Guided Notes and Practice Problems Part 2

QSIS 285: Business Statistics

Lisa Over

ABSTRACT

This packet includes guided notes, examples, and exercises for chapters 7 through 10 and chapters 12, 14, 15, and 17 from the course text "Business Statistics: Communicating with Numbers" Second Edition by Jaggia/Kelly.

Contents

1 Samples and Sampling Distributions

1.1 Objectives

- Explain common sample biases.
- Describe various sampling methods.
- Describe the sampling distribution of the sample mean (proportion).
- Explain the importance of the Central Limit Theorem.
- Determine if the sampling distribution of sample means (proportions) is normally distributed.
- Calculate, or infer, the expected value and standard error of the sampling distribution of sample means (proportions)
- Find probabilities for sample means (proportions) using properties of the sampling distribution of sample means (proportions).

1.2 Sampling

A **sampling frame** is a group of accessible subjects from which to draw a sample.

Example: Frequent shopper list - a store does not have a list of all of its regular shoppers but it does have a list of shoppers who signed up for the "frequent shopper" program.

Caution: When the sampling frame differs from the population, which it almost always does, you must consider and deal with the differences.

- Are the opinions of those who signed up as frequent shoppers different from the rest of the shoppers?
- What about customers who signed up for the frequent shopper program but who don't shop there anymore?

Basic Definitions and Concepts	
Definition	**Example**
A **population** consists of all items of interest in a statistical problem. A **sample** is a subset of the population. A valid sample contains information that is representative of the population.	Population – all American high school teachers Sample – 30 high school teachers who, as a group, represent the population
Representative –	A survey about home water treatment systems selects participants who own homes.
Random –	The selection of manufacturers for a survey about future expectations does not favor participants based on type of manufacturing, location, size, or sales revenue.
Biased –	A survey about an upcoming election selects participants who answer calls from unknown caller IDs.

1.2.1 Bias

Bias refers to the tendency of sample statistic to systematically over- or underestimate a population parameter.

Types of Bias	
Definition	**Example**
Selection bias refers to a… …, i.e., portions of the population are excluded from the sample.	An Internet poll asks respondents to enter the number of computers they have in their homes. People without a computer or with limited access to a computer are excluded from the survey.
Nonresponse bias respondents… …from non-respondents.	An email survey asked people how receptive they are to receiving email solicitations. People who don't like email solicitations will not respond.
Response bias (also called survey **bias**) is the tendency of a person to answer questions on a survey…	The leading nature of the following survey question may pressure respondents to answer 'no.' "Many people believe this playground is too small and in need of repair. Do you think the playground should be repaired and expanded even if that means imposing an entrance fee to the park?"

1.2.2   Sampling Methods

Bad Sampling Methods	
Definition	**Example**
Convenience sample –	A manager wants to know what people think about the selection of stores and restaurants in the mall and surveys people who frequent the mall. What do people who don't go to the mall think about the stores and restaurants?
Voluntary response sample –	A request that travelers who have used the local airport visit a survey site to report on their experiences is much more likely to hear from those who had long waits, cancelled flights, and lost luggage than from those who had flights that were on time and carefree. These samples tend to be biased toward those with strong opinions or who are strongly motivated – especially from those with negative opinions.

Valid Sampling Methods		
Definition	**Example**	**Visual**
Simple Random Sample (SRS) Randomly select *n* observations from the population (sampling frame) such that the resulting group had the same probability of being selected from the population as any other sample of *n* observations.	A population contains 10 members under the age of 25 and 20 members over the age of 25. The sample will include six people chosen at random, without regard to age.	
Stratified Random Sampling 1. Define mutually exclusive and collectively exhaustive groups, called strata, based on members' shared attributes or characteristics. 2. Divide the population (sampling frame) into the strata. 3. Randomly select observations from each stratum that are proportional to the stratum's size. 4. Combine the selected outcomes into one stratified sample.	A population contains 10 members under the age of 25 and 20 members over the age of 25. The sample will include two people chosen at random under the age of 25 and four people chosen at random over 25. Bias can occur if information from the sample overemphasizes a particular stratum of the population.	
Cluster Sampling 1. Define mutually exclusive and collectively exhaustive groups, called clusters, based on members' shared attributes or characteristics. 2. Divide the population (sampling frame) into the clusters. 3. Randomly select one or more clusters and collect all observations from each selected cluster. 4. Combine the selected outcomes into one cluster sample.	A population contains 5,000,000 members divided approximately equally among 5 Northeastern states. The sample will include all members from one or two randomly selected states.	
Systematic Sample Shuffle the database to make sure certain groups are not excluded and take every nth subject.	A sample consists of every 10th employee from a population of 500 employees.	

1.3 Sampling Distributions

A sampling distribution demo is in the Excel file:
Sampling Distribution Demo.xlsx

A central limit theorem example is in the Excel file:
Penny Sampling.xlsx

Penny population…
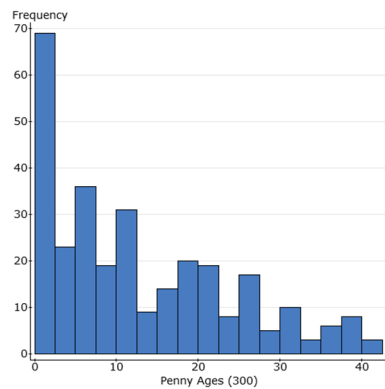
Summary statistics:

	Mean	Variance	Std. dev.	Median	Range	Skewness
Penny Ages (300)	12.71	120.1063	10.9593	10	41	0.791928
Sample Averages n=5	12.70	22.8884	4.7842	12.4	27.2	0.349977
Sample Averages n=15	12.68	7.5891	2.75482	12.6667	17.13	0.177568
Sample Averages n=30	12.70	3.6023	1.898	12.6	12.3	0.132682

Sampling Distributions…

14
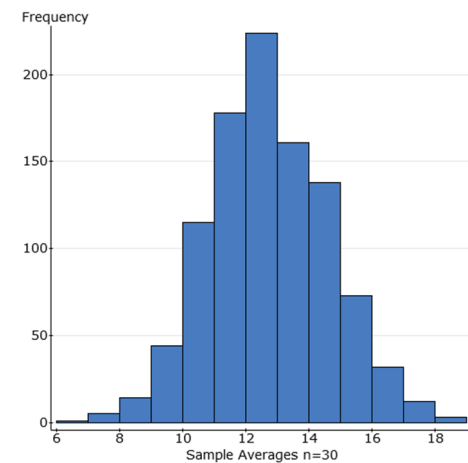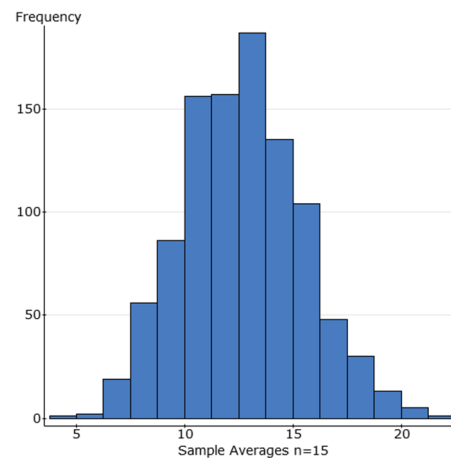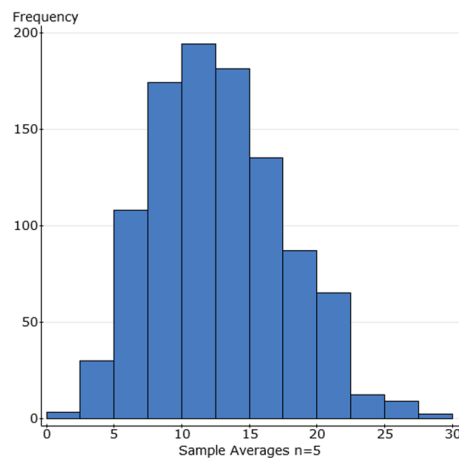
1.4 Sampling Distribution Notation, Definitions, and Formulas

	Sampling Distribution of Sample Proportions	Sampling Distribution of Sample Means
Notation		
Central Limit Theorem	The **Central Limit Theorem** isn't about the distribution of individual values from the sample. It is about the sample *proportions* or sample *means* of many different random samples drawn from the same population.	
	For any population proportion p, the sampling distribution of $\bar{P}$ is **approximately normal if the sample size n is sufficiently large**.	For any population mean μ, the sampling distribution of $\bar{X}$ is **approximately normal if the sample size n is sufficiently large**.*
Criteria for Assuming Normality		*Note:* When the population is known to be normally distributed, the sampling distribution of sample means is normally distributed for any size n.
Expected Value		
Standard Error		
Z, standard normal value (means when σ is known)		

1.5 Individual Data Values Vs. Sample Means and Proportions

Standard Normal Probabilities with Individual Data Values
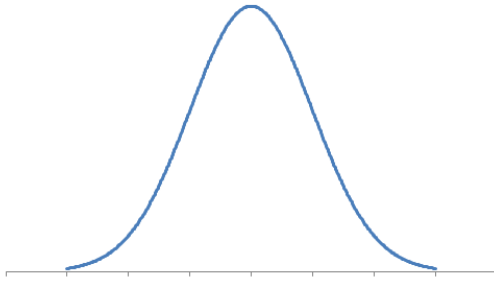
Standard Normal Probabilities with Sample Means	Standard Normal Probabilities with Sample Proportions
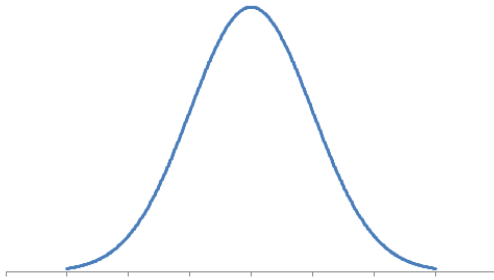
16
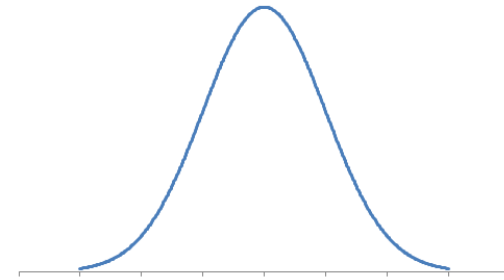
1.6 Sampling Distribution Example Problem

Suppose that, on average, beginning teachers earn approximately $33,800 a year in the United States. The distribution of starting salaries is a normally distributed random variable with standard deviation $3,500.

a. What is the probability that a randomly selected first year teacher earns more than $35,000? (**Round "z" value to 2 decimal places, and final answer to 4 decimal places.**)
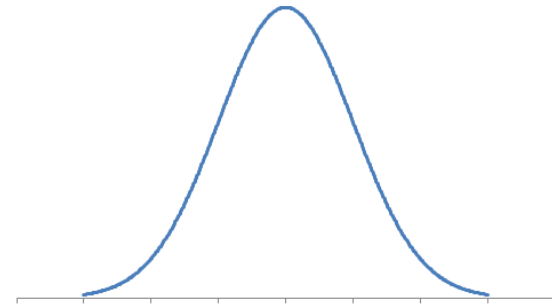
b. What is the probability that the average salary of four randomly selected first-year teachers is more than $35,000? (**Round "z" value to 2 decimal places, and final answer to 4 decimal places.**)
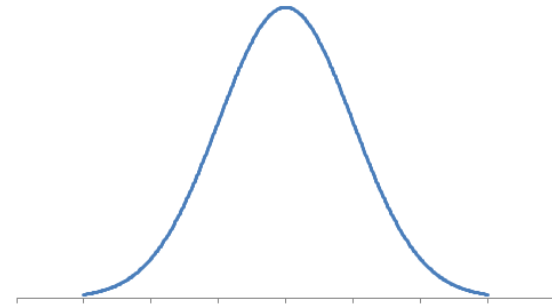
c. If four first-year teachers are randomly selected, what is the probability that all of the teachers earn more than $35,000?(**Round "z" value to 2 decimal places, and final answer to 4 decimal places.**)

3. 0.0179
2. 0.2465
1. 0.3658

1.7 Sampling Distribution Practice Problems (Means)

The following exercises are also in the Excel file:
Sampling Distribution Practice Problems.xlsx

Answers to the following exercises are in the Excel file:
Sampling Distribution Practice Problems KEY.xlsx

1.7.1 Exercise 1: Elementary School Absences

Over the entire six years that students attend an Ohio elementary school, they are absent, on average, 27 days due to influenza. Assume that the standard deviation over this time period is σ = 8 days. Upon graduation from elementary school, a random sample of 34 students is taken and asked how many days of school they missed due to influenza.

a. What is the expected value for the sampling distribution of the number of school days missed due to influenza?

b. What is the standard deviation (standard error) for the sampling distribution of the number of school days missed due to influenza?

c. The probability that the sample mean is less than 30 school days is _____.

d. The probability that the sample mean is between 24 and 30 school days is _____.

1.7.2 Exercise 2: Electricians' Salary

Suppose that, on average, electricians earn approximately μ = $58,000 per year in the United States. Assume that the distribution for electricians' yearly earnings is normally distributed and that the standard deviation is σ = $14,000.

a. Given a sample of four electricians, what is the standard deviation (standard error) for the sampling distribution of the sample mean?

b. What is the probability that the average salary of four randomly selected electricians exceeds $60,000?

c. What is the probability that the average salary of four randomly selected electricians is less than $50,000?

d. What is the probability that the average salary of four randomly selected electricians is more than $50,000 but less than $60,000?

1.7.3 Exercise 3: Bowling Score

Susan has been on a bowling team for 14 years. After examining all of her scores over that period of time, she finds that they follow a normal distribution. Her average score is 220, with a standard deviation of 10.

    a. What is the probability that in a one-game playoff, her score is more than 225?

    b. If during a typical week Susan bowls 16 games, what is the probability that her average score is more than 225?

    c. If during a typical week Susan bowls 16 games, what is the probability that her average score for the week is between 216 and 224?

    d. If during a typical month Susan bowls 64 games, what is the probability that her average score in this month is above 224?

1.7.4 Exercise 4: Professor Elderman

Professor Elderman has given the same multiple-choice final exam in his Principles of Microeconomics class for many years. After examining his records from the past 10 years, he finds that the scores have a mean of 75 and a standard deviation of 8.

    a. What is the probability that a class of 15 students will have a class average greater than 70 on Professor Elderman's final exam?

    b. What is the probability that a class of 36 students will have an average greater than 70 on Professor Elderman's final exam?

    c. Professor Elderman offers his class of 36 a pizza party if the class average is above 80. What is the probability that he will have to deliver on his promise? (i.e., What is the probability the class will earn a pizza party?)

    d. What is the probability Professor Elderman's class of 36 has a class average below 77?

4. Cannot determine, 0.9999, 0.0001, 0.9332
3. 0.3085, 0.0228, 0.8904, 0.0007
2. $7,000, 0.3859, 0.1271, 0.4870
1. 27 days, 1.37 days, 0.9857, 0.9714

1.8 Sampling Distribution Practice Problems (Proportions)

The following exercises are also in the Excel file:
Sampling Distribution Practice Problems.xlsx

Answers to the following exercises are in the Excel file:
Sampling Distribution Practice Problems KEY.xlsx

1.8.1 Exercise 1: University Administrator

A university administrator expects that 20% of students in a core course will receive an A. He looks at the grades assigned to 66 students.

a. What are the expected value and the standard error for the proportion of students that receive an A?

b. The probability that the proportion of students that receive an A is 0.15 or less is _____.

c. The probability that the proportion of students who receive an A is between 0.15 and 0.30 is _____.

d. The probability that the proportion of students who receive an A is *not* between 0.15 and 0.25 is _____.

1.8.2 Exercise 2: Labor Force

The labor force participation rate is the number of people in the labor force divided by the number of people in the country who are of working age and not institutionalized. The BLS reported in February 2012 that the labor force participation rate in the United States was 63.7% (Calculatedrisk.com). A marketing company asks 110 working-age people if they either have a job or are looking for a job, or, in other words, whether they are in the labor force.

    a. What are the expected value and the standard error for a labor participation rate in the company's sample?

    b. For the company's sample, the probability that the proportion of people who are in the labor force is greater than 0.67 is _____.

    c. What is the probability that fewer than 60% of those surveyed are members of the labor force?

    d. What is the probability that between 58% and 62.5% of those surveyed are members of the labor force?

1.8.3   Exercise 3: Super Bowl XLVI

Super Bowl XLVI was played between the New York Giants and the New England Patriots in Indianapolis. Due to a decade-long rivalry between the Patriots and the city's own team, the Colts, most Indianapolis residents were rooting heartily for the Giants. Suppose that 95% of Indianapolis residents wanted the Giants to beat the Patriots.

a. What is the probability that, of a sample of 100 Indianapolis residents, at least 12% were rooting for the Patriots in Super Bowl XLVI?

b. What is the probability that from a sample of 100 Indianapolis residents, fewer than 90% were rooting for the Giants in Super Bowl XLVI?

c. What is the probability that from a sample of 40 Indianapolis residents, fewer than 90% were rooting for the Giants in Super Bowl XLIV?

d. What is the probability that from a sample of 200 Indianapolis residents, fewer than 185 were rooting for the Giants in Super Bowl XLIV?

1.8.4 Exercise 4: Conservative States

According to the 2011 Gallup daily tracking polls (www.gallup.com, February 3, 2012), Mississippi is the most conservative U.S. state, with 53.4 percent of its residents identifying themselves as conservative.

a. What is the probability that at least 58% of a random sample of 200 Mississippi residents identify themselves as conservative?

b. What is the probability that at least 90 but fewer than 110 respondents of a random sample of 200 Mississippi residents identify as conservative?

c. What is the probability that at least 60 respondents of a random sample of 100 Mississippi residents do *not* identify themselves as conservative?

d. What is the probability that fewer than 50 respondents of a random sample of 100 Mississippi residents do *not* identify themselves as conservative?

[This page intentionally left blank]

2 Interval Estimation

2.1 Objectives

- Explain a point and an interval estimator.
- Calculate a confidence interval for the population mean when the population standard deviation is unknown.
- Discuss features of the t distribution.
- Calculate a confidence interval for the population proportion.
- Describe factors that influence the width of a confidence interval.

2.2 Point Estimate and Confidence Interval

For any given sample, how confident are you that the sample you evaluate represents the true population value?

The sample mean is a **point estimate** and, by itself, is limited in estimating the population mean because a new sample of the same size would yield a different mean.

A **confidence interval** provides…

The most typical form of a calculated interval is…

The **margin of error**, ME, is a certain number of standard errors added to and subtracted from the point estimate. The number of standard errors is a critcal value based on the level of confidence.

The **probability of making an error**, α, is 1 minus the level of confidence.
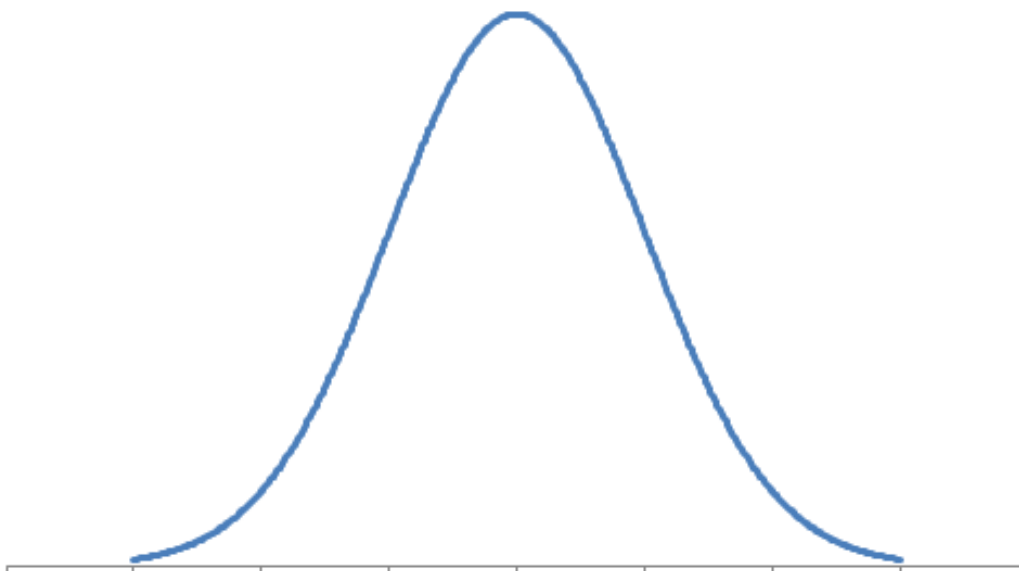
The following image illustrates 20 confidence intervals created from 20 sample means. The center of each interval is the sample mean. The black line is the population mean. It is apparent that not all sample means equal the population mean. But it is equally apparent that 19 out of the 20 confidence intervals (95%) contain the true population mean.
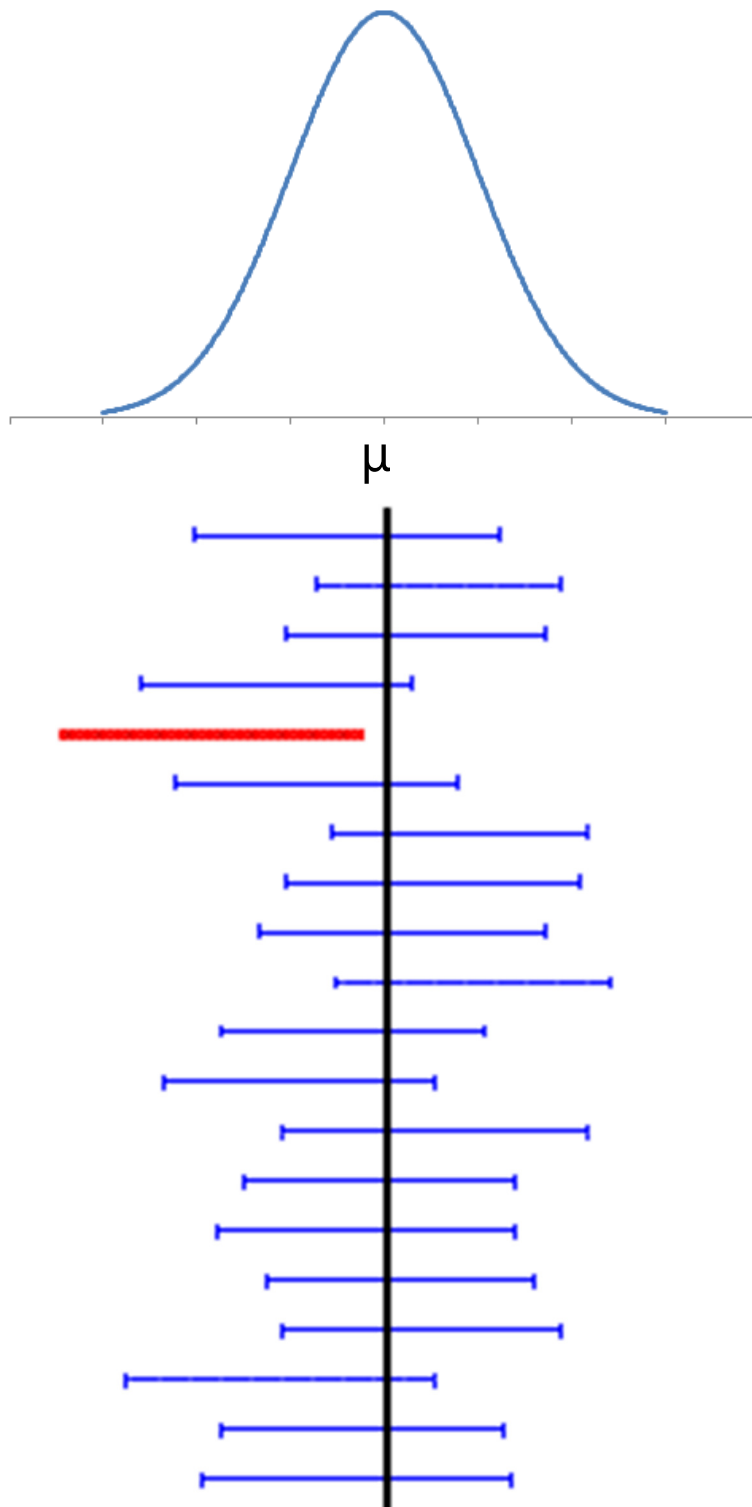
2.3 Conditions for Constructing Confidence Intervals

1. The sampled values must be…
2. The sampling process must be…
3. The sample size, n, must be no larger than…
4. The sampling distribution must be normally distributed
    o   For proportions, …

    o   For means, …

2.4 Point and Standard Error Estimates

Means		Proportions
Estimate the population mean, μ, with…	**Estimate the population mean, μ, with…**	**Estimate the population proportion, p, with…**
Standard Error Equation	**Standard Error Equation**	**Standard Error Equation**
When the population standard deviation is known, use sigma in the standard error equation.	When the population standard deviation is unknown, use the sample mean, s, in the standard error equation.	Use the sample proportion in the standard error formula.
Use a standard normal z value as the critical value	**Use a Student's t value as the critical value**	**Use a standard normal z value as the critical value**

2.5 Steps to Construct a Confidence Interval

1. Choose the level of confidence and the corresponding z* or t* critical values.
2. Compute the standard error of the sampling distribution.
3. Multiply the standard error by the critical value. This is called the **margin of error**.
4. Subtract the margin of error from the point estimate. This is the lower value of the estimate.
5. Add the margin of error to the point estimate. This is the upper value of the estimate.
6. Report the interval.

2.6 Construct a Confidence Interval for a Sample Mean, σ is known

The daily revenue from the sale of fried dough at a local street vendor in Boston is known to be normally distributed with a known standard deviation of $120. The revenue on each of the last 25 days is noted, and the average is computed as $550. Construct a 95% confidence interval for the population mean of the sale of fried dough by this vendor.

1.	Choose the level of confidence and the corresponding z* critical value.	
2.	Compute the standard error of the sampling distribution.	
3.	Multiply the standard error by the critical value. This is called the **margin of error**.	
4.	Subtract the margin of error from the point estimate. This is the lower value of the estimate.	
5.	Add the margin of error to the point estimate. This is the upper value of the estimate.	
6.	Report the interval.	

Critical Values for Common Confidence Levels

90% CI $z^*_{0.05} = 1.645$

95% CI $z^*_{0.025} = 1.96$

98% CI $z^*_{0.01} = 2.33$

99% CI $z^*_{0.005} = 2.58$

2.7 The Student's t

The sample standard deviation, s, is a biased estimator, which means it's value is different from the true value of the parameter being estimated. When s is used to estimate the population standard deviation, the sampling distribution is the Student's t distribution with (n-1) degrees of freedom. The Student's t…

- Is symmetric, unimodal, bell shaped
    - Resembles a <u>normally distributed</u> variable with mean 0 and variance 1
    - Has thicker tails, is wider and lower
    - Approaches the normal distribution with mean 0 and variance 1 as the number of degrees of freedom grows.
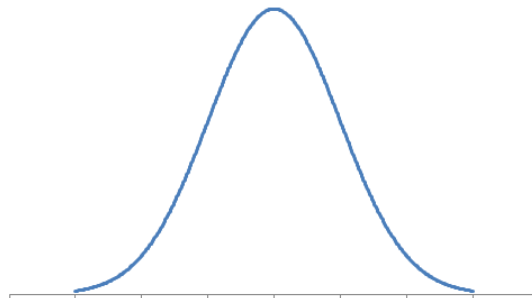
2.8 Construct a Confidence Interval for a Sample Mean, σ is unknown

At a particular academically challenging high school, the average GPA of a high school senior is known to be normally distributed. After a sample of 20 seniors is taken, the average GPA is found to be 2.8 and the variance is determined to be 0.25. Find a 90% confidence interval for the population mean GPA.

1.	Choose the level of confidence and the corresponding t* critical value.	
2.	Compute the standard error of the sampling distribution.	
3.	Multiply the standard error by the critical value. This is called the **margin of error**.	
4.	Subtract the margin of error from the point estimate. This is the lower value of the estimate.	
5.	Add the margin of error to the point estimate. This is the upper value of the estimate.	
6.	Report the interval.	

2.9 Construct a Confidence Interval for a Sample Proportion

A sample of 1,400 American households was asked if they planned to buy a new car next year. Of the respondents, 41% indicated they planned to buy a new car next year. Construct a 99% confidence interval of the proportion of American households who expect to buy a new car next year.

1.	Choose the level of confidence and the corresponding z* critical value.	
2.	Compute the standard error of the sampling distribution.	
3.	Multiply the standard error by the critical value. This is called the **margin of error**.	
4.	Subtract the margin of error from the point estimate. This is the lower value of the estimate.	
5.	Add the margin of error to the point estimate. This is the upper value of the estimate.	
6.	Report the interval.	

Critical Values for Common Confidence Levels

90% CI $z^*_{0.05} = 1.645$

95% CI $z^*_{0.025} = 1.96$

98% CI $z^*_{0.01} = 2.33$

99% CI $z^*_{0.005} = 2.58$

2.10 Interval Width, Confidence, and Sample Size

Statisticians like precision in their interval estimates, which means a narrow interval with a high level of confidence. A low margin of error is needed to reduce the width of the interval.

There are two ways to reduce the margin of error:

1.
2.

Which of the above supports a high level of confidence?

-

-

A larger sample size reduces the margin of error without reducing the level of confidence.

2.11 Interval Estimates Practice Problems (Sample Size)

2.11.1 Exercise 1: Larger Sample Size

An analyst takes a random sample of 25 firms in the telecommunications industry and constructs a confidence interval for the mean return for the prior year. Holding all else constant, if he increased the sample size to 30 firms, how are the standard error of the mean and the width of the confidence interval affected?

	Standard error of the mean	Width of confidence interval
A	Increases	Becomes wider
B	Increases	Becomes narrower
C	Decreases	Becomes wider
D	Decreases	Becomes narrower

2.11.2 Exercise 2: Greater Confidence

A 90% confidence interval is constructed for the population mean. If a 95% confidence interval had been constructed instead (everything else remaining the same), the width of the interval would have been _____ and the probability of making an error would have been _____.

2. Wider, smaller
1. Decreases, becomes narrower

2.12 Interval Estimates Practice Problems (Means)

The following exercises are also in the Excel file:
Interval Estimates Practice Problems.xlsx

Answers to the following exercises are in the Excel file:
Interval Estimates Practice Problems KEY.xlsx

2.12.1 Exercise 1: Students' T 95/15

What is the critical value for a 95% confidence interval of the population mean based on a sample of 15 observations?

2.12.2 Exercise 2: Students' T 99/25

What is the critical value for a 99% confidence interval of the population mean based on a sample of 25 observations?

2.12.3 Exercise 3: CI 95

Given a sample mean of 27 and a sample standard deviation of 3.5 computed from a sample of size 36, find a 95% confidence interval on the population mean.

2.12.4 Exercise 4 CI 99

Given a sample mean of 12.5—drawn from a normal population, a sample of size 25, and a sample variance of 2.4—find a 99% confidence interval for the population mean.

2.12.5 Exercise 5: HS GPA

At a particular academically challenging high school, the average GPA of a high school senior is known to be normally distributed. After a sample of 20 seniors is taken, the average GPA is found to be 2.8 and the variance is determined to be 0.25. Find a 90% confidence interval for the population mean GPA.

2.12.6 Exercise 6: Holiday Spending

In an examination of holiday spending (known to be normally distributed) of a sample of 16 holiday shoppers at a local mall, an average of $53 was spent per hour of shopping. Based on the current sample, the standard deviation is equal to $20. Find a 90% confidence interval for the population mean level of spending per hour.

2.12.7 Exercise 7: Accounting Professors 95/25

Professors of accountancy are in high demand at American universities. A random sample of 25 new accounting professors found the average salary was $135 thousand with a standard deviation of $16 thousand. Assume the distribution is normally distributed. Construct a 95% confidence interval for the salary of new accounting professors. Answers are in thousands of dollars.

2.12.8 Exercise 8: Accounting Professors 90/25

Professors of accountancy are in high demand at American universities. A random sample of 25 new accounting professors found the average salary was $135 thousand with a standard deviation of $16 thousand. Assume the distribution is normally distributed. Construct a 90% confidence interval for the salary of new accounting professors. Answers are in thousands of dollars.

2.12.9 Exercise 9: Accounting Professors 95/35

Professors of accountancy are in high demand at American universities. A random sample of 35 new accounting professors found the average salary was $135 thousand with a standard deviation of $16 thousand. Assume the distribution is normally distributed. Construct a 95% confidence interval for the salary of new accounting professors. Answers are in thousands of dollars.

2.12.10 Exercise 10: Compare Accounting Professor Salary Estimates

Point Estimate	Confidence	Sample Size	Lower CI Limit	Upper CI Limit

10. Smaller interval with larger sample or lower confidence
9. [$129.5, $140.5]
8. [$129.5, $140.5]
7. [$128.4, $141.6]
6. [$44.24, $61.77]
5. [2.61, 2.99]
4. [11.63, 13.37]
3. [25.82, 28.18]
2. 2.797
1. 2.145

2.13 Interval Estimates Practice Problems (Proportions)

2.13.1 Exercise 1: Parents Helping Adult Children

According to a report in *USAToday* (February 1, 2012), more and more parents are helping their young adult children get homes. Suppose eight persons in a random sample of 40 young adults who recently purchased a home in Kentucky received help from their parents. You have been asked to construct a 95% confidence interval for the population proportion of all young adults in Kentucky who received help from their parents. What is the margin of error for a 95% confidence interval for the population proportion?

2.13.2 Exercise 2: Opposing Candidates 95/300

Candidate A is facing two opposing candidates in a mayoral election. In a recent poll of 300 residents, she has garnered 53% support. Construct a 95% confidence interval on the population proportion for the support of candidate A in the following election.

2.13.3 Exercise 3: Opposing Candidates 90/300

Candidate A is facing two opposing candidates in a mayoral election. In a recent poll of 300 residents, 160 supported her. Construct a 90% confidence interval on the population proportion for the support of candidate A in the following election.

2.13.4 Exercise 4: Opposing Candidates 95/400

Candidate A is facing two opposing candidates in a mayoral election. In a recent poll of 400 residents, she has garnered 53% support. Construct a 95% confidence interval on the population proportion for the support of candidate A in the following election.

2.13.5 Exercise 5: Compare Opposing Candidates Support Estimates

Point Estimate	Confidence	Sample Size	Lower CI Limit	Upper CI Limit

2.13.6 Exercise 6: Americans and China

A sample of 2,007 American adults was asked how they viewed China, with 17% of respondents calling the country "unfriendly" and 6% of respondents indicating the country was "an enemy". Construct a 95% confidence interval of the proportion of American adults who viewed China as "an enemy."

2.13.7 Exercise 7: Opposing Candidates 98/300

Candidate A is facing two opposing candidates in a mayoral election. In a recent poll of 300 residents, 102 supported candidate B and 54 supported candidate C. Construct a 98% confidence interval on the population proportion for the support of candidate A in the following election.

2.13.8 Exercise 8: Buy a New Car

A sample of 1,400 American households was asked if they planned to buy a new car next year. Of the respondents, 41% indicated they planned to buy a new car next year. Construct a 99% confidence interval of the proportion of American households who expect to buy a new car next year.

8. [0.376, 0.444]
7. [0.413, 0.547]
6. [0.0496, 0.1096]
5. Smaller interval with larger sample or lower confidence
4. [0.481, 0.579]
3. [0.483, 0.577]
2. [0.474, 0.587]
1. 0.124

3 Hypothesis Testing

3.1 Objectives

- Define the null hypothesis and the alternative hypothesis.
- Distinguish between Type I and Type II errors.
- Calculate test statistics for means (population standard deviation unknown) and proportions.
- Conduct a hypothesis test using the p-value approach.
- Conduct a hypothesis test using the critical value approach.

3.2 Introduction to Hypothesis Testing

Hypothesis

Hypothesis Test

Rare Event Rule for Inferential Statistics

If, under a given assumption, the probability of a particular observed event is extremely small, we conclude that this assumption is probably not correct, i.e., we reject explanations when they are based on extremely small probabilities.

3.3 Hypothesis Testing Example Problem Part 1

3.3.1 *Informal* Hypothesis Test

A small hair salon in Denver, Colorado, averages about 80 customers on weekdays with a standard deviation of 6. It is safe to assume that the underlying distribution is normal. In an attempt to increase the number of weekday customers, the manager offers a $5 discount on 5 consecutive weekdays. She reports that her strategy has worked since the sample mean of customers during this 5-weekday period jumps to 88.

a. What is the probability of getting a sample average of 88 or more customers if the manager had not offered the discount? *Answer: 0.0014*

b. Do you feel confident that the manager's discount strategy has worked?
    Answer: Yes, there is only a small chance (less than 5%) of getting 88 or more customers without the discount.

Assumption: The assumption is the status quo, i.e., **the average (80) and standard deviation (6) before the discount.**
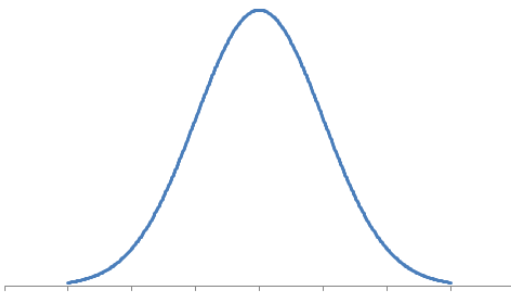
Under the assumption that the hair salon serves an average of 80 customers during a 5-weekday period with a standard deviation of 6, the probability of seeing 88 customers is _____

_____.

We _____ the status quo. The data _____ that the average number of customers is greater than 80, and we conclude that _____

_____.

3.4 Components of a Formal Hypothesis Test

1. Null and Alternative Hypotheses
2. Significance Level and Critical Value(s)
3. Test Statistic and P-value
4. Conclusion

3.4.1 **Null and Alternative Hypotheses**

Null Hypothesis

a) The null hypothesis is denoted H_0.

b) The null hypothesis is a statement that…

c) By definition, this statement contains a statement of equality:

d) We make a decision to either reject H_0 or to fail to reject H_O. We never "accept" H_0.

e) The hypothesis test begins by assuming H_0 is true. The null hypothesis is what is tested.

Alternative Hypothesis

a) The alternative hypothesis is denoted H_A.

b) The alternative hypothesis is a statement that…

c) By definition, this statement is the complement of H_0. The two hypotheses are mutually exclusive.

3.4.2   Significance Level and Critical Values

Significance Level

The significance level, α, is the probability of rejecting the null hypothesis when it is true. For example, a significance level of 0.05 indicates a 5% chance of concluding that the status quo is not correct when the truth is that the status quo is correct.

Possibility of Error*

Decision	Null Hypothesis is True	Null Hypothesis is False
Reject the null hypothesis		
Fail to reject the null hypothesis		

*Reducing one type of error increases the other. The only way to reduce both types of error is to collect more evidence, i.e., increase the sample size, n.

A Type I error is committed when we reject the null hypothesis when it is actually true. A Type II error is made when we do not reject the null hypothesis when it is actually false.

The objective of hypothesis testing is to reject a null hypothesis when it is false and to fail to reject a null hypothesis when it is true.

Critical Value(s)

The critical value is a _____ or _____ score that corresponds to the significance level. The critical value also depends on the alternative hypothesis, i.e., whether the alternative hypothesis involves ≠, <, or >.
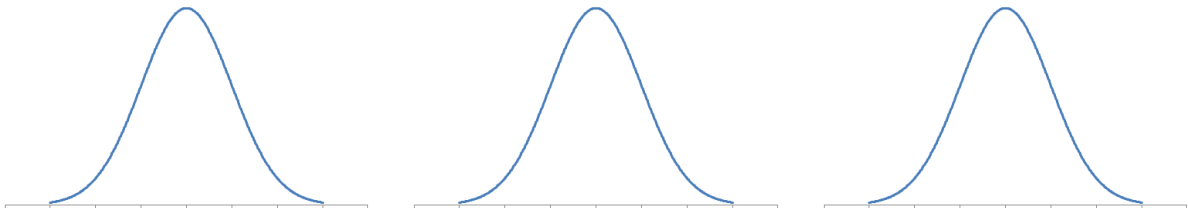
3.4.3   Test Statistic and P-value

Test Statistic

The test statistic is the standardized sample statistic, $\bar{x}\ or\ \bar{p}$, under the assumption that the null hypothesis is true, i.e., the sample statistic is standardized using the mean of the status quo.

P-value

The p-value is the probability of observing the given sample statistic or of observing one with a more extreme value.
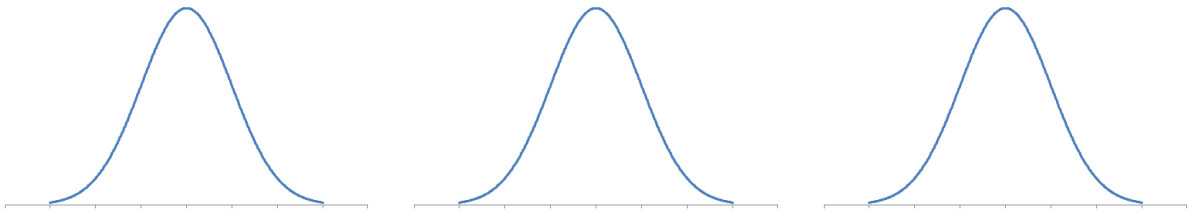
3.4.4   Conclusion

The conclusion states the results of the test, i.e., to reject or fail to reject the null, and includes an interpretation of the results in terms of the alternative hypothesis and in terms of the data.

3.5 Steps of a Hypothesis Test

1. Define the two hypotheses: Null and Alternative.

   Three Steps to Formulate Hypotheses

   a.

   b.

   c.

H_0	H_A	Test Type
		Two-tail
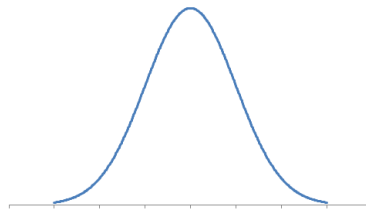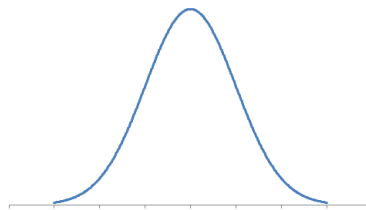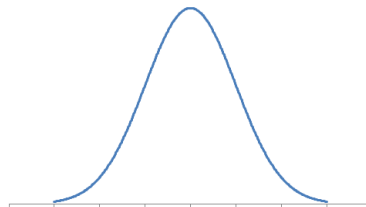
| | | One-tail, Left-tail |

| | | One-tail, Right-tail |

2. Specify the significance level, α, i.e., the probability of making a type I error. Find the critical values associated with α for a one-tailed test or $\frac{\alpha}{2}$ for a two-tailed test.

3. Calculate the value of a test statistic and the p-value.

Test statistic for a Proportion		p_0 is the "status quo" population proportion specified in the null hypothesis
		$\bar{p}$ is the sample proportion
Test statistic for a Mean		μ_0 is the "status quo" population mean specified in the null hypothesis
		$\bar{x}$ is the sample mean
Calculate the p-value	Find the probability of observing the test statistic, or a value more extreme, using Excel. This value is also available in the computer output after running a statistical test.	

4. State the conclusion and interpret the results.
   a. State the conclusion as "_____" or "_____."
   b. Interpret the results in terms of the data.

3.6 Hypothesis Testing Example Problem Part 2

3.6.1 *Formal* Hypothesis Test

A small hair salon in Denver, Colorado, averages about 80 customers on weekdays. It is safe to assume that the underlying distribution is normal. In an attempt to increase the number of weekday customers, the manager offers a $5 discount on 5 consecutive weekdays. She reports that her strategy has worked since the sample mean of customers during this 5 weekday period jumps to 88 with a sample standard deviation of 5.3. **Can the manager conclude that the average number of customers exceeds 80 in a 5 weekday period? Test the hypothesis at a 5% level of significance using the p-value approach and the critical value approach.**

1. Define the two hypotheses: Null and Alternative.

   Three Steps to Formulate Hypotheses

   a. Identify the relevant population parameter of interest.

   b. Determine whether it is a one- or a two-tailed test.

   c. Include some form of the equality sign in H_0 and use H_A to establish a claim.

H_0	H_A	Test Type
$=$	$\neq$	Two-tail
$\geq$	$<$	One-tail, Left-tail
$\leq$	$>$	One-tail, Right-tail

   Hypotheses…

2. Specify the significance level, α, i.e., the probability of making a type I error. Find the critical values associated with α for a one-tailed test or $\frac{\alpha}{2}$ for a two-tailed test.

Level of Significance and Critical Values…

3. Calculate the value of a test statistic and the p-value.

Test statistic for a Proportion	$z = \dfrac{\bar{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$	p_0 is the "status quo" population proportion specified in the null hypothesis $\bar{p}$ is the sample proportion
Test statistic for a Mean σ *unknown*	$t_{df} = \dfrac{\bar{x} - \mu_0}{\dfrac{s}{\sqrt{n}}}$	μ_0 is the "status quo" population mean specified in the null hypothesis $\bar{x}$ is the sample mean
Calculate the p-value	Find the probability of observing the test statistic, or a value more extreme, using Excel. This value is also available in the computer output after running a statistical test.	

Test Statistic and P-value…
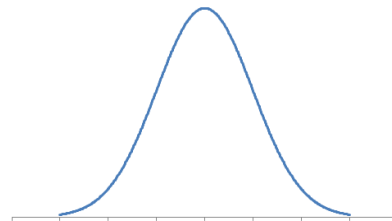
4. State the conclusion and interpret the results.
   a. State the conclusion as "Reject the null hypothesis" or "Fail to reject the null hypothesis."
   b. Interpret the results in terms of the data.

Conclusion…

3.7 Parameter of Interest Practice Problems

3.7.1 Exercise 1: Type of Data for Proportions

What type of data would necessitate using a hypothesis test of the population proportion rather than a test of the population mean?

    A. Ratio
    B. Interval
    C. Qualitative
    D. Quantitative

3.7.2 Exercise 2: Reading Comprehension Test

The national average for an eighth-grade reading comprehension test is 73. A school district claims that its eighth-graders outperform the national average. In testing the school district's claim, how does one define the population parameter of interest?

    A. The mean score on the eighth-grade reading comprehension test
    B. The number of eighth graders who took the reading comprehension test
    C. The standard deviation of the score on the eighth-grade reading comprehension test
    D. The proportion of eighth graders who scored above 73 on the reading comprehension test

3.7.3 Exercise 3: Local Courier Service

A local courier service advertises that its average delivery time is less than 6 hours for local deliveries. When testing the two hypotheses, $H_o{:}\mu \geq 6$ and $H_A{:}\mu < 6$, μ stands for _____.

    A. The standard deviation of the delivery time
    B. The number of deliveries that took less than 6 hours
    C. The proportion of deliveries that took less than 6 hours
    D. The mean delivery time

3.7.4 Exercise 4: Texas Babies

It is generally believed that no more than 0.50 of all babies in a town in Texas are born out of wedlock. A politician claims that the proportion of babies born out of wedlock is increasing. When testing the two hypotheses, $H_0{:}\ p \leq 0.50$ and $H_A{:}\ p > 0.50$, p stands for _____.

    A. The mean number of babies born out of wedlock
    B. The current proportion of babies born out of wedlock
    C. The number of babies born out of wedlock
    D. The general belief that the proportion of babies born out of wedlock is no more than 0.50

3.7.5 Exercise 5: Fast-Food Franchise

A fast-food franchise is considering building a restaurant at a busy intersection. A financial advisor determines that the site is acceptable only if, on average, more than 300 automobiles pass the location per hour. If the advisor tests the hypotheses H_o: $\mu \leq 300$ versus H_A: $\mu > 300$, μ stands for _____.

    A. The number of automobiles that pass the intersection per hour
    B. The proportion of automobiles that pass the intersection per hour
    C. The average number of automobiles that pass the intersection per hour
    D. The standard deviation of the number of automobiles that pass the intersection per hour

3.7.6 Exercise 6: University Honors Program GPA 3.5

A university is interested in promoting graduates of its honors program by establishing that the mean GPA of these graduates exceeds 3.50. A sample of 36 honors students is taken and is found to have a mean GPA equal to 3.60. The population standard deviation is assumed to equal 0.40. The parameter to be tested is _____.

    A. The mean GPA of all university students
    B. The mean GPA of 3.60 for the 36 selected honors students
    C. The proportion of honors students with a GPA exceeding 3.50
    D. The mean GPA of the university honors students

3.7.7 Exercise 7: University Honors Program GPA 3.0

A university interested in tracking its honors program believes that the proportion of graduates with a GPA of 3.00 or below is less than 0.20. In a sample of 200 graduates, 30 students have a GPA of 3.00 or below. In testing the university's belief, how does one define the population parameter of interest?

    A. It's the mean number of honors graduates with a GPA of 3.00 or below.
    B. It's the proportion of honors graduates with a GPA of 3.00 or below.
    C. It's the number of honors graduates with a GPA of 3.00 or below.
    D. It's the standard deviation of the number of honors graduates with a GPA of 3.00 or below.

3.7.8 Exercise 8: Car Dealership

The owner of a large car dealership believes that the financial crisis decreased the number of customers visiting her dealership. The dealership has historically had 800 customers per day. The owner takes a sample of 100 days and finds the average number of customers visiting the dealership per day was 750. Assume that the population standard deviation is 350. The population parameter to be tested is _____.

    A. The mean number of customers visiting the dealership per day
    B. The average number of 750 customers per day
    C. The proportion of customers visiting the dealership per day
    D. The standard deviation of the number of customers visiting the dealership per day

3.8 Define Hypotheses Practice Problems

3.8.1 Exercise 1: Two-tailed, Right-tailed, or Left-tailed

Expedia would like to test if the average round-trip airfare between Philadelphia and Dublin is less than $1,200. Which of the following hypothesis tests should be performed?

    A. Two-tailed
    B. Right-tailed
    C. Left-tailed

3.8.2 Exercise 2: Define Null and Alternative Hypotheses

Define the null and alternative hypotheses for the following tests:

    a. Test if the mean weight of cereal in a cereal box differs from 18 ounces.

    b. Test if the stock price increases on more than 60% of the trading days.

    c. Test if Americans get an average of less than seven hours of sleep.

Which of the following statements are valid null and alternative hypotheses?

    A. $H_0: \bar{x} \leq 210$; $H_A: \bar{x} > 210$

    B. $H_0: \mu = 120$; $H_A: \mu \neq 120$

    C. $H_0: p \leq 0.24$; $H_A: p > 0.24$

    D. $H_0: \mu < 252$; $H_A: \mu > 252$

Define the null and alternative hypotheses for the following claims.

    a. "I am going to get the majority of votes to win this election."

    b. "I suspect that your 10" pizzas are, on average, less than 10" in size."

3.8.3 Exercise 3: LED Streetlights

Many cities around the United States are installing LED streetlights, in part to combat crime by improving visibility after dusk. An urban police department claims that the proportion of crimes committed after dusk will fall below the current level of 0.84 if LED streetlights are installed. Define the null and alternative hypotheses to test the police department's claim.

3.8.4 Exercise 4: Expedia Round-trip Airfare

Expedia would like to test if the average round-trip airfare between Philadelphia and Dublin is less than \$1,200. The correct hypothesis statement would be _____.

3.8.5 Exercise 5: Texas Babies

It is generally believed that no more than 0.50 of all babies in a town in Texas are born out of wedlock. A politician claims that the proportion of babies born out of wedlock is increasing. Identify the correct null and alternative hypotheses to test the politician's claim.

3.8.6 Exercise 6: Short Sale homes

You want to test if more than 20% of homes in a neighborhood have recently been sold through a short sale, at a foreclosure auction, or by the bank following an unsuccessful foreclosure auction. You take a sample of 60 homes from this neighborhood and find that 14 fit your criteria. The appropriate null and alternative hypotheses are _____.

6. $H_0: p \leq 0.2; H_A: p > 0.2$
5. $H_0: p \leq 0.5; H_A: p > 0.5$
4. $H_0: \mu \geq \$1,200; H_A: \mu < \$1,200$
3. $H_0: p \geq 0.84; H_A: p < 0.84$
$H_0: \mu \geq 10; H_A: \mu < 10$
$H_0: p \leq 0.5; H_A: p > 0.5$
B, C
$H_0: \mu \geq 7; H_A: \mu < 7$
$H_0: p \leq 0.6; H_A: p > 0.6$
2. $H_0: \mu = 18; H_A: \mu \neq 18$
1. C

3.9 Level of Significance and Critical Values Practice Problems

3.9.1 Exercise 1: Probability of Rejecting the Null

If the chosen significance level is $\alpha = 0.05$, then _____.

    A. There is a 5% probability of accepting a true null hypothesis
    B. There is a 5% probability of rejecting a true null hypothesis
    C. There is a 5% probability of rejecting a false null hypothesis
    D. There is a 5% probability of accepting a false null hypothesis

3.9.2 Exercise 2: Compare Different Significance Levels

If the null hypothesis is rejected at a 1% significance level, then _____.

    A. The alternative hypothesis will be rejected at a 5% significance level
    B. The null hypothesis will not be rejected at a 5% significance level
    C. The null hypothesis will be rejected at a 5% significance level
    D. The alternative hypothesis will not be rejected at a 5% significance level

3.9.3 Exercise 3: Critical Values with a Two-tailed Test

A two-tailed hypothesis test of the population mean or population proportion has _____.

    A. Two critical values, both positive
    B. Two critical values, both negative
    C. Only one critical value
    D. Two critical values, one positive and one negative

3.9.4 Exercise 4: Critical Values with a One-tailed Test

A one-tailed hypothesis test of the population mean has _____.

    A. Only one critical value
    B. Two critical values, both positive
    C. Two critical values, both negative
    D. Two critical values, one positive and one negative

3.9.5 Exercise 5: Find the Critical Values

Find the critical value(s) and specify the degrees of freedom (if applicable) for the following hypothesis tests:

    a. $H_0: \mu \leq 4.5$; $H_A: \mu > 4.5$; $\alpha = 0.05$; $n = 24$

    b. $H_0: \mu \geq 4.5$; $H_A: \mu < 4.5$; $\alpha = 0.05$; $n = 24$

    c. $H_0: \mu = 4.5$; $H_A: \mu \neq 4.5$; $\alpha = 0.05$; $n = 24$

    d. $H_0: p \leq 0.2$; $H_A: p > 0.2$; $\alpha = 0.05$

    e. $H_0: p \geq 0.2$; $H_A: p < 0.2$; $\alpha = 0.05$

    f. $H_0: p = 0.2$; $H_A: p \neq 0.2$; $\alpha = 0.05$

3.9.6 Exercise 4: Private Emails at Work

A company decided to test the hypothesis that the average time a company's employees are spending to check their private e-mails at work is more than 6 minutes. A random sample of 40 employees were selected and they averaged 6.6 minutes with a standard deviation of 1.7 minutes. The α is set to 0.05. The critical value for this hypothesis test would be _____.

6. one-tail, $t^*_{0.05,39} = 1.685$

$z^*_{0.05} = -1.645$, $z^*_{0.025} = \mp 1.96$

$t^*_{0.025,23} = \mp 2.069$, $z^*_{0.05} = 1.645$,

5. $t^*_{0.05,23} = 1.714$, $t^*_{0.05,23} = -1.714$,

4. A

3. D

2. C

1. B

3.10 Test Statistic and P-value Practice Problems

The p-value exercises are also in the Excel file:
Hypothesis Testing Practice Problems.xlsx

Answers to the p-value exercises are in the Excel file:
Hypothesis Testing Practice Problems KEY.xlsx

3.10.1 Exercise 1: Calculate the Test Statistic

Calculate the test statistic and specify the degrees of freedom (if applicable) for the following tests:

a. $H_0: \mu \leq 4.5$; $H_A: \mu > 4.5$ 　　　　　　$\bar{x} = 4.8$; $s = 0.8$; $n = 24$

b. $H_0: p = 0.2$; $H_A: p \neq 0.2$ 　　　　　　$\bar{p} = 0.23$; $n = 30$

c. $H_0: \mu \geq 200$; $H_A: \mu < 200$ 　　　　　　$\bar{x} = 196$; $s = 9.8$; $n = 26$

3.10.2 Exercise 2: Luxury Cars Dealer

A car dealer who sells only late-model luxury cars recently hired a new salesperson and believes that this salesperson is selling at lower markups. He knows that the long-run average markup in his lot is $5,600. He takes a random sample of 16 of the new salesperson's sales and finds an average markup of $5,000 and a standard deviation of $800. Assume the markups are normally distributed. What is the value of an appropriate test statistic for the car dealer to use to test his claim?

3.10.3 Exercise 3: Texas Babies

It is generally believed that no more than 0.50 of all babies in a town in Texas are born out of wedlock. A politician claims that the proportion of babies born out of wedlock is increasing. In testing the politician's claim, you draw a sample of 200 newborn and find that 55% of them were born out of wedlock. What is the value of an appropriate test statistic to determine if the proportion of babies born out of wedlock has increased?

3.10.4 Exercise 4: Calculate the Test Statistic and P-value

A company decided to test the hypothesis that the average time a company's employees are spending to check their private e-mails at work is more than 6 minutes. A random sample of 40 employees were selected and they averaged 6.6 minutes with a standard deviation of 1.7 minutes. The α is set to 0.05. The *p*-value for this hypothesis test would be _____.

3.10.5 Exercise 5: Approximate the p-value

Consider the following hypotheses:

$H_0: \mu \leq 210;\ H_A: \mu > 210$

Approximate the p-value for this test based on the following sample information.

    a. $\bar{x} = 216;\ s = 26;\ n = 40$

    b. $\bar{x} = 216;\ s = 26;\ n = 80$

    c. $\bar{x} = 216;\ s = 16;\ n = 40$

Consider the following hypotheses:

$H_0: \mu = 12;\ H_A: \mu \neq 12$

Approximate the p-value for this test based on the following sample information.

    d. $\bar{x} = 11;\ s = 3.2;\ n = 36$
    e. $\bar{x} = 11;\ s = 2.8;\ n = 36$
    f. $\bar{x} = 11;\ s = 2.8;\ n = 49$

Consider the following hypotheses:

$H_0: p \leq 0.5;\ H_A: p > 0.5$

Approximate the p-value for this test based on the following sample information.

    g. $\bar{p} = 0.55;\ n = 50$

    h. $\bar{p} = 0.55;\ n = 200$

3.11 Draw a Conclusion Practice Problems

3.11.1 Exercise 1: Rejection Criteria

To test if the mean returns on a major index have changed from the historic monthly average of 1.2%, a sample of 36 recent monthly returns is used to calculate the value of the relevant t_{df} test statistic. At the 5% level of significance, we reject the null hypothesis if this value is _____.

  A. Greater than 1.69
  B. Greater than 1.645
  C. Greater than 2.03 or less than −2.03
  D. Greater than 1.96 or less than −1.96

3.11.2 Exercise 2: Interpret the Critical Value/Test Statistic

An analyst conducts a hypothesis test to check whether the mean return for a particular fund differs from 10%. He assumes that returns are normally distributed and sets up the following competing hypotheses: $H_o{:}\mu = 10$, $H_A{:}\mu \neq 10$. Over the past 10 years the fund has had an average annual return of 13.4% with a standard deviation of 2.6%. The value of the test statistic is 4.14 and the critical values at the 5% significance level are−2.262 and 2.262.The correct decision is to

_____.

  A. Reject H_o; we cannot conclude that the mean differs from 10%
  B. Reject H_o; we can conclude that the mean differs from 10%
  C. Not reject H_o; we can conclude that the mean differs from 10%
  D. Not reject H_o; we cannot conclude that the mean differs from 10%

3.11.3 Exercise 3: P-value Decision Rule

What is the decision rule when using the *p*-value approach to hypothesis testing?

  A. Reject H_o if the *p*-value < α.
  B. Reject H_o if the *p*-value > α.
  C. Do not reject H_o if the *p*-value < $1 - \alpha$.
  D. Do not reject H_o if the *p*-value > $1 - \alpha$.

3.11.4 Exercise 4: Correct Conclusion Based on P-value 0.027; α = 0.05

If the *p*-value for a hypothesis test is 0.027 and the chosen level of significance is α = 0.05, then the correct conclusion is to _____.

  A. Reject the null hypothesis if $\sigma = 10$
  B. Not reject the null hypothesis if $\sigma = 10$
  C. Reject the null hypothesis
  D. Not reject the null hypothesis

3.11.5 Exercise 5: Correct Conclusion Based on P-value 0.07; α = 0.05

If the p-value for a hypothesis test is 0.07 and the chosen level of significance is $\alpha = 0.05$, then the correct conclusion is to _____.

    A. Reject the null hypothesis
    B. Reject the null hypothesis if $\sigma = 10$
    C. Not reject the null hypothesis
    D. Not reject the null hypothesis if $\sigma = 10$

3.11.6 Exercise 6: Calculate and Interpret the P-value

Consider the following competing hypotheses: $H_o:p = 0.2$, $H_A:p \neq 0.2$. The value of the test statistic is $z = -1.38$. If we choose a 5% significance level, then we _____.

    A. Do not reject the null hypothesis and conclude that the population proportion is significantly different from 0.2.
    B. Do not reject the null hypothesis and conclude that the population proportion is not significantly different from 0.2.
    C. Reject the null hypothesis and conclude that the population mean is significantly different from 0.2.
    D. Reject the null hypothesis and conclude that the population mean is not significantly different from 0.2.

3.11.7 Exercise 7: Test if the Mean IQ is Greater than 100

To test if the mean IQ of employees in an organization is greater than 100, a sample of 30 employees is taken and the value of the test statistic is computed as $t_{29} = 2.42$ If we choose a 5% significance level and conduct a test with $H_0: \mu \leq 100$; $H_A: \mu > 100$, we _____.

    A. Do not reject the null hypothesis and conclude that the mean IQ is not greater than 100
    B. Do not reject the null hypothesis and conclude that the mean IQ is greater than 100
    C. Reject the null hypothesis and conclude that the mean IQ is not greater than 100
    D. Reject the null hypothesis and conclude that the mean IQ is greater than 100

3.11.8 Exercise 8: Test if Americans are Retiring Later

A recent report claimed that Americans are retiring later in life (*U.S. News & World Report*, August 17, 2011). An economist wishes to determine if the mean retirement age has increased from 62. To conduct the relevant test $H_0: \mu \leq 62$; $H_A: \mu > 62$, she takes a random sample of 38 Americans who have recently retired and computes the value of the test statistic as $t_{37} = 1.92$. With $\alpha = 0.05$, she _____.

    A. Does not reject the null hypothesis and does not conclude that the mean retirement age has not increased
    B. Does not reject the null hypothesis and does not conclude that the mean retirement age has increased
    C. Rejects the null hypothesis and concludes that the mean retirement age has not increased
    D. Rejects the null hypothesis and concludes that the mean retirement age has increased

3.12 Types of Error Practice Problems

3.12.1 Exercise 1: Type I Error

A Type I error occurs when we _____.

    A. Do not reject the null hypothesis when it is actually true
    B. Do not reject the null hypothesis when it is actually false
    C. Reject the null hypothesis when it is actually true
    D. Reject the null hypothesis when it is actually false

3.12.2 Exercise 2: Type II Error

A Type II error occurs when we _____.

    A. Do not reject the null hypothesis when it is actually true
    B. Do not reject the null hypothesis when it is actually false
    C. Reject the null hypothesis when it is actually true
    D. Reject the null hypothesis when it is actually false

3.12.3 Exercise 3: Reject a False Null Hypothesis

When we reject the null hypothesis when it is actually false, we have committed _____.

    A. A Type I error
    B. A Type II error
    C. A Type I error and a Type II error
    D. No error

3.12.4 Exercise 4: Incorrect Decisions

When conducting a hypothesis test, which of the following decisions represents an error?

    A. Rejecting the null hypothesis when it is false.
    B. Rejecting the null hypothesis when it is true.
    C. Not rejecting the null hypothesis when it is true.
    D. Rejecting the null hypothesis when it is false and not rejecting the null hypothesis when it is true.

3.12.5 Exercise 5: Calculate Error (Polygraph)

A polygraph (lie detector) is an instrument used to determine if the individual is telling the truth. These tests are considered to be 95% reliable. In other words, if an individual lies, there is a 0.95 probability that the test will detect a lie. Let there also be a 0.005 probability that the test erroneously detects a lie even when the individual is actually telling the truth. Consider the null hypothesis, "the individual is telling the truth," to answer the following questions.

    1. What is the probability of Type I error?

    2. What is the probability of Type II error?

3.12.6 Exercise 6: Calculate Error (Steroids Test)

A professional sports organization is going to implement a test for steroids. The test gives a positive reaction in 94% of the people who have taken the steroid. However, it erroneously gives a positive reaction in 4% of the people who have not taken the steroid. What is the probability of Type I and Type II errors giving the null hypothesis "the individual has not taken steroids."

3.12.7 Exercise 7: Calculate Error (Cheating)

A statistics professor works tirelessly to catch students cheating on his exams. He has particular routes for his teaching assistants to patrol, an elevated chair to ensure an unobstructed view of all students, and even a video recording of the exam in case additional evidence needs to be collected. He estimates that he catches 95% of students who cheat in his class, but 1% of the time that he accuses a student of cheating he is actually incorrect. Consider the null hypothesis, "the student is not cheating." What is the probability of a Type I error?

3.12.8 Exercise 8: Sample Size, n

For a given sample size n, _____.

  A. Decreasing the probability of a Type I error α will increase the probability of a Type II error β
  B. Decreasing the probability of a Type I error α will decrease the probability of a Type II error β
  C. Changing the probability of a Type I error α will have no impact on the probability of a Type II error β
  D. Increasing the probability of a Type I error α will increase the probability of a Type II error β as long as σ is known

3.12.9 Exercise 9: Fast-Food Franchise Type I Error

A fast-food franchise is considering building a restaurant at a busy intersection. A financial advisor determines that the site is acceptable only if, on average, more than 300 automobiles pass the location per hour. The advisor tests the following hypotheses:

H_o: $\mu \leq 300$; H_A: $\mu > 300$.

The consequences of committing a Type I error would be that_____.

  A. The franchiser does not build on an acceptable site
  B. The franchiser does not build on an unacceptable site
  C. The franchiser builds on an acceptable site
  D. The franchiser builds on an unacceptable site

3.12.10 Exercise 10: Fast-Food Franchise Type II Error

A fast-food franchise is considering building a restaurant at a busy intersection. A financial advisor determines that the site is acceptable only if, on average, more than 300 automobiles pass the location per hour. The advisor tests the following hypotheses:

H_o: $\mu \leq 300$.

H_A: $\mu > 300$.

The consequences of committing a Type II error would be that_____.

    A. The franchiser does not build on an acceptable site
    B. The franchiser does not build on an unacceptable site
    C. The franchiser builds on an acceptable site
    D. The franchiser builds on an unacceptable site

3.12.11 Exercise 11: Company Manager Concerns

A company has developed a new diet that it claims will lower one's weight by more than 10 pounds. Health officials decide to conduct a test to validate this claim. The manager of the company _____.

    A. Is not concerned at all
    B. Is more concerned about Type I error
    C. Is more concerned about Type II error
    D. Is concerned about both Type I and Type II errors

3.12.12 Exercise 12: Consumer Concerns

A company has developed a new diet that it claims will lower one's weight by more than 10 pounds. Health officials decide to conduct a test to validate this claim. The consumers should be _____.

    A. More concerned about Type II error
    B. More concerned about Type I error
    C. Concerned about both Type I and Type II errors
    D. Is not concerned at all

1. C
2. B
3. D
4. B
5. Type I: 0.005, Type II: 0.05
6. Type I: 4%, Type II: 6%
7. 1%
8. A
9. D
10. A
11. C
12. B

62

3.13 Hypothesis Testing Practice Problems

3.13.1 Exercise 1: Population Mean

A hypothesis test regarding the population mean is based on _____.

    A. The sampling distribution of the sample variance
    B. The sampling distribution of the sample mean
    C. The sampling distribution of the sample standard deviation
    D. The sampling distribution of the sample proportion

3.13.2 Exercise 2: Population Proportion

A hypothesis test regarding the population mean is based on _____.

    A. The sampling distribution of the sample variance
    B. The sampling distribution of the sample mean
    C. The sampling distribution of the sample standard deviation
    D. The sampling distribution of the sample proportion

3.13.3 Exercise 3: Confidence Interval Test

A 99% confidence interval for the population mean yields the following results: $[-3.79, 5.86]$. At the 1% significance level, what decision should be made regarding the following hypothesis test with $H_o : \mu = 0$, $H_A : \mu \neq 0$?

    A. Do not reject H_o; we can conclude that the mean differs from zero.
    B. Do not reject H_o; we cannot conclude that the mean differs from zero.
    C. Reject H_o; we can conclude that the mean differs from zero.
    D. Reject H_o; we cannot conclude that the mean differs from zero.

3.13.4 Exercise 4: University Honors Program

A university is interested in promoting graduates of its honors program by establishing that the mean GPA of these graduates exceeds 3.50. A sample of 36 honors students is taken and is found to have a mean GPA equal to 3.60 and a standard deviation equal to 0.25.

    a. Define the appropriate hypotheses to establish whether the mean GPA exceeds 3.50.

    b. Calculate the value of the test statistic.

    c. Determine the critical value(s) for a 5% level of significance.

    d. Using the critical value approach, can you conclude that the mean GPA is greater than 3.50?

3.13.5 Exercise 5: Car Dealership

The owner of a large car dealership believes that the financial crisis decreased the number of customers visiting her dealership. The dealership has historically had 800 customers per day. The owner takes a sample of 100 days and finds the average number of customers visiting the dealership per day was 725 with a standard deviation of 350.

    a. Define the appropriate hypotheses to establish to determine whether there has been a decrease in the average number of customers visiting the dealership daily.

    b. Calculate the value of the test statistic.

    c. Determine the critical value(s) for a 5% level of significance.

    d. Using the critical value approach, can you conclude that there has been a decrease in the average number of customers visiting the dealership daily?

3.13.6 Exercise 6: Boston Public Schools

The Boston public school district has had difficulty maintaining on-time bus service for its students ("A Year Later, School Buses Still Late," *Boston Globe*, October 5, 2011). Suppose the district develops a new bus schedule to help combat chronic lateness on a particularly woeful route. Historically, the bus service on the route has been, on average, 12 minutes late. After the schedule adjustment, the first 36 runs were an average of 8 minutes late with a standard deviation of 10 minutes. As a result, the Boston public school district claimed that the schedule adjustment was an improvement—students were not as late.

    a. Define the appropriate hypotheses to establish to determine whether the schedule adjustment reduced the average lateness time of 12 minutes.

    b. Calculate the value of the test statistic.

    c. Calculate the p-value.

    d. At the 5% significance level, does the evidence support the Boston public school district's claim?

    e. At the 1% significance level, does the evidence support the Boston public school district's claim?

3.13.7 Exercise 7: Department of Education

The Department of Education would like to test the hypothesis that the average debt load of graduating students with a bachelor's degree differs from $17,000. A random sample of 34 students had an average debt load of $18,200 with a standard deviation of $4,200. The α is set to 0.05.

    a. The confidence interval for this hypothesis test would be _____.

    b. The p-value for this hypothesis test would be _____.

3.13.8 Exercise 8: GPA Below 3.00

A university interested in tracking its honors program believes that the proportion of graduates with a GPA of 3.00 or below is less than 0.20. In a sample of 200 graduates, 30 students have a GPA of 3.00 or below.

    a. In testing the university's belief, the appropriate hypotheses are _____.

    b. The value of the test statistic and its associated p-value are _____.

    c. At a 5% significance level, the decision is to…

3.13.9 Exercise 9: Prime Viewing Time

A television network is deciding whether or not to give its newest television show a spot during prime viewing time at night. If this is to happen, it will have to move one of its most viewed shows to another slot. The network conducts a survey asking its viewers which show they would rather watch. The network receives 863 responses, of which 467 indicate they would like to see the new show in the lineup.

a. Define the appropriate hypotheses to test if the television network should give its newest show a spot during prime time at night?

b. The test statistic for this hypothesis would be _____.

c. At the 1% significance level, the critical value that marks the rejection region(s) for this hypothesis would be _____.

d. At the 1% significance level, does the data support the network airing the new show in the prime time slot?

3.13.10 Exercise 10: Institute of Education Sciences

The Institute of Education Sciences measures the high school dropout rate as the percentage of 16-through 24-year-olds who are not enrolled in school and have not earned a high school credential. In 2009, the high school dropout rate was 8.1%. A polling company recently took a survey of 1,000 people between the ages of 16 and 24 and found that 7.0% of them are high school dropouts. The polling company would like to determine whether the dropout rate has decreased.

a. When testing whether the dropout rate has decreased, define the appropriate hypotheses.

b. The value of the test statistic and p-value are…

c. At a 5% significance level, the decision is to _____.

1. B

2. D

3. B

4. $H_0: \mu \leq 3.5$; $H_A: \mu > 3.5$

   $t_{35}=2.38$

   $t^*_{0.05,35}=1.69$

   Reject H_0; conclude the mean GPA is significantly greater than 3.5.

5. $H_0: \mu \geq 800$; $H_A: \mu < 800$

   $t_{99}=-2.14$

   $t^*_{0.05,99}=-1.66$

   Reject H_0; conclude the mean number of customers visiting the dealership is significantly less than 800.

6. $H_0: \mu \geq 12$; $H_A: \mu < 12$

   $t_{35}=-2.395$

   p-value=0.0111 [=t.dist(-2.395, 35, 1)]

   5% significance: Reject H_0; the data support the district's claim that the schedule adjustment reduced the average lateness time of 12 minutes.

   1% significance: Do not reject H_0; the data do not support the district's claim that the schedule adjustment reduced the average lateness time of 12 minutes.

7. $t^*_{0.025,33}=2.035$

   [$15,534.2, $18,465.8]

   $t_{33}=1.67$

   p-value=0.1044 [=T.DIST.2T(1.67, 33)]

   Do not reject H_0, do not conclude that student debt load differs from $17,000.

8. $H_0: p \geq 0.2$; $H_A: p < 0.2$

   z=-0.88

   p-value=0.1894 [=NORM.S.DIST(-0.88, 1)]

   Do not reject H_0; do not conclude that the proportion of graduates with a GPA of 3.00 or below is significantly less than 0.20.

9. $H_0: p \leq 0.5$; $H_A: p > 0.5$

   z=2.41

   z*=2.33

   Reject H_0: conclude that the proportion of viewers who would like to see the new show is greater than 0.5. The data support the network airing the new show in the prime time slot.

10. $H_0: p \geq 0.081$; $H_A: p < 0.081$

    z=-1.28

    p-value=0.0319 [=NORM.S.DIST(-1.28, 1)]

    Do not reject H_0; do not conclude that the high school dropout rate has decreased.

3.14 More Hypothesis Testing Practice Problems

The following exercises are in the Excel file:
Hypothesis Testing Practice Problems.xlsx

Answers to the following exercises are in the Excel file:
Hypothesis Testing Practice Problems KEY.xlsx

3.14.1 Fixed Mortgages

Rates on 30-year fixed mortgages continue to be at historic lows (*Chron Business News,* September 23, 2010). According to Freddie Mac, the average rate for 30-year fixed loans for the week was 4.27%. An economist wants to test if there is any change in the mortgage rates in the following week. She searches the Internet for 30-year fixed loans in the following week and reports the rates offered by seven banks as: 4.25%, 4.125%, 4.375%, 4.50%, 4.75%, 4.375%, and 4.875%. Assume that rates are normally distributed.

At a 5% level of significance, test if the average mortgage rate differs from 4.27%.

3.14.2 Monthly Sales

An entrepreneur examines monthly sales (in $1,000s) for 40 convenience stores in Rhode Island. Test whether average sales differ from $130,000 at a 5% level of significance.

3.14.3 Cell Phone Use

A Pew Research study finds that 25% of Americans use only a cell phone, and no land line, for making phone calls (*The Wall Street Journal*, October 14, 2010). A year later, a researcher samples 200 Americans and finds that 60 of them use only cell phones for making phone calls.

At a 5% level of significance, test whether the proportion of Americans who solely use cell phones to make phone calls differs from 25%.

3.14.4 Proportion of Women

According to a report on workforce diversity, about 60% of the employees in high-tech firms in Silicon Valley are white and about 20% are Asian (http://moneycnn.com, November 9, 2011). Women, along with blacks and Hispanics, are highly underrepresented. Just about 28% of all employees are women, with blacks and Hispanics, accounting for only about 15% of the workforce. Tara Jones is a recent college graduate, working for a large high-tech firm in Silicon Valley. She wants to determine if her firm faces the same diversity as in the report. She collects gender and ethnicity information on 50 employees in her firm. A random sample of the data is shown in the Excel file.

At the 5% level of significance, determine if the proportion of women in Tara's firm is different from 30%.

At the 5% level of significance, determine if the proportion of whites in Tara's firm is more than 50%.

3.14.5 Highway Speeds

The speed limit on this portion of Interstate 95 is 64 mph. Define the competing hypotheses in order to determine if the average speed is greater than the speed limit.

At α = 0.10, are the officer's concerns warranted?

3.14.6 Computer Prices

Small, energy-efficient, Internet-centric, new computers are increasingly gaining popularity (*New York Times*, July 20, 2008). These computers, often called netbooks, have scant onboard memory and are intended largely for surfing websites and checking e-mail. Some of the biggest companies are wary of the new breed of computers because their low price could threaten PC makers' already thin profit margins. An analyst comments that the larger companies have a cause for concern since the mean price of these small computers has fallen below $350. She examines six popular brands of these small computers and records their retail prices.

What assumption regarding the distribution of the price of small computers is necessary to test the analyst's claim?

At the 5% significance level, should the larger computer companies be concerned?

3.14.7 Retailer Services

A retailer is looking to evaluate its customer service. Management has determined that if the retailer wants to stay competitive, then it will have to have at least an 91% satisfaction rate among its customers. Management will take corrective actions if the satisfaction rate falls below 91%. A survey of 1,450 customers showed that 1,305 were satisfied with their customer service.

At a 5% level of significance, test if the retailer needs to improve its services.

3.14.8 Movie Viewers

A movie production company is releasing a movie with the hopes of many viewers returning to see the movie in the theater for a second time. Their target is to have 30 million viewers, and they want more than 30% of the viewers to return to see the movie again. They show the movie to a test audience of 200 people, and after the movie they asked them if they would see the movie in theaters again. Of the test audience, 65 people said they would see the movie again.

At a 5% level of significance, test if more than 30% of the viewers will return to see the movie again.

Repeat the analysis at a 10% level of significance.

4 Statistical Inference Concerning Two Populations

4.1 Objectives

- Make inferences about the difference between two population means based on independent sampling.
- Make inferences about the difference between two population proportions based on independent sampling.

4.2 Confidence Intervals for the Difference Between Two Populations

	CI for the Difference Between Two Means Independent Sampling with Quantitative Data	CI for the Difference Between Two Proportions Independent Sampling with Qualitative Data
Example Questions	Does the mean GPA among freshman differ from the mean GPA among sophomores at a university?	Does the incidence of tooth decay among three-year-old Pit Bulls differ from the incidence of tooth decay among golden retrievers?
Similarities with one sample confidence intervals…	The process is the same for all confidence intervals… 1. Determine the point estimate. 2. Find the critical value related to the desired confidence. 3. Calculate the standard error. 4. Calculate the margin of error. 5. Calculate the interval with the point estimate as the center; subtract/add the margin of error.	
Differences from one sample confidence intervals…	**Calculate the point estimate…** **Standard error and degrees of freedom…**	**Calculate the point estimate…** **Standard error…**

4.3 Testing the Difference Between Two Populations

	Testing the Difference Between Two Means **Independent Sampling with Quantitative Data**	**Testing the Difference Between Two Proportions** **Independent Sampling with Qualitative Data**
Example Questions	Which of two restaurants, owned by the same person, generates more revenue? Of two countries, country 1 and country 2, is the life expectancy in country 1 less than that of country 2?	Compare the likelihood of two candidates in two different elections winning their elections. Is there a higher incident of smoking among women than among men in a neighborhood?
Similarities with One Sample Hypothesis Testing	The components and process are the same for all hypothesis tests: 1. Define null and alternative hypotheses 2. Determine the level of significance, α. 3. Calculate the critical value (critical value approach). 4. Calculate the test statistic. 5. Calculate the p-value (p-value approach). 6. Report the conclusion.	

	Testing the Difference Between Two Means Independent Sampling with Quantitative Data	Testing the Difference Between Two Proportions Independent Sampling with Qualitative Data
Differences from One Sample Hypothesis Testing	**Test statistic and degrees of freedom for the difference between two means…**	**Test statistic for the difference between two proportions…**
	Null and Alternative Hypotheses for the difference between two means…	**Null and Alternative Hypotheses for the difference between two proportions…**

	Testing the Difference Between Two Means **Independent Sampling with Quantitative Data**	**Testing the Difference Between Two Proportions** **Independent Sampling with Qualitative Data**
Conditions and Assumptions	**Observations in one sample are independent of the observations in the other sample.** Two or more random samples are considered independent if the process that generates one sample is completely separate from the process that generates the other sample. **The Sampling Distribution $\bar{X}_1 - \bar{X}_2$ is normally distributed.** If _____ _____ meets normality requirements, then $\bar{X}_1 - \bar{X}_2$ can be assumed normal, i.e., either both populations are normal or the size of each sample is greater than or equal to 30: **Variances equal or unequal?** There are two ways to test the difference between two means. One way is to assume the variances of the two populations are equal, and the other is to assume the variances are unequal. This class is concerned only with problems where the variances are assumed equal.	**Observations in one sample are independent of the observations in the other sample.** Two or more random samples are considered independent if the process that generates one sample is completely separate from the process that generates the other sample. **The Sampling Distribution $\bar{P}_1 - \bar{P}_2$ is normally distributed.** If _____ _____ meets normality requirements, then $\bar{P}_1 - \bar{P}_2$ can be assumed normal, i.e., all of the following must be true:

[This page intentionally left blank]

4.4 Two Population Confidence Interval Practice Problems

4.4.1 Example 1: SAT Math Score

A university wants to compare out-of-state applicants' mean SAT math scores (μ_1) to in-state applicants' mean SAT math scores (μ_2). The university looks at 35 in-state applicants and 35 out-of-state applicants. The mean SAT math score for in-state applicants was 540, with a standard deviation of 20. The mean SAT math score for out-of-state applicants was 555, with a standard deviation of 25. It is reasonable to assume the corresponding population standard deviations are equal.

a. To calculate the confidence interval for the difference $\mu_1 - \mu_2$, what is the number of degrees of freedom of the appropriate probability distribution?

b. Calculate a 90% confidence interval for the difference $\mu_1 - \mu_2$.

c. At the 10% significance level, can the university conclude that the mean SAT math score for in-state students and out-of-state students differ?

4.4.2 Exercise 2: Tooth Decay

A veterinarian wants to know if pit bulls or golden retrievers have a higher incidence of tooth decay at the age of three. The vet surveys 120 three-year-old pit bulls and finds 30 of them have tooth decay. The vet then surveys 160 three-year-old golden retrievers and finds 32 of them have tooth decay. Number the population of pit bulls and golden retrievers by 1 and 2, respectively.

a. Calculate a 95% confidence interval for the difference in the population proportion of pit bulls and golden retrievers that have tooth decay.

b. At the 5% significance level, can the veterinarian conclude the proportion of pit bulls that have tooth decay is different than the proportion of golden retrievers that have tooth decay?

4.4.3 Exercise 3: Calcium

Calcium is an essential nutrient for strong bones and for controlling blood pressure and heart beat. Because most of the body's calcium is stored in bones and teeth, the body withdraws the calcium it needs from the bones. Over time, if more calcium is taken out of the bones than is put in, the result may be thin, weak bones. This is especially important for women who are often recommended a calcium supplement. A consumer group activist assumes that calcium content in two popular supplements are normally distributed with the same unknown population variance, and uses the following information obtained under independent sampling:

Supplement 1	Supplement 2
$\bar{x}_1$ = 1,000 mg	$\bar{x}_2$ = 1,016 mg
s_1 = 23 mg	s_2 = 24 mg
n_1 = 12	n_2 = 15

   a. Let μ_1 and μ_2 denote the corresponding population means. Construct a 98% confidence interval for the difference $\mu_1 - \mu_2$.

   b. Let μ_1 and μ_2 denote the corresponding population means. Can we conclude that the average calcium content of the two supplements differs at the 98% confidence level?

1. Point Estimate = -15 points, df=68, s_p^2=512.5, $e(\bar{X}_1 - \bar{X}_2)$ = 5.41 $t^*_{0.05,68}$ = 1.667,
ME = 9.02
[-24.02, -5.98]
Yes, because the confidence interval does not contain zero.

2. Point Estimate = 0.05, $z^*_{0.025}$ = 1.96, $se(\hat{P}_1 - \hat{P}_2)$ = 0.0506, ME = 0.0992
[-0.0492, 0.1492]
No, because the confidence interval contains zero.

3. Point Estimate = 16 mg, df=25, s_p^2=555.32, $e(\bar{X}_1 - \bar{X}_2)$ = 9.1268 $t^*_{0.01,25}$ = 2.485,
ME = 22.68
[-38.68, 6.68]
No, because the 98% confidence interval contains the hypothesized value of zero.

4.5 Two Population Hypothesis Test Practice Problems (Basic)

4.5.1 Example 1: Mean GPA

Suppose you want to perform a test to compare the mean GPA of all freshmen with the mean GPA of all sophomores in a college? What type of sampling is required for this test?

    A. Independent sampling with qualitative data
    B. Independent sampling with quantitative data
    C. Dependent sampling with qualitative data
    D. Dependent sampling with quantitative data

4.5.2 Example 2: Likelihood of Winning

What type of data should be collected when examining a situation in which two candidates running in different elections are being compared in their likelihood of winning their elections?

    A. Dependent sampling with qualitative data.
    B. Dependent sampling with quantitative data.
    C. Independent sampling with qualitative data.
    D. Independent sampling with quantitative data.

4.5.3 Example 3: Hypotheses

Which of the following pairs of hypotheses are used to test if the mean of the first population is smaller than the mean of the second population, using independent random sampling?

    A. $H_0: p_1 - p_2 \leq 0$, $H_A: p_1 - p_2 > 0$
    B. $H_0: p_1 - p_2 \geq 0$, $H_A: p_1 - p_2 < 0$
    C. $H_0: \mu_1 - \mu_2 \leq 0$, $H_A: \mu_1 - \mu_2 > 0$
    D. $H_0: \mu_1 - \mu_2 \geq 0$, $H_A: \mu_1 - \mu_2 < 0$

4.5.4 Exercise 4: Loan Modification Programs

A particular bank has two loan modification programs for distressed borrowers: Home Affordable Modification Program (HAMP) modifications, where the federal government pays the bank $1,000 for each successful modification, and non-HAMP modifications, where the bank does not receive a bonus from the federal government. To qualify for a HAMP modification, borrowers must meet a set of financial suitability criteria. What type of hypothesis test should we use to test whether borrowers from this particular bank who receive HAMP modifications are more likely to re-default than those who receive non-HAMP modifications?

    A. A hypothesis test for p_1 - p_2.
    B. A hypothesis test for μ_1 - μ_2.
    C. A hypothesis test for p with a sample proportion.
    D. A hypothesis test for μ with a sample mean.

4.5.5 Example 5: Loan Modification Programs

A particular bank has two loan modification programs for distressed borrowers: Home Affordable Modification Program (HAMP) modifications, where the federal government pays the bank $1,000 for each successful modification, and non-HAMP modifications, where the bank does not receive a bonus from the federal government. To qualify for a HAMP modification, borrowers must meet a set of financial suitability criteria. Define the null and alternative hypotheses to test whether borrowers who receive HAMP modifications default less than borrowers who receive non-HAMP modifications. Let p_1 and p_2 represent the proportion of borrowers who received HAMP and non-HAMP modifications that did not re-default, respectively.

4.5.6 Example 6: Two Restaurants

A restaurant chain has two locations in a medium-sized town and, believing that it has oversaturated the market for its food, is considering closing one of the restaurants. The manager of the restaurant with a downtown location claims that his restaurant generates more revenue than the sister restaurant by the freeway. The CEO of this company, wishing to test this claim, randomly selects 36 monthly revenue totals for each restaurant. The revenue data from the downtown restaurant have a mean of $360,000 and a standard deviation of $50,000, while the data from the restaurant by the freeway have a mean of $340,000 and a standard deviation of $40,000. Let μ_1 and μ_2 denote the mean monthly revenue of the downtown restaurant and the restaurant by the freeway, respectively. Define the hypotheses that should be used to test the manager's claim.

4.5.7 Example 7: Smoking Among Women

You would like to determine if there is a higher incidence of smoking among women than among men in a neighborhood. Let women and men be represented by populations 1 and 2, respectively. Define the hypotheses that are relevant to this claim.

4.5.8 Example 8: Smoking Among Men

You would like to determine if there is a higher incidence of smoking among women than among men in a neighborhood. Let men and women be represented by populations 1 and 2, respectively. Define the hypotheses that are relevant to this claim.

1. B
2. C
3. D
4. A
5. $H_0: p_1 - p_2 \geq 0,\ H_A: p_1 - p_2 < 0$
6. $H_0: \mu_1 - \mu_2 \leq 0,\ H_A: \mu_1 - \mu_2 > 0$
7. $H_0: p_1 - p_2 \leq 0,\ H_A: p_1 - p_2 > 0$
8. $H_0: p_1 - p_2 \geq 0,\ H_A: p_1 - p_2 < 0$

4.6 Two Population Hypothesis Test Practice Problems

4.6.1 Exercise 1: University Student Senate

The student senate at a local university is about to hold elections. A representative from the women's sports program and a representative from the men's sports program must both be elected. Two candidates, an incumbent and a challenger, are vying for each position and early polling results are presented next. A hypothesis test must be performed to determine whether the percentages of supporting votes are different between the two incumbent candidates. In a sample of 100 voters, 69 said that they would vote for the women's incumbent candidate. In a separate sample of 100 voters, 54 said they would vote for the men's incumbent candidate. Let p_1 and p_2 be the proportions of supporting votes for the incumbent candidates representing women's and men's sports programs, respectively.

a. Define the competing hypotheses to determine whether the percentages of supporting votes are different between the two incumbent candidates.

b. Specify the critical value(s) of the appropriate test for a 5% level of significance?

c. Calculate the test statistic.

d. At a 5% level of significance, can you conclude that the votes differ between the two incumbent candidates?

4.6.2 Exercise 2: Lost Luggage

A consumer magazine wants to figure out which of two major airlines lost a higher proportion of luggage on international flights. The magazine surveyed Standard Air (population 1) and Down Under airlines (population 2). Standard Air lost 56 out of 600 bags. Down Under airlines lost 34 of 500 bags. Does Standard Air have a higher population proportion of lost bags on international flights?

    a. Define the correct competing hypotheses.

    b. Calculate the test statistic.

    c. Calculate the p-value.

    d. At a 5% level of significance, can you conclude that Standard Air have a higher population proportion of lost bags on international flights?

4.6.3 Exercise 3: Right-to-Cure Period

In August 2010, Massachusetts enacted a 150-day right-to-cure period that mandates that lenders give homeowners who fall behind on their mortgage an extra five months to become current before beginning foreclosure proceedings. Policymakers claimed that the policy would result in a higher proportion of delinquent borrowers becoming current on their mortgages. To test this claim, researchers took a sample of 244 homeowners in danger of foreclosure in the time period surrounding the enactment of this law. Of the 100 who fell behind just before the law was enacted, 27 were able to avoid foreclosure, and of 144 who fell behind just after the law was enacted, 50 were able to avoid foreclosure. Let p_1 and p_2 represent the proportion of delinquent borrowers who avoid foreclosure just before and just after the right-to-cure law is enacted, respectively.

  a. Define the competing hypotheses that will test the policymakers' claim.

  b. Calculate the test statistic.

  c. Calculate the p-value.

  d. Assuming $\alpha = 0.05$, does the evidence support the policymakers' claim? Explain.

4.6.4　Exercise 4: Household Chores

A new study has found that, on average, 6- to 12-year-old children are spending less time on household chores today compared to 1981 levels. Suppose two samples representative of the study's results report the following summary statistics for the two periods. Assume the unknown population variances are equal.

1981 Levels	2008 Levels
$\bar{x}_1 = 28$ minutes	$\bar{x}_2 = 25$ minutes
$s_1 = 3.8$ minutes	$s_2 = 3.6$ minutes
$n_1 = 30$	$n_2 = 30$

a. Define the appropriate competing hypotheses?

b. Calculate the test statistic.

c. Specify the critical values for a 5% level of significance.

d. Using a 5% significance level, do children spend less time on household chores today compared to 1981 levels?

1. $H_0: p_1 - p_2 = 0$, $H_A: p_1 - p_2 \neq 0$
   $z^*_{0.025} = \mp 1.96$
   $z = 2.18$
   Reject $p_1 - p_2 = 0$; conclude that the proportions of supporting votes for the incumbent candidates differ.

2. $H_0: p_1 - p_2 \leq 0$, $H_A: p_1 - p_2 > 0$
   $z = 1.5$
   p-value=0.0668
   Fail to reject H_0; we cannot conclude that the proportion of lost luggage is higher for Standard Air than for Down Under airlines.

3. $H_0: p_1 - p_2 \geq 0$, $H_A: p_1 - p_2 < 0$
   $z = -1.32$
   p-value=0.0934
   No, because the p-value is greater than the significance level.

4. $H_0: \mu_1 - \mu_2 \leq 0$, $H_A: \mu_1 - \mu_2 > 0$
   df=58, s_p^2=13.7, $t_{58} = 3.14$, $t^*_{0.05,58} = 1.671$
   Reject $\mu_1 - \mu_2 \leq 0$, conclude that children spend less time on household chores today compared to 1981 levels.

4.7 Two Population Hypothesis Test Practice Problems (Excel)

The following exercises are in the Excel file:
Two Pop Means Practice Problems.xlsx

Answers to the exercises are in the Excel file:
Two Pop Means Practice Problems KEY.xlsx

4.7.1 Exercise 1: Website Searches

"The See Me" marketing agency wants to determine if time of day for a television advertisement influences website searches for a product. They have extracted the number of website searches occurring during a one-hour period after an advertisement was aired for a random sample of 26 day and 26 evening advertisements. The data and guidelines are in the Excel file.

4.7.2 Exercise 2: Different Diets

According to a study published in the New England Journal of Medicine, overweight people on low-carbohydrate and Mediterranean diets lost more weight and got greater cardiovascular benefits than people on a conventional low-fat diet (Boston Globe, July 17, 2008). A nutritionist wishes to verify these results and documents the weight loss (in pounds) of 22 dieters on the low-carbohydrate and Mediterranean diets and 22 dieters on the low-fat diet. Let Low-carb or Mediterranean and Low-fat diets represent populations 1 and 2, respectively. The data and guidelines are in the Excel file.

4.7.3 Exercise 3: Nicknames

Baseball has always been a favorite pastime in America and is rife with statistics and theories. In a recent paper, researchers showed that major league players who have nicknames live 2½ years longer than those without them (The Wall Street Journal, July 16, 2009). You do not believe in this result and decide to collect data on the lifespan of 25 baseball players along with a nickname variable that equals 1 if the player had a nickname and 0 otherwise. The data and guidelines are in the Excel file.

5 Chi-Square Test for Independence

5.1 Objectives

- Conduct a Chi-Square Test for Independence

5.2 Recall

Two events are **independent** if the occurrence of one event does not affect the probability of the other event occurring, i.e., $P(A|B) = P(A)$.

The **joint probability**, $P(A \cap B)$, is the product $P(A)P(B)$.

A **contingency table**, or cross-tabulation table, displays the frequencies of two qualitative variables in a matrix format and is used to study the association between the two variables.

5.3 Chi-Square Test of Independence

The Chi-Square Test of Independence is used to test if two categorical variables are independent of each other. The idea is to calculate the frequencies that we would expect to see if the two variables were independent and to compare those to the observed frequencies.

Conditions

-
-
-

5.4 Example Problem: Chi-Square Test of Independence

The following example is in the Excel file:
Chi-Sq Guided Notes Example.xlsx

Does the age of a customer affect the brand of digital camera he or she purchases? Using the Chi-Square Test of Independence, we can find out if a pattern exists between the age of the customer and the brand of camera he or she chooses. Use a 5% level of significance, $\alpha = 0.05$.

OBSERVED	CAMERA BRAND			
AGE GROUP	CANON	NIKON	SONY	Total
18-34	30	16	8	54
35-51	22	25	19	66
52 and older	8	9	13	30
Total	60	50	40	150

Step 1: Identify the Null and Alternative Hypotheses

H_0:

H_A:

Step 2: Calculate Expected Frequencies

$$f_e = \frac{(Row\ Total)(Column\ Total)}{Total\ Number\ of\ Observations}$$

OBSERVED (f_o)	CAMERA BRAND			
AGE GROUP	CANON	NIKON	SONY	Total
18-34	30	16	8	54
35-51	22	25	19	66
52 and older	8	9	13	30
Total	60	50	40	150

EXPECTED (f_e)	CAMERA BRAND			
AGE GROUP	CANON	NIKON	SONY	Total
18-34				54
35-51				66
52 and older				30
Total	60	50	40	150

Step 3: Calculate the Chi-Square Test Statistic

The test statistic for testing independence follows a new distribution: the chi-square distribution. The x^2 test statistic is always positive because it involves squared deviations.

The logic of the Ch-Square test statistic is as follows:

- If the observed frequencies are close to the expected frequencies, the term $(f_o - f_e)^2$ will be small. This will result in a small chi-square statistic, which increases the likelihood that there is support for the null hypothesis that the two variables are independent.

■ α

x^2
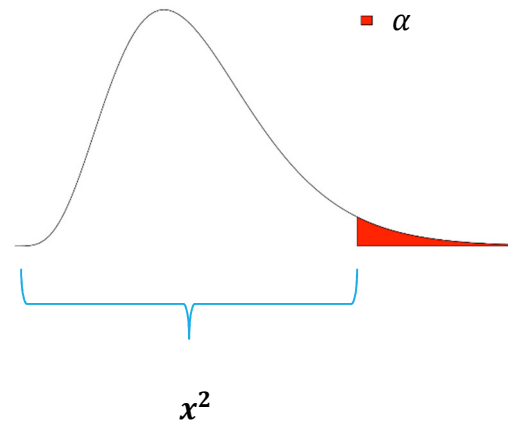
- If the observed frequencies are far apart from the expected frequencies, the term $(f_o - f_e)^2$ will be large. This will result in large chi-square statistic, which increases the likelihood of rejecting the null hypothesis and thus concluding that the two variables are not independent.

■ α

a
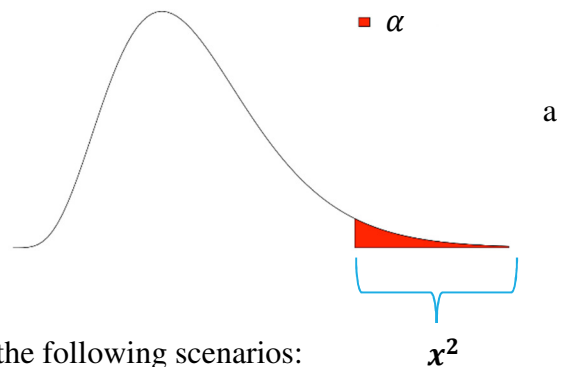
- The x^2 test statistic increases under either of the following scenarios: x^2
    o When observed frequencies are _____ _____ expected frequencies
    o When observed frequencies are _____ _____ expected frequencies

Step 4: Calculate the Ch-Square Critical Value at $\alpha = 0.05$, $x^2_{\alpha,df}$

The x^2 distribution has degrees of freedom:

where r is the number of rows and c is the number of columns in the contingency table.
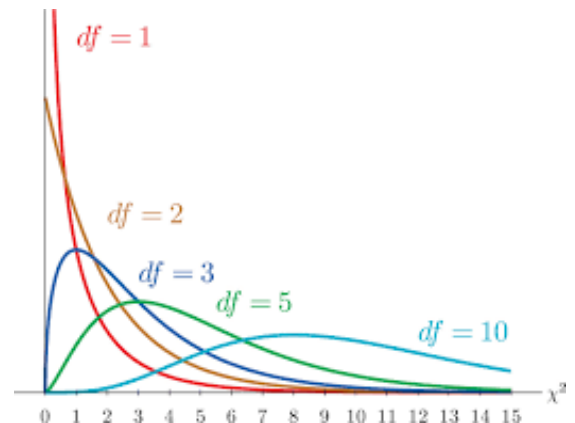
$x^2_{\alpha,df} = \text{CHISQ. INV. RT}(0.05, df) = $ _____

- $\alpha = 0.05$
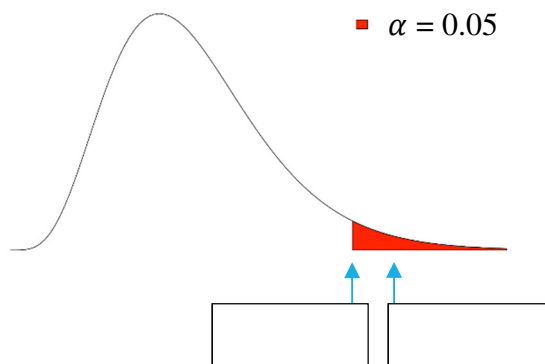
Because x^2_{df} _____, which is _____ the critical value $x^2_{0.05,df} = 9.488$, _____ the null hypothesis.

Step 5: Determine the p-value

p-value $= \text{CHISQ. DIST. RT}($ _____ , ____ $) = $ _____

Because the p-value, _____, is _____ the level of significance, $\alpha = 0.05$, _____ the null hypothesis.

Step 6: State Your Conclusion

5.5 Chi-Square Test of Independence Practice Problems

The following exercises are in the Excel file:
 Chi-Sq Practice Problems.xlsx

Answers to the exercises are in the Excel file:
Chi-Sq Practice Problems KEY.xlsx

5.5.1 Venders and Shipment Quality

The following sample data reflect shipments received by a large firm from three different vendors and the quality of those shipments.

Step 1: Define the Hypotheses

H_0: Quality and source of shipment (vendor) are independent.
H_A: Quality and source of shipment (vendor) are dependent.

Step 2: Calculate the Expected Frequencies

Observed

Vendor	Defective	Acceptable	Total
1	14	112	126
2	10	70	80
3	22	150	172
Total	46	332	378

Expected

Vendor	Defective	Acceptable	TOTAL
1			126
2			80
3			172
TOTAL	46	332	378

Step 3: Calculate the Test Statistic and Degrees of Freedom

Vender	Quality	Obs – Exp	(Obs – Exp)^2	(Obs – Exp)^2/Exp
			Sum	

Degrees of Freedom	
Number of Rows (r)	
Number of Columns (c)	
(r – 1)(c – 1)	

Step 4: Specify the Critical Value and Decision Rule at a 1% Significance Level

=CHISQ.INV.RT(α, df)

Reject H0 if chi-sq test stat > _____.

Step 5: Calculate the p-value.

=CHISQ.DIST.RT(test stat, df)

Step 6: State the Conclusion and answer the question: Should the firm be concerned about the source of the shipment?

1. H0: Quality and source of shipment (vendor) are independent; HA: Quality and source of shipment (vendor) are dependent.
2. 15.33, 9.74, 20.93, 110.67, 70.26, 151.07
3. 0.2; 2
4. Reject H0 if chi-sq test stat > 9.21
5. 0.9048
6. Do not reject H0; there is not enough evidence to support the claim that quality and source of shipment are dependent. The firm should not be concerned about the source of shipments.

5.5.2   Gender and Candidate Preference

In the following table, likely voters' preferences of two candidates are cross-classified by gender.

Step 1: Define the Hypotheses

H_0: Gender and candidate preference are independent.
H_A: Gender and candidate preference are dependent.

Step 2: Calculate the Expected Frequencies

Observed

	Male	Female	Total
Candidate A	155	135	
Candidate B	95	115	
Total			

	Male	Female	Total
Candidate A			
Candidate B			
Total			

Step 3: Calculate the Test Statistic and Degrees of Freedom

$$x^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$df = (r - 1)(c - 1)$$

Step 4: Specify the Critical Value and Decision Rule at a 5% Significance Level

Step 5: Calculate the p-value.

Step 6: State the Conclusion

1. H_0: Gender and candidate preference are independent; H_A: Gender and candidate preference are dependent.
2. 145, 105, 145, 105
3. 3.28; 1
4. Reject H_0 if chi-sq test stat > 3.84
5. 0.0701
6. Do not reject H_0: there is not enough evidence to support the claim that gender and candidate preference are dependent.

5.6 More Chi-Square Test of Independence Practice Problems

The following exercises are in the Excel file:
Chi-Sq Practice Problems.xlsx

Answers to the exercises are in the Excel file:
Chi-Sq Practice Problems KEY.xlsx

5.6.1 Additional Practice Problems

- Gender and Car Color
- Shift and Product Quality
- Race and Seniority
- Age and Income Level
- Income and Credit Score

6 Regression Analysis

6.1 Objectives

- Interpret a scatterplot for linearity requirements
- Calculate and interpret the correlation coefficient
- Estimate the simple linear regression model and interpret the coefficients
- Estimate the multiple regression model and interpret the coefficients
- Calculate predicted values and residuals using the sample regression equation
- Calculate and interpret the standard error of the estimate
- Calculate and interpret the coefficient of determination, R^2
- Differentiate between R^2 and Adjusted R^2

6.2 Example Problem: Simple Linear Regression

ACE Corporation, owner of the 7-Eleven convenience stores chain, completed a marketing test study to identify potential site locations for new stores.

Twenty (20) of its current stores in operation were selected for study. All 20 stores were nearly identical on variables that contribute to sales in a particular location (demographic data on surrounding residential/commercial area, store square footage, amount of parking, and traffic access patterns into and out of the store location). ***Results of this study will be used to identify potential sites for constructing new convenience stores.***

Store n=20	Avg. Daily Auto Traffic (000s)	Annual Sales ($000s)
1	62	1121
2	35	766
3	36	701
4	72	1304
5	41	832
6	39	782
7	49	977
8	25	503
9	41	773
10	39	839
11	35	893
12	27	588
13	55	957
14	38	703
15	24	497
16	28	657
17	53	1209
18	55	997
19	33	844
20	29	883

6.2.1 Scatterplot

Annual Sales by Avg. Daily Auto Traffic

$y = 14.717x + 240.86$
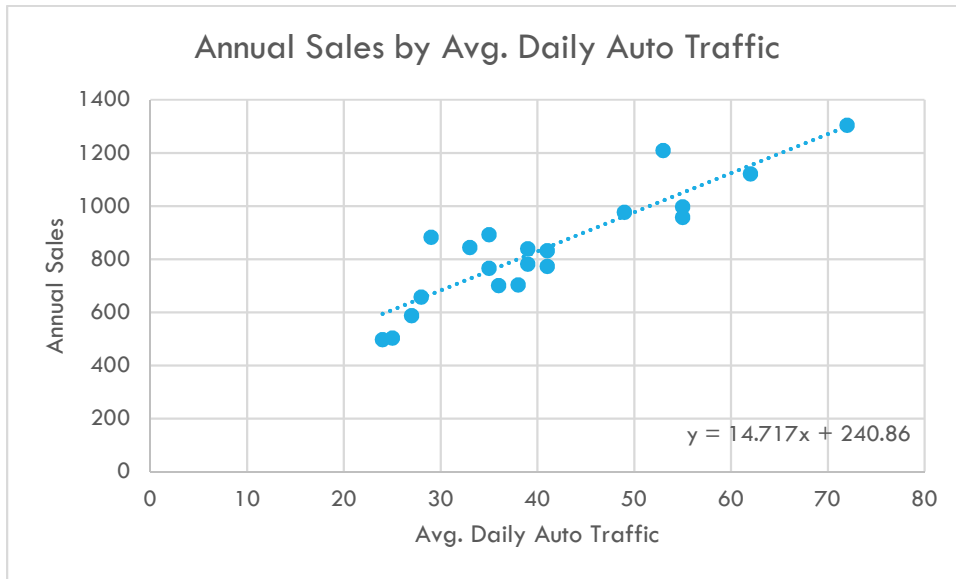
Briefly describe the relationship (form, direction, strength) in terms of the data measured and how it may be useful to management.

6.2.2 Correlation

The necessary summary statistics are given below:

Column	n	Mean	Std. dev.
Avg. Daily Auto Traffic (000s)	20	40.8	13.0449
Annual Sales ($000s)	20	841.3	214.2134

The covariance between Avg. Daily Auto Traffic (000s) and Annual Sales ($000s) is **2504.33**.

r =

$$r = \frac{Covariance}{(s_x)(s_y)} =$$

6.3 Example Problem: Simple Linear Regression

If there is a linear relationship between the value of x and the value of y, then for a given value of x, we can use the regression line to predict the value of y (y-hat).

6.3.1   Variables in a Regression Model

Response variable measures an outcome of study, i.e., the variable you are interested in drawing conclusions about - also called the **outcome variable** or the **dependent variable, y**

Explanatory variable(s) may explain or affect the response variable, i.e., it/they may influence or change the value of the response variable - also called the **independent variable, x**. Sometimes called **predictor** variables because, if the effect on the response variable is strong enough, we can use **x** to predict **y**.

6.3.2   Identify the response and predictor variables in this model:

- Predictor –

- Response –

6.3.3   Simple Linear Regression Equation

Find the equation for the 'best fit' line through the data, $\hat{y} = b_0 + b_1(x)$, where

$$b_1 = r(\frac{s_y}{s_x}) =$$

$$b_0 = \bar{y} - b_1(\bar{x}) =$$

Therefore, the expected (or predicted) value of y for a given value of x is given by the linear equation.

Predicted annual sales $(\hat{y})$ =

6.3.4   What does the slope tell you?

On average, for every unit increase in the predictor, x, the response, y, increases or decreases by <the slope of the line>.

The slope does **not** indicate how strong a relationship is. The correlation indicates the strength of the relationship.

6.3.5   The intercept is not statistically meaningful.

The intercept is necessary to make accurate predictions, however, it has no statistical meaning.

Using the regression line for prediction far outside the range of values (upper and lower values) of the explanatory variable x that you used to obtain the line is called _____.

6.3.6   Facts about Least Squares Regression

1. The distinction between explanatory and response variables is essential.

2. The slope b and correlation r always have the same sign.

3. The least-squares regression line always passes through the point $(\bar{x}, \bar{y})$. This is simply a consequence of calculating the least-squares regression line.

4. The correlation, r, describes the strength of a linear relationship, **not** the slope of the line.

6.3.7   Residuals

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line:

Values of the response, y, that are below the least-squares regression line have negative residuals, i.e., their value is smaller than the corresponding predicted ŷ.

Values of y that are above the line have positive residuals, i.e., their value is greater than the corresponding predicted ŷ.

Use the estimated linear regression equation to predict annual sales for a location that has an average daily auto traffic noted in problems 1 through 3 below and compute the residual error (observed – predicted) of each estimated (predicted) value:

1. x = 35,000 cars
   Predicted annual sales = $_____

   Residual, e = $_____

2. x = 45,000 cars
   Predicted annual sales = $_____

   Residual, e = $_____

3. x = 55,000 cars
   Predicted annual sales = $_____

   Residual, e = $_____; e = $_____

6.3.8   Regression Output for 7-Eleven (Excel)

Summary Output

Regression Statistics	
Multiple R	0.896199726
R Square	0.803173948
Adjusted R Square	0.792239168
Standard Error	97.64015589
Observations	20

Analysis of Variance

ANOVA

	df	SS	MS	F	Significance F
Regression	1	700255.3992	700255.3992	73.45130865	9.06631E-08
Residual	18	171604.8008	9533.600042		
Total	19	871860.2			

Parameter Estimates

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	240.8565508	73.38346911	3.282163595	0.004141128	86.68360315	395.0294984	86.68360315	395.0294984
Avg. Daily Auto Traffic (000s)	14.71675121	1.717165598	8.570373892	9.06631E-08	11.10912016	18.32438226	11.10912016	18.32438226

6.4 Multiple Linear Regression

A multiple linear regression model involves two or more explanatory variables so the effect of multiple explanatory variables on the response variable can be studied.

6.4.1 Multiple Linear Regression Equation

$\hat{y} = $ _____

6.4.2 Residuals

$e = $ _____

6.4.3 Multiple Linear Regression Output

The following linear model is used to predict winning percentages for NFL teams based on points scored and points against…

$Winning\ \widehat{Percentage} = $ _____ $+$ _____$(PointsFor) - $ _____$(PointsAgainst)$

Summary Output

Regression Statistics	
Multiple R	0.940408305
R Square	0.884367781
Adjusted R Square	0.876393145
Standard Error	0.072981621
Observations	32

Analysis of Variance

ANOVA

	df	SS	MS	F	Significance F
Regression	2	1.181351277	0.590675638	110.8975758	2.59803E-14
Residual	29	0.154463192	0.005326317		
Total	31	1.335814469			

Parameter Estimates

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.417223037	0.139448041	2.991960548	0.005610565	0.13201977	0.702426304	0.13201977	0.702426304
PointsFor	0.001766203	0.000186994	9.445224145	2.37046E-10	0.001383757	0.002148649	0.001383757	0.002148649
PointsAgainst	-0.001526842	0.000275065	-5.550836431	5.49585E-06	-0.002089413	-0.00096427	-0.002089413	-0.00096427

6.4.4 Interpret the Slope

The slope coefficients of a multiple regression equation are interpreted differently from how the slope of a simple linear regression equation is interpreted.

6.5 Goodness-of-Fit Measures

How well does the model fit the data? Use the **Standard Error of the Estimate**, the **Coefficient of Determination** (R^2), and/or **Adjusted R^2.**

ANOVA Table of Regression Results

ANOVA					
	df	**Sum of Squares (SS)**	**Mean Square (MS)**	**F**	**Significance F**
Regression	k	$SSR = \sum (\hat{y}_i - \bar{y})^2$	$MSR = {SSR}/{k}$	$F_{df_1, df_2} = \dfrac{MSR}{MSE}$	p-value for the F test of joint significance
Residual	$n - k - 1$	$SSE = \sum (y_i - \hat{y}_i)^2$	$MSE = {SSE}/{(n - k - 1)}$		
Total	$n - 1$	$SST = \sum (y_i - \bar{y})^2$			

k is the number of independent variables
$df_1 = k$ *and* $df_2 = n - k - 1$

ANOVA Table of Simple Linear Regression to Predict Annual Sales from Average Daily Automobile Traffic

ANOVA					
	df	**Sum of Squares (SS)**	**Mean Square (MS)**	**F**	**Significance F**
Regression	1	700255.3992	700255.3992	73.45130865	9.06631E-08
Residual	18	171604.8008	9533.600042		
Total	19	871860.2			

Annual Sales

Unexplained Difference: $(y_i - \hat{y}_i)$, the residual

Total Difference: $(y_i - \bar{y})$

Explained Difference: $(\hat{y}_i - \bar{y})$

1200

1000

800

600

30   40   50   60   70
Avg Daily Auto Traffic
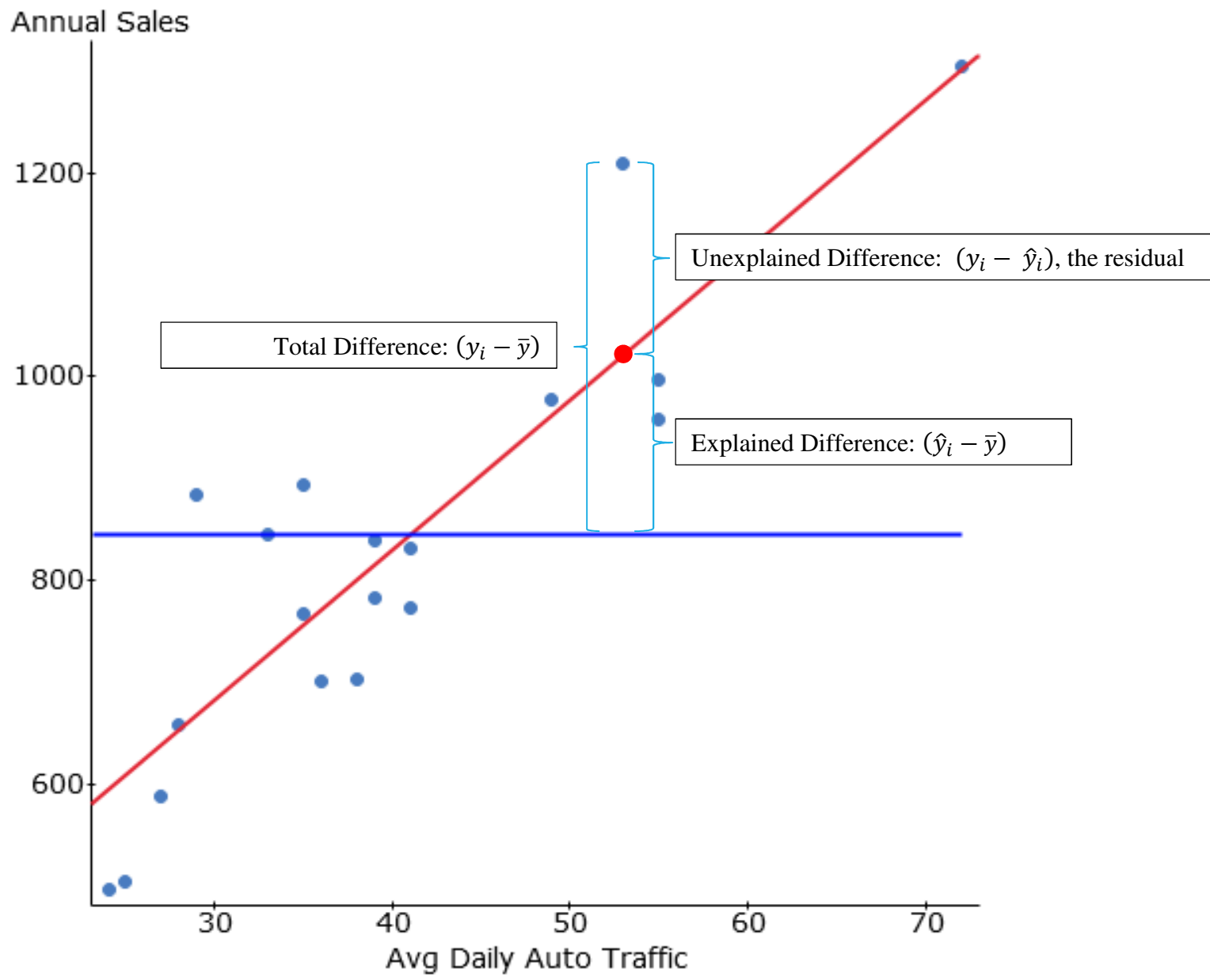
6.6 Standard Error of the Estimate

The MSE, also s_e^2, is the variance of the regression residuals. The square root of the MSE is the **Standard Error of the Estimate**, s_e.

6.7 The Coefficient of Determination, R^2

The **Coefficient of Determination** is the proportion of the sample variation in the response variable that is explained by the sample regression equation.

- R^2 falls between 0 and 1, the closer the value is to 1, the _____.
- R^2 is the square of the correlation coefficient, $r_{y,\hat{y}}$, which is the sample correlation coefficient between y and $\hat{y}$. The value $r_{y,\hat{y}}$ is also called **Multiple R**.
- To interpret R^2, convert the proportion to a percentage and report as follows…

6.8 Adjusted R^2

The Coefficient of Determination never decreases as more variables are added to the model. Also, this measure can increase when variables are added that have no economic or intuitive value in the model. These phenomena make R^2 an inaccurate measure of variability in multiple regression models, i.e., when a model has more than one explanatory variable.

An adjustment is made to R^2 for multiple regression models, and this is referred to as the **Adjusted R^2**.

The interpretation of **Adjusted R^2** is the same as the interpretation of R^2.

6.9 Summary Output for the Simple Linear Regression to Predict Annual Sales from Average Daily Automobile Traffic

Regression Statistics	
Multiple R	0.896199726
R Square	**(1)**
Adjusted R Square	**(2)**
Standard Error	**(3)**
Observations	20

ANOVA

	df	SS	MS	F	Significance F
Regression	1	700255.3992	700255.3992	73.45130865	9.06631E-08
Residual	18	171604.8008	9533.600042		
Total	19	871860.2			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	240.8565508	73.38346911	3.282163595	0.004141128	86.68360315	395.0294984	86.68360315	395.0294984
Avg. Daily Auto Traffic (000s)	14.71675121	1.717165598	8.570373892	9.06631E-08	11.10912016	18.32438226	11.10912016	18.32438226

1.

2.

3.

* Software output reports both R^2 and $Adjusted\ R^2$ for every model you run. You must know when to use one over the other.

6.10 Compare Regression Models

Finding the best fitting model usually requires comparing many different models.

The most basic comparison involves the model $y = \bar{y}$ and a simple linear model $y = b_0 + b_1x$.

MODEL 1	MODEL 2
$y = \bar{y}$	$\hat{y} = b_0 + b_1x$
The Standard Deviation of the Response, s_y	**Standard Error of the Estimate, s_e**
Represents the average distance that the observed values are from the **mean of the response, $\bar{y}$**. This measure tells you how wrong the **model $y = \bar{y}$** is on average using the units of the response variable. Smaller values are better because it indicates that the observations are closer to the **mean**.	Represents the average distance that the observed values are from the **regression line, $\hat{y} = b_0 + b_1x$**. This measure tells you how wrong the **regression model** is on average using the units of the response variable. Smaller values are better because it indicates that the observations are closer to the **fitted line**.
ACE 7-11 Stores Average Annual Sales (000s), $\widehat{Annual\ Sales} = \$841.3$	**ACE 7-11 Stores Linear Model (000s),** $\widehat{Annual\ Sales} = 240.86 + 14.72x$
$s_y = \$_____$	$s_e = \$_____$

$s_e < s_y$, therefore, the model explains more variability in the response than does the mean of the response.

6.11 Guidelines for Comparing Models

Models for Comparison	Criteria for Determining the Best Fitting Model
Simple Linear Regression vs. Mean of the Response Multiple Linear Regression vs. Mean of the Response	• Compare s_e to s_y
Simple Linear Regression vs. Simple Linear Regression	• Compare R^2 of one model to R^2 of the other model • Compare s_e of one model to s_e of the other model
Simple Linear Regression vs. Multiple Linear Regression	• Compare $Adjusted\ R^2$ of one model to $Adjusted\ R^2$ of the other model • Compare s_e of one model to s_e of the other model
Multiple Linear Regression vs. Multiple Linear Regression	• Compare $Adjusted\ R^2$ of one model to $Adjusted\ R^2$ of the other model • Compare s_e of one model to s_e of the other model

When the best fitting model is identified, evaluate R^2 or $Adjusted\ R^2$ for the selected model to determine if the given model explains enough of the variability in the response to justify using the model; this often depends on the context of the data.

[This page intentionally left blank]

7 Conditions for Regression

7.1.1 Condition One

The relationship between the independent variable and each dependent variable is linear.

A scatterplot with the dependent variable on the vertical axis and the independent variable on the horizontal axis reveals the relationship between the two quantitative variables. Data that meets the linearity condition shows a linear trend.

7.1.2 Condition Two

The residuals are independent of one another, i.e., the deviation of one point from the line does not influence the deviation of the others.

7.1.3 Condition Three

The variation of the dependent variable is constant (the same) across all values of the independent variable, which is known as homoscedasticity. If the variance of the error term is not the same for all observations, we cannot conduct test of significance.

A residual plot displays the residuals on the vertical axis and the independent variable on the horizontal axis. Data that meets conditions two and three show no pattern in the residual plot, i.e., the points are scattered randomly above and below zero and in a band of relatively constant width.

7.1.4 Condition Four

The residuals are normally distributed.

A Normal Probability Plot displays the fitted or predicted values against a theoretical normal distribution. Data that meets the normality condition form a straight line. Any curvature in the middle or at either or both ends of the plot indicate the residuals are not normally distributed.

7.1.5 Condition Five

There exists **no** multicollinearity. Multicollinearity means two or more of the predictors in the regression model are highly correlated, i.e., there is a strong linear relationship among two or more explanatory variables. The text uses the phrase "perfect multicollinearity." There can be no perfect multicollinearity.

In multiple regression, when two or more predictors are correlated, the redundancy of data causes unexpected and erroneous results. One or more variables may have an unexpected sign or the model may have insignificant variables with a high R^2. If regression results are unexpected, check for multicollinearity among two or more of the predictors. Remove one of the collinear variables.

7.2 Example Problems: Conditions

7.2.1 Example 1: NFL Winning Percentages Data

The following linear model is used to predict winning percentages for NFL teams based on points scored…

$$Winning \widehat{Percentage} = -0.283 + \ 0.0023(PointsFor)$$

The scatterplot in Figure 1 shows how that both explanatory variables meet the linearity condition for regression analysis.
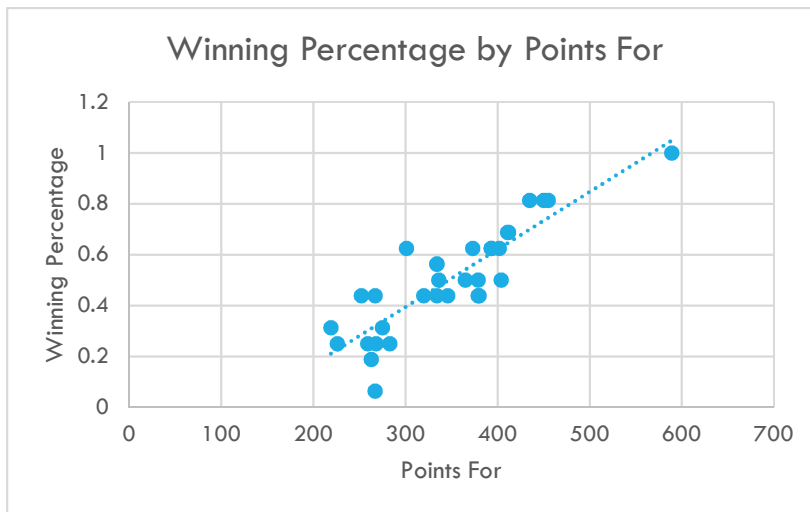
Figure 1 Scatterplot of Winning Percentage by Points For

The residual plot in Figure 2 shows the residuals are roughly evenly dispersed both above and below zero (horizontal axis) and form a band of relatively constant width. This random pattern meets the "no pattern" and "constant variance" conditions.
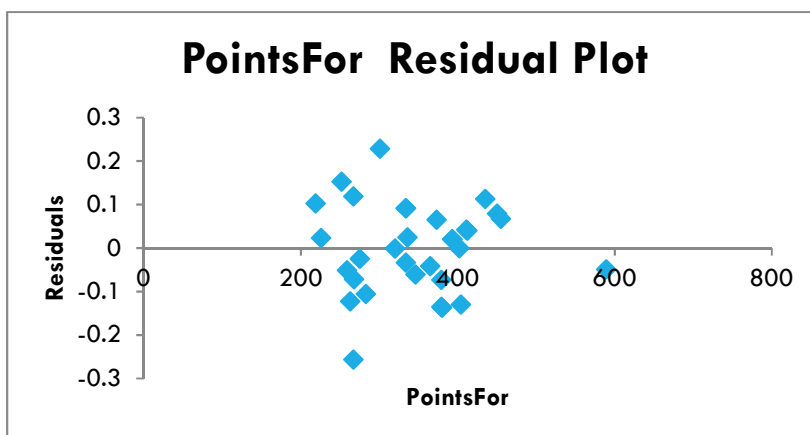
Figure 2 Residual Plot: Residuals Vs. Points For

The relatively straight line in the normal probability plot in Figure 3 below indicates that the residuals are normally distributed. A couple of points deviate from the line.
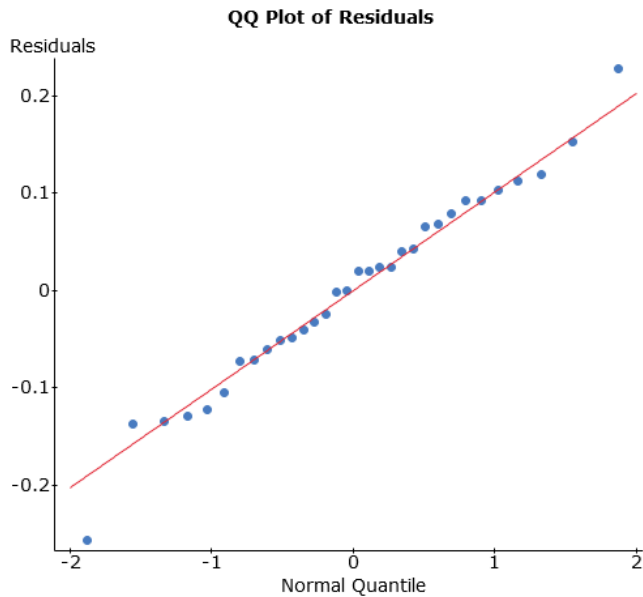
Figure 3 Normal Probability Plot (QQ Plot) for Residuals of Winning Percentage Model

7.2.2   Example 2: Age and Happiness Data

The following linear model **cannot be used** to predict happiness from age…

$$\widehat{happiness} = 56.18 + 0.28(Age)$$

The scatterplot in Figure 4 shows the explanatory variable **does not** meet the linearity condition for regression analysis.
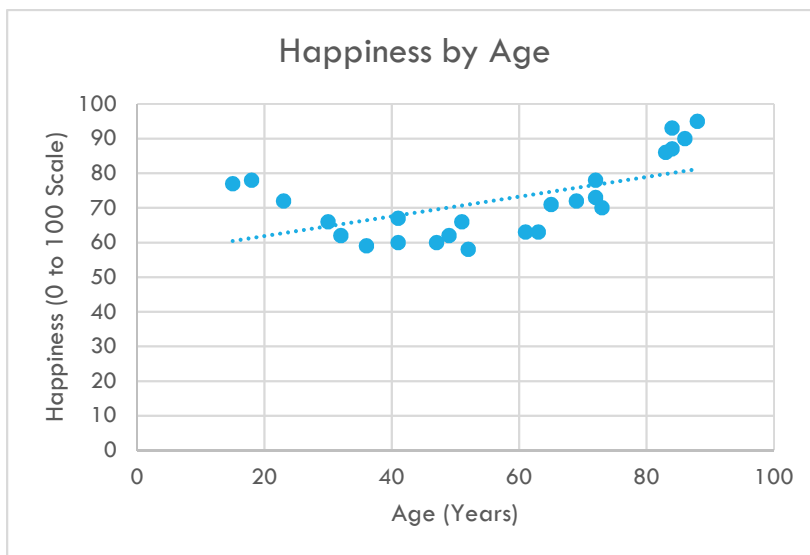
Figure 4 Scatterplot of Happiness by Age

The residuals plot in Figure 5 reveals a pattern that **fails** both the independence and constant variance conditions.
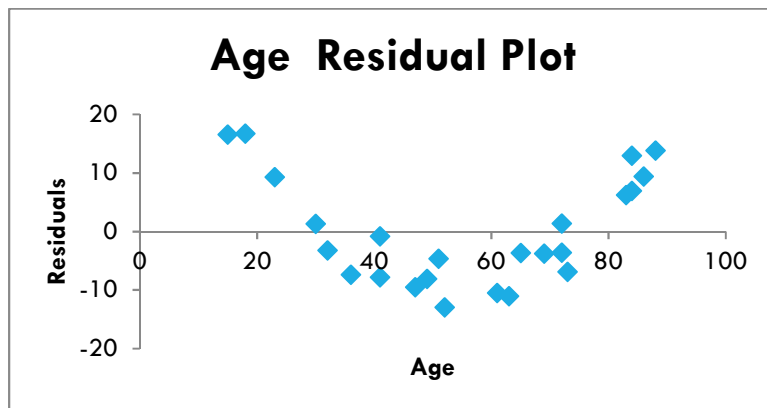
Figure 5 Residual Plot: Residuals Vs. Age

The line in the normal probability plot in Figure 6 below **deviates from the line** indicating that residuals are not normally distributed.
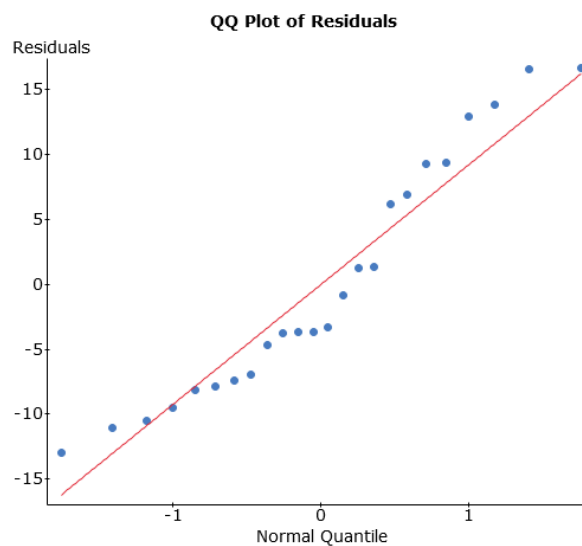
Figure 6 Normal Probability Plot (QQ Plot) for Residuals of Happiness Model

7.2.3 Example 3: Salary and Work Experience Data

The following linear model **cannot be used** to predict salary from work experience…

$$\widehat{salary} = 26234.77 + 3672.21(Experience)$$

The scatterplot in Figure 7 shows the explanatory variable **does not** meet the linearity condition for regression analysis.

Figure 7 Scatterplot of Salary by Work Experience

The residuals plot in Figure 8 reveals a pattern that **fails** both the independence and constant variance conditions.
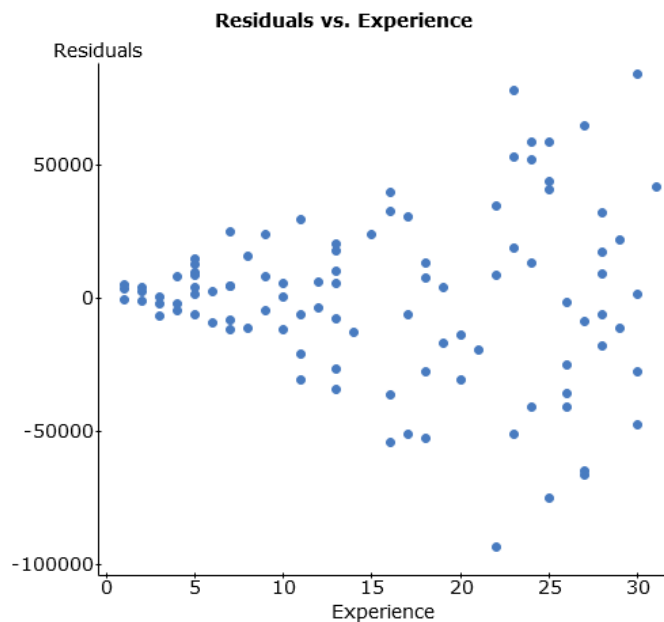
Figure 8 Residual Plot: Residuals Vs. Experience

The line in the normal probability plot in Figure 9 below **deviates from the line** indicating that residuals are not normally distributed.
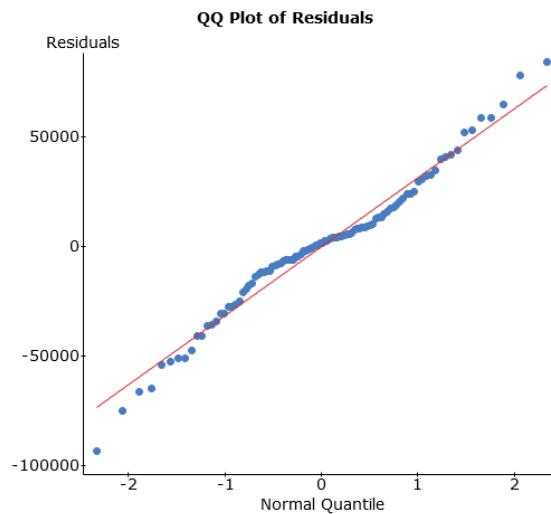
Figure 9 Normal Probability Plot (QQ Plot) for Residuals of Happiness Model

The regression output below shows that a computer software program will run a linear model even if it is not appropriate. The graphs above show that the linear model is **not** appropriate for this data, yet the following results show a correlation of 0.73 and the regression equation.

Regression Statistics	
Multiple R	0.72574951
R Square	0.526712352
Adjusted R Square	0.521882886
Standard Error	31762.56584
Observations	100

ANOVA

	df	SS	MS	F	Significance F
Regression	1	1.10029E+11	1.1E+11	109.0622387	1.32357E-17
Residual	98	98868337695	1.01E+09		
Total	99	2.08897E+11			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	26234.76877	6381.355874	4.111159	8.18802E-05	13571.17572	38898.36182	13571.17572	38898.36182
Experience	3672.213547	351.6338894	10.44329	1.32357E-17	2974.407537	4370.019556	2974.407537	4370.019556

8 Inference in Regression

8.1 Objectives

- Conduct tests of individual significance
- Calculate and interpret confidence intervals for parameter estimates
- Conduct a test of joint significance

8.2 Tests of Individual Significance

If a slope coefficient is zero, $\beta_j = 0$, the explanatory variable x_j, drops out of the equation, i.e., there is not a linear relationship between x_j and y.

If $\beta_j \neq 0$, x_j influences y, i.e., x_j and y are **linearly related**.

To test if a variable influences y, perform a T test on the slope coefficient…

8.2.1 Example 1: Use the P-value to Test Individual Significance

When estimating a multiple linear regression model based on 30 observations, the following results were obtained.

	Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
Intercept	152.27	119.70	1.27	0.2142	−93.34	397.87
x_1	12.91	2.68	4.81	5.06E-05	7.40	18.41
x_2	2.74	2.15	1.28	0.2128	−1.67	7.14

Determine whether x_1 and y are linearly related…

1. State the hypotheses

    H_0:

    H_A:

2. At the 5% significance level, use the p-value approach to determine if x_1 and y linearly related.

$$t_{df} = \frac{b_1 - 0}{se(b_1)} = \frac{b_1}{se(b_1)} = \underline{\hspace{2cm}} = \underline{\hspace{2.5cm}}$$

$$Pvalue = T.DIST.2T(t_{df}, df) = T.DIST.2T(4.81, 27) = \underline{\hspace{1.5cm}}$$
$$= 0.000051$$

$$df = n - k - 1 = 30 - 2 - 1 = 27$$

CONCLUSION:

8.2.2 Example 2: Use a Confidence Interval to Test Individual Significance

When estimating a multiple linear regression model based on 30 observations, the following results were obtained.

	Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
Intercept	152.27	119.70	1.27	0.2142	−93.34	397.87
x_1	12.91	2.68	4.81	5.06E-05	7.40	18.41
x_2	2.74	2.15	1.28	0.2128	−1.67	7.14

Use a 95% Confidence Interval for β_2 to determine if x_2 is significant in explaining y

[_____, _____]

$$b_2 \pm \text{critical value} \times se(b_2)$$

$$t^*_{0.025,27} = T.INV.2T(0.05, 27) = 2.05$$

$$ME = 2.05 \times 2.15 = 4.41$$

$$[2.74 - 4.41, 2.74 + 4.41]$$

$$[-1.67, 7.14]$$

CONCLUSION:

8.3 The Test of Joint Significance (F Statistic)

The Test of Joint Significance is a test of the overall usefulness of a regression model.

$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$

$H_A: At\ least\ one\ \beta_j \neq 0$

$F_{df_1, df_2} = \dfrac{MSR}{MSE}$

where $df_1 = k\ and\ df_2 = n - k - 1$

The ratio of the **mean square regression**, MSR, to the **mean square error**, MSE, follows an F distribution.

ANOVA

	df	SS	MS	F	Significance F
Regression	2	1.181351277	0.590675638	110.8975758	2.59803-14
Residual	29	0.154463192	0.005326317		
Total	31	1.335814469			

CONCLUSION

[This page intentionally left blank]

9   Regression Analysis Practice Problems

The following table shows the number of cars sold last month by six dealers at Centreville Nissan dealership and their number of years of sales experience.

Years of Experience	Sales
1	7
2	9
2	9
4	8
5	14
8	14

1. If a simple linear regression is performed to estimate monthly car sales using the number of years of sales experience, what are the response and explanatory variables?

Response (outcome, independent):

Explanatory (predictor, independent):

The following ANOVA table shows the partial output for a regression to estimate monthly car sales using the number of years of sales experience.

ANOVA

	df	SS	MS	F	Significance F
Regression	1	33.333333	33.333333	9.876543	0.034757
Residual	4	13.5	MSE		
Total	5	46.833333			

2. What is the total sum of squares? What is the error sum of squares? What is the regression sum of squares?

3. Calculate the mean square error (MSE), the variance of the residuals.

4. Calculate the standard error of the estimate (standard deviation of the residuals, the average difference between the observed and predicted response values).

5. Calculate the coefficient of determination, R^2, the proportion of variation in the response variable that is explained by the model (explained variation).

6. Calculate the proportion of unexplained variation in the response variable.

7. Interpret the coefficient of determination.

The following Coefficients table shows the partial output for regression of monthly car sales by years of sales experience.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	6.5	1.3869932	4.6863966	0.0094029	2.6490897	10.3509103	2.6490897	10.3509103
Years of Experience	1	0.3181981	**B₁ t Stat**	0.0347574	0.1165406	1.8834594	0.1165406	1.8834594

8. What are the intercept and slope coefficients, b_0 and b_1, with respect to the problem scenario?

9. Calculate the test statistic for the slope coefficient.

10. What are the null and alternative hypotheses for an individual test of significance to determine if years of sales experience and monthly car sales are linearly related? Test whether the slope is not equal to zero.

11. Use the P-value and a 5% level of significance to evaluate whether the slope coefficient is statistically significant, i.e., are years of sales experience and monthly car sales linearly related at $\alpha=0.05$? State your conclusion and explain.

12. Use the 95% confidence interval to evaluate whether the slope coefficient is statistically significant? State your conclusion and explain.

13. Use the P-value and a 1% level of significance to evaluate whether the slope coefficient is statistically significant? State your conclusion and explain.

14. Write the sample regression equation.

15. Interpret the slope, Years of Experience.

16. Predict monthly car sales for a salesperson with 3 years of experience.

American football is the highest paying sport on a per-game basis. The quarterback, considered the most important player on the team, is appropriately compensated. A sports statistician wants to use 2009 data to estimate a multiple linear regression model that links the quarterback's salary with his pass completion percentage (PCT), total touchdowns scored (TD), and his age. A portion of the data is shown in the accompanying table.

Number	Player	Salary (in $million)	PCT	TD	Age
1	Philip Rivers	25.5566	65.2	28	27
2	Jay Cutler	22.0441	60.5	27	26
3	Eli Manning	20.5000	62.3	27	28
4	Kurt Warner	19.0047	66.1	26	38
5	Matt Schaub	17.0000	67.9	29	28
6	Matt Cassel	15.0052	55	16	27

1. If a multiple linear regression is performed to estimate a quarterback's salary using his PCT, TD, and age, what are the response and explanatory variables?

Response (outcome, independent):

Explanatory (predictor, independent):

122

The following ANOVA table shows the partial output for a regression to estimate a quarterback's salary using his PCT, TD, and age.

ANOVA

	df	SS	MS	F	Significance F
Regression	3	548.34964	182.78321	**F Stat**	0.00262
Residual	28	846.67514	**MSE**		
Total	31	1395.02479			

2. What is the total sum of squares? What is the error sum of squares? What is the regression sum of squares?

3. Calculate the mean square error (MSE), the variance of the residuals.

4. Calculate the standard error of the estimate (standard deviation of the residuals, the average difference between the observed and predicted response values).

5. Calculate the coefficient of determination for multiple linear regression, Adjusted R^2, the proportion of variation in the response variable that is explained by the model (explained variation).

6. Calculate the proportion of unexplained variation in the response variable.

7. Interpret the coefficient of determination.

8. Calculate the F Statistic for a test of joint significance.

9. Write the null and alternative hypotheses for a test of joint significance?

10. What are the degrees of freedom for the F Statistic? What is the critical value at the 5% level of significance?

11. Use the critical value approach to determine if the explanatory variables are jointly significant. Explain.

12. Use the P-value approach to determine if the explanatory variables are jointly significant. Explain.

The following Coefficients table shows the partial output for regression of monthly car sales by years of sales experience.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	32.786679	20.730998	1.581529	0.124989	-9.678845	75.252203	-9.678845	75.252203
PCT	-0.826472	0.438338	B_1 t Stat	0.069782	-1.724366	0.071421	-1.724366	0.071421
TD	0.784974	0.249271	B_2 t Stat	0.003872	0.274365	1.295583	0.274365	1.295583
Age	0.386183	0.253262	B_3 t Stat	0.138517	-0.132601	0.904968	-0.132601	0.904968

13. What are the intercept and slope coefficients, b_0, b_1, b_2, and b_3 with respect to the problem scenario?

14. Calculate the test statistic for the slope coefficients.

15. What are the null and alternative hypotheses for an individual test of significance to determine if quarterback salary and PCT are linearly related? Test whether the PCT coefficient is not equal to zero.

16. Use the P-value and a 5% level of significance to evaluate whether the PCT coefficient is statistically significant, i.e., are quarterback salary and PCT linearly related at $\alpha=0.05$? State your conclusion and explain.

17. Use the 95% confidence interval to evaluate whether the PCT coefficient is statistically significant? State your conclusion and explain.

18. Use the P-value and a 1% level of significance to evaluate whether the PCT coefficient is statistically significant? State your conclusion and explain.

19. Answer questions 3-6 for TD and age.

20. Write the sample regression equation.

21. Predict quarterback salary for a player with PCT=60, TD=30, and age=26.

22. Interpret the slopes, PCT, TD, and age.

Simple Linear Regression – Monthly Car Sales by Years of Experience

1. Response: Monthly Car Sales; Explanatory: Years of Experience
2. SST=46.83; SSE=13.5; SSR=33.3
3. $MSE = \frac{SSE}{n-k-1} = \frac{13.5}{4} = 3.375$
4. $se = \sqrt{MSE} = \sqrt{3.375} = 1.837$
5. $R^2 = \frac{SSR}{SST} = \frac{33.33}{46.83} = 0.7117$
6. Unexplained variance is $\frac{SSE}{SST} = \frac{13.5}{46.83} = 0.2883$ (Note this is $1 - R^2$)
7. On average, 71.2% of the variation in monthly car sales is explained by years of experience.
8. $b_0 = 6.5;\ b_1 = 1$
9. $t_{df} = \frac{b_1 - 0}{se(b_1)} = \frac{b_1}{se(b_1)} = \frac{1}{0.3182} = 3.14$
10. $H_0: \beta_1 = 0;\ H_A: \beta_1 \neq 0$
11. Reject the null. Monthly sales and years of experience are linearly related with a p-value of 0.03 which is less than 0.05.
12. The interval [0.1165, 1.8835] does not contain zero so the slope coefficient for years of experience is likely not equal to zero.
13. Reject the null. Monthly sales and years of experience are linearly related with a p-value of 0.03 which is greater than 0.01.
14. $\hat{y} = 6.5 + 1 * Years\ of\ Experience$
15. For every one year increase in years of experience, monthly car sales increase by 1 car.
16. $\hat{y} = 9.5$

Multiple Linear Regression – Quarterback's Salary by PCT, TD, and Age

1. Response: Quarterback's Salary; Explanatory: PCT, TD, and Age
2. SST=1395.02; SSE=846.68; SSR=548.35
3. $MSE = \frac{SSE}{n-k-1} = \frac{846.68}{28} = 30.24$
4. $se = \sqrt{MSE} = \sqrt{30.24} = 5.5$
5. $R^2 = \frac{SSR}{SST} = \frac{548.35}{1395.02} = 0.39;\ Adjusted\ R^2 = 1 - (1-R^2)\left(\frac{n-1}{n-k-1}\right) = 1 - (1-0.39)\left(\frac{31}{28}\right) = 0.3246$
6. Unexplained variance is $1 - $ Adjusted $R^2 = 0.6754$
7. On average, 36.8% of the variation in quarterback's salary is explained by the model with PCT, TD, and age.
8. $F_{3,28} = \frac{MSR}{MSE} = \frac{182.78}{30.24} = 6.04$
9. $H_0: \beta_1 = \beta_2 = \beta_3 = 0;\ H_A: At\ least\ one\ coefficient\ does\ not\ equal\ zero$
10. $F^*_{3,28} = 2.95$
11. Reject the null. At least one coefficient does not equal zero.
12. Reject the null. At least one coefficient does not equal zero.
13. $b_0 = 32.79;\ b_1 = -0.83;\ b_2 = 0.78;\ b_3 = 0.39$
14. $t_{df} = \frac{b_1 - 0}{se(b_1)} = \frac{b_1}{se(b_1)} = \frac{-0.83}{0.4383} = 1.89;\ t_{df} = \frac{b_1 - 0}{se(b_1)} = \frac{b_1}{se(b_1)} = \frac{0.78}{0.2493} = 3.13;\ t_{df} = \frac{b_1 - 0}{se(b_1)} = \frac{b_1}{se(b_1)} = \frac{0.39}{0.2533} = 1.54$
15. $H_0: \beta_1 = 0;\ H_A: \beta_1 \neq 0$
16. Do not reject the null. Quarterback's salary and PCT are not linearly related with a p-value of 0.06 which is greater than 0.05.
17. The interval [-1.72, 0.07] contains zero so the slope coefficient for PCT is likely equal to zero.
18. Do not reject the null. Quarterback's salary and PCT are not linearly related with a p-value of 0.06 which is greater than 0.01.
19. TD – $H_0: \beta_2 = 0;\ H_A: \beta_2 \neq 0$; Reject $H_0: \beta_2 = 0$; The interval [0.27, 1.3] does not contain zero.
Age – $H_0: \beta_3 = 0;\ H_A: \beta_3 \neq 0$; Do not reject $H_0: \beta_3 = 0$; The interval [-0.13, 0.9] contains zero.
20. $\hat{y} = 32.79 - 0.83 * PCT + 0.78 * TD + 0.39 * Age$
21. $\hat{y} = 16.53\ million$
22. For every unit increase in PCT, quarterback's salary decreases by $0.83 million.
For every unit increase in TD, quarterback's salary increases by $0.78 million.
For every year increase in age, quarterback's salary increases by $0.39 million.

[This page intentionally left blank]

10 Dummy Variables in Regression

10.1 Objectives

- Use dummy variables to represent qualitative explanatory variables.
- Test the differences between the categories of a qualitative variable.

10.2 Regression with Qualitative Variables

The regression models in the previous sections have involved one or more explanatory variables that are quantitative, e.g., income, years of experience, GPA, average daily automobile traffic, etc.

The explanatory variables used in a regression can be quantitative or qualitative. When qualitative variables are used in a regression model, one or more dummy variables are assigned to represent the categories.

10.2.1 Qualitative Variables with Two Categories

When a qualitative variable has two categories, one dummy variable that takes on either a value of 0 or 1 is used where 0 represent one category and 1 represents the other.

The following data includes a partial dataset with professor Salary ($1000s), Exper (experience in years), Gender (male or female), Age (under or over 65), GenderD (0 for females and 1 for males), and AgeD (0 for under 60 and 1 for over 60).

The sample regression equation for predicting professor salary from years of experience, gender, and age is $\hat{y} = b_0 + b_1x + b_2d_1 + b_3d_2$.

Individual	Salary	Exper	Gender	Age	GenderD	AgeD
1	67.50	14	Male	Under	1	0
2	53.51	6	Male	Under	1	0
3	50.05	2	Female	Under	0	0
4	111.88	34	Male	Over	1	1
5	63.68	21	Male	Under	1	0
6	75.56	35	Female	Over	0	1
7	65.50	14	Male	Under	1	0
8	61.04	3	Male	Under	1	0
9	64.30	20	Female	Under	0	0
10	58.38	4	Male	Under	1	0
11	88.29	28	Male	Under	1	0
12	88.61	26	Male	Under	1	0
13	52.23	6	Female	Under	0	0
14	59.65	16	Female	Under	0	0
15	63.49	22	Female	Under	0	0

10.2.2 Tests of Individual Significance with Dummy Variables

In the sample regression equation, $\hat{y} = b_0 + b_1 x + b_2 d_1 + b_3 d_2$, tests of individual significance can be performed on dummy variables in the same way these tests are performed on quantitative variables. If the coefficient is significant, i.e., it is not zero, then there is a difference in the predicted value between the two categories of the qualitative variable.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	40.6060	3.6919	10.9987	0.0000	33.1322	48.0799
Exper	1.1279	0.1790	6.3000	0.0000	0.7655	1.4904
GenderD	13.9240	2.8667	4.8572	0.0000	8.1207	19.7272
AgeD	4.3428	4.6436	0.9352	0.3556	-5.0577	13.7434

According to the regression output for the professor salary data…

-

-

-

The interpretation of dummy variable coefficients is different from the interpretation of quantitative variable coefficients.

10.2.3 Interpret the Dummy Variable Coefficients

In the sample regression equation, $\widehat{professor\ salary} = 40.61 + 1.13(Exper) + 13.92 d_1 + 4.34 d_2$, the interpretation of the coefficient for the quantitative variable, b_1, is "For every year increase in experience, professor salary increases by \$1,130." The interpretation of dummy variable coefficients is different.

Interpret b_2 and b_3

Categories	b_2	b_3	Equation
Male professor under 60	1	0	
Male professor over 60	1	1	
Female professor under 60	0	0	
Female professor over 60	0	1	

The slope of the line does not change. Only the intercept changes when b_2 or b_3 change. The above equations are four parallel lines distinguished by the intercept and the combination of gender and age categories.

10.2.4 Test of Joint Significance, R^2, s_e with Dummy Variables

The test of joint significance and the calculation and interpretation of the coefficient of determination and standard error of the estimate are the same as with a sample regression equation that includes only quantitative variables.

10.2.5 Qualitative Variables with More then Two Categories

Qualitative variables may be defined by more than two categories. In such cases, use multiple dummy variables to capture all categories. The number of dummy variables representing a qualitative variable should be **one less than the number of categories** of the variable.

The following data includes a partial dataset with SAT scores and a four-category qualitative variable, Ethnicity. The categories include White, Black, Asian, and Hispanic. Note that Hispanic is not listed as a column in the table. The rows that have all zeros (colored gray) are the subjects who are Hispanic, i.e., White=0, Black=0, and Asian=0. This is called the reference category because the intercept is b_0 alone. Each of the other three categories are compared against this reference category.

SAT	White	Black	Asian
1515	1	0	0
1530	0	0	0
1524	1	0	0
1490	1	0	0
1724	0	0	1
1582	0	0	1
1544	0	0	1
1850	0	0	1
1373	0	0	0
1594	1	0	0
1595	1	0	0
1720	1	0	0
1389	0	0	0
1363	0	0	0

Interpret b_1, b_2, and b_3

Categories	b_1	b_2	b_3	Equation
White	1	0	0	
Black	0	1	0	
Asian	0	0	1	
Hispanic	0	0	0	

With Hispanic as the reference category, each of the other categories is compared to Hispanic.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	1388.8919	9.3567	148.4384	0.0000	1370.4392	1407.3446
White	201.1447	12.9056	15.5859	0.0000	175.6931	226.5963
Black	-31.4544	22.1913	-1.4174	0.1579	-75.2188	12.3101
Asian	264.8581	17.8584	14.8310	0.0000	229.6388	300.0775

-

-

-

With Asian as the reference category, each of the other categories is compared to Asian.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	1653.75	15.2110	108.7203	0.0000	1623.7517	1683.7483
White	-63.7134	17.6177	-3.6164	0.0004	-98.4580	-28.9689
Black	-296.3125	25.2247	-11.7469	0.0000	-346.0591	-246.5659
Hispanic	-264.8581	17.8584	-14.8310	0.0000	-300.0775	-229.6388

-

-

-

11 Excel: Data Analysis Toolpak

1. Open Excel → File → Options
   → Add-Ins (On a Mac, go to
   Tools→ Add-Ins)

2. On the Manage drop down menu
   at the bottom of the Excel
   Options dialog box, select Excel
   Add-ins → click the Go…
   button
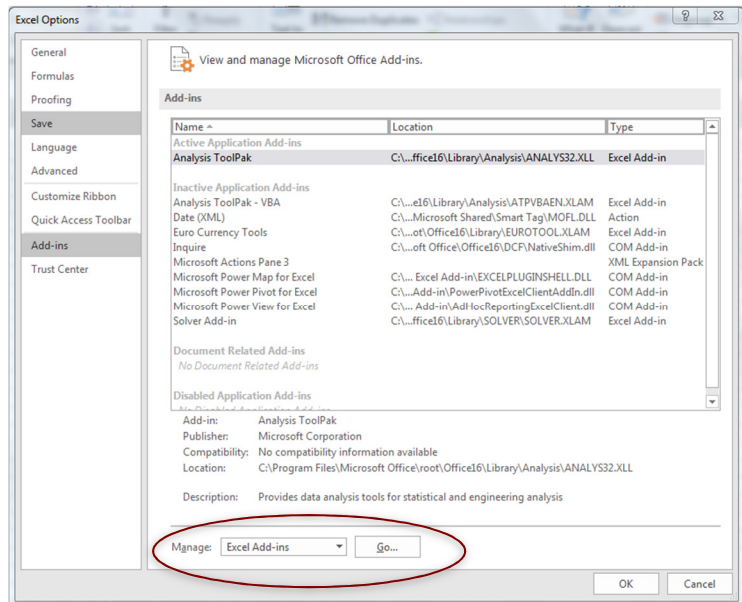
3. In the Add-ins dialog box, check Analysis ToolPak →
   click OK
4. The Data Analysis options appear on the far right of the
   Data tab
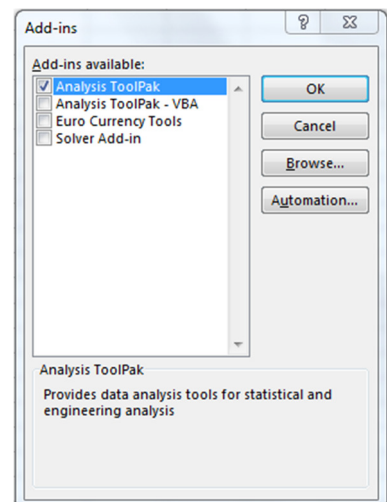
[This page intentionally left blank]

12 Summary of Statistical Definitions and Formulas with Excel Functions

12.1 Binomial and Normal Probabilities

Measure	Formula	Excel Formula
Binomial Probability $X \sim \text{Bin}(n, p)$	$P(X = k) = \ nCk \ \times p^k \\ \qquad\qquad \times (1 - p)^{n-k}$	=BINOM.DIST(k, n, p, 1) → cumulative =BINOM.DIST(k, n, p, 0) → exact
Standard Normal $Z \sim N(0, 1)$	$z = \dfrac{x - \mu}{\sigma}$	=NORM.S.DIST(z, 1) =NORM.S.INV(probability) → Reverse Lookup
Normal $X \sim N(\mu, \sigma^2)$	$x = z * \sigma + \mu$	=NORM.DIST(x, mean, standard deviation, 1) =NORM.INV(probability, mean, standard deviation) → Reverse Lookup

12.2 Sampling Distributions

Measure	Formula for Means	Formula for Proportions
Expected Value	$EX(\bar{X}) = \mu_{\bar{X}} = \mu$	$EX(\bar{P}) = \mu_{\bar{p}} = p$
Standard Error	$se(\bar{X}) = \sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$	$se(\bar{P}) = \sigma_{\bar{p}} = \sqrt{\dfrac{p(1-p)}{n}}$
Z, standard normal value when σ is known	$Z = \dfrac{\bar{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$	$Z = \dfrac{\bar{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$

Measure	Excel Formula
Standard Normal: Z scores and Probability	=NORM.S.DIST(z, 1) =NORM.S.INV(probability)

12.3 Confidence Intervals for One Sample

Measure	Formula for Means Sigma, σ_X, is Unknown	Formula for Proportions
Standard Error	$se(\bar{X}) = \dfrac{s}{\sqrt{n}}$	$se(\bar{P}) = \sqrt{\dfrac{\bar{p}(1-\bar{p})}{n}}$
Critical Value	$t^*_{\frac{\alpha}{2}, df}$	$z^*_{\frac{\alpha}{2}}$
Degrees of Freedom	$df = n - 1$	N/A
Margin of Error (ME)	$ME = t^*_{\frac{\alpha}{2}, df} \times se(\bar{X})$	$ME = z^*_{\frac{\alpha}{2}} \times se(\bar{P})$

Measure	Excel Formula
Standard Normal: Critical Values	=NORM.S.DIST(z, 1) =NORM.S.INV$\left(\dfrac{\alpha}{2}\right)$
T Distribution: Critical Values	=T.INV.2T(α, df) =T.INV($\dfrac{\alpha}{2}$, df)

12.4 Hypothesis Testing for One Sample

Measure	Formula for Means Sigma, σ_X, is Unknown	Formula for Proportions
Critical Value	$t^*_{\frac{\alpha}{2},\,df}$ for two-tailed hypothesis tests $t^*_{\alpha,\,df}$ for one-tailed hypothesis tests	$z^*_{\frac{\alpha}{2}}$ for two-tailed hypothesis tests z^*_{α} for one-tailed hypothesis tests
Test Statistic	$$t_{df} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$	$$z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$
Degrees of Freedom	$df = n - 1$	N/A

Measure	Excel Formula
Standard Normal: P-value and Critical Values	=NORM.S.DIST(z, 1) =NORM.S.INV$\left(\frac{\alpha}{2}\right)$ or =NORM.S.INV(α)
T Distribution: P-value and Critical Values	=T.DIST(t, df, 1) =T.DIST.2T(t, df) =T.DIST.RT(t, df) =T.INV.2T$(\alpha$, df) =T.INV$(\frac{\alpha}{2}$, df) or =T.INV$(\alpha$, df)

12.5 Confidence Intervals for the Difference Between Two Populations

Measure	Formula for Two Means	Formula for Two Proportions
Point Estimate	$\bar{x}_1 - \bar{x}_2$	$\bar{p}_1 - \bar{p}_2$
Standard Error	$se(\bar{X}_1 - \bar{X}_2) = \sqrt{s_p^2 \left(\dfrac{1}{n_1} + \dfrac{1}{n_2} \right)}$ $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$	$se(\bar{P}_1 - \bar{P}_2) = \sqrt{\dfrac{\bar{p}_1(1 - \bar{p}_1)}{n_1} + \dfrac{\bar{p}_2(1 - \bar{p}_2)}{n_2}}$
Critical Value	$t^*_{\frac{\alpha}{2}, df}$	$z^*_{\frac{\alpha}{2}}$
Degrees of Freedom	$df = (n_1 - 1) + (n_2 - 1)$	N/A
Margin of Error (ME)	$ME = t^*_{\frac{\alpha}{2}, df} \times se(\bar{X}_1 - \bar{X}_2)$	$ME = z^*_{\frac{\alpha}{2}} \times se(\bar{P}_1 - \bar{P}_2)$

Measure	Excel Formula
Standard Normal: Critical Values	=NORM.S.DIST(z, 1) =NORM.S.INV$\left(\dfrac{\alpha}{2}\right)$
T Distribution: Critical Values	=T.INV.2T(α, df) =T.INV($\dfrac{\alpha}{2}$, df)

12.6 Testing the Difference Between Two Populations

Measure	Formula for Two Means	Formula for Two Proportions
Critical Value	$t^*_{\frac{\alpha}{2}, df}$ for two-tailed hypothesis tests $t^*_{\alpha, df}$ for one-tailed hypothesis tests	$z^*_{\frac{\alpha}{2}}$ for two-tailed hypothesis tests z^*_{α} for one-tailed hypothesis tests
Test Statistic	$$t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$ $$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$	$$z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)_0}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$ $$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$
Degrees of Freedom	$$df = (n_1 - 1) + (n_2 - 1)$$	N/A

Measure	Excel Formula
Standard Normal: P-value and Critical Values	=NORM.S.DIST(z, 1) =NORM.S.INV$\left(\frac{\alpha}{2}\right)$ or =NORM.S.INV(α)
T Distribution: P-value and Critical Values	=T.DIST(t, df, 1) =T.DIST.2T(t, df) =T.DIST.RT(t, df) =T.INV.2T$(\alpha$, df) =T.INV($\frac{\alpha}{2}$, df) or =T.INV$(\alpha$, df)

12.7 Chi Square Test of Independence

Measure	Formulas
Critical Value	$\chi^{2*}_{\alpha, df}$ for right-tailed hypothesis test
Expected Frequency, f_e	$$f_e = \frac{(row\ total)(column\ total)}{sample\ size}$$
Test Statistic	$$\chi^2_{df} = \sum \frac{(f_o - f_e)^2}{f_e}$$
Degrees of Freedom	$$df = (r-1) \times (c-1)$$ where r=number of rows and c=number of columns

Measure	Excel Formula
Chi Square Distribution: P-value and Critical Values	$= \text{CHISQ. DIST. RT}(\chi^2, df)$ $= \text{CHISQ. INV. RT}(\alpha, df)$

12.8 Simple Linear Regression

Measure	Formula
Correlation	$r = \dfrac{covariance}{s_x s_y}$
Slope	$b_1 = r\left(\dfrac{s_y}{s_x}\right)$
Intercept	$b_0 = \bar{y} - b_1 \times \bar{x}$
Regression Equation	$\hat{y} = b_0 + b_1(x)$
Residual (observed – predicted)	$e = y_i - \hat{y}_i$

Measure	Excel Formula
Covariance	=COVARIANCE.S(<data range>, <data range>)
Correlation	=CORREL(<data range>, <data range>)
Slope	=SLOPE(<y data range>, <x data range>)

12.9 Multiple Linear Regression

Measure	Formula
Regression Equation	$\hat{y} = b_0 + b_1(x_1) + b_2(x_2) + b_3(x_3) \cdots b_k(x_k)$
Residual (observed – predicted)	$e = y_i - \hat{y}_i$
Adjusted R^2	$Adjusted\ R^2 = 1 - (1 - R^2)\left(\dfrac{n-1}{n-k-1}\right)$

12.10 Dummy Variables with Regression

Measure	Formula
Regression Equation	$\hat{y} = b_0 + b_1(x) + b_2(d_1) + b_3(d_2)*$
Residual (observed – predicted)	$e = y_i - \hat{y}_i$
Reference Category	The category where the value of all the dummy variables (that represent one qualitative variable) are zero. Each category is compared against this category.

*The number of dummy variables depends on the number of qualitative variables and the number of categories represented by each qualitative variable. The number of dummy variables used to represent each variable must be one less than the number of categories.

[This page intentionally left blank]
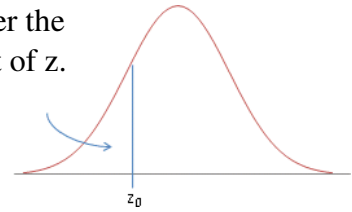
13 Z Table

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
-3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

14 Student's T Table

	0.20	0.10	0.05	0.02	0.01
Two-tailed test	0.20	0.10	0.05	0.02	0.01
One-tailed test	0.10	0.05	0.025	0.01	0.005
df					
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
Confidence	80%	90%	95%	98%	99%

	Significance Level				
Two-tailed test	0.20	0.10	0.05	0.02	0.01
One-tailed test	0.10	0.05	0.025	0.01	0.005
df					
32	1.309	1.694	2.037	2.449	2.738
34	1.307	1.691	2.032	2.441	2.728
36	1.306	1.688	2.028	2.434	2.719
38	1.304	1.686	2.024	2.429	2.712
40	1.303	1.684	2.021	2.423	2.704
42	1.302	1.682	2.018	2.418	2.698
44	1.301	1.680	2.015	2.414	2.692
46	1.300	1.679	2.013	2.410	2.687
48	1.299	1.677	2.011	2.407	2.682
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
70	1.294	1.667	1.994	2.381	2.648
80	1.292	1.664	1.990	2.374	2.639
90	1.291	1.662	1.987	2.368	2.632
100	1.290	1.660	1.984	2.364	2.626
120	1.289	1.658	1.980	2.358	2.617
150	1.287	1.655	1.976	2.351	2.609
200	1.286	1.653	1.972	2.345	2.601
300	1.284	1.650	1.968	2.339	2.592
400	1.284	1.649	1.966	2.336	2.588
500	1.283	1.648	1.965	2.334	2.586
600	1.283	1.647	1.964	2.333	2.584
z	1.282	1.645	1.960	2.326	2.576
Confidence	80%	90%	95%	98%	99%

15 Chi Square Critical Values
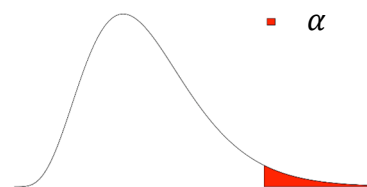
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169