) = MST$ | MST/MSE |
| Error | SSE | $n - k$ | $SSE/(n - k) = MSE$ | |
| Total | SS total | $n - 1$ | | |

$$\text{SS total} = \sum(X - \bar{X}_G)^2$$

where:

X is each sample observation.

\bar{X}_G is the overall or grand mean.

$$\text{SSE} = \sum(X - \bar{X}_c)^2$$

where:

\bar{X}_c is the sample mean for treatment c .

Comparing Means of Two or More Populations – Example

$$\bar{X}_G = \frac{1,664}{22} = 75.64$$

| American | Delta | United | US Airways | Total |
|----------|-------|--------|------------|-------------|
| 94 | 75 | 70 | 68 | |
| 90 | 68 | 73 | 70 | |
| 85 | 77 | 76 | 72 | |
| 80 | 83 | 78 | 65 | |
| | 88 | 80 | 74 | |
| | | 68 | 65 | |
| | | 65 | | |
| Column | | | | |
| total | 349 | 391 | 510 | 414 1,664 |
| n | 4 | 5 | 7 | 6 22 |
| Mean | 87.25 | 78.20 | 72.86 | 69.00 75.64 |

Computing SS Total and SSE

$$(X - \bar{X}_G)^2$$

| Eastern | TWA | Allegheny | Ozark |
|----------|-------|-----------|------------|
| American | Delta | United | US Airways |
| 14.36 | -7.64 | -2.64 | -5.64 |
| 9.36 | 1.36 | 0.36 | -3.64 |
| 4.36 | 7.36 | 2.36 | -10.64 |
| | 12.36 | 4.36 | -1.64 |
| | | -7.64 | -10.64 |
| | | | -10.64 |

$$(X - \bar{X}_G)^2$$

| Eastern | TWA | Allegheny | Ozark | Total |
|----------|--------|-----------|------------|----------|
| American | Delta | United | US Airways | |
| 206.21 | 58.37 | 6.97 | 51.81 | |
| 87.61 | 1.85 | 0.13 | 13.25 | |
| 19.0 | 54.17 | 5.57 | 113.21 | |
| | 152.77 | 19.01 | 2.69 | |
| | | 58.37 | 113.21 | |
| | | | 113.21 | |
| Total | 649.91 | 267.57 | 235.07 | 332.54 |
| | | | | 1,485.09 |

$$\text{SS total} = \sum(X - \bar{X}_G)^2$$

$$(X - \bar{X}_c)^2$$

| Eastern | TWA | Allegheny | Ozark |
|----------|-------|-----------|------------|
| American | Delta | United | US Airways |
| 2.75 | -10.2 | 0.14 | 1 |
| -2.25 | -1.2 | 3.14 | 3 |
| -7.25 | 4.8 | 5.14 | -4 |
| | 9.8 | 7.14 | 5 |
| | | -4.86 | -4 |
| | | | -7.86 |

$$(X - \bar{X}_c)^2$$

| Eastern | TWA | Allegheny | Ozark | Total |
|----------|----------|-----------|------------|--------|
| American | Delta | United | US Airways | |
| 7.5625 | 104.04 | 0.02 | 1 | |
| 5.0625 | 1.44 | 9.86 | 9 | |
| 52.5625 | 23.04 | 26.42 | 16 | |
| | 96.04 | 50.98 | 25 | |
| | | 23.62 | 16 | |
| | | | 61.78 | |
| Total | 110.7500 | 234.80 | 180.86 | 68 |
| | | | | 594.41 |

$$\text{SSE} = \sum(X - \bar{X}_c)^2$$

Computing SST

$$\text{SST} = \text{SS total} - \text{SSE} = 1,485.09 - 594.41 = 890.68.$$

| ANOVA Table | | | | |
|---------------------|----------------|--------------------|-----------------------------------|---------|
| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | F |
| Treatments | SST | $k - 1$ | $\text{SST}/(k - 1) = \text{MST}$ | MST/MSE |
| Error | SSE | $n - k$ | $\text{SSE}/(n - k) = \text{MSE}$ | |
| Total | SS total | $n - 1$ | | |

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | F |
|---------------------|----------------|--------------------|-------------|------|
| Treatments | 890.68 | 3 | 296.89 | 8.99 |
| Error | 594.41 | 18 | 33.02 | |
| Total | 1,485.09 | 21 | | |

The computed value of F is 8.99, which is greater than the critical value of 5.09, so the **null hypothesis is rejected**.

Conclusion: The population means are not all equal. The mean scores are **not the same** for the four airlines; at this point we can **only conclude there is a difference in the treatment means**. We cannot determine which treatment groups differ or how many treatment groups differ.

Excel

TextbookData.xlsx - Microsoft Excel

Home Insert Page Layout Formulas Data Review View Acrobat

Cut Copy Format Painter Paste Clipboard

Font Alignment Number

G22 f_x

| | A | B | C | D | E | F |
|----|----------------------|-------------|--------|-------------|-------------|-------------------|
| 1 | | | | | | |
| 2 | American | Delta | United | US Airways | | |
| 3 | 94 | 75 | 70 | 68 | | |
| 4 | 90 | 68 | 73 | 70 | | |
| 5 | 85 | 77 | 76 | 72 | | |
| 6 | 80 | 83 | 78 | 65 | | |
| 7 | | 88 | 80 | 74 | | |
| 8 | | | 68 | 65 | | |
| 9 | | | 65 | | | |
| 10 | | | | | | |
| 11 | | | | | | |
| 12 | | | | | | |
| 13 | Anova: Single Factor | | | | | |
| 14 | | | | | | |
| 15 | SUMMARY | | | | | |
| 16 | Groups | Count | Sum | Average | Variance | |
| 17 | American | 4 | 349 | 87.250 | 36.917 | |
| 18 | Delta | 5 | 391 | 78.200 | 58.700 | |
| 19 | United | 7 | 510 | 72.857 | 30.143 | |
| 20 | US Airways | 6 | 414 | 69.000 | 13.600 | |
| 21 | | | | | | |
| 22 | | | | | | |
| 23 | ANOVA | | | | | |
| 24 | Source of Variation | SS | df | MS | F | P-value F crit |
| 25 | Between Groups | 890.6837662 | 3 | 296.8945887 | 8.990643302 | 0.000743 3.159908 |
| 26 | Within Groups | 594.4071429 | 18 | 33.02261905 | | |
| 27 | Total | 1485.090909 | 21 | | | |
| 28 | | | | | | |
| 29 | | | | | | |
| 30 | | | | | | |

Data Analysis

Analysis Tools

- Anova: Single Factor
- Anova: Two-Factor With Replication
- Anova: Two-Factor Without Replication
- Correlation
- Covariance
- Descriptive Statistics
- Exponential Smoothing
- F-Test Two-Sample for Variances
- Fourier Analysis
- Histogram

OK Cancel Help

LO6 Develop confidence intervals for the difference in treatment means and interpret the results.

Confidence Interval for the Difference Between Two Means

When we reject the null hypothesis that the means are equal, we may want to know which treatment means differ. One of the simplest procedures is through the use of confidence intervals.

$$(\bar{X}_1 - \bar{X}_2) \pm t \sqrt{MSE \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

where

\bar{X}_1 is the mean of the first sample.

\bar{X}_2 is the mean of the second sample.

t is obtained from Appendix B.2. The degrees of freedom is equal to $n - k$.

MSE is the mean square error term obtained from the ANOVA table [$SSE/(n - k)$].

n_1 is the number of observations in the first sample.

n_2 is the number of observations in the second sample.

Confidence Interval for the Difference Between Two Means - Example

From the previous example, develop a 95% confidence interval for the difference in the mean between American and US Airways. Can we conclude that there is a difference between the two airlines' ratings?

$$\begin{aligned}(\bar{X}_E - \bar{X}_O) \pm t\sqrt{\text{MSE}\left(\frac{1}{n_E} + \frac{1}{n_O}\right)} &= (87.25 - 69.00) \pm 2.101\sqrt{33.0\left(\frac{1}{4} + \frac{1}{6}\right)} \\ &= 18.25 \pm 7.79\end{aligned}$$

The 95 percent confidence interval ranges from 10.46 up to 26.04. Both endpoints are positive; hence, we can conclude these treatment means differ significantly. That is, passengers on American rated service significantly different from those on United Airways.

Confidence Interval for the Difference Between Two Means – Minitab Output

From the previous example, develop a 95% confidence interval for the difference in the mean between American and US Airways. Can we conclude that there is a difference between the two airlines' ratings?

Approximate results can also be obtained directly from the MINITAB output.

| | Individual 95% CIs For Mean Based on Pooled StDev | | | |
|------------|---------------------------------------------------|------|---------------|------|
| Level | -----+-----+-----+ | | | |
| American | | | (-----*-----) | |
| Delta | | | (-----*-----) | |
| United | | | (-----*-----) | |
| US Airways | (-----*-----) | | | |
| | -----+-----+-----+-----+ | | | |
| | 64.0 | 72.0 | 80.0 | 88.0 |

LO7 Carry out a test of hypothesis among treatment means using a blocking variable and understand the results.

Two-Way Analysis of Variance

- For the two-factor ANOVA we test whether there is a significant difference between the **treatment effect** and whether there is a difference in the **blocking effect**. Let B_r be the block totals (r for rows)
- Let SSB represent the sum of squares for the blocks where:

$$SSB = k \sum (\bar{x}_b - \bar{x}_G)^2$$

k is the number of treatments.

b is the number of blocks.

\bar{x}_b is the sample mean of block b .

\bar{x}_G is the overall or grand mean.

Two-Way Analysis of Variance - Example



WARTA, the Warren Area Regional Transit Authority, is expanding bus service from the suburb of Starbrick into the central business district of Warren. There are four routes being considered from Starbrick to downtown Warren: (1) via U.S. 6, (2) via the West End, (3) via the Hickory Street Bridge, and (4) via Route 59.

WARTA conducted several tests to determine whether there was a difference in the mean travel times along the four routes. Because there will be many different drivers, the test was set up so each driver drove along each of the four routes. Next slide shows the travel time, in minutes, for each driver-route combination. At the .05 significance level, is there a difference in the mean travel time along the four routes? If we remove the effect of the drivers, is there a difference in the mean travel time?

Two-Way Analysis of Variance - Example

Sample Data

| Driver | Travel Time From Starbrick to Warren (minutes) | | | |
|----------|------------------------------------------------|----------|-------------|---------|
| | U.S. 6 | West End | Hickory St. | Rte. 59 |
| Deans | 18 | 17 | 21 | 22 |
| Snaverly | 16 | 23 | 23 | 22 |
| Ormson | 21 | 21 | 26 | 22 |
| Zollaco | 23 | 22 | 29 | 25 |
| Filbeck | 25 | 24 | 28 | 28 |

Two-Way Analysis of Variance - Example

Step 1: State the null and alternate hypotheses.

$$H_0: \mu_u = \mu_w = \mu_h = \mu_r$$

H_1 : Not all treatment means are the same

Reject H_0 if $F > F_{\alpha, k-1, n-k}$

Step 2: State the level of significance.

The .05 significance level is stated in the problem.

Step 3: Find the appropriate test statistic.

Because we are comparing means of more than two groups,
use the F statistic

Step 4: State the decision rule.

Reject H_0 if $F > F_{\alpha, v1, v2}$

$F > F_{.05, k-1, n-k}$

$F > F_{.05, 4-1, 20-4}$

$F > F_{.05, 3, 16}$

$F > 3.24$

Two-Way Analysis of Variance - Example

| Travel Time From Starbrick to Warren (minutes) | | | | | | |
|------------------------------------------------|--------|----------|-------------|---------|-------------|--------------|
| Driver | U.S. 6 | West End | Hickory St. | Rte. 59 | Driver Sums | Driver Means |
| Deans | 18 | 17 | 21 | 22 | 78 | 19.5 |
| Snaverly | 16 | 23 | 23 | 22 | 84 | 21 |
| Ormson | 21 | 21 | 26 | 22 | 90 | 22.5 |
| Zollaco | 23 | 22 | 29 | 25 | 99 | 24.75 |
| Filbeck | 25 | 24 | 28 | 28 | 105 | 26.25 |

$$SSB = k \sum (\bar{x}_b - \bar{x}_G)^2$$

where

k is the number of treatments.

b is the number of blocks.

\bar{X}_b is the sample mean of block b .

\bar{X}_G is the overall or grand mean.

$$\begin{aligned}
 SSB &= k \sum (\bar{X}_b - \bar{X}_G)^2 \\
 &= 4(19.5 - 22.8)^2 + 4(21.0 - 22.8)^2 + 4(22.5 - 22.8)^2 \\
 &\quad + 4(24.75 - 22.8)^2 + 4(26.25 - 22.8)^2 \\
 &= 119.7
 \end{aligned}$$

Two-Way Analysis of Variance - Example

SUM OF SQUARES ERROR, TWO-WAY

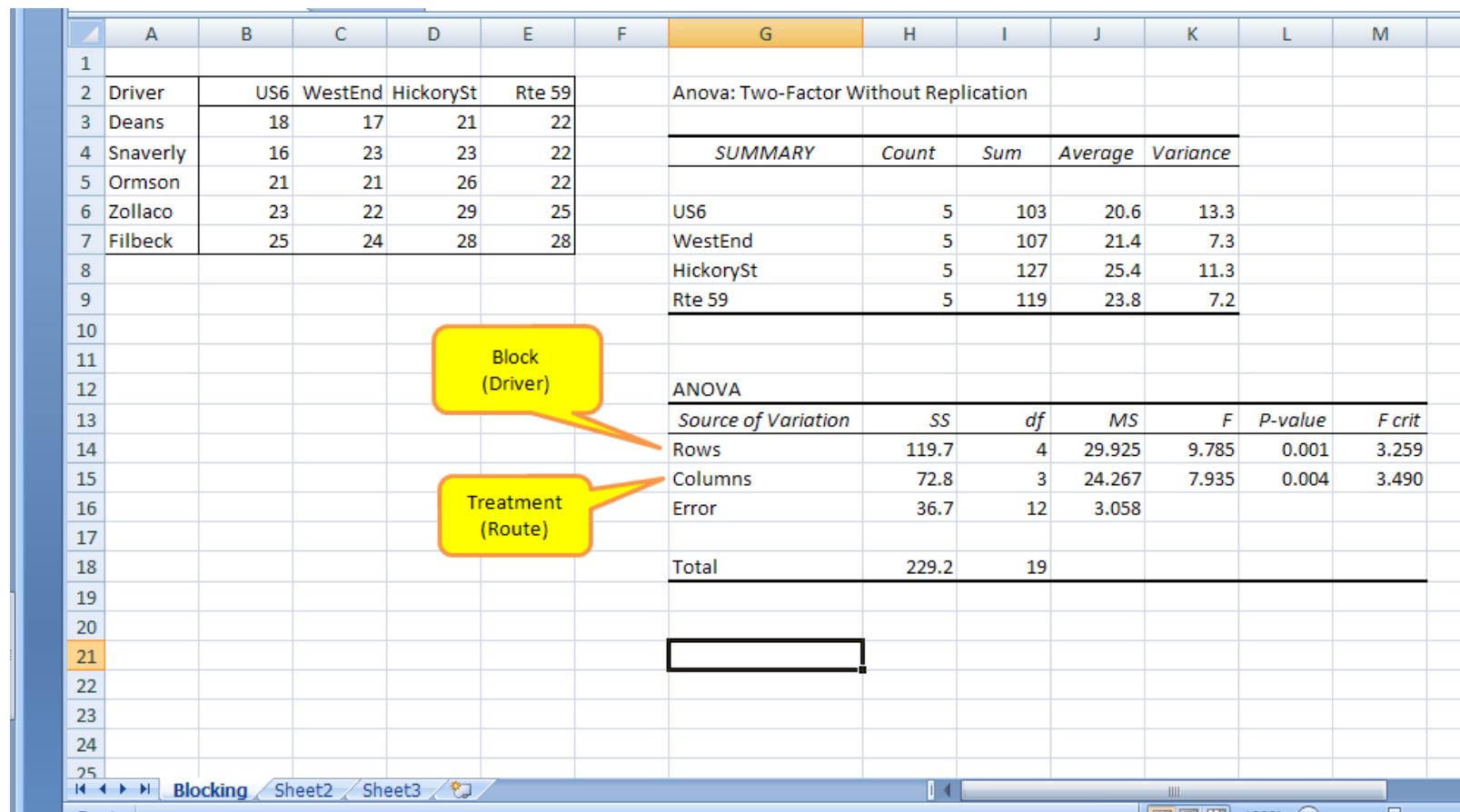
$$\text{SSE} = \text{SS total} - \text{SST} - \text{SSB}$$

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | F |
|---------------------|-----------------|---------------------------|------------------------------------------|---------|
| Treatments | SST | $k - 1$ | $\text{SST}/(k - 1) = \text{MST}$ | MST/MSE |
| Blocks | SSB | $b - 1$ | $\text{SSB}/(b - 1) = \text{MSB}$ | MSB/MSE |
| Error | SSE | $(k - 1)(b - 1)$ | $\text{SSE}/(k - 1)(b - 1) = \text{MSE}$ | |
| Total | <u>SS total</u> | <u>$n - 1$</u> | | |

$$\text{SSE} = \text{SS total} - \text{SST} - \text{SSB} = 229.2 - 72.8 - 119.7 = 36.7$$

| Source of Variation | (1) Sum of Squares | (2) Degrees of Freedom | (3) Mean Square (1)/(2) |
|---------------------|-----------------------|---------------------------|-------------------------------|
| Treatments | 72.8 | 3 | 24.27 |
| Blocks | 119.7 | 4 | 29.93 |
| Error | 36.7 | 12 | 3.06 |
| Total | 229.2 | 19 | |

Two-Way Analysis of Variance – Excel Example



Using Excel to perform the calculations, we conclude:

- (1) The mean time is not the same for all drivers
- (2) The mean times for the routes are not all the same

Two-way ANOVA with Interaction

INTERACTION The effect of one factor on a response variable differs depending on the value of another factor.

- In the previous section, we studied the separate or independent effects of two variables, routes into the city and drivers, on mean travel time.
- There is another effect that may influence travel time. This is called an interaction effect between route and driver on travel time. For example, is it possible that one of the drivers is especially good driving one or more of the routes?
- The **combined effect** of driver and route may also explain differences in mean travel time.
- To measure interaction effects it is necessary to have at least two observations in each cell.

Interaction Effect

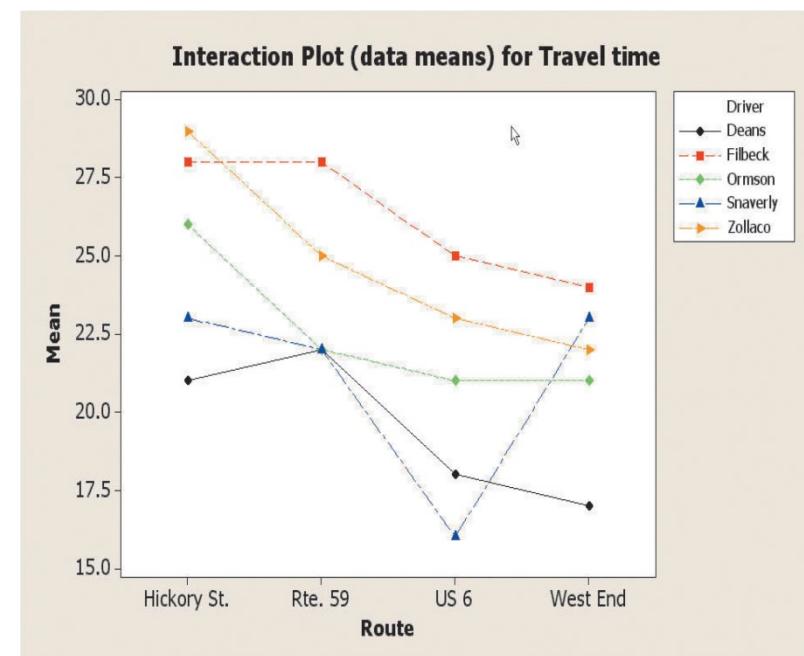
- When we use a two-way ANOVA to study interaction, we now call the two variables as **factors** instead of blocks
- Interaction occurs if the combination of two factors has some effect on the variable under study, in addition to each factor alone.
- The variable being studied is referred to as the **response variable**.
- One way to study interaction is by plotting factor means in a graph called an **interaction plot**.

Graphical Observation of Mean Times

Our graphical observations show us that interaction effects are possible. The next step is to conduct statistical tests of hypothesis to further investigate the possible interaction effects. In summary, our study of travel times has several questions:

- Is there really an interaction between routes and drivers?
- Are the travel times for the drivers the same?
- Are the travel times for the routes the same?

Of the three questions, we are most interested in the test for interactions. To put it another way, **does a particular route/driver combination result in significantly faster (or slower) driving times?** Also, the results of the hypothesis test for interaction affect the way we analyze the route and driver questions.



Example – ANOVA with Replication

Suppose the WARTA blocking experiment discussed earlier is repeated by measuring two more travel times for each driver and route combination with the data shown in the Excel worksheet.

| | | L8 | = | | | | | |
|----|--|----------|------|----------|------------|----------|--|--|
| 1 | | | US 6 | West End | Hickory St | Route 59 | | |
| 2 | | Deans | 18 | 14 | 20 | 19 | | |
| 3 | | Deans | 15 | 17 | 21 | 22 | | |
| 4 | | Deans | 21 | 20 | 22 | 25 | | |
| 5 | | Snaverly | 19 | 20 | 24 | 24 | | |
| 6 | | Snaverly | 15 | 24 | 23 | 22 | | |
| 7 | | Snaverly | 14 | 25 | 22 | 20 | | |
| 8 | | Ormson | 19 | 23 | 25 | 23 | | |
| 9 | | Ormson | 21 | 21 | 29 | 23 | | |
| 10 | | Ormson | 23 | 19 | 24 | 20 | | |
| 11 | | Zollaco | 24 | 20 | 30 | 26 | | |
| 12 | | Zollaco | 20 | 24 | 28 | 25 | | |
| 13 | | Zollaco | 25 | 22 | 29 | 24 | | |
| 14 | | Filbeck | 27 | 24 | 28 | 28 | | |
| 15 | | Filbeck | 25 | 24 | 28 | 30 | | |
| 16 | | Filbeck | +23 | 24 | 28 | 26 | | |
| 17 | | | | | | | | |
| 18 | | | | | | | | |

Three Tests in ANOVA with Replication

The ANOVA now has three sets of hypotheses to test:

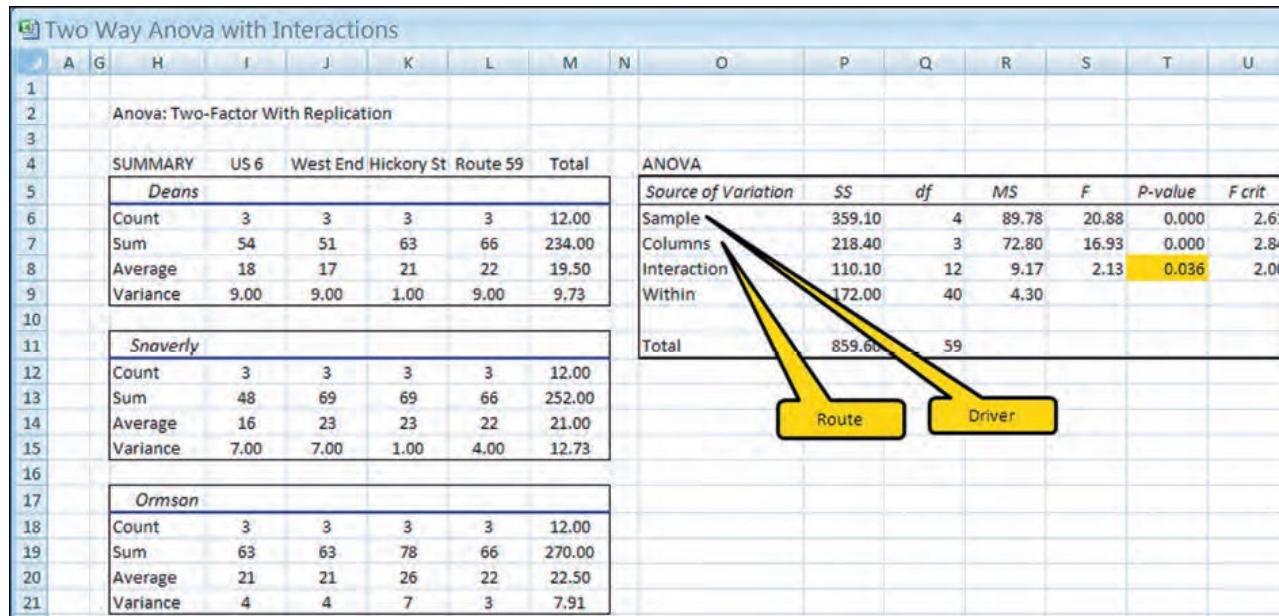
1. H_0 : There is no interaction between drivers and routes.
 H_1 : There is interaction between drivers and routes.

2. H_0 : The driver means are the same.
 H_1 : The driver means are *not* the same.

3. H_0 : The route means are the same.
 H_1 : The route means are *not* the same.

ANOVA Table

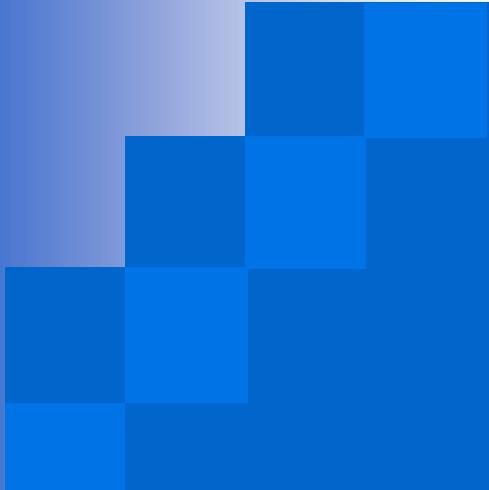
| Source | Sum of Squares | df | Mean Square | F |
|-------------|----------------|------------------|----------------------------|---------|
| Route | Factor A | $k - 1$ | $SSA/(k - 1) = MSA$ | MSA/MSE |
| Driver | Factor B | $b - 1$ | $SSB/(b - 1) = MSB$ | MSB/MSE |
| Interaction | SSI | $(k - 1)(b - 1)$ | $SSI/(k - 1)(b - 1) = MSI$ | MSI/MSE |
| Error | SSE | $n - kb$ | $SSE/(n - kb) = MSE$ | |
| Total | SS total | $n - 1$ | | |



One-way ANOVA for Each Driver

H_0 : Route travel times are equal.

| Deans: H_0: Route travel times are equal. | | | | | | Snaiverly: H_0: Route travel times are equal. | | | | | |
|-----------------------------------------------------------------|----|--------|-------|------|-------|-------------------------------------------------------------------|----|--------|-------|------|-------|
| Source | DF | SS | MS | F | P | Source | DF | SS | MS | F | P |
| Dean RTE | 3 | 51.00 | 17.00 | 2.43 | 0.140 | SN RTE | 3 | 102.00 | 34.00 | 7.16 | 0.012 |
| Error | 8 | 56.00 | 7.00 | | | Error | 8 | 38.00 | 4.75 | | |
| Total | 11 | 107.00 | | | | Total | 11 | 140.00 | | | |
| Ormson: H_0: Route travel times are equal. | | | | | | Zollaco: H_0: Route travel times are equal. | | | | | |
| Source | DF | SS | MS | F | P | Source | DF | SS | MS | F | P |
| Ormson RTE | 3 | 51.00 | 17.00 | 3.78 | 0.059 | Z-RTE | 3 | 86.25 | 28.75 | 8.85 | 0.006 |
| Error | 8 | 36.00 | 4.50 | | | Error | 8 | 26.00 | 3.25 | | |
| Total | 11 | 87.00 | | | | Total | 11 | 112.25 | | | |
| Filbeck: H_0: Route travel times are equal. | | | | | | | | | | | |
| Source | DF | SS | MS | F | P | | | | | | |
| Filbeck RTE | 3 | 38.25 | 12.75 | 6.38 | 0.016 | | | | | | |
| Error | 8 | 16.00 | 2.00 | | | | | | | | |
| Total | 11 | 54.25 | | | | | | | | | |



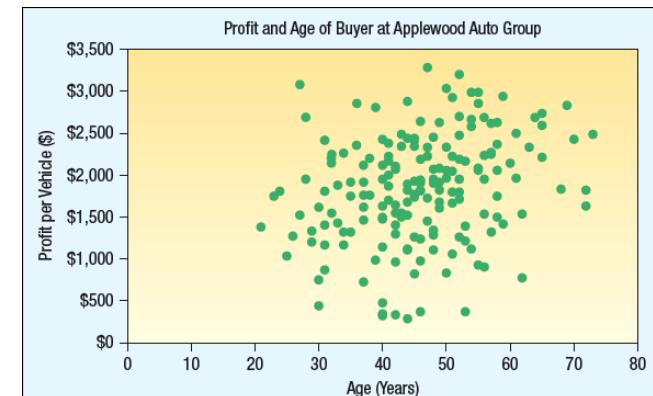
Correlation and Linear Regression

Chapter 13



Correlation Analysis – Measuring the Relationship Between Two Variables

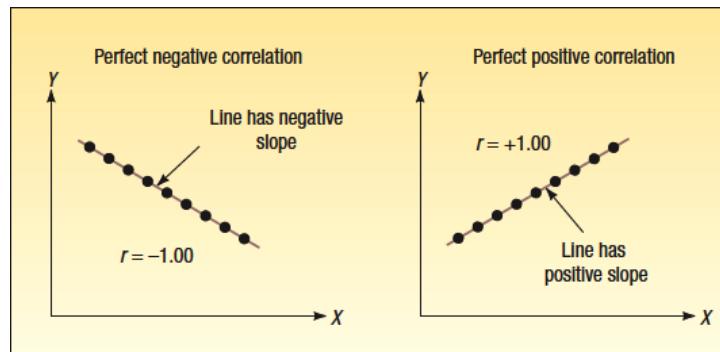
- Analyzing relationships between two quantitative variables.
- The basic hypothesis of correlation analysis: Does the data indicate that there is a relationship between two quantitative variables?
- For the Applewood Auto sales data, the data is displayed in a scatter graph.
- Are profit per vehicle and age correlated?



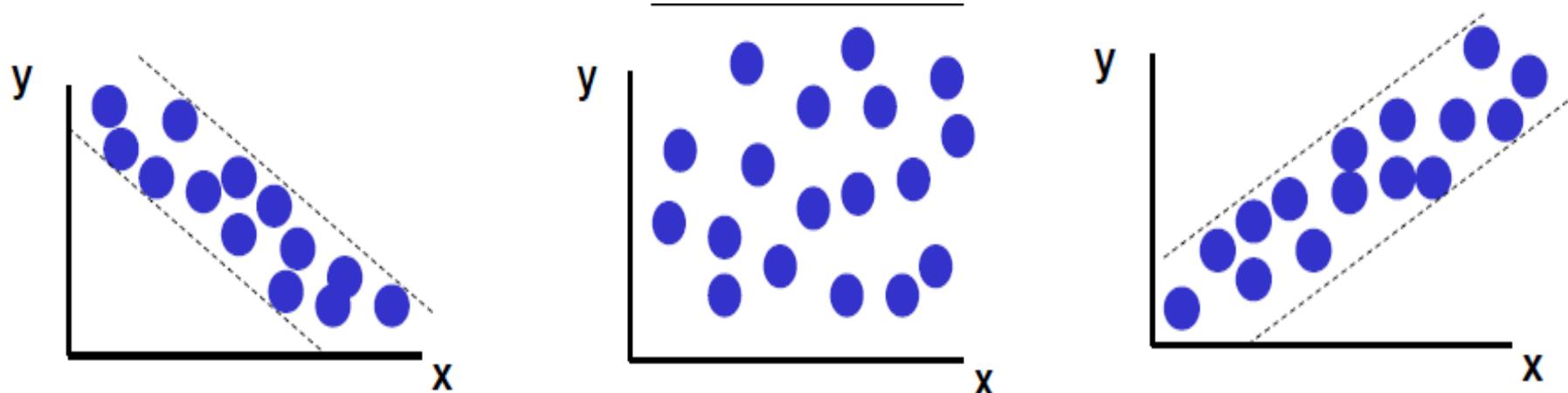
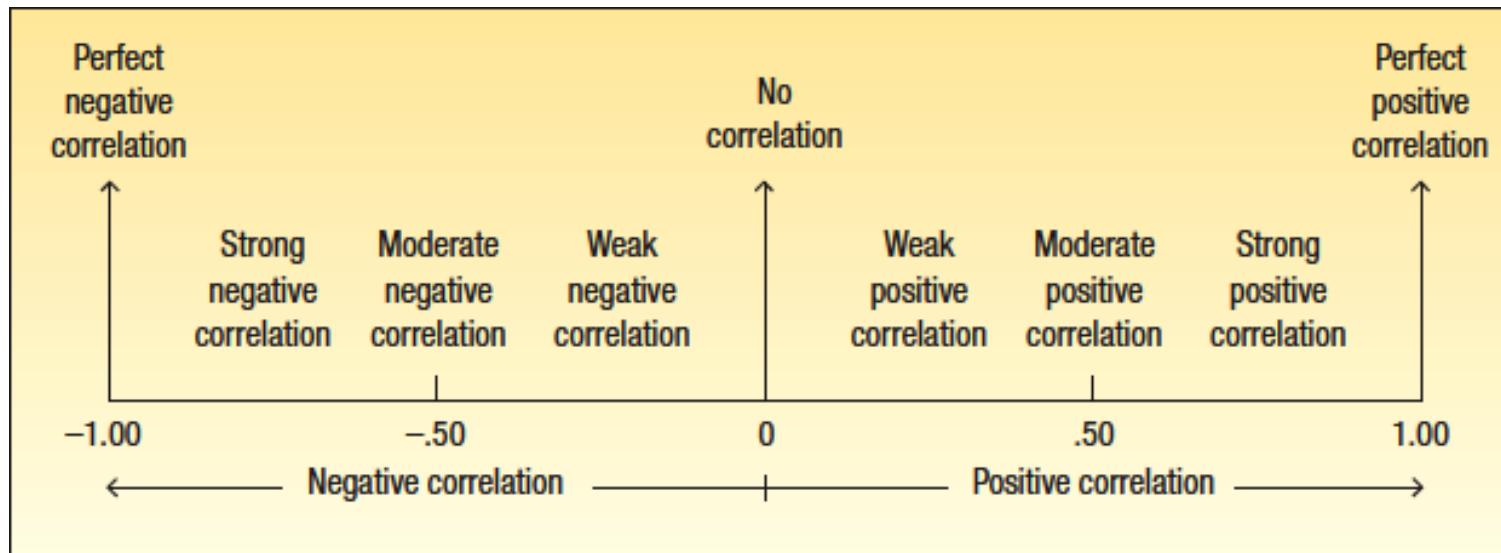
Correlation Analysis – Measuring the Relationship Between Two Variables

The **Coefficient of Correlation (r)** is a measure of the strength of the relationship between two variables.

- It shows the direction and strength of the linear relationship between two interval- or ratio-scale variables.
- It ranges from -1 up to and including +1.
- A value near 0 indicates there is little linear relationship between the variables.
- A value near +1 indicates a direct or positive linear relationship between the variables.
- A value near -1 indicates an inverse or negative linear relationship between the variables.



Correlation Analysis – Measuring the Relationship Between Two Variables



Correlation Analysis – Measuring the Relationship Between Two Variables

- Computing the Correlation Coefficient:

CORRELATION COEFFICIENT

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y}$$

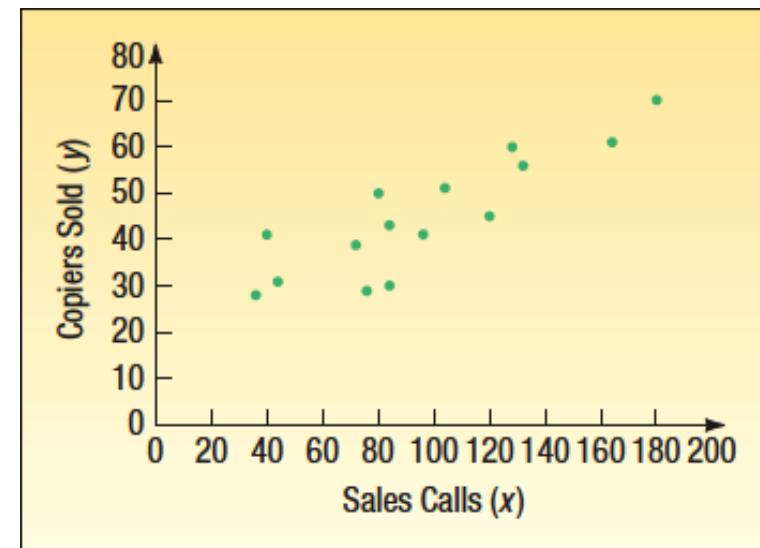
[13-1]

Correlation Analysis – Example

The sales manager of Copier Sales of America has a large sales force throughout the United States and Canada and wants to determine whether there is a **relationship between the number of sales calls made** in a month and the **number of copiers sold that month**. The manager selects a random sample of 15 representatives and determines the number of sales calls each representative made last month and the number of copiers sold.

Determine if the number of sales calls and copiers sold are correlated.

| Sales Representative | Sales Calls | Copiers Sold |
|----------------------|-------------|--------------|
| Brian Virost | 96 | 41 |
| Carlos Ramirez | 40 | 41 |
| Carol Saia | 104 | 51 |
| Greg Fish | 128 | 60 |
| Jeff Hall | 164 | 61 |
| Mark Reynolds | 76 | 29 |
| Meryl Rumsey | 72 | 39 |
| Mike Kiel | 80 | 50 |
| Ray Snarsky | 36 | 28 |
| Rich Niles | 84 | 43 |
| Ron Broderick | 180 | 70 |
| Sal Spina | 132 | 56 |
| Soni Jones | 120 | 45 |
| Susan Welch | 44 | 31 |
| Tom Keller | 84 | 30 |



Correlation Coefficient – Example

CORRELATION COEFFICIENT

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y}$$

[13-1]

$$\bar{x} = 96; \bar{y} = 45; s_x = 42.76; s_y = 12.89$$

| Sales Representative | Sales Calls (x) | Copiers Sold (y) | $x - \bar{x}$ | $y - \bar{y}$ | $(x - \bar{x})(y - \bar{y})$ |
|----------------------|-----------------|------------------|---------------|---------------|------------------------------|
| Brian Virost | 96 | 41 | 0 | -4 | 0 |
| Carlos Ramirez | 40 | 41 | -56 | -4 | 224 |
| Carol Saia | 104 | 51 | 8 | 6 | 48 |
| Greg Fish | 128 | 60 | 32 | 15 | 480 |
| Jeff Hall | 164 | 61 | 68 | 16 | 1,088 |
| Mark Reynolds | 76 | 29 | -20 | -16 | 320 |
| Meryl Rumsey | 72 | 39 | -24 | -6 | 144 |
| Mike Kiel | 80 | 50 | -16 | 5 | -80 |
| Ray Snarsky | 36 | 28 | -60 | -17 | 1,020 |
| Rich Niles | 84 | 43 | -12 | -2 | 24 |
| Ron Broderick | 180 | 70 | 84 | 25 | 2,100 |
| Sal Spina | 132 | 56 | 36 | 11 | 396 |
| Soni Jones | 120 | 45 | 24 | 0 | 0 |
| Susan Welch | 44 | 31 | -52 | -14 | 728 |
| Tom Keller | 84 | 30 | -12 | -15 | 180 |
| Totals | 1440 | 675 | 0 | 0 | 6,672 |

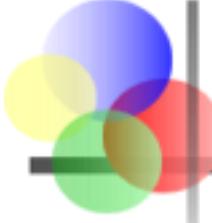
Numerator

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y} = \frac{6672}{(15 - 1)(42.76)(12.89)} = 0.865$$

Regression Analysis

Correlation Analysis tests for the strength and direction of the relationship between two quantitative variables.

Regression Analysis evaluates and “measures” the relationship between two quantitative variables with a linear equation. This equation has the same elements as any equation of a line, that is, a slope and an intercept.



Population Linear Regression

The population regression model:

$$y = \alpha + \beta x + \varepsilon$$

Annotations for the population regression model:

- Dependent Variable
- Population y intercept
- Population Slope Coefficient
- Independent Variable
- Random Error term, or residual

The equation is divided into two main components by a blue brace at the bottom:

- Linear component: $\alpha + \beta x$
- Random Error component: ε

Regression Analysis

$$Y = a + bX$$

Estimate of the regression
intercept

Estimate of the
regression slope

Regression Analysis

EXAMPLES

- Assuming a linear relationship between the size of a home, measured in square feet, and the cost to heat the home in January, how does the cost vary relative to the size of the home?
- In a study of automobile fuel efficiency, assuming a linear relationship between miles per gallon and the weight of a car, how does the fuel efficiency vary relative to the weight of a car?

Regression Analysis: Variables

$$Y = a + b X$$

- **Y** is the **Dependent Variable**. It is the variable being predicted or estimated.
- **X** is the **Independent Variable**. For a regression equation, it is the variable used to estimate the dependent variable, **Y**. **X** is the predictor variable.

Examples of dependent and independent variables:

- How does the size of a home, measured in number of square feet, relate to the cost to heat the home in January? We would use the home size as, **X**, the independent variable to predict the heating cost, and **Y** as the dependent variable.

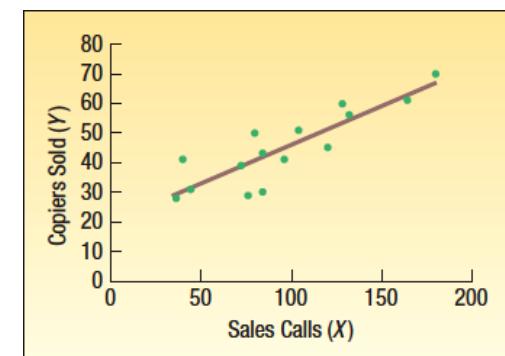
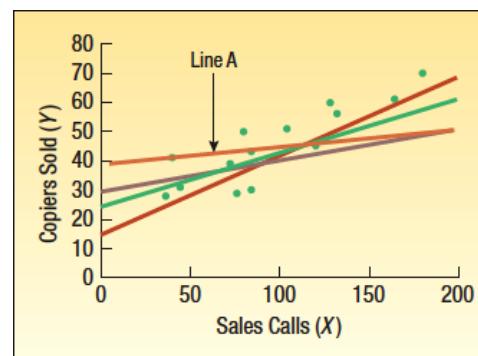
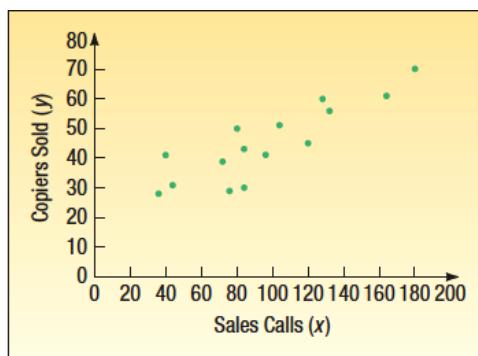
Regression equation: Heating cost = a + b (home size)

- How does the weight of a car relate to the car's fuel efficiency? We would use car weight as, **X**, the independent variable to predict the car's fuel efficiency, and **Y** as the dependent variable.

Regression equation: Miles per gallon = a + b (car weight)

Regression Analysis – Example

- Regression analysis estimates **a** and **b** by fitting a line to the observed data.
- Each line ($Y = a + bX$) is defined by values of **a** and **b**.

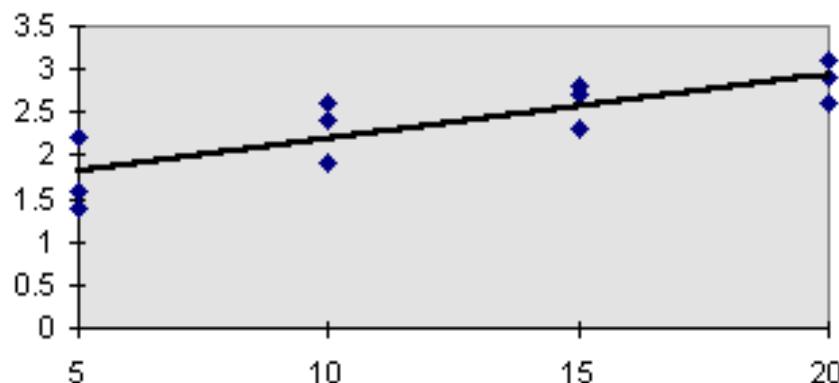


A way to find the line of “best fit” to the data is the:

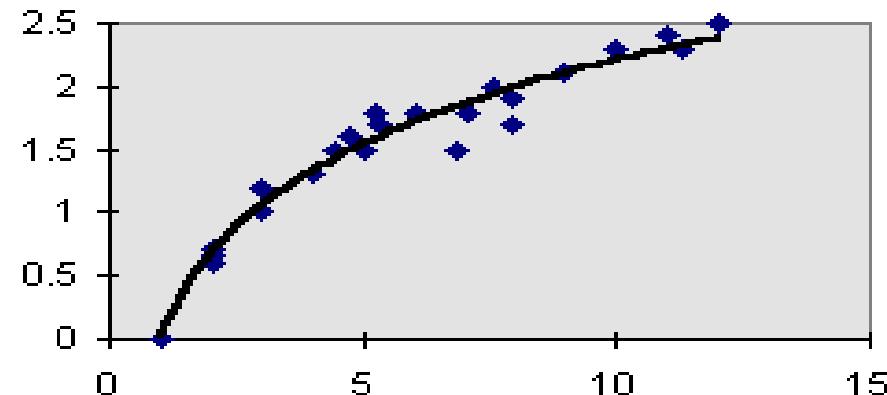
LEAST SQUARES PRINCIPLE Determining a regression equation by minimizing the sum of the squares of the vertical distances between the actual **Y** values and the predicted values of **Y**.

Types of Regression Models

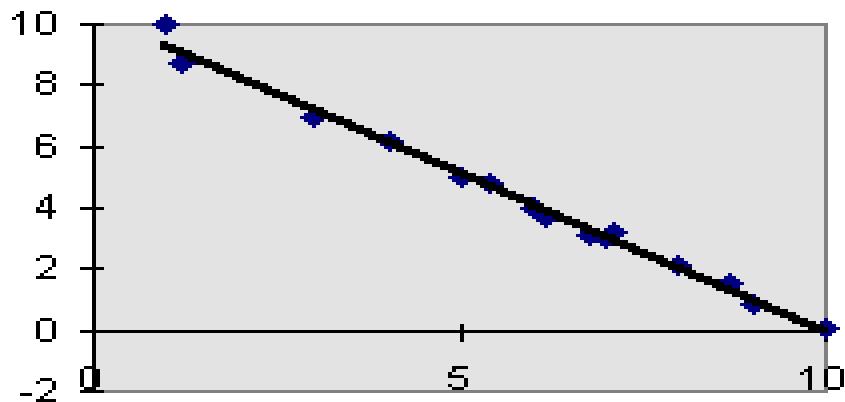
Positive Linear Relationship



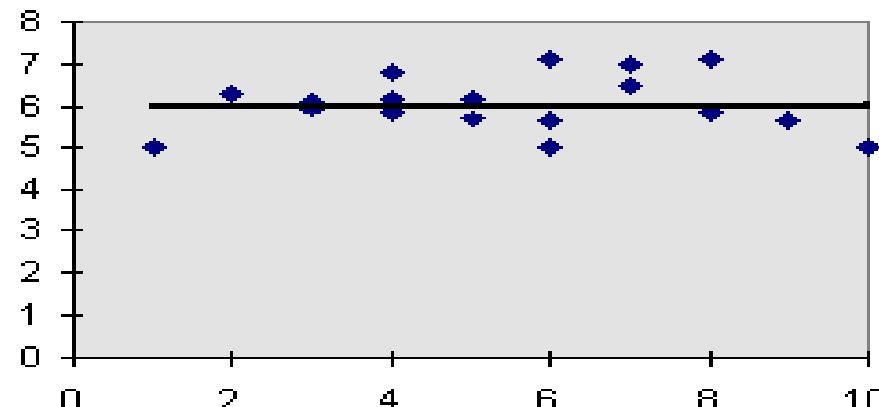
Relationship NOT Linear



Negative Linear Relationship



No Relationship



Regression Analysis – Example

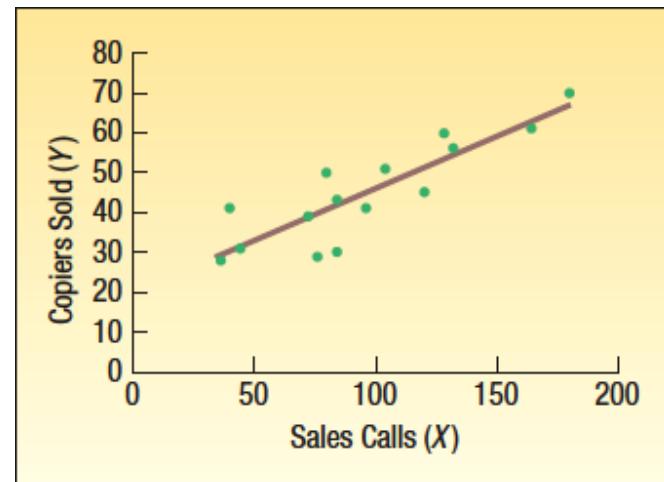
Recall the example involving Copier Sales of America. The sales manager gathered information on the number of sales calls made and the number of copiers sold for a random sample of 15 sales representatives. Use the least squares method to determine a linear equation to express the relationship between the two variables.

In this example, the number of sales calls is the independent variable, X , and the number of copiers sold is the dependent variable, Y .

$$\text{Number of Copiers Sold} = a + b (\text{ Number of Sales Calls})$$

What is the expected number of copiers sold by a representative who **made 20 calls?**

| Sales Representative | Sales Calls | Copiers Sold |
|----------------------|-------------|--------------|
| Brian Virost | 96 | 41 |
| Carlos Ramirez | 40 | 41 |
| Carol Saia | 104 | 51 |
| Greg Fish | 128 | 60 |
| Jeff Hall | 164 | 61 |
| Mark Reynolds | 76 | 29 |
| Meryl Rumsey | 72 | 39 |
| Mike Kiel | 80 | 50 |
| Ray Snarsky | 36 | 28 |
| Rich Niles | 84 | 43 |
| Ron Broderick | 180 | 70 |
| Sal Spina | 132 | 56 |
| Soni Jones | 120 | 45 |
| Susan Welch | 44 | 31 |
| Tom Keller | 84 | 30 |
| Total | 1440 | 675 |



Regression Analysis – Example

Correlation coefficient:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y} = \frac{6672}{(15 - 1)(42.76)(12.89)} = 0.865$$

Regression Analysis - Example

Step 1: Find the slope (b) of the line.

SLOPE OF THE REGRESSION LINE

$$b = r \left(\frac{s_y}{s_x} \right)$$

[13-4]

where:

r is the correlation coefficient.

s_y is the standard deviation of y (the dependent variable).

s_x is the standard deviation of x (the independent variable).

$$b = r \left(\frac{s_y}{s_x} \right) = .865 \left(\frac{12.89}{42.76} \right) = 0.2608$$

Step 2: Find the y -intercept (a).

Y-INTERCEPT

$$a = \bar{y} - b\bar{x}$$

[13-5]

where:

\bar{y} is the mean of y (the dependent variable).

\bar{x} is the mean of x (the independent variable).

$$a = \bar{y} - b\bar{x} = 45 - .2608(96) = 19.9632$$

Regression Analysis - Example

Step 3: Create the regression equation.

$$\text{Number of Copiers Sold} = 19.9632 + 0.2608 (\text{ Number of Sales Calls})$$

Step 4: What is the predicted number of sales if someone makes 20 sales calls?

$$\text{Number of Copiers Sold} = 25.1792 = 19.9632 + 0.2608(20)$$

Regression Analysis (Excel) – Example

| SUMMARY OUTPUT | | | | | |
|-----------------------|-------|--------------|----------------|----------|----------------|
| Regression Statistics | | | | | |
| Multiple R | 0.865 | | | | |
| R Square | 0.748 | | | | |
| Adjusted R Square | 0.728 | | | | |
| Standard Error | 6.720 | | | | |
| Observations | 15 | | | | |
| ANOVA | | | | | |
| | df | SS | MS | F | Significance F |
| Regression | 1 | 1738.89 | 1738.89 | 38.50313 | 3.19277E-05 |
| Residual | 13 | 587.11 | 45.16231 | | |
| Total | 14 | 2326 | | | |
| | | Coefficients | Standard Error | t Stat | P-value |
| Intercept | | 19.9800 | 4.389675533 | 4.551589 | 0.000544 |
| Sales calls (x) | | 0.2606 | 0.042001817 | 6.205089 | 3.19E-05 |

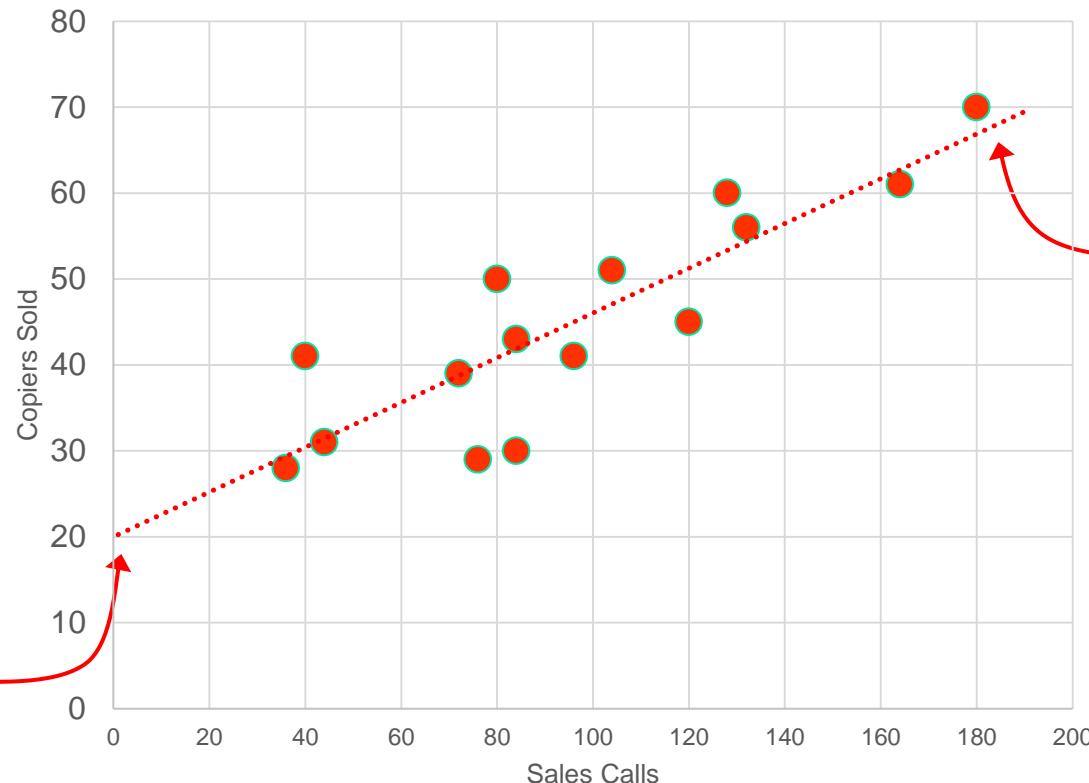
a

b

$$\text{Number of Copiers Sold} = 19.9800 + 0.2606 (\text{ Number of Sales Calls})$$

* Note that the Excel differences in the values of a and b are due to rounding.

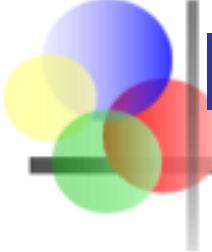
Graphical Presentation



Intercept
= 19.98

Slope = 0.26

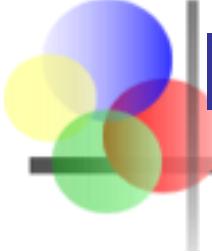
$$\widehat{\text{Copiers Sold}} = 19.98 + 0.26 \text{ (calls)}$$



Interpretation of the Intercept, a

$$\text{Copiers Sold} = 19.98 + 0.26 \text{ (calls)}$$

- b is the estimated average value of Y when the value of X is zero (if $x = 0$ is in the range of observed x values)
 - Here, it is the estimated average copiers sold when there is exactly zero calls.



Interpretation of the Slope, b

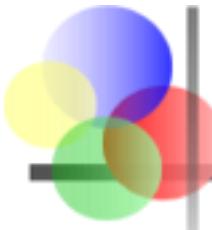
$$\text{Copiers Sold} = 19.98 + 0.26 \text{ (calls)}$$

- b measures the estimated change in the average value of Y as a result of a one-unit change in X
 - Here, b=0.26 tells us that the average value of a copiers sold increases by 0.26, on average, for each additional call

Regression Analysis: Coefficient of Determination

The **coefficient of determination** (r^2) is the proportion of the total variation in the dependent variable (Y) that is explained or accounted for by the variation in the independent variable (X). It is the square of the coefficient of correlation.

- It ranges from 0 to 1.
- It does not provide any information on the direction of the relationship between the variables.



Explained and Unexplained Variation

- Total variation is made up of two parts:

$$SST = SSE + SSR$$

Total sum of Squares

Sum of Squares Error

Sum of Squares Regression

$$SST = \sum (y - \bar{y})^2$$

$$SSE = \sum (y - \hat{y})^2$$

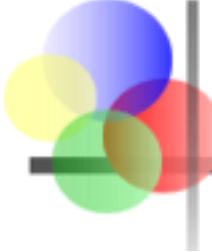
$$SSR = \sum (\hat{y} - \bar{y})^2$$

where:

\bar{y} = Average value of the dependent variable

y = Observed values of the dependent variable

\hat{y} = Estimated value of y for the given x value



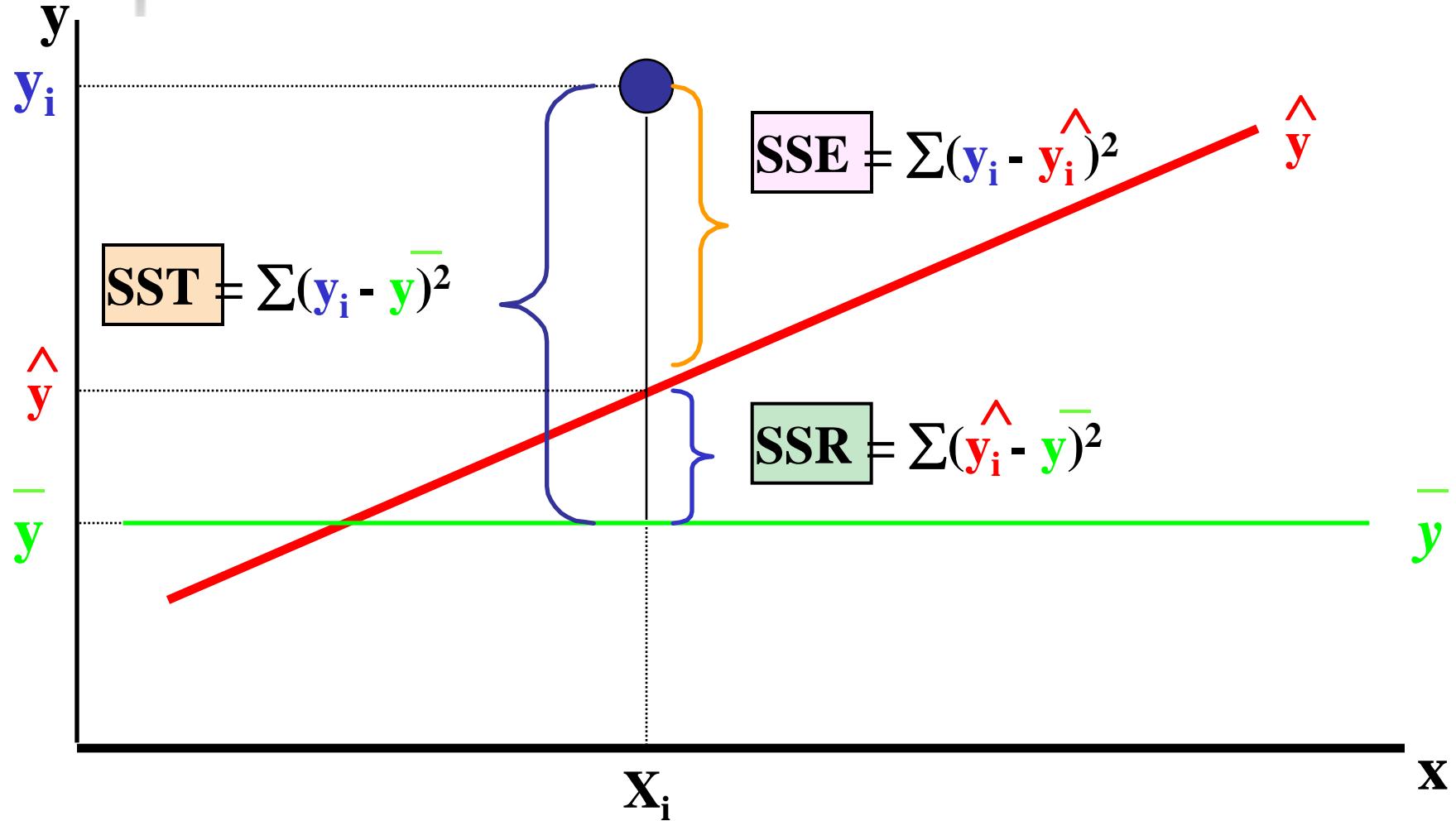
Explained and Unexplained Variation

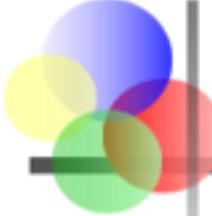
(continued)

- **SST = total sum of squares**
 - Measures the variation of the y_i values around their mean y
- **SSE = error sum of squares**
 - Variation attributable to factors other than the relationship between x and y
- **SSR = regression sum of squares**
 - Explained variation attributable to the relationship between x and y

Explained and Unexplained Variation

(continued)





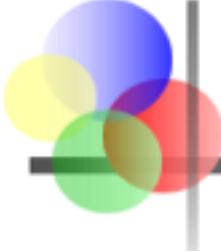
Coefficient of Determination, R²

- 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 R²

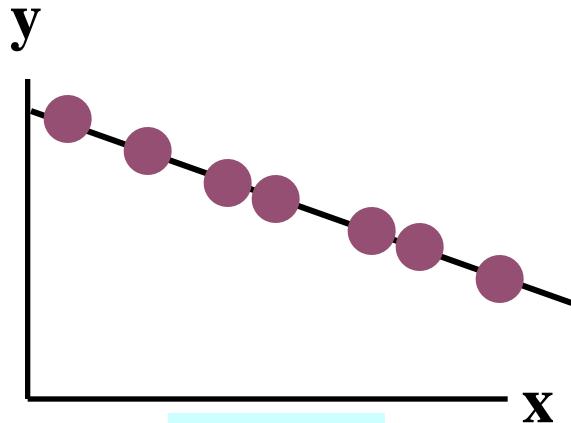
$$R^2 = \frac{SSR}{SST}$$

where

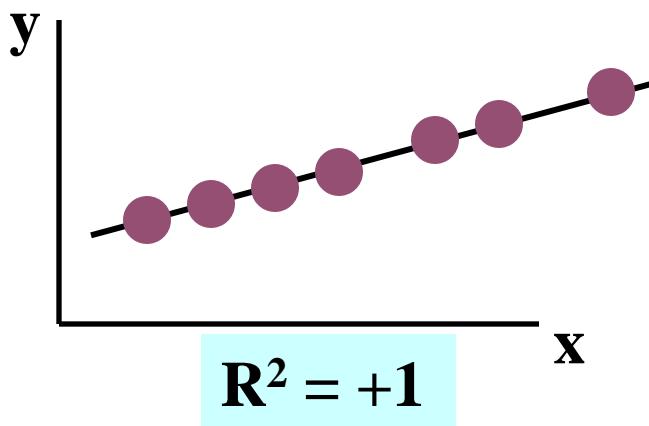
$$0 \leq R^2 \leq 1$$



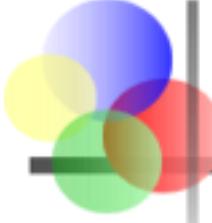
Examples of Approximate R^2 Values



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



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



Coefficient of Determination, R²

(continued)

Coefficient of determination

$$R^2 = \frac{SSR}{SST} = \frac{\text{sum of squares explained by regression}}{\text{total sum of squares}}$$

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

$$R^2 = r^2$$

where:

R² = Coefficient of determination

r = Simple correlation coefficient

Regression Analysis: Coefficient of Determination – Example

- The coefficient of determination, r^2 , is 0.748. It can be computed as the correlation coefficient, squared: $(0.865)^2$.
- The coefficient of determination is expressed as a proportion or percent; we say that **74.8 percent of the variation in the number of copiers sold is explained**, or accounted for, by **the variation in the number of sales calls**.

Regression Analysis: Coefficient of Determination – Example

The Coefficient of Determination can also be computed based on its definition. We can divide the Regression Sum of Squares (the variation in the dependent variable explained by the regression equation) divided by the Total Sum of Squares (the total variation in the dependent variable).

| SUMMARY OUTPUT | | | | | |
|-----------------------|--------------|-------------------------|----------|----------|----------------|
| Regression Statistics | | | | | |
| Multiple R | 0.865 | Correlation Coefficient | | | |
| R Square | 0.748 | | | | |
| Adjusted R Square | 0.728 | | | | |
| Standard Error | 6.720 | | | | |
| Observations | 15 | | | | |
| ANOVA | | | | | |
| | df | SS | MS | F | Significance F |
| Regression | 1 | 1738.89 | 1738.89 | 38.50313 | 3.19277E-05 |
| Residual | 13 | 587.11 | 45.16231 | | |
| Total | 14 | 2326 | | | |
| | Coefficients | Standard Error | t Stat | P-value | |
| Intercept | 19.9800 | 4.389675533 | 4.551589 | 0.000544 | |
| Sales calls (x) | 0.2606 | 0.042001817 | 6.205089 | 3.19E-05 | |

$$R^2 = \frac{\text{RegressionSumofSquares}}{\text{TotalSumofSquares}} = \frac{1738.89}{2326} = 0.748$$

Regression Analysis: Testing the Significance of the Slope – Example

Step 1: State the null and alternate hypotheses.

$H_0: \beta = 0$ (the slope of the regression equation is 0)

$H_1: \beta \neq 0$ (the slope of the regression equation is not 0)

Step 2: Select a level of significance.

We select a .05 level of significance.

Step 3: Identify the test statistic.

To test a hypothesis about the slope of a regression equation, we use the t -statistic. For this analysis, there will be $n-2$ degrees of freedom.

Regression Analysis: Testing the Significance of the Slope – Example

Step 4: Formulate a decision rule.

Reject H_0 if:

$$t > t_{\alpha/2, n-2} \text{ or } t < -t_{\alpha/2, n-2}$$

$$t > t_{0.025, 13} \text{ or } t < -t_{0.025, 13}$$

$$t > 1.771 \text{ or } t < -1.771$$

Regression Analysis: Testing the Significance of the Slope – Example

Step 5: Take a sample, using Excel, arrive at a decision.

| SUMMARY OUTPUT | | | | | |
|-----------------------|-------------|---------|-------------|----------|----------------|
| Regression Statistics | | | | | |
| Multiple R | 0.864631791 | | | | |
| R Square | 0.747588134 | | | | |
| Adjusted R Square | 0.728171837 | | | | |
| Standard Error | 6.720290745 | | | | |
| Observations | 15 | | | | |
| ANOVA | | | | | |
| | df | SS | MS | F | Significance F |
| Regression | 1 | 1738.89 | 1738.89 | 38.50313 | 0.000 |
| Residual | 13 | 587.11 | 45.16230769 | | |
| Total | 14 | 2326 | | | |
| Coefficients | | | | | |
| Intercept | 19.98 | 4.39 | 4.55 | 0.001 | |
| Sales Calls | 0.26 | 0.04 | 6.21 | 0.000 | |

TEST FOR THE
SLOPE

$$t = \frac{b - 0}{s_b}$$

with $n - 2$ degrees of freedom

[13-6]

Decision: Reject the null hypothesis that the slope of the regression equation is equal to zero.

Inferences about the Slope: t Test Example

Test Statistic: $t = 6.20$

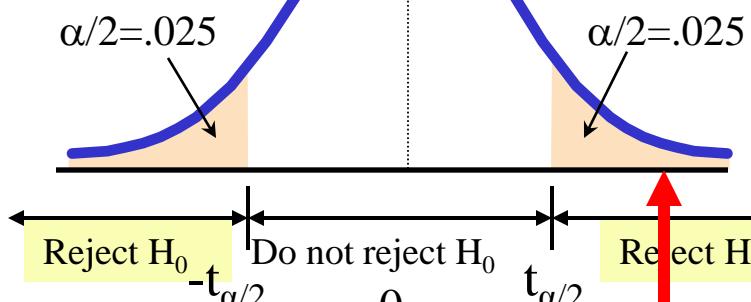
$$H_0: \beta = 0$$

$$H_A: \beta \neq 0$$

From Excel output:

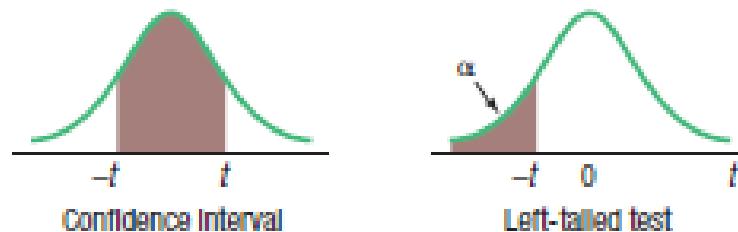
| | Coefficients | Standard Error | t Stat | P-value |
|-------------|--------------|----------------|--------|---------|
| Intercept | 19.98 | 4.388 | 4.55 | 0.0001 |
| Square Feet | 0.2606 | 0.042 | 6.20 | 0.0000 |

$$d.f. = 15 - 2 = 13$$



6.20

Decision: Reject H_0
Conclusion: There is sufficient evidence that calls effect copiers sold.



| df | Confidence Intervals, c | | | | | |
|------|-----------------------------------------------------|-------|--------|--------|--------|---------|
| | Level of Significance for One-Tailed Test, α | | | | | |
| | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.0005 |
| | Level of Significance for Two-Tailed Test, α | | | | | |
| 0.20 | 0.10 | 0.05 | 0.02 | 0.01 | 0.005 | 0.001 |
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 636.619 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 31.599 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 12.924 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 8.610 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 6.869 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.959 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 5.408 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 5.041 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.781 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.587 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.437 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 4.318 |
| 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 4.221 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 4.140 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 4.073 |
| 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 4.015 |
| 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.965 |
| 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.922 |
| 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.883 |
| 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.850 |

Regression Analysis: Testing the Significance of the Slope – Example

Step 6: Interpret the result. For the regression equation that predicts the number of copier sales based on the number of sales calls, the data indicate that the slope, (0.2606), is not equal to zero.

As in correlation analysis, please note that this statistical analysis does not provide any evidence of a causal relationship. Another type of study is needed to test that hypothesis.

Multiple Regression Analysis

Chapter 14



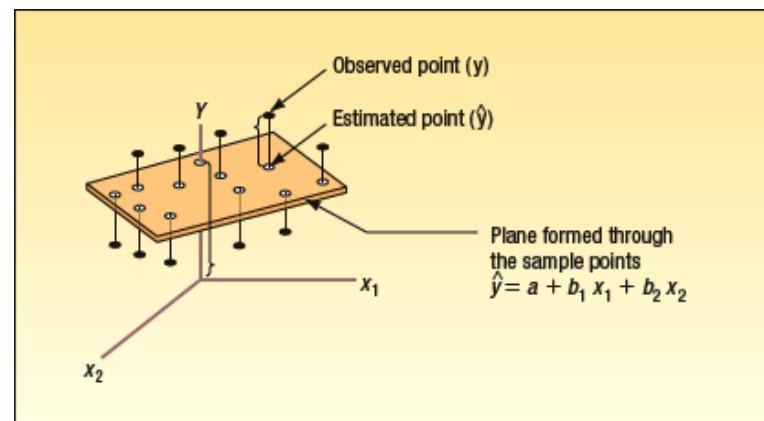
Multiple Regression Analysis

The general multiple regression equation with k independent variables is given by:

**GENERAL MULTIPLE
REGRESSION EQUATION**

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_kx_k \quad [14-1]$$

- $X_1 \dots X_k$ are the independent variables.
- a is the y -intercept
- b_1 is the net change in Y for each unit change in X_1 holding $X_2 \dots X_k$ constant.
- Determining b_1 , b_2 , etc. is very tedious. A software package is recommended.
- The least squares criterion is used to develop this equation.



Multiple Regression Analysis- Example

Salsberry Realty sells homes along the east coast of the United States. One of the questions most frequently asked by prospective buyers is: If we purchase this home, how much can we expect to pay to heat it during the winter? The research department at Salsberry has been asked to develop some guidelines regarding heating costs for single-family homes.

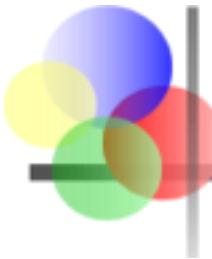


| Home | Heating Cost (\$) | Mean Outside Temperature (°F) | Attic Insulation (inches) | Age of Furnace (years) |
|------|-------------------|-------------------------------|---------------------------|------------------------|
| 1 | \$250 | 35 | 3 | 6 |
| 2 | 360 | 29 | 4 | 10 |
| 3 | 165 | 36 | 7 | 3 |
| 4 | 43 | 60 | 6 | 9 |
| 5 | 92 | 65 | 5 | 6 |
| 6 | 200 | 30 | 5 | 5 |
| 7 | 355 | 10 | 6 | 7 |
| 8 | 290 | 7 | 10 | 10 |
| 9 | 230 | 21 | 9 | 11 |
| 10 | 120 | 55 | 2 | 5 |
| 11 | 73 | 54 | 12 | 4 |
| 12 | 205 | 48 | 5 | 1 |
| 13 | 400 | 20 | 5 | 15 |
| 14 | 320 | 39 | 4 | 7 |
| 15 | 72 | 60 | 8 | 6 |
| 16 | 272 | 20 | 5 | 8 |
| 17 | 94 | 58 | 7 | 3 |
| 18 | 190 | 40 | 8 | 11 |
| 19 | 235 | 27 | 9 | 8 |
| 20 | 139 | 30 | 7 | 5 |

Three variables are thought to relate to the heating costs: (1) the mean daily outside temperature, (2) the number of inches of insulation in the attic, and (3) the age in years of the furnace.
The Multiple Linear regression equation is:

$$\text{Heating Cost} = a + b_1(\text{Mean Outside Temp}) + b_2(\text{Inches of Attic Insulation}) + b_3(\text{Age of the Furnace})$$

To investigate, Salsberry's research department selected a random sample of 20 recently sold homes and measured all four variables.

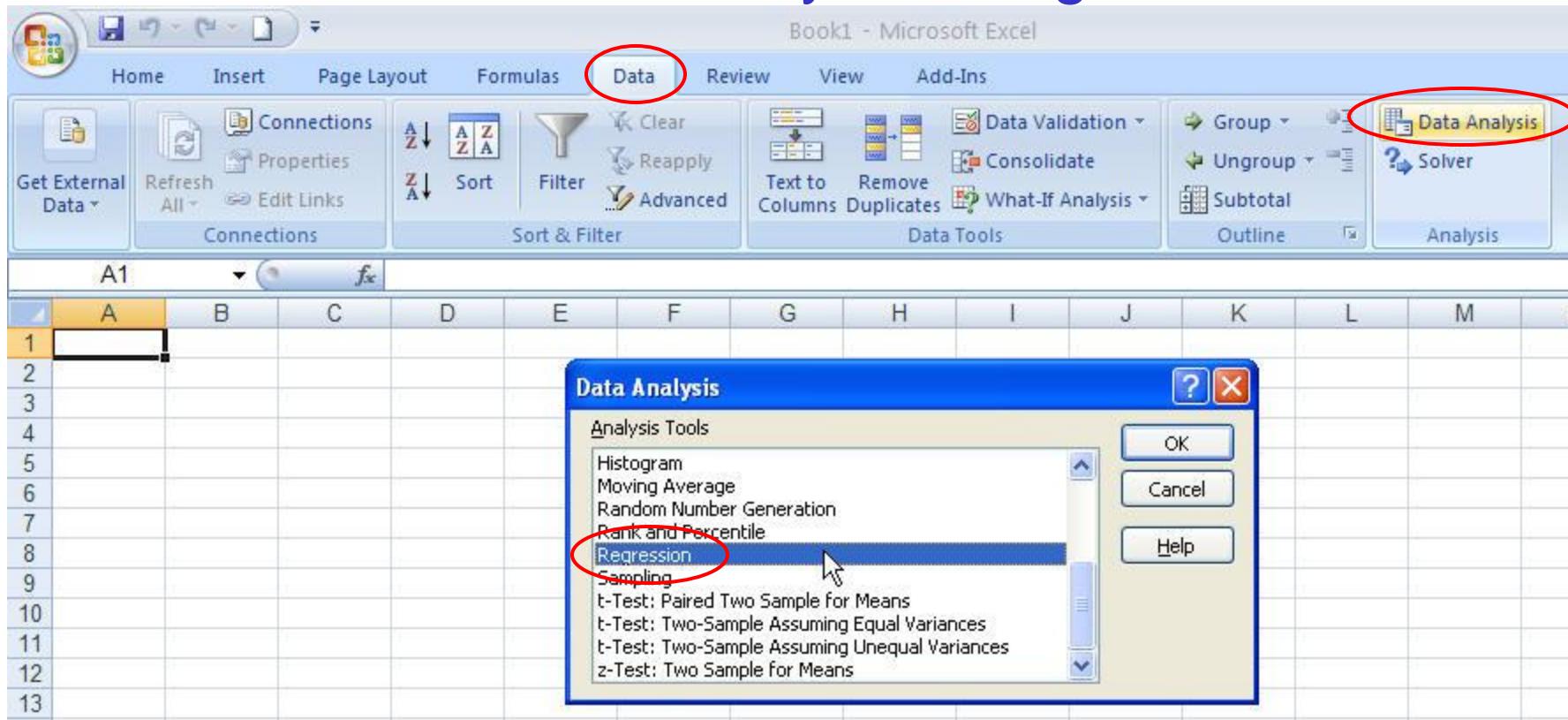


Estimating a Multiple Linear Regression Equation

- Computer software is generally used to generate the coefficients and measures of goodness of fit for multiple regression
- Excel:
 - Data / Data Analysis / Regression

Estimating a Multiple Linear Regression Equation

- Excel:
 - Data / Data Analysis / Regression



Multiple Regression Analysis— Example: Results from EXCEL

| | A | B | C | D | E | F | G | H | I | J | K | |
|----|------|------|-------|-----|------------------------------------------------|---|---|---|---|---|---|--|
| 1 | Cost | Temp | Insul | Age | SUMMARY OUTPUT | | | | | | | |
| 2 | 250 | 35 | 3 | 6 | <i>Regression Statistics</i> | | | | | | | |
| 3 | 360 | 29 | 4 | 10 | Multiple R 0.897 | | | | | | | |
| 4 | 165 | 36 | 7 | 3 | R Square 0.804 | | | | | | | |
| 5 | 43 | 60 | 6 | 9 | Adjusted R Square 0.767 | | | | | | | |
| 6 | 92 | 65 | 5 | 6 | Standard Error 51.049 | | | | | | | |
| 7 | 200 | 30 | 5 | 5 | Observations 20 | | | | | | | |
| 8 | 355 | 10 | 6 | 7 | | | | | | | | |
| 9 | 290 | 7 | 10 | 10 | | | | | | | | |
| 10 | 230 | 21 | 9 | 11 | <i>ANOVA</i> | | | | | | | |
| 11 | 120 | 55 | 2 | 5 | df | | | | | | | |
| 12 | 73 | 54 | 12 | 4 | Regression 3 171220.473 57073.491 21.901 0.000 | | | | | | | |
| 13 | 205 | 48 | 5 | 1 | Residual 16 41695.277 2605.955 | | | | | | | |
| 14 | 400 | 20 | 5 | 15 | Total 19 212915.750 | | | | | | | |
| 15 | 320 | 39 | 4 | 7 | | | | | | | | |
| 16 | 72 | 60 | 8 | 6 | <i>Coefficients</i> | | | | | | | |
| 17 | 272 | 20 | 5 | 8 | Intercept 427.194 59.601 7.168 0.000 | | | | | | | |
| 18 | 94 | 58 | 7 | 3 | Temp -4.583 0.772 -5.934 0.000 | | | | | | | |
| 19 | 190 | 40 | 8 | 11 | Insul -14.831 4.754 -3.119 0.007 | | | | | | | |
| 20 | 235 | 27 | 9 | 8 | Age 6.101 4.012 1.521 0.148 | | | | | | | |
| 21 | 139 | 30 | 7 | 5 | | | | | | | | |

Using the regression coefficients, the multiple regression equation to estimate heating cost is:

$$\hat{y} = 427.194 - 4.583x_1 - 14.831x_2 + 6.101x_3$$

Heating Cost = 427.194 – 4.583 (Mean Outside Temp) – 14.831(Inches of Attic Insulation) + 6.101(Age of the Furnace)



Multiple Regression Analysis-Example

$$\hat{y} = 427.194 - 4.583x_1 - 14.831x_2 + 6.101x_3$$

Heating Cost = 427.194 – 4.583 (Mean Outside Temp) – 14.831(Inches of Attic Insulation) + 6.101(Age of the Furnace)

Interpreting the Regression Coefficients

- The coefficient for x_1 (mean outside temp) is -4.583: For every unit increase in temperature, holding the other two independent variables constant, monthly heating cost is expected to decrease by \$4.583 .
- The coefficient for (attic insulation variable) x_2 is -14.831: For each additional inch of insulation, the cost to heat the home is expected to decline by \$14.83 per month.
- The coefficient for (age of the furnace) x_3 is 6.1: For each additional year older the furnace is, the cost is expected to increase by \$6.10 per month.
- How about the intercept?

Multiple Regression Analysis – Example: Estimating the Dependent Variable

$$\hat{y} = 427.194 - 4.583x_1 - 14.831x_2 + 6.101x_3$$

Heating Cost = $427.194 - 4.583$ (Mean Outside Temp) – 14.831 (Inches of Attic Insulation) + 6.101 (Age of the Furnace)

Applying the Model for Estimation

What is the estimated heating cost for a home if the mean outside temperature is 30 degrees, there are 5 inches of insulation in the attic, and the furnace is 10 years old?

Substituting the values for each of the independent variable, we can calculate the estimated cost to heat the home.

$$\hat{y} = 427.194 - 4.583(30) - 14.831(5) + 6.101(10) = 276.56$$

In this case, the predicted cost to heat is \$276.56 per month.

Multiple Regression Analysis: Evaluation -Coefficient of Multiple Determination (R^2)

COEFFICIENT OF MULTIPLE DETERMINATION The percent of variation in the dependent variable, y , explained by the set of independent variables, $x_1, x_2, x_3, \dots x_k$.

The characteristics of the coefficient of multiple determination are:

- It is symbolized by a capital R squared. In other words, it is written as R^2 because it behaves like the square of a correlation coefficient.
- It can range from 0 to 1. A value near 0 indicates little association between the set of independent variables and the dependent variable. A value near 1 means a strong association.
- It cannot assume negative values. Any number that is squared or raised to the second power cannot be negative.
- It is easy to interpret.

Multiple Regression Analysis: Evaluation

-Coefficient of Multiple Determination

$$R^2 = \frac{SSR}{SST} = \frac{\text{Sum of squares regression}}{\text{Total sum of squares}}$$

| ANOVA | | | | | |
|------------|----|------------|-----------|--------|----------------|
| | df | SS | MS | F | Significance F |
| Regression | 3 | 171220.473 | 57073.491 | 21.901 | 0.000 |
| Residual | 16 | 41695.277 | 2605.955 | | |
| Total | 19 | 212915.750 | | | |

$$R^2 = \frac{SSR}{SS \text{ total}} = \frac{171,220.473}{212,915.750} = .804$$

We interpret the R^2 of 0.804 as the percent of variation in heating cost that is explained or accounted for by the three independent variables. Specifically, 80.4% of the variation in heating cost is explained by the three independent variables. The remaining 19.6% is random error and variation from independent variables not included in the equation.

Multiple Regression Analysis: Evaluating Individual Regression Coefficients ($\beta_i = 0$)

These tests are used to determine which independent variables are significantly related to the dependent variable.

Step 1: State the hypotheses: $H_0: \beta_i = 0$; $H_1: \beta_i \neq 0$

Step 2: Decide on a level of significance: $\alpha = 0.05$

Step 3: Select the appropriate test statistic:

The test statistic is the t distribution with $n-(k+1)$ degrees of freedom.

Step 4: Formulate a decision rule:

Reject H_0 if $t > t_{\alpha/2, n-(k+1)}$ or $t < -t_{\alpha/2, n-(k+1)}$

Reject H_0 if $t > 2.120$ or $t < -2.120$

Multiple Regression Analysis: Evaluating Individual Regression Coefficients ($\beta_i = 0$)

Step 5: Take a sample, do the analysis, arrive at a decision.

| TESTING INDIVIDUAL REGRESSION COEFFICIENTS | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------|----------------|-----------|---------|----------------|----------------|--|--|--|--|--|-----------------------|--|--|--|--|--|------------|-------|--|--|--|--|----------|-------|--|--|--|--|-------------------|-------|--|--|--|--|----------------|--------|--|--|--|--|--------------|----|--|--|--|--|-------|--|--|--|--|--|--|----|----|----|---|----------------|------------|---|------------|-----------|--------|-------|----------|----|-----------|----------|--|--|-------|----|------------|--|--|--|--|--------------|----------------|--------|---------|--|-----------|---------|--------|-------|-------|--|------|--------|-------|--------|-------|--|-------|---------|-------|--------|-------|--|-----|-------|-------|-------|-------|--|
| $t = \frac{b_i - 0}{s_{b_i}}$ [14-6] | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th colspan="6" style="text-align: left;">SUMMARY OUTPUT</th> </tr> <tr> <th colspan="6" style="text-align: left;">Regression Statistics</th> </tr> </thead> <tbody> <tr> <td>Multiple R</td> <td>0.897</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>R Square</td> <td>0.804</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>Adjusted R Square</td> <td>0.767</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>Standard Error</td> <td>51.049</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>Observations</td> <td>20</td> <td></td> <td></td> <td></td> <td></td> </tr> </tbody> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th colspan="6" style="text-align: left;">ANOVA</th> </tr> <tr> <th></th> <th style="text-align: center;">df</th> <th style="text-align: center;">SS</th> <th style="text-align: center;">MS</th> <th style="text-align: center;">F</th> <th style="text-align: center;">Significance F</th> </tr> </thead> <tbody> <tr> <td>Regression</td> <td style="text-align: center;">3</td> <td style="text-align: center;">171220.473</td> <td style="text-align: center;">57073.491</td> <td style="text-align: center;">21.901</td> <td style="text-align: center;">0.000</td> </tr> <tr> <td>Residual</td> <td style="text-align: center;">16</td> <td style="text-align: center;">41695.277</td> <td style="text-align: center;">2605.955</td> <td></td> <td></td> </tr> <tr> <td>Total</td> <td style="text-align: center;">19</td> <td style="text-align: center;">212915.750</td> <td></td> <td></td> <td></td> </tr> </tbody> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th></th> <th style="text-align: center;">Coefficients</th> <th style="text-align: center;">Standard Error</th> <th style="text-align: center;">t Stat</th> <th style="text-align: center;">P-value</th> <th></th> </tr> </thead> <tbody> <tr> <td>Intercept</td> <td style="text-align: center;">427.194</td> <td style="text-align: center;">59.601</td> <td style="text-align: center;">7.168</td> <td style="text-align: center;">0.000</td> <td></td> </tr> <tr> <td>Temp</td> <td style="text-align: center;">-4.583</td> <td style="text-align: center;">0.772</td> <td style="text-align: center;">-5.934</td> <td style="text-align: center;">0.000</td> <td></td> </tr> <tr> <td>Insul</td> <td style="text-align: center;">-14.831</td> <td style="text-align: center;">4.754</td> <td style="text-align: center;">-3.119</td> <td style="text-align: center;">0.007</td> <td></td> </tr> <tr> <td>Age</td> <td style="text-align: center;">6.101</td> <td style="text-align: center;">4.012</td> <td style="text-align: center;">1.521</td> <td style="text-align: center;">0.148</td> <td></td> </tr> </tbody> </table> | | | | | | SUMMARY OUTPUT | | | | | | Regression Statistics | | | | | | Multiple R | 0.897 | | | | | R Square | 0.804 | | | | | Adjusted R Square | 0.767 | | | | | Standard Error | 51.049 | | | | | Observations | 20 | | | | | ANOVA | | | | | | | df | SS | MS | F | Significance F | Regression | 3 | 171220.473 | 57073.491 | 21.901 | 0.000 | Residual | 16 | 41695.277 | 2605.955 | | | Total | 19 | 212915.750 | | | | | Coefficients | Standard Error | t Stat | P-value | | Intercept | 427.194 | 59.601 | 7.168 | 0.000 | | Temp | -4.583 | 0.772 | -5.934 | 0.000 | | Insul | -14.831 | 4.754 | -3.119 | 0.007 | | Age | 6.101 | 4.012 | 1.521 | 0.148 | |
| SUMMARY OUTPUT | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Regression Statistics | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Multiple R | 0.897 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| R Square | 0.804 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Adjusted R Square | 0.767 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Standard Error | 51.049 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Observations | 20 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| ANOVA | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | df | SS | MS | F | Significance F | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Regression | 3 | 171220.473 | 57073.491 | 21.901 | 0.000 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Residual | 16 | 41695.277 | 2605.955 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Total | 19 | 212915.750 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | Coefficients | Standard Error | t Stat | P-value | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Intercept | 427.194 | 59.601 | 7.168 | 0.000 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Temp | -4.583 | 0.772 | -5.934 | 0.000 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Insul | -14.831 | 4.754 | -3.119 | 0.007 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Age | 6.101 | 4.012 | 1.521 | 0.148 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| For temperature: $H_0: \beta_1 = 0$ ← $t = \frac{-4.583}{0.772} = -5.934, \text{reject}$ $H_1: \beta_1 \neq 0$ <hr/> For insulation: $H_0: \beta_2 = 0$ ← $t = \frac{-14.831}{4.754} = -3.119, \text{reject}$ $H_1: \beta_2 \neq 0$ <hr/> For furnace age: $H_0: \beta_3 = 0$ ← $t = \frac{6.101}{4.012} = 1.521, \text{fail.to.reject}$ $H_1: \beta_3 \neq 0$ | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Multiple Regression Analysis: Evaluating Individual Regression Coefficients ($\beta_i = 0$)

Step 6: Interpret the results.

- The tests for each of the individual regression coefficients show that two independent variables, mean outside temperature and inches of insulation, are significantly related to heating cost. The results also show that the age of the furnace is not related to heating cost.
- In fact, the results suggest that we can eliminate the “age of furnace” variable from the analysis. We then redo the multiple regression analysis without the “age of furnace” variable and calculate a new equation using only the temperature and inches of insulation variables.

Multiple Regression Analysis: Evaluating Individual Regression Coefficients ($\beta_i = 0$)

Step 6: Interpret the results continued without furnace age.

| SUMMARY OUTPUT | | | | | |
|-----------------------|--------------|----------------|-----------|---------|----------------|
| Regression Statistics | | | | | |
| Multiple R | 0.881 | | | | |
| R Square | 0.776 | | | | |
| Adjusted R Square | 0.749 | | | | |
| Standard Error | 52.982 | | | | |
| Observations | 20 | | | | |
| ANOVA | | | | | |
| | df | SS | MS | F | Significance F |
| Regression | 2 | 165194.521 | 82597.261 | 29.424 | 0.000 |
| Residual | 17 | 47721.229 | 2807.131 | | |
| Total | 19 | 212915.750 | | | |
| | Coefficients | Standard Error | t Stat | P-value | |
| Intercept | 490.286 | 44.410 | 11.040 | 0.000 | |
| Temp | -5.150 | 0.702 | -7.337 | 0.000 | |
| Insul | -14.718 | 4.934 | -2.983 | 0.008 | |

What happened when furnace age is removed? Using all the previous steps in multiple regression analysis we find that:

- The Global Test is rejected; at least one of the independent variables is significantly related to heating cost.
- The Individual Tests of Regression Coefficients (Temperature and Insulation) reject the null hypotheses and we conclude that both mean temperature and inches of insulation are significantly related to heating cost.
- The multiple regression equation is now:

$$\text{Heating Cost} = 490.286 - 5.150 \text{ (mean outside temperature)} - 14.718 \text{ (inches of insulation)}$$

- The adjusted R^2 indicates the two independent variables account for 74.9% of the variation in heating cost.

Nonparametric Methods: Chi-Square Applications



Chapter 15

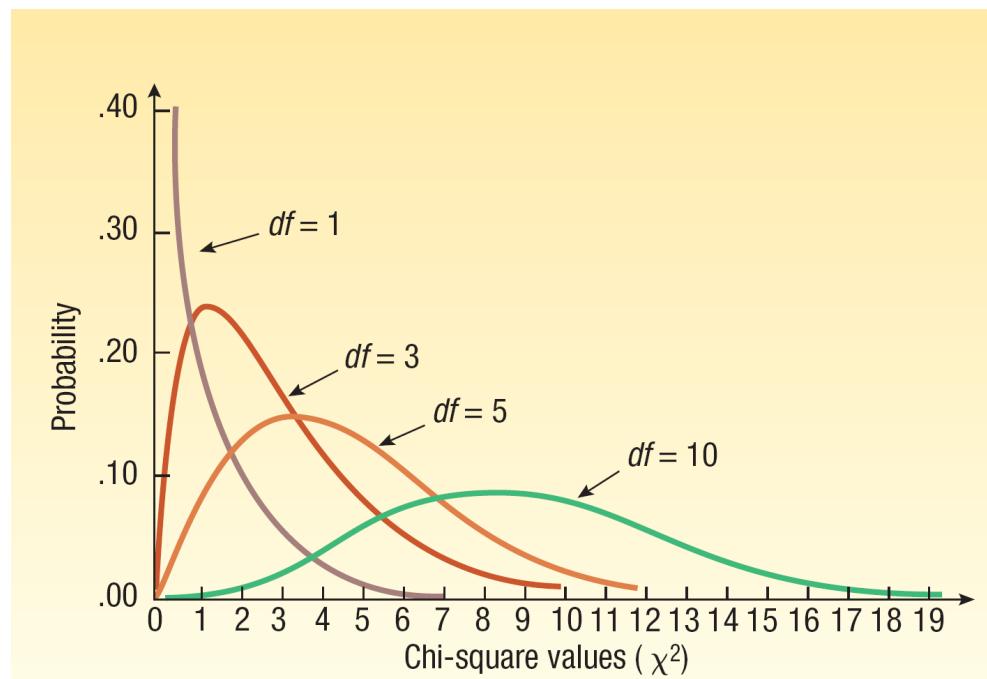
GOALS

1. List the characteristics of the *chi-square distribution*.
2. Conduct a test of hypothesis comparing an observed set of frequencies to an expected distribution.
3. Conduct a test of hypothesis to determine whether two classification criteria are related.

Characteristics of the Chi-Square Distribution

The major characteristics of the chi-square distribution

- It is positively skewed.
- It is non-negative.
- It is based on degrees of freedom.
- When the degrees of freedom change a new distribution is created.



Goodness-of-Fit Test: Equal Expected Frequencies

- Let f_0 and f_e be the observed and expected frequencies respectively.
- H_0 : There is no difference between the observed and expected frequencies.
- H_1 : There is a difference between the observed and the expected frequencies.

Goodness-of-fit Test: Equal Expected Frequencies

The test statistic is:

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

The critical value is a chi-square value with $(k-1)$ degrees of freedom, where k is the number of categories

Goodness-of-Fit Example

Ms. Jan Kilpatrick is the marketing manager for a manufacturer of sports cards. She plans to begin selling a series of cards with pictures and playing statistics of former Major League Baseball players. One of the problems is the selection of the former players. At a baseball card show at Southwyck Mall last weekend, she set up a booth and offered cards of the following six Hall of Fame baseball players: Tom Seaver, Nolan Ryan, Ty Cobb, George Brett, Hank Aaron, and Johnny Bench. At the end of the day she sold a total of 120 cards. The number of cards sold for each old-time player is shown in the table on the right.

Can she conclude the sales are not the same for each player? Use 0.05 significance level.



| Player | Cards Sold |
|--------------|------------|
| Tom Seaver | 13 |
| Nolan Ryan | 33 |
| Ty Cobb | 14 |
| George Brett | 7 |
| Hank Aaron | 36 |
| Johnny Bench | 17 |
| Total | 120 |

Goodness-of-Fit Example

Step 1: State the null hypothesis and the alternate hypothesis.

H_0 : there is no difference between f_o and f_e

H_1 : there is a difference between f_o and f_e

Step 2: Select the level of significance.

$\alpha = 0.05$ as stated in the problem

Step 3: Select the test statistic.

The test statistic follows the chi-square distribution, designated as χ^2

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

Goodness-of-Fit Example

Step 4: Formulate the decision rule.

Reject H_0 if $\chi^2 > \chi^2_{\alpha, k-1}$

$$\sum \left[\frac{(f_o - f_e)^2}{f_e} \right] > \chi^2_{\alpha, k-1}$$

$$\sum \left[\frac{(f_o - f_e)^2}{f_e} \right] > \chi^2_{.05, 6-1}$$

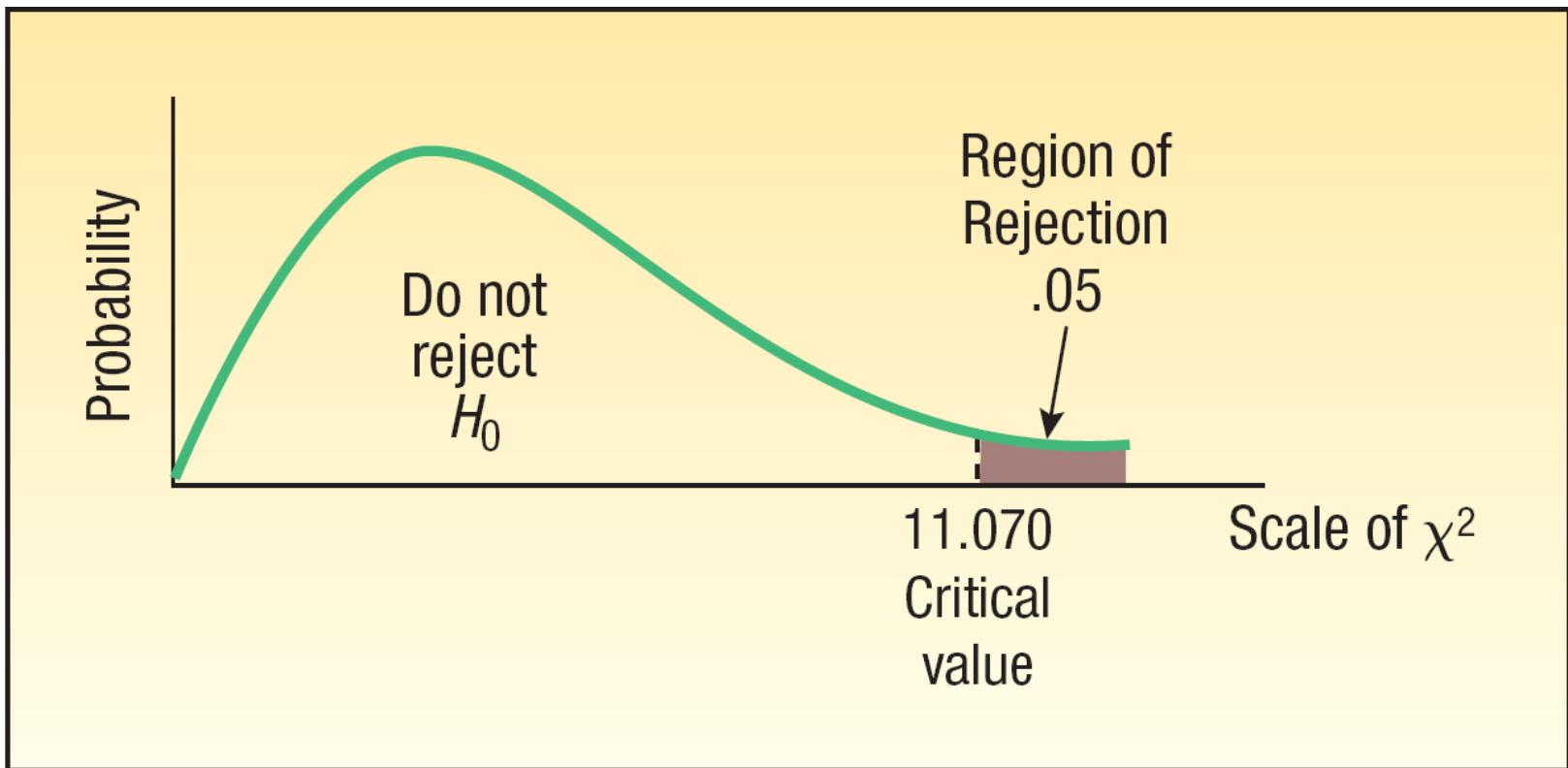
$$\sum \left[\frac{(f_o - f_e)^2}{f_e} \right] > \chi^2_{.05, 5}$$

$$\sum \left[\frac{(f_o - f_e)^2}{f_e} \right] > 11.070$$

A Portion of the Chi-Square Table

| Degrees of Freedom <i>df</i> | Right-Tail Area | | | | |
|---------------------------------|-----------------|--------|--------|--------|--|
| | .10 | .05 | .02 | .01 | |
| 1 | 2.706 | 3.841 | 5.412 | 6.635 | |
| 2 | 4.605 | 5.991 | 7.824 | 9.210 | |
| 3 | 6.251 | 7.815 | 9.837 | 11.345 | |
| 4 | 7.779 | 9.488 | 11.668 | 13.277 | |
| 5 | 9.236 | 11.070 | 13.388 | 15.086 | |

Goodness-of-Fit Example



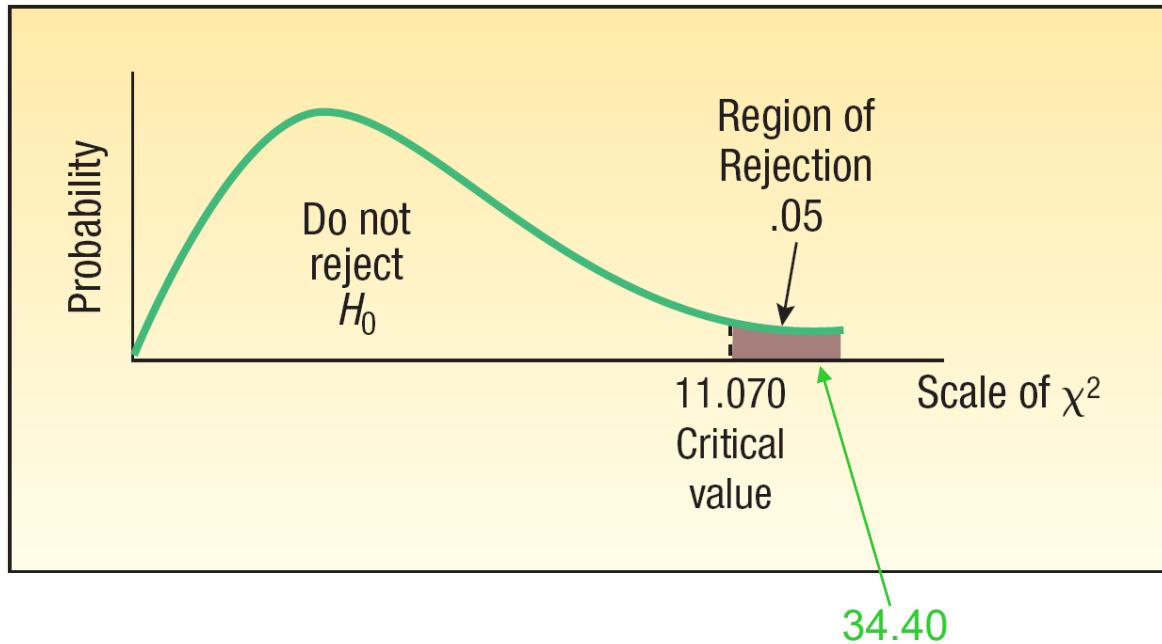
Goodness-of-Fit Example

Step 5: Compute the value of the Chisquare statistic and make a decision

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

| Baseball Player | f_o | f_e | (1) $(f_o - f_e)$ | (2) $(f_o - f_e)^2$ | (3) $\frac{(f_o - f_e)^2}{f_e}$ |
|-----------------|-------|-------|----------------------|------------------------|------------------------------------|
| Tom Seaver | 13 | 20 | -7 | 49 | 49/20 = 2.45 |
| Nolan Ryan | 33 | 20 | 13 | 169 | 169/20 = 8.45 |
| Ty Cobb | 14 | 20 | -6 | 36 | 36/20 = 1.80 |
| George Brett | 7 | 20 | -13 | 169 | 169/20 = 8.45 |
| Hank Aaron | 36 | 20 | 16 | 256 | 256/20 = 12.80 |
| Johnny Bench | 17 | 20 | -3 | 9 | 9/20 = 0.45 |
| | | | 0 | | 34.40 |
| Must be | | | | χ^2 | |

Goodness-of-Fit Example



The computed χ^2 of 34.40 is in the rejection region, beyond the critical value of 11.070. The decision, therefore, is to reject H_0 at the .05 level .

Conclusion: The difference between the observed and the expected frequencies is not due to chance. Rather, the differences between f_o and f_e are large enough to be considered significant. It is unlikely that card sales are the same among the six players.

Chisquare - MegaStat

The screenshot shows a Microsoft Excel window titled "Microsoft Excel - Book1". The menu bar includes File, Edit, View, Insert, Format, Tools, MegaStat, Data, Window, and Help. The ribbon tabs show "Anal" (Analysis) selected. The worksheet contains the following data:

| | A | B | C | D | E | F | G | H | I | J |
|----|----------------------|----------|---------|--------------------------|------------|---|---|---|---|---|
| 1 | | | | | | | | | | |
| 2 | Goodness of Fit Test | | | | | | | | | |
| 3 | | | | | | | | | | |
| 4 | observed | expected | O - E | (O - E) ² / E | % of chisq | | | | | |
| 5 | 13 | 20.000 | -7.000 | 2.450 | 7.12 | | | | | |
| 6 | 33 | 20.000 | 13.000 | 8.450 | 24.56 | | | | | |
| 7 | 14 | 20.000 | -6.000 | 1.800 | 5.23 | | | | | |
| 8 | 7 | 20.000 | -13.000 | 8.450 | 24.56 | | | | | |
| 9 | 36 | 20.000 | 16.000 | 12.800 | 37.21 | | | | | |
| 10 | 17 | 20.000 | -3.000 | 0.450 | 1.31 | | | | | |
| 11 | 120 | 120.000 | 0.000 | 34.400 | 100.00 | | | | | |
| 12 | | | | | | | | | | |
| 13 | | | | | | | | | | |
| 14 | | | | | | | | | | |
| 15 | | | | | | | | | | |
| 16 | | | | | | | | | | |
| 17 | | | | | | | | | | |
| 18 | | | | | | | | | | |
| 19 | | | | | | | | | | |
| 20 | | | | | | | | | | |
| 21 | | | | | | | | | | |
| 22 | | | | | | | | | | |
| 23 | | | | | | | | | | |
| 24 | | | | | | | | | | |

Below the table, the output summary is:

34.40 chi-square
5 df
1.98E-06 p-value

The status bar at the bottom shows "Ready", "Microsoft Excel - Book1", "Chapter15-shots", "54%", "3:06 PM", and a battery icon.

Goodness-of-Fit Test: Unequal Expected Frequencies

- Let f_0 and f_e be the observed and expected frequencies respectively.
- H_0 : There is no difference between the observed and expected frequencies.
- H_1 : There is a difference between the observed and the expected frequencies.

Goodness-of-Fit Test: Unequal Expected Frequencies - Example

The American Hospital Administrators Association (AHAA) reports the following information concerning the number of times senior citizens are admitted to a hospital during a one-year period. Forty percent are not admitted; 30 percent are admitted once; 20 percent are admitted twice, and the remaining 10 percent are admitted three or more times.

A survey of 150 residents of Bartow Estates, a community devoted to active seniors located in central Florida, revealed 55 residents were not admitted during the last year, 50 were admitted to a hospital once, 32 were admitted twice, and the rest of those in the survey were admitted three or more times.

Can we conclude the survey at Bartow Estates is consistent with the information suggested by the AHAA? Use the .05 significance level.

Goodness-of-Fit Test: Unequal Expected Frequencies - Example

Step 1: State the null hypothesis and the alternate hypothesis.

H_0 : There is no difference between local and national experience for hospital admissions.

H_1 : There is a difference between local and national experience for hospital admissions.

Step 2: Select the level of significance.

$\alpha = 0.05$ as stated in the problem

Step 3: Select the test statistic.

The test statistic follows the chi-square distribution, designated as χ^2

CHI-SQUARE TEST STATISTIC

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

Goodness-of-Fit Test: Unequal Expected Frequencies - Example

Step 4: Formulate the decision rule.

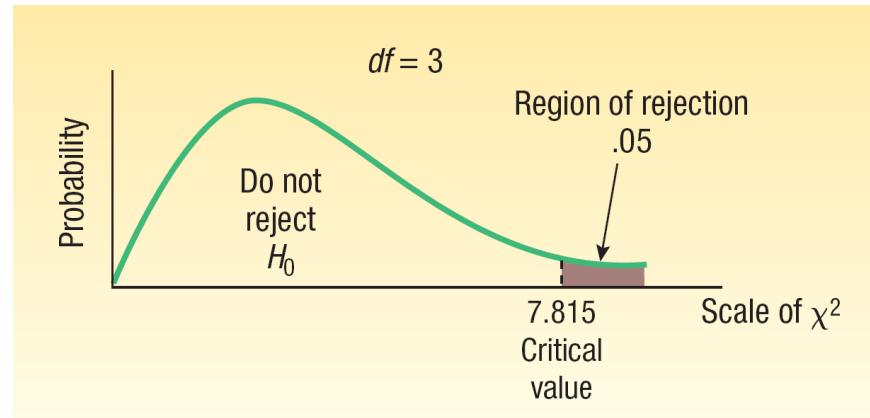
Reject H_0 if $\chi^2 > \chi^2_{\alpha, k-1}$

$$\sum \left[\frac{(f_o - f_e)^2}{f_e} \right] > \chi^2_{\alpha, k-1}$$

$$\sum \left[\frac{(f_o - f_e)^2}{f_e} \right] > \chi^2_{.05, 4-1}$$

$$\sum \left[\frac{(f_o - f_e)^2}{f_e} \right] > \chi^2_{.05, 3}$$

$$\sum \left[\frac{(f_o - f_e)^2}{f_e} \right] > 7.815$$



Goodness-of-Fit Test: Unequal Expected Frequencies - Example

| Number of Times Admitted | AHAA Percent of Total | Number of Bartow Residents (f_o) | Expected Number of Residents (f_e) |
|--------------------------|-----------------------|--------------------------------------|----------------------------------------|
| 0 | 40 | 55 | 60 |
| 1 | 30 | 50 | 45 |
| 2 | 20 | 32 | 30 |
| 3 or more | 10 | 13 | 15 |
| Total | 100 | $\frac{150}{150}$ | $\frac{150}{150}$ |

Computation of f_e
 $0.40 \times 150 = 60$
 $0.30 \times 150 = 45$
 $0.30 \times 150 = 30$
 $0.10 \times 150 = 15$

Goodness-of-Fit Test: Unequal Expected Frequencies - Example

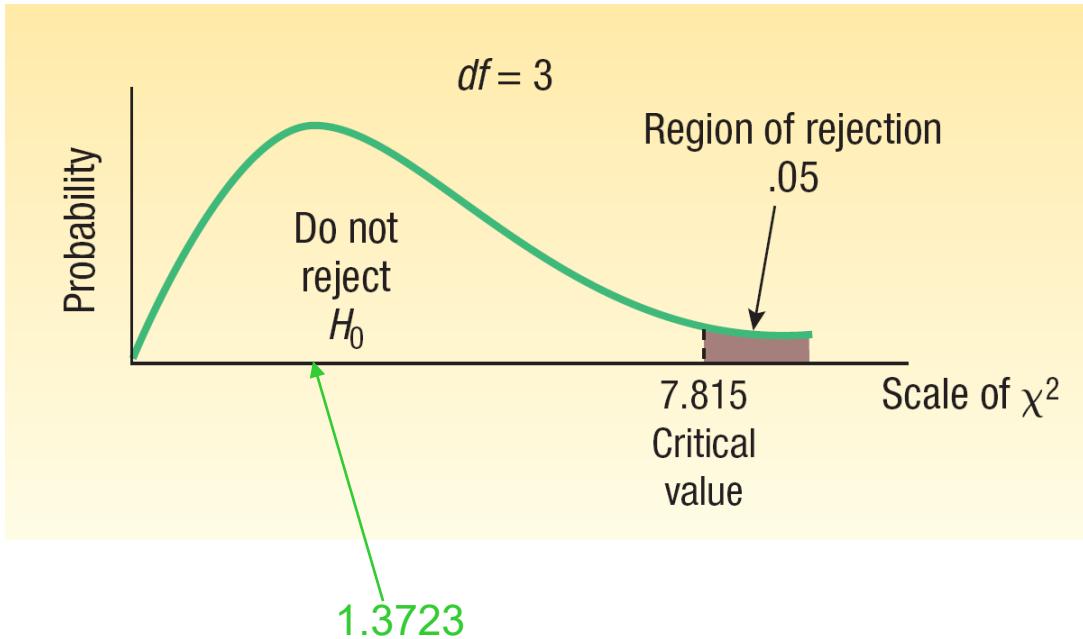
Step 5: Compute the value of the Chisquare statistic and make a decision

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

| Number of Times Admitted | (f_o) | (f_e) | $f_o - f_e$ | $(f_o - f_e)^2/f_e$ |
|--------------------------|---------|---------|-------------|---------------------|
| 0 | 55 | 60 | -5 | 0.4167 |
| 1 | 50 | 45 | 5 | 0.5556 |
| 2 | 32 | 30 | 2 | 0.1333 |
| 3 or more | 13 | 15 | -2 | 0.2667 |
| Total | 150 | 150 | 0 | 1.3723 |

Computed χ^2

Goodness-of-Fit Test: Unequal Expected Frequencies - Example



The computed χ^2 of 1.3723 is in the “Do not reject H_0 ” region. The difference between the observed and the expected frequencies is due to chance.

We conclude that there is no evidence of a difference between the local and national experience for hospital admissions.

Contingency Table Analysis

A **contingency table** is used to investigate whether two traits or characteristics are related. Each observation is classified according to two criteria. We use the usual hypothesis testing procedure.

- The **degrees of freedom** is equal to:
(number of rows-1)(number of columns-1).
- The **expected frequency** is computed as:

EXPECTED FREQUENCY

$$f_e = \frac{(\text{Row total})(\text{Column total})}{\text{Grand total}}$$

Contingency Analysis

We can use the chi-square statistic to formally test for a relationship between two nominal-scaled variables. To put it another way, Is one variable *independent* of the other?

- Ford Motor Company operates an assembly plant in Dearborn, Michigan. The plant operates three shifts per day, 5 days a week. The quality control manager wishes to compare the quality level on the three shifts. Vehicles are classified by quality level (acceptable, unacceptable) and shift (day, afternoon, night). Is there a difference in the quality level on the three shifts? That is, is the quality of the product related to the shift when it was manufactured? Or is the quality of the product independent of the shift on which it was manufactured?
- A sample of 100 drivers who were stopped for speeding violations was classified by gender and whether or not they were wearing a seat belt. For this sample, is wearing a seatbelt related to gender?
- Does a male released from federal prison make a different adjustment to civilian life if he returns to his hometown or if he goes elsewhere to live? The two variables are adjustment to civilian life and place of residence. Note that both variables are measured on the nominal scale.

Contingency Analysis - Example

The Federal Correction Agency is investigating the “Does a male released from federal prison make a different adjustment to civilian life if he returns to his hometown or if he goes elsewhere to live?” To put it another way, is there a relationship between adjustment to civilian life and place of residence after release from prison? Use the .01 significance level.

Contingency Analysis - Example

The agency's psychologists interviewed 200 randomly selected former prisoners. Using a series of questions, the psychologists classified the adjustment of each individual to civilian life as outstanding, good, fair, or unsatisfactory.

The classifications for the 200 former prisoners were tallied as follows. Joseph Camden, for example, returned to his hometown and has shown outstanding adjustment to civilian life. His case is one of the 27 tallies in the upper left box (circled).

| Residence after Release from Prison | Adjustment to Civilian Life | | | |
|-------------------------------------|-----------------------------|---------------------|---------------------|----------------|
| | Outstanding | Good | Fair | Unsatisfactory |
| Hometown | | | | |
| | | | | |
| Not hometown | | | | |

| Residence after Release from Prison | Adjustment to Civilian Life | | | | Total |
|-------------------------------------|-----------------------------|------|------|----------------|-------|
| | Outstanding | Good | Fair | Unsatisfactory | |
| Hometown | 27 | 35 | 33 | 25 | 120 |
| Not hometown | 13 | 15 | 27 | 25 | 80 |
| Total | 40 | 50 | 60 | 50 | 200 |

Contingency Analysis - Example

Step 1: State the null hypothesis and the alternate hypothesis.

H_0 : There is no relationship between adjustment to civilian life and where the individual lives after being released from prison.

H_1 : There is a relationship between adjustment to civilian life and where the individual lives after being released from prison.

Step 2: Select the level of significance.

$\alpha = 0.01$ as stated in the problem

Step 3: Select the test statistic.

The test statistic follows the chi-square distribution, designated as χ^2

CHI-SQUARE TEST STATISTIC

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

Contingency Analysis - Example

Step 4: Formulate the decision rule.

Reject H_0 if $\chi^2 > \chi^2_{\alpha,(r-1)(c-1)}$

$$\sum \left[\frac{(f_o - f_e)^2}{f_e} \right] > \chi^2_{\alpha,(2-1)(4-1)}$$

$$\sum \left[\frac{(f_o - f_e)^2}{f_e} \right] > \chi^2_{.01,(1)(3)}$$

$$\sum \left[\frac{(f_o - f_e)^2}{f_e} \right] > \chi^2_{.01,3}$$

$$\sum \left[\frac{(f_o - f_e)^2}{f_e} \right] > 11.345$$

Computing Expected Frequencies (f_e)

EXPECTED FREQUENCY

$$f_e = \frac{(\text{Row total})(\text{Column total})}{\text{Grand total}}$$

| Residence after Release from Prison | Adjustment to Civilian Life | | | | | | | | | | Total | |
|-------------------------------------------|-----------------------------|-------|-------|-------|-------|-------|----------------|-------|-------|-------|-------|--|
| | Outstanding | | Good | | Fair | | Unsatisfactory | | | | | |
| | f_o | f_e | f_o | f_e | f_o | f_e | f_o | f_e | f_o | f_e | | |
| Hometown | 27 | 24 | 35 | 30 | 33 | 36 | 25 | 30 | 120 | 120 | | |
| Not hometown | 13 | 16 | 15 | 20 | 27 | 24 | 25 | 20 | 80 | 80 | | |
| Total | 40 | 40 | 50 | 50 | 60 | 60 | 50 | 50 | 200 | 200 | | |

$\frac{(120)(50)}{200}$

$\frac{(80)(50)}{200}$

Must be equal

Computing the Chisquare Statistic

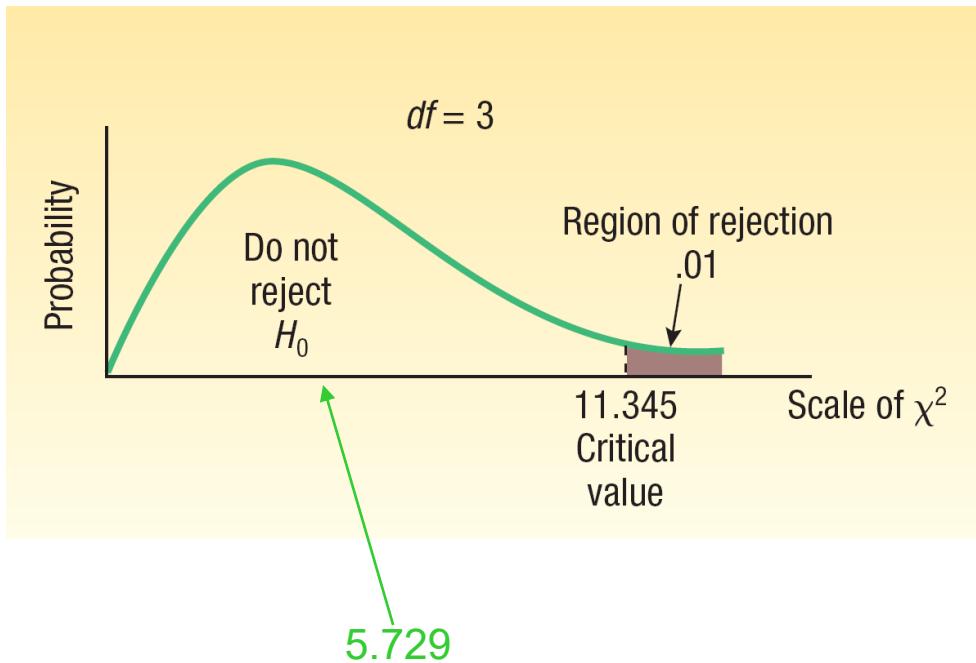
| Residence after Release from Prison | Adjustment to Civilian Life | | | | | | | | Total | |
|-------------------------------------------|-----------------------------|-------|-------|-------|-------|-------|----------------|-------|-------|-------|
| | Outstanding | | Good | | Fair | | Unsatisfactory | | | |
| | f_o | f_e | f_o | f_e | f_o | f_e | f_o | f_e | f_o | f_e |
| Hometown | 27 | 24 | 35 | 30 | 33 | 36 | 25 | 30 | 120 | 120 |
| Not hometown | 13 | 16 | 15 | 20 | 27 | 24 | 25 | 20 | 80 | 80 |
| Total | 40 | 40 | 50 | 50 | 60 | 60 | 50 | 50 | 200 | 200 |

Starting with the upper left cell:

$$\chi^2 = \Sigma \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

$$\begin{aligned}
 \chi^2 &= \frac{(27 - 24)^2}{24} + \frac{(35 - 30)^2}{30} + \frac{(33 - 36)^2}{36} + \frac{(25 - 30)^2}{30} \\
 &\quad + \frac{(13 - 16)^2}{16} + \frac{(15 - 20)^2}{20} + \frac{(27 - 24)^2}{24} + \frac{(25 - 20)^2}{20} \\
 &= 0.375 + 0.833 + 0.250 + 0.833 + 0.563 + 1.250 + 0.375 + 1.250 \\
 &= 5.729
 \end{aligned}$$

Conclusion



The computed χ^2 of 5.729 is in the “Do not rejection H_0 ” region. The null hypothesis is not rejected at the .01 significance level.

We conclude there is no evidence of a relationship between adjustment to civilian life and where the prisoner resides after being released from prison. For the Federal Correction Agency’s advisement program, adjustment to civilian life is not related to where the ex-prisoner lives.

Contingency Analysis - Minitab

The screenshot shows the Minitab software interface. The main window displays the results of a Chi-Square Test for a 2x4 contingency table. The test compares observed counts against expected counts, with contributions printed below the expected values. The Minitab session window also shows a portion of a worksheet titled 'C1-T'.

Chi-Square Test: Outstanding, Good, Fair, POOR

Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

| | Outstanding | Good | Fair | POOR | Total |
|-------|-------------|-------|-------|-------|-------|
| 1 | 27 | 35 | 33 | 25 | 120 |
| | 24.00 | 30.00 | 36.00 | 30.00 | |
| | 0.375 | 0.833 | 0.250 | 0.833 | |
| 2 | 13 | 15 | 27 | 25 | 80 |
| | 16.00 | 20.00 | 24.00 | 20.00 | |
| | 0.563 | 1.250 | 0.375 | 1.250 | |
| Total | 40 | 50 | 60 | 50 | 200 |

Chi-Sq = 5.729, DF = 3, P-Value = 0.126

Worksheet 1 ***

| | C1-T | C2 | C3 | C4 | C5 | |
|---|--------------|-------------|------|------|------|--|
| | Residence | Outstanding | Good | Fair | POOR | |
| 1 | Hometown | 27 | 35 | 33 | 25 | |
| 2 | Not Hometown | 13 | 15 | 27 | 25 | |
| 3 | | | | | | |
| 4 | | | | | | |
| 5 | | | | | | |
| 6 | | | | | | |

Current Worksheet: Worksheet 1 4:20 PM

Start Chapter17x MINITAB - Untitled Address <> 4:20 PM