

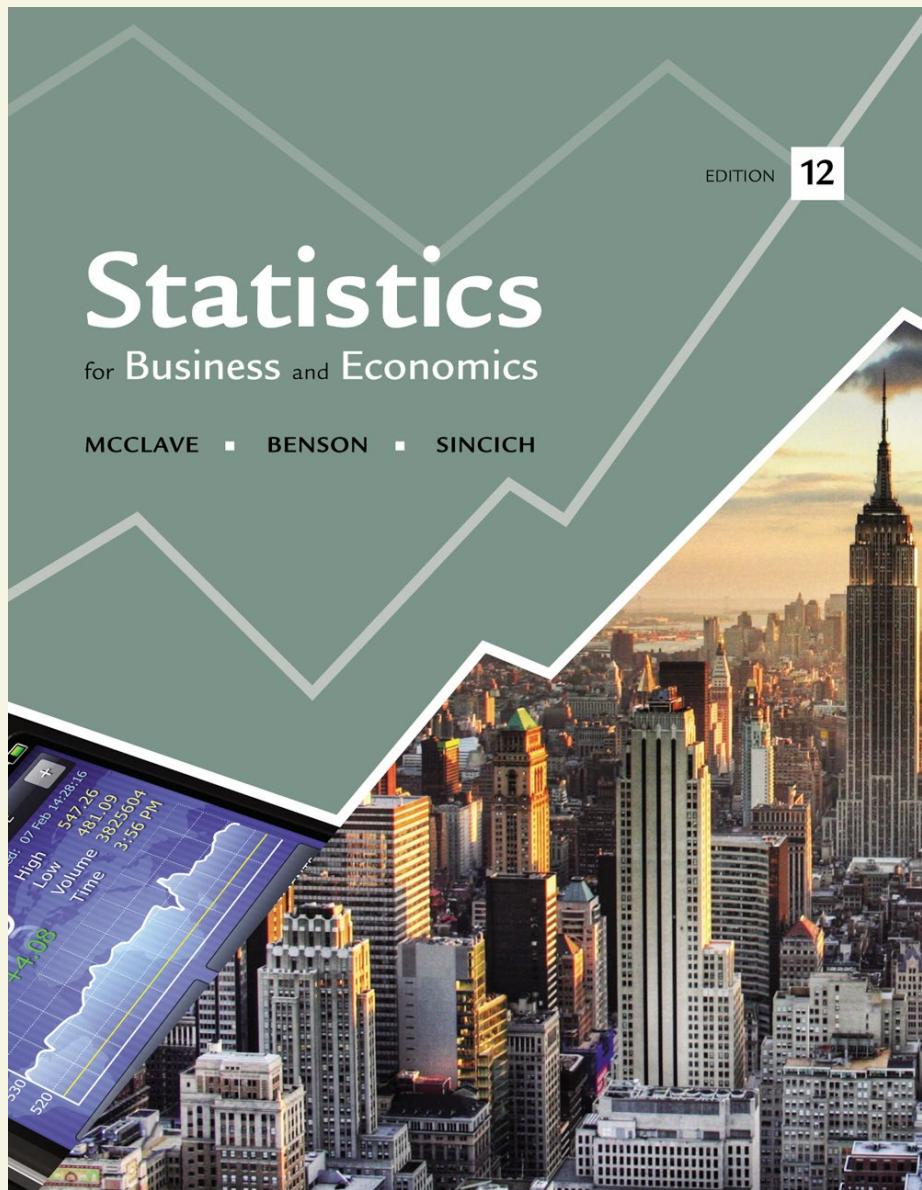
EDITION

12

# Statistics

for Business and Economics

MCCLAVE ■ BENSON ■ SINCICH



# **Statistics for Business and Economics**

## **Chapter 2**

### **Methods for Describing Sets of Data**

# Contents

1. Describing Qualitative Data
2. Graphical Methods for Describing Quantitative Data
3. Numerical Measures of Central Tendency
4. Numerical Measures of Variability
5. Using the Mean and Standard Deviation to Describe Data

# Contents

6. Methods for Detecting Outliers: Box Plots and  $z$ -scores
7. Graphing Bivariate Relationships
8. The Time Series Plot
9. Distorting the Truth with Descriptive Techniques

# 2.1

## Describing Qualitative Data

# Data Presentation

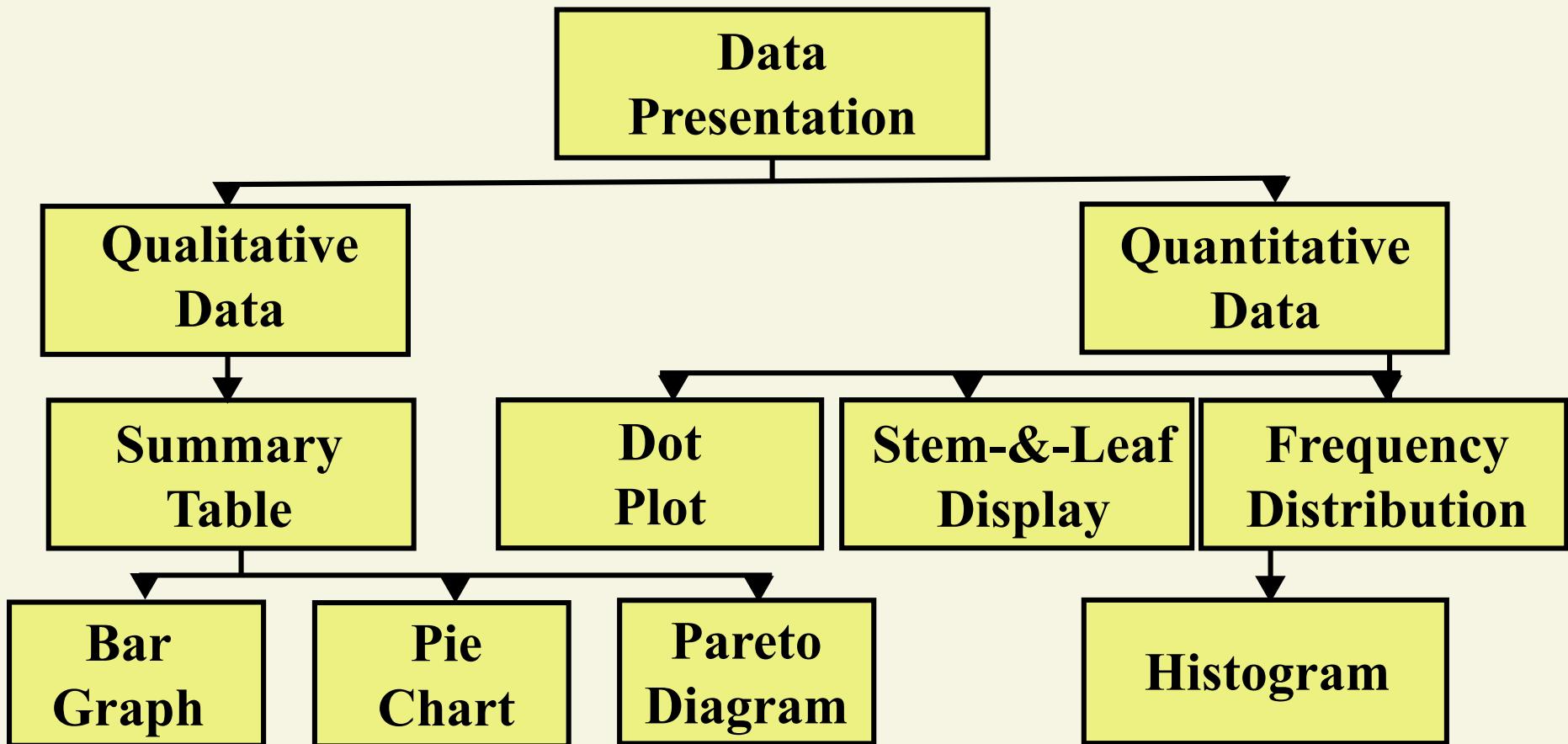


Table 2.1 Data on 40 Best-Paid Executives

	CEO	Company	Salary (\$ millions)	Age	Degree
1	Hemsley, Stephen	UnitedHealth Group	101.96	58	Bachelor's
2	Mueller, Edward	Qwest Communications	65.80	64	MBA
3	Iger, Robert	Walt Disney	53.32	60	Bachelor's
4	Paz, George	Express Scripts	51.52	56	Bachelor's
5	Frankfort, Lew	Coach	49.45	65	MBA
6	Lauren, Ralph	Polo Ralph Lauren	43.00	71	None
7	Martin, John	Gilead Sciences	42.72	59	PhD
8	Hackett, James	Anadarko Petroleum	38.94	57	MBA
9	Chambers, John	Cisco Systems	37.90	61	MBA
10	Seidenberg, Ivan	Verizon Commun	36.75	64	MBA
11	Pyott, David	Allergan	33.76	58	Master's
12	Lucier, Gregory	Life Technologies	33.75	46	MBA
13	Davidson, Charles	Noble Energy	33.44	61	Master's
14	Hammergren, John	McKesson	32.46	52	MBA
15	Tucci, Joseph	EMC	31.63	63	MBA
16	Huang, Jen-Hsun	Nvidia	31.41	48	Master's
17	Boyce, Gregory	Peabody Energy	30.66	56	Bachelor's
18	Merelli, F H	Cimarex Energy	30.53	75	None
19	Palmisano, Samuel	IBM	30.32	59	Bachelor's
20	Camilleri, Louis	Philip Morris Intl	30.09	56	Bachelor's
21	Watford, Michael	Ultra Petroleum	30.04	57	MBA
22	Schultz, Howard	Starbucks	29.73	57	Bachelor's
23	Novak, David	Yum Brands	29.67	58	Bachelor's
24	Thiry, Kent	DaVita	29.52	55	MBA
25	Farr, David	Emerson Electric	28.93	56	MBA
26	Cutler, Alexander	Eaton	28.47	59	MBA
27	Solomon, Howard	Forest Labs	27.10	83	Law
28	Moonves, Leslie	CBS	26.42	62	Bachelor's
29	Adkerson, Richard	Freeport Copper	25.30	64	MBA

# Summary Table

1. Lists categories & number of elements in category
2. Obtained by tallying responses in category
3. May show frequencies (counts), % or both

Row Is  
Category  
or Class

		DEGREE			
Valid	Category	Frequency	Percent	Valid Percent	Cumulative Percent
	Bachelors	13	32.5	32.5	32.5
	Law	1	2.5	2.5	35.0
	Masters	5	12.5	12.5	47.5
	MBA	15	37.5	37.5	85.0
	None	4	10.0	10.0	95.0
	PhD	2	5.0	5.0	100.0
	Total	40	100.0	100.0	

# Key Terms

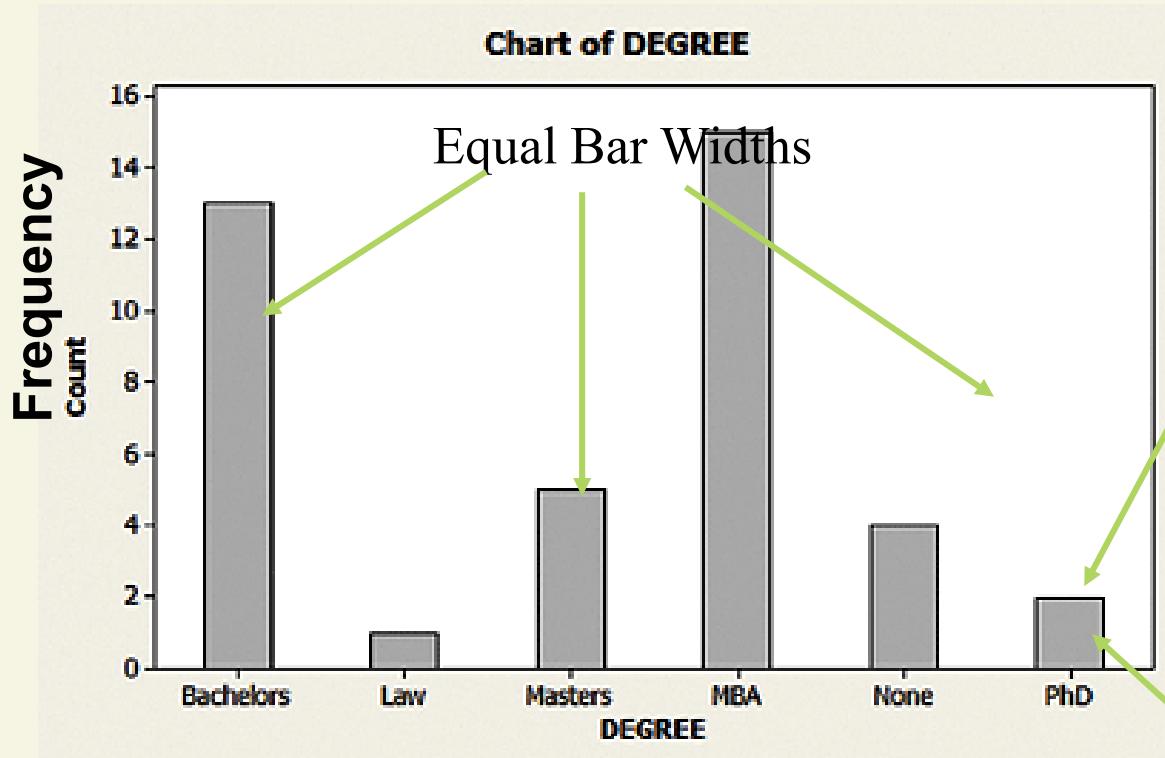
A **class** is one of the categories into which qualitative data can be classified.

The **class frequency** is the number of observations in the data set falling into a particular class.

The **class relative frequency** is the class frequency divided by the total numbers of observations in the data set.

# Bar Graph

Percent  
Used  
Also

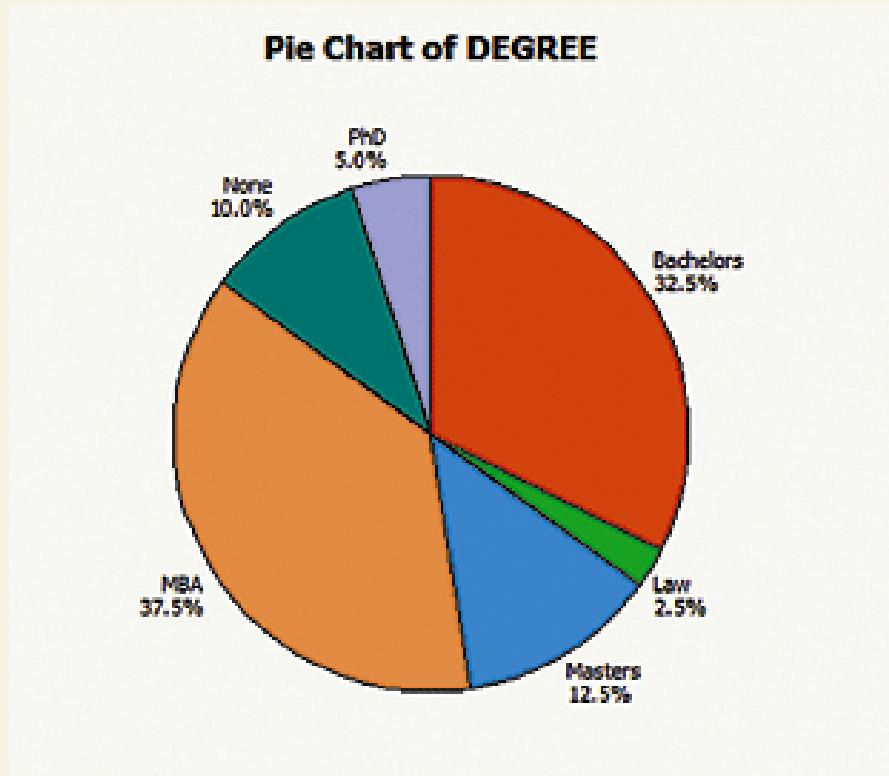


Bar Height  
Shows  
Frequency or %

Vertical Bars  
for Qualitative  
Variables

# Pie Chart

1. Shows breakdown of total quantity into categories
2. Useful for showing relative frequencies



# Summary

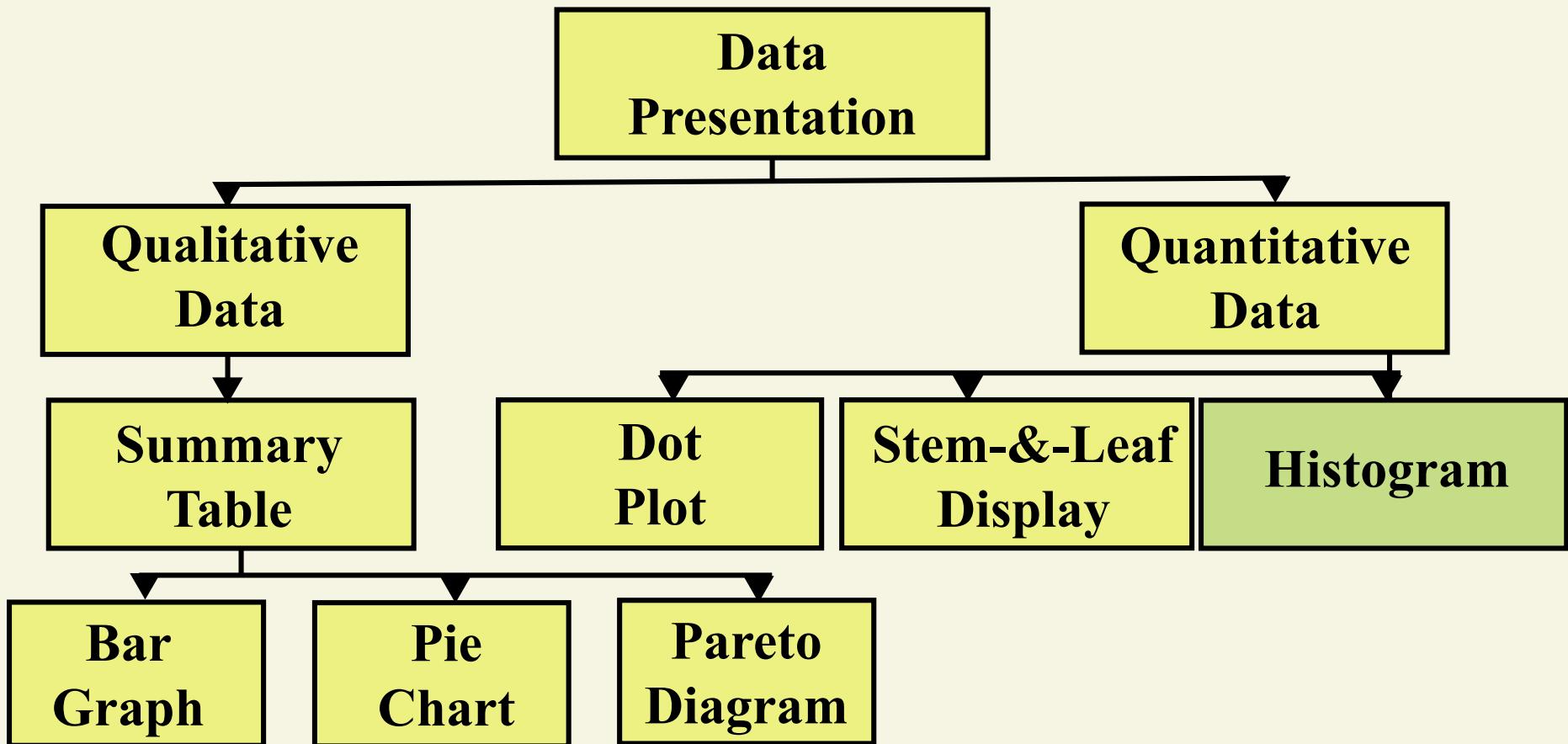
**Bar graph:** The categories (classes) of the qualitative variable are represented by bars, where the height of each bar is either the class frequency, class relative frequency, or class percentage.

**Pie chart:** The categories (classes) of the qualitative variable are represented by slices of a pie (circle). The size of each slice is proportional to the class relative frequency.

# **2.2**

# **Graphical Methods for Describing Quantitative Data**

# Data Presentation



# Histogram

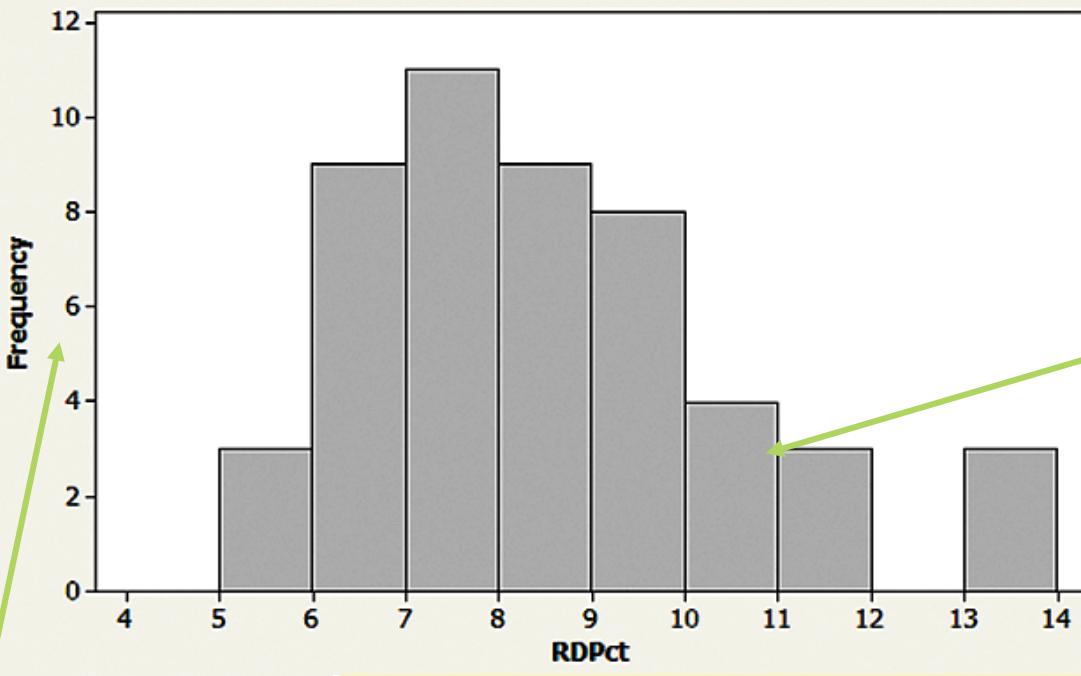
A financial analyst is interested in the amount of resources spent by computer hardware and software companies on R&D. She samples 50 of these high-technology firms and calculates the amount each spent last year on R&D as a percentage of their total revenue.

Table 2.2 Percentage of Revenues Spent on Research and Development

Company	Percentage	Company	Percentage	Company	Percentage	Company	Percentage
1	13.5	14	9.5	27	8.2	39	6.5
2	8.4	15	8.1	28	6.9	40	7.5
3	10.5	16	13.5	29	7.2	41	7.1
4	9.0	17	9.9	30	8.2	42	13.2
5	9.2	18	6.9	31	9.6	43	7.7
6	9.7	19	7.5	32	7.2	44	5.9
7	6.6	20	11.1	33	8.8	45	5.2
8	10.6	21	8.2	34	11.3	46	5.6
9	10.1	22	8.0	35	8.5	47	11.7
10	7.1	23	7.7	36	9.4	48	6.0
11	8.0	24	7.4	37	10.5	49	7.8
12	7.9	25	6.5	38	6.9	50	6.5
13	6.8	26	9.5				

# Histogram

Histogram of RDPct



Frequency

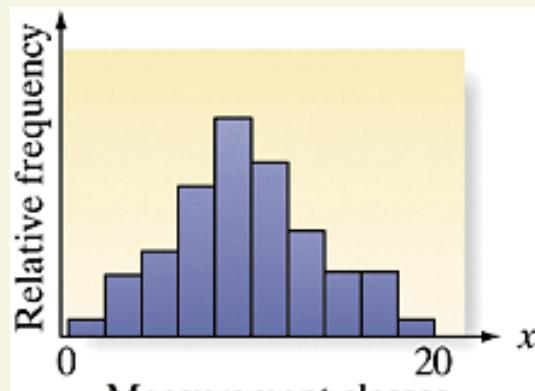
Relative  
Frequency

Percent

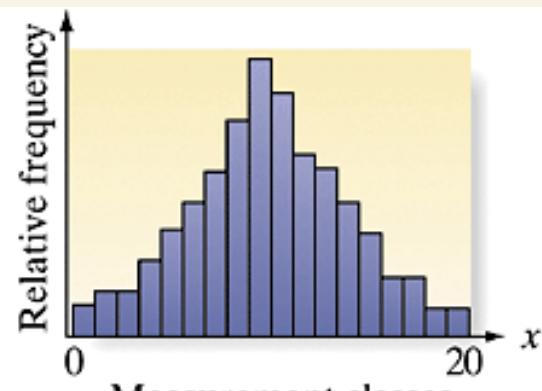
Bars  
Touch

Class	Class Interval	Class Frequency	Class Relative Frequency
1	5.0–6.0	3	$3/50 = .06$
2	6.0–7.0	9	$9/50 = .18$
3	7.0–8.0	11	$11/50 = .22$
4	8.0–9.0	9	$9/50 = .18$
5	9.0–10.0	8	$8/50 = .16$
6	10.0–11.0	4	$4/50 = .08$
7	11.0–12.0	3	$3/50 = .06$
8	12.0–13.0	0	$0/50 = .00$
9	13.0–14.0	3	$3/50 = .06$
Totals		50	1.00

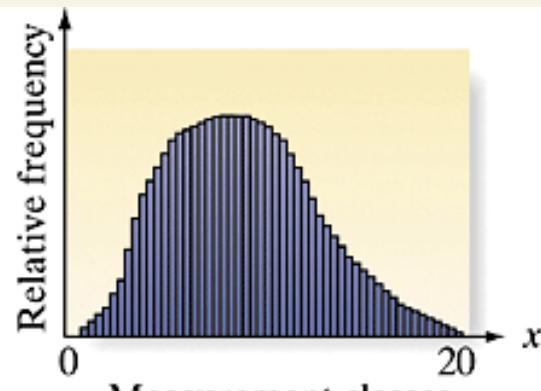
# Histogram



a. Small data set



b. Larger data set



c. Very large data set

# Summary

**Histogram:** The possible numerical values of the quantitative variable are partitioned into class intervals, where each interval has the same width. These intervals form the scale of the horizontal axis. The frequency or relative frequency of observations in each class interval is determined. A horizontal bar is placed over each class interval, with height equal to either the class frequency or class relative frequency.

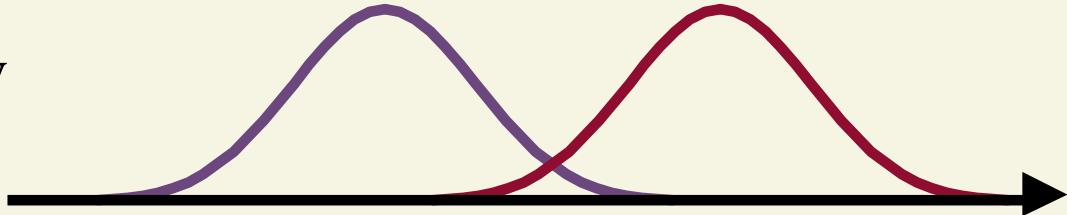
# **2.3**

## **Numerical Measures of Central Tendency**

# Two Characteristics

The **central tendency** of the set of measurements—that is, the tendency of the data to cluster, or center, about certain numerical values.

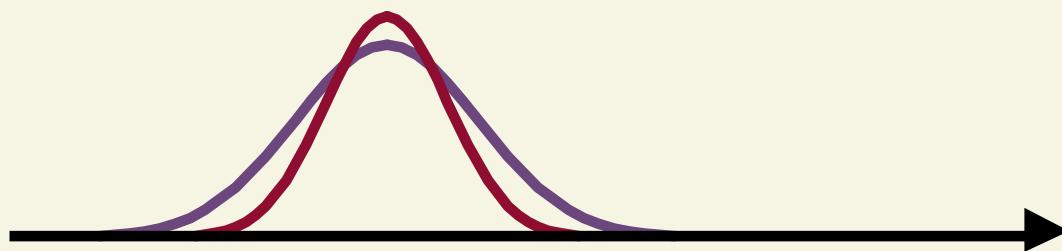
**Central Tendency  
(Location)**



# Two Characteristics

The **variability** of the set of measurements—that is, the spread of the data.

**Variation  
(Dispersion)**



# Standard Notation

Measure	Sample	Population
Mean	$\bar{X}$	$\mu$
Size	$n$	$N$

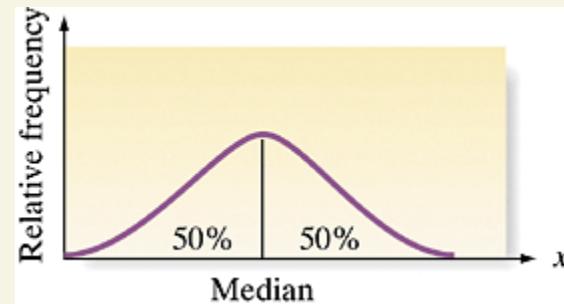
# Mean

1. Most common measure of central tendency
2. Acts as ‘balance point’
3. Affected by extreme values (‘outliers’)
4. Denoted  $\bar{x}$  where

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

# Median

1. Measure of central tendency
2. Middle value in ordered sequence
  - If  $n$  is odd, middle value of sequence
  - If  $n$  is even, average of 2 middle values
3. Position of median in sequence
$$\text{Positioning Point} = \frac{n + 1}{2}$$
4. Not affected by extreme values



# **2.4**

## **Numerical Measures of Variability**

# Variance & Standard Deviation

1. Measures of dispersion
2. Most common measures
3. Consider how data are distributed
4. Show variation about mean ( $\bar{x}$  or  $\mu$ )

# Standard Notation

Measure	Sample	Population
Mean	$\bar{x}$	$\mu$
Standard Deviation	$s$	$\sigma$
Variance	$s^2$	$\sigma^2$
Size	$n$	$N$

# Sample Variance and Standard Deviation

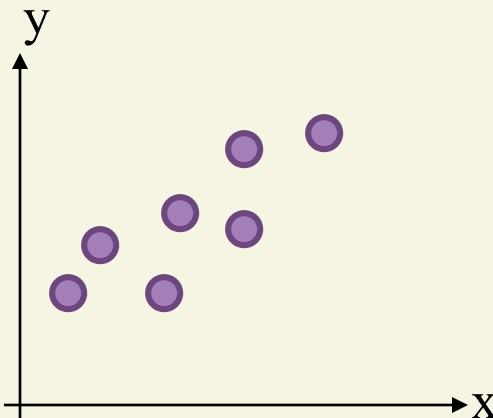
$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$
$$= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$
$$s = \sqrt{s^2}$$

# 2.8

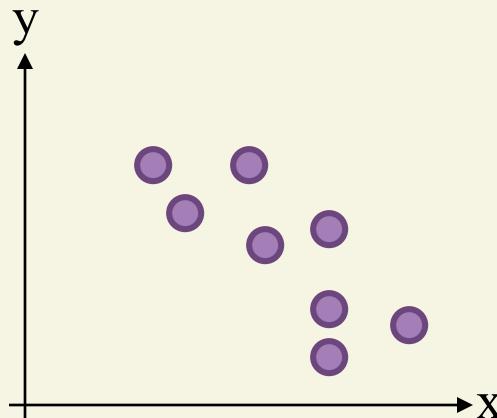
## Graphing Bivariate Relationships

# Graphing Bivariate Relationships

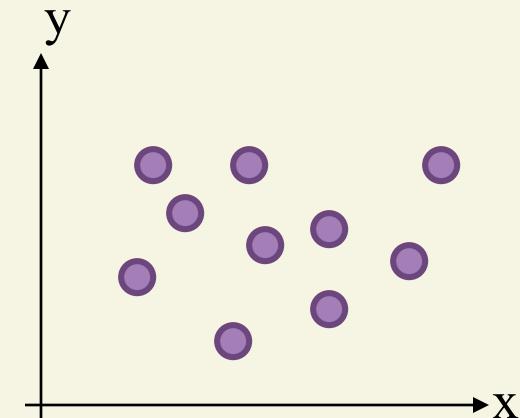
- Describes a relationship between two quantitative variables
- Plot the data in a scattergram (or scatterplot)



Positive  
relationship



Negative  
relationship



No  
relationship

# 2.9

## The Time Series Plot

# Time Series Plot

- Used to graphically display data produced over time
- Shows trends and changes in the data over time
- Time recorded on the horizontal axis
- Measurements recorded on the vertical axis
- Points connected by straight lines

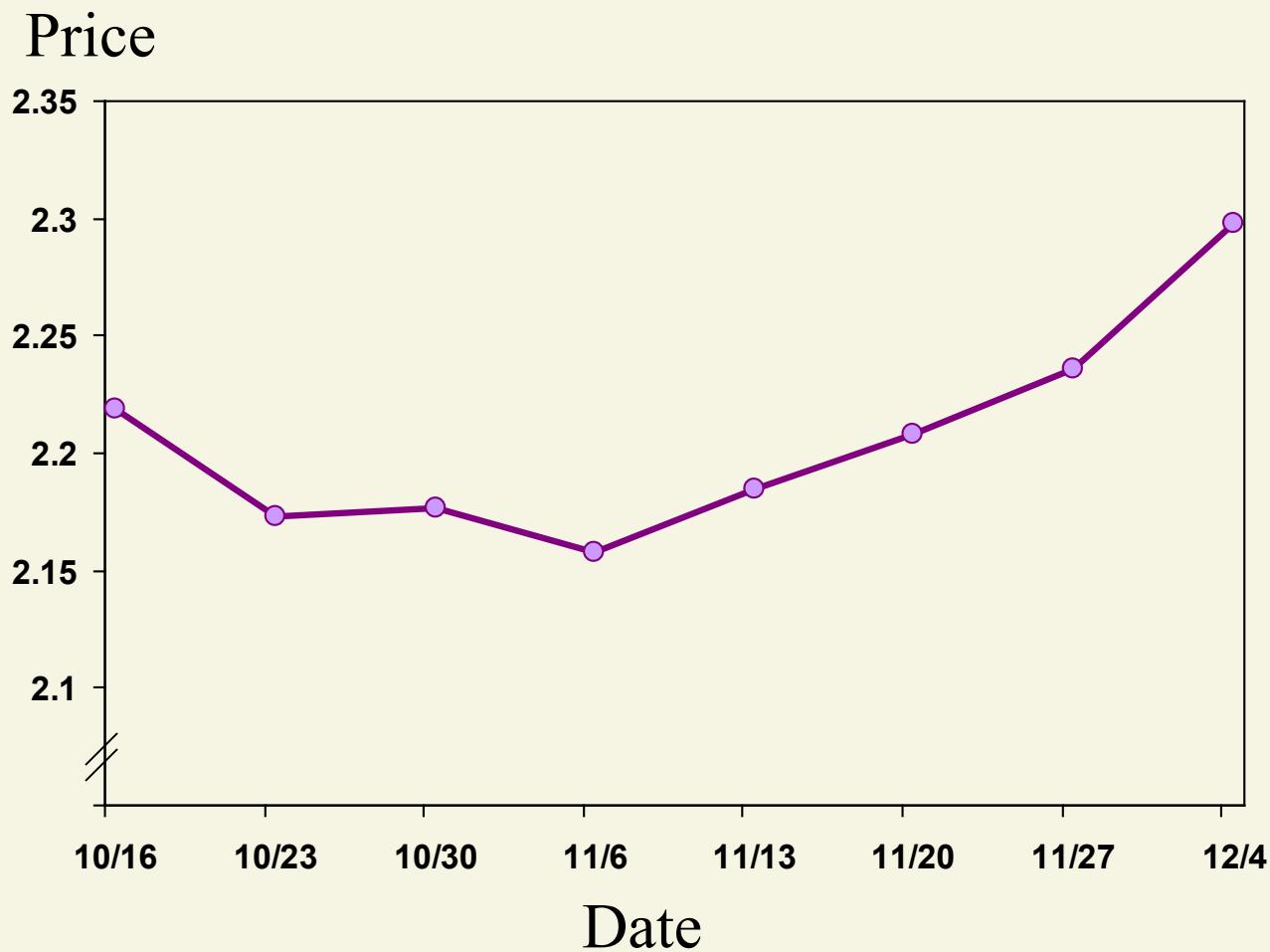
# Time Series Plot Example

- The following data shows the average retail price of regular gasoline in New York City for 8 weeks in 2006.
- Draw a **time series plot** for this data.



Date	Average Price
Oct 16, 2006	\$2.219
Oct 23, 2006	\$2.173
Oct 30, 2006	\$2.177
Nov 6, 2006	\$2.158
Nov 13, 2006	\$2.185
Nov 20, 2006	\$2.208
Nov 27, 2006	\$2.236
Dec 4, 2006	\$2.298

# Time Series Plot Example



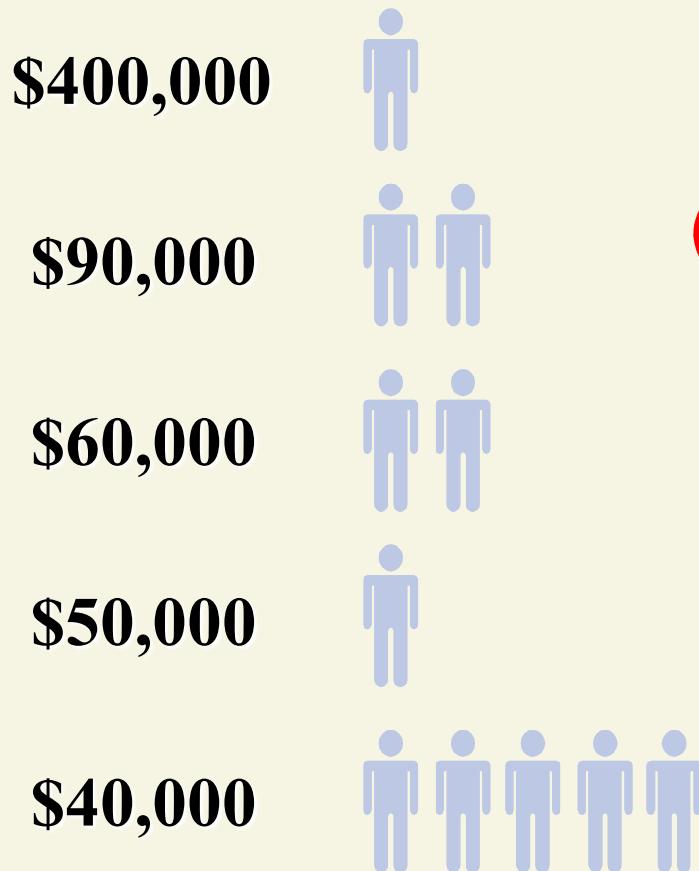
# 2.10

## Distorting the Truth with Descriptive Statistics

# Your Average Salary

- You're inquiring about the salary you could expect to earn if joining a company.
- You receive two answers:
  - The president tells you that the average salary is \$90,000 (mean).
  - One of the workers tells you that an “average employee” earns \$40,000 (median)
- Which answer can you believe?

# Thinking Challenge



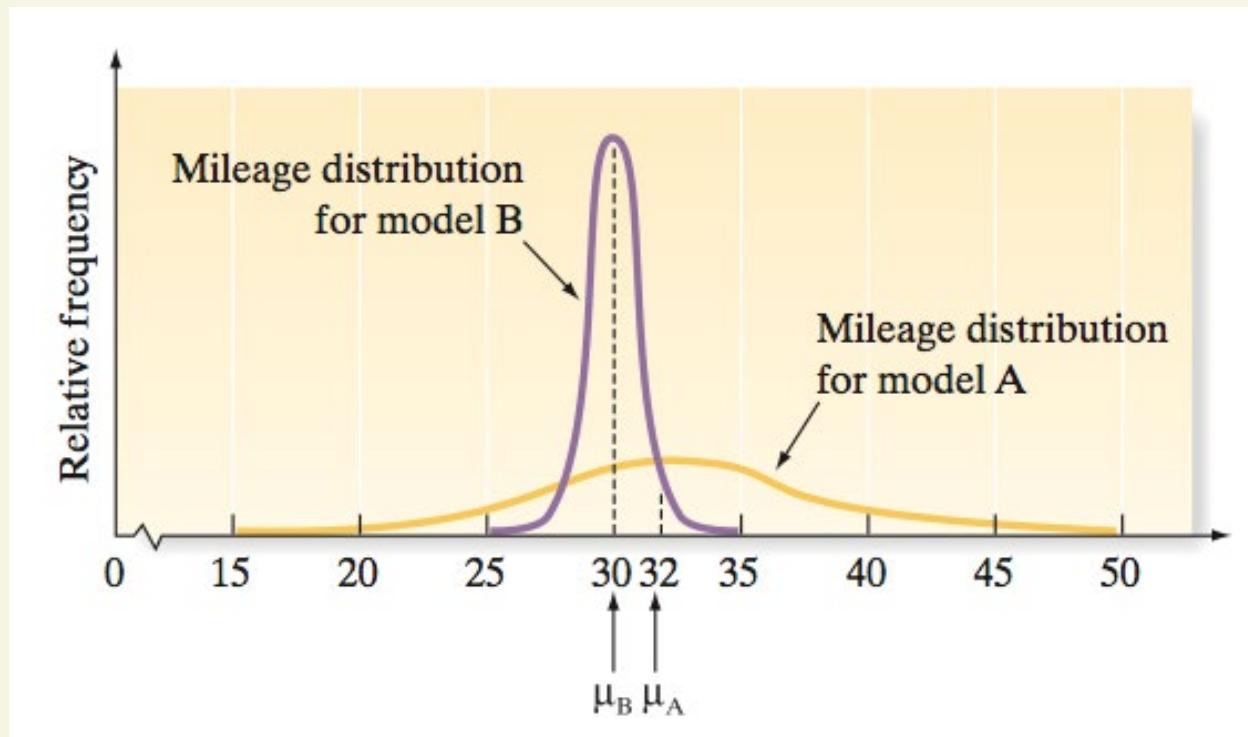
**Strike!**

... employees cite low pay --  
most workers earn only  
\$40,000.

... President claims average  
pay is \$90,000!

# Knowing only central tendency

- A: 32 MPG
- B: 30 MPG



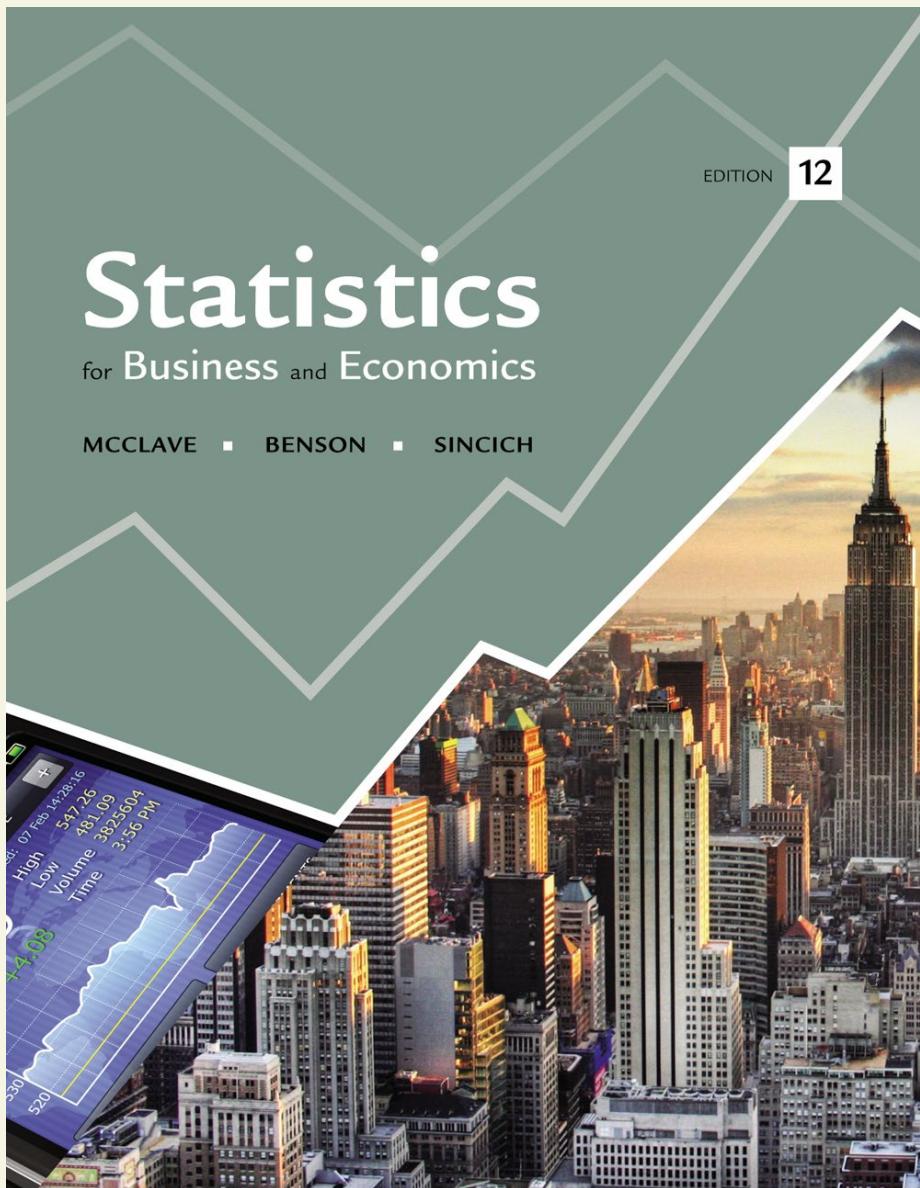
EDITION

12

# Statistics

for Business and Economics

MCCLAVE ■ BENSON ■ SINCICH



# **Statistics for Business and Economics**

## **Chapter 4** **Normal Distribution**

# 4.6

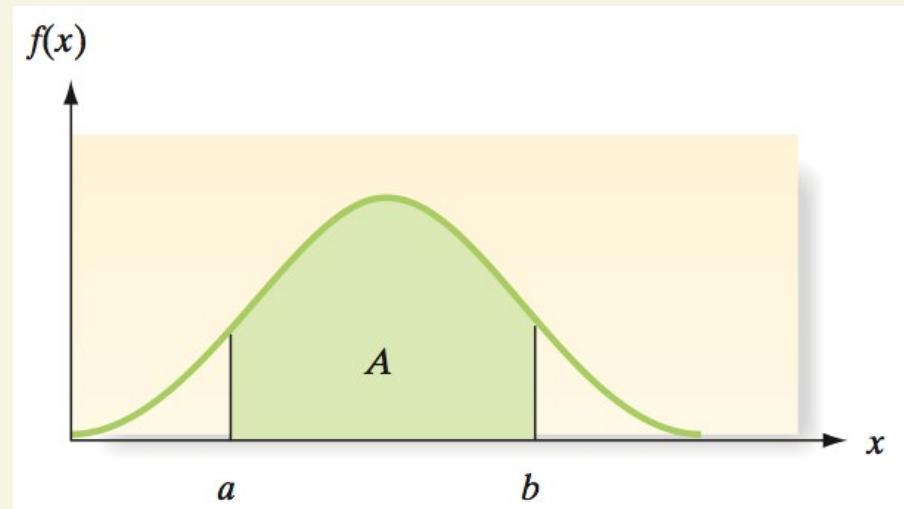
## The Normal Distribution

# Probability Density Function

**Probability Density Function (pdf)** is the graphical form of the probability distribution for a variable.

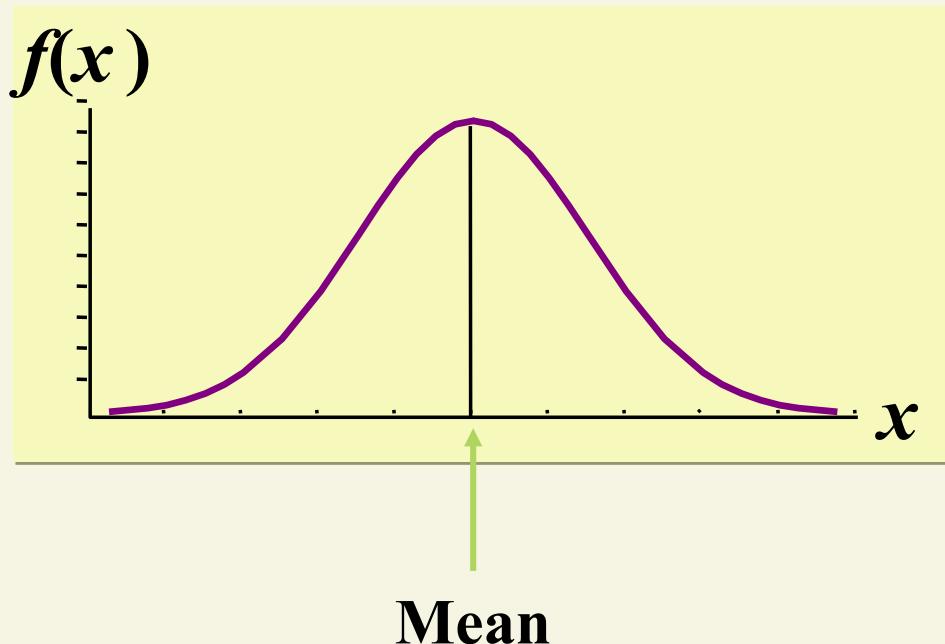
The areas under the pdf curve correspond to probabilities.

- $P(a < x < b) = \text{area } A$
- $P(x = c) = 0$  for any value of  $c$



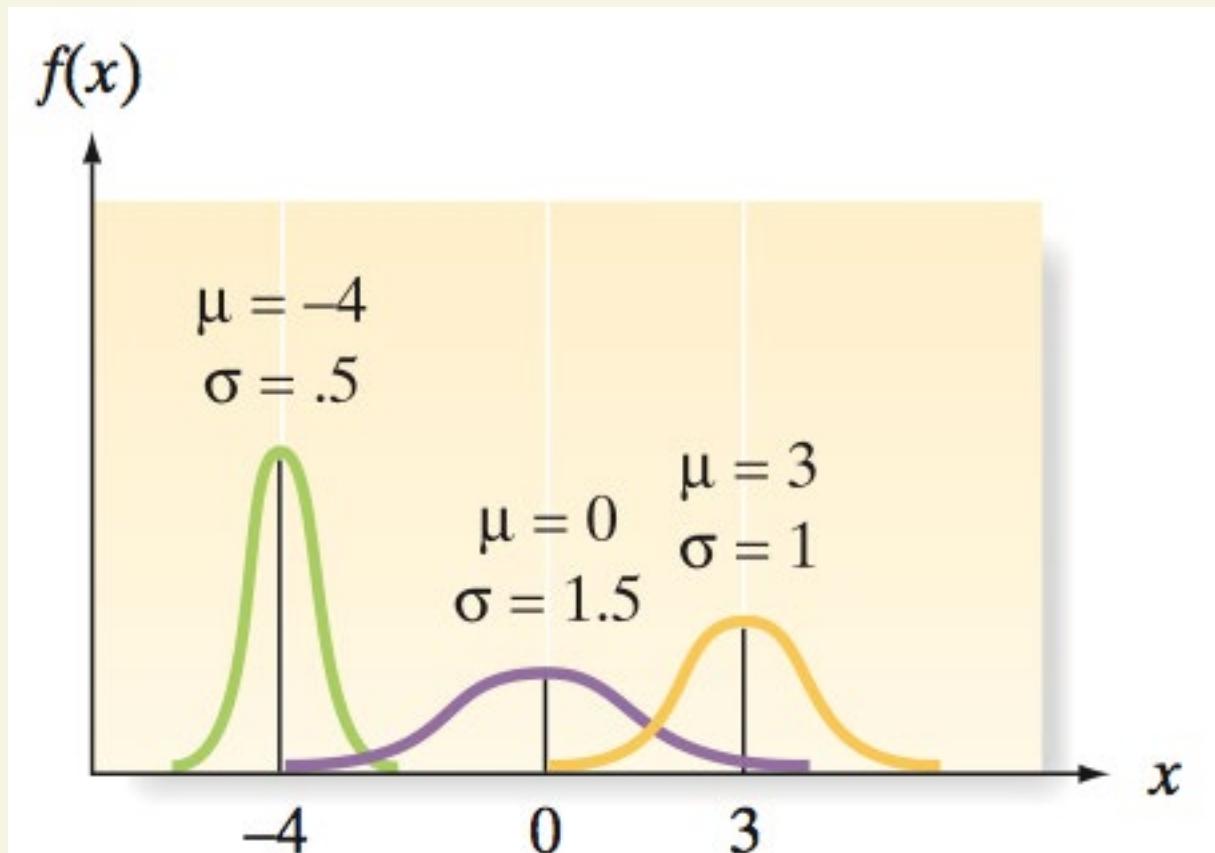
# Normal Distribution

1. 'Bell-shaped'
2. symmetrical around the mean



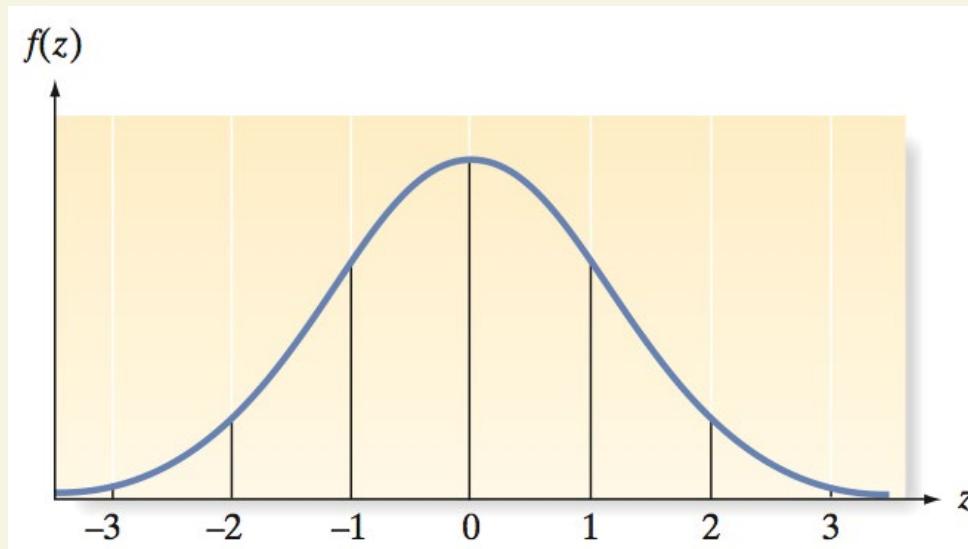
# Effect of Varying Parameters ( $\mu$ & $\sigma$ )

1. The mean ( $\mu$ ) determines the position.
2. The standard deviation ( $\sigma$ ) determines the shape.



# Standard Normal Distribution

The **standard normal distribution** is a normal distribution with  $\mu = 0$  and  $\sigma = 1$ . A random variable with a standard normal distribution, denoted by the symbol  $z$ , is called a standard normal random variable.

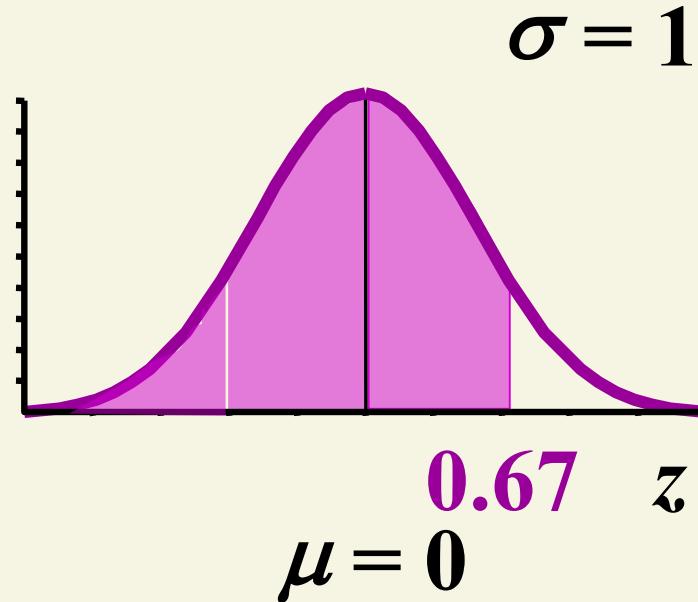


# The Standard Normal Probability

$$P(z \leq 0.67)$$

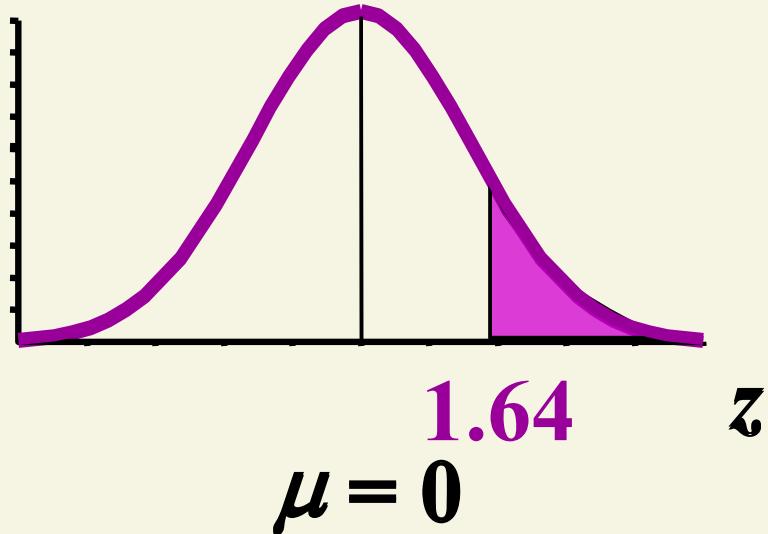
$$P(z \leq 0.67)$$

$$= .7486$$



Excel: =NORM.S.DIST(0.67, TRUE)

# The Standard Normal Probability



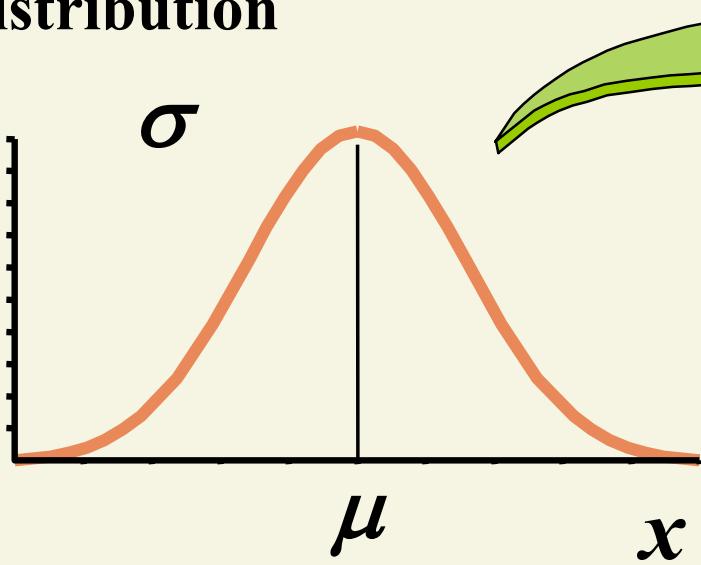
$$\sigma = 1$$

$$P(z > 1.64) \\ = .0505$$

Excel: =1-NORM.S.DIST(1.64, TRUE)

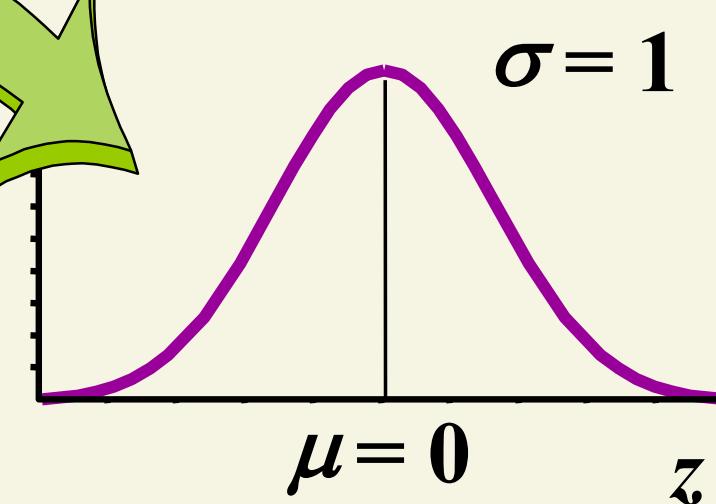
# Standardize the Normal Distribution

Normal  
Distribution



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

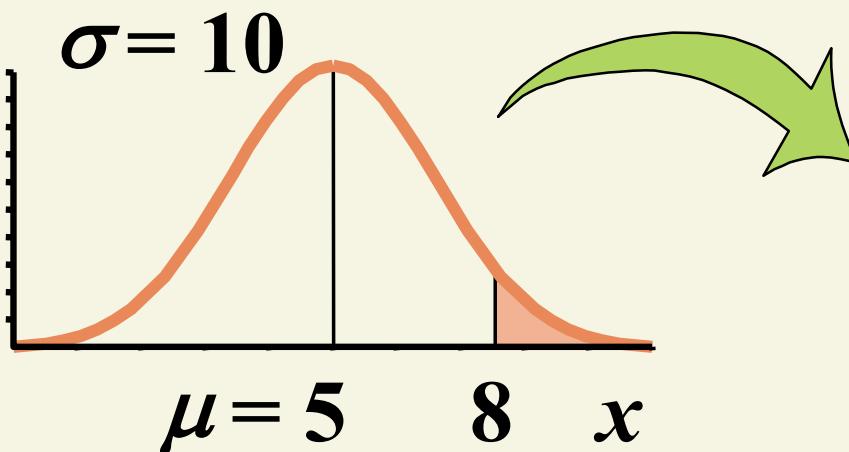
Standardized Normal  
Distribution



# Non-standard Normal $\mu = 5$ , $\sigma = 10$ :

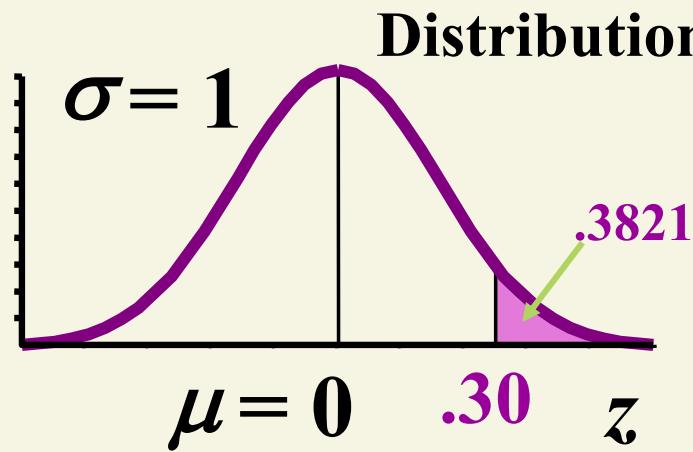
$$P(x \geq 8)$$

Normal Distribution



$$z = \frac{x - \mu}{\sigma} = \frac{8 - 5}{10} = .30$$

Standardized Normal Distribution



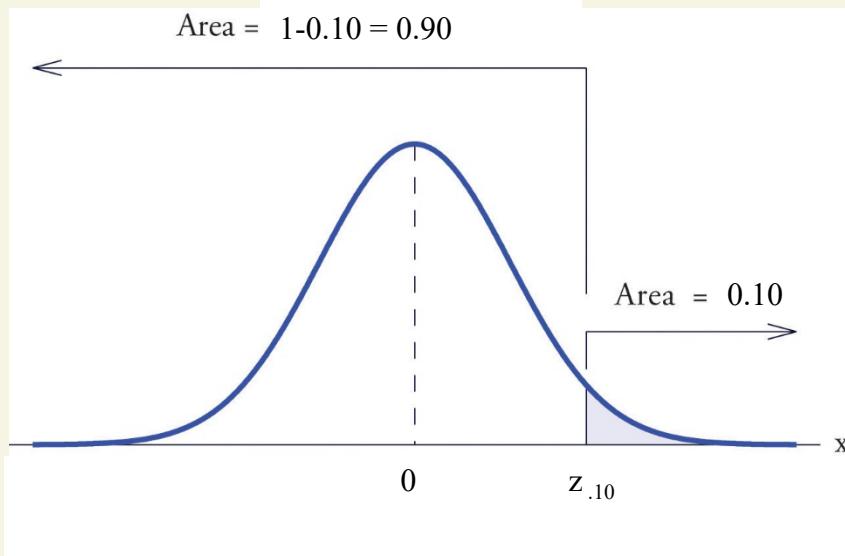
Excel: =1- NORM.DIST(8, 5, 10, TRUE)=0.3821

Or

=1 - NORM.S.DIST(.30, TRUE)

# Finding z-Values for Known Probabilities

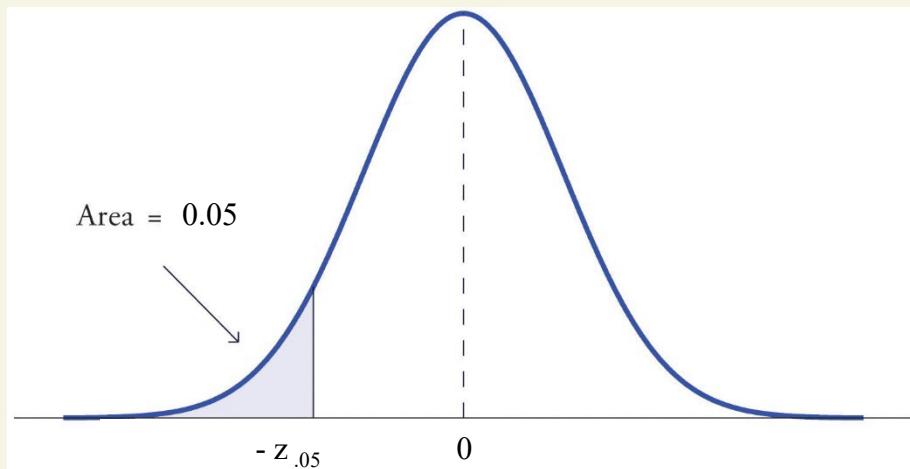
Find the value of  $z$ , call it  $z_\alpha$ , in the standard normal distribution that will be exceeded only 10% of the time—that is, find  $z_\alpha$  such that  $P(z > z_\alpha) = .10$ . In this case,  $z_\alpha$  is also denoted  $z_{.10}$ .



Solution:  $z_{0.10}=1.285$   
 $= -\text{NORM.S.INV}(.10)$

# Finding z-Values for Known Probabilities

Find  $-z_{.05}$  such that  $P(z < -z_{.05}) = .05$ .



Solution:  $-z_{.05} = -1.96$   
 $= \text{NORM.S.INV}(.05)$

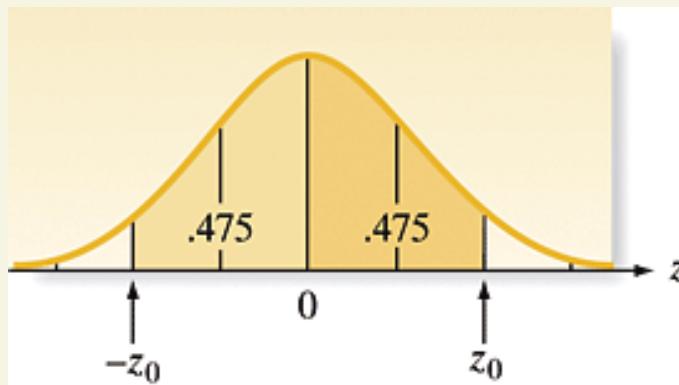
# Finding z-Values for Known Probabilities

Find the value of  $z_\alpha$  such that 95% of the standard normal z values lie between  $-z_\alpha$  and  $z_\alpha$ ; that is,  $P(-z_\alpha < z < z_\alpha) = .95$ .

The subscript  $\alpha$  always refer to the area to in the tail. Therefore in our case,  $\alpha=0.025$ .

Solution:  $z_{.025} = 1.96$ ,  $-z_{.025} = -1.96$

Excel:- = -NORM.S.INV(.025)



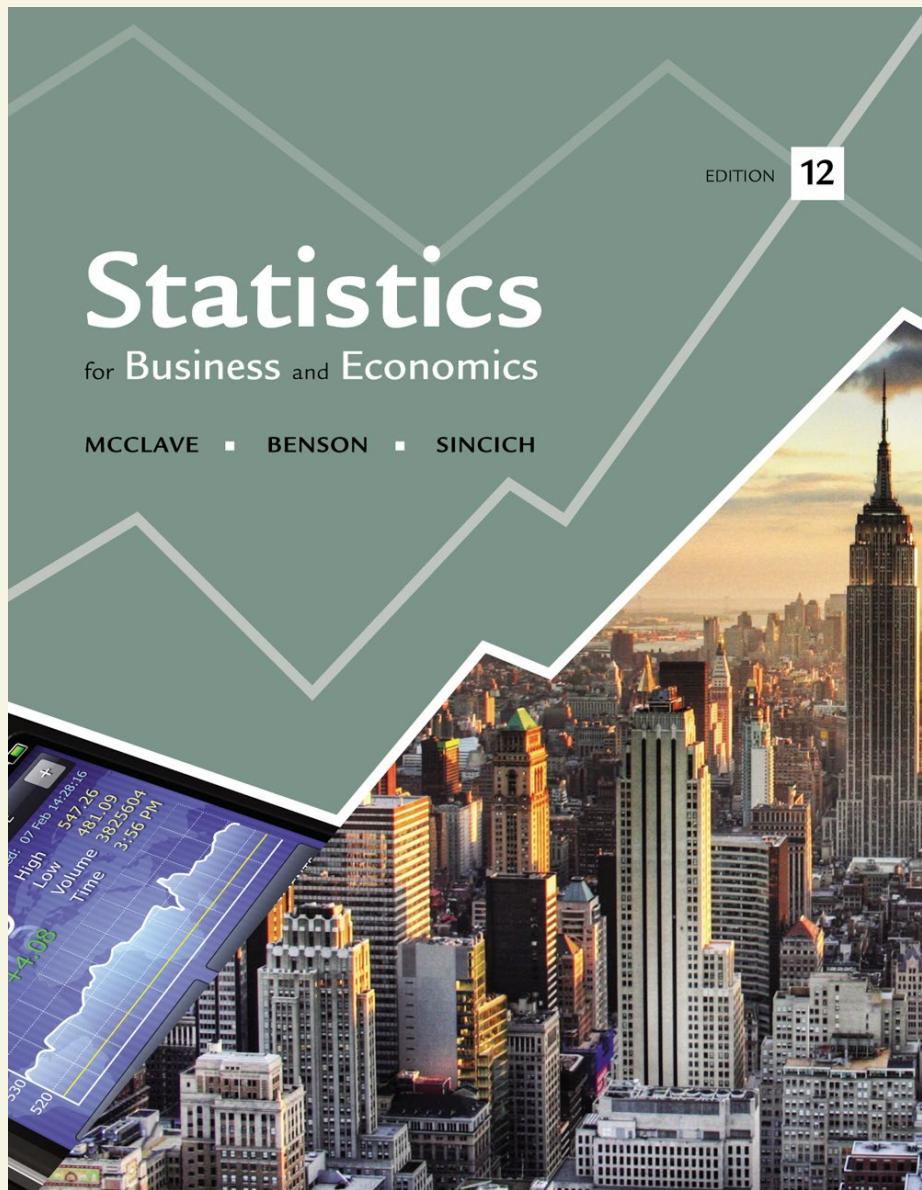
EDITION

12

# Statistics

for Business and Economics

MCCLAVE ■ BENSON ■ SINCICH



# **Statistics for Business and Economics**

## **Chapter 5** **Central Limit Theorem**

# Content

1. The Concept of a Sampling Distribution
2. The Sample Distribution of the Sample Mean and the Central Limit Theorem

# **5.1**

# **The Concept of a Sampling Distribution**

# Parameter & Statistic

A **parameter** is a numerical descriptive measure of a population. Because it is based on all the observations in the population, its value is almost always unknown.

A **sample statistic** is a numerical descriptive measure of a sample. It is calculated from the observations in the sample.

# Common Statistics & Parameters

	Sample Statistic	Population Parameter
Mean	$\bar{x}$	$\mu$
Standard Deviation	$s$	$\sigma$
Variance	$s^2$	$\sigma^2$
Binomial Proportion	$\hat{p}$	$p$

# **5.3**

# **The Central Limit Theorem (CLT)**

# Theorem

If we keep drawing samples of the same size from a population, how will the sample means be distributed?

- If samples are drawn from a normal population, the sampling distribution of  $\bar{x}$  will be a normal distribution.
- If samples are drawn from a non-normal population, the sampling distribution of  $\bar{x}$  will be a normal distribution if the sample size is larger than 30.

# Theorem

1. Mean of the sampling distribution (sample mean) equals of population mean, that is,

$$\mu_{\bar{x}} = E(\bar{x}) = \mu.$$

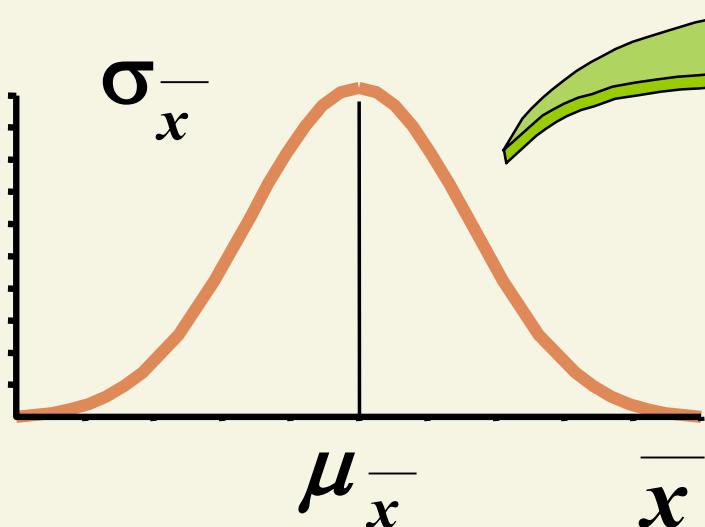
2. Standard deviation of the sampling distribution (sample mean) equals

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

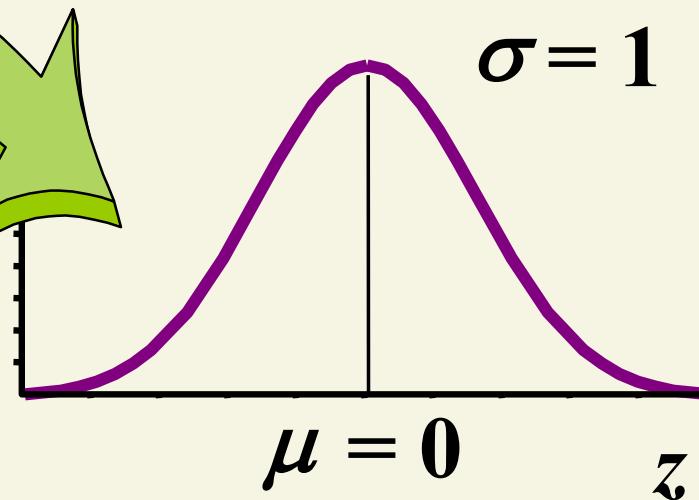
# Standardizing the Sampling Distribution of $\bar{x}$

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

Sampling Distribution



Standardized Normal Distribution



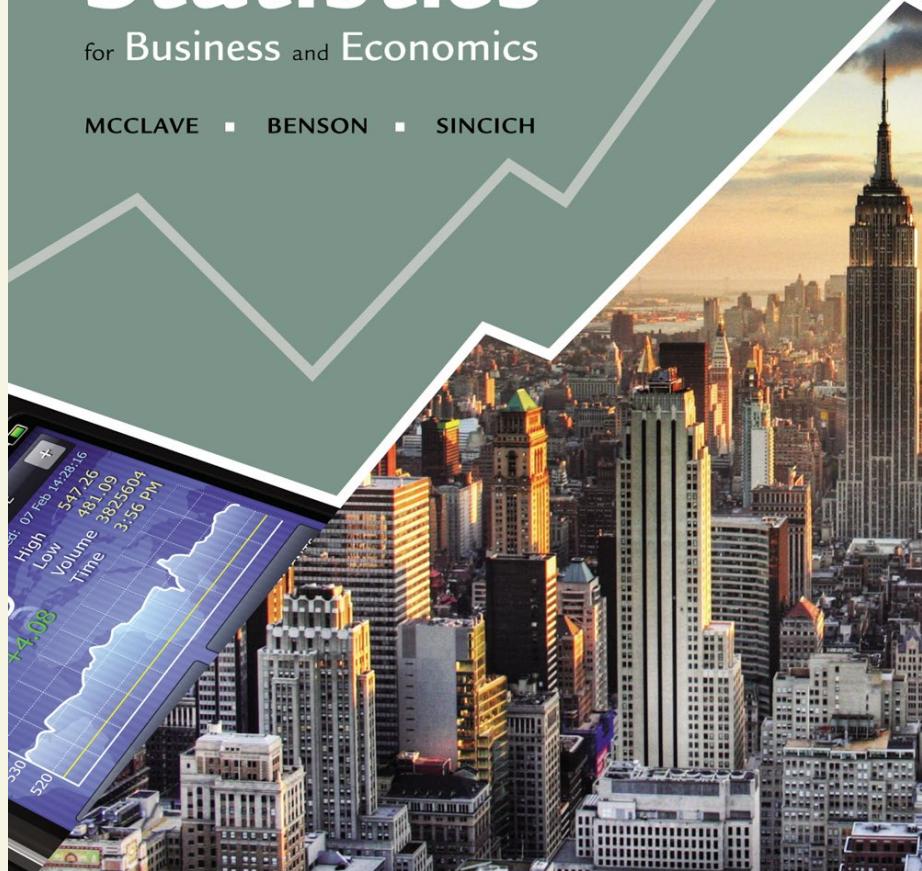
EDITION

12

# Statistics

for Business and Economics

MCCLAVE ■ BENSON ■ SINCICH



# **Statistics for Business and Economics**

## **Chapter 7**

### **Inferences Based on a Single Sample: Tests of Hypotheses**

# Content

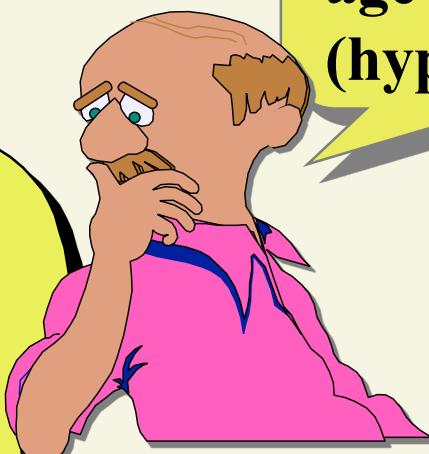
1. The Elements of a Test of Hypothesis
2. Formulating Hypotheses and Setting Up the Rejection Region
3. Observed Significance Levels:  $p$ -Values
4. Test of Hypothesis about a Population Mean: Normal ( $z$ ) Statistic
5. Test of Hypothesis about a Population Mean: Student's  $t$ -Statistic

# **7.1**

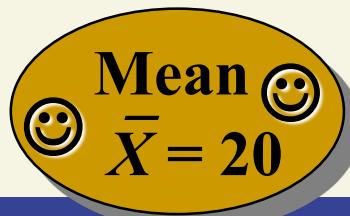
## **The Elements of a Test of Hypothesis**

# Hypothesis Testing

Population



Random sample



I believe the population mean age is 50 (hypothesis).

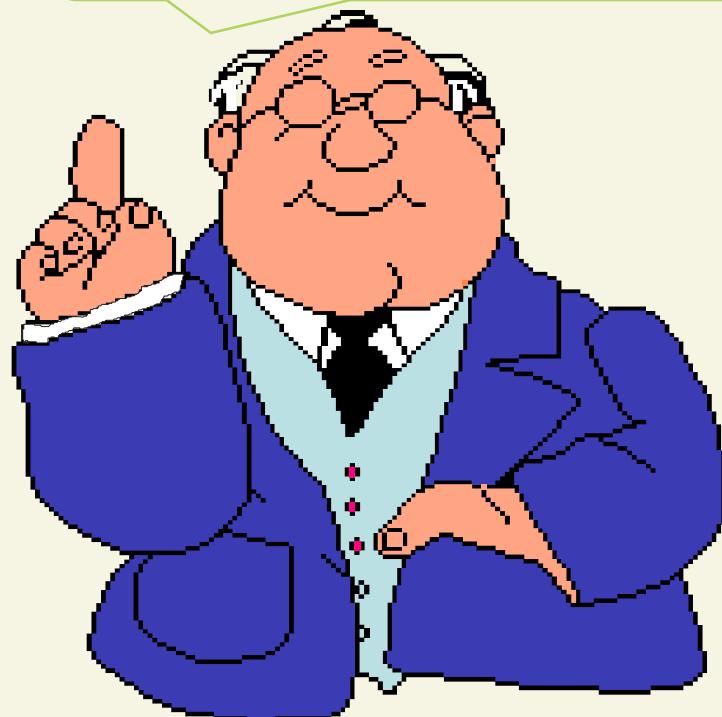
Reject hypothesis!  
Not close.



# What's a Hypothesis?

A statistical **hypothesis** is a statement about the numerical value of a population parameter.

I believe the mean GPA of this class is 3.5!



© 1984-1994 T/Maker Co.

# Null Hypothesis

## Null hypothesis,

- Denoted  $H_0$ ,
- Represents the hypothesis that will not be rejected (~~be accepted~~) unless the data provide convincing evidence that it is false.
- Represents the “status quo” or an assumption about the population parameter (such as  $\mu$ ) that the researcher wants to test (against) and hope to reject.

# Alternative Hypothesis

## Alternative (research) hypothesis:

- Denoted  $H_a$ ,
- Represents the hypothesis that will be proved to be true (~~accepted~~) only if the data provide convincing evidence in support of it.
- This usually represents the statement about a population parameter for which the researcher wants to gather evidence to support.

# Identifying Hypothesis

1. Null hypothesis:  $H_0: \mu = \text{(some value)}$
2. Alternative hypothesis, stated in one of the following forms

$H_a: \mu \neq \text{(some value)}$

$H_a: \mu < \text{(some value)}$

$H_a: \mu > \text{(some value)}$

# Test of Hypothesis

- Suppose building specifications in a certain city require that the average breaking strength of residential sewer pipe be more than 2,400 pounds per foot of length. Each manufacturer who wants to sell pipe in this city must demonstrate that its product meets the specification.
- We are interested in making an inference about the mean of a population. We are less interested in estimating the value of  $\mu$  than we are in testing a statement about its value
  - We want to decide whether the mean breaking strength of the pipe exceeds 2,400 pounds per linear foot.

# What Are the Hypotheses?

We want to know if the mean braking strength of the pipe exceeds 2,400 pounds per linear foot?

- State the alternative hypothesis:  $H_a: \mu > 2,400$
- State the opposite statistically:  $\mu \leq 2,400$
- State the null hypothesis:  $H_0: \mu = 2,400$

# Type I Error

A **Type I error** occurs if the researcher rejects the null hypothesis when, in fact,  $H_0$  is true. The probability of committing a Type I error is denoted by  $\alpha$ .

# Type II Error

A **Type II error** occurs if the researcher accepts the null hypothesis when, in fact,  $H_0$  is false. The probability of committing a Type II error is denoted by  $\beta$ .

# Conclusions and Consequences for a Test of Hypothesis

		True State of Nature
Conclusion	$H_0$ True	$H_a$ True
Fail to reject $H_0$ (Assume $H_0$ True)	Correct decision	Type II error (probability $\beta$ )
Reject $H_0$ (Assume $H_a$ True)	Type I error (probability $\alpha$ )	Correct decision

# Evidence in support of $H_a$

- How can the city decide when enough evidence exists to conclude that the manufacturer's pipe meets specifications?
- The city can select a sample of the manufacturer's pipe and measure the average breaking strength.
- The city can conclude that the pipe meets specifications only when the sample mean  $\bar{X}$  convincingly indicates that the population mean exceeds 2,400 pounds per linear foot.
- “Convincing” evidence in favor of the alternative hypothesis will exist when the value of  $\bar{X}$  exceeds 2,400 by an amount that cannot be readily attributed to sampling variability.

# Evidence in support of $H_a$

- Suppose the city tested 50 sections of sewer pipe and find the mean and standard deviation for these 50 measurements to be
  - $\bar{x} = 2460$  pounds per linear foot
  - $s = 200$  pounds per linear foot
- Is that enough evidence?
- Or does a sample mean of 2460 exceeds 2,400 by an amount that cannot be readily attributed to sample variability?

# Test Statistic

The **test statistic** is a sample statistic, computed using a sample, that we use to decide whether enough evidence is found in support of  $H_a$ .

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Assuming for the moment that the true mean is 2400. Then the sampling distribution of  $\bar{X}$  follows a normal distribution with  $\mu_0 = 2,400$ .

Test statistic:  $z = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{2460-2,400}{200/\sqrt{50}} = 2.12$

# ***p*-Value**

- The ***p*-value**, is the probability (assuming  $H_0$  is true) of observing a value of the test statistic that is at least as extreme as the one calculated from the sample.

# Example continued

- $H_0: \mu = 2,400$  vs  $H_a: \mu > 2,400$
- Test statistic

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{2460 - 2,400}{200/\sqrt{50}} = 2.12$$

- P-value =  $P(\bar{x} > 2460) = P(z > 2.12) = 0.017$
- The interpretation of the p-value is the following:
  - While assuming the population mean is 2400, the likelihood of observing a sample mean that is 2460 (or higher) is 0.017 (very unlikely).

# How to Decide Whether to Reject $H_0$

**Level of significant:** denoted by  $\alpha$ , the probability of type I error.

Choose the maximum value of  $\alpha$  that you are willing to tolerate.

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

For our example, we choose  $\alpha = 5\%$ .

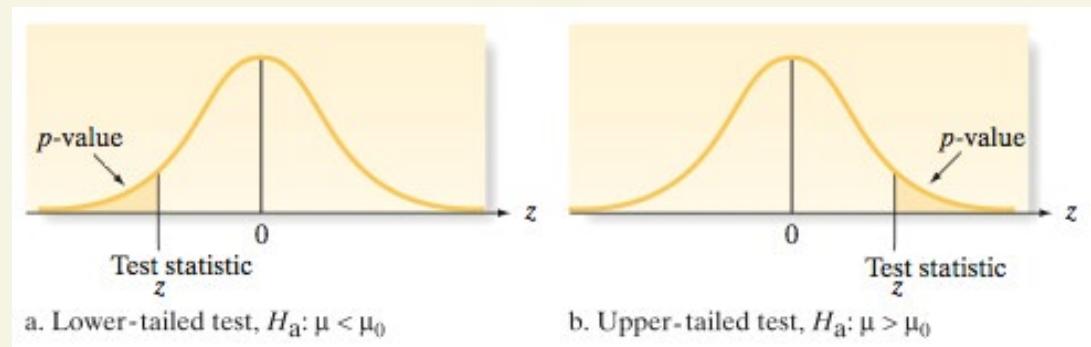
Since  $p\text{-value}=0.017$ , which is less than 5% (the significance level), therefore we can reject  $H_0$ .

# Conclusion

- Since we reject the null hypothesis, there is enough evidence to conclude that the average breaking strength of residential sewer pipe is more than 2,400 pounds per foot of length.

# Steps for Calculating the *p*-Value for a Test of Hypothesis

1. Determine the value of the test statistic  $z$ .
2. One-tail test: the sign in  $H_a$  is “ $<$ ” or “ $>$ ”
  - If the sign is “ $<$ ”, then  $p\text{-value} = P(Z < \text{test statistic})$
  - If the sign is “ $>$ ”, then  $p\text{-value} = P(z > \text{test statistic})$

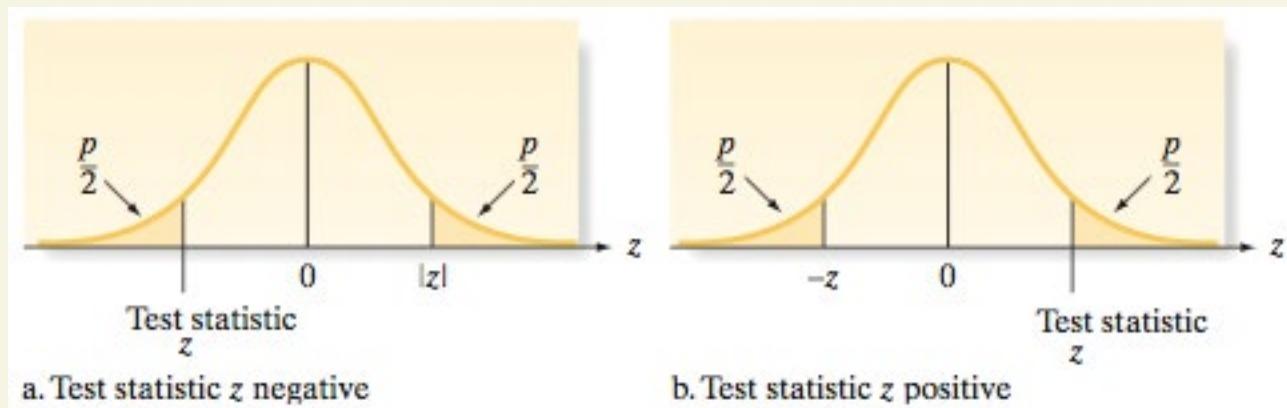


# Steps for Calculating the $p$ -Value for a Test of Hypothesis

3. Two-tail test: the sign in  $H_a$  is “ $\neq$ ”.

- P-value =  $P(z < \text{negative test statistic}) + P(z > \text{positive test statistic})$

→  $2P(z < \text{test stat})$  if test stat is negative  
 $2P(z < -\text{test stat})$  if test stat is positive

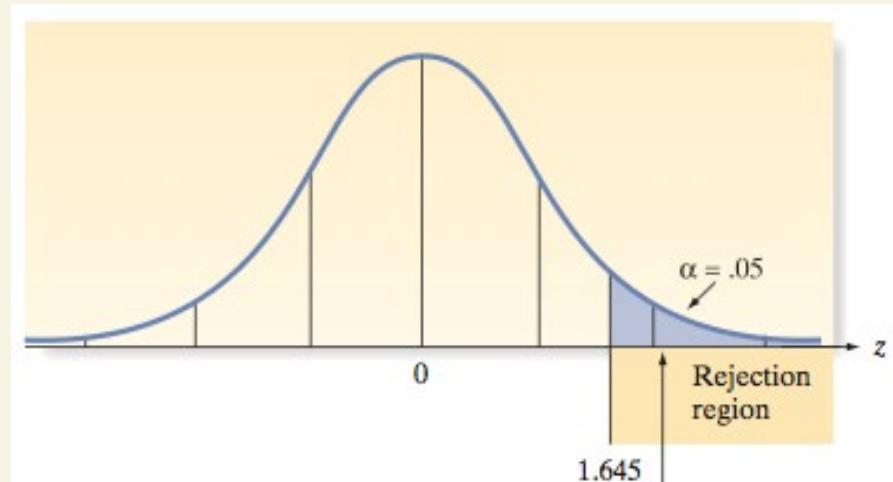


# Rejection Region Approach (Optional)

The **rejection region** of a statistical test is the set of possible values of the test statistic for which the researcher will reject  $H_0$  in favor of  $H_a$ .

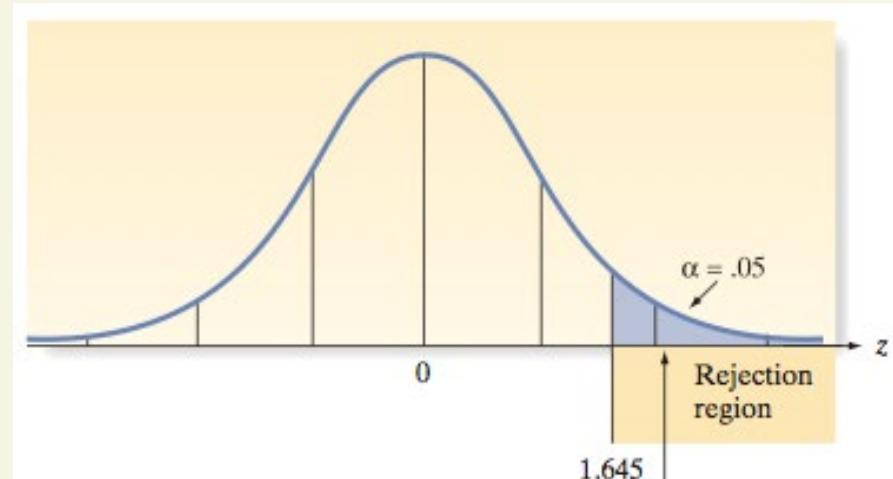
# Rejection Region

- Suppose we pick a 5% level of significance. Then the rejection region for our example is the right tail area (because of the sign in  $H_a$ ) such that the area under the normal curve is 0.05 (highlighted in blue)
- The rejection region is denoted as  $z > 1.645$ .
- Here 1.645 is called the **critical value**.



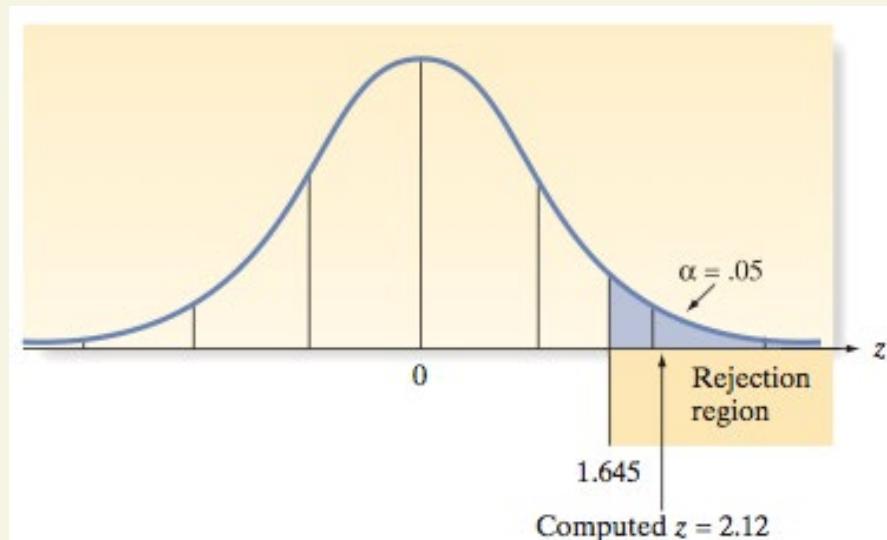
# Test Statistic vs Rejection Region

- For a test statistic that falls in the rejection region, we reject the null hypothesis and conclusion that enough evidence is found in support of  $H_a$ ;
- otherwise we say we fail to reject the null hypothesis and thus not enough evidence is found in support of  $H_a$ .



# Test Statistic vs Rejection Region

- For our example here, the test statistic is 2.12, which is higher than 1.645 and it falls in the rejection region.
- Thus we reject the null hypothesis and conclude that
  - enough evidence is found in support of the alternative hypothesis.
  - the average breaking strength of residential sewer pipe is more than 2,400 pounds per foot of length



# **7.4**

## **Test of Hypotheses about a Population Mean: Normal (z) Statistic**

# Example: two-tail test

A manufacturer of cereal wants to test the performance of one of its filling machines. The machine is designed to discharge a mean amount of 12 ounces per box, and the manufacturer wants to detect any departure from this setting. This quality study calls for randomly sampling 100 boxes from today's production run and determining whether the mean fill for the run is 12 ounces per box. The mean and standard deviation in the sample are 11.851 and 0.512, respectively.

- a. Set up a test of hypothesis for this study, using  $\alpha = .01$ .
- b. Calculate the p-value and interpret the results

# Solution

## Part a

- $H_0: \mu=12$
- $H_a: \mu\neq12$

## Part b

- Test statistic

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{11.851 - 12}{.512/\sqrt{100}} = -2.91$$

- For a two-tail test (denoted by “ $\neq$ ” in  $H_a$ ), the p-value is given by
- $P\text{-value}=2*P(z<-2.91) = 2(.0018)=.0036$

# Solution Part b

- P-value=0.0036
- The p-value tells us that if the machine were meeting specifications ( $\mu=12$ ), we would observe a sample mean as small as 11.85 or even smaller with a chance as small as 0.0036.
- Conclusion: since  $p\text{-value}<5\%$ , we reject the null hypothesis and conclude that the mean fill for the run is NOT 12 ounces per box.
- It is a strong indication that the machine is not filling the boxes correctly.

# Test of Hypothesis about $\mu$

## One-Tailed Test

$$H_0: \mu = \mu_0$$

$$H_a: \mu < \mu_0$$

(or  $H_a: \mu > \mu_0$ )

*Test Statistic:*

Large sample size

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

## Two-Tailed Test

$$H_0: \mu = \mu_0$$

$$H_a: \mu \neq \mu_0$$

*Test Statistic:*

small sample size

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

# Test of Hypothesis about $\mu$

## One-Tailed Test

P-value:  $P(z < \text{test stat})$  when  $H_a: \mu < \mu_0$

$P(z > \text{test stat})$  when  $H_a: \mu > \mu_0$ )

## Two-Tailed Test

*P-value:*

$2P(z < \text{test stat})$  if test stat is negative

$2P(z < -\text{test stat})$  if test stat is positive

# 7.5

## **Test of Hypothesis about a Population Mean: Student's *t*-Statistic**

# Example: *t* Test

- A major car manufacturer wants to test a new engine to determine whether it meets new air pollution standards. The mean emission of all engines of this type must be less than 20 parts per million of carbon. Ten engines are manufactured for testing purposes, and the emission level of each is determined. The data (in parts per million) are listed below.
- Do the data supply sufficient evidence to allow the manufacturer to conclude that this type of engine meets the pollution standard? Assume that the production process is stable and the manufacturer is willing to risk a Type I error with probability  $\alpha = .01$ .

**Table 7.5** Emission Levels for Ten Engines

15.6	16.2	22.5	20.5	16.4	19.4	19.6	17.9	12.7	14.9
------	------	------	------	------	------	------	------	------	------

# Solution

- $H_0: \mu = 20$
- $H_a: \mu < 20$
- $\alpha = 0.01$

**Test Statistic:**

$$\bar{x} = 17.57; s = 2.95$$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{17.57 - 20}{2.95/\sqrt{10}} = -2.6$$

$$\text{P-value} = P(t < -2.6) = 0.014$$

**Decision:**

**do not reject  $H_0$**

**Conclusion:** there is not enough evidence in support of the claim that the type of engine meets the standard.

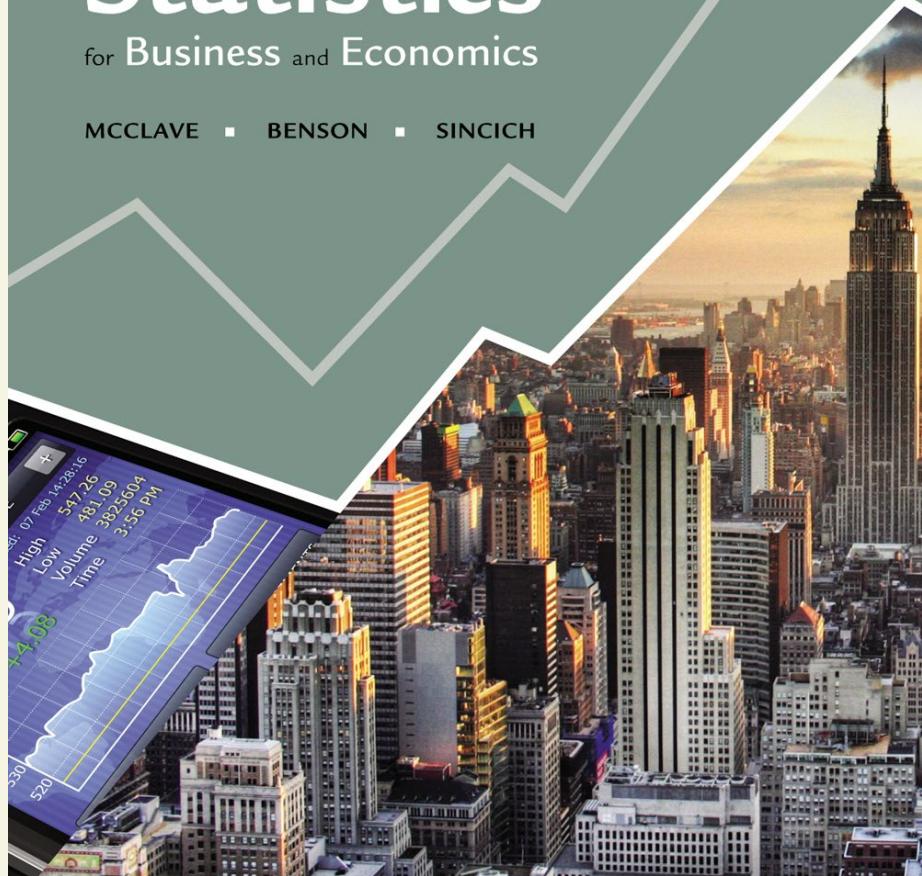
EDITION

12

# Statistics

for Business and Economics

MCCLAVE ■ BENSON ■ SINCICH



# **Statistics for Business and Economics**

## **Chapter 12 Multiple Regression and Model Building**

# Content

1. Multiple Regression Models
2. Least Squares Estimation
3. Evaluating Model Validity and Utility
4. Making Inferences about Individual Coefficients
5. Residual Analysis: Checking the Regression Assumptions

# Content

5. Using the Model for Estimation and Prediction
7. Qualitative (Dummy) Variable Models
8. Models with Both Quantitative and Qualitative Variables
9. 10. Some Pitfalls: Estimability, Multicollinearity, and Extrapolation

# **12.1**

## **Multiple Regression Models**

# The General Multiple Linear Regression Model

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

$y$ : the dependent variable

$x_1, x_2, \dots, x_k$ : the independent variables

$\beta_i$ : population coefficient

$\varepsilon$ : the error term

# Analyzing a Multiple Regression Model

- Step 1** Hypothesize the multiple linear regression model. This involves the choice of the independent variables to be included in the model.
- Step 2** Use the sample data to estimate the unknown model parameters  $\beta_0, \beta_1, \beta_2, \dots, \beta_k$  in the model.

# Analyzing a Multiple Regression Model

- Step 3** Check that the assumptions on  $\varepsilon$  are satisfied and make model modifications if necessary.
- Step 4** Statistically evaluate the usefulness of the model.
- Step 5** When satisfied that the model is useful, use it for prediction, estimation, and other purposes.

# Assumptions for Random Error $\varepsilon$

For any given set of values of  $x_1, x_2, \dots, x_k$ , the random error  $\varepsilon$  has a probability distribution with the following properties:

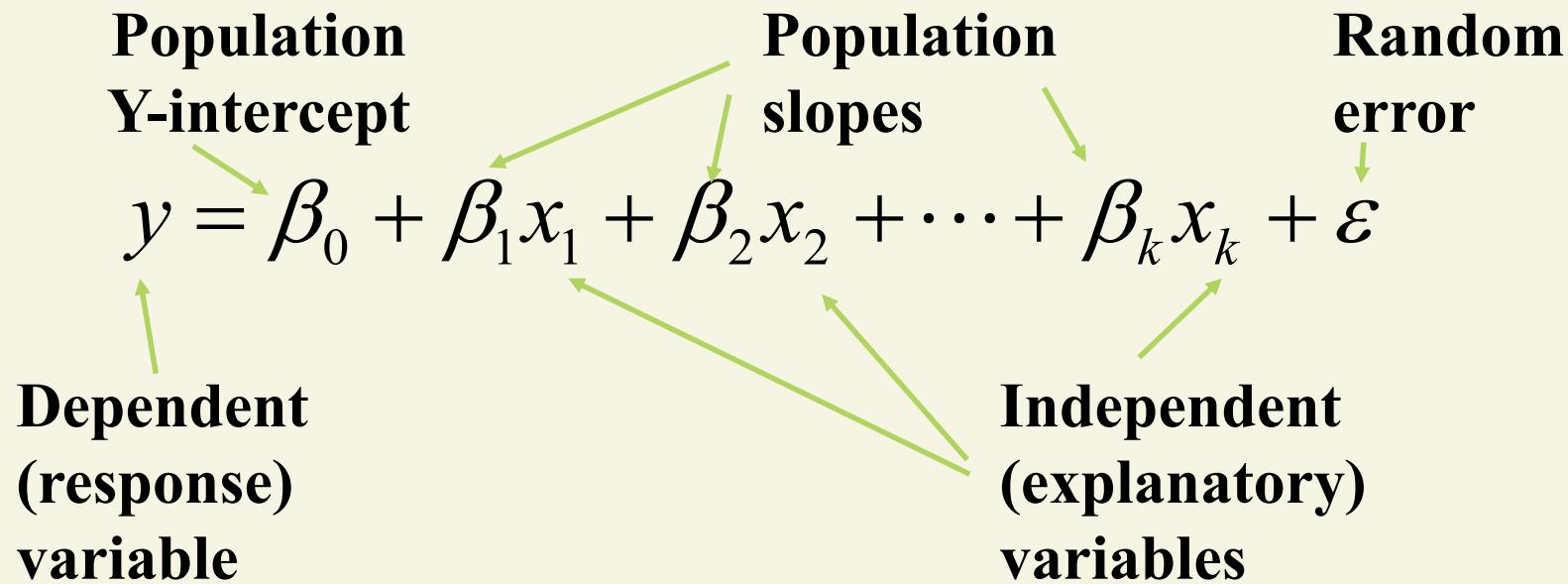
1. Mean equal to 0
2. Constant Variance
3. Normal distribution
4. Random errors are independent

# **12.2**

## **Least Squares Estimation**

# First-Order Multiple Regression Model

Relationship between 1 dependent and 2 or more independent variables is a linear function



# Example

A collector of antique grandfather clocks sold at auction believes that the price received for the clocks depends on both the age of the clocks ( $x_1$ ) and the number of bidders ( $x_2$ ) at the auction. Data file: CLOCKS

- a. Use scatterplots to plot the sample data. Interpret the plots.
- b. Use the method of least squares to estimate the model.

# Example

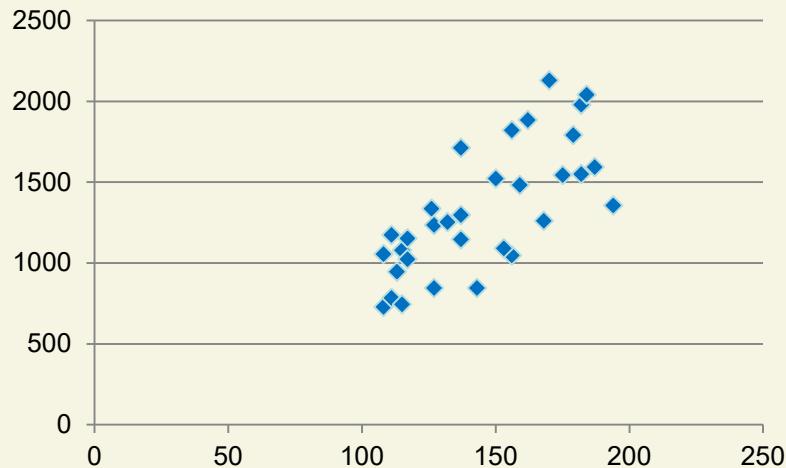
Table 12.1

Auction Price Data

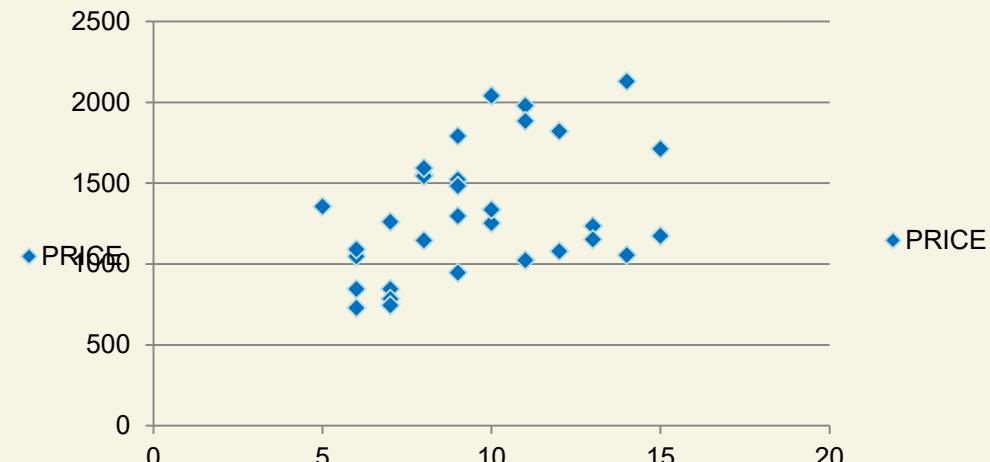
Age, $x_1$	Number of Bidders, $x_2$	Auction Price, $y$	Age, $x_1$	Number of Bidders, $x_2$	Auction Price, $y$
127	13	\$1,235	170	14	\$2,131
115	12	1,080	182	8	1,550
127	7	845	162	11	1,884
150	9	1,522	184	10	2,041
156	6	1,047	143	6	845
182	11	1,979	159	9	1,483
156	12	1,822	108	14	1,055
132	10	1,253	175	8	1,545
137	9	1,297	108	6	729
113	9	946	179	9	1,792
137	15	1,713	111	15	1,175
117	11	1,024	187	8	1,593
137	8	1,147	111	7	785
153	6	1,092	115	7	744
117	13	1,152	194	5	1,356
126	10	1,336	168	7	1,262

# Solution Part a

**PRICE vs Age**



**PRICE vs Bidders**



Both variables, age ( $x_1$ ) and number of bidders ( $x_2$ ) appears to be positively related to auction price, though the relationship for age is stronger.

# Excel Output

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.94464							
R Square	0.892344							
Adjusted R Square	0.884919							
Standard Error	133.4847							
Observations	32							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	2	4283063	2141531	120.1882	9.22E-15			
Residual	29	516726.5	17818.16					
Total	31	4799790						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1338.95	173.8095	-7.70356	1.71E-08	-1694.43	-983.471	-1694.43	-983.471
AGE	12.74057	0.90474	14.08202	1.69E-14	10.89017	14.59098	10.89017	14.59098
NUMBIDS	85.95298	8.728523	9.847368	9.34E-11	68.10115	103.8048	68.10115	103.8048

# Solution Part b

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1338.95	173.8095	-7.70356	1.71E-08	-1694.43	-983.471	-1694.43	-983.471
AGE	12.74057	0.90474	14.08202	1.69E-14	10.89017	14.59098	10.89017	14.59098
NUMBIDS	85.95298	8.728523	9.847368	9.34E-11	68.10115	103.8048	68.10115	103.8048

The least squares estimates of the  $\beta$  parameters (highlighted) are  $\hat{\beta}_0 = -1,339$ ,  $\hat{\beta}_1 = 12.74$ , and  $\hat{\beta}_2 = 85.95$ .

The estimated regression line is

$$\hat{y} = -1,339 + 12.74x_1 + 85.95x_2$$

# 12.3

## Evaluating Overall Model Utility

# Example

- Refer to the previous example, in which an antique collector modeled the auction price ( $y$ ) of grandfather clocks as a function of the age of the clock ( $x_1$ ) and the number of bidders ( $x_2$ ).
  - c. Conduct the global F-test of model validity at the  $\alpha = .05$  level of significance.
  - d. Evaluate the overall fitness of the model: find and interpret the adjusted coefficient of determination.

# Testing Global Usefulness of the Model: The analysis of Variance *F*-Test

## Testing Global Usefulness of the Model: The Analysis of Variance *F*-Test

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$  (All model terms are unimportant for predicting  $y$ .)

$H_a:$  At least one  $\beta_i \neq 0$  (At least one model term is useful for predicting  $y$ .)

$$\begin{aligned} \text{Test statistic: } F &= \frac{(\text{SS}_{yy} - \text{SSE})/k}{\text{SSE}/[n - (k + 1)]} = \frac{R^2/k}{(1 - R^2)/[n - (k + 1)]} \\ &= \frac{\text{Mean Square (Model)}}{\text{Mean Square (Error)}} \end{aligned}$$

where  $n$  is the sample size and  $k$  is the number of terms in the model.

*Rejection region:*  $F > F_\alpha$ , with  $k$  numerator degrees of freedom and  $[n - (k + 1)]$  denominator degrees of freedom.

# Solution Part C

ANOVA		df	SS	MS	F	Significance F
Regression	2	4283063	2141531	120.1882	9.22E-15	
Residual	29	516726.5	17818.16			
Total	31	4799790				

- $H_0: \beta_1 = \beta_2 = 0$
- $H_a:$  At least one  $\beta$  is not zero
- $\alpha = .05$
- P-value= 9.22E-15  $\approx 0$
- Decision: Reject  $H_0$
- Conclusion: 1. The model is valid; 2. At least one  $\beta$  is not zero; 3. At least one x is significantly linearly related to price.

# The Coefficient of Determination, $R^2$

$$R^2 = 1 - \frac{\text{SSE}}{\text{SS}_{yy}} = \frac{\text{SS}_{yy} - \text{SSE}}{\text{SS}_{yy}} = \frac{\text{Explained Variability}}{\text{Total Variability}}$$

- Proportion of variation in  $y$  ‘explained’ by all  $x$  variables **taken together**
- Never decreases when new  $x$  variable is added to model

# The Adjusted Coefficient of Determination

$$R_a^2 = 1 - \left[ \frac{n-1}{n-(k+1)} \right] \left( \frac{\text{SSE}}{\text{SS}_{yy}} \right)$$
$$= 1 - \left[ \frac{n-1}{n-(k+1)} \right] (1 - R^2)$$

Note:  $R_a^2 \leq R^2$

- Takes into account sample size and number of parameters
- Similar interpretation to  $R^2$

# Solution Part d

Regression Statistics	
Multiple R	0.94464
R Square	0.892344
Adjusted R Square	0.884919
Standard Error	133.4847
Observations	32

- The adjusted  $R^2$  (highlighted) is .885.
- This implies that the least squares model has explained about 88.5% of the total sample variation in y values (auction prices), after adjusting for sample size and number of independent variables in the model.

# 12.4

## Making Inferences about Individual Coefficients

# Test of an Individual Parameter Coefficient in the Multiple Regression Model

## One-Tailed Test

$$H_0: \beta_i = 0$$

$$H_a: \beta_i < 0 \text{ [or } H_a: \beta_i > 0]$$

## Two-Tailed Test

$$H_0: \beta_i = 0$$

$$H_a: \beta_i \neq 0$$

*Test statistic:*  $t = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$

*Rejection region:*  $t < -t_\alpha$

[or  $t > t_\alpha$  when  $H_a: \beta_i > 0$ ]

where  $t_\alpha$  and  $t_{\alpha/2}$  are based on  $n - (k + 1)$  degrees of freedom and

$n$  = Number of observations

$k + 1$  = Number of  $\beta$  parameters in the model

*Rejection region:*  $|t| > t_{\alpha/2}$

# Example

Refer to Examples 12.1 and 12.2. The collector of antique grandfather clocks knows that the price ( $y$ ) received for the clocks increases linearly with the age ( $x_1$ ) of the clocks and the number of bidders ( $x_2$ ) increases.

- d. Test the hypothesis that the mean auction price of a clock increases as the age of the clocks increases when number of bidders is held constant. Use  $\alpha = .05$ .
- e. Test the hypothesis that the mean auction price of a clock increases as the number of bidders increases when age is held constant. Use  $\alpha = .05$ .
- f. Form 95% confidence intervals for  $\beta_1$  and  $\beta_2$ .

# Solution

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1338.95	173.8095	-7.70356	1.71E-08	-1694.43	-983.471	-1694.43	-983.471
AGE	12.74057	0.90474	14.08202	1.69E-14	10.89017	14.59098	10.89017	14.59098
NUMBIDS	85.95298	8.728523	9.847368	9.34E-11	68.10115	103.8048	68.10115	103.8048

d.  $H_0: \beta_1 = 0$

$H_a: \beta_1 > 0$

P-value= 1.69E-14≈0

Decision: reject  $H_0$

Conclusion: Age is significantly and positively related to bidding price, and thus is useful in predicting price.

e.  $H_0: \beta_2 = 0$

$H_a: \beta_2 > 0$

P-value=9.34E-11 ≈0

Decision: reject  $H_0$

Conclusion: Number of bids is significantly and positively related to bidding price, and thus is useful in predicting price.

# Solution

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1338.95	173.8095	-7.70356	1.71E-08	-1694.43	-983.471	-1694.43	-983.471
AGE	12.74057	0.90474	14.08202	1.69E-14	10.89017	14.59098	10.89017	14.59098
NUMBIDS	85.95298	8.728523	9.847368	9.34E-11	68.10115	103.8048	68.10115	103.8048

f. The 95% confidence intervals for  $\beta_1$  and  $\beta_2$  are (10.89, 14.59) and (68.10, 103.80), respectively.

# Interpretation of Estimated Coefficients

## 1. Slope ( $\hat{\beta}_k$ )

- Estimated  $y$  changes by  $\hat{\beta}_k$  for each 1 unit increase in  $x_k$  ***holding all other variables constant***

## 2. $y$ -Intercept ( $\hat{\beta}_0$ )

- Average value of  $y$  when  $x_k = 0$

# Interpretations of coefficients

- $\hat{\beta}_0 = -1,339$ : no meaningful interpretation in this example.
- $\hat{\beta}_1 = 12.74$ : the mean auction price  $E(y)$  of an antique clock to increase \$12.74 for every 1-year increase in age ( $x_1$ ) when the number of bidders ( $x_2$ ) is held fixed.
- $\hat{\beta}_2 = 85.95$ : We estimate the mean auction price  $E(y)$  of an antique clock to increase \$85.95 for every 1-bidder increase in the number of bidders ( $x_2$ ) when age ( $x_1$ ) is held fixed.

# 12.5

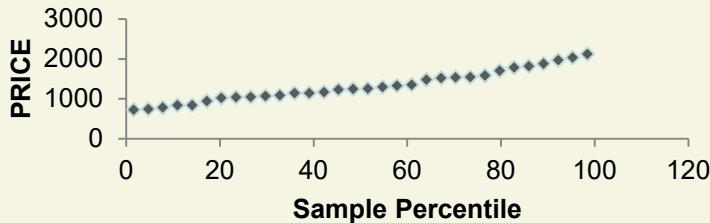
## Residual Analysis: Checking the Regression Assumptions

# Assumptions

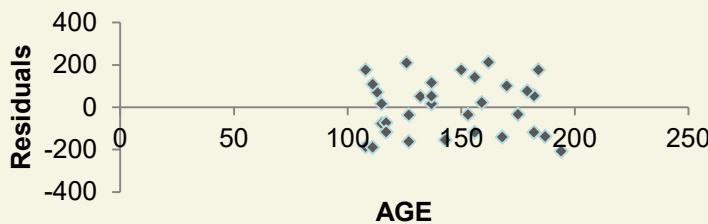
1. All assumptions on the error term discussed in the simple linear regression case hold here.
2. The independent variables should not be correlated with each other. If this assumption is violated, we say there is multicollinearity.

# Example

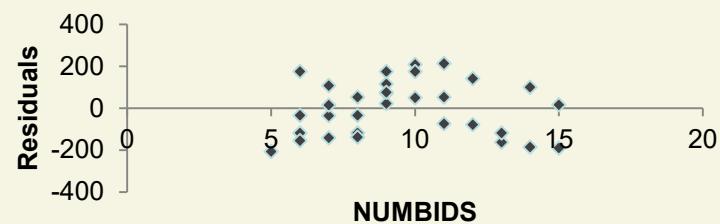
**Normal Probability Plot**



**AGE Residual Plot**



**NUMBIDS Residual Plot**



Conditions are satisfied: residuals are normally distributed; residuals show constant variance and independence.

# Multicollinearity

**Multicollinearity** exists when two or more of the independent variables used in regression are correlated.

## Consequences:

- Nonsignificant t–tests for all of the individual parameters when the F-test for overall model is significant.
- Sign opposite from what is expected in the estimated parameters.

# Using the Correlation Coefficient $r$ to Detect Multicollinearity

- Extreme multicollinearity:  $|r| \geq .8$
- Moderate multicollinearity:  $.2 \leq |r| < .8$
- Low multicollinearity:  $|r| < .2$

# Detecting Multicollinearity

1. Significant correlations between pairs of independent variables
2. Nonsignificant  $t$ -tests for all of the individual  $\beta$  parameters when the  $F$ -test for overall model adequacy is significant
3. Sign opposite from what is expected in the estimated  $\beta$  parameters

# Example

	AGE	NUMBIDS	PRICE
AGE	1		
NUMBIDS	-0.25375	1	
PRICE	0.729631	0.395204	1

- No multicollinearity detected.

# Example-Multicollinearity

- Data file: Multicollinearity
- Y: company revenue
- X: Household income, GDP, and price index

	Coefficients	Standard Error	t Stat	P-value
Intercept	15159.31	8403.206	1.803991	0.074073
Income	-6.5603	2.254371	-2.91003	0.004405
Price index	957.8601	840.7944	1.139232	0.257174
GDP	7.258794	1.116654	6.500485	2.68E-09

- Household income and revenue negatively and significantly related, which is counter intuitive.
- Cause: multicollinearity

	Revenue	Income	Price index	GDP
Revenue	1			
Income	0.932407	1		
Price index	-0.17019	-0.21388	1	
GDP	0.947986	0.993734	-0.21469	1

# Solutions to Some Problems Created by Multicollinearity in Regression

Drop one or more of the correlated independent variables from the model. One way to decide which variables to keep in the model is to employ stepwise regression.

# Example-Multicollinearity

- If we remove GDP:

	Coefficients	Standard Error	t Stat	P-value
Intercept	-36785.1	3185.018	-11.5494	1.82E-20
Income	8.01715	0.308332	26.00165	8.28E-48
Price index	851.3107	993.3756	0.856988	0.393385

- Income is now positively and significantly related to revenue.

# **12.6**

## **Using the Model for Estimation and Prediction**

# Example Estimation and Prediction

A collector of antique grandfather clocks sold at auction knows that the price  $y$  received for the clocks increases linearly with the age  $x_1$  of the clocks and the number of bidders  $x_2$ . The estimated regression equation is:

$$\hat{y} = -1,339 + 12.74x_1 + 85.95x_2$$

$y$  = auction price of grandfather clock

$x_1$  = age of clock

$x_2$  = number of bidders

Predict the auction price for a 150-year old clock sold at an auction with 10 bidders.

# Solution Part a

- $\hat{y} = -1,339 + 12.74x_1 + 85.95x_2$   
 $= -1,339 + 12.74 * 150 + 85.95 * 10 = 1431.7$

Predicted Values for New Observations

New

Obs	Fit	SE Fit	95% CI	95% PI
1	1431.7	24.6	(1381.4, 1481.9)	(1154.1, 1709.3)

Values of Predictors for New Observations

New

Obs	AGE	NUMBIDS
1	150	10.0

Confidence interval (CI): prediction interval for the average of multiple clocks.

Prediction interval (PI): prediction interval for a single clock.

# Example Estimation and Prediction

Suppose you want to predict the auction price for one clock that is 50 years old and has 2 bidders. How should you proceed?

Both values of  $x_1$  and  $x_2$  fall well *outside* their respective ranges in the data. Doing so may lead to an unreliable prediction.

# 12.7

## Qualitative (Dummy) Variable Models

# Dummy Variables

- Qualitative (categorical) independent variables can be included in regression using dummy variables.
- We code the categories of the qualitative variable as numbers.
- Example: suppose a female executive at a certain company claims that male executives earn higher salaries, on average, than female executives with the same education, experience, and responsibilities. To support her claim, she wants to model the salary  $y$  of an executive using a qualitative independent variable representing the gender of an executive (male or female).

# Qualitative Independent Variable with two Levels

The dummy variable used to describe gender could be coded as follows:

$$x = \begin{cases} 1 & \text{if male} \\ 0 & \text{otherwise (if female)} \end{cases}$$

The model takes the following form:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Males ( $x=1$ ):  $y_M = \beta_0 + \beta_1 x + \varepsilon = \beta_0 + \beta_1(1) + \varepsilon$

$$= \beta_0 + \beta_1 + \varepsilon$$

Females ( $x=0$ ):  $y_F = \beta_0 + \beta_1 x + \varepsilon$

$$= \beta_0 + \beta_1(0) + \varepsilon = \beta_0 + \varepsilon$$

# Qualitative Independent Variable with $k$ Levels

For a qualitative variable with  $k$  levels, a total of  $k-1$  dummy variables are needed to represent the first  $k-1$  levels.

$$x_i = \begin{cases} 1 & \text{if } y \text{ is observed at the } i\text{th level} \\ 0 & \text{otherwise} \end{cases}$$

The  $k$ th level (last category) is called the omitted or default category, or base level. It is defined when

$$x_1 = x_2 = \cdots = x_{k-1} = 0$$

# Interpreting Dummy-Variable Model Equation

Given:  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

$y$  = Starting salary of college graduates

$x_1$  = GPA

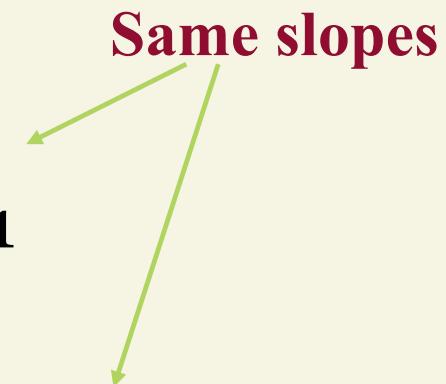
$x_2 = \begin{cases} 0 & \text{if Male} \\ 1 & \text{if Female} \end{cases}$

Male ( $x_2 = 0$ ):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2(0) = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

Female ( $x_2 = 1$ ):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2(1) = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 x_1$$



# Dummy-Variable Model Example

Computer Output:  $\hat{y} = 3 + 5x_1 + 7x_2$

$$x_2 = \begin{cases} 0 & \text{if Male} \\ 1 & \text{if Female} \end{cases}$$

Male ( $x_2 = 0$ ):

$$\hat{y} = 3 + 5x_1 + 7(0) = 3 + 5x_1$$

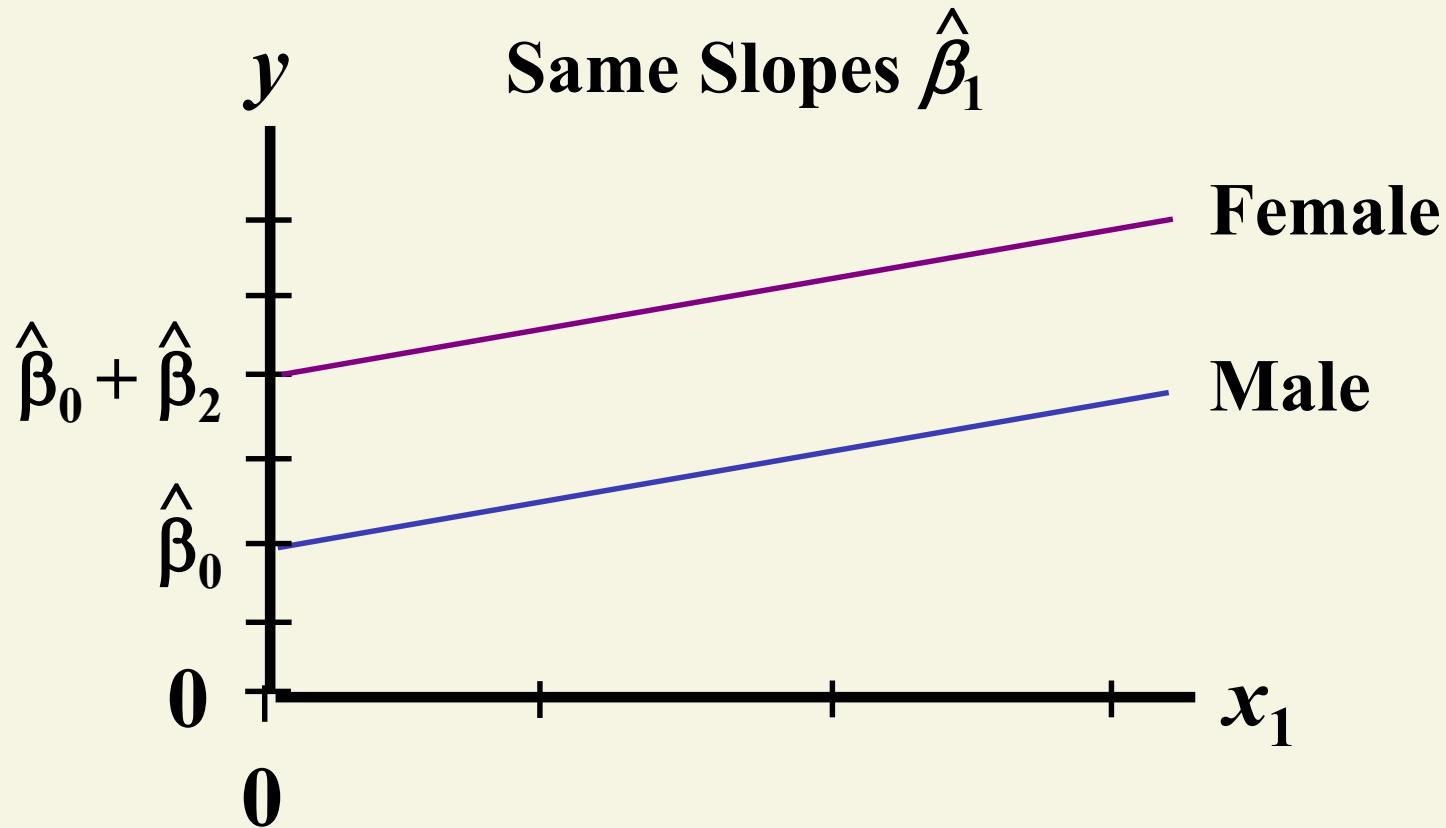
Female ( $x_2 = 1$ ):

$$\hat{y} = 3 + 5x_1 + 7(1) = (3 + 7) + 5x_1$$

Same slopes



# Dummy-Variable Model Relationships



# Example

- The value of a used car is related to its odometer reading. To examine this issue, a used-car dealer randomly selected 100 three-year-old Toyota Camrys that were sold at auction during the past month. Each car was in top condition and equipped with all the features that come standard with this car. The dealer recorded the price (\$1,000), the number miles (thousands) on the odometer and the color (white, silver, other) of the car. The dealer wants to find the regression line.
- Data file: Toyota

# Solution

Since the qualitative variable has three levels, we define the following independent variables.

$$x_1 = \text{mileage}$$

$$x_2 = \begin{cases} 1 & \text{if the car is white} \\ 0 & \text{otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if the car is silver} \\ 0 & \text{otherwise} \end{cases}$$

Note: other color is defined when  $x_2 = x_3 = 0$

The regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

# Data

A screenshot of a Microsoft Excel spreadsheet titled "Price". The table has three columns: "Price", "Odometer", and "Color". The "Price" column contains numerical values ranging from 14.1 to 15.7. The "Odometer" column contains numerical values ranging from 24.0 to 48.6. The "Color" column contains categorical values 1, 2, and 3. The first row is a header row.

	A	B	C	D	E	F	G	H	I
1	Price	Odometer	Color						
2	14.6	37.4	1						
3	14.1	44.8	1						
4	14.0	45.8	3						
5	15.6	30.9	3						
6	15.6	31.7	2						
7	14.7	34.0	2						
8	14.5	45.9	1						
9	15.7	19.1	3						
10	15.1	40.1	1						
11	14.8	40.2	1						
12	15.2	32.4	2						
13	14.7	43.5	1						
14	15.6	32.7	1						
15	15.6	34.5	2						
16	14.6	37.7	2						
17	14.6	41.4	1						
18	15.7	24.5	3						
19	15.0	35.8	1						
20	14.7	48.6	1						
21	15.4	24.0	1						

# Excel Output

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.837135							
R Square	0.700794							
Adjusted R Square	0.691444							
Standard Error	0.304258							
Observations	100							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	3	20.81492	6.938306	74.9498	4.65E-25			
Residual	96	8.886981	0.092573					
Total	99	29.7019						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Intercept	16.83725	0.197105	85.42255	2.28E-92	16.446	17.2285	16.446	17.2285
Odometer	-0.05912	0.005065	-11.6722	4.04E-20	-0.06918	-0.04907	-0.06918	-0.04907
x2	0.091131	0.072892	1.250224	0.214257	-0.05356	0.235819	-0.05356	0.235819
x3	0.330368	0.08165	4.046157	0.000105	0.168294	0.492442	0.168294	0.492442

# Interpretation

- The estimated regression model is
- $\hat{y} = 16.84 - .059x_1 + .09x_2 + .33x_3$
- For a white car,  $x_2 = 1$  and  $x_3 = 0$ . The regression line is

$$\begin{aligned}\hat{y} &= 16.84 - .059x_1 + .09(1) + .33(0) \\ &= 16.93 - .059x_1\end{aligned}$$

- For a silver car,  $x_2 = 0$  and  $x_3 = 1$ . The regression line is

$$\begin{aligned}\hat{y} &= 16.84 - .059x_1 + .09(0) + .33(1) \\ &= 17.167 - .059x_1\end{aligned}$$

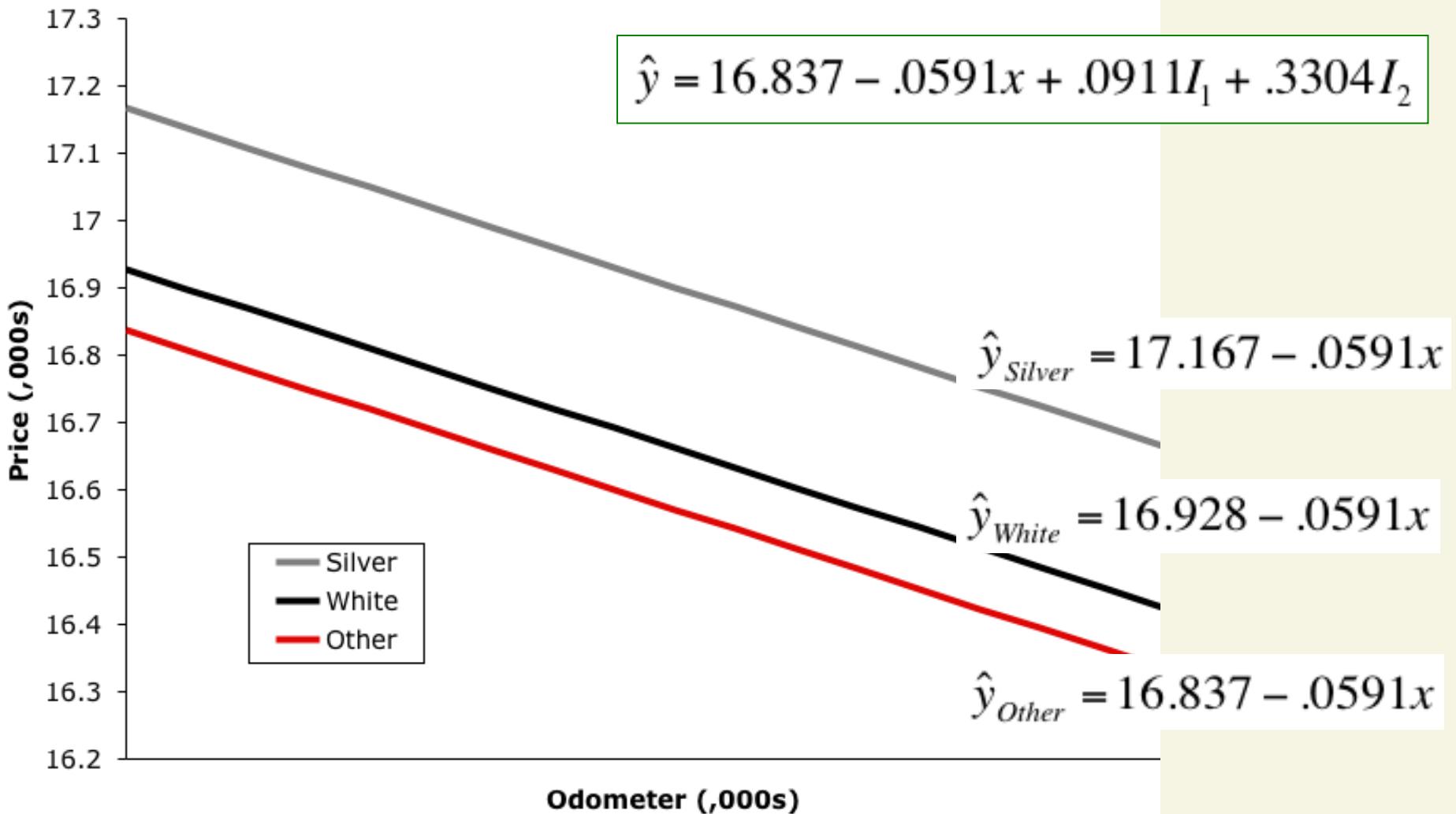
# Interpretation

- For a nonwhite nonsilver car,  $x_2 = 0$  and  $x_3 = 0$ .  
The regression line is

$$\begin{aligned}\hat{y} &= 16.84 - .059x_1 + .09(0) + .33(0) \\ &= 16.84 - .059x_1\end{aligned}$$

- $\hat{\beta}_3 = .33$ :
  - On average, a silver car sells for \$330.40 more than other colors with the same odometer reading

# Graphically



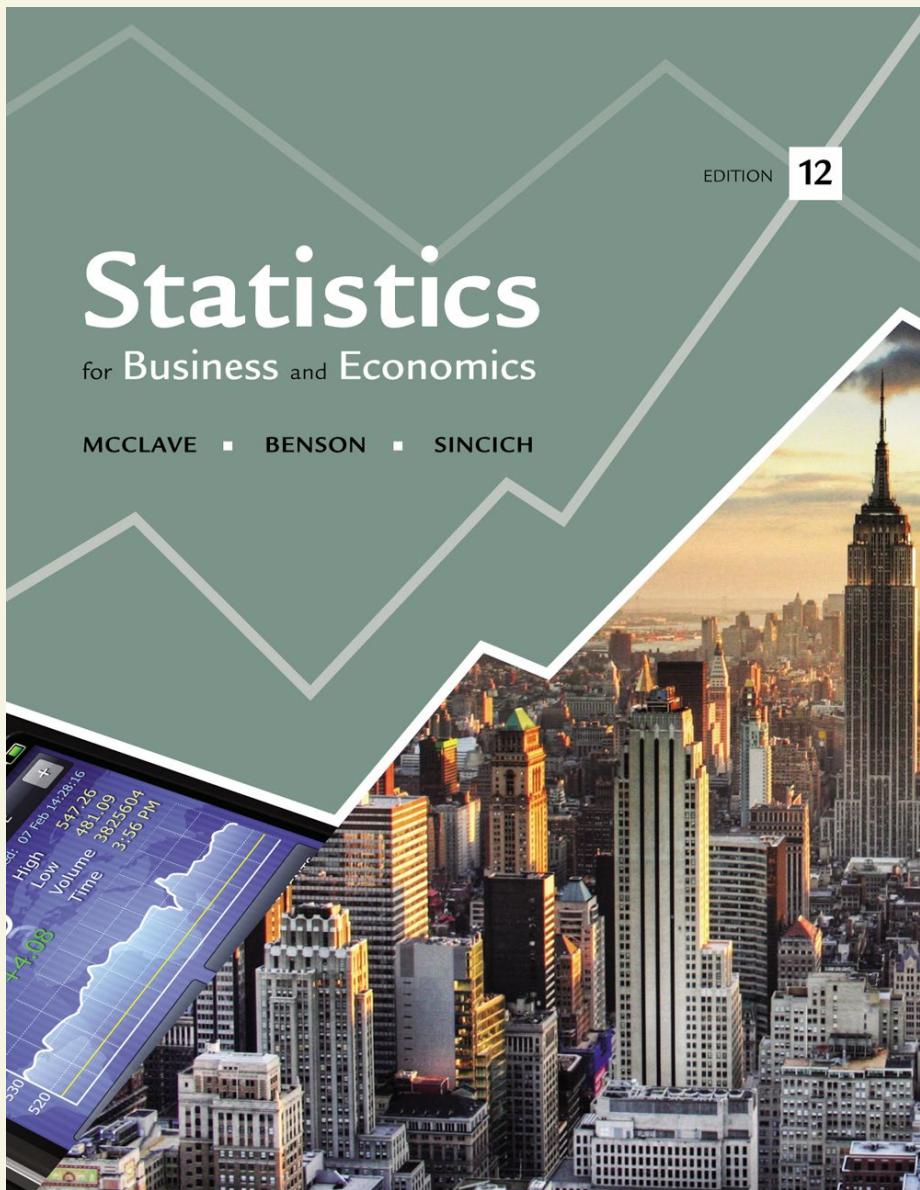
EDITION

12

# Statistics

for Business and Economics

MCCLAVE ■ BENSON ■ SINCICH



# **Statistics for Business and Economics**

## **Chapter 11 Simple Linear Regression**

# Contents

1. Probabilistic Models
2. Fitting the Model: The Least Squares Approach
3. Model Assumptions
4. Making Inferences about the Slope  $\beta_1$

# Contents

5. The Coefficients of Correlation and Determination
6. Using the Model for Estimation and Prediction
7. A Complete Example

# Regression Analysis - Uses

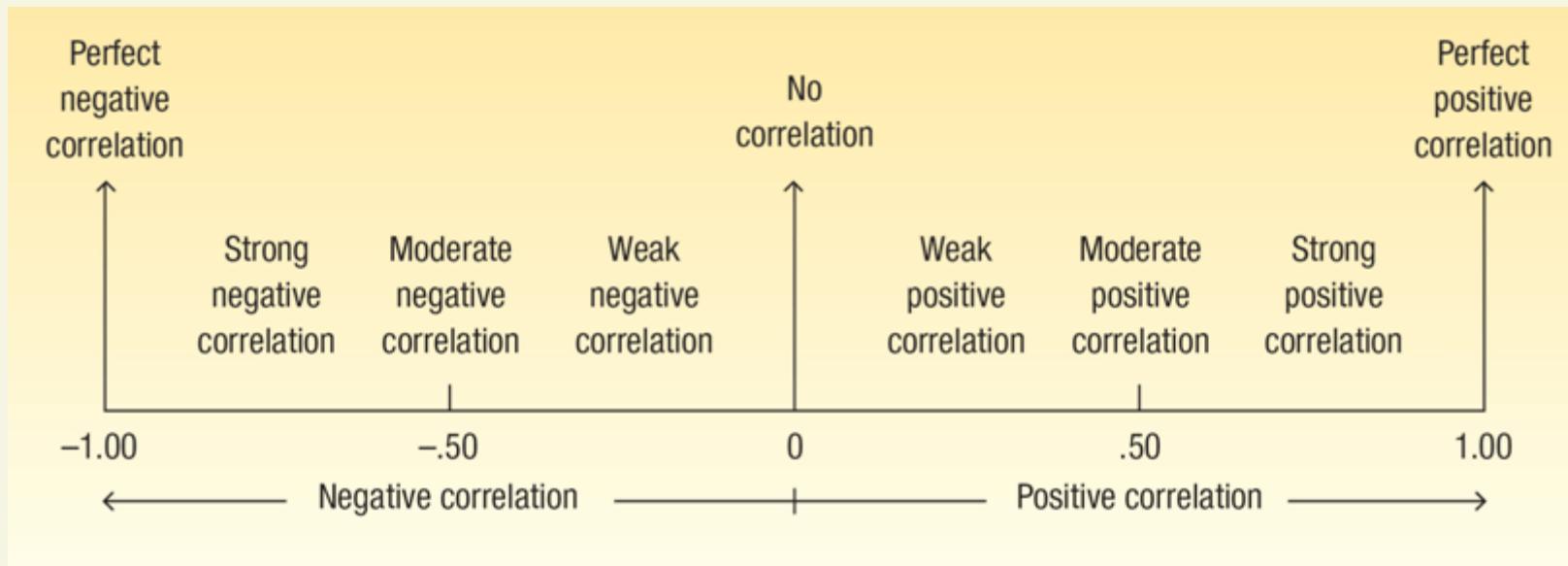
Some examples.

- Is there a relationship between the amount Healthtex spends per month on advertising and its sales in the month?
- Can we base an estimate of the cost to heat a home in January on the number of square feet in the home?
- Is there a relationship between the miles per gallon achieved by large pickup trucks and the size of the engine?
- Is there a relationship between the number of hours that students studied for an exam and the score earned?

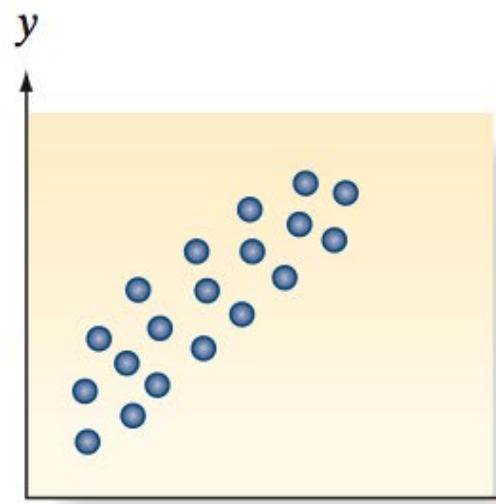
# Correlation Models

- Answers ‘How strong is the **linear** relationship between two variables?’
- Measure the association between two variables.
- Coefficient of correlation
  - Sample correlation coefficient denoted  $r$
  - Values range from  $-1$  to  $+1$
  - Measures degree of association
  - Does not indicate cause–effect relationship

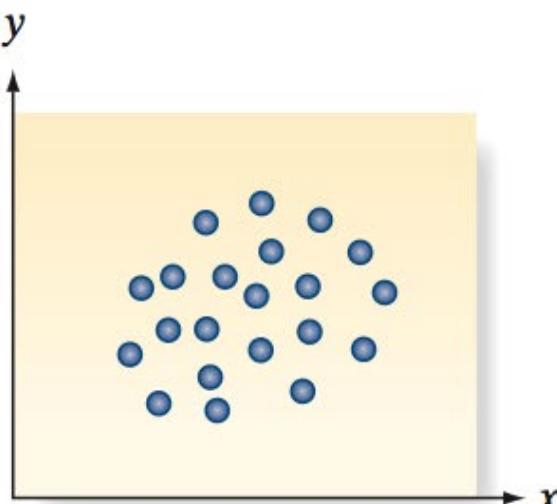
# Correlation Coefficient



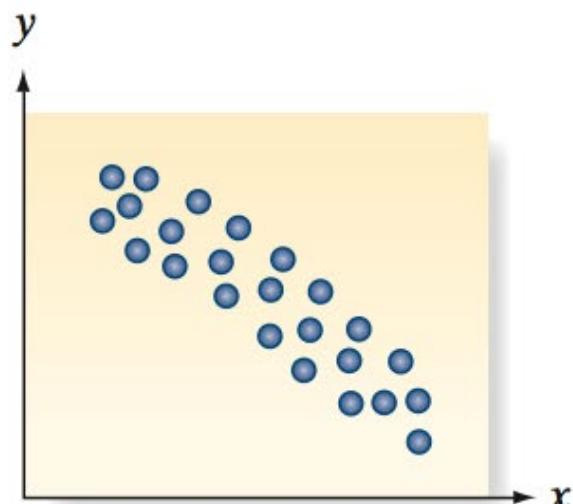
# Coefficient of Correlation



a. Positive  $r$ :  $y$  increases as  $x$  increases



b.  $r$  near 0: little or no relationship between  $y$  and  $x$



c. Negative  $r$ :  $y$  decreases as  $x$  increases

# 11.1

## Probabilistic Models

# Probabilistic Models

- Hypothesize two components
  - Deterministic
  - Random error
- Example: sales volume ( $y$ ) is 10 times advertising spending ( $x$ ) + random error
  - $y = 10x + \varepsilon$
  - Random error may be due to factors other than advertising

# A Simple Linear Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where

$y$  = **Dependent or response variable**

(variable to be modeled)

$x$  = **Independent or predictor variable**

(variable used as a predictor of  $y$ )

$\varepsilon$  (epsilon) = Random error component

# A Simple Linear Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$\beta_0$  (beta zero) = **y-intercept of the line**

$\beta_1$  (beta one) = **slope of the line.** Denotes the change (amount of increase or decrease) in  $y$  for every 1-unit increase in  $x$

- The values of  $\beta_0$  and  $\beta_1$  are population parameters and thus will be unknown in almost all practical applications of regression analysis.
- We need to use sample data to estimate  $\beta_0$  and  $\beta_1$ .

# Estimation Steps

- Step 1: Set up the regression that relates the dependent variable  $y$ , to the independent variable  $x$ .
- Step 2: Use the sample data to estimate unknown parameters in the model.
- Step 3: Examine assumptions of the error term
- Step 4: Statistically evaluate the usefulness of the model.
- Step 5: When satisfied that the model is useful, use it for prediction, estimation, and other purposes.

# **11.2**

## **Fitting the Model: The Least Squares Approach**

# Estimation

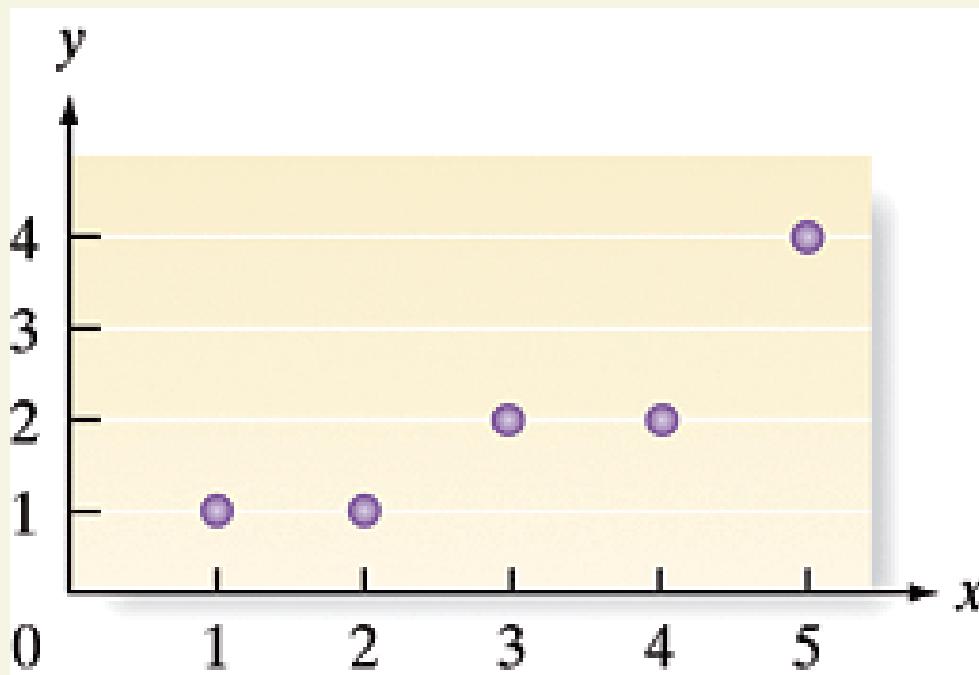
- Suppose an appliance store conducts a 5-month experiment to determine the effect of advertising on sales revenue. The results are shown here. Data file: ADSALE

Table 11.1 Advertising-Sales Data		
Month	Advertising Expenditure, $x$ (\$100s)	Sales Revenue, $y$ (\$1,000s)
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

- Consider the straight-line model,  $y = \beta_0 + \beta_1 x + \varepsilon$ , where  $y$  = sales revenue (thousands of dollars) and  $x$  = advertising expenditure (hundreds of dollars).

# Scatterplot

1. Plot of all  $(x_i, y_i)$  pairs
2. Suggests how well model will fit



# Fitting the model

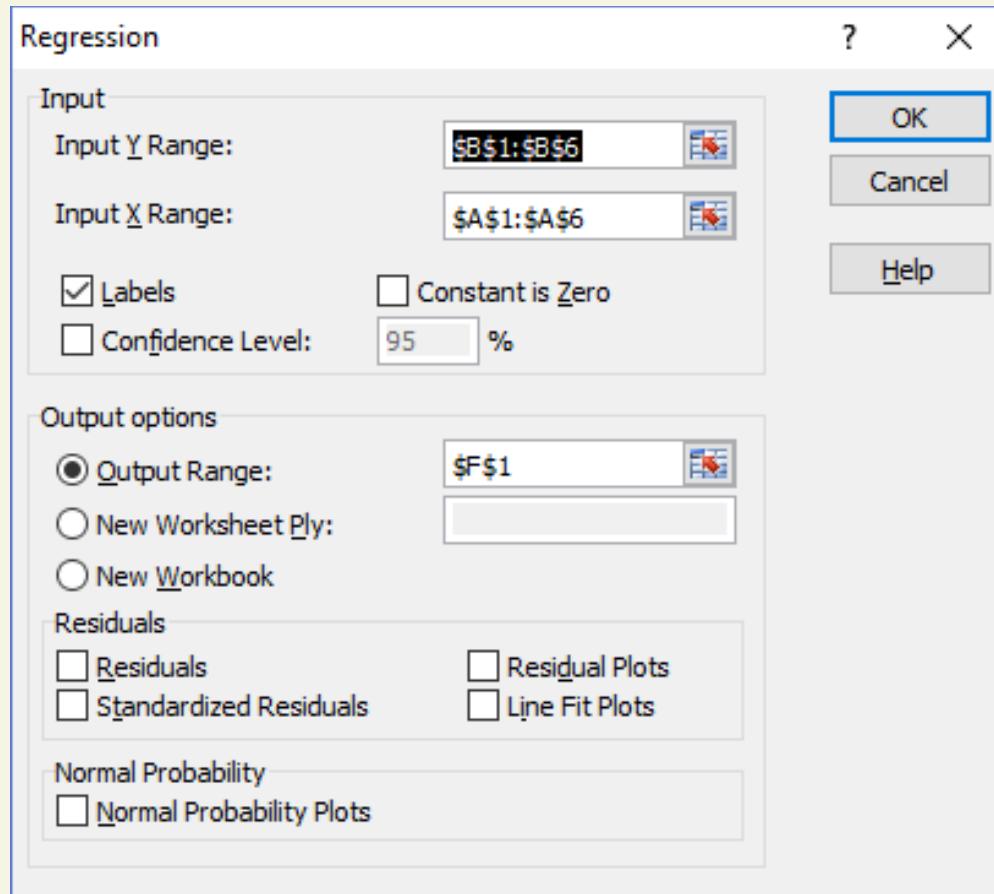
$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Errors of prediction
  - Measures the extent to which the data points deviate from the line (our prediction)
  - It is the vertical differences between the observed and the predicted values of  $y$  using the fitted line
  - Sum of the errors equals 0

# Least Squares Line

- It is the line that best fits the data.
- The sum of squared errors (SSE) is smaller than for any other straight-line model.

# Excel Analysis



Excel: Data->Data Analysis->Regression

# Excel Output

## SUMMARY OUTPUT

### Regression Statistics

Multiple R	0.90
R Square	0.82
Adjusted R Square	0.76
Standard Error	0.61
Observations	5.00

### ANOVA

	df	SS	MS	F	Significance F
Regression	1.00	4.90	4.90	13.36	0.04
Residual	3.00	1.10	0.37		
Total	4.00	6.00			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-0.10	0.64	-0.16	0.88	-2.12	1.92	-2.12	1.92
ADVEXP_X	0.70	0.19	3.66	0.04	0.09	1.31	0.09	1.31

# Least Squares Line

- In general, the estimated least squares line is written as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- $\hat{\beta}_0$  and  $\hat{\beta}_1$  denotes the estimated  $\beta_0$  and  $\beta_1$  that are obtained using the sample data.
- For our example,  $\hat{\beta}_0 = -0.1$ ,  $\hat{\beta}_1 = 0.7$ , so the least squares line is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -.1 + .7x$$

# Interpreting the Estimates of $\beta_0$ and $\beta_1$ in Simple Linear Regression

- *y-intercept:*  $\hat{\beta}_0$  represents the predicted value of  $y$  when  $x = 0$ 
  - Caution: This value will not be meaningful if the value  $x = 0$  is nonsensical or outside the range of the sample data.
- *slope:*  $\hat{\beta}_1$  represents the increase (or decrease) in  $y$  for every 1-unit increase in  $x$ 
  - Caution: This interpretation is valid only for  $x$ -values within the range of the sample data.

# Interpretation

- $\hat{\beta}_0 = -.1$ : the y-intercept is not within the range of the sampled values of x and thus is meaningless.
- $\hat{\beta}_1 = .7$ : this implies that for every unit increase of x, the mean value of y is estimated to increase by .7 unit.
- In terms of this example, for every \$100 increase in advertising, the mean sales revenue is estimated to increase by \$700.

# 11.4

## Assessing the Utility of the Model: Making Inferences about the Slope $\beta_1$

# Inference on $\beta_1$

- If the slope coefficient  $\beta_1$  is zero, then

$$y = \beta_0 + \varepsilon$$

- The simple linear regression is invalid/useless.
- We need to perform hypothesis test on  $\beta_1$  to check if  $\beta_1$  is different from zero.
- The point estimator of  $\beta_1$  is  $\hat{\beta}_1$ , which can be used to test the hypothesis  $\beta_1 = 0$

# A Test of Model Utility: Simple Linear Regression

## A Test of Model Utility: Simple Linear Regression

### One-Tailed Test

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 < 0 \text{ (or } H_a: \beta_1 > 0\text{)}$$

### Two-Tailed Test

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

$$\text{Test statistic: } t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}}$$

One-tail test p-value is obtained by dividing the p-value shown in Excel by 2

# Example

- $H_0: \beta_1 = 0$
- $H_a: \beta_1 \neq 0$
- P-value=0.04, which is less than 5% (default significance level); the null is rejected.
- What can we say:
  - The slope coefficient is significant(ly different from zero).
  - Sales revenue and advertising expenditure are linearly related.
  - The model is valid.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-0.10	0.64	-0.16	0.88	-2.12	1.92	-2.12	1.92
ADVEXP_X	0.70	0.19	3.66	0.04	0.09	1.31	0.09	1.31

# 11.5

## The Coefficients of Determination

# Coefficient of Determination: $R^2$

It represents the proportion of the total sample variability in  $y$  that is explained by the linear relationship between  $y$  and  $x$ .

$$r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

The higher the coefficient of determination is, the better the model fits the data.

$$0 \leq r^2 \leq 1$$



# Example

- Calculate the coefficient of determination for the advertising-sales example. Interpret the result.

Regression Statistics	
Multiple R	0.90
R Square	0.82
Adjusted R Square	0.76
Standard Error	0.61
Observations	5.00

**Interpretation:** About 82% of the sample variation in Sales ( $y$ ) can be explained by using Ad \$ ( $x$ ) to predict Sales ( $y$ ) in the linear model.

# **11.6**

## **Using the Model for Estimation and Determination**

# Prediction and Intervals

- Confidence interval: inferences on the mean value of  $y$  for multiple months
  - Estimate the mean sales for all months during which \$400 ( $x = 4$ ) is expended on advertising
- Prediction interval: inference on the value of  $y$  for one month
  - If we expend \$400 in advertising next month, we want to predict the sales revenue for that month

# Example

Refer to the sales-appraisal simple linear regression in previous examples. The estimated regression equation is

$$\hat{y} = -.1 + .7x$$

- a. Predict the monthly sales when the store spends \$400 on advertising.
- b. Find a 95% confidence interval for the mean monthly sales when the store spends \$400 on advertising.
- c. Predict the monthly sales for next month using a 95% prediction interval if \$400 is spent on advertising.

# Minitab Output

## Predicted Values for New Observations

New

Obs	Fit	SE Fit	95% CI	95% PI
1	2.700	0.332	(1.645, 3.755)	(0.503, 4.897)

## Values of Predictors for New Observations

New

Obs	ADVEXP_X
1	4.00

- a. Point estimator:

$$\hat{y} = -.1 + .7x = -.1 + .7 * 4 = 2.7$$

- a. 95% CI: (1.645, 3.755)  
b. 95% PI: (0.503, 4.897)

# 11.7

## A Complete Example

# Example

Suppose a fire insurance company wants to relate the amount of fire damage in major residential fires to the distance between the burning house and the nearest fire station. The study is to be conducted in a large suburb of a major city; a sample of 15 recent fires in this suburb is selected. The amount of damage,  $y$ , and the distance between the fire and the nearest fire station,  $x$ , are recorded for each fire. Data file: RFIRES

# Example

## Fire Damage Data

Distance from Fire Station, $x$ (miles)	Fire Damage, $y$ (thousands of dollars)
3.4	26.2
1.8	17.8
4.6	31.3
2.3	23.1
3.1	27.5
5.5	36.0
.7	14.1
3.0	22.3
2.6	19.6
4.3	31.3
2.1	24.0
1.1	17.3
6.1	43.2
4.8	36.4
3.8	26.1



Data Set: FIREDAM

# Example

**Step 1:** First, we hypothesize a model to relate fire damage,  $y$ , to the distance from the nearest fire station,  $x$ . We hypothesize a simple linear regression model:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

# Example

**Step 2:** Use a statistical software package to estimate the unknown parameters in the deterministic component of the hypothesized model.

# Example

Regression Analysis						
Regression Statistics						
Multiple R	0.960977715					
R Square	0.923478169					
Adjusted R Square	0.917591874					
Standard Error	2.316346184					
Observations	15					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	841.766358	841.766358	156.8861596	1.2478E-08	
Residual	13	69.75097535	5.365459643			
Total	14	911.5173333				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	10.27792855	1.420277811	7.236562082	6.58556E-06	7.209605476	13.34625162
DISTANCE	4.919330727	0.392747749	12.52542054	1.2478E-08	4.070850963	5.767810491

$$\hat{\beta}_1 = 4.919331$$

$$\hat{\beta}_0 = 10.277929$$

Least Squares Equation:  $\hat{y} = 10.278 + 4.919x$

# Example

Interpreting the coefficients:

$\hat{\beta}_1 = 4.919$ : the estimated mean damage increases by \$4,919 for each additional mile from the fire station.

$\hat{\beta}_0 = 10.278$ : Because  $x = 0$  is outside the range in this case, the y-intercept has no practical interpretation.

# Example

**Step 3:** First, test the slope  $\beta_1$  to if fire damage increases as the distance increases.

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 > 0$$

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	10.28	1.42	7.24	6.5856E-06	7.21	13.35	7.21	13.35
DISTANCE	4.92	0.39	12.53	1.2478E-08	4.07	5.77	4.07	5.77

P-value for the test is virtually zero. Therefore we reject the null hypothesis and conclude that fire damage increases as the distance increases.

# Example

The coefficient of determination is  $r^2 = .9235$ , which implies that about 92% of the sample variation in fire damage ( $y$ ) is explained by the distance ( $x$ ) between the fire and the fire station.

# Example

**Step 4:** We are now prepared to use the least squares model. Suppose the insurance company wants to predict the fire damage if a major residential fire were to occur 3.5 miles from the nearest fire station. A 95% confidence interval and prediction interval when  $x = 3.5$  are shown on the Minitab printout on the next slide.

# Example

## Predicted Values for New Observations

New

Obs	Fit	SE Fit	95% CI	95% PI
1	27.496	0.604	(26.190, 28.801)	(22.324, 32.667)

## Values of Predictors for New Observations

New

Obs	DISTANCE
1	3.50

**Figure 10.27**

Minitab confidence and prediction interval for fire damage regression

# Example

The predicted value (highlighted on the printout) is  $\hat{y} = 27.496$ , while the 95% prediction interval (also highlighted) is (22.3239, 32.6672). Therefore, with 95% confidence we predict fire damage in a major residential fire 3.5 miles from the nearest station to be between \$22,324 and \$32,667.

Zikmund

Babin

Carr

Griffin

ninth edition

# The Role of Business Research





1. Understand how research contributes to business success
2. Know how to define business research
3. Understand the difference between basic and applied business research
4. Understand how research activities can be used to address business decisions
5. Know when business research should and should not be conducted
6. Appreciate the way technology and internationalization are changing business research

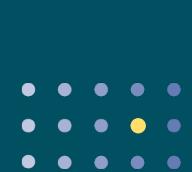




# ESPN Hits a Home Run

- ESPN has information in many databases.
- Business research integrated it so they could learn more about how fans use their media.
- Gaining intelligence had bottom-line implications for their own revenue and their advertisers

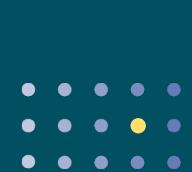




# Business Research Defined

- Business research is the application of the scientific method in searching for the truth about business phenomena.
  - *Financial managers, personnel managers; marketing managers.*
- “It ain’t the things we don’t know that gets in trouble. It’s the things we know that ain’t so.”  
—Artemus Ward
  - *Accurate and objective information*
  - *Test idea*



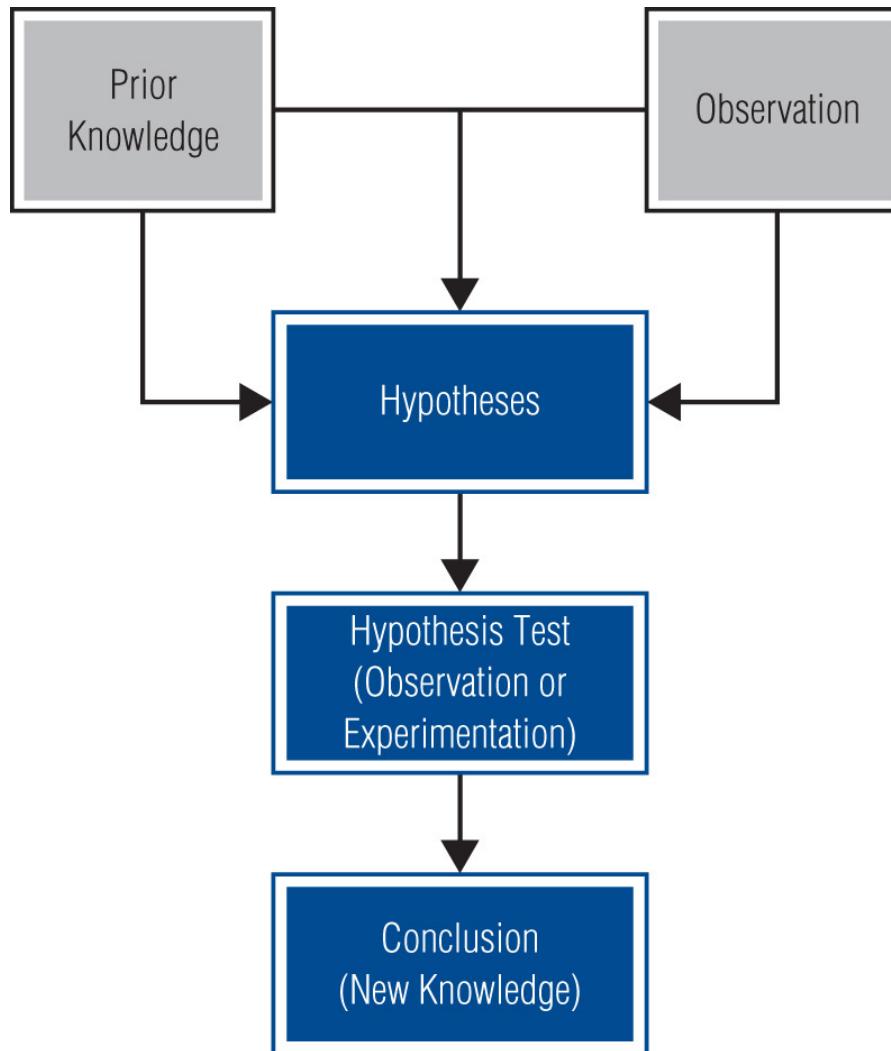


# Business Research Defined

- The process includes:
  - *idea and theory development*
  - *problem definition*
  - *searching for and collecting information*
  - *analyzing data*
  - *communicating the findings and their implications*



# The Scientific Method





# Managerial Value of Business Research

- Business Strategies
- Product-oriented: technical superiority
  - *technicians and experts in the field*
- Production-oriented: production process
  - *Input from workers, engineers, finance and accounting*
- Marketing-oriented: consumer value
  - *Focus on customer: desires, beliefs and attitude*
  - *Yoplait Go-Gurt*



# Business Class Success?

- Business-class travelers want comfort, good food, and convenient boarding, but the price is hefty.
- Two start-ups offered “discount” business-class-only airlines but failed.
- Could more effective research have determined that these were not feasible business ventures?





# Harley-Davidson Goes Abroad

- 
- Consumers in different countries have different preferences.
  - Even if consumers want it, government regulations can make it prohibitive (e.g., India).
  - Harley is pursuing the U.S. women's market for bikes.



Zikmund

Babin

Carr

Griffin

ninth edition

# The Business Research Process: An Overview





## Learning Outcomes

1. Define decision making and understand the role research plays in making decisions
2. Classify business research as either exploratory research, descriptive research, or causal research
3. List the major phases of the research process and the steps within each
4. Explain the difference between a research project and a research program





# Introduction

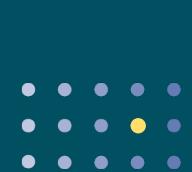
- Business research contributes to decision making.
- Firms seek potential opportunities or deal with business problems.
- Decision making:
  - *process of developing and deciding among alternative ways or resolving a problem or choosing from among alternative opportunities*



# Decision Making

- Business opportunity
  - *A situation that makes some potential competitive advantage possible.*
  - *Ebay and Garage sale*
- Business problem
  - *A situation that makes some significant negative consequence more likely.*
- Symptoms
  - *Observable cues that serve as a signal of a problem because they are caused by that problem.*





# Decision Making

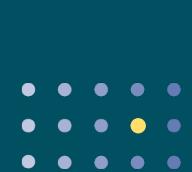
- Research's role in the decision making process
  - *Recognizing the nature of the problem or opportunity.*
  - *Identifying how much information is currently available and how reliable it is.*
  - *Determining what information is needed to better deal with the situation.*



# Conditions Affecting Decision Making

- Certainty
  - *The decision maker has all information needed to make an optimal decision.*
  - *Business research not needed.*
- Uncertainty
  - *The manager grasps the general nature of desired objectives, but the information about alternatives is incomplete.*
  - *Eg: firm refinancing*
- Ambiguity
  - *The nature of the problem itself is unclear such that objectives are vague and decision alternatives are difficult to define.*
  - *Eg: fast-food restaurant sales*





# Types of Business Research

- Exploratory
- Descriptive
- Causal





# Exploratory Research

- Exploratory Research
  - *Conducted to clarify ambiguous situations or discover ideas that may be potential business opportunities.*
  - *Initial research conducted to clarify and define the nature of a problem.*
    - ▷ Does not provide conclusive evidence
    - ▷ Subsequent research expected
  - *Eg: Sony and Honda researching on robot*



# Cute, Funny, or Sexy? What Makes a Mascot Tick?

- Questions:
  - *Pillsbury Doughboy ever changed?*
  - *How old should be brawny man be?*
  - *What should M&M be named?*
- Mr. Peanut in Bermuda shorts?
- M&M characters are referred to by their color.
- Women want a sexy Brawny man!



# Descriptive Research

- Describes characteristics of objects, people, groups, organizations, or environments.
  - *Paint a picture: who, what, when, where, why, and how questions.*
  - *Considerable understanding of the nature of the problem exists.*
  - *Accuracy is important*
  - *Eg: Wholefood Markets; online MBA*
- Diagnostic analysis
  - Seeks to diagnose reasons for market outcomes and focuses specifically on the beliefs and feelings consumers have about and toward a competing products.





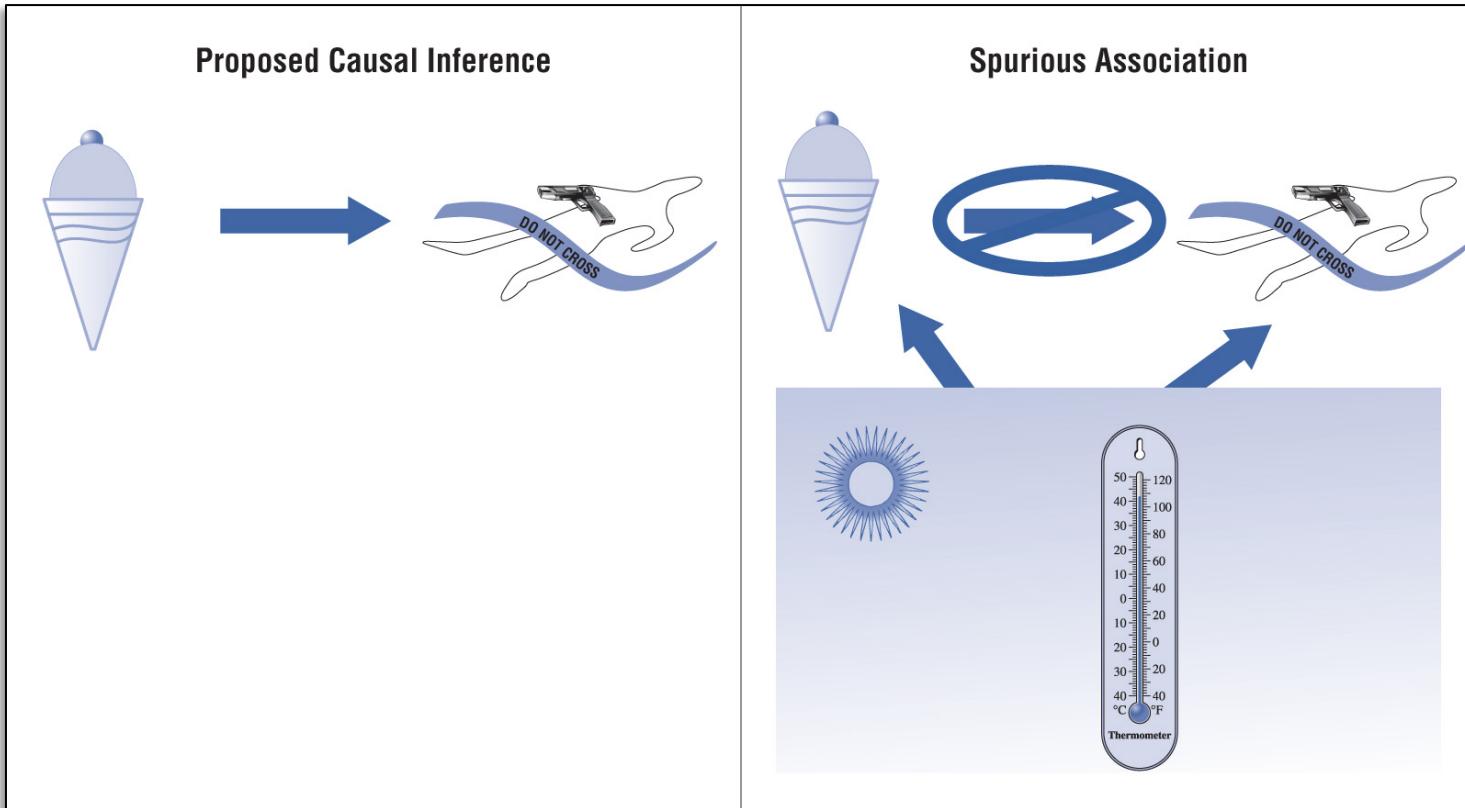
# Causal Research

- Research conducted to identify cause and effect relationships (inferences).
- Evidence of Causality:
- Temporal sequence-the appropriate causal order of events.
- Concomitant variation-two phenomena vary together.
- Nonspurious association-an absence of alternative plausible explanations.



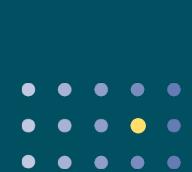
EXHIBIT 4.2

# The Spurious Effect of Ice Cream



<http://www.tylervigen.com/spurious-correlations>





# Degrees of Causality

- Absolute Causality
  - *The cause is necessary and sufficient to bring about the effect.*
- Conditional Causality
  - *A cause is necessary but not sufficient to bring about an effect.*
- Contributory Causality
  - *A cause need be neither necessary nor sufficient to bring about an effect.*
  - *Weakest form of causality*





# Experiments

- Experiment
  - *A carefully controlled study in which the researcher manipulates a proposed cause and observes any corresponding change in the proposed effect.*
- Experimental variable
  - *Represents the proposed cause and is controlled by the researcher by manipulating it.*
- Manipulation
  - *The researcher alters the level of the variable in specific increments.*
- Test-market
  - *An experiment that is conducted within actual market conditions.*

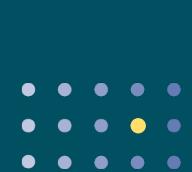


# Wee Box example

Wee Box Sales by Condition		
	High Price	Low Price
Specialty Distribution	Peoria, Illinois: Retail Price: \$200 Retail Store: Best Buy	Des Moines, Iowa: Retail Price: \$100 Retail Store: Best Buy
General Distribution	St. Louis, Missouri: Retail Price: \$200 Retail Store: Big Cheap-Mart	Kansas City: Missouri: Retail Price: \$100 Retail store: Big Cheap-Mart

Assuming that Wee Box consumers are the same in each of these cities, the extent to which price and distribution cause sales can be examined by comparing sales results in each of these four conditions





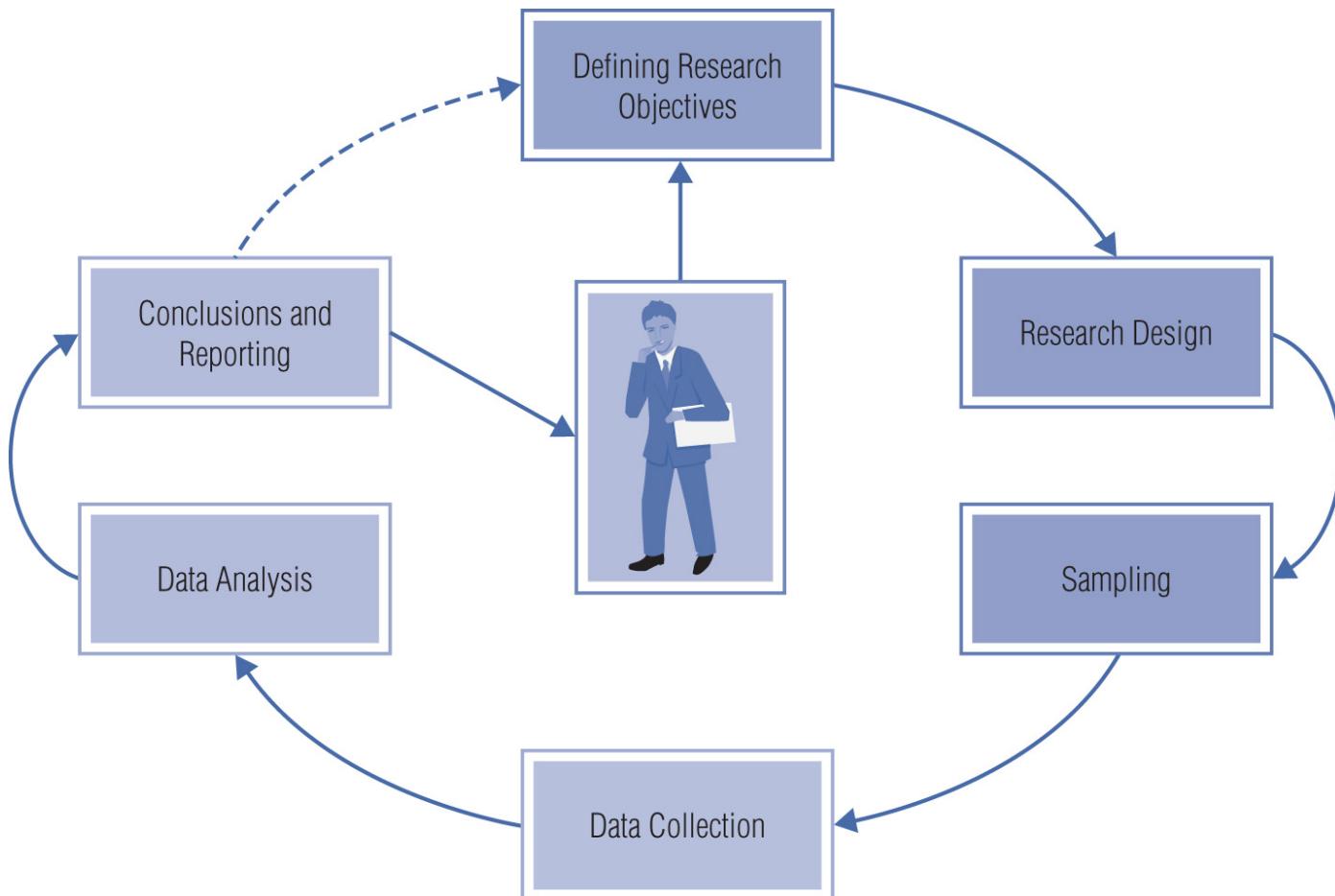
# Stages in the Research Process

- Process stages:
  1. *Defining the research objectives*
  2. *Planning a research design*
  3. *Planning a sample*
  4. *Collecting the data*
  5. *Analyzing the data*
  6. *Formulating the conclusions and preparing the report*



EXHIBIT 4.5

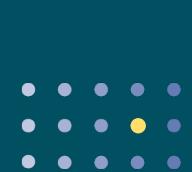
# Stages of the Research Process



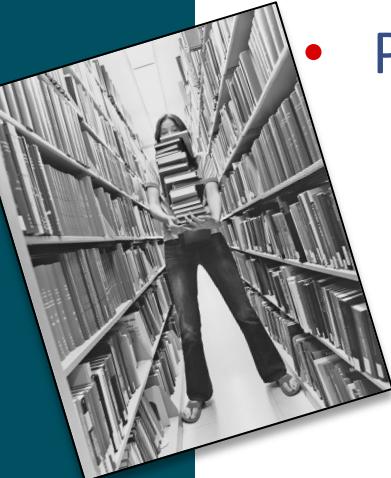
# Defining the Research Objectives

- Research objectives
  - *The goals to be achieved by conducting research.*
- Deliverables
  - *The consulting term used to describe research objectives to a research client.*
  - **Research proposal:** involves managers and researchers
- “A problem well defined is a problem half solved”





# Exploratory Research Techniques



- Previous Research

- *Literature review*

- A directed search of published works, including periodicals and books, that discusses theory and presents empirical results that are relevant to the topic at hand.

- *Pilot Studies*

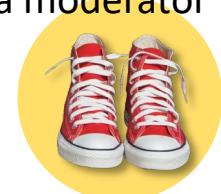
- A small-scale research project that collects data from respondents similar to those to be used in the full study.

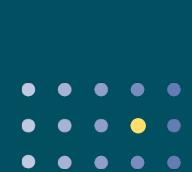
- Pretest

- ◆ A small-scale study in which the results are only preliminary and intended only to assist in design of a subsequent study.

- Focus Group

- ◆ A small group discussion about some research topic led by a moderator who guides discussion among the participants.





# Planning the Research Design

- Research Design
  - *A master plan that specifies the methods and procedures for collecting and analyzing the needed information.*
  - *Basic design techniques for descriptive and causal research:*
    - ▷ Surveys: interview; questionnaire
    - ▷ Experiments: laboratory; field
    - ▷ Secondary data
    - ▷ Observation





# Sampling

- Sampling
- Involves any procedure that draws conclusions based on measurements of a portion of the population.
- Sampling decisions



Zikmund

Babin

Carr

Griffin

ninth edition

# Problem Definition: The Foundation of Business Research



# Deland Trucking Has a “Recruitment” Problem

- “Why are our recruiting costs so high?”
- Costs for driver selection and recruitment have not gone up.
- The company has to do so many orientation and hiring sessions.
- The researcher needs to put together a proposal.



# Good Decisions Start with a Good Problem Definition

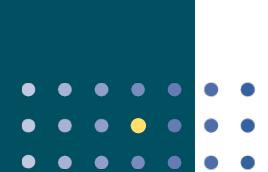
- Translate business decision situation into specific research objectives
- Research serves management



# Good Answers, Bad Questions?

- “New Coke”:
  - *Blind taste test*
  - *taste or brand*
- Ford’s Edsel:
  - *Styling;*
  - *Poor workmanship;*
  - *failed to understand consumer;*
  - *The name, Edsel, was never tested*
- Smokeless cigarettes:
  - *Smokeless and smoke never compared*





# The Problem mean gaps

- Business performance is worse than expected business performance
  - *Sales, profits, margins below targets set by management*
- Actual business performance is less than possible business performance.
  - *Opportunity seeking*
- Expected business performance is greater than possible business performance.
  - *Total market potential*

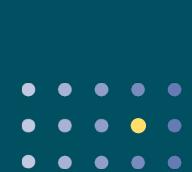




## EXHIBIT 6.2

# The Problem-Definition Process





# Understand the Business Decision

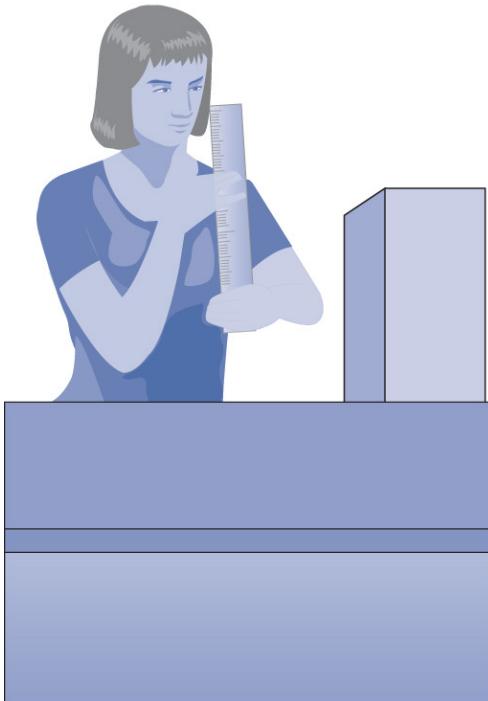
- Situation analysis
  - *Gather background information*
- Interview process
  - *Interview all relevant parties*
  - *Develop many alternative problem statements*
  - *Think about possible solutions to the problem*
  - *Make lists of ideas*
  - *Be open-minded*
- Identifying all key symptoms





## EXHIBIT 6.3

# What Has Changed?



Question: What changes have occurred recently?

Probe: Tell me about this change.

Probe: What has brought this about?

Problem: How might this be related to your problem.

Question: What other changes have occurred recently (i. e. competitors, customers, environment, pricing, promotion, suppliers, employees, etc.)?

Continue Probing



Firm's Situation	Symptoms	Likely Problem(s)	Decision Statement
22-year-old neighborhood swimming association seeks research help	<ul style="list-style-type: none"> <li>Declining membership for six years</li> <li>Increased attendance at new water park</li> <li>Less frequent usage among members</li> </ul>		
Manufacturer of palm-sized computer with wireless Internet access believes that B2B sales are too low	<ul style="list-style-type: none"> <li>Distributors complain prices are too high</li> <li>Business users still use larger computers for displaying information to customers or smartphones for other purposes</li> </ul>		
A new microbrewery is trying to establish itself	<ul style="list-style-type: none"> <li>Consumers seem to prefer national brands over the local microbrew products</li> <li>Many customers order national brands within the microbrew itself</li> <li>Some customers hesitate to try new microbrew flavors</li> </ul>	Is there a negative flavor gap? Do consumers appreciate the microbrew approach and the full beer tasting (as opposed to drinking) experience?	<p>How can we encourage more consumers to come to the microbrew and try our products?</p> <p>Should we redesign the brewery to be more inviting?</p>

# Clarity in Research Questions and Hypotheses

- Research Questions—be specific
  - “*Is advertising copy 1 better than advertising copy 2*”
  - *Managerial Action Standard: provide input that can be used as a standard for selecting from among alternative solutions.*
    - “If sales is higher than X, management will do plan A; otherwise plan B”
- Hypotheses
  - *Statements that can be empirically tested.*





# Translating Decision Statements into Research Objectives/Questions

## Writing Research Hypothesis



Decision Statements	Research Objectives	Research Questions	Research Hypotheses
What things can be done to energize new markets and create a more favorable attitude toward the association?			
What things can be done to improve competitive positioning of the new product in B2B markets?			
How can we encourage more consumers to come to the microbrew and try our products? Should we redesign the brewery to be more inviting?	<p>Describe how situational factors influence beer consumption and consumer attitudes toward beer products.</p> <p>List factors that will improve attitudes toward the microbrewery.</p>	<p>Do situational factors (such as time of day, food pairings, or environmental factors) relate to taste perceptions in beer?</p>	<p>Microbrew beer is preferred when consumed with food.</p> <p>An exciting atmosphere will improve consumer attitudes toward the microbrew.</p>



# Determine the Relevant Variable

- What is a Variable?
  - *Anything that varies or changes from one instance to another.*
- What is a Constant?
  - *Something that does not change; is not useful in addressing research questions.*



# Types of Variables

- Continuous variable
  - *Can take on a range of quantitative values.*
- Categorical variable
  - *Indicates membership in some group.*
  - *Also called classificatory variable.*
- Dependent variable
  - *A process outcome or a variable that is predicted and/or explained by other variables.*
- Independent variable
  - *A variable that is expected to influence the dependent variable in some way.*





## How Much Time Should Be Spent on Problem Definition?

- Budget constraints usually influence how much effort is spent on problem definition.
- The more important the decision faced by management, the more resources should be allocated toward problem definition.
- The time taken to identify the correct problem is usually time well spent.





# The Research Proposal

- Research Proposal
  - *A written statement of the research design.*
- Uses for the Proposal
  - *As a planning tool*
  - *As a contract with management*
- Funded Business Research
- Consult pages 124-125 in your textbook when writing your research proposal but do follow the layout in the guideline I provided.





# Types of Variables

- Continuous variable
  - *Can take on a range of quantitative values.*
- Categorical variable
  - *Indicates membership in some group.*
  - *Also called classificatory variable.*
- Dependent variable
  - *A process outcome or a variable that is predicted and/or explained by other variables.*
- Independent variable
  - *A variable that is expected to influence the dependent variable in some way.*



# Example Dummy Table

	Standardized Regression Coefficient	Rank (Importance)
Increase cents mile		1
Number of long-haul routes (per month)		2
Days off (per month)		3
Vehicle quality		4
Benefits provided		5



Zikmund

Babin

Carr

Griffin

ninth edition

# Survey Research: An Overview

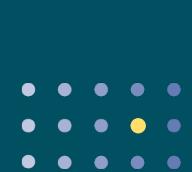




# Survey Research

- Respondents
  - *People who verbally answer an interviewer's questions or provide answers to written questions.*
- Sample Survey
  - *A survey that emphasizes contacting respondents who are a representative sample of the target population.*
- Advantages
  - *Quick*
  - *Inexpensive*
  - *Efficient*





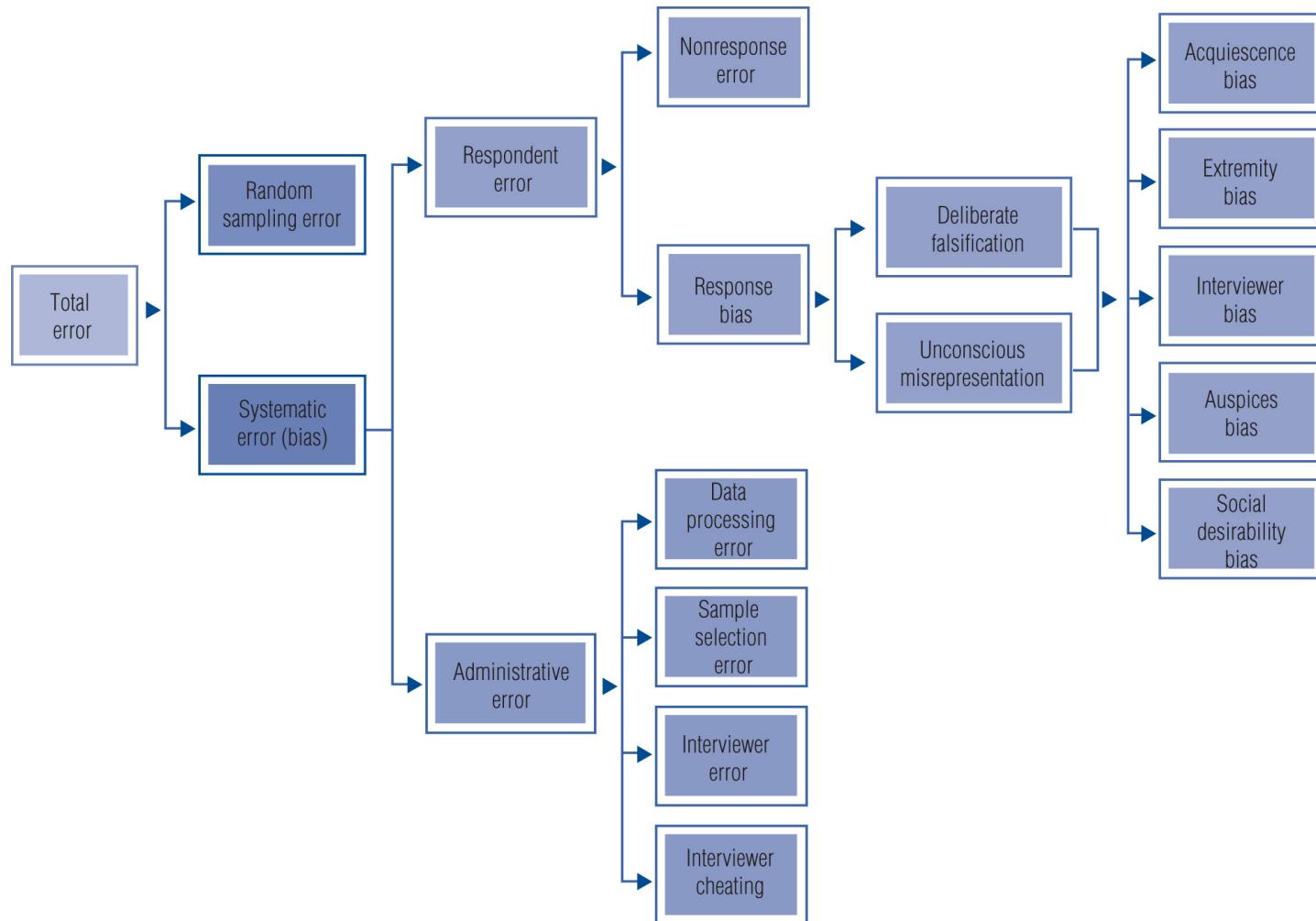
# Errors in Survey Research

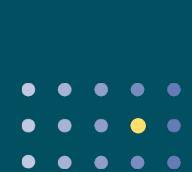
- Random Sampling Error
  - *Respondents are not selected randomly*
  - *Chances of selection vary across respondents.*
  - *1969 Vietnam Draft Lottery*
- Sample Bias
  - *A persistent tendency for the results of a sample to deviate in one direction from the true value of the population parameter.*



EXHIBIT 9.1

# Categories of Survey Errors





# Respondent Error

- Sample bias resulting from respondent side
- Nonresponse Error
  - **No contacts:** people who are not at home or who are otherwise inaccessible on the first and second contact.
  - **Refusals:** People who are unwilling to participate in a research project.
- Self-Selection Bias
  - A bias that occurs because people who feel strongly about a subject are more likely to respond to survey questions than people who feel indifferent about it.



# Overestimating Patient Satisfaction

- On satisfaction surveys, do responses represent a cross section of customers?
- Researchers studied patient satisfaction surveys and found that more-satisfied patients were more likely to complete and return the survey.
- Overestimated satisfaction.
- Sample bias



# Response Bias

- Deliberate Falsification
  - *Occasionally people deliberately give false answers.*
    - ▷ Misrepresent answers to appear intelligent
    - ▷ Conceal personal information
    - ▷ Avoid embarrassment
  - *Average-person hypothesis:*
    - ▷ Individuals may prefer to be viewed as average, so they alter their responses to conform more closely to their perception of the average person.





# Response Bias

- Unconscious Misrepresentation
  - *When a respondent is consciously trying to be truthful and cooperative, response bias can arise from the question format, the question content, or some other stimulus that affects their response to a question.*
  - *Sources of misrepresentation:*
    - ▶ Misunderstanding the question
    - ▶ Unable to recall details
    - ▶ Unprepared response to an unexpected question
    - ▶ Inability to translate feelings into words
    - ▶ After-event underreporting



# Types of Response Bias

- Acquiescence Bias
  - *A tendency to agree with all or most questions.*
- Extremity Bias
  - *The tendency of some Individuals to use extremes when responding to questions.*
- Interviewer Bias
  - *The presence of the interviewer influences respondents' answers.*
- Social Desirability Bias
  - *Bias in responses caused by respondents' desire, either conscious or unconscious, to gain prestige or appear in a different social role.*



# Administrative Error

- An error caused by the improper administration or execution of the research task.
  - ***Data-processing error:*** *incorrect data entry, incorrect computer programming, or other procedural errors during data analysis.*
  - ***Sample selection error:*** *improper sample design or sampling procedure execution.*
  - ***Interviewer error:*** *mistakes made by interviewers failing to record survey responses correctly.*
  - ***Interviewer cheating:*** *filling in fake answers or falsifying questionnaires by an interviewer.*





# Classifying Survey Research Methods

- **Cross-sectional study:**

- *various segments of a population are sampled and data are collected at a single moment in time.*

- **Longitudinal study:**

- **Tracking study:** *uses successive samples to compare trends and identify changes in variables such as consumer satisfaction, brand image, or advertising awareness.*
  - **Consumer panel:** *a survey of the same sample of individuals or households to record (in a diary) their attitudes, behavior, or purchasing habits over time.*

- National Longitudinal Surveys (NLS) by Bureau of Labor Statistics
    - <https://www.bls.gov/nls/nlsy79.htm>



Zikmund

Babin

Carr

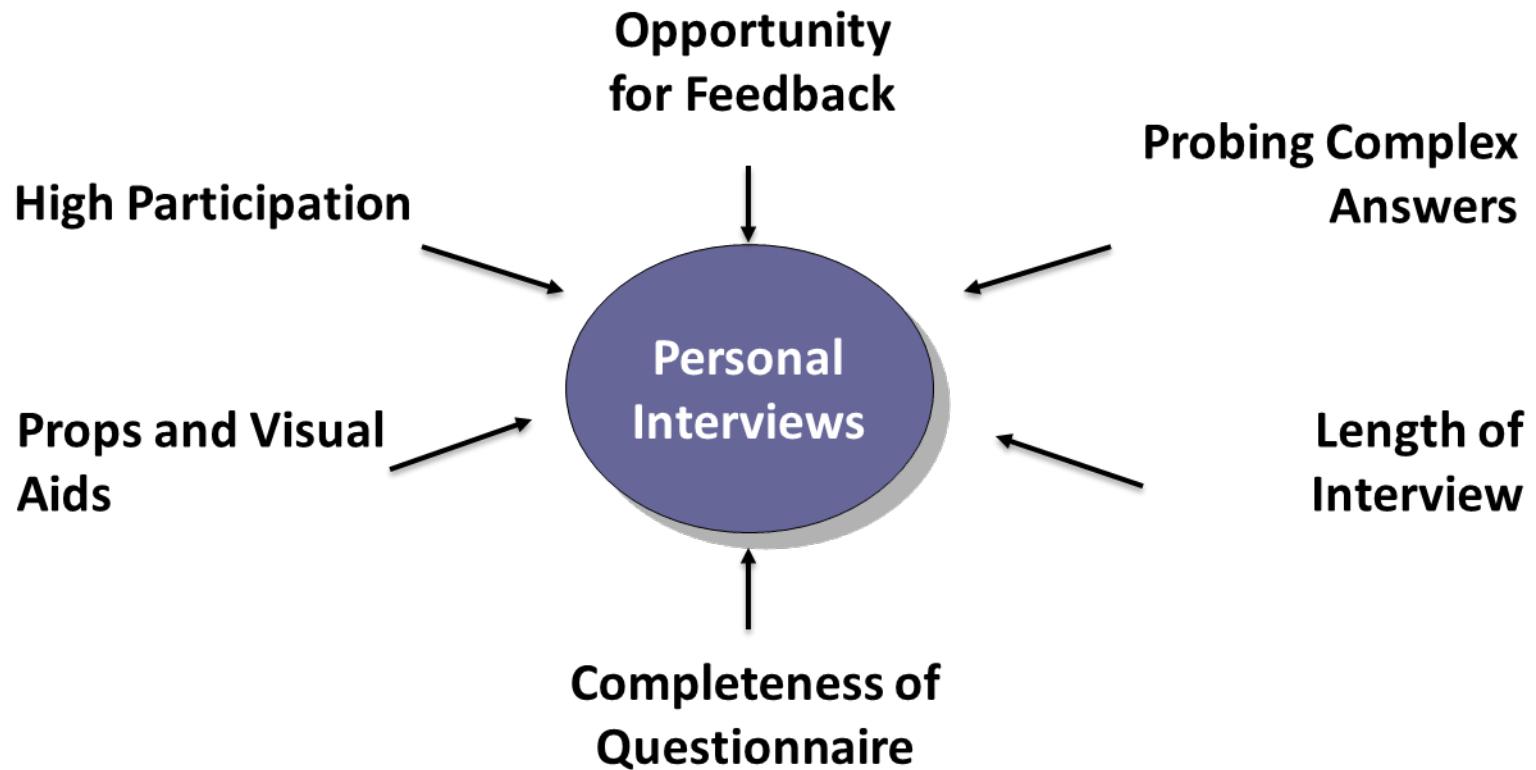
Griffin

ninth edition

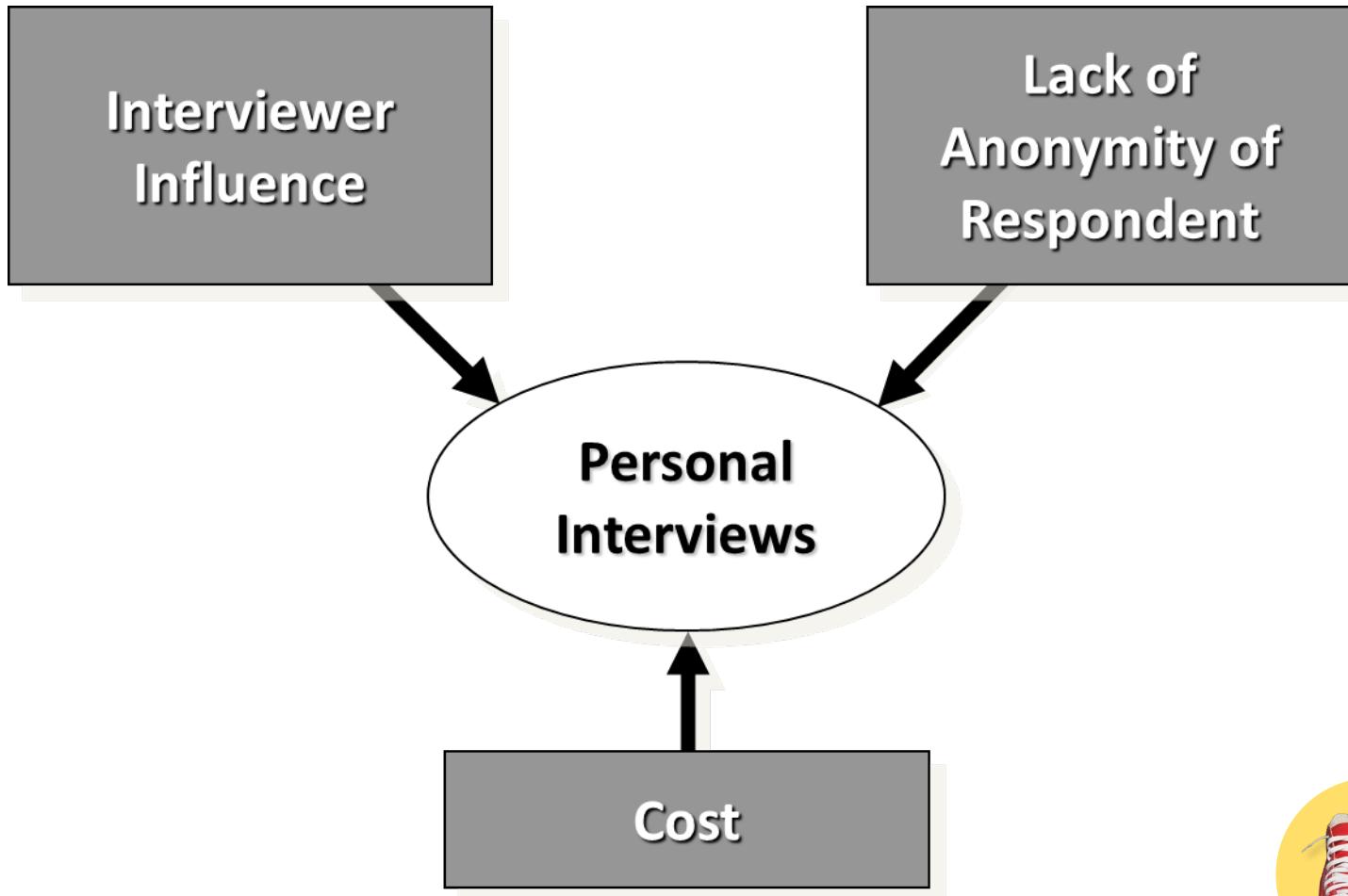
# Survey Research: Communicating with Respondents

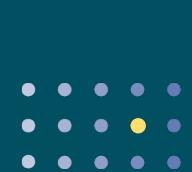


# Personal Interviews—advantages



# Disadvantages of Personal Interviews





# Types of personal interviews

- Door-to-door personal interviews
- Mall intercept interviews



# Telephone Interviews

- Telephone Interviews
  - *Personal interviews conducted by telephone.*
  - *The mainstay of commercial survey research.*
  - *“No-call” legislation has limited this capacity.*

**National Do Not Call Registry**

[REGISTER HOME](#)   [REGISTER A PHONE NUMBER](#)   [VERIFY A REGISTRATION](#)   [MORE INFORMATION](#)   [EN ESPAÑOL](#)   [FILE A COMPLAINT](#)   [PRIVACY AND SECURITY](#)

**REGISTER YOUR HOME OR MOBILE PHONE NUMBER**

Follow the registration steps below. Click here for [detailed registration instructions](#).

1. Enter up to three phone numbers and your email address. Click Submit.  
2. Check for errors. Click Register.  
3. Check your email for a message from Register@donotcall.gov. Open the email and click on the link to complete your registration.

If you share any of these telephone numbers with others, please remember that you are registering for everyone who uses these lines.

**STEP ONE**

Area Code:  Phone:

Email Address:

Confirm Email Address:

Your email address MUST be correct to process your registration. Learn why your email address is required. If you do not receive the verification email within a few minutes, please check your spam filter or junk email folder.

Enter phone numbers with or without a dash. Do not use spaces or periods.

**Submit**





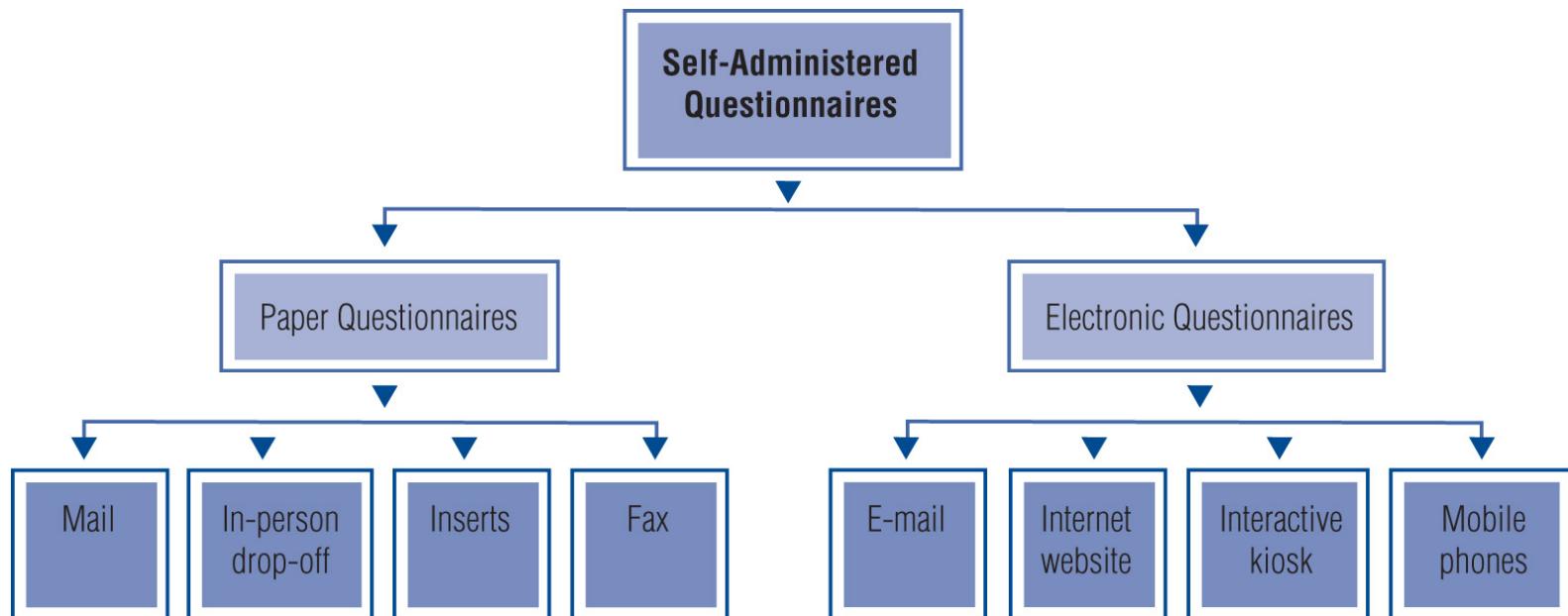
# Phone Interview Characteristics

- Pros
- Cons



EXHIBIT 10.1

## Self-Administered Questionnaires Can Be Either Printed or Electronic





# Mail Questionnaires

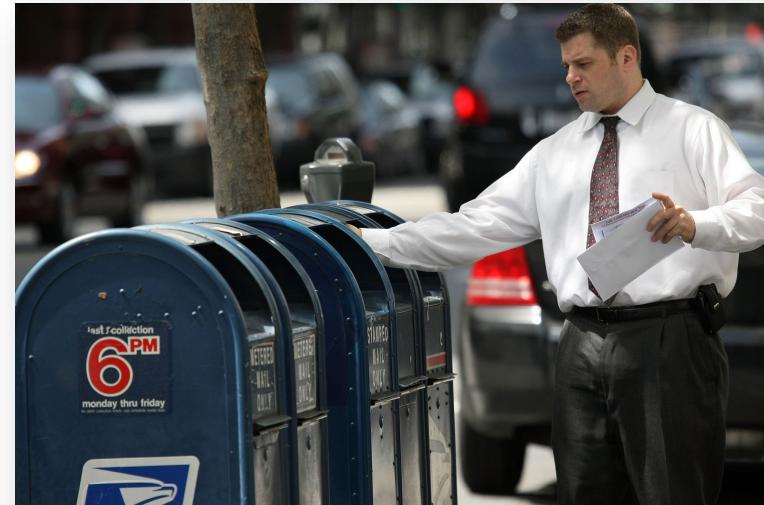
- Pros

- Cons



# Increasing Response Rates for Mail Surveys

- Cover letter
- Money helps
- Interesting questions
- Follow-ups
- Advance notification





# E-Mail/Internet Surveys

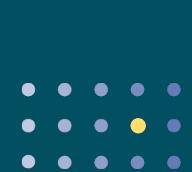
- Advantages
- Disadvantage



# Selecting the Appropriate Survey Approach

- Questions to be answered:
  - *Is the assistance of an interviewer necessary?*
  - *Are respondents interested in the issues being investigated?*
  - *Will cooperation be easily attained?*
  - *How quickly is the information needed?*
  - *Will the study require a long and complex questionnaire?*
  - *How large is the budget?*





# Pretesting Survey Instruments

- Pretesting
  - *Screening procedure that involves a trial run with a group of respondents to iron out fundamental problems in the survey design.*
- Basic Ways to Pretest:
  - *Screen the questionnaire with other research professionals.*
  - *Have the client or the research manager review the finalized questionnaire.*
  - *Collect data from a small number of respondents.*



Zikmund

Babin

Carr

Griffin

ninth edition

# Questionnaire Design





## Learning Outcomes

1. Explain the significance of decisions about questionnaire design and wording
2. Define alternatives for wording open-ended and fixed-alternative questions
3. Summarize guidelines for questions that avoid mistakes in questionnaire design
4. Describe how the proper sequence of questions may improve a questionnaire
5. Discuss how to design a questionnaire layout
6. Describe criteria for pretesting and revising a questionnaire and for adapting it to global markets



# J.D. Power Asks: It's Interesting, But Do You Really Want It?



- Car makers have to evaluate what features consumers want but will actually buy.
- J.D. Power surveyed consumers on what features they were familiar with and if they were willing to pay the price.
- Knowing price information changed consumers' interest levels.





# Questionnaire Quality and Design: Basic Considerations

- Questionnaire design is one of the most critical stages in the survey research process.
  - *A questionnaire (survey) is only as good as the questions it asks—ask a bad question, get bad results.*
  - *Composing a good questionnaire appears easy, but it is usually the result of long, painstaking work.*
  - *The questions must meet the basic criteria of relevance and accuracy.*





# Decisions in Questionnaire Design

1. What should be asked?
2. How should questions be phrased?
3. In what sequence should the questions be arranged?
4. What questionnaire layout will best serve the research objectives?
5. How should the questionnaire be pretested? Does the questionnaire need to be revised?



# What Should Be Asked?

- Questionnaire Relevancy
  - *All information collected should address a research question in helping the decision maker in solving the current business problem.*
- Questionnaire Accuracy
  - *Increasing the reliability and validity of respondent information requires that:*
    - ▶ Questionnaires should use simple, understandable, unbiased, unambiguous, and nonirritating words.
    - ▶ Questionnaire design should facilitate recall and motivate respondents to cooperate.
    - ▶ Proper question wording and sequencing to avoid confusion and biased answers.





# Wording Questions

- Open-ended Response Questions
  - *Pose some problem and ask respondents to answer in their own words.*
  - *Advantages:*
    - Are most beneficial in exploratory research, especially when the range of responses is not known.
    - May reveal unanticipated reactions toward the product.
    - Are good first questions because they allow respondents to warm up to the questioning process.
  - *Disadvantages:*
    - High cost of administering open-ended response questions.
    - The possibility that interviewer bias will influence the answer.
    - Bias introduced by articulate individuals' longer answers.



# Corporate Reputations: Consumers Put Johnson & Johnson, Microsoft, and Google on Top

- The Harris Reputation Quotient (RQ) has been used to assess the reputations of the 60 most visible companies in the U.S.
- Two stages
  - *Stage one: open-ended questions to determine the 60 companies*
  - *Stage two: fixed-alternative questions to evaluate each of the 60 companies*
- J&J has been top-ranked for the first seven years of the survey.

<https://theharrispoll.com/axios-harrispoll-100/>





# Wording Questions (cont'd)

- Fixed-alternative Questions
  - *Questions in which respondents are given specific, limited-alternative responses and asked to choose the one closest to their own viewpoint.*
  - *Advantages:*
    - ▶ Require less interviewer skill
    - ▶ Take less time to answer
    - ▶ Are easier for the respondent to answer
    - ▶ Provides comparability of answers
  - *Disadvantages:*
    - ▶ Lack of range in the response alternatives
    - ▶ Tendency of respondents to choose convenient alternative



# Types of Fixed-Alternative Questions

- Simple-dichotomy (dichotomous) Question
  - *Requires the respondent to choose one of two alternatives (e.g., yes or no).*
- Determinant-choice Question
  - *Requires the respondent to choose one response from among multiple alternatives (e.g., A, B, or C).*
- Frequency-determination Question
  - *Asks for an answer about general frequency of occurrence (e.g., often, occasionally, or never).*
- Checklist Question
  - *Allows the respondent to provide multiple answers to a single question by checking off items.*



# Cautions about Dichotomous or multiple-choice alternatives

- Totally exhaustive
- Mutually exclusive

Examples:

\$10,000-\$30,000

\$30,000-\$50,000

\$50,000-70,000

\$70,000-90,000

\$90,000-110,000

Over \$110,000

VS

*Less than \$10,000*

\$10,000-\$29,999

\$30,000-\$49,999

\$50,000-\$69,000

\$70,000-\$89,000

\$90,000-\$109,999

*Over \$110,000*





# Phrasing Questions for Self-Administered, Telephone, and Personal Interview Surveys

- Influences on Question Phrasing:
  - *The means of data collection—telephone interview, personal interview, self-administered questionnaire—will influence the question format and question phrasing.*
    - Questions for mail, Internet, and telephone surveys must be less complex than those used in personal interviews.
    - Questionnaires for telephone and personal interviews should be written in a conversational style.



# Guidelines for Constructing Questions

- Avoid complexity: Use simple, conversational language.
  - *Avoid words such as “Brand image,” “positioning,” “marginal analysis”*
- Avoid leading and loaded questions.
  - *Do accounting graduates who attended state university, such Washington State University, make better auditors?*
  - *What most influences your vote in major elections?*
- Avoid ambiguity: Be as specific as possible.
  - *What media do you rely on most?*
- Avoid double-barreled items.
  - *Do you feel our hospital emergency room waiting area is clean and comfortable?*



# Guidelines for Constructing Questions

- Avoid making assumptions.
  - *Should General Electric continue to pay its outstanding quarterly dividends?*
- Avoid burdensome questions that may tax the respondent's memory.
- Make certain questions generate variance
  - *Did you have any overnight travel for work-related activities last month?*
  - *Yes-no vs specific numbers*





# What Citizens Think About Climate Change



- A survey of 5,000 Australians found that there is considerable confusion about the underlying causes of climate change.



# What Is the Best Question Sequence?

- *Order bias*
  - ▷ Bias caused by the influence of earlier questions in a questionnaire or by an answer's position in a set of answers.
    - ◆ Election: first name listed tends to receive more votes
- *Funnel technique*
  - ▷ Asking general questions before specific questions in order to obtain unbiased responses.
    - ◆ Severity of the issue as an environmental problem: air pollution from automobile exhausts, air pollution from open burning... air pollution from industry
- *Filter question*
  - ▷ A question that screens out respondents who are not qualified to answer a second question.
- *Pivot question*
  - ▷ A filter question used to determine which version of a second question will be asked.
    - ◆ Do you prefer public transportation or driving your own car to commute to work?





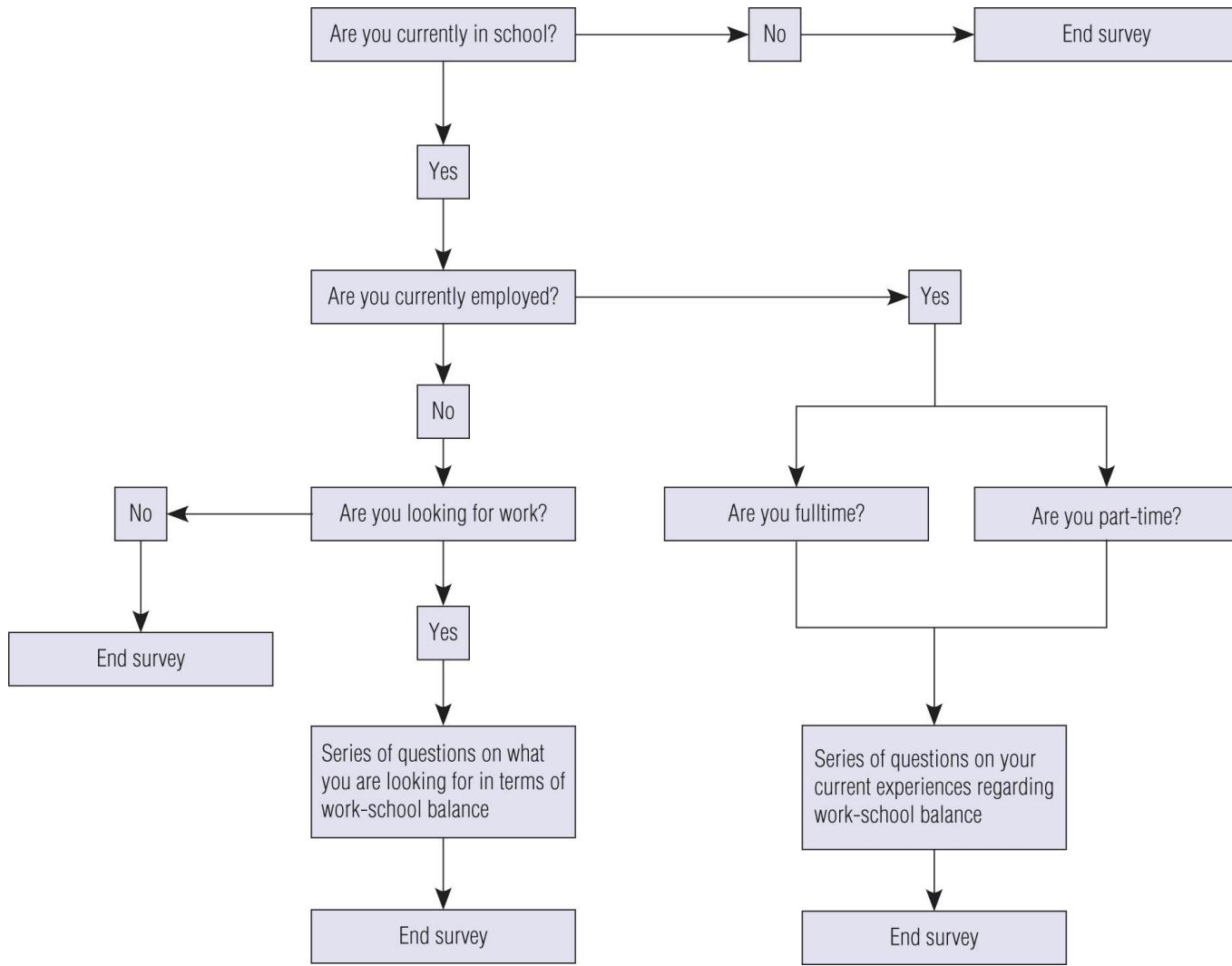
# What Is the Best Layout?

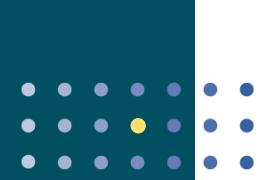
- Traditional Questionnaires
  - *Multiple-grid question*
    - Several similar questions arranged in a grid format.
  - *The title of a questionnaire should be phrased carefully:*
    - To capture the respondent's interest, underline the importance of the research
    - Emphasize the interesting nature of the study
    - Appeal to the respondent's ego
    - Emphasize the confidential nature of the study
    - To not bias the respondent in the same way that a leading question might



EXHIBIT 15.2

## Flow of Questions to Determine the Role of Work-school Balance for University Students





# Internet Questionnaires

- Survey software
  - *Surveymonkey.com*
  - *Doodle.com*
- Layout Issues
  - *Paging layout - going from screen to screen.*
  - *Scrolling layout – entire questionnaire appears on one page and respondent has the ability to scroll down.*





# Pretesting and Revising Questionnaires

- Pretesting Process
  - *Seeks to determine whether respondents have any difficulty understanding the questionnaire and whether there are any ambiguous or biased questions.*
- Preliminary Tabulation
  - *A tabulation of the results of a pretest to help determine whether the questionnaire will meet the objectives of the research.*



# Pretesting the CAHPS Hospital Survey

- The federal government makes hospital performance information available to all.
- The Consumer Assessment of Health Providers and Systems (CAHPS) Hospital Survey underwent several levels of pretesting.
  - *First version of 68 questions given to 18 individuals*
  - *Survey modified and tested on 13 more people*
  - *A draft survey with 66 items tested with almost 50,000 patients*
  - *Questionnaire reduced to 32 items*
  - *Further tested at several hospitals*
  - *Final version consists of 27 items.*



Zikmund

Babin

Carr

Griffin

ninth edition

# Sampling Designs and Sampling Procedures



# Changing Pocketbook Problems for Today's Families

- Public perceptions of financial concerns of the family.
- Each quarter, the Gallup Corporation develops a representative sample of approximately 1,000 U.S. adults.
- The most important problem facing families can often change over time.
  - *October 2007: healthcare costs (19%)*
  - *July 2008: gas prices (29%)*
  - *July 2011: low wages (17%)*
  - *May 2019: Cost of Healthcare (17%)*
  - *April 2021: Lack of money/Low wages*





# Sampling Terminology

- Population (universe)
  - *Any complete group of entities that a researcher is interested in.*
- Census
  - *An investigation of all the individual elements that make up a population.*
- Sample
  - *A subset, or some part, of a larger population.*



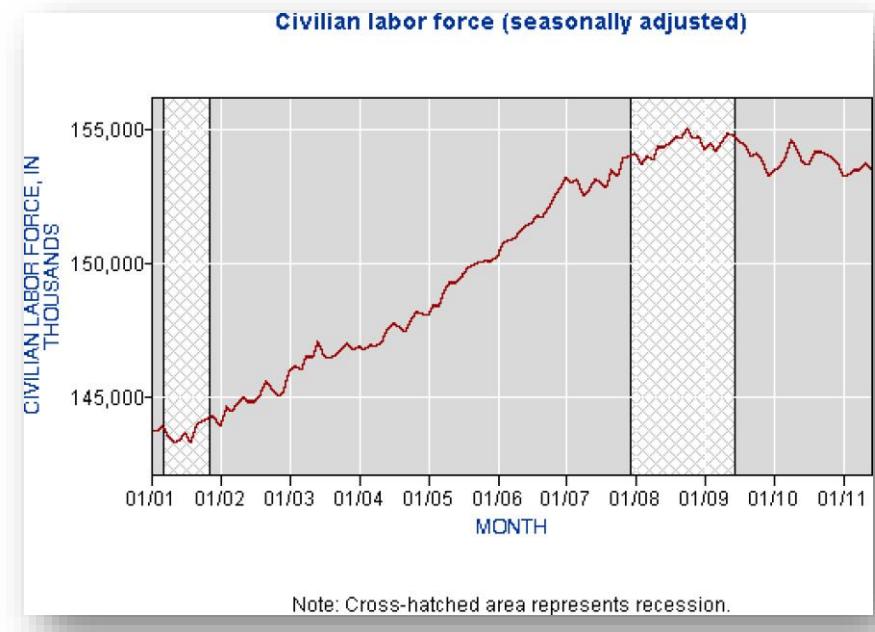
# Why Sample?

- Pragmatic Reasons
  - *Budget and time constraints.*
  - *Limited access to total population.*
- Accurate and Reliable Results
  - *Samples can yield reasonably accurate information if there is strong similarities in between population and sample—good sampling*
  - *Sampling may be more accurate than a census.*
- Destruction of Test Units
  - *Sampling reduces the costs of research in finite populations.*
  - *Manufacturer of firecrackers*



# Finding Out about Work Is a Lot of Work!

- The U.S. Census Bureau and the Bureau of Labor Statistics conduct the Current Population Survey (CPS).
- Uses a scientifically derived panel sample of 60,000 households.
- Surveyed each month
- Sophisticated and detailed.



# Practical Sampling Concepts

- Defining the Target Population
  - *What is the relevant population?*
  - *Whom do we want to talk to?*
    - Baby food manufacturer
    - Comic book
- The Sampling Frame
  - *A list of elements from which a sample may be drawn; also called working population.*
- Sampling Frame Error
  - Occurs when certain sample elements are not listed or are not accurately represented in a sampling frame.



# EXHIBIT 16.2 Mailing List Directory Page

## ***Lists Available - Alphabetical***

S.I.C. Code	List Title	United States		Canadian	S.I.C. Code	List Title	United States		Canadian					
		Total Count	State Count Page	Count			Total Count	State Count Page	Count					
<b>A</b>														
5122-02	Abdominal Supports .....	201	‡	28	7313-03	Advertising-Radio .....	2866	59	247					
8399-03	Abortion Alternatives Organizations` .....	946	‡	*	7311-07	Advertising-Shoppers' Guides .....	392	‡	4					
8093-04	Abortion Information & Services .....	551	‡	*	5199-17	Advertising-Specialties .....	12827	52	1648					
5085-23	Abrasives .....	1811	‡	277	7389-12	Advertising-Telephone .....	120	‡	*					
5169-04	Absorbents .....	145	‡	*	7313-05	Advertising-Television .....	1746	‡	102					
6541-03	Abstracters .....	4057	58	*	7319-02	Advertising-Transit & Transportation .....	179	‡	38					
6411-06	Accident & Health Insurance .....	2113	‡	9	0721-03	Aerial Applicators (Service) .....	1479	‡	61					
8748-52	Accident Reconstruction Service .....	125	‡	*	3999-01	Aerosols .....	158	‡	*					
8721-01	Accountants .....	127392	64	6933	3812-01	Aerospace Industries .....	426	‡	*					
8721-02	Accounting & Bookkeeping General Svc .....	27996	64	2072	5191-04	Agricultural Chemicals .....	549	‡	210					
5044-08	Accounting & Bookkeeping Machines/Suppls .....	889	‡	50	8748-20	Agricultural Consultants .....	1047	‡	474					
5044-01	Accounting & Bookkeeping Systems .....	624	‡	1230	9999-32	Air Balancing .....	353	‡	*					
8711-02	Acoustical Consultants .....	381	‡	91	5084-64	Air Brushes .....	219	‡	*					
1742-02	Acoustical Contractors .....	3063	47	433	4512-02	Air Cargo Service .....	6005	48	*					
1742-01	Acoustical Materials .....	878	‡	210	5075-01	Air Cleaning & Purifying Equipment .....	2055	‡	342					
8999-10	Actuaries .....	1185	‡	*	5084-02	Air Compressors .....	4358	50	717					
8049-13	Acupuncture (Acupuncturists) .....	2921	62	493	1711-17	Air Conditioning Contractors & Systems .....	50951	47	2667					
5044-02	Adding & Calculating Machines/Supplies .....	5524	49	648	***Available By Brands Sold**									
5044-09	Addressing Machines & Supplies .....	345	‡	29	Airtemp (A) .....									
5169-12	Adhesives & Glues .....	1187	‡	4	Amana (B) .....									
3579-02	Adhesives & Gluing Equipment .....	170	‡	204	Arco Aire (2) .....									
6411-02	Adjusters .....	6164	57	8357	Armstrong/Magic Chef (C) .....									
6411-01	Adjusters-Public .....	161	‡	*	Arvin (4) .....									
8322-07	Adoption Agencies .....	1621	‡	32	Bryant (D) .....									
8059-03	Adult Care Facilities .....	596	‡	*	Carrier (E) .....									
8361-08	Adult Congregate Living Facilities .....	170	‡	*	Coleman (5) .....									
7319-03	Advertising-Aerial .....	337	‡	26	Comfortmaker/Singer (O) .....									
7311-01	Advertising-Agencies & Counselors .....	27753	59	2552	Day & Night (Z) .....									
7336-05	Advertising-Art Layout & Production Svc .....	457	‡	101	Fedders (H) .....									
7331-05	Advertising-Direct Mail .....	6347	59	540	Heli/Quaker (3) .....									
7311-03	Advertising-Directory & Guide .....	2465	‡	124	Janitrol (7) .....									
7319-01	Advertising-Displays .....	3441	59	571	Kero-Sun (W) .....									
7319-11	Advertising-Indoor .....	209	‡	63	Lennox (K) .....									
7311-05	Advertising-Motion Picture .....	143	‡	11	Luxaire (L) .....									
7311-06	Advertising-Newspaper .....	4274	59	404	Payne (M) .....									
7312-01	Advertising-Outdoor .....	3052	59	297	*									
7311-08	Advertising-Periodical .....	817	‡	78	*									



# Practical Sampling Concepts (cont'd)

- Sampling services (list brokers)
  - *Provide lists or databases of the names, addresses, phone numbers, and e-mail addresses of specific populations.*
  - *Subscriptions, credit card application, warranty card registrations*
  - *Reverse directory*
    - A directory that lists by city and street address or by phone number





# Random Sampling and Nonsampling Errors

- Random Sampling Error
  - *The difference between the sample result and the result of a census conducted using identical procedures.*
  - *A statistical fluctuation that occurs because of chance variations in the elements selected for a sample.*
- Systematic Sampling Error
  - *Systematic (nonsampling) error results from nonsampling factors, primarily the nature of a study's design and the correctness of execution.*
    - It is *not* due to chance fluctuation.





# Probability versus Nonprobability Sampling

- Probability Sampling
  - *A sampling technique in which every member of the population has a known, nonzero probability of selection.*
- Nonprobability Sampling
  - *A sampling technique in which units of the sample are selected on the basis of personal judgment or convenience.*



# Nonprobability Sampling

- Convenience Sampling
  - *Obtaining those people or units that are most conveniently available.*
    - Person-on-the-street interviews
    - Interception at shopping center
- Judgment (Purposive) Sampling
  - *An experienced individual selects the sample based on personal judgment.*
    - Consumer price index
    - Fashion manufacturer
- Quota Sampling
  - *Ensures that various subgroups of a population will be represented.*





# American Kennel Club Tries to Keep Pet Owners out of the Doghouse

- The American Kennel Club (AKC) used quota sampling in a Dog Ownership Study.
- Sample size: 1000
- Set quotas for age, sex, and geographic categories.
- Sampled 500 dog owners and 500 non-owners.



# Nonprobability Sampling (cont'd)

- Possible Sources Of Bias
  - *Respondents chosen because they were:*
    - ▷ Similar to interviewer
    - ▷ Easily found
    - ▷ Willing to be interviewed
    - ▷ Middle-class
  - Advantages of Quota Sampling
    - *Speed of data collection*
    - *Lower costs*
    - *Convenience*



# Probability Sampling

- Simple Random Sampling
  - *Assures each element in the population of an equal chance of being included in the sample.*
- Systematic Sampling
  - *A starting point is selected by a random process and then every nth number on the list is selected.*
- Stratified Sampling
  - *Population is divided into subgroups*
  - *Simple random sampling in each subgroups*
  - *Eg: sample physicians about a certain prescribed drug. MDs and ODs*





# Cluster Sampling

- Population divided into groups (clusters)
- A simple random sample of the groups is selected
- Within a group/cluster, a random sample of individuals are selected
- Sample employees and self-employed works for a downtown project.
  - Classify target population as business and government
  - Select a random sample of firms
  - Select a simple random sample of individual workers





# What is the Appropriate Sample Design?

- Degree of accuracy
- Resources
- Time
- Advanced knowledge of the population
- National versus local project

