Chapter 11 Comparing Two Population Parameters

Section 11.1A/11.1 Comparing Two Population Proportions

Objectives

❶Distinguish between independent and dependent sampling
Use randomization to compare two population proportions
❷Test hypotheses regarding two proportions from independent samples
❸Construct and interpret confidence intervals for the difference between two population proportions
❹Determine the sample size necessary for estimating the difference between two population proportions

❶ Distinguish between Independent and Dependent Sampling

A sampling method is **independent** when an individual selected for one sample does not dictate which individual is to be in a second sample. A sampling method is **dependent** when an individual selected to be in one sample is used to determine the individual in the second sample. Dependent samples are often referred to as **matched-pairs** samples. It is possible for an individual to be matched against him- or herself.

Example Distinguish between Independent and Dependent Sampling

For each of the following, determine whether the sampling method is independent or dependent.

A researcher wants to know whether the price of a one night stay at a Holiday Inn Express is less than the price of a one night stay at a Red Roof Inn. She randomly selects 8 towns where the location of the hotels is close to each other and determines the price of a one night stay.

Do a higher proportion of men or women vote in presidential elections?

Use Randomization to Compare Two Population Proportions

In Chapter 10, we focused on determining whether a sample may come from a population with a specified characteristic (such as a specified proportion of the population with a given characteristic, or a specified population mean).

Now, we shift gears to comparing the characteristics of two groups. Comparison of two groups may occur through either observational studies (such as comparing the proportion of Republicans to the proportion of Democrats in favor of some policy) or through designed experiments. Recall in a completely randomized design, a group of individuals is randomly assigned to two or more treatment groups, the treatment is imposed on the individuals, and a response variable is measured. The methods of this section apply where there are two levels of the treatment (or two distinct groups) and the response variable is qualitative with two possible outcomes.

To help understand the logic behind the approach that may be used to compare two groups, we lay out a scenario. A "flipped" classroom is one in which students learn material through video and other activities at home, while class-time is dedicated to group learning and peer-to-peer instruction. Professor Lyne wanted to know if Introductory Statistics students taught using a "flipped" classroom resulted in a higher proportion of students passing over her traditional lecture-based model of teaching. One semester, she taught one section of Introductory Statistics using a "flipped" model and one section using a traditional lecture model. At the end of the semester, she recorded whether a student passed or did not pass the class. Any student who withdrew was recorded as a student who did not pass. Results are shown in Table 1 below.

Table 1

	Flipped Course	**Traditional Course**	**Total**
Pass	25	21	46
Did Not Pass	7	11	18
Total	32	32	64

Does this evidence suggest the flipped course is superior to the traditional course as measured by pass rates?

Notice that the response variable in the study is whether the student passed, or not. This is a qualitative variable with two possible outcomes. There are two courses being compared – the flipped course or traditional course. The experimental units are the 64 students. The assumption is that the students randomly selected one of the two courses and were similar in terms of academic ability, attendance patterns, and other variables that may affect whether a student passes, or not prior to the course beginning. One variable that potentially could confound the results is the time the class meets. Professor Lyne, obviously, cannot teach each course at the same time. Even if one course is offered MWF at 8 am and the other at 9 am, class time could be a variable that affects the response variable. Unfortunately, there is nothing we can do about this (other than acknowledge the potential shortcoming of the study).

In this study, the proportion of students in the flipped course who passed is

$$\hat{p}_F = \frac{25}{32} = 0.781$$

and the proportion of students in the traditional course who passed is

$$\hat{p}_T = \frac{21}{32} = 0.656$$

The difference in sample proportions is 0.781 – 0.656 = 0.125. The questions we ask are:

(1) Is the difference in sample proportions statistically significant, suggesting the flipped course has a higher pass rate than the traditional course?
(2) Or, is it possible that the difference is due to random chance and there is no difference in the pass rates of the two courses?" In other words, is it possible the 46

students who passed would have passed regardless of which course they were enrolled in, and a higher proportion of these students just happened to enroll in the flipped course?

There are two possibilities here:

(1) The flipped course is not more effective and the higher pass rate in the flipped course was due to random chance. That is, the proportion of students who pass the flipped course equals the proportion who pass the traditional course.
(2) The flipped course is more effective and this explains the difference in pass rates. That is, the proportion of students who pass the flipped course is greater than the proportion of students who pass the traditional course.

We can state these two possibilities using the notation of hypothesis tests.

$$H_0 : p_F = p_T \quad \text{This is Statement (1)}$$
$$H_1 : p_F > p_T \quad \text{This is Statement (2)}$$

We could also write each of these statements as a difference in proportions:

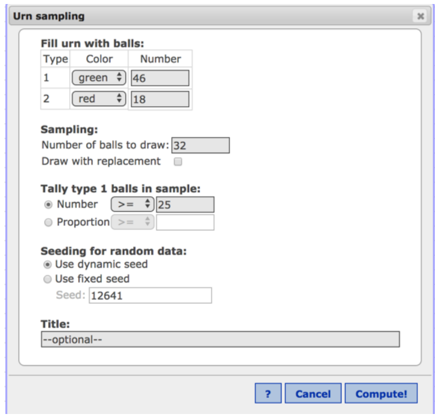
$$H_0 : p_F - p_T = 0 \quad \text{This is Statement (1)}$$
$$H_1 : p_F - p_T > 0 \quad \text{This is Statement (2)}$$

To answer the questions posed, assume that Statement (1), the null hypothesis, is true because this is the statement of "no effect" or "no difference". Use this statement to build the null model. To develop a conceptual understanding for building the null model, we use an urn. Let 46 green balls represent the 46 students who passed the course and let 18 red balls represent the 18 students who did not pass. Mix the 64 balls in the urn and randomly choose 32 balls. These 32 balls will represent the 32 students who enrolled in the flipped course. Notice that this random assignment is done under the assumption the statement in the null hypothesis is true because each ball has an equally likely chance of going to the flipped course or the traditional course. Note: We are not saying that passing versus not passing is equally likely, just that the likelihood of a "passing" student going to the flipped course is the same as that student going to the traditional course.

Rather than physically working with an urn, we can use StatCrunch. Figure 1 shows the dialogue box to build the urn applet in StatCrunch. Notice that we are drawing 32 balls (the 32 students). In addition, we want the applet to record how often we observe 25 or more green balls. This is analogous to determining how many of the students in the flipped course pass. The value of 25 serves as the **test statistic** for this hypothesis test. Note that we could also have used 0.781, the sample proportion of students who passed in the flipped course, as the test statistic.

Click "1 run". How many of the students passed the flipped course? Click "1 run" a second time. How many of the students passed the flipped course?

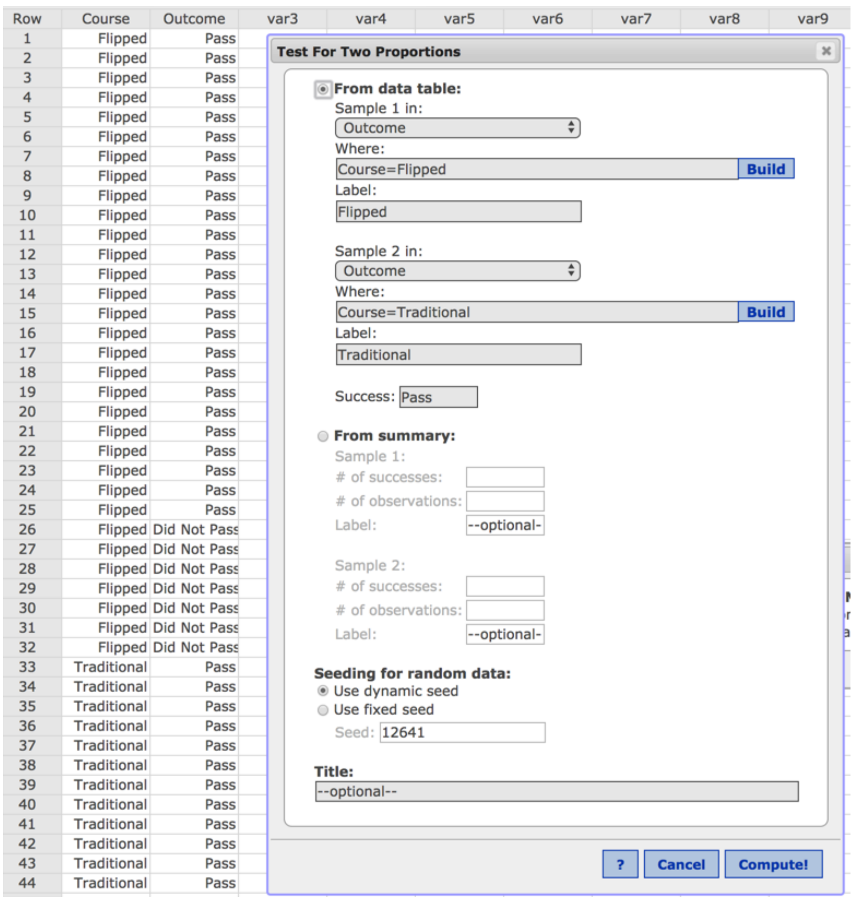
Figure 1



To determine a long-run pattern of outcomes, we can repeat this random assignment many times by clicking "1000 runs" in the applet. Click "1000 runs" five times (for a total of 5002 random assignments). What proportion of the random assignments resulted in 25 or more flipped students passing (assuming there is no difference in the pass rates of the two courses)? What do you conclude?

An urn applet is not the only way to randomly assign individuals to estimate a *P*-value. We could also use the “Randomization test for two proportions” applet in StatCrunch. To use this applet, you may either have raw data from a data table or summary statistics. Figure 3 shows the raw data from Professor Lyne’s study entered into the StatCrunch spreadsheet. The file is Lyne in the SullyStats group. In StatCrunch, select Applets > Resampling > Randomization test for two proportions. Figure 3 also shows how to enter results into the dialogue box. Now click Compute!.

FIGURE 3



In the applet, click “1000 times” five times (for a total of 5000 random assignments). Notice that this applet uses the difference of sample proportions, 0.125, as the test statistic. In the 5000 runs, how many resulted in a difference in sample proportions of 0.125 or higher? What is the *P*-value? Is it close to the results of the urn? What do you conclude?

Notice the basic shape of the outcomes. Also, notice the difference in proportions is centered at 0, which is what we would expect if the population proportions are equal. The dispersion about 0 is due to the nature of random assignment – the dispersion reflects the natural variability in the outcomes of the two classes.

The following steps may be used to test hypotheses regarding two population proportions using random assignment.

Testing Hypotheses Regarding Two Proportions Using Random Assignment

Step 1. Verify that the explanatory variable (or treatment) has two levels and the response variable in the study is qualitative with two possible outcomes.

Step 2. Determine the null and alternative hypotheses. The hypotheses can be structured in one of three ways.

Two-Tailed	Left-Tailed	Right-Tailed
H_0: $p_1 = p_2$	H_0: $p_1 = p_2$	H_0: $p_1 = p_2$
H_1: $p_1 \neq p_2$	H_1: $p_1 < p_2$	H_1: $p_1 > p_2$

NOTE
p_1 is the population proportion for population 1, and p_2 is the population proportion for population 2.

Step 3. Build a null model that randomly assigns the individuals in the study under the assumption the statement in the null hypothesis is true.

Step 4. Estimate the P-value from the model in Step 3.

Step 5. State the conclusion.

Example Believe in Ghosts?

In a recent Harris Interactive Survey, individuals were asked, "Do you believe in ghosts?". Among the 510 teens (aged 13 – 17 years of age), 224 responded "yes." Among the 2463 adults (aged 18 years or older), 1010 responded "yes." Use randomization to judge whether there a difference in the belief of ghosts between the two groups.

❷ Test Hypotheses Regarding Two Population Proportions from Independent Samples (Using the Normal Model)

To perform inference comparing two population proportions from independent samples using a model, we first need to know the sampling distribution for the difference between two proportions from independent samples. The shape of the outcomes from random assignment in the Ghosts example suggests the sampling distribution is approximately normal.

Sampling Distribution of the Difference between Two Proportions (Independent Sample)

Suppose a simple random sample of size n_1 is taken from a population where x_1 of the individuals have a specified characteristic, and a simple random sample of size n_2 is independently taken from a different population where x_2 of the individuals have a specified characteristic. The sampling distribution of $\hat{p}_1 - \hat{p}_2$, where $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$, is approximately normal, with mean $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$ and standard deviation $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$, provided that $n_1\hat{p}_1(1-\hat{p}_1) \geq 10$ and $n_2\hat{p}_2(1-\hat{p}_2) \geq 10$ and each sample size is no more than 5% of the population size. The standardized version of $\hat{p}_1 - \hat{p}_2$ is then written as

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

which has an approximate standard normal distribution.

When comparing two population proportions, the null hypothesis is a statement of "no difference" (as always), so H_0: $p_1 = p_2$. Because the null hypothesis is assumed to be true, the test assumes that $p_1 = p_2$, or $p_1 - p_2 = 0$. Because we assume that both p_1 and p_2 equal p, where p is the common population proportion, substitute p into Equation (1), and obtain

$$z_0 = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \textbf{(2)}$$

In Other Words
The pooled estimate of p is obtained by summing the number of individuals in the two samples that have a certain characteristic and dividing this result by the sum of the two sample sizes.

We need a point estimate of p because it is unknown. The best point estimate of p is called the **pooled estimate of p**, denoted $\hat{p}$, where

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Substituting the pooled estimate of p into Equation (2), we obtain

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sigma_{\hat{p}_1 - \hat{p}_2}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

From the applet where the Ghost data was analyzed using randomization, click Analyze. This exports the difference in sample proportions for all the random assignments. Find the mean and standard deviation of the difference in sample proportions. Compare these to the theoretical mean and standard deviation.

Hypothesis Test Regarding the Difference between Two Population Proportions

To test hypotheses regarding two population proportions, p_1 and p_2, use the steps that follow, provided that

- the samples are independently obtained using simple random sampling or through a completely randomized experiment with two levels of treatment,
- $n_1 \hat{p}_1(1 - \hat{p}_1) \geq 10$ and $n_2 \hat{p}_2(1 - \hat{p}_2) \geq 10$, and
- $n_1 \leq 0.05N_1$ and $n_2 \leq 0.05N_2$ (the sample size is no more than 5% of the population size); this requirement ensures the independence necessary for a binomial experiment.

(continued)

Step 1 Determine the null and alternative hypotheses. The hypotheses can be structured in one of three ways:

Two-Tailed	Left-Tailed	Right-Tailed
$H_0: p_1 = p_2$	$H_0: p_1 = p_2$	$H_0: p_1 = p_2$
$H_1: p_1 \neq p_2$	$H_1: p_1 < p_2$	$H_1: p_1 > p_2$

Note: p_1 is the population proportion for population 1, and p_2 is the population proportion for population 2.

Step 2 Select a level of significance, α, depending on the seriousness of making a Type I error.

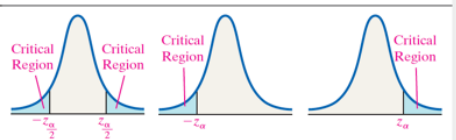
Classical Approach

Step 3 Compute the **test statistic**

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ where } \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Use Table V to determine the critical value.

	Two-Tailed	Left-Tailed	Right-Tailed
Critical value	$-z_{\frac{\alpha}{2}}$ and $z_{\frac{\alpha}{2}}$	$-z_\alpha$	z_α
Critical region			



Step 4 Compare the critical value to the test statistic.

Two-Tailed	Left-Tailed	Right-Tailed
If $z_0 < -z_{\frac{\alpha}{2}}$ or $z_0 > z_{\frac{\alpha}{2}}$, reject the null hypothesis.	If $z_0 < -z_\alpha$, reject the null hypothesis.	If $z_0 > z_\alpha$, reject the null hypothesis.

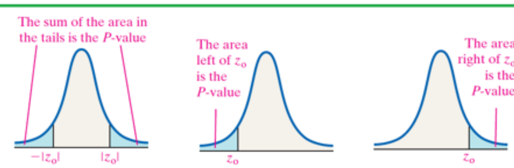
P-Value Approach

By Hand Step 3 Compute the **test statistic**

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ where } \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Use Table V to determine the *P*-value.

Two-Tailed	Left-Tailed	Right-Tailed
P-value $= 2P(z > \lvert z_0 \rvert)$	P-value $= P(z < z_0)$	P-value $= P(z > z_0)$



Technology Step 3 Use a statistical spreadsheet or calculator with statistical capabilities to obtain the *P*-value. The directions for obtaining the *P*-value using the TI-83/84 Plus graphing calculators, Excel, Minitab, and StatCrunch are in the Technology Step-by-Step on page 533.

Step 4 If *P*-value $< \alpha$, reject the null hypothesis.

Step 5 State the conclusion.

Example Believe in Ghosts?

In a recent Harris Interactive Survey, individuals were asked, “Do you believe in ghosts?”. Among the 510 teens (aged 13 – 17 years of age), 224 responded “yes.” Among the 2463 adults (aged 18 years or older), 1010 responded “yes.” Is there a difference in the belief of ghosts between the two groups?

③ Construct and Interpret Confidence Intervals for the Difference between Two Population Proportions

Constructing a $(1-\alpha)\cdot 100\%$ Confidence Interval for the Difference between Two Population Proportions (Independent Samples)

To construct a $(1-\alpha)\cdot 100\%$ confidence interval for the difference between two population proportions from independent samples, the following requirements must be satisfied:

1. The samples are obtained independently, using simple random sampling or from a randomized experiment.
2. $n_1\hat{p}_1(1-\hat{p}_1) \geq 10$ and $n_2\hat{p}_2(1-\hat{p}_2) \geq 10$.
3. $n_1 \leq 0.05N_1$ and $n_2 \leq 0.05N_2$ (the sample size is no more than 5% of the population size); this ensures the independence necessary for a binomial experiment.

Provided that these requirements are met, a $(1-\alpha)\cdot 100\%$ confidence interval for $p_1 - p_2$ is given by

$$\text{Lower bound: } (\hat{p}_1-\hat{p}_2) - z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$\text{Upper bound: } (\hat{p}_1-\hat{p}_2) + z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \quad \textbf{(3)}$$

Example   Position on Divorce

In a recent Harris Interactive Survey, respondents were asked, “How acceptable is divorce to you personally?”. Among the 970 individuals who considered themselves religious, 834 considered divorce to be acceptable. Among the 1285 individuals who did not consider themselves religious, 1157 considered divorce to be acceptable. Construct and interpret a 95% confidence interval for the difference in the two proportions.

❹ Determine the Sample Size Necessary for Estimating the Difference between Two Population Proportions

Sample Size for Estimating $p_1 - p_2$

The sample size required to obtain a $(1-\alpha)\cdot 100\%$ confidence interval with a margin of error, E, is given by

$$n = n_1 = n_2 = [\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)]\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 \qquad \textbf{(4)}$$

rounded up to the next integer, if prior estimates of p_1 and p_2, $\hat{p}_1$ and $\hat{p}_2$, are available. If prior estimates of p_1 and p_2 are unavailable, the sample size is

$$n = n_1 = n_2 = 0.5\left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 \qquad \textbf{(5)}$$

rounded up to the next integer.

The margin of error should always be expressed as a decimal when using Formulas (4) and (5).

Example Determining Sample Size

A doctor wants to estimate the difference in the proportion of 15-19 year old mothers that received prenatal care and the proportion of 30-34 year old mothers that received prenatal care. What sample size should be obtained if she wished the estimate to be within 2 percentage points with 95% confidence assuming:

A. she uses the results of the National Vital Statistics Report results in which 98% of the 15-19 year old mothers received prenatal care and 99.2% of 30-34 year old mothers received prenatal care.
B. she does not use any prior estimates.

11.2 Inference about Two Means: Dependent Samples

Objectives

1. Test hypotheses for a population mean from matched-pairs data
2. Construct and interpret confidence intervals about the population mean difference of matched-pairs data

The statistical inference methods on matched-pairs data use the same methods as inference on a single population mean, except that the differences are analyzed.

Testing Hypotheses Regarding the Difference of Two Means Using a Matched-Pairs Design

To test hypotheses regarding the mean difference of data obtained from a dependent sample (matched-pairs data), use the following steps, provided that

- the sample is obtained by simple random sampling or the data result from a matched-pairs design experiment.
- the sample data are dependent (matched pairs).
- the differences are normally distributed with no outliers or the sample size, n, is large $(n \geq 30)$.
- the sampled values are independent (sample size is no more than 5% of population size).

Step 1 Determine the null and alternative hypotheses. The hypotheses can be structured in one of three ways, where μ_d is the population mean difference of the matched-pairs data.

Two-Tailed	Left-Tailed	Right-Tailed
H_0: $\mu_d = 0$	H_0: $\mu_d = 0$	H_0: $\mu_d = 0$
H_1: $\mu_d \neq 0$	H_1: $\mu_d < 0$	H_1: $\mu_d > 0$

Step 2 Select a level of significance, α, depending on the seriousness of making a Type I error.

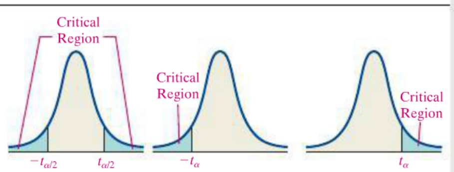
Classical Approach

Step 3 Compute the **test statistic**

$$t_0 = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}} = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

which follows Student's t-distribution with $n - 1$ degrees of freedom. The values of $\bar{d}$ and s_d are the mean and standard deviation of the differenced data. Use Table VII to determine the critical value.

	Two-Tailed	Left-Tailed	Right-Tailed
Critical value	$-t_{\frac{\alpha}{2}}$ and $t_{\frac{\alpha}{2}}$	$-t_\alpha$	t_α
Critical region			



Step 4 Compare the critical value to the test statistic.

Two-Tailed	Left-Tailed	Right-Tailed
If $t_0 < -t_{\frac{\alpha}{2}}$ or $t_0 > t_{\frac{\alpha}{2}}$, reject the null hypothesis.	If $t_0 < -t_\alpha$, reject the null hypothesis.	If $t_0 > t_\alpha$, reject the null hypothesis.

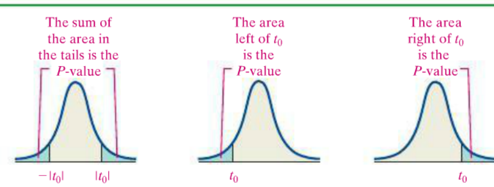
P-Value Approach

By Hand Step 3 Compute the **test statistic**

$$t_0 = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}} = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

which follows Student's t-distribution with $n - 1$ degrees of freedom. The values of $\bar{d}$ and s_d are the mean and standard deviation of the differenced data. Use Table VII to approximate the P-value.

Two-Tailed	Left-Tailed	Right-Tailed



P-value $= 2P(t > \lvert t_0 \rvert)$	P-value $= P(t < t_0)$	P-value $= P(t > t_0)$

Technology Step 3 Use a statistical spreadsheet or calculator with statistical capabilities to obtain the P-value. The directions for obtaining the P-value using the TI-83/84 Plus graphing calculators, Minitab, Excel, and StatCrunch are given in the Technology Step-by-Step on page 542.

Step 4 If P-value $< \alpha$, reject the null hypothesis.

Step 5 State the conclusion.

Example Testing a Claim Regarding Matched-Pairs Data

The following data represent the cost of a one night stay in Hampton Inn Hotels and La Quinta Inn Hotels for a random sample of 10 cities. Does the sample evidence suggest that Hampton Inn is more expensive than La Quinta? Use the $\alpha = 0.05$ level of significance.

City	Hampton Inn	La Quinta
Dallas	129	105
Tampa Bay	149	96
St. Louis	149	159
Seattle	189	149
San Diego	109	119
Chicago	160	125
New Orleans	149	92
Phoenix	129	89
Atlanta	129	119
Orlando	119	113

Confidence Interval for Matched-Pairs Data

A $(1-\alpha) \cdot 100\%$ confidence interval for μ_d is given by

$$\text{Lower bound: } \bar{d} - t_{\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}} \qquad \text{Upper bound: } \bar{d} + t_{\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}} \tag{1}$$

The critical value $t_{\alpha/2}$ is determined using $n-1$ degrees of freedom. The values of $\bar{d}$ and s_d are the mean and standard deviation of the differenced data.

Note: The interval is exact when the population is normally distributed and approximately correct for nonnormal populations, provided that n is large.

Example

16. Braking Distance An automotive researcher wanted to estimate the difference in distance required to come to a complete stop while traveling 40 miles per hour on wet versus dry pavement. Because car type plays a role, the researcher used eight different cars with the same driver and tires. The braking distance (in feet) on both wet and dry pavement is shown in the table below. Construct a 95% confidence interval for the mean difference in braking distance on wet versus dry pavement where the differences are computed as "wet minus dry." Interpret the interval. **Note:** A normal probability plot and boxplot of the data indicate that the differences are approximately normally distributed with no outliers.

Car	1	2	3	4	5	6	7	8
Wet	106.9	100.9	108.8	111.8	105.0	105.6	110.6	107.9
Dry	71.8	68.8	74.1	73.4	75.9	75.2	75.7	81.0

11.3A/11.3 Inference about Two Means: Independent Samples

Section 11.1 dealt with using random assignment to compare two independent proportions. We can also use random assignment to compare two independent means. The data in this scenario is collected from a completely randomized design with two levels of treatment and a quantitative response variable. Or, the data is collected using an observation study in which there are two distinct groups and the variable of interest is quantitative.

It is common for teachers to print exams on different colored paper as a deterrent to cheating. Is this practice fair? That is, does the color of the paper affect exam performance? To test this an instructor printed the same multiple choice exam on white paper and marine blue paper (distractor choices were scrambled, but other than that the exams were identical). The instructor randomly handed the exam to students as they entered the classroom and did not allow the same color paper to sit next to each other. Table 1 shows the exam results for an introductory psychology course. *Note*: This scenario is based on a study conducted by researchers at the University of New Mexico. "Effect of Paper Color and Question Order on Exam Performance" by Ilanit R. Tal, Katherine G. Akers, and Gordon K. Hodge. *Teaching of Psychology* 35:25-28, 2008.

Table 1

White Paper

77	78	74	91
85	87	83	61
84	86	74	70
72	81	77	78

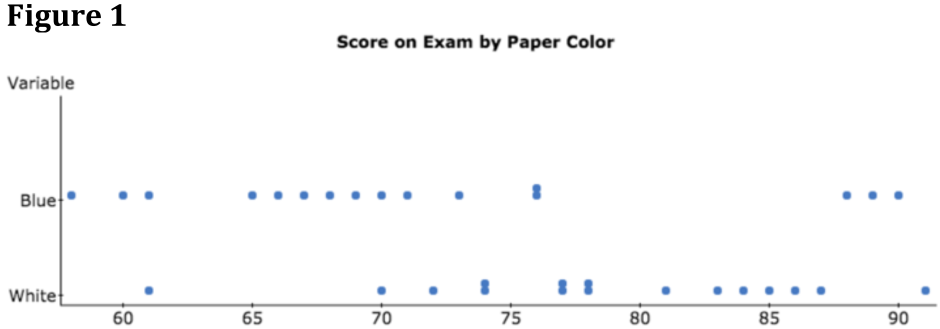
Blue Paper

65	90	71	66
68	60	88	58
67	70	76	73
69	61	76	89

It is important to recognize that this data was obtained through a completely randomized design with two levels of treatment: white paper or blue paper. The variable of interest is score on the exam, which is a quantitative variable.

Figure 1 shows a dot plot by color of the data in Table 1.

Figure 1



The sample mean score for the 16 white paper exams is 78.6 while the sample mean score for the 16 blue exams is 71.7. The sample means suggest that white paper exam scores are higher than blue paper exam scores. However, is this difference in exam scores significant? Or, is it possible, simply due to random chance that we obtained a sample mean difference of 78.6 – 71.7 = 6.9 and the color of the paper does not affect exam scores. The value of 6.9 is the test statistic for this study. Is the difference in sample means of 6.9 statistically significant, suggesting that white paper results in higher exam scores than blue paper? Or, is it possible that the difference is due to random chance and there is no difference in the exam scores for the two colors?" There are two possibilities:

(1) There is no association between exam scores and paper color and the difference in exam scores shown in Table 1 is due to random chance. That is, exam score is not related to paper color.
(2) There is an association between exam scores and paper color. Namely, white paper, on average, yields higher exam scores and this explains the difference in sample means. That is, exam score is related to paper color.

We can state these two possibilities using the language of hypothesis testing. We want to know if exams on white paper result in higher exam scores than exams on blue paper, on average. If true, then the mean score on white paper exams would exceed the mean score on blue paper exams, or $\mu_W > \mu_B$. The statement of "no change" or "no difference" would be that $\mu_W = \mu_B$. From this, we have the following null and alternative hypotheses.

$H_o : \mu_W = \mu_B$ This is Statement (1)
$H_1 : \mu_W > \mu_B$ This is Statement (2)

As always, the statement in the null hypothesis is used to build the null model. The logic behind building the null model is the same as it was in Section 11.1A. Statement (1) suggests that the exam scores observed in Table 1 are independent of paper color. The null model is based on this statement. Under the assumption that Statement (1) is true, we can randomly assign a paper color to each exam score (that is, the individual would have obtained the score actually obtained whether the paper color was white or blue). One way this can be accomplished physically is by writing the exam scores on 32 index cards. Shuffle the cards and deal 16 to the white paper group. The remaining 16 cards are part of blue paper group. For each group compute the sample mean and then compute the sample mean difference. It is important to recognize that the random assignment conforms to Statement (1) – exam score is not associated with paper color. Table 2 shows the results we obtained when doing this random assignment once.

Table 2

White Paper

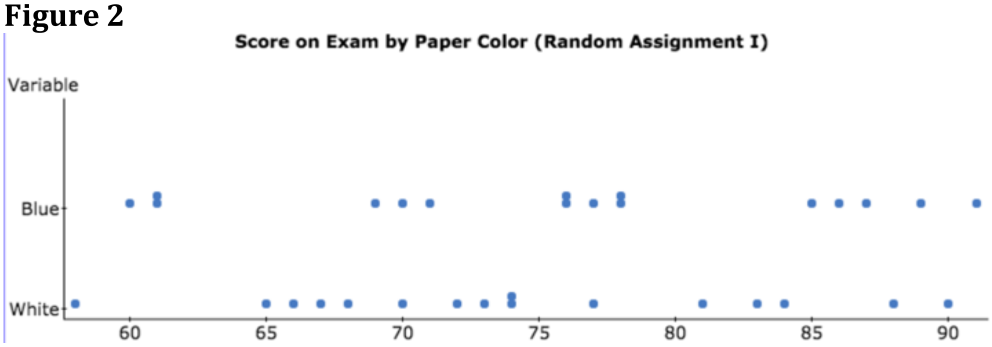
58	65	66	67
68	70	72	73
74	74	77	81
83	84	88	90

Blue Paper

60	61	61	69
70	71	76	76
77	78	78	85
86	87	89	91

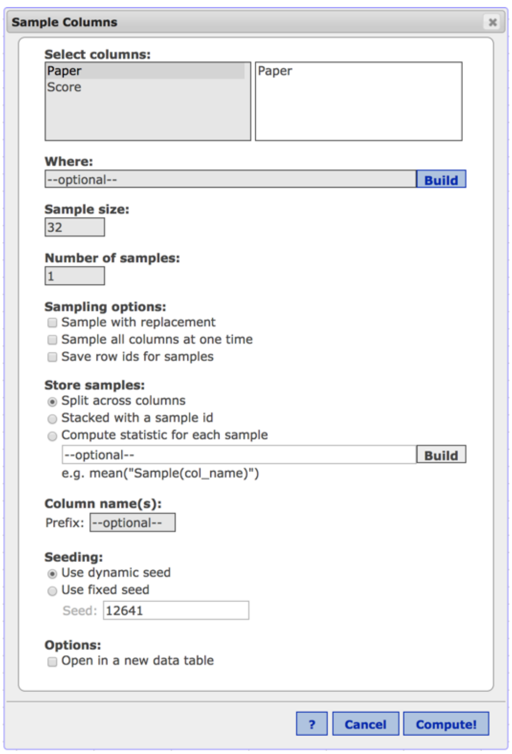
Figure 2 shows a dot plot of the randomly assigned data in Table 2. The sample mean score for white paper is 74.4 and the sample mean for blue paper is 75.9. The sample mean difference (white minus blue) is 74.4 – 75.9 = -1.5. This sample mean difference would become part of the null model.

Figure 2



Rather than physically doing the random assignment, we could use StatCrunch. To do this, first enter the paper color and exam score into a StatCrunch spreadsheet. Now, select Data, then highlight Sample. Select the column “Paper”, make the sample size 32 (for the 32 students). Let “Number of samples:” equal 1. Click Compute!. StatCrunch essentially randomly assigns a gender to a study time. See Figure 3.

Figure 3



Paper	Score	Sample(Paper)
Blue	65	Blue
Blue	90	White
Blue	71	Blue
Blue	66	Blue
Blue	68	White
Blue	60	White
Blue	88	Blue
Blue	58	White
Blue	67	White
Blue	70	White
Blue	76	Blue
Blue	73	White
Blue	69	White
Blue	61	White
Blue	76	Blue
Blue	89	White
White	77	Blue
White	78	Blue
White	74	White
White	91	Blue
White	85	Blue
White	87	White
White	83	Blue
White	61	Blue
White	84	White
White	86	White
White	74	Blue
White	70	White
White	72	White
White	81	Blue
White	77	Blue
White	78	Blue

The third column, Sample(Paper), represents the results of the random assignment, which is summarized in Table 3.

Table 3

White Paper

90	68	60	58
67	70	73	69
61	89	74	87
84	86	70	72

Blue Paper

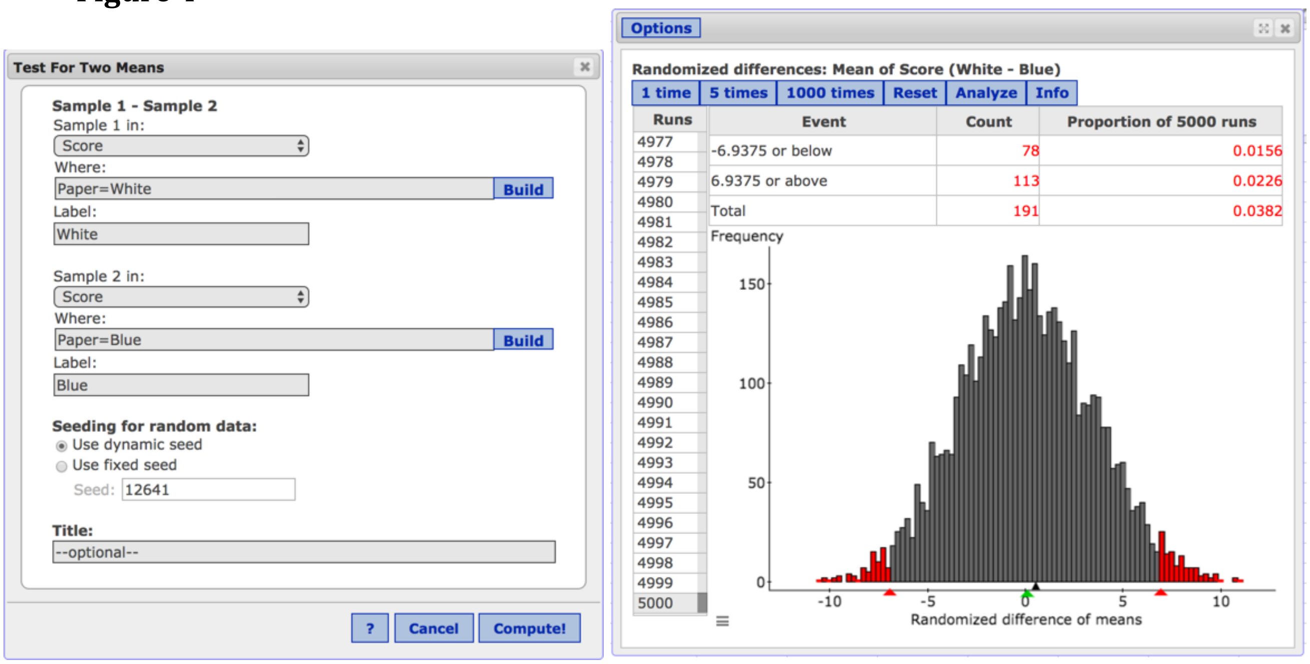
65	71	66	88
76	76	77	78
91	85	83	61
74	81	77	78

The sample mean exam score for white paper is 73.6 and the sample mean exam score for blue paper is 76.7. The sample mean difference (white minus blue) is 73.6

– 76.7 = -3.1.   This sample mean difference would also become part of the null model.

Of course, to get a proxy for the null model, the random assignment process would need to be repeated many, many times to get a sense of the distribution of sample mean differences under the assumption there is no association between exam score and paper color. That is, Statement (1) is true. The null model may be obtained easily using the "Randomization test for two means" applet in StatCrunch. Figure 4(a) shows how to fill in the dialogue box for this applet.

Figure 4



(a) (b)

Notice that a sample mean difference of 6.9 or higher was observed in 113 out of 5000 random assignments of paper color to exam score. Therefore, the *P*-value for this hypothesis test is 0.023. The likelihood of observing a sample mean difference of 6.9 or greater under the assumption that paper color is not associated with exam score is 0.023. Put another way, we would expect a sample mean difference of 6.9 or greater in about 2 of every 100 repetitions of this study *if* paper color does not impact exam score.   Because the observed results are unusual under the assumption of no association, we reject the statement in the null hypothesis. The conclusion is that the sample evidence suggests that exam scores on white paper are greater than exam scores on blue paper.

It is worth noting the shape of the distribution of randomized differences is bell-shaped and that the distribution is centered at 0. The fact that the center is near 0 should not be surprising because the statement in the null hypothesis is that the means of the two groups are the same (so the difference in the means should equal 0). Therefore, over the long-term we would expect half the random assignments to result in a mean difference in exam scores less than zero, and half to result in a mean difference in exam scores greater than 0.

The comparison of two means with unequal (and unknown) population variances is called the Behrens–Fisher problem. While an exact method for performing inference on the equality of two means with unequal population standard deviations does not exist, an approximate solution is available. The approach that we use is known as Welch's approximate *t*, in honor of English statistician Bernard Lewis Welch (1911–1989).

Sampling Distribution of the Difference of Two Means: Independent Samples with Population Standard Deviations Unknown (Welch's t)

Suppose that a simple random sample of size n_1 is taken from a population with unknown mean μ_1 and unknown standard deviation σ_1. In addition, a simple random sample of size n_2 is taken from a second population with unknown mean μ_2 and unknown standard deviation σ_2. If the two populations are normally distributed or the sample sizes are sufficiently large ($n_1 \geq 30$ and $n_2 \geq 30$), then

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \qquad \textbf{(1)}$$

approximately follows Student's t-distribution with the smaller of $n_1 - 1$ or $n_2 - 1$ degrees of freedom, where $\bar{x}_1$ is the sample mean and s_1 is the sample standard deviation from population 1, and $\bar{x}_2$ is the sample mean and s_2 is the sample standard deviation from population 2.

Click export from the randomization applet in Figure 4(b). Compare the mean and standard deviation from the randomization applet to the theoretical values from above.

❶ Test Hypotheses Regarding the Difference of Two Independent Means

Testing Hypotheses Regarding the Difference of Two Means

To test hypotheses regarding two population means, μ_1 and μ_2, with unknown population standard deviations, we can use the following steps, provided that

- the samples are obtained using simple random sampling or through a completely randomized experiment with two levels of treatment.
- the samples are independent.
- the populations from which the samples are drawn are normally distributed or the sample sizes are large ($n_1 \geq 30$ and $n_2 \geq 30$).
- for each sample, the sample size is no more than 5% of the population size.

Step 1 Determine the null and alternative hypotheses. The hypotheses are structured in one of three ways:

Two-Tailed	Left-Tailed	Right-Tailed
H_0: $\mu_1 = \mu_2$	H_0: $\mu_1 = \mu_2$	H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$	H_1: $\mu_1 < \mu_2$	H_1: $\mu_1 > \mu_2$

Note: μ_1 is the population mean for population 1, and μ_2 is the population mean for population 2.

Step 2 Select a level of significance α, depending on the seriousness of making a Type I error.

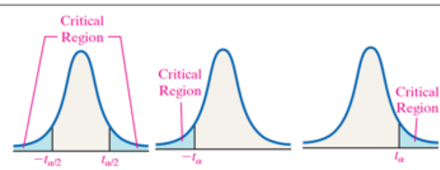
Classical Approach

Step 3 Compute the **test statistic**

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

which approximately follows Student's t-distribution. Use Table VII to determine the critical value using the smaller of $n_1 - 1$ or $n_2 - 1$ degrees of freedom.

	Two-Tailed	Left-Tailed	Right-Tailed
Critical value(s)	$-t_{\frac{\alpha}{2}}$ and $t_{\frac{\alpha}{2}}$	$-t_\alpha$	t_α
Critical region			



Step 4 Compare the critical value to the test statistic.

Two-Tailed	Left-Tailed	Right-Tailed
If $t_0 < -t_{\frac{\alpha}{2}}$ or $t_0 > t_{\frac{\alpha}{2}}$, reject the null hypothesis.	If $t_0 < -t_\alpha$, reject the null hypothesis.	If $t_0 > t_\alpha$, reject the null hypothesis.

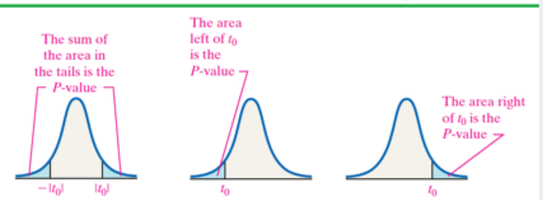
P-Value Approach

By Hand Step 3 Compute the **test statistic**

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

which approximately follows Student's t-distribution. Use Table VII to approximate the P-value using the smaller of $n_1 - 1$ or $n_2 - 1$ degrees of freedom.

Two-Tailed	Left-Tailed	Right-Tailed



Two-Tailed	Left-Tailed	Right-Tailed
P-value $= 2P(t > \lvert t_0 \rvert)$	P-value $= P(t < t_0)$	P-value $= P(t > t_0)$

Technology Step 3 Use a statistical spreadsheet or calculator with statistical capabilities to obtain the P-value. The directions for obtaining the P-value using the TI-83/84 Plus graphing calculators, Excel, Minitab, and StatCrunch are in the Technology Step-by-Step on pages 553–554.

Step 4 If P-value $< \alpha$, reject the null hypothesis.

Step 5 State the conclusion.

Example Testing Hypotheses Regarding Two Means

It is common for teachers to print exams on different colored paper as a deterrent to cheating. Is this practice fair? That is, does the color of the paper affect exam performance? To test this an instructor printed the same multiple-choice exam on white paper and marine blue paper (distractor choices were scrambled, but other than that the exams were identical). The instructor randomly handed the exam to students as they entered the classroom and did not allow the same color paper to sit next to each other. Table 1 shows the exam results for an introductory psychology course. Does the sample evidence suggest that students score higher on exams written on white paper than exams written on blue paper? *Note*: This scenario is based on a study conducted by researchers at the University of New Mexico. "Effect of Paper Color and Question Order on Exam Performance" by Ilanit R. Tal, Katherine G. Akers, and Gordon K. Hodge. *Teaching of Psychology* 35:25-28, 2008.

Table 1

White Paper

77	78	74	91
85	87	83	61
84	86	74	70
72	81	77	78

Blue Paper

65	90	71	66
68	60	88	58
67	70	76	73
69	61	76	89

The degrees of freedom used to determine the critical value(s) in the Classical Approach and by-hand P-value presented in Example 1 are conservative (this means we would require more evidence against the statement in the null hypothesis than if we use the formula for degrees of freedom given below). Results that are more accurate can be obtained by using the following formula for degrees of freedom:

$$\text{df} = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \qquad \textbf{(2)}$$

❷ Construct and Interpret Confidence Intervals Regarding the Difference of Two Independent Means

Constructing a $(1 - \alpha) \cdot 100\%$ Confidence Interval for the Difference of Two Means

A simple random sample of size n_1 is taken from a population with unknown mean μ_1 and unknown standard deviation σ_1. Also, a simple random sample of size n_2 is taken from a population with unknown mean μ_2 and unknown standard deviation σ_2. If the two populations are normally distributed or the sample sizes are sufficiently large ($n_1 \geq 30$ and $n_2 \geq 30$), a $(1 - \alpha) \cdot 100\%$ confidence interval about $\mu_1 - \mu_2$ is given by

$$\text{Lower bound:} \quad (\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and **(3)**

$$\text{Upper bound:} \quad (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $t_{\frac{\alpha}{2}}$ is computed using the smaller of $n_1 - 1$ or $n_2 - 1$ degrees of freedom or Formula (2).

12. Government Waste In a Gallup poll 513 national adults aged 18 years or older who consider themselves to be Republican were asked, "Of every tax dollar that goes to the federal government in Washington, D.C., how many cents of each dollar would you say are wasted?" The mean wasted was found to be 54 cents with a standard deviation of 2.9 cents. The same question was asked of 513 national adults aged 18 years or older who consider themselves to be Democrat. The mean wasted was found to be 41 cents with a standard deviation of 2.6 cents. Construct a 95% confidence interval for the mean difference in government waste, $\mu_R - \mu_D$. Interpret the interval.

CAUTION!

We would use the pooled two-sample *t*-test when the two samples come from populations that have the same variance. Pooling refers to finding a weighted average of the two sample variances from the independent samples. It is difficult to verify that two population variances might be equal based on sample data, so we will always use Welch's *t* when comparing two means.

8. Sugary Beverages It has been reported that consumption of sodas and other sugar-sweetened beverages cause excessive weight gain. Researchers conducted a randomized study in which 224 overweight and obese adolescents who regularly consumed sugar-sweetened beverages were randomly assigned to experimental and control groups. The experimental groups received a one-year intervention designed to decrease consumption of sugar-sweetened beverages, with follow-up for an additional year without intervention. The response variable in the study was body mass index (BMI—the weight in kilograms divided by the square of the height in meters). Results of the study appear in the following table. *Source:* Cara B. Ebbeling, PhD and associates, "A Randomized Trial of Sugar-Sweetened Beverages and Adolescent Body Weight" *N Engl J Med* 2012;367:1407–16. DOI: 10.1056/NEJMoal2Q3388

	Experimental Group ($n = 110$)	**Control Group ($n = 114$)**
Start of Study	Mean BMI = 30.4 Standard Deviation BMI = 5.2	Mean BMI = 30.1 Standard Deviation BMI = 4.7
After One Year	Mean Change in BMI = 0.06 Standard Deviation Change in BMI = 0.20	Mean Change in BMI = 0.63 Standard Deviation Change in BMI = 0.20
After Two Years	Mean Change in BMI = 0.71 Standard Deviation Change in BMI = 0.28	Mean Change in BMI = 1.00 Standard Deviation Change in BMI = 0.28

(a) What type of experimental design is this? Completely randomiz design
(b) What is the response variable? What is the explanatory variable? Change in BMI; Intervention method
(c) One aspect of statistical studies is to verify that the subjects in the various treatment groups are similar. Does the sample evidence support the belief that the BMIs of the subjects in the experimental group is not different from the BMIs in the control group at the start of the study? Use an $\alpha = 0.05$ level of significance.
(d) One goal of the research was to determine if the change in BMI for the experimental group was less than that for the control group after one year. Conduct the appropriate test to see if the evidence suggests this goal was met. Use an $\alpha = 0.05$ level of significance. What does this result suggest?
(e) Does the sample evidence suggest the change in BMI is less for the experimental group than the control group after two years? Use an $\alpha = 0.05$ level of significance. What does this result suggest?
(f) To what population do the results of this study apply?

For each of the following scenarios, state the type of inference that should be performed.

1. In an emergency situation in the wilderness, the ability to start a fire quickly is critical. A scoutmaster wanted to determine if the time required for scouts in the troop to start a fire was faster using (i) flint and steel or (ii) a battery and steel wool. Each scout built one fire using each of the two methods. The order in which they started their fires was randomized. The time required to start the fire was recorded for each trial.

2. In a Gallup poll published 25 January 2012, a sample of Americans were asked if they thought the economic system of the United States was fair. 37% of Democrats said the system was fair. 55% of the Republicans surveyed said the system was fair. (*Note to instructor: It is interesting to note that when the respondents were asked if the system was fair to them personally, 68% of the Democrats and 63% of the Republicans said it was fair to them personally.*)

3. A student wanted to assess if there is a difference in the wages of male and female college students. A random sample of students was selected, and they were asked to report the hourly wage they currently earned. What type of test should be performed?

4. You want to invest in a stock and have narrowed your choices down to either Citibank (C) or Wells Fargo (WFC). To help you decide which stock to invest in, you decide to compare weekly rates of return for the two stocks. Which would be a better sampling plan: (a) randomly choose 15 weeks and determine the rate of return of Citibank and then independently randomly choose 15 weeks and determine the rate of return of Wells Fargo or (b) randomly choose 15 weeks and determine the rate of return of both companies and compare the rate of return. For both scenarios, explain the inferential method you would use and justify your sampling plan.