ECON 262: Principles of Statistics II

Aina Katsikas, M.S.

Department of Economics, University of Nevada, Reno

Spring 2023

These lecture notes are adapted from Business Statistics: Communicating with Numbers by Jaggia & Kelly 4th edition.

Contents

1 Week 1: Jan. 23 - Jan. 27

1.1 Introduction to Hypothesis Testing

We all make decisions based on our beliefs. Each of us hold conjectures, or hypotheses, about certain things, and we base our life decisions on these hypotheses. For example, I believe that Mexico will be a popular spring break destination this year. At some point in time, our personal

hypotheses come face to face with evidence, and we are either confirmed to be true or false. In this case, evidence would be me actually traveling to Mexico to confirm if it is indeed a popular spring break destination.

For every hypothesis we have, there is the potential for an alternative scenario that contradicts our personal hypothesis. In this case, the alternative scenario would be that Mexico is not a popular spring break destination this year. **Our personal hypothesis is the null hypothesis and the contradicting scenario is the alternative hypothesis. Our goal is to determine if the null hypothesis false and therefore, should be rejected.**

If evidence is inconsistent with the null, we reject the null hypothesis. For example, if evidence finds that Mexico will not be a popular destination, we reject the null. If evidence is not inconsistent with the null, then we do not reject the null. Here is where it gets tricky, it's not entirely accurate to say "we accept the null hypothesis." Why? Because even if the null is not false, that does not mean that it's 100% true. When we say we "do not reject" the null that "Mexico will be a popular destination," what we're really saying is that it's possible for Mexico to be a popular destination. We are NOT establishing that Mexico absolutely WILL be a popular destination. We are simply saying that there is a very strong possibility that it will be. It's kind of like in court, you can have enough evidence to show that someone is NOT innocent, but you may not have enough to show that they are absolutely guilty. Even if we prove that Mexico is NOT an UNpopular destination, we can't entirely prove that it will ABSOLUTELY be a popular destination. All we can do is accept that it may potentially be a popular destination. The null hypothesis is written as H_0 : Mexico will be a popular destination. The alternative is written as H_A : Mexico will not be a popular destination.

There are some rules on how we can structure these hypotheses statements.

We can't be too broad in statistics. For example, the hypothesis: "Mexico will be a popular spring break destination this year" is too vague. We need to be specific.

H_0: The Mexico daily average tourist population will = 200,000 people. The matching alternative, H_A: The Mexico daily average tourist population will $\neq$ 200,000 people. This is an example of a two-tailed test. We can reject the null hypothesis (that Mexico tourist population = 200,000 people) if the actual tourist population is >200,000 or <200,000. It's called a two-tailed test because we can reject the null hypothesis on either side of the hypothesized value.

We can also frame our hypothesis to be in the form a one-tailed test. H_0: The Mexico daily average tourist population will be $\leq$ 200,000 people. The matching alternative, H_A: The Mexico daily average tourist population will be $>$ 200,000 people. This is a one-tailed test. We can only reject the null hypothesis (that the daily average tourist population will be $\leq$ 200,000 people) if the reality is that the daily average tourist population is $>$ 200,000 people. In a one-tailed test, we can reject the null hypothesis only on one side of the hypothesized value. In this case, it is a right one-tailed test because we can only reject the null hypothesis on the right side of the hypothesized value.

An easy trick to distinguish between right-tailed, left-tailed and two-tailed hypotheses has

to do with the mathematical sign used in the alternative hypothesis. The table below demonstrates this idea.

Type	Signs	Null Example	Alternative Example
Right-Tailed	$\geq, >$	The daily mean tourist pop. < 200,000.	The daily mean tourist pop. $\geq$ 200,000.
Left-Tailed	$\leq, <$	The daily mean tourist pop. > 200,000.	The daily mean tourist pop. $\leq$ 200,000.
Two-Tailed	$=, \neq$	The daily mean tourist pop. = 200,000.	The daily mean tourist pop. $\neq$ 200,000.

The difference between the null and alternative hypotheses.

The null is the status quo, it represents what we currently believe. The alternative is the statement that we are testing to see if it's true. The alternative hypothesis challenges the status quo. The alternative is always the opposite of the null and is mutually exclusive from the null. This means there is no overlap between the null and alternative.

For example, if we had a set of hypotheses that looked like:

- H_0 : The daily mean tourist pop. < 200,000.
- H_A : The daily mean tourist pop. < 200,000.

These are NOT correct because the alternative is identical to the null. It would still be incorrect if our hypotheses looked like these:

- H_0 : The daily mean tourist pop. < 200,000.
- H_A : The daily mean tourist pop. > 200,000.

These are incorrect because we need to make sure there is no "gray" area for rejecting the null hypothesis. What if the evidence shows that the daily mean tourist pop is exactly equal to 200,000? This value of 200,000 doesn't belong to the null or alternative hypotheses. We need to fix the hypotheses so they are completely distinct of each other, with no "gray" area in between:

- H_0 : The daily mean tourist pop. $\leq$ 200,000.
- H_A : The daily mean tourist pop. > 200,000.

When reading a word problem, the alternative hypothesis usually starts with "test if" or "test the claim" or "wants to determine if."

There is another element to the structure of these hypotheses that we need to understand and that is whether to use the μ (pronounced "mew") or p symbol. The μ represents the population mean. The p represents the population proportion. In most of the hypothesis tests we'll learn in this class, we are evaluating either a mean or proportion value. In the example above, we were evaluating the mean value of the tourist population. Therefore, if I were to rewrite the null and alternative hypotheses from above using the correct notation, they would look like this:

Null Example	Alternative Example
$\mu \leq$ 200,000	$\mu >$ 200,000
μ = 200,000	$\mu \neq$ 200,000

Notice that we use the term "mean" instead of "average." Although the two are synonymous, "mean" is a more accurate term and more commonly used.

Let's do some examples.

Example 1

A consulting firm currently hypothesizes that the mean value of back to school spending per family will be $600 this year. They want to test if the mean value of back to school spending per family will differ from this amount. Specify the null and alternative hypotheses.

In this case, our null is that mean value of back to school spending $=$ $600 and the alternative is that mean value of back to school spending $\neq$ $600.To write this out in formal statistics lingo, our null and alternative would look like this:

$$H_0\text{: } \mu = \$600$$
$$H_A\text{: } \mu \neq \$600$$

Example 2

An advertisement for a popular weight-loss clinic suggests that participants in its program experience a mean weight loss of more than 10 pounds. A consumer activist wants to test if the advertisement's claim is valid. Specify the null and alternative hypotheses to validate the advertisement's claim.

We are hypothesizing about the mean value, μ. The consumer activist wants to test if people actually do weigh more than the mean value of 10 pounds. (Remember that "test if" language indicates this is our alternative.)

$$H_0\text{: } \mu \leq 10 \text{ pounds}$$
$$H_A\text{: } \mu > 10 \text{ pounds}$$

Example 3

A television analyst wishes to test a claim that more than 50% of the households will tune in for a TV episode. Specify the null and alternative hypotheses to test the claim.

In this case, the analyst believes that less than 50% of households tune in, but wants to test if the alternative is true. Because we are testing for a proportion, (50% of households,) we use p in our fancy statistics lingo:

$$H_0\text{: } p \leq 0.5$$
$$H_A\text{: } p > 0.5$$

Example 4

It is generally believed that 60% or more of all Reno residents are happy. A sociologist wants to test if this value has dropped to below 60%. Specify the null and alternative hypotheses to test if the sociologist's claim is valid.

This is another one-tailed test for a proportion value. We can see that "test if" in there, indicating the alternative is the sociologist's claim that less than 60% of Reno residents are happy.

$$H_0\text{: } p \geq 0.6$$
$$H_A\text{: } p < 0.6$$

1.2 Type I & II Errors

Statistics is not always perfect, and we are bound to make errors. There are two main types of errors that we should look out for:

Type I Error	Rejecting a true null
Type II Error	Failing to reject a false null

A **Type I error** is when we mistakenly reject the null hypothesis when in reality, it's actually true. We are rejecting a true null hypothesis. A good example of this is a false negative covid test given the null hypothesis is that we have covid. Let's say we do an at-home covid test and it gives us a negative result. Therefore, we reject our null hypothesis. However, in reality, the test was wrong and we actually DO have covid. Therefore, we rejected our null hypothesis that was actually true. We rejected a true null hypothesis. This is a Type I error, rejecting a true null.

A **Type II error** is when we mistakenly do not reject the null hypothesis when in reality, we should reject the null because it's false. We are not rejecting a false null hypothesis. A good example of this is a false positive covid test given the null hypothesis is that we have covid. Let's say we do an at-home covid test and it gives us a positive result. Therefore, we do not reject our null hypothesis. However, in reality, the test was wrong and we actually do not have covid. Therefore, we did not reject our null hypothesis that was actually false. We FAILED to reject a false null hypothesis. This is a Type II error, failing to reject a false null.

Let's do some practice.

Example 1

H_0: A person does not have the flu.
H_A: A person has the flu.

What are examples of Type I vs. Type II errors?

A type I error occurs when the test indicates that the person has the disease but in reality, they do not. Therefore, a type I caused us to reject a true null hypothesis. A type II error occurs when the test indicates that the person does not have the disease, but in reality, they do have the disease. A type II error describes failure to reject a false null hypothesis.

Example 2

H_0: An accused person is innocent.
H_A: An accused person is guilty.

What are examples of Type I vs. Type II errors?

A Type I error occurs when the court finds the accused person to be guilty when in reality they are innocent. A Type II error occurs when the court finds the accused person to be innocent when in reality they are guilty. In situations like these, statistical errors can be costly.

Example 3

This example is very similar to your homework.

The screening process for detecting a rare disease is not perfect. Researchers have developed a blood test that is considered fairly reliable. It gives a positive reaction in 98% of the people who have that disease. However, it erroneously gives a positive reaction in 3% of the people who do not have the disease. Consider the null hypothesis "the individual does not have the disease" to answer the following questions.

What is the probability of a Type I and Type II error?

A Type I error is rejecting a true null hypothesis. In this case, the null is the individual does not have the disease. In a Type I scenario, this would occur if the test says the individual DOES have the disease when in reality, the individual does not have the disease. Out of the pool of people who DON'T have the disease, the test wrongly shows up positive for 3% of these people. **The probability of a Type I error is 3%.**

A Type II error is failing to reject a false null hypothesis. In this case, the null is that the individual does not have the disease. In a Type II scenario, this would occur if the test says the individual does not have the disease when in reality, the individual has the disease. Out of the pool of people who have the disease, the test only shows up positive for 98% of these people, meaning that the remaining 2% are given false negative results. **The probability of a Type II error is 2%.**

1.3 Hypothesis Test for the Population Mean when Population Standard Deviation(σ) is Known

Now that we are professionals in constructing hypotheses, we can conduct statistical testing to determine if we can reject the null hypothesis. There are two basic principles to hypothesis testing we need to know before we get started:

1. A hypothesis test regarding the population mean μ (pronounced "mew") based on the sample mean, $\bar{X}$ (pronounced "X bar") assumes that the underlying population is normally distributed. If the population is normally distributed, then this implies that the sample is also normally distributed.

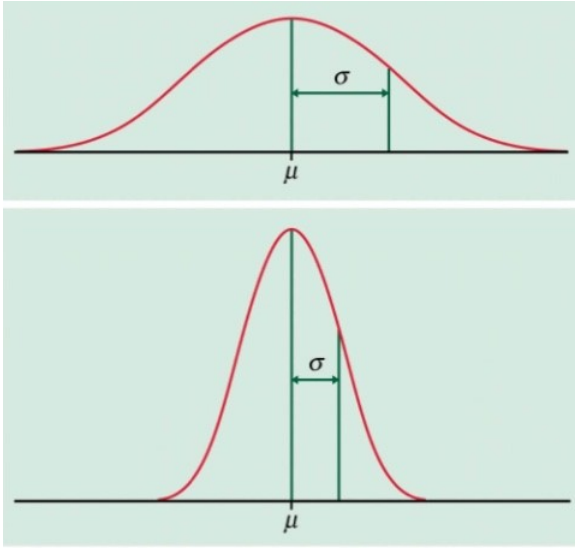
   What does this mean? It means that our large dataset (or our "population") must be normally distributed for us to move forward with testing. Otherwise, our results will be off. The easiest way to tell if your data is normally distributed is if the mean equals the median. IF the population dataset is not normally distributed, then we need the sample size to be large enough, which is n$\geq$ 30.

2. We always start by assuming that the null hypothesis is true and then go through hypothesis testing to determine if it is not true.

   This is very similar to the judicial system, the accused individual is innocent until proven guilty.

The first test we are learning is a hypothesis test for the population mean when population standard deviation (σ) is known. Remember from ECON 261, the standard deviation of a dataset is a measure of how spread out the data is from the mean. Looking at the figure below, we can see that the top dataset has a larger standard deviation compared to the bottom dataset. This is because the top dataset is more spread out and away from the mean. The bottom dataset has most of the data points close to the mean.

Figure 1.1: Standard Deviation of Two Different Datasets



Now, we can proceed with using the P Value approach to conduct this hypothesis test.

The P Value Approach in 6 Steps

1. Write out the null and alternative hypotheses.

2. Choose an α value (.01,.05, or .10). This is the allowed probability of making a Type I error.

3. Verify the population is normally distributed (mean=median).

   If it's not normally distributed, verify that our sample size is ≥ 30.

4. Calculate test statistic (also known as z value).

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

   - $\bar{x}$ = mean of the sample data
   - μ_0 (pronounced "mew-not") = hypothesized value of the population mean
   - σ = standard deviation of our population
   - $\sqrt{n}$ = square root of the number of observations in the sample data

5. Using the test statistic, calculate p value (using Excel). The Excel command differs whether it's a right tailed, left tailed or two-tailed test. Remember, a GREATER (>)than sign in the alternative hypothesis means we're doing a RIGHT-tailed test. A LESS (<) than sign in the alternative hypothesis means we're doing a LEFT-tailed test. A $\neq$ sign in the alternative means we are doing a TWO tailed test.

Operator in Alternative Hypothesis	Type of Test
> or $\geq$	Right-tailed Test
< or $\leq$	Left-tailed Test
$\neq$	Two-tailed Test

   The Excel commands for each type of test are below:

   For a right tailed, $= 1 - NORM.DIST(teststatistic, 0, 1, TRUE)$.

   For a left tailed, $= NORM.DIST(teststatistic, 0, 1, TRUE)$.

   For a two tailed, $= 2 * (1 - NORM.DIST(ABS(teststatistic), 0, 1, TRUE)$. For a two tailed test, you need to use the absolute value of the test statistic in your Excel command. For a right or left tailed test, you do not need to use the absolute value in your Excel command.

6. Using the p and α values, either reject or do not reject the null hypothesis.

   Reject the null if p-value < α.

   Do not reject the null if p-value $\geq \alpha$.

Let's do an example. A sociologist wants to determine if the mean retirement age is greater than 67 because she assumes right now that the mean retirement age is less than 67. Assuming that the data we're working with is normally distributed and we know that the population standard deviation is 9 (σ=9). To test this hypothesis, we take a sample of 25 retirees from the population of Reno. We find that the mean retiring age of this sample of 25 individuals is 71.

Step one, write out the null and alternative hypotheses. Because we assume right now that the mean age is ≤ 67, this is our null hypothesis. This is a right tailed test because we are using the $>$ sign in our hypotheses.

$$H_0\text{: } \mu \leq 67.$$
$$H_A\text{: } \mu > 67.$$

Step two, choose an α value of .01, .05 or .10. These are also called "significance levels" of 1%, 5% and 10%. These just indicate how accurate we want our results to be. The smaller the α value, the less room for error we are allowing. Choosing an α of .01 is holding our results to the highest standards. In most of these problems, I will give you the α value. In the real world, I present my research results at all three levels of significance and let my critics decide. Most economists present results in this format. For this problem, we'll use an $\alpha = 0.01$.

Step three, verify the population is normally distributed. We can assume it's normally distributed because it was stated in the question.

Step four, calculate the test statistic.

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

Let's remember the definition of each of these terms:

- $\bar{x}$ = mean of the sample data
- μ_0 (pronounced "mew-not") = hypothesized value of the population mean
- σ = population standard deviation
- $\sqrt{n}$ = square root of the number of observations in the sample data

So, let's define each of these terms in this example:

1. $\bar{x} = 71$
2. $\mu_0 = 67$
3. $\sigma = 9$
4. $\sqrt{n} = \sqrt{25}$

We plug them into the equation to look like this:

$$z = \frac{71 - 67}{9/\sqrt{25}} =$$

$$z = \frac{4}{9/5} =$$

$$z = \frac{4}{1.8} =$$

$$z = \frac{4}{1.8} = 2.2222$$

Step five, using the test statistic of 2.2222, find the p value using Excel. Given the fact that this is a right tailed test because $>$ is used in the alternative hypothesis, we enter $= 1 - NORM.DIST(2.2222, 0, 1, TRUE)$ into Excel and it spits out 0.0131. This is the p-value. This means that if we choose to reject the null hypothesis, there is a 1.31% chance that we are wrong. The p-value is the observed probability of making a Type I error. This seems like a small chance but is it small enough to allow us to reject the null hypothesis? This is where the α values come into play. The α value is the ALLOWED probability of making a Type I error.

Step six, using the p and α values, either reject or do not reject the null hypothesis. Reject the null if p-value < α. Do not reject the null if p-value $\geq \alpha$. In this case, our p-value is 0.0131 and our chosen α is 0.01. Therefore, 0.0131 > 0.01 so we do not reject the null hypothesis at the 1% significance level.

The last implicit step is to write a conclusion statement. Because we DO NOT reject the null hypothesis, we are saying there is a very good probability that $H_0 : \mu \leq 67$ is true. Therefore, we can take these results back to the sociologist and state that "The mean retirement age for the population in Reno is less than or equal to 67." Therefore, at the 1% significance level, we cannot conclude that the mean retirement age for the population of Reno is greater than 67.

You can look back on your α and p-values and reflect on this scenario. The p-value of 0.0131 is the observed probability of making a Type I error. Therefore, IF we chose to reject the null, there is a 1.31% chance that we are mistakenly rejecting a true null hypothesis. Because our α value that we chose was so low (0.01), we determined that the probability of making a Type I error (1.31%) was way too high for us, and we would rather NOT reject the null. Therefore, we do NOT reject the null when the observed probability of making a Type I error (p-value) is greater than the allowed probability of making a Type I error (α).

2 Week 2: Jan. 30 - Feb. 3

2.1 Hypothesis Test for the Population Mean When Standard Deviation(σ) is Unknown

We've been talking about how to conduct a hypothesis test for the population mean (μ) when we KNOW σ. But in the real world, we usually don't know σ. Therefore, we'll learn how to conduct a hypothesis test when we DON'T KNOW the population standard deviation. We follow the same 6 steps, but step 3 looks different when computing the test statistic.

In your homework, if you are only given the sample standard deviation and not the POPULATION standard deviation, that means σ is UNKNOWN and therefore you should follow these steps.

The P Value Approach in 6 Steps

1. Write out the null and alternative hypotheses.
2. Choose an α value (.01,.05, or .10). This is the allowed probability of making a Type I error.
3. Verify the population is normally distributed (mean=median).

   If we don't know if it's normally distributed, verify that our sample is ≥ 30.
4. Calculate test statistic (also known as a t value).

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

   Let me break it down what all these terms mean:

   - $\bar{x}$ = mean value of our sample data
   - μ_0 (pronounced "mew-not") = hypothesized value of the population mean
   - s = standard deviation of the sample
   - n = number of observations in the sample
   - $df = n - 1$ = number of observations in the sample - 1. This is also known as the "degrees of freedom" which I will talk more about down below.
5. Using the test statistic, calculate p value (using Excel). The Excel command differs whether it's a right tailed, left tailed or two-tailed test. Remember, a GREATER (>)than sign in the alternative hypothesis means we're doing a RIGHT-tailed test. A LESS (<) than sign in the alternative hypothesis means we're doing a LEFT-tailed test. A $\neq$ sign in the alternative means we are doing a TWO tailed test.

Operator in Alternative Hypothesis	Type of Test
> or $\geq$	Right-tailed Test
< or $\leq$	Left-tailed Test
$\neq$	Two-tailed Test

For a right tailed, $= T.DIST.RT(teststatistic, df)$.

For a left tailed, $= 1 - T.DIST.RT(teststatistic, df)$.

For a two tailed, $= 2 * T.DIST.RT(ABS(teststatistic), df)$.

6. Using the p and α values, either reject or do not reject the null hypothesis.

   Reject the null if p-value < α.

   Do not reject the null if p-value $\geq \alpha$.

Let's do an example. A university dean thinks that students study $\geq$ 24 hrs/week. She is curious if they study less than that. She randomly selects a sample of 35 students and interviews them on their quantity of study hours per week. From these interviews, she calculates a sample mean of 16.3714 hours and a sample standard deviation of 7.2155. **Remember, this is only the standard deviation of the sample, NOT the entire university population.** In statistics, s is used to denote the standard deviation of the sample and σ denotes the standard deviation of the population data.

Step one, write out the null and alternative hypotheses. Because we assume right now that the mean study time $\geq$ 24hrs/week, this is our null hypothesis. This is a left tailed test because we are using the < sign in our alternative hypothesis.

H_0: $\mu \geq 24$.
H_A: $\mu < 24$.

Step two, choose an α value of .01, .05 or .10. These are also called "significance levels" of 1%, 5% and 10%. These just indicate how accurate want our results to be. The smaller the α value, the less room for error we are allowing. For this example, let's choose an $\alpha = 0.05$.

Step three, verify that the population is normally distributed. OR verify that the sample size is sufficiently large, which is $\geq$ 30. In this case, our sample size is 35, and 35>30.

Step four, calculate the test statistic.

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

Let's recall what each of these terms mean:

- $\bar{x}$ = mean value of our sample data

- μ_0 (pronounced "mew-not") = hypothesized value of the population mean
- s = standard deviation of the sample
- n = number of observations in the sample
- $df = n - 1$ = number of observations in the sample - 1. This is also known as the "degrees of freedom." The degrees of freedom are a fancy way of measuring how many pieces went into computing the sample mean. It is important to note that the degrees of freedom vary depending on the test. For example, some df are computed as $n - 2$.

So, let's define each of these terms in this example:

- $\bar{x} = 16.3714$
- $\mu_0 = 24$
- $s = 7.2155$
- $\sqrt{n} = \sqrt{35}$
- $n - 1 = 35 - 1 = 34$

We plug them into the equation to look like this:

$$t_{df} = \frac{16.3714 - 24}{7.2155/\sqrt{35}} =$$

$$t_{df} = \frac{-7.6286}{7.2155/5.9161} =$$

$$t_{df} = \frac{-7.6286}{1.2196} = -6.2550$$

Step five, using the test statistic of -6.2550, find the p value using Excel. We enter $= 1 - T.DIST.RT(-6.2550, 34)$ into Excel and it spits out 2.013E-07. This translates to .0000002013, a very tiny number. This is the p-value. This means that if we choose to reject the null hypothesis, there is only a .00002013% chance that we are wrong.

Step six, using the p and α values, either reject or do not reject the null hypothesis. Reject the null if p-value < α. Do not reject the null if p-value $\geq \alpha$. In this case, our p-value is .0000002013 and our chosen α is .05. Therefore, .0000002013 < .05 and we reject the null hypothesis at the 5% significance level.

The last implicit step is to write a conclusion statement. Because we rejected the null hypothesis, we are saying there is no way IN HELL that $H_0 : \mu \geq 24$. It means we CANNOT accept the statement: "University students study, on average, more than 24 hours a week."

2.2 Hypothesis Test for the Population Proportion

If you're still reading at this point in the course, I'm really really proud of you. At this point, I think I completely tuned out of statistics altogether when I was an undergraduate.

Up until now, we've only been doing hypothesis tests for population MEANS. In the most recent example, we were testing for the mean number of hours studied for university students. Now, we'll talk about testing for a population PROPORTION, p. We will follow similar steps as we've done before, but two things differ: validating the normal distribution assumption looks different and computing the test statistic looks different.

The P Value Approach in 6 Simple Steps

1. Write out the null and alternative hypotheses.

2. Choose an α value (.01,.05, or .10). This is the allowed probability of making a Type I error.

3. Verify that the population data is normally distributed OR that the sample size is large enough.

   To verify the sample size is large enough, we need to check that $n * p_0$>=5, where n = sample size and p_0 = hypothesized proportion value.

4. Calculate test statistic (also known as z value).

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

   Let's break down the definition of each of these terms:

   - $\bar{p}$ (pronounced "p-bar") = proportion of the sample
   - p_0 (pronounced "p-not") = hypothesized value of the population proportion
   - n = number of observations in the sample

5. Using the test statistic, calculate p value (using Excel). The Excel command differs whether it's a right tailed, left tailed or two-tailed test. Remember, a GREATER (>)than sign in the alternative hypothesis means we're doing a RIGHT-tailed test. A LESS (<) than sign in the alternative hypothesis means we're doing a LEFT-tailed test. A $\neq$ sign in the alternative means we are doing a TWO tailed test.

Operator in Alternative Hypothesis	Type of Test
> or $\geq$	Right-tailed Test
< or $\leq$	Left-tailed Test
$\neq$	Two-tailed Test

For a right tailed, $= 1 - NORM.DIST(teststatistic, 0, 1, TRUE)$.

For a left tailed, $= NORM.DIST(teststatistic, 0, 1, TRUE)$.

For a two tailed, $= 2 * (1 - NORM.DIST(ABS(teststatistic), 0, 1, TRUE))$.

6. Using the p and α values, either reject or or do not reject the null hypothesis.

   Reject the null if p-value < α.

   Do not reject the null if p-value $\geq \alpha$.

Let's do an example. A magazine believes that more than 40% of households in the United States have changed their lifestyles because of environmental concerns. A recent survey of 180 households finds that 67 households have made lifestyle changes due to environmental concerns.

Step one, write out the null and alternative hypotheses. Because we assume right now that the status quo for the population is that more than 40% of households in the United States have changed their lifestyles because of environmental concerns, this is the null hypothesis. This is a left tailed test because we are using the < sign in our alternative hypothesis.

H_0: $p \geq .40$.
H_A: $p < .40$.

Step two, choose an α value of .01, .05 or .10. These are also called "significance levels" of 1%, 5% and 10%. These just indicate how accurate want our results to be. The smaller the α value, the less room for error we are allowing. For this example, let's choose an $\alpha = 0.05$.

Verify that the population data is normally distributed OR that the sample size is large enough. We have not been given the population data in this case. We only have access to the sample data. Therefore, we need to verify that the sample size is large enough.To do this, we just need to check that $np_0 \geq 5$, where n = sample size and p_0 = hypothesized proportion value. In this case, n = 180 and p_0 = 40% or 0.4. So, $np_0 = 180 * .40 = 72 \geq 5$.

Step four, calculate the test statistic.

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Let me break it down what all these terms mean:

- $\bar{p}$ (pronounced "p-bar") is the proportion of our sample (67/180)
- p_0 (pronounced "p-not") is our hypothesized value of the population average (we think more than 40% of households make lifestyle changes due to environmental concerns.)
- n is the number of observations in our sample (they interviewed 180 people).

So, let's define each of these terms in this example:

- $\bar{p}$ = 67/180 = 0.3722
- p_0 = 0.4
- n = 180

We plug them into the equation to look like this:

$$z = \frac{0.3722 - 0.4}{\sqrt{0.4(1-0.4)/180}} =$$

$$z = \frac{-0.0278}{\sqrt{0.4(0.6)/180}} =$$

$$z = \frac{-0.0278}{\sqrt{0.24/180}} =$$

$$z = \frac{-0.0278}{\sqrt{0.0013}} =$$

$$z = \frac{-0.0278}{0.0365} =$$

$$z = \frac{-0.0278}{0.0361} = -0.7613$$

Step five, using the test statistic of -0.7613, find the p value using Excel. We enter $= NORM.DIST(-0.7613, 0, 1, TRUE)$ into Excel and it spits out 0.2232. This means that if we choose to reject the null hypothesis, there is a 22.32% chance that we are wrong.

Step five, using the p and α values, either reject or do not reject the null hypothesis. Reject the null if p-value < α. Do not reject the null if p-value $\geq \alpha$. In this case, our p-value is 0.2232 and our chosen α is .05. Therefore, 0.2232 > .05 and we do NOT reject the null hypothesis at the 5% significance level.

The last implicit step is to write a conclusion statement. Because we DO NOT reject the null hypothesis, we are saying there is a good possibility that H_0: $p \geq .40$ is an accurate statement. Therefore, the magazine's claim that fewer than 40% of households in the US have changed their lifestyles because of environmental concerns is not justified by the sample data.

3 Week 3: Feb. 6 - Feb. 10

3.1 Inference of Difference Between Two Means Part. 1

Lately, we've been testing a hypothesis to determine the mean value of one sample. In this new hypothesis test, we will compare the mean values of two independent random samples and see if

they are statistically different. For example, let's say we collect a sample of 100 male salaries and a sample of 100 womens' salaries. We take the mean of each and compare. The mean of the male salaries is $50k annually, compared to the mean of the womens' salaries of $60k. We will apply this new hypothesis test to determine if this difference is true for just these samples or are true for the population data.

I mentioned the phrase **independent random samples** which is just a fancy term for "you picked your data samples out of a hat, completely blind." The process that generated the samples for men was completely separate and independent from the process that generated the samples for women. In other words, you picked your samples out of a hat without knowing what the salary values were.

There are two methods we'll learn to test if the population means are statistically different, and those are the confidence interval and p-value approaches.

The Confidence Interval Approach in 4 steps:

1. Write out the null and alternative hypotheses.

   This is much simpler now than it was before. Our null is simply that the two means are exactly the same (Mean1 = Mean2). In statistics, we like to write this as (μ_1 - μ_2 = 0) in other words, there is no difference between the mean values of the two populations. The alternative is that (Mean 1 $\neq$ Mean 2). Or that, (μ_1 - $\mu_2 \neq 0$). Our hypothesized value will always be 0 in this type of hypothesis test. The hypotheses in the confidence interval approach will resemble the following:

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

2. Choose an α value (.01,.05, or .10). This is the allowed probability of making a Type I error.

3. Verify the data is normally distributed or that $n_1 \geq 30$ and $n_2 \geq 30$. These two inequalities mean that both sample sizes are ≥ 30.

4. Compute the confidence interval at 1-α.

   This is where it gets slightly complicated. The equation we use to calculate the confidence interval depends on one thing: do we know the population variances of each sample?

   If we know the population variances:

$$(\bar{x_1} - \bar{x_2}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

   Let me define each of these terms for you:

   - $\bar{x_1}$ is the mean of the first sample

- $\bar{x_2}$ is the mean of the second sample
- n_1 is the size of the first sample
- n_2 is the size of the second sample
- σ_1^2 is the population variance of the first sample
- σ_2^2 is the population variance of the second sample
- $z_{\alpha/2}$ is the z-scores at α/2.

  We learned about z-scores back in ECON 261. Our α/2 = 0.05/2 = 0.025. The z-score is going to be one of 3 values. Refer to the table below to determine the z-score value.

Figure 3.1: Z Score at Various Confidence Levels

Percentage Confidence	z*-Value
80	1.28
90	1.645
95	1.96
98	2.33
99	2.58

If we DON'T know the population variances, but can assume them to be equal:

$$(\bar{x_1} - \bar{x_2}) \pm t_{\alpha/2,df}\sqrt{S_p^2(\frac{1}{n_1}) + \frac{1}{n_2}}$$

where,

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

where $df = n_1 + n_2 - 2$ are the corresponding sample variances (also known as the squared values of the sample standard deviations).
Let me define each of these terms for you:

- $\bar{x_1}$ is the mean of the first sample
- $\bar{x_2}$ is the mean of the second sample
- n_1 is the size of the first sample
- n_2 is the size of the second sample
- s_1^2 is the sample variance of the first sample

- s_2^2 is the sample variance of the second sample
- $t_{\alpha/2,df}$ is the t value at α/2 and degrees of freedom.

  We learned t values back in ECON 261. Our α/2 = 0.05/2 = 0.025 and degrees of freedom in this case = n_1+ $n_2 - 2$, where n_1 and n_2 are the number of observations in each group. Instead of looking at a t-table to find this value, I recommend using this calculator online, where you can directly enter the value of α without having to divide it by 2.
- S_p^2 is the pooled estimate of the population variance. This is just a fancy term for "educated guess of the population variance." We are making an educated guess because we do not know the population variance. We use the formula above to calculate S_p^2.

If we DON'T know the population variances, and CANNOT assume them to be equal:

$$(\bar{x_1} - \bar{x_2}) \pm t_{\alpha/2,df}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where,

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

.

- $\bar{x_1}$ is the mean of the first sample
- $\bar{x_2}$ is the mean of the second sample
- n_1 is the size of the first sample
- n_2 is the size of the second sample
- s_1^2 is the sample variance of the first sample (the square of the sample standard deviation)
- s_2^2 is the sample variance of the second sample (the square of the sample standard deviation)
- $t_{\alpha/2,df}$ is the t value at α/2 and degrees of freedom.

  We learned t values back in ECON 261. Our α/2 = 0.05/2 = 0.025 and degrees of freedom in this case is that long, complicated formula above where $df = ...$ Instead of looking at a t-table to find this value, I recommend using this calculator online, where you can directly enter the value of α without having to divide it by 2.

5. Choose to reject or do not reject the null hypothesis.

   If the confidence interval "goes through" or includes the hypothesized value of 0, we do not reject the null hypothesis. This means that the difference between the two means could very well be equal to 0.

If the confidence interval does NOT "go through" or does not include the hypothesized value of 0, we reject the null hypothesis. This means that the difference between the two means is $\neq 0$.

Let's do an example. A consumer advocate analyzes the nicotine content in two brands of cigarettes. A sample of 20 cigarettes of Brand A resulted in an average nicotine content of 1.68 milligrams with a standard deviation of 0.22 milligrams; 25 cigarettes of Brand B yielded an average nicotine content of 1.95 milligrams with a standard deviation of 0.24 milligrams. Nicotine content is normally distributed. The population variances here are unknown but assumed to be equal. Specify the competing hypotheses to determine whether the average nicotine levels are statistically different between Brand A and Brand B using the 95% confidence interval.

Step one, state the null and alternative hypotheses. In these cases where we are testing the difference between two population means, the null is always that the there is no statistical difference, or that Mean1 = Mean2. Remember, the fancy notation for "mean" is μ.

$$H_0\text{: } \mu_1 \text{ - } \mu_2 = 0$$
$$H_A\text{: } \mu_1 \text{ - } \mu_2 \neq 0$$

Step two, choose the α value. We are instructed to construct a confidence interval at 95%, which means we 1-α = .95, in other words, α = .05.

Step three, verify the data is normally distributed, or verify that the sample size is large enough. The problem states that "nicotine content is normally distributed."

Step four, compute the confidence interval. In this case, the population variances are unknown but assumed to be equal. So we use this equation:

$$(\bar{x_1} - \bar{x_2}) \pm t_{\alpha/2,df} \sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2}}$$

where,

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

where $df = n_1 + n_2 - 2$ and s_1^2 and s_2^2 are the corresponding sample variances (also known as the squared values of the sample standard deviations).

Let me break it down what all these terms mean:

1. $\bar{x_1}$ is the mean of the first sample
2. $\bar{x_2}$ is the mean of the second sample

3. n_1 is the size of the first sample

4. n_2 is the size of the second sample

5. s_1^2 is the sample variance of the first sample (the square of the sample standard deviation)

6. s_2^2 is the sample variance of the second sample (the square of the sample standard deviation)

7. $t_{\alpha/2,df}$ is the t value at α/2 and degrees of freedom.

   We learned t values back in ECON 261. Our α/2 = 0.05/2 = 0.025 and degrees of freedom in this case = n_1+ $n_2 - 2$, where n_1 and n_2 are the number of observations in each group. Instead of looking at a t-table to find this value, I recommend using this calculator online, where you can directly enter the value of α without having to divide it by 2.

8. S_p^2 is the pooled estimate of the population variance. This is just a fancy term for "educated guess of the population variance." We are making an educated guess because we do not know the population variance.

   We use the formula below to calculate S_p^2:

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

   where $df = n_1 + n_2 - 2$ and s_1^2 and s_2^2 are the corresponding sample variances (also known as the squared values of the sample standard deviations).

So, let's define each of these terms in this example:

1. $\bar{x_1}$ = 1.68

2. $\bar{x_2}$ = 1.95

3. n_1 = 20

4. n_2 = 25

5. s_1^2 is the sample variance of the first sample (the square of the sample standard deviation)

6. s_2^2 is the sample variance of the second sample (the square of the sample standard deviation)

7. $t_{\alpha/2,df}$ is the t value at α/2 and degrees of freedom

   $\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$

   = $n_1 + n_2 - 2 = 20 + 25 - 2 = 43$

   Using an online calculator,, $t_{.025,43}$ = 2.0167. In this calculator, enter the α value as it is (0.05) without dividing it by 2. Because this is a two tailed test (the $\neq$ sign is in the alternative).

8. Knowing what each of these terms mean, we can compute S_p^2. We use the equation,

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

where $df = n_1 + n_2 - 2$ and s_1^2 and s_2^2 are the corresponding sample variances (also known as the squared values of the sample standard deviations).

n_1 = 20

n_2 = 25

s_1^2 = 0.22^2

s_2^2 = 0.24^2

Using all of these values, we can plug in and solve for S_p^2:

$$S_p^2 = \frac{(20 - 1)(0.22)^2 + (25 - 1)(0.24)^2}{20 + 25 - 2}$$

$$S_p^2 = \frac{(19)(0.0484) + (24)(0.0576)}{43}$$

$$S_p^2 = \frac{.9196 + 1.3824}{43}$$

$$S_p^2 = \frac{2.302}{43}$$

$$S_p^2 = 0.0535$$

Now we can plug S_p^2 into our huge formula:

$$(1.68 - 1.95) \pm 2.0167\sqrt{.0535 * (\frac{1}{20} + \frac{1}{25})}$$

$$-0.27 \pm 2.0167\sqrt{.0535 * .09}$$

$$-0.27 \pm 2.0167\sqrt{.0048}$$

$$-0.27 \pm 2.0167 * .0694$$

$$-0.27 \pm 0.1398$$

$$-0.27 + 0.1398 = -0.1302$$

$$-.27 - 0.1398 = -0.4098$$

Therefore, our confidence interval ranges from -0.4098 to -0.1302.

Step five, reject or do not reject the null hypothesis. If the confidence interval "goes through" or includes the hypothesized value of 0, we do not reject the null hypothesis. If the confidence interval does NOT "go through" or does not include the hypothesized value of 0, we reject the null hypothesis. In this case, the confidence interval does NOT go through the value of 0, so we REJECT the null hypothesis. Therefore, there is no way in HELL that H_0: $\mu_1 - \mu_2 = 0$ is true. Therefore, at the 5% significance level, we support the conclusion that the difference between two means is NOT equal to 0. The means of each population of nicotine content are different.

3.2 Inference of Difference Between Two Means Part. 2

Now, let's move onto the second method to do this kind of hypothesis test, using the p-value approach.

The P Value Approach in 6 Steps

1. Write out the null and alternative hypotheses. For these hypotheses, the hypothesized value will also be 0. We can use any of the three operators in our hypotheses: $\geq, \leq, \neq$. Therefore, the hypotheses may look like one of the following:

   - $H_0 = \mu_1 - \mu_2 = 0$
   - $H_A = \mu_1 - \mu_2 \neq 0$
   - $H_0 = \mu_1 - \mu_2 > 0$
   - $H_A = \mu_1 - \mu_2 \leq 0$
   - $H_0 = \mu_1 - \mu_2 < 0$
   - $H_A = \mu_1 - \mu_2 \geq 0$

2. Choose an α value (.01,.05, or .10). This is the allowed probability of making a Type I error.

3. Verify that the population data is normally distributed, or verify that the sample size is large enough.

   If we don't know whether the population data is normally distributed, then we need to verify that $n_1 \geq 30$ and $n_2 \geq 30$.

4. Calculate the test statistic.

   This is where it gets slightly complicated. The equation we use to calculate the test statistic depends on one thing: do we know the population variances of each sample?

   If we know the population variances:

$$(z = \frac{\bar{x_1} - \bar{x_2} - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}})$$

Let me break it down what all these terms mean:

- $\bar{x_1}$ is the mean of the first sample
- $\bar{x_2}$ is the mean of the second sample
- d_0 is the hypothesized mean value, which is 0 in this case
- σ_1^2 is the population variance of population 1
- σ_2^2 is the population variance of population 2
- n_1 is the size of sample 1
- n_2 is the size of sample 2

If we DON'T know the population variances, but can assume them to be equal:

$$(t_{df} = \frac{\bar{x_1} - \bar{x_2} - d_0}{S_p^2\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}})$$

where,

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

Let me break it down what all these terms mean:

- $\bar{x_1}$ is the mean of the first sample
- $\bar{x_2}$ is the mean of the second sample
- d_0 is the hypothesized mean value, which is 0 in this case
- n_1 is the size of sample 1
- n_2 is the size of sample 2
- s_1^2 is the sample variance of sample 1
- s_2^2 is the sample variance of sample 2
- $df = n_1 + n_2 - 2$

If we DON'T know the population variances and CANNOT assume them to be equal:

$$t_{df} = \frac{\bar{x_1} - \bar{x_2} - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where,

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

Let me break it down what all these terms mean:

- $\bar{x_1}$ is the mean of the first sample
- $\bar{x_2}$ is the mean of the second sample
- d_0 is the hypothesized mean value, which is 0 in this case
- n_1 is the size of sample 1
- n_2 is the size of sample 2
- s_1^2 is the sample variance of sample 1
- s_2^2 is the sample variance of sample 2

5. Using the test statistic, calculate p value (using Excel). The Excel command differs depending on whether we used the first formula in computing the test statistic. That is, did we use the formula for when we know the population variances? If we did, these are the Excel commands:

   The Excel command differs whether it's a right tailed, left tailed or two-tailed test. Remember, a GREATER (>) than sign in the alternative hypothesis means we're doing a RIGHT-tailed test. A LESS (<) than sign in the alternative hypothesis means we're doing a LEFT-tailed test. A $\neq$ sign in the alternative means we are doing a TWO tailed test.

Operator in Alternative Hypothesis	Type of Test
> or $\geq$	Right-tailed Test
< or $\leq$	Left-tailed Test
$\neq$	Two-tailed Test

   For a right tailed, $= 1 - NORM.DIST(teststatistic, 0, 1, TRUE)$.

   For a left tailed, $= NORM.DIST(teststatistic, 0, 1, TRUE)$.

   For a two tailed, $= 2 * (1 - NORM.DIST(ABS(teststatistic), 0, 1, TRUE))$.

   If we used the second or third formulas to calculate the test statistics, then we use the following Excel commands. The Excel command differs whether it's a right tailed, left tailed or two-tailed test. Remember, a GREATER (>)than sign in the alternative hypothesis means we're doing a RIGHT-tailed test. A LESS (<) than sign in the alternative hypothesis means we're doing a LEFT-tailed test. A $\neq$ sign in the alternative means we are doing a TWO tailed test.

Operator in Alternative Hypothesis	Type of Test
> or $\geq$	Right-tailed Test
< or $\leq$	Left-tailed Test
$\neq$	Two-tailed Test

   For a right tailed, $= T.DIST.RT(teststatistic, df)$.

   For a left tailed, $= 1 - T.DIST.RT(teststatistic, df)$.

   For a two tailed, $= 2 * T.DIST.RT(ABS(teststatistic), df)$.

6. Using the p and α values, either reject or do not reject the null hypothesis.

   Reject the null if p-value < α.

   Do not reject the null if p-value $\geq \alpha$.

Let's do an example. An economist currently believes the average weekly food expenses for households in City 1 is less than or equal to the average weekly food expenses for households in City 2. She surveys 35 households in City 1 that yield an average weekly food expense of \$164. She surveys 30 households in City 2 that yield an average weekly food expense of \$159. The standard deviation for the population of City 1 households is \$12.5 and for city 2 is \$9.25. At the 5% significance level, can we support the economist's claim that City 1 has reduced expenses than City 2?

Step one, state the null and alternative hypotheses. We believe that city 2 has greater average food expenses than city 1, so therefore subtracting Mean1 - Mean2 would yield a negative value because we believe Mean2 > Mean1.

$$H_0\text{: } \mu_1 \text{ - } \mu_2 \leq 0.$$
$$H_A\text{: } \mu_1 \text{ - } \mu_2 > 0.$$

Step two, choose the α value. We are given, $\alpha = .05$.

Step three, verify that the population data is normally distributed or that the sample size is big enough. We don't have access to the data, but both $n_1 \geq 30$ and $n_2 \geq 30$, so we can proceed.

Step four, Compute the test statistic. In this case, we know the population standard deviations (and population variance is just the squared value of population standard deviation,) and so we use this equation:

$$(z = \frac{\bar{x_1} - \bar{x_2} - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}})$$

Let me break it down what all these terms mean:

- $\bar{x_1}$ is the mean of the first sample
- $\bar{x_2}$ is the mean of the second sample
- d_0 is the hypothesized mean value, which is 0 in this case
- σ_1^2 is the squared value of the standard deviation of sample 1
- σ_2^2 is the squared value of the standard deviation of sample 2
- n_1 is the size of sample 1

- n_2 is the size of sample 2

So, let's define each of these terms in this example:

1. $\bar{x_1}$ = 164
2. $\bar{x_2}$ = 159
3. d_0 = 0
4. σ_1^2 = 12.50^2
5. σ_2^2 = 9.25^2
6. n_1 = 35
7. n_2 = 30

And plug it into our equation:

$$z = \frac{(164 - 159) - 0}{\sqrt{\frac{(12.50)^2}{35} + \frac{(9.25)^2}{30}}}$$

$$z = \frac{5}{\sqrt{\frac{156.25}{35} + \frac{85.5625}{30}}}$$

$$z = \frac{5}{\sqrt{4.4643 + 2.8521}}$$

$$z = \frac{5}{\sqrt{7.3164}}$$

$$z = \frac{5}{2.7049}$$

$$z = 1.8485$$

Step five, compute the p value using excel. Because we used the first formula to calculate the test statistic and the $>$ sign in our alternative hypothesis, this is a right tailed test. To find the p value for this right tailed test, we enter $1 - NORM.DIST(1.8485, 0, 1, TRUE)$ into Excel, and get 0.0323.

Step six, reject or do not reject the null hypothesis. The p value is less than 0.05, so we reject the null. This means that there is no way in HELL that H_0: $\mu_1 - \mu_2 \leq 0$. Therefore, we can conclude that the food expenses in city 1 is higher than those of city 2.

4 Week 4: Feb. 13 - Feb. 17

4.1 Inference of Mean Differences Part. 1

We know how to conduct a hypothesis test for the difference of the means of two independently random samples. But what if our samples are not independently random? Remember, independently random samples just means that the process that generated observations in one sample is independent of the process that generated the other sample. This means we "pulled our test subjects out of a hat" and they were completely random. What if they are not random? What if we wanted to compare statistics between siblings who all come from the same parents. This is where we get into **matched pairs sampling**. There are two types of matched pairs sampling.

1. **Before & After** sampling is like if we were to deliver antibiotics to a sick patient and measure their illness status before and after the "intervention" of antibiotics. We are evaluating the status of the same patient at two points in time (before and after the intervention.) The "matched pair" in this case is the patient's illness status before and after the antibiotics.

2. **Control vs. Treatment Groups** sampling is like if we were to deliver antibiotics to one of TWO sick patients and compare their illness statuses after the "intervention" of antibiotics. The first patient gets the antibiotics, this is the "treatment" group. The second patient gets a placebo medicine, this is the "control" group. We are evaluating the status of both patients at one point in time, after the intervention.The "matched pair" in this case is BOTH patients' illness statuses after the antibiotics.

In these kinds of experiments, we are interested in seeing if the differences between each matched pair is statistically significant, or are the differences just due to silly chance? If they are just due to chance, than we wouldn't find the same pattern in the population data. However, if they are statistically significant, then that means these samples are accurately reflecting trends in the population data. For example, let's say in the Control vs. Treatment experiment I mentioned above, let's say the patient that received the antibiotic got better and recovered from the illness. The patient that received the placebo did not recover. How do we know if this recovery is due to the power of the antibiotics or just random chance? This is why we do sadistical testing. There are two methods to test these kinds of experiments: **Confidence Intervals** and **P-Value Approach**.

First, we'll do the **Confidence Interval Approach**:

1. State the null and alternative hypotheses. In this case, we'll use μ_D to denote the "mean difference," which is the difference between the two means.

   Hypotheses in these confidence interval approaches for this specific test are standardized and will always be the same:

   - $H_0 : \mu_D = 0$
   - $H_A : \mu_D \neq 0$

2. Choose the α value of .01, .05, or .10. This is the allowed probability of making a Type I error.

3. Verify that the population data is normally distributed OR verify that the sample is large enough, n $\geq$ 30.

4. Compute the confidence interval at the 1 - α level using the equation below:

$$\bar{d} \pm t_{\alpha/2,df} S_D/\sqrt{n}$$

- $\bar{d}$ = mean of the sample differences
- S_D = standard deviation of the sample differences
- n = sample size
- $t_{\alpha/2,df}$ = t value at $\alpha/2$ and df. We'll use this t value calculator to do this. This calculatror makes it easy, enter the α value without dividing by 2. Use the t value that is "two tailed." On the test, I will give you the t value in a case like this.

  df = n - 1

  If you would like to be traditional, you can use this t table distribution and manually interpret the table. But the focus of this class is not interpreting these tables, so I am fine with you using online t value calculators.

5. Choose to reject or do not reject the null hypothesis.

   If the confidence interval "goes through" or includes the hypothesized value of 0, we do not reject the null hypothesis. This means that the difference between the two means could very well be equal to 0.

   If the confidence interval does NOT "go through" or does not include the hypothesized value of 0, we reject the null hypothesis. This means that the difference between the two means is $\neq$ 0.

Let's do an example. A manger wants to improve the productivity of her plant by changing the layout of a workstation. She tracks the same 10 workers in the time it takes them to individually complete a task before and after the intervention. This is an example of a "Before & After" matched pairs sampling. She currently believes there is no difference in the time to complete a task before and after the intervention because she's pessimistic. She finds that the difference between the mean times to complete at task for each group is 8.5 and the standard deviation is 11.38. Compute the confidence interval at the 95% level to see if this difference in means (μ_D) is due to statistical significance or random chance. Assume that the population data is normally distributed.

Step one, state the null and alternative hypotheses. She currently believes there is no difference before and after the intervention, or that the difference in means(μ_D) is 0.

1. $H_0 : \mu_D = 0$
2. $H_a : \mu_D \neq 0$

Step two, Verify that our data is normally distributed OR verify that the sample is large enough, (≥ 30). The problem said we can assume that our data is normally distributed.

Step three, Choose the α value of .01, .05, or .10. Here, we are told to compute a confidence interval at the 95% level which means we must have an α value of .05.

Step four, Compute the confidence interval at the 1 - α level using the equation below.

$$\bar{d} \pm t_{\alpha/2,df} S_D/\sqrt{n}$$

Let's break down what each of these terms mean.

- $\bar{d}$ = mean of the sample differences
- S_D = standard deviation of the sample differences
- n = sample size
- $t_{\alpha/2,df}$ = t value at $\alpha/2$ and *df* We'll use this t value calculator to do this.

  df = n - 1

  If you would like to be traditional, you can use this t table distribution and manually interpret the table. But the focus of this class is not interpreting these tables, so I am fine with you using online t value calculators. On the test, I will give you the t value in a case like this.

Now let's define each of these terms in this example.

1. $\bar{d}$ = 8.5
2. S_D = 11.38
3. n = 10
4. $t_{\alpha/2,df}$ = t value at $\alpha_{/2}$ and $df = n - 1$

   $\alpha/2$ = $.05/2 = .025$

   $df = 10 - 1 = 9$

   Use this t value calculator, and we get, $t_{.025,9}$ = 2.262.

Now we can plug this all into our equation:

$$8.5 \pm 2.262 * (\frac{11.38}{\sqrt{10}})$$

$$8.5 \pm 2.262 * (\frac{11.38}{3.1622})$$

$$8.5 \pm 2.262 * (3.5987)$$

$$8.5 \pm 8.1403$$

$$8.5 + 8.1403 = 16.6403$$

$$8.5 - 8.1403 = 0.3597$$

So, our confidence interval ranges from 0.3597 to 16.6403.

Step five, reject or do not reject the null. Our confidence interval ranges from 0.3597 to 16.6403 which does NOT contain the hypothesized value of a mean difference of 0. So, we reject the null hypothesis. This means there is no way in HELL that $H_0 : \mu_D = 0$ is true. This means that the difference in the mean time to complete a task before and after the intervention is NOT 0. This is good news for the plant manager because it means that the difference is not due to just random chance, but is in fact due to the new layout.

4.2 Exam 1 Review in Class on Feb. 15

5 Week 5: Feb. 20 - Feb. 24

5.1 NO CLASS MON, FEB. 20

5.2 Exam 1 on Feb. 22 via Canvas

6 Week 6: Feb. 27 - Mar. 3

6.1 Inference of Mean Differences Part. 2

Now we'll do the **P-Value Approach**.

1. State the null and alternative hypotheses. In this case, we'll use μ_D to denote the "mean difference," which is the difference between the two means.

2. Choose the α value of .01, .05, or .10. This is the allowed probability of making a Type I error.

3. Verify that our data is normally distributed OR verify that the sample is large enough, n $\geq$ 30, where n is the number of sample differences.

4. Compute the test statistic using the equation below.

$$t_{df} = \frac{\bar{d} - d_0}{S_D/\sqrt{n}}$$

Let's break down what each of these terms mean.

- $\bar{d}$ = mean of the sample differences
- d_0 = hypothesized value of the mean difference
- S_D = standard deviation of the sample differences
- n = number of sample differences
- $df = n - 1$

5. Compute the p-value in Excel using the test statistic. The Excel command differs whether it's a right tailed, left tailed or two-tailed test. Remember, a GREATER (>)than sign in the alternative hypothesis means we're doing a RIGHT-tailed test. A LESS (<) than sign in the alternative hypothesis means we're doing a LEFT-tailed test. A $\neq$ sign in the alternative means we are doing a TWO tailed test.

Operator in Alternative Hypothesis	Type of Test
> or $\geq$	Right-tailed Test
< or $\leq$	Left-tailed Test
$\neq$	Two-tailed Test

For a right tailed, $= T.DIST.RT(teststatistic, df)$.

For a left tailed, $= 1 - T.DIST.RT(teststatistic, df)$.

For a two tailed, $= 2 * T.DIST.RT(ABS(teststatistic), df)$.

6. Using the p and α values, either reject or do not reject the null hypothesis.

Reject the null if p-value < α.

Do not reject the null if p-value $\geq \alpha$.

Let's do an example. A nutritionist is evaluating if an intervention in a restaurant will change how many calories consumers drink. The intervention is posting the calories in each drink. The nutritionist tracks 40 consumers before and after the intervention. She is pessimistic and currently believes that consumers on average consume more calories after the intervention. The average of the differences, ($\bar{d}$) is 2.1. The standard deviation is 8.1549. Conduct a hypothesis test at the 5% significance level to evaluate if the intervention reduced the average calories consumed in drinks.

Step one, state the null and alternative hypotheses. She currently believes people drank more calories after the intervention. Let's assume that Mean2 is the average calories consumed after the intervention. Mean1 is the average calories consumed before the intervention. We can write this as Mean1 - Mean2 ≤ 0 or that the difference in means(μ_d) is ≤ 0.

1. $H_0 : \mu_d \leq 0$
2. $H_a : \mu_d > 0$

Step two, verify that our data is normally distributed OR verify that the sample is large enough. Our sample size is large enough $n \geq 30$.

Step three, choose the α value of .01, .05, or .10. Here, we are told to test at an α of 5%.

Step four, compute the test statistic using the equation below.

$$t_{df} = \frac{\bar{d} - d_0}{S_D/\sqrt{n}}$$

Let's break down what each of these terms mean.

- $\bar{d}$ = mean of the sample differences
- d_0 = hypothesized value of the mean difference
- S_D = standard deviation of the sample differences
- n = number of sample differences
- $df = n - 1$

Now let's define each of these terms in the context of the problem.

- $\bar{d}$ = 2.10
- d_0 = 0
- S_D = 8.1549
- n = 40
- $df = 40 - 1 = 39$

Now we can plug this into our equation to compute the test statistic.

$$t_{39} = \frac{2.10 - 0}{8.1549/\sqrt{40}} =$$

$$t_{39} = \frac{2.10}{8.1549/6.3246} =$$

$$t_{39} = \frac{2.10}{1.2894} =$$

$$t_{39} = \frac{2.10}{1.2894} = 1.6287$$

Step five, compute the p-value in Excel using the test statistic. Because we use the > sign in the alternative hypothesis, we know this is a right tailed test. Our df = 40 - 1 =39. Therefore, we enter $T.DIST.RT(1.6287, 39)$ into Excel and get 0.0557 in return.

Step six, reject or do not reject the null hypothesis. In this case, our p value is greater than the alpha value, 0.0557 > 0.05, so we do NOT reject the null hypothesis. Our null hypothesis was that consumers drink more calories after the intervention than before. Because we cannot reject this null hypothesis, it's possible that people are drinking more calories after the intervention. Therefore, we can't say that the intervention was successful in reducing caloric consumption from drinks.

6.2 Inference of the Difference Between 2 Proportions

In this unit of week 6, we'll learn a hypothesis test for the difference between two proportions. An example of this would be if we wanted to test if the proportion of a product purchased by women vs. men differs. Is this difference in proportions due to statistical significance or pure chance? If this is due to statistical significance, then we would expect to find the same patterns beyond just the sample data, but in the population data as well.

There are two approaches for this kind of hypothesis test: **Confidence Interval Approach** and **P-Value Approach**.

First, we'll do the **Confidence Interval Approach**:

1. State the null and alternative hypotheses. Our hypotheses will use the p value, because we are testing for a proportion. The set of hypotheses will always look like one of these three, with the hypothesized value being 0.

   - $H_0 : p_1 - p_2 = 0$
   - $H_A : p_1 - p_2 \neq 0$

   - $H_0 : p_1 - p_2 \geq 0$

- $H_A : p_1 - p_2 < 0$
- $H_0 : p_1 - p_2 \leq 0$
- $H_A : p_1 - p_2 > 0$

2. Choose the α value of .01, .05, or .10. This is the allowed probability of making a Type I error.

3. Verify that our data is normally distributed OR verify that the sample is large enough, which in this case means:

   $n_1\bar{p_1} \geq 5$ AND
   n_1(1-$\bar{p_1}$)≥ 5 AND
   $n_2\bar{p_2} \geq 5$ AND
   n_2(1-$\bar{p_2}$)≥ 5 AND

   We have to check that each of these is ≥ 5 to make sure the sample is large enough, where n_1 is the first sample size, n_2 is the second sample size and p_1 is the proportion of the first sample and p_2 is the proportion of the second sample.

4. Compute the confidence interval at the 1 - α level using the equation below:

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

Let's break down what each of these terms mean.

- n_1 = first sample size
- n_2 = second sample size
- $\bar{p_1}$ = first proportion
- $\bar{p_2}$ = second proportion
- $z_{\alpha/2,df}$ = z value at $\alpha/2$ and df where df = number of samples

  To get this z value, we need to read a z table. But to make your lives easier, here are the z values for corresponding significance levels. We are only ever going to use confidence intervals at these levels.

Percentage Confidence	z*-Value
80	1.28
90	1.645
95	1.96
98	2.33
99	2.58

5. Choose to reject or do not reject the null hypothesis.

   If the confidence interval "goes through" or includes the hypothesized value of 0, we do not reject the null hypothesis. This means that the difference between the two means could very well be equal to 0.

   If the confidence interval does NOT "go through" or does not include the hypothesized value of 0, we reject the null hypothesis. This means that the difference between the two means is $\neq 0$.

Let's do an example. Candidate A appears to have gained support among voters in the People Choice Awards for best reality star. Three months ago, in a survey of 120 individuals, 55 people said that they would vote for Candidate A. Today, in a survey of 80 individuals, 41 said that they would vote for Candidate A. Construct the 95% confidence interval for the difference between the two population proportions given this sample data. We hypothesize that there is no difference between the two proportions.

Step one, state the null and alternative hypotheses. We currently believe there is no difference between the two proportions, or that p_1 - p_2 = 0. This is our null.

1. $H_0 : p_1$ - p_2 = 0
2. $H_A : p_1$ - $p_2 \neq 0$

Step two, choose the α value of .01, .05, or .10. Here, we are told to compute a confidence interval at the 95% level which means we must have an α value of .05.

Step three, verify that our data is normally distributed OR verify that the sample is large enough. We need to check these four things:

1. $n_1\bar{p_1} \geq 5$ AND
2. $n_1(1 - \bar{p_1}) \geq 5$ AND
3. $n_2\bar{p_2} \geq 5$ AND
4. $n_2(1 - \bar{p_2}) \geq 5$

Remember that n_1 is our first sample size (80), n_2 is our second sample size (120), $\bar{p_1}$ is the first proportion (41/80), and $\bar{p_2}$ is the second proportion (55/120).

It doesn't matter which is your "first" or "second" samples, so long as you keep it consistent throughout your calculations. In this case, I'm going to assign the survey from three months ago as our "second sample."

We can verify these four equations are in fact, true:

1. $80 * (41/80) = 41 \geq 5$ AND
2. $80 * (1 - (41/80)) = 39 \geq 5$ AND
3. $120 * (55/120) = 55 \geq 5$ AND

4. $120 * (1 - (55/120)) = 65.004 \geq 5$

Step four, compute the confidence interval at the 1 - α level using the equation below.

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

Let's break down what each of these terms mean.

1. n_1 = first sample size
2. n_2 = second sample size
3. $\bar{p_1}$ = first proportion
4. $\bar{p_2}$ = second proportion
5. $z_{\alpha/2,df}$ = z value at $\alpha/2$ and df where df = number of samples

   To get this z value, we need to read a z table. But to make your lives easier, here are the z values for corresponding significance levels. We are only ever going to use confidence intervals at these levels.

Percentage Confidence	z*-Value
80	1.28
90	1.645
95	1.96
98	2.33
99	2.58

Now let's define them in this example:

1. n_1 = 80
2. n_2 = 120
3. $\bar{p_1}$ = 41/80 = 00.5125
4. $\bar{p_2}$ = 55/120 = 00.4583
5. $z_{\alpha/2,df}$ = z value at .05/2 = .025 and $df = 2$

   Because we are computing this confidence interval at the 95% level, the corresponding z value is, $z_{.025,2} = 1.96$

Now we can plug this all into our equation:

$$(00.5125 - 00.4583) \pm 1.96 * \sqrt{\frac{00.5125(1 - 00.5125)}{80} + \frac{00.4583(1 - 00.4583)}{120}}$$

$$0.0542 \pm 1.96 * \sqrt{\frac{00.5125(00.4875)}{80} + \frac{00.4583(0.5417)}{120}}$$

$$0.0542 \pm 1.96 * \sqrt{\frac{0.2498}{80} + \frac{0.2483}{120}}$$

$$0.0542 \pm 1.96 * \sqrt{0.0031 + 0.0021}$$

$$0.0542 \pm 1.96 * \sqrt{00.0052}$$

$$0.0542 \pm 1.96 * 00.0721$$

$$0.0542 \pm 0.1413$$

$$0.0542 + 0.1413 = 0.1955$$

$$0.0542 - 0.1413 = -0.0871$$

So, our confidence interval ranges from -0.0871 to 0.1955.

Step five, reject or do not reject the null.

Our confidence interval ranges from -0.0871 to 0.1955 which DOES go through the hypothesized value of a difference of 0. So, we do NOT reject the null hypothesis. Therefore, given the sample data, we cannot conclude that the two proportions are different and we cannot provide evidence that the support for Candidate A has changed.

Now we'll do the **P-Value Approach**.

1. State the null and alternative hypotheses. Our hypotheses will use the p value, because we are testing for a proportion.

   The hypotheses will always look like one of these three, with the hypothesized value being 0.

   - $H_0 : p_1 - p_2 = 0$
   - $H_A : p_1 - p_2 \neq 0$

   - $H_0 : p_1 - p_2 \geq 0$

- $H_A : p_1 - p_2 < 0$

- $H_0 : p_1 - p_2 \leq 0$
- $H_A : p_1 - p_2 > 0$

2. Verify that our data is normally distributed OR verify that the sample is large enough, which in this case means:

   $n_1\bar{p_1} \geq 5$ AND
   n_1(1-$\bar{p_1}$) ≥ 5 AND
   $n_2\bar{p_2} \geq 5$ AND
   n_2(1-$\bar{p_2}$) ≥ 5

   We have to check that each of these is ≥ 5 to make sure the sample is large enough, where n_1 is the first sample size, n_2 is the second sample size and p_1 is the proportion of the first sample and p_2 is the proportion of the second sample.

3. Choose the α value of .01, .05, or .10. This is the probability of making a Type I error.

4. Compute the test statistic using one of the two equations below.

   Use this formula if d_0 = 0

$$z = \frac{\bar{p_1} - \bar{p_2}}{\sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Let's break down what each of these terms mean.

- n_1 = first sample size
- n_2 = second sample size
- $\bar{p_1}$ = first proportion = x_1/n_1
- $\bar{p_2}$ = second proportion = x_2/n_2
- $\bar{p}$ = total proportion = $(x_1 + x_2)/(n_1 + n_2)$

Use this formula if $d_0 \neq 0$:

$$z = \frac{\bar{p_1} - \bar{p_2} - d_0}{\sqrt{\frac{\bar{p_1}(1-\bar{p_1})}{n_1} + \frac{\bar{p_2}(1-\bar{p_2})}{n_2}}}$$

Let's break down what each of these terms mean.

- n_1 = first sample size
- n_2 = second sample size
- $\bar{p_1}$ = first proportion = x_1/n_1
- $\bar{p_2}$ = second proportion = x_2/n_2
- d_0 = hypothesized value of the difference (this value is from our hypotheses)

5. Using the test statistic, calculate p value (using Excel). The Excel command differs depending on if it's a right tailed, left tailed or two-tailed test. Remember, a GREATER (>) than sign in the alternative hypothesis means we're doing a RIGHT-tailed test. A LESS (<) than sign in the alternative hypothesis means we're doing a LEFT-tailed test. A $\neq$ sign in the alternative means we are doing a TWO tailed test.

Operator in Alternative Hypothesis	Type of Test
> or $\geq$	Right-tailed Test
< or $\leq$	Left-tailed Test
$\neq$	Two-tailed Test

   For a right tailed, $= 1 - NORM.DIST(teststatistic, 0, 1, TRUE)$.

   For a left tailed, $= NORM.DIST(teststatistic, 0, 1, TRUE)$.

   For a two tailed, $= 2 * (1 - NORM.DIST(ABS(teststatistic), 0, 1, TRUE))$.

6. Reject or do not reject the null hypothesis.

   If the p value < α, reject the null.

   If the p value > α, do not reject the null.

Let's do an example, using the same scenario that we used in the Confidence Interval Approach. Candidate A appears to have gained support among voters in the People Choice Awards for best reality star. Three months ago, in a survey of 120 individuals, 55 said that they would vote for Candidate A. Today, in a survey of 80 individuals, 41 said that they would vote for Candidate A. At the 5% significance level, test if there is a difference between the two proportions. Take the hint here that "test if" language indicates the following statement is the alternative hypothesis.

Step one, state the null and alternative hypotheses. We currently believe there is no difference between the two proportions, or that p_1 - p_2 = 0. This is our null.

1. H_0 : p_1 - p_2 = 0
2. H_a : p_1 - $p_2 \neq 0$

Step two, Choose the α value of .01, .05, or .10. Here, we are told to compute at the 5% significance level.

Step three, verify that our data is normally distributed OR verify that the sample is large enough. We need to check these four things:

1. $n_1\bar{p_1} > 5$ AND
2. $n_1(1-\bar{p_1}) > 5$ AND
3. $n_2\bar{p_2} > 5$ AND
4. $n_2(1-\bar{p_2}) > 5$

Remember that n_1 is our first sample size (80), n_2 is our second sample size (120), $\bar{p_1}$ is the first proportion (41/80), and $\bar{p_2}$ is the second proportion (55/120).

It doesn't matter which is your "first" or "second" samples, so long as you keep it consistent throughout your calculations. In this case, I'm going to assign the survey from three months ago as our "second sample size."

We can verify these four equations are in fact, true:

1. $80 * (41/80) = 41 \geq 5$ AND
2. $80 * (1 - (41/80)) = 39 \geq 5$ AND
3. $120 * (55/120) = 55 \geq 5$ AND
4. $120 * (1 - (55/120)) = 65.004 \geq 5$

Step four, compute the test statistic.

In this case, our hypothesized difference between the two proportions (d_0) is 0, so we use this formula:

$$z = \frac{\bar{p_1} - \bar{p_2}}{\sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Let's break down what each of these terms mean.

1. n_1 = first sample size
2. n_2 = second sample size
3. $\bar{p_1}$ = first proportion = $\frac{x_1}{n_1}$
4. $\bar{p_2}$ = second proportion = $\frac{x_2}{n_2}$
5. $\bar{p}$ = total proportion = $\frac{x_1+x_2}{n_1+n_2}$

Now let's define them in this example

1. n_1 = 80
2. n_2 = 120
3. $\bar{p_1}$ = 41/80 = 00.5125
4. $\bar{p_2}$ = 55/120 = 00.4583
5. $\bar{p}$ = $\frac{41+55}{80+120}$ = $\frac{96}{200}$ = 00.48

Now we can plug this all into our equation:

$$z = \frac{(0.5125 - 0.4583)}{\sqrt{0.48(1-0.48)(\frac{1}{80} + \frac{1}{120})}}$$

$$z = \frac{(0.5125 - 0.4583)}{\sqrt{0.48(1-0.48)(0.0125 + 0.0083)}}$$

$$z = \frac{(0.0542)}{\sqrt{0.48(0.52)(0.0208)}}$$

$$z = \frac{(0.0542)}{\sqrt{0.0052}}$$

$$z = \frac{0.0542}{0.0721}$$

$$z = 0.7517$$

Step five, compute the p-value in Excel using the test statistic. Because we use the $\neq$ sign in the alternative hypothesis, we know this is a two tailed test. Therefore, we enter $2 * (1 - NORM.DIST(ABS(0.7517), 0, 1, TRUE))$ into Excel and get 0.4522 in return.

Step six, reject or do not reject the null.

In this case, our p value is greater than the alpha value, 0.4522 > .05, so we do not reject the null hypothesis. So, we do NOT reject the null hypothesis. Therefore, given the sample data, we cannot conclude that the two proportions are different and we cannot provide evidence that the support for Candidate A has changed.

7 Week 7: Mar. 6 - Mar. 10

7.1 Goodness-of-Fit Test For a Multinomial Experiment

In this section, we are testing to see if two OR MORE population proportions differ from what we hypothesize them to be. This type of test is called a **goodness of fit test for a multinomial experiment**. It is commonly referred to in the stats world as a "chi-square test for a multinomial experiment." A good example of this is a political contest. We may want to test the null hypothesis that candidates A,B, and C will receive 70%, 20% and 10% of the vote, respectively. So the null and alternative hypotheses in this case would look like:

H_0: p_A = .70 p_B = .20 p_C = .10
H_a: not all population proportions equal their hypothesized values

You have seen a multinomial experiment before in ECON 261. To recap your memory, an example of a multinomial experiment is: a consumer rates service at a restaurant as excellent, good, fair or poor (k = 4). Although this doesn't sound like a genuine Bill-Nye style science "experiment," we refer to it as a multinomial experiment in statistics because there are various possible outcomes. There are 4 possible outcomes: excellent, good, fair or poor. In a multinomial experiment, there are many possible outcomes, each has their own probability and all probabilities sum to 1.

Now back to a goodness of fit test for a multinomial experiment. When conducting this type of test, we take a random sample and test if the proportions of this sample (Candidate A,B,C) match our hypothesized population proportions (70%, 20% and 10%).

6 steps to conduct a goodness of fit test for a multinomial experiment:

1. State the null and alternative hypotheses.

   Remember that your null includes all the hypothesized proportion values, which sum to one. The alternative is simply that the population proportions do not equal their hypothesized values.

2. Choose the α value of .01, .05, or .10. This is the allowed probability of making a Type I error.

3. Verify that the expected frequencies for each category ≥ 5 to make sure our sample size is large enough.

   The expected frequency is $e_i = np_i$ where n is the sample size and p_i is the hypothesized proportion for each of the categories (whatever values we have in the hypothesis.) From the candidate example above, the hypothesized proportions are 70% for candidate A, 20% for candidate B and 10% for candidate C. Therefore, the expected frequencies is the product of these proportions and the size of the sample (n).

4. Compute the test statistic using the equation below, where $df = k - 1$ and k is the number of potential outcomes in the experiment.

$$\chi^2_{df} = \Sigma \frac{(o_i - e_i)^2}{e_i}$$

Let's explain each of these terms.

- o_i is the observed frequency of category i, this is what we see in the survey results
- e_i is the expected frequency of category i, this is what we expect it to be.

  This is computed as $e_i = np_i$, where p_i is the past frequency. These are the same expected frequencies we computed in step three.
- χ^2_{df} is the test statistic we compute, which is called the "Chi-squared" value at the degrees of freedom.

  $df = k - 1$ where k is the number of possible outcomes
- $\sum$ is the mathematical symbol that instructs us to sum across each i, we'll understand this better when I write it out below

5. Compute the p-value using Excel.

   These goodness of fit tests are all going to be right-tailed.

   In Excel, enter $= CHISQ.DIST.RT(teststatistic, df)$ where $df = k - 1$ and k is the number of potential outcomes in the experiment (if there are 3 candidates: A,B,C, then k = 3)

6. Reject or do not reject the null hypothesis.

   If the p value < α, reject the null.

   If the p value > α, do not reject the null.

Let's do an example.

A restaurant currently believes that the percentage of people that would rate the restaurant as Excellent, Good, Fair, or Poor are listed in the following table:

Excellent	Good	Fair	Poor
15%	30%	45%	10%

Recently, the restaurant surveyed a random sample of 250 people to rate the restaurant as Excellent, Good, Fair, or Poor. The number of respondents that selected each rating is shown below;

Excellent	Good	Fair	Poor
46	83	105	16

At the 5% significance level, we want to determine whether there is any difference between the hypothesized population proportions and the observed sample proportion values.

Step one, state the null and alternative hypotheses. Denote p_1, p_2, p_3 and p_4 as the hypothesized population proportions of the ratings for Excellent, Good, Fair, or Poor, respectively.

H_0: p_1 = .15 p_2 = .30 p_3 = .45 p_4 = .10
H_a: not all population proportions equal their hypothesized values

Step two, choose the α value of .01, .05, or .10. The problem states we are testing at the 5% significance level.

Step three, verify that the expected frequencies for each category are five or more. We do this by using the formula $e_i = np_i$ where n is the sample size and p_i is the hypothesized proportion for each of the categories (excellent, good, fair, poor)

- $e_{excellent}$ = 250 * .15 = 37.5
- e_{good} = 250 * .3 = 75
- e_{fair} = 250 * .45 = 112.5
- e_{poor} = 250 * .10 = 25

Yes, the expected frequencies are each ≥ 5.

Step four, compute the test statistic using the equation below.

$$\chi^2_{df} = \Sigma \frac{(o_i - e_i)^2}{e_i}$$

Let's explain each of these terms.

- o_i is the observed frequency of category i, this is what we see in the survey results
- e_i is the expected frequency of category i, this is what we expect it to be.

  This is computed as $e_i = np_i$, where p_i is the past frequency. These are the same expected frequencies we computed in step three.
- χ^2_{df} is the test statistic we compute, which is called the "Chi-squared" value at the degrees of freedom.

  $df = k - 1$ where k is the number of possible outcomes
- $\sum$ is the mathematical symbol that instructs us to sum across each i, we'll understand this better when I write it out below

Let's explain each of these terms in the example problem.

- $o_{excellent}$ = 46
- o_{good} = 83
- o_{fair} = 105

- o_{poor} = 16
- $e_{excellent}$ = 250 * .15 = 37.5
- e_{good} = 250 * .3 = 75
- e_{fair} = 250 * .45 = 112.5
- e_{poor} = 250 * .10 = 25
- df = 4 - 1 = 3 (we have 4 categories: excellent, good, fair, poor)

So when I plug all of these numbers in, this is the structure I'm going to follow:

$$\chi^2_3 = \frac{(o_{excellent} - e_{excellent})^2}{e_{excellent}} + \frac{(o_{good} - e_{good})^2}{e_{good}} + \frac{(o_{fair} - e_{fair})^2}{e_{fair}} + \frac{(o_{poor} - e_{poor})^2}{e_{poor}}$$

Now we can plug these into our summation sequence.

$$\chi^2_3 = \frac{(46 - 37.5)^2}{37.5} + \frac{(83 - 75)^2}{75} + \frac{(105 - 112.5)^2}{112.5} + \frac{(16 - 25)^2}{25}$$

$$\chi^2_3 = \frac{72.25}{37.5} + \frac{64}{75} + \frac{56.25}{112.5} + \frac{81}{25}$$

$$\chi^2_3 = 1.9267 + 0.8533 + 0.5000 + 3.2400$$

$$\chi^2_3 = 6.5200$$

Step five, compute the p-value using Excel. In Excel, we enter $= CHISQ.DIST.RT(6.520, 3)$ and get 0.0889 in return.

Step six, reject or do not reject the null hypothesis. The p value is greater than α because 0.0889 > 0.05, therefore we do NOT reject the null hypothesis. Therefore, we cannot conclude that the observed sample survey proportions differ from the hypothesized survey population proportions.

7.2 Chi-Square Test for Independence

In this section, we are doing a Chi-square test for independence, which is testing for the relationship between two categorical variables. A good example of this would be an attorney in a discrimination lawsuit trying to prove that a person's sex and promotion status are related. Sex and promotion status are two categorical variables, and this attorney is testing if they are related. The hypotheses in these tests would look like:

H_0: The two categorical variables (sex and promotion status) are independent.
H_a: The two categorical variables (sex and promotion status) are dependent.

6 steps to conduct a Chi-square test for independence:

1. State the null and alternative hypotheses.

   The null is always that the two categorical variables are independent.

2. Choose the α value of .01, .05, or .10. This is the allowed probability of making a Type I error.

3. Verify that the expected frequencies for each category ≥ 5 to make sure our sample size is large enough.

   Computing these frequencies takes a lot of work, I'll go into detail below.

4. Compute the test statistic using the equation below.

$$\chi^2_{df} = \Sigma\Sigma \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

   Let's explain each of these terms.

   - o_{ij} is the observed frequency of row i & column j

     This is what we observe based on the collected data.

   - e_{ij} is the expected frequency of row i & column j

     These are the same expected frequencies we computed in step 3.

   - χ^2_{df} is the test statistic we compute, which is called the "Chi-squared" value at the degrees of freedom.

     $df = (r-1)(c-1)$ where r is the number of rows and c is the number of columns.

5. Compute the p-value using Excel.

   Note in this test for independence, the test is always right-tailed.

   In Excel, enter $= CHISQ.DIST.RT(teststatistic, df)$ where $df = (r-1)(c-1)$ where r is the number of rows and c is the number of columns in the data table. As a student, you will be given these tables.

6. Reject or do not reject the null hypothesis.

   If the p value < α, reject the null.

   If the p value > α, do not reject the null.

Let's do an example.

We want to determine whether the likelihood of gym enrollment depends on the age of attendees. In other words, are age and gym enrollment related? We will conduct this test at the 5% significance level. The variable "Enrollment Outcome" has two possible categories: Enroll (E) and Not Enroll (N). The variable "Age Group" has three possible categories: Under 30 (U), Between 30 and 50 (B), and Over 50 (O). Each cell in this table represents an observed frequency o_{ij}, where the subscript ij refers to the ith row and the jth column. For example, o_{13} refers to the cell in the first row and the third column. Here, o_{13} = 44, or, equivalently, there are 44 individuals who are over 50 years of age and enrolled in the gym. In your homework, this is referred to as a "contingency table."

Enrollment Outcome	Age Group		
	Under 30 (*U*)	Between 30 and 50 (*B*)	Over 50 (*O*)
Enroll (*E*)	24	72	44
Not Enroll (*N*)	84	88	88

Step one, state the null and alternative hypotheses.

H_0: Age Group & Enrollment Outcome are independent.
H_a: Age Group & Enrollment Outcome are dependent.

Step two, choose the α value of .01, .05, or .10. The problem states that we are testing at the 5% significance level.

Step three, verify that the expected frequencies for each cell are five or more.

Computing expected frequencies requires more work than it did for the Chi-square test for a multinomial experiment. We essentially now need to predict what we think the frequencies are based on our null assumption: that there is no relationship between age group and enrollment. The formula we use for this:

$$e_{ij} = \frac{(Row\ i\ total)(Column\ j\ total)}{Sample\ Size}$$

The first step is to sum each of the two rows and each of the three columns, as well as find the total sample size. We can find the total sample size by summing up all the values in the table entirely. Each of the two row totals are under the "Row Total" column. Each of the three column totals are in the "Column Total" row. The total sample size here is 400.

Category	Under 30	Between 30 and 50	Over 50	Row Total
Enroll	24	72	44	140
Not enroll	84	88	88	260
Column Total	108	160	132	400

Now, we can multiply the row total and the column total for each row and column, and divide this product by the total sample size, which is 400. The total sample size is always in the bottom right corner of this table.

Category	Under 30	Between 30 and 50	Over 50	Row Total
Enroll	$\frac{(140)(108)}{400}$	$\frac{(140)(160)}{400}$	$\frac{(140)(132)}{400}$	140
Not enroll	$\frac{(260)(108)}{400}$	$\frac{(260)(160)}{400}$	$\frac{(260)(132)}{400}$	260
Column Total	108	160	132	400

Simplified, our table of expected frequencies looks like:

Category	Under 30	Between 30 and 50	Over 50
Enroll	37.80	56	46.2
Not enroll	70.2	104	85.8

We can confirm that each expected frequency is ≥ 5.

Step four, compute the test statistic using the formula below.

$$\chi^2_{df} = \Sigma\Sigma\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Let's explain each of these terms.

- o_{ij} is the observed frequency of row i & column j

  This is what we observe based on the collected data.

- e_{ij} is the expected frequency of row i & column j

  These are the same expected frequencies we computed in step 3.

- χ^2_{df} is the test statistic we compute, which is called the "Chi-squared" value at the degrees of freedom.

  $df = (r-1)(c-1)$ where r is the number of rows and c is the number of columns.

Let's recall what our observed frequencies are. These are the frequencies we were given in the problem, this is what we actually observed while collecting data at the gym. The table below is the table of observed frequencies.

Category	Under 30	Between 30 and 50	Over 50
Enroll	24	72	44
Not enroll	84	88	88

The table below is the expected frequencies. These are the numbers we would EXPECT based on our null hypothesis that age and gym enrollment are independent of each other.

Category	Under 30	Between 30 and 50	Over 50
Enroll	37.80	56	46.2
Not enroll	70.2	104	85.8

Now we can calculate our test statistic, using this formula:

$$\chi^2_{df} = \Sigma\Sigma \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Don't let the double summation operators give you anxiety. All this equation is telling us is that we need to subtract each expected frequency from the observed frequency, square this difference, and divide it by the expected frequency. We'll get 6 of these, since there are 6 cells in our table, then we'll calculate the sum of all 6. And this final sum is our test statistic. I'll do it in the table to make it easier to visualize.

Category	Under 30	Between 30 and 50	Over 50
Enroll	$\frac{(24-37.8)^2}{37.8}$	$\frac{(72-56)^2}{56}$	$\frac{(44-46.2)^2}{46.2}$
Not enroll	$\frac{(84-70.2)^2}{70.2}$	$\frac{(88-104)^2}{104}$	$\frac{(88-85.8)^2}{85.8}$

Now after we simplify all of these calculations, we get:

Category	Under 30	Between 30 and 50	Over 50
Enroll	5.0381	4.5714	0.1048
Not enroll	2.7128	2.4615	0.0564

Now we add all 6 of these values:

$$5.0381 + 4.5714 + 0.1048 + 2.7128 + 2.4615 + 0.0564 = 14.9450$$

Our test statistic is 14.9450.

Step five, compute the p-value using Excel. In Excel, enter $= CHISQ.DIST.RT(14.945, df)$ where $df = (r-1)(c-1)$ where r is the number of rows and c is the number of columns in the data table. This can also be interpreted as r is the number of categories for the first variable and c is the number of categories for the second variable. Therefore, $df = (2-1)(3-1) = 1 * 2 = 2$. So we enter $= CHISQ.DIST.RT(14.9450, 2)$ in Excel and get 0.0006 in return.

Step six, reject or do not reject the null hypothesis. The p-value is less than alpha in this case because 0.0006 < 0.05, therefore we reject the null hypothesis. Because the null is that age and gym enrollment are independent of each other, we are stating there is no way in HELL that age and gym enrollment are not related. At the 5% significance level, we conclude that there IS a relationship between these two variables.

8 Week 8: Mar. 13 - Mar. 17

8.1 One-Way ANOVA Test

In this section, we'll cover a One-Way Analysis of Variance (ANOVA) test. This tests for a statistical difference between the means of three or more populations under independent sampling. A good real-life application of this is if we had a group of volunteers for a study, split them into three groups (A,B C) randomly, and then gave each group a different type of running shoe. Each group received instructions to run 1 mile. The outcome variable we measure is each individual's mile time. In the end, we took the average time for each group and compared them. We would conduct a one-way ANOVA test to determine if the differences between these three means is due to pure chance or rather, some statistical significance. If these differences are due to statistical significance, then we would expect to find these patterns beyond just the sample data. We would find these differences in the population data as well.

The one-way ANOVA test is used to test c population means given the following assumptions:

1. The populations are normally distributed.
2. The population standard deviations are unknown but assumed equal.
3. The samples are selected independently.

If you remember from early on, we covered the Inference of Difference Between Two Means where we tested for the difference between two population means. So why can't we just do that test to compare the difference between various combinations of two of the three running groups? Why not compare running group A vs. B then B vs. C then A vs. C? It's because every time we test for the difference between two means, we run the risk of committing a Type I error. So every time we run a test using the same population and sample data, we increase the risk of a Type I error. A one-way ANOVA test makes it easy to test between three or more population means simultaneously without increasing the risk of a Type I error.

10 Steps to Conduct a One-Way ANOVA test:

1. State the null and alternative hypotheses.

   The null is that the means of the populations are equal to each other. The alternative is always that the means are not equal.

2. Choose the α value of .01, .05, or .10. This is the allowed probability of making a Type I error.

3. Calculate the grand mean. The grand mean is the average of all observations across all treatment groups. Like in the running example, this is the average running time for runners across all 3 groups, denoted as $\bar{\bar{x}}$.

4. Calculate the Sum of Squares due to Treatments (SSTR) using the following formula:

$$SSTR = \Sigma_{i=1}^{c} n_i(\bar{x}_i - \bar{\bar{x}})^2$$

Let me break down what each of these terms mean:

- c = number of treatment groups
- n_i = number of observations in each sample
- $\bar{x}_i$ = mean of each sample
- $\bar{\bar{x}}$ = grand mean

5. Calculate the Between-Treatments Estimate, also known as the Mean Square for Treatments (MSTR). This measures how much "action" or variance, we have going on between each group of runners. We use the following formula:

$$MSTR = \frac{SSTR}{c-1}$$

6. Calculate the error sum of squares, also known as the SSE. We use the following formula for this:

$$SSE = \Sigma_{i=1}^{c}(n_i - 1)s_i^2$$

Let me break down what each of these terms mean:

- c = number of treatment groups
- n_i = number of observations in each sample
- s_i^2 = variance for one of the sample groups

7. Calculate the Within-Treatments Estimate, also known as the Mean Square Error (MSE). This measures how much "action" or variance, we have going on WITHIN each group of runners. We use the following formula to compute the MSE:

$$MSE = \frac{SSE}{n_T - c}$$

Where n_T represents the total number of observations.

8. Compute the test statistic using the formula below:

$$F_{df_1,df_2} = \frac{MSTR}{MSE}$$

Where $df_1 = c - 1$ and $df_2 = n_T - c$ where n_T is the total sample size and c is the number of treatment groups.

9. Compute the p-value using Excel. One-way ANOVA tests are always right-tailed.

   In Excel, enter $= F.DIST.RT(teststatistic, df_1, df_2)$ where $df_1 = c - 1$ and $df_2 = n_T - c$

10. Reject or do not reject the null hypothesis.

    If the p value < α, reject the null.

    If the p value > α, do not reject the null.

It is important to understand how statisticians visually organize their data in a one-way ANOVA table. The table below is a demonstration of how to organize all of the components of your one-way ANOVA test.

Format of a One-way ANOVA table					
Source of Variation	**SS**	**df**	**MS**	**F**	**p-value**
Between Groups	SSTR	c-1	MSTR = SSTR/(c-1)	MSTR/MSE	F.DIST.RT(test statistic, DF1, DF2)
Within Groups	SSE	nT-c	MSE = SSE/(nT - c)		
Total	SST	nT-1			
			DF1 = c-1		
			DF2 = nT - c		

Let's do an example. We want to determine whether differences exist in the mean monthly sales among these three store layouts at the 5% significance level. Assume that the population data is normally distributed. There are three samples with 10 stores in each, meaning we have a total population of 30 observations.

Table 13.1 outlines monthly sales.

TABLE 13.1 Monthly Sales (in $ millions)

Layout 1	Layout 2	Layout 3
1.3	2.0	2.3
1.8	2.2	2.3
⋮	⋮	⋮
2.0	1.8	2.2
$\bar{x}_1 = 1.92$ $s_1^2 = 0.0973$	$\bar{x}_2 = 2.08$ $s_2^2 = 0.1062$	$\bar{x}_3 = 2.42$ $s_3^2 = 0.0373$

Step one, state the null and alternative hypotheses. We are comparing the mean sales for each of the three layouts.

H_0: $\mu_1 = \mu_2 = \mu_3$
H_a: Not all population means are equal.

Step two, choose the α value of .01, .05, or .10. The problem stated an α of .05.

Step three, calculate the grand mean. The grand mean is the average of all observations across all treatment groups. Like in the running example, this is the average running time for runners across all 3 groups, denoted as $\bar{\bar{x}}$. An easy way to do this is to sum across the averages of each sample, and divide by 3.

$$\bar{\bar{x}} = \frac{1.92 + 2.08 + 2.42}{3} = 2.14$$

Step four, calculate the Sum of Squares due to Treatments (SSTR). We use the following formula:

$$SSTR = \Sigma_{i=1}^{c} n_i(\bar{x_i} - \bar{\bar{x}})^2$$

Let me break down what each of these terms mean:

- c = number of treatment groups
- n_i = number of observations in each sample
- $\bar{x_i}$ = mean of each sample
- $\bar{\bar{x}}$ = grand mean

So, because we have 3 groups, that means the formula will look like this:

$$SSTR = n_1(\bar{x_1} - \bar{\bar{x}})^2 + n_2(\bar{x_2} - \bar{\bar{x}})^2 + n_3(\bar{x_3} - \bar{\bar{x}})^2$$

Now that we can visualize what the whole thing will look like, I'm going to define the value of each individual term.

- n_1 = 10
- n_2 = 10

- n_3 = 10
- $\bar{x_1}$ = 1.92
- $\bar{x_2}$ = 2.08
- $\bar{x_3}$ = 2.42
- $\bar{\bar{x}}$ = 2.14

Ok now we have all of our puzzle pieces, so we can plug them into the formula and solve.

$$SSTR = 10(1.92 - 2.14)^2 + 10(2.08 - 2.14)^2 + 10(2.42 - 2.14)^2 =$$

$$SSTR = 10(-0.22)^2 + 10(-0.06)^2 + 10(0.28)^2 =$$

$$SSTR = 10(0.0484) + 10(0.0036) + 10(0.0784) =$$

$$SSTR = (0.484) + (0.036) + (0.784) = 1.304$$

Step five, calculate the Between-Treatments Estimate, also known as the Mean Square for Treatments (MSTR). We use the following formula:

$$MSTR = \frac{SSTR}{c-1} =$$

$$MSTR = \frac{1.304}{3-1} =$$

$$MSTR = \frac{1.304}{2} = 0.652$$

Step six, calculate the Error Sum of Squares (SSE). We use the following formula:

$$SSE = \sum_{i=1}^{c}(n_i - 1)s_i^2$$

Let me break down what each of these terms mean:

- c = number of treatment groups
- n_i = number of observations in each sample

- s_i^2 = variance for one of the sample groups

So, because we have 3 groups, that means the formula will look like this:

$$SSE = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2$$

Now that we can visualize what the whole thing will look like, I'm going to define the value of each individual term.

- n_1 = 10
- n_2 = 10
- n_3 = 10
- s_1^2 = 0.0973
- s_2^2 = 0.1062
- s_3^2 = 0.0373

Ok, now we have all of our puzzle pieces, so we can plug in.

$$SSE = (10 - 1)0.0973 + (10 - 1)0.1062 + (10 - 1)0.0373 =$$

$$SSE = 9 * 0.0973 + 9 * 0.1062 + 9 * 0.0373 =$$

$$SSE = 0.876 + 0.956 + 0.336 = 2.167$$

$$SSE = 2.167$$

Step seven, calculate the Within-Treatments Estimate, also known as the Mean Square Error (MSE). We use the following formula:

$$MSE = \frac{SSE}{n_T - c}$$

$$MSE = \frac{2.167}{30 - 3} = 0.0803$$

$$MSE = \frac{2.167}{27} = 0.0803$$

Step eight, compute the test statistic. We use the formula below:

$$F_{df_1,df_2} = \frac{MSTR}{MSE}$$

$$F_{df_1,df_2} = \frac{0.625}{0.0803}$$

$$F_{df_1,df_2} = 8.1196$$

$$F_{2,27} = 8.1196$$

Where $df_1 = c - 1 = 3 - 1$ and $df_2 = 30 - 3 = 27$ where n_T is the total sample size and c is the number of treatment groups.

Step nine, compute the p-value using Excel. We enter $= F.DIST.RT(8.1196, 2, 27)$ into Excel and get a p value of 0.0017 in return.

Step ten, reject or do not reject the null hypothesis. Our p value is less than alpha because 0.0017 < 0.005, therefore we reject the null hypothesis. Therefore, at the 5% significance level, we conclude that the mean monthly sales differ between the three store layouts.

8.2 Multiple Comparison Methods

As noted earlier, while the ANOVA test determines that not all population means are equal, it does not indicate which ones differ. To find out which population means differ requires further analysis. In the grocery store example, we concluded that the means were not all equal. This means that all three were not the same, but it's possible that two of them were the same and only one differed. To determine which one differs, we can do one of two tests:

1. Fisher's Least Significant Difference (LSD) Method
2. Tukey's Honestly Significant Difference (HSD) Method

We'll learn about both.

It is important to note that we can ONLY implement Fisher's or Tukey's methods IF the ANOVA test determines that not all population means are equal. If the ANOVA test does not reject the null hypothesis that the means are equal, then we do not implement Fisher's or Tukey's methods.

We are learning the confidence interval approach for both methods.

4 Steps of Fisher's Least Significant Difference (LSD) Method using the Confidence Interval Approach:

1. State the null and alternative hypotheses.

   Because we are testing for the differences between each sample mean for every treatment group, we'll have MULTIPLE null and alternatives. Our sets of hypotheses would look like this if we're testing for statistical differences between three population means:

   - $H_0 : \mu_1 - \mu_2 = 0$
   - $H_A : \mu_1 - \mu_2 \neq 0$
   - $H_0 : \mu_1 - \mu_3 = 0$
   - $H_A : \mu_1 - \mu_3 \neq 0$
   - $H_0 : \mu_2 - \mu_3 = 0$
   - $H_A : \mu_2 - \mu_3 \neq 0$

2. Choose the α value of .01, .05, or .10. This is the allowed probability of making a Type I error.

3. Compute the confidence intervals at 1 - α for each set of hypotheses using the formula below.

$$(\bar{x_i} - \bar{x_j}) \pm t_{\alpha/2, n_T - c} \sqrt{MSE(\frac{1}{n_1} + \frac{1}{n_j}}$$

   Let's break down what each of these terms mean:

   (a) $\bar{x_i}$ is the mean of one group

   (b) $\bar{x_j}$ is the mean of the second group

   (c) $t_{\alpha/2, n_T - c}$ is the t value at $\alpha/2$ and $n_T - c$ where n_T is the total number of observations and c is the number of treatment groups.

   We'll use this t value calculator to do this.

   If you would like to be traditional, you can use this t table distribution and manually interpret the table. But the focus of this class is not interpreting these tables, so I am fine with you using online t value calculators. On the test, I will give you the t value in a case like this.

   (d) MSE is the mean squared error that we learned to calculate in the one-way ANOVA test, I'll copy the formula here.

$$MSE = \frac{SSE}{n_T - c}$$

(e) n_i is the number of observations in one group

(f) n_j is the number of observations in the second group

4. Reject or do not reject the null hypothesis for each set of hypotheses.

   If the confidence interval goes through the hypothesized value of 0, do not reject the null.

   If the confidence interval DOES NOT go through the hypothesized value of 0, reject the null.

Let's do an example. We'll use the same grocery store example from the previous section, I'll copy the information here. We want to determine whether some differences exist in the mean monthly sales of a grocery store depending on one of three possible store layouts. Table 13.1 outlines monthly sales. There are three samples with 10 stores in each, meaning we have a total population of 30 observations. Use the information in the table to calculate 95% confidence intervals for the difference between all possible pairings.

TABLE 13.1 Monthly Sales (in $ millions)

Layout 1	Layout 2	Layout 3
1.3	2.0	2.3
1.8	2.2	2.3
⋮	⋮	⋮
2.0	1.8	2.2
$\bar{x}_1 = 1.92$ $s_1^2 = 0.0973$	$\bar{x}_2 = 2.08$ $s_2^2 = 0.1062$	$\bar{x}_3 = 2.42$ $s_3^2 = 0.0373$

Step one, state the null and alternative hypotheses. In this case, we have 3 samples and so we'll have 3 sets of matching hypotheses. Each set needs to compare two means, because we are trying to figure out which means differ from each other. So, we want to compare μ_1 to μ_2, and then μ_1 to μ_3 and then μ_2 to μ_3.

- $H_0 : \mu_1 - \mu_2 = 0$
- $H_A : \mu_1 - \mu_2 \neq 0$
- $H_0 : \mu_1 - \mu_3 = 0$
- $H_A : \mu_1 - \mu_3 \neq 0$
- $H_0 : \mu_2 - \mu_3 = 0$
- $H_A : \mu_2 - \mu_3 \neq 0$

Step two, choose the α value of .01, .05, or .10. The problem states that we'll compute 95% confidence intervals, therefore we must have an α of .05.

Step three, compute the confidence intervals for each set of hypotheses using the equation below.

$$(\bar{x_i} - \bar{x_j} \pm t_{\alpha/2,n_T-c}\sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})}$$

Let's break down what each of these terms mean:

1. $\bar{x_i}$ is the mean of one group
2. $\bar{x_j}$ is the mean of the second group
3. $t_{\alpha/2,n_T-c}$ is the t value at $\alpha/2$ and $n_T - c$ where n_T is the total number of observations and c is the number of treatment groups.

   We'll use this t value calculator to do this.

   If you would like to be traditional, you can use this t table distribution and manually interpret the table. But the focus of this class is not interpreting these tables, so I am fine with you using online t value calculators. On the test, I will give you the t value in a case like this.
4. MSE is the mean squared error that we learned to calculate in section 13.1, I'll copy the formula here.

$$MSE = \frac{SSE}{n_T - c}$$

5. n_i is the number of observations in one group
6. n_j is the number of observations in the second group

Let's define each of these terms in our problem. We need to compute three confidence intervals for each set of hypotheses, so let's deal with the first set now, that is:

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

These hypotheses are looking at the mean monthly sales for store layout 1 and store layout 2. Now I'll define each of the terms that we need to compute the confidence interval for the first set of hypotheses:

- $\bar{x_1} = 1.92$
- $\bar{x_2} = 2.08$
- n_1 = 10
- n_2 = 10
- $t_{\alpha/2,nT-c}$ is the t value at $.05/2 = .025$ and $n_T - c = 30 - 3 = 27$ where n_T is the total number of observations and c is the number of treatment groups. There are 3 treatments groups in this case, and 10 observations in each group.

  Using the t value calculator, enter 0.05 for α and 27 for degrees of freedom. From the resulting options, use the two-tailed value, which is 2.0518.
- MSE is the mean squared error that we learned to calculate previously in the following equation. Because the problem is the same, we can look back at our notes and see that the MSE was 0.0803.

Now we can plug all of these values into our equation for a confidence interval:

$$(1.92 - 2.08) \pm 2.0518\sqrt{.0803(\frac{1}{10} + \frac{1}{10})}$$
$$-0.16 \pm 2.0518\sqrt{0.0803(0.2)}$$
$$-0.16 \pm 2.0518\sqrt{.01606}$$
$$-0.16 \pm 2.0518 * 0.1267$$
$$-.16 \pm .26$$
$$-0.16 + 0.26 = 0.10$$
$$-0.16 - 0.26 = -0.42$$

Therefore, our confidence interval ranges from -0.42 to 0.10.

Now we can do the same process for the remaining two hypotheses.

$$\mu_1 - \mu_3 = (1.92 - 2.42) \pm 2.0518\sqrt{.0803(\frac{1}{10} + \frac{1}{10})}$$
$$\mu_1 - \mu_3 = -.5 \pm .26 = [-.76, -.24]$$
$$\mu_2 - \mu_3 = (2.08 - 2.42) \pm 2.0518\sqrt{.0803(\frac{1}{10} + \frac{1}{10})}$$
$$\mu_2 - \mu_3 = -.34 \pm .26 = [-.6, -.08]$$

So the three confidence intervals we have are:

1. $\mu_1 - \mu_2 = [-0.42, 0.10]$
2. $\mu_1 - \mu_3 = [-0.76, -0.24]$
3. $\mu_2 - \mu_3 = [-0.6, -0.08]$

Step four, do not reject or reject the null hypothesis for each set of hypotheses. We do not reject the null hypothesis if the confidence interval contains the hypothesized value of 0.

1. $\mu_1 - \mu_2 = 0$

   Do not reject this null because the confidence interval contains 0.

2. $\mu_1 - \mu_3 = 0$

   Reject this null because the confidence interval does not contain 0.

3. $\mu_2 - \mu_3 = 0$

   Reject this null because the confidence interval does not contain 0.

The only null hypothesis we did not reject was that $H_0 : \mu_1 - \mu_2 = 0$. Therefore, we cannot conclude that mean monthly sales differ between Layout 1 and Layout 2.

Now, we move onto **Tukey's Honestly Significant Difference (HSD) Method**.

4 Steps of Tukey's Honestly Significant Difference (HSD) Method using the Confidence Interval Approach:

1. State the null and alternative hypotheses.

   Because we are testing for the differences between each sample mean for every treatment group, we'll have several null and alternatives. Our sets of hypotheses would look like this if we're testing for statistical differences between three population means:

   - $H_0 : \mu_1 - \mu_2 = 0$
   - $H_A : \mu_1 - \mu_2 \neq 0$
   - $H_0 : \mu_1 - \mu_3 = 0$
   - $H_A : \mu_1 - \mu_3 \neq 0$
   - $H_0 : \mu_2 - \mu_3 = 0$
   - $H_A : \mu_2 - \mu_3 \neq 0$

2. Choose the α value of .01, .05, or .10. This is the allowed probability of making a Type I error.

3. Compute the confidence intervals at 1 - α for each set of hypotheses using one of the equations below.

   We use this equation if our data is balanced. Balanced data is when there are the same number of samples in each treatment group.

$$(\bar{x_i} - \bar{x_j}) \pm q_{\alpha,(c,n_T,c)} \sqrt{\frac{MSE}{n}}$$

   We use this equation if our data is not balanced. Unbalanced data is when there are a different number of samples in each treatment group.

$$(\bar{x_i} - \bar{x_j}) \pm q_{\alpha,(c,n_T,c)} \sqrt{\frac{MSE}{2}(\frac{1}{n_i} + \frac{1}{n_j})}$$

   Let's break down what each of these terms mean:

   (a) $\bar{x_i}$ is the mean of one group

   (b) $\bar{x_j}$ is the mean of the second group

   These correspond to each set of hypotheses.

   (c) $q_{\alpha,(c,n_T-c)}$ is the q value at α and (c, n_T-c) where n_T is the total number of observations and c is the number of treatment groups.

   We'll use this online q value calculator to find the q value.

   Or, if you want to be traditional, you can use this q table. On the test, I will give you the q value in a case like this.

   (d) MSE is the mean squared error that we learned to calculate before.

   (e) n_i is the number of observations in one group

   (f) n_j is the number of observations in the second group

   Because we have the same number of items in each group, we just have one n = 10.

4. Do not reject or reject the null hypothesis for each set of hypotheses.

   If the confidence interval goes through the hypothesized value of 0, do not reject the null.

   If the confidence interval does not go through the hypothesized value of 0, reject the null.

Let's do an example. We'll use the same grocery store example from the previous section, I'll copy the information here. We want to determine whether some differences exist in the mean monthly sales of a grocery store depending on one of three possible store layouts. Table 13.1 outlines monthly sales. There are three samples with 10 stores in each, meaning we have a total population of 30 observations. Use the information in the table to calculate 95% confidence intervals for the difference between all possible pairings.

TABLE 13.1 Monthly Sales (in $ millions)

Layout 1	Layout 2	Layout 3
1.3	2.0	2.3
1.8	2.2	2.3
⋮	⋮	⋮
2.0	1.8	2.2
$\bar{x}_1 = 1.92$ $s_1^2 = 0.0973$	$\bar{x}_2 = 2.08$ $s_2^2 = 0.1062$	$\bar{x}_3 = 2.42$ $s_3^2 = 0.0373$

Step one, state the null and alternative hypotheses. We should have one set of hypotheses for each sample group. In this case, we have three sample groups so we'll have three sets of null and alternative hypotheses.

- $H_0 = \mu_1 - \mu_2 = 0$
- $H_A = \mu_1 - \mu_2 \neq 0$
- $H_0 = \mu_1 - \mu_3 = 0$
- $H_A = \mu_1 - \mu_3 \neq 0$
- $H_0 = \mu_2 - \mu_3 = 0$
- $H_A = \mu_2 - \mu_3 \neq 0$

Step two, choose the α value of .01, .05, or .10. The problem states that we'll compute 95% confidence intervals, therefore we must have an α of .05.

Step three, compute the confidence intervals for each set of hypotheses. Because we have the same number of observations in each sample (10 stores per layout,) we have balanced data and will use the first equation. We need to compute three confidence intervals, so let's deal with the first set of hypotheses that is testing for differences between the first and second store layouts:

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

We will use this formula to compute each interval:

$$(\bar{x_i} - \bar{x_j}) \pm q_{\alpha,(c,n_T,c)}\sqrt{\frac{MSE}{n}}$$

Let's break down what each of these terms mean:

1. $\bar{x_i}$ is the mean of one group
2. $\bar{x_j}$ is the mean of the second group

   These correspond to each set of hypotheses.
3. MSE is the mean squared error that we learned to calculate before.
4. n is the number of observations in each group
5. $q_{\alpha,(c,n_T-c)}$ is the q value at α and $(c, n_T - c)$ where n_T is the total number of observations and c is the number of treatment groups.

   We'll use this online q value calculator to find the q value.

   Or, if you want to be traditional, you can use this q table. On the test, I will give you the q value in a case like this.

Using the formula for the confidence interval for balanced data, let's break down what each of these terms mean in our problem.

1. $\bar{x_1} = 1.92$
2. $\bar{x_2} = 2.08$
3. $n = 10$
4. $q_{\alpha,(c,n_T-c)}$ is the q value at .05 and $(c, n_T - c) = (3, 30 - 3)$ where n_T is the total number of observations and c is the number of treatment groups.

   We'll use this online q value calculator to find the q value. The calculator gives you two values: 3.51 and 4.49. The 3.51 is correct because we are using an $\alpha = 0.05$.
5. MSE is the mean squared error that we learned to calculate before. Because we are using the same word problem, we can look back at our notes and see that the MSE is 0.0803.

Now we can plug in all of these values to compute confidence intervals for each set of hypotheses, using the formula for balanced data:

$$(\bar{x_i} - \bar{x_j}) \pm q_{\alpha,(c,n_T,c)}\sqrt{\frac{MSE}{n}}$$

$$(1.92 - 2.08) \pm 3.51\sqrt{\frac{.0803}{10}}$$
$$-.16 \pm 3.51\sqrt{.00803}$$
$$-.16 \pm 3.51 * .0896$$
$$-.16 \pm 0.314$$
$$-.16 + 0.314 = 0.15$$
$$-.16 - 0.314 = -0.47$$
$$-.16 \pm .314 = [-0.47, 0.15]$$

Now we can do the same process for the remaining two hypotheses.

$$\mu_1 - \mu_3 = (1.92 - 2.42) \pm 3.51\sqrt{\frac{.0803}{10}}$$
$$\mu_1 - \mu_3 = -.5 \pm .314 = [-0.81, -0.18]$$
$$\mu_2 - \mu_3 = (2.08 - 2.42) \pm 3.51\sqrt{\frac{.0803}{10}}$$
$$\mu_2 - \mu_3 = -.34 \pm .314 = [-0.65, -0.02]$$

So the three confidence intervals we have are:

1. $\mu_1 - \mu_2 = [-0.47, 0.15]$
2. $\mu_1 - \mu_3 = [-0.81, -0.18]$
3. $\mu_2 - \mu_3 = [-0.65, -0.02]$

Step four, reject or do not reject the null hypothesis for each set of hypotheses. We do not reject the null hypothesis if the confidence interval contains the hypothesized value of 0.

1. $\mu_1 - \mu_2 = 0$

   Do not reject this null because the confidence interval goes through 0.

2. $\mu_1 - \mu_3 = 0$

   Reject this null because the confidence interval does not go through 0.

3. $\mu_2 - \mu_3 = 0$

   Reject this null because the confidence interval does not go through 0.

Therefore, the only null hypothesis we do not reject is $H_0 : \mu_1 - \mu_2 = 0$. Therefore, at the 5% significance level, we cannot conclude that mean monthly sales differ between Layout 1 and Layout 2. However, we can conclude that the mean monthly sales differ between Layouts 1 and 3, and also between Layouts 2 and 3.

9 Week 9: Mar. 20 - Mar. 24

9.1 NO CLASS SPRING BREAK

10 Week 10: Mar. 27 - Mar. 31

10.1 Two-Way ANOVA Test: No Interaction

A one-way ANOVA test is used to compare population means based on one categorical variable or one factor. For instance, we can use a one-way ANOVA test to determine whether differences exist in average miles per gallon depending on the brand name of hybrid cars. In this section, we are learning about a **two-way ANOVA test**, which extends the analysis by measuring the effects of two factors simultaneously. Suppose we want to determine if the brand of a hybrid car (Hyundai, Toyota, Honda) AND the octane level (78,89,91) of gasoline influence the gas economy of the vehicle (measured in average miles per gallon). A one-way ANOVA test is able to assess EITHER the brand effect OR the octane-level effect in isolation, a two-way ANOVA test is able to assess the effect of both factors in action at the same time. An added requirement for a two-way ANOVA test is that all groups must have the same sample size. **A two-way ANOVA test is used to simultaneously examine the effect of two factors on the population mean.**

This hypothesis test is an absolute pain in the ass to do manually. So, I am going to do it all manually for you ONCE below, so you can see the work. Then, I'm going to show you how to do it in Excel and interpret the results. For the homework and test, I will only expect you to know how to do it in Excel and then understand the results.

12 Steps to Conduct a Two-Way ANOVA test without using Excel

1. State the null and alternative hypotheses.

   We'll have two sets of null and alternative hypotheses for each factor. The alternative is always that the means of each of the groups are not equal. A good example of hypotheses in using the hybrid car example above looks like this:

   - $H_0 : \mu_{Hyundai} = \mu_{Toyota} = \mu_{Honda}$
   - H_A : not all population means are equal
   - $H_0 : \mu_{Octane78} = \mu_{Octane89} = \mu_{Octane91}$
   - H_A : not all population means are equal

2. Choose the α value of .01, .05, or .10. This is still the probability of making a Type I error.

3. Calculate the grand mean, $\bar{\bar{x}}$. The "grand mean" is the average of all observations across all treatment groups.

4. Calculate the sample mean for each of the 6 sample groups (referring back to the car example), $x_{Hyundai}$, x_{Toyota}, x_{Honda}, $x_{Octane78}$, $x_{Octane89}$, $x_{Octane91}$.

5. Calculate the Sum of Squares for Factor A (SSA) and Factor B (SSB). This measures how much "action" or variance, we have going on between each car brand (factor A) and each octane level (factor B).

$$SSA = r\sum_{i=1}^{c}(\bar{x}_i - \bar{\bar{x}})^2$$

$$SSB = c\sum_{j=1}^{r}(\bar{x}_j - \bar{\bar{x}})^2$$

Where r is the number of categories in factor B and c is the number of categories in factor A.

6. Calculate the Mean Square for Factor A (MSA) and Factor B (MSB):

$$MSA = \frac{SSA}{c-1}$$

$$MSB = \frac{SSB}{r-1}$$

Where c is the number of categories in factor A and r is the number of categories in factor B.

7. Calculate the Total Sum of Squares (SST) using the following equation:

$$SST = \sum_{i=1}^{c}\sum_{j=1}^{r}(x_{ij} - \bar{\bar{x}})^2$$

This is just a fancy equation that wants you to take each value in your table (each x_{ij} and subtract the grand mean, $\bar{\bar{x}}$, then square each of these differences. Then, add them all together. The most important thing to understand the SST value is that it measures how much "action" we have going on in our dataset. It measures how much each data point deviates from the mean of the entire dataset. A high SST indicates there is high variability across the data.

8. Calculate the Error Sum of Squares (SSE). This measures how much "action" there is within each factor group. It measures how much each data point within each factor group deviates from the mean of that factor group.

$$SSE = SST - (SSA + SSB)$$

9. Compute the Mean Square Error (MSE) using the formula below:

$$MSE = \frac{SSE}{n_T - c - r + 1}$$

Where n_T is the total number of observations in our data, c is the number of categories in factor A and r is the number of categories in factor B.

10. Compute the test statistics using the equations below for factor groups A and B.

$$F_{df_1, df_2} = \frac{MSA}{MSE}$$

$$\text{Where } df_1 = c - 1 \text{ and } df_2 = n_T - c - r + 1$$

$$F_{df_1, df_2} = \frac{MSB}{MSE}$$

$$\text{Where } df_1 = r - 1 \text{ and } df_2 = n_T - c - r + 1$$

11. Compute the p-value using Excel.

    Note in these ANOVA tests, the test is always right-tailed.

    In Excel, enter $= F.DIST.RT(teststatistic, df_1, df_2)$.

12. Reject or do not reject the null hypothesis.

    If the p value < α, reject the null.

    If the p value > α, do not reject the null.

Now that we've seen the brutal two-way ANOVA process, we'll learn how to do this in Excel using an example.

Julia Hayes is an undergraduate who is completely undecided as to what career she should pursue. To help in her decision process, she wants to determine whether or not there are significant differences in annual incomes depending on the field of employment AND education level. She evaluates the following three fields: journalism, finance and medical. She randomly interviews an 4 individuals in each industry. Each of the 4 individuals have different levels of education (no high school, high school, bachelor's and master's). Please use the Excel file on Canvas, titled "Income by industry education" to apply this example. The data in this table is in the $1,000s.

The goal of the analysis is to answer the following two questions:

1. At the 5% significance level, does annual income differ by field of employment?

2. At the 5% significance level, does field of employment differ by annual income?

Step one, state the null and alternative hypotheses. Because we are comparing the means for factor A (employment field) and means for factor B (education level,) we'll have two sets of hypotheses.

1. $H_0 : \mu_{education} = \mu_{financial} = \mu_{medical}$

2. H_A not all population means are equal

3. $H_0 : \mu_{NoHighSchool} = \mu_{HighSchool} = \mu_{Bachelors} = \mu_{Masters}$

4. H_A not all population means are equal

Step two, choose the α value of .01, .05, or .10. The problem states that we'll use an α = .05 for both questions.

Step three, make sure your Excel is prepared to analyze data. Click on the "Data" tab on the top, then click on "Analysis Tools." If you have a Mac, you need to check the "Analysis ToolPak" box. If you have a PC, you need to "Browse" for the "Analysis ToolPak." Once you have this installed, you're ready to go.

Step four, run the two way Anova test. To do this, click on the "Data" tab on the top, then click on "Data Analysis." From this list, select "Anova: Two-Factor Without Replication." For the input range, select the entire table of values (B3:E7) and click the "Labels" box to indicate that we have included the labels of each category in our selection. The alpha should be 0.05. Then, click "ok."

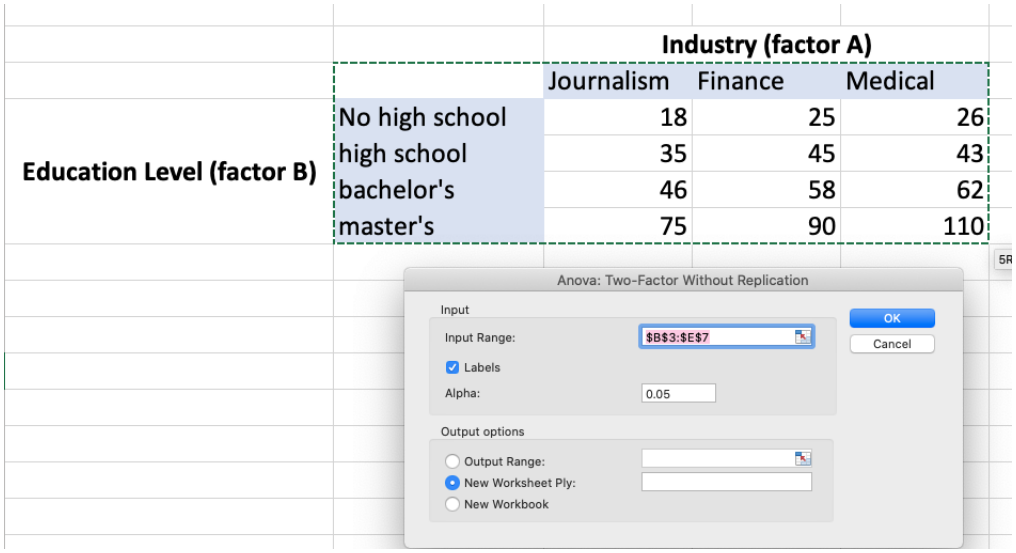


Figure 10.1: Your screen should look like this.

Step five, interpret and understand the Excel output.

Anova: Two-Factor Without Replication						
SUMMARY	*Count*	*Sum*	*Average*	*Variance*		
No high scho	3	69	23	19		
high school	3	123	41	28		
bachelor's	3	166	55.333333	69.333333		
master's	3	275	91.666667	308.33333		
Journalism	4	174	43.5	573.66667		
Finance	4	218	54.5	744.33333		
Medical	4	241	60.25	1316.25		
ANOVA						
ırce of Variati	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	7632.9167	3	2544.3056	56.575046	8.595E-05	4.7570627
Columns	579.5	2	289.75	6.442866	0.0320666	5.1432528
Error	269.83333	6	44.972222			
Total	8482.25	11				

Figure 10.2: Your Excel output should look like this.

To understand what each cell in this table represents, refer to the table below:

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	SSB	r-1	MSB	test statistic for factor B	p value for factor B	irrelevant
Columns	SSA	c-1	MSA	test statistic for factor A	p value for factor A	irrelevant
Error	SSE	nT-c-r+1	MSE			
Total	SST	nT-1				

Figure 10.3: Your Excel output can be understood like this.

Therefore, based on our table, 7632.9167 is the SSB value and 2544.3056 is the MSB value. The p value for each factor's hypothesis test is 0.00008595 and 0.0320666.

Step six, reject or do not reject the null hypothesis.

For Factor Group A, we reject the null because $p < \alpha$, $0.0320666 < .05$. Therefore, average annual salaries do differ by field of employment at the 5% significance level.

For Factor Group B, we reject the null because $p < \alpha$, $0.00008595 < .05$. At the 5% significance level, average annual incomes differ by education level.

It is much easier to do this test in Excel. However, if you are interested in seeing how the math works out for all of these calculations done by hand, check out my work below:

Step one, state the null and alternative hypotheses. Because we are comparing the means for factor A (employment field) and means for factor B (education level,) we'll have two sets of hypotheses.

1. $H_0 : \mu_{education} = \mu_{financial} = \mu_{medical}$
2. H_A not all population means are equal
3. $H_0 : \mu_{NoHighSchool} = \mu_{HighSchool} = \mu_{Bachelors} = \mu_{Masters}$
4. H_A not all population means are equal

Step two, choose the α value of .01, .05, or .10. The problem states that we'll use an α = .05 for both questions.

Step three, calculate the grand mean, $\bar{\bar{x}}$. The "grand mean" is the average of all observations across all treatment groups.

$$\bar{\bar{x}} = \frac{18 + 35 + 46 + 75 + 25 + 45 + 58 + 90 + 26 + 43 + 62 + 110}{12}$$
$$\bar{\bar{x}} = 52.75$$

Step four, calculate the sample mean for each group for both categories.

$$\bar{x}_{journalism} = \frac{18 + 35 + 46 + 75}{3} = 43.5$$
$$\bar{x}_{finance} = \frac{25 + 45 + 58 + 90}{3} = 54.5$$
$$\bar{x}_{medical} = \frac{26 + 43 + 62 + 110}{3} = 60.25$$
$$\bar{x}_{NoHS} = \frac{18 + 25 + 26}{3} = 23.00$$
$$\bar{x}_{HS} = \frac{35 + 45 + 43}{3} = 41.00$$
$$\bar{x}_{Bachelors} = \frac{46 + 58 + 62}{3} = 55.33$$
$$\bar{x}_{Masters} = \frac{75 + 90 + 110}{3} = 91.67$$

Step five, Calculate the Sum of Squares for Factor A (SSA) and Factor B (SSB). This measures how much "action" or variance, we have going on between each industry (factor A) and each education level (factor B).

First, I'll do it for factor A (industry).

$$SSA = r\sum_{i=1}^{c}(\bar{x}_i - \bar{\bar{x}})^2$$

$$SSA = 4[(43.5 - 52.75)^2 + (54.5 - 52.75)^2 + (60.25 - 52.75)^2]$$

$$SSA = 579.5$$

Now we do the same for factor B, education.

$$SSB = c\sum_{j=1}^{r}(\bar{x}_j - \bar{\bar{x}})^2$$

$$SSB = 3[(23 - 52.75)^2 + (41 - 52.75)^2 + (55.33 - 52.75)^2 + (91.67 - 52.75)^2]$$

$$SSB = 7632.92$$

Step six, calculate the Mean Square for Factor A (MSA) and Factor B (MSB).

First, we do it for Factor A. Remember, c is the number of columns.

$$MSA = \frac{SSA}{c-1}$$

$$MSA = \frac{579.5}{3-1}$$

$$MSA = 289.75$$

Now we do the same for Factor B. Remember, r is the number of rows.

$$MSB = \frac{SSB}{r-1}$$

$$MSB = \frac{7632.92}{4-1}$$

$$MSB = 2544.3056$$

Step seven and eight, calculate the Total Sum of Squares (SST) and the Error Sum of Squares (SSE). But to do that, we first need to compute the SST. Know that x_{ij} represents each of our observations, that is 18, 35, 46, and so on.

$$SST = \sum_{i=1}^{c}\sum_{j=1}^{r}(\bar{x_{ij}} - \bar{\bar{x}})^2$$

$$SST = (18-52.75)^2+(35-52.75)^2+(46-52.75)^2+(75-52.75)^2+(25-52.75)^2+(45-52.75)^2$$
$$+(58-52.75)^2+(90-52.75)^2+(26-52.75)^2+(43-52.75)^2+(62-52.75)^2+(110-52.75)^2$$
$$SST = 8482.25$$

Now we can compute SSE:

$$SSE = SST - (SSA + SSB)$$
$$SSE = 8482.25 - (579.5 + 7632.92)$$
$$SSE = 269.83$$

Step nine, compute the Mean Square Error (MSE). Remember, n_T is the number of observations. We have 12 here.

$$MSE = \frac{SSE}{n_T - c - r + 1}$$
$$MSE = \frac{269.83}{12 - 3 - 4 + 1}$$
$$MSE = 44.97$$

Step ten, compute the test statistics using the equations below.

$$F_{df_1,df_2} = \frac{MSA}{MSE}$$
$$\text{Where } df_1 = c - 1 \text{ and } df_2 = n_T - c - r + 1$$
$$F_{(3-1,12-3-4+1)} = \frac{289.75}{44.97}$$

$$F_{(2,6)} = \frac{289.75}{44.97}$$
$$F_{(2,6)} = 6.443$$

$$F_{df_1,df_2} = \frac{MSB}{MSE}$$

Where $df_1 = r - 1$ and $df_2 = n_T - c - r + 1$

$$F_{df_1,df_2} = \frac{2544.3056}{44.97}$$

$$F_{(3,6)} = 56.575$$

Step eleven, compute the p-value using Excel.

For Factor Group A: We enter $= F.DIST.RT(6.443, 2, 6)$ into Excel and get a p value of 0.032 in return.

For Factor Group B: We enter $= F.DIST.RT(56.575, 3, 6)$ into Excel and get a p value of 8.519E-05 in return, which translates to .00008.

Step twelve, reject or do not reject the null hypothesis.

For Factor Group A, we reject the null because p < α, .032 < .05. Therefore, average annual salaries do differ by field of employment at the 5% significance level.

For Factor Group B, we reject the null because p < α, .00008 < .05. At the 5% significance level, average annual incomes differ by education level.

10.2 Exam 2 Review in Class on Mar. 29

11 Week 11: Apr. 3 - Apr. 7

11.1 Exam 2 on April 3 via Canvas

11.2 Hypothesis Test for the Correlation Coefficient

The sample correlation coefficient gauges the direction and the strength of the linear relationship between two variables x and y. We calculate the sample correlation coefficient r_{xy} using the formula below:

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

Let me define each of these terms for you:

- S_{xy} is the sample covariance
- s_x is the the sample standard deviation for the x variable
- s_y is the the sample standard deviation for the y variable

The sample correlation coefficient r_{xy} is unit-free and its value falls between -1 and 1. If r_{xy} equals 1, then a perfect positive linear relationship exists between x and y. Similarly, a perfect negative linear relationship exists if r_{xy} equals -1. If r_{xy} equals zero, then no linear relationship exists between x and y. Other values for r_{xy} must be interpreted with reference to -1, 0, and 1. As the absolute value of r_{xy} approaches 1, the linear relationship grows stronger. For instance, r_{xy} = 0.80 indicates a strong negative linear relationship, whereas r_{xy} = 0.12 indicates a weak positive linear relationship.

We can conduct a hypothesis test to determine whether the sample correlation is representative of the population correlation. Let ρ_{xy} denote the population correlation coefficient.

The 5 Steps to a Hypothesis Test for a Correlation Coefficient:

1. State the null and alternative hypotheses. When testing whether the population correlation coefficient differs from zero, is greater than zero, or is less than zero, the competing hypotheses will take one of the following forms:

   - $H_0 : \rho_{xy} = 0$
   - $H_A : \rho_{xy} \neq 0$

   - $H_0 : \rho_{xy} \leq 0$
   - $H_A : \rho_{xy} > 0$

   - $H_0 : \rho_{xy} \geq 0$
   - $H_A : \rho_{xy} < 0$

2. Choose the α value of .01, .05, or .10. This is the allowed probability of making a Type I error.

3. Use Excel to calculate the sample correlation coefficient.

   In excel, enter: =CORREL(column of variable 1, column of variable 2).

   The "column" is going to be the column of data in excel that pertains to one of the two variables that we are computing the correlation coefficient for.

4. Calculate the test statistic.

$$t_{df} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-(r_{xy})^2}}$$

Where $df = n - 2$

5. Using the test statistic, calculate p value (using Excel). The Excel command differs depending on if it's a right tailed, left tailed or two-tailed test. Remember, a GREATER (>) than sign in the alternative hypothesis means we're doing a RIGHT-tailed test. A LESS (<) than sign in the alternative hypothesis means we're doing a LEFT-tailed test. A $\neq$ sign in the alternative means we are doing a TWO tailed test.

Operator in Alternative Hypothesis	Type of Test
> or $\geq$	Right-tailed Test
< or $\leq$	Left-tailed Test
$\neq$	Two-tailed Test

   For a right tailed test, enter $= T.DIST.RT(teststatistic, df)$

   For a left tailed test, enter $= 1 - T.DIST.RT(teststatistic, df)$

   For a two tailed test, enter $= 2 * T.DIST.RT(ABS(teststatistic), df)$ Note that you need to enter the absolute value of the test statistic in the two tailed test.

6. Reject or do not reject the null hypothesis.

   If the p value < α, reject the null.

   If the p value > α, do not reject the null.

Let's do an example. Use the Debt Payments data file on Canvas to solve the following problems. Calculate and interpret the correlation coefficient between Debt and Income. We want to test if the correlation coefficient differs from 0. At the 5% significance level, determine whether the correlation coefficient is significant.

Step one, state the null and alternative hypotheses.

1. $\rho_{xy} = 0$
2. $\rho_{xy} \neq 0$

Step two, choose the α value of .01, .05, or .10. The problem states that we will use an $\alpha = .05$.

Step three, use Excel to calculate the sample correlation coefficient. We enter $= CORREL(B2 : B27, C2 : C27)$ and get 0.8675 in return. This is the r_{xy} value.

Step four, calculate the test statistic using the formula below:

$$t_{df} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-(r_{xy})^2}}$$

$$\text{Where } df = n-2$$

$$t_{26-2} = \frac{0.8675\sqrt{26-2}}{\sqrt{1-0.8675^2}}$$

$$t_{24} = \frac{0.8675\sqrt{24}}{\sqrt{1-0.7526}}$$

$$t_{24} = \frac{0.8675 * 4.8990}{\sqrt{0.2474}}$$

$$t_{24} = \frac{4.2499}{0.4974}$$

$$t_{24} = 8.544$$

Step five, compute the p value. Because this is a two tailed test, we enter $= 2{*}T.DIST.RT(ABS(8.544), 24)$ into Excel and get a p value of 9.66E-09 in return, which translates to 0.00000000966.

Reject or do not reject the null hypothesis. We reject the null because p < α, .00000000966 < .05. At the 5% significance level, we conclude that the population correlation coefficient between Debt and Income differs from zero.

12 Week 12: Apr. 10 - Apr. 14

12.1 Linear Regression Model

Regression analysis is one of the most widely used statistical methodologies in business, engineering, and the social sciences. It evaluates how one variable, (the response variable), is influenced by other variables, (the explanatory variables.) A good example of this is trying to predict the selling price of a house (response variable) on the basis of its size (explanatory variable) and location (explanatory variable). Regression models are known to perform well for making predictions.

Correlation and regression analyses are related in a sense that they both measure some form of association between variables. The correlation coefficient measures the strength of the linear relationship between two variables, whereas regression extends the analysis to capture the relationship between the response variable and multiple explanatory variables.

Before we get into the details of linear regressions, we first need to understand **deterministic vs. stochastic relationships.**

A **deterministic** relationship is where NO randomness is involved. For example, momentum is a product of mass and velocity. We know that only mass and velocity affect momentum, there are no random factors that would affect momentum other than mass and velocity.

A **stochastic** relationship involves some randomness or some unknown element and therefore is not included as one of the explanatory variables. For example, if we are trying to predict how education affects income, our predictions may not be accurate because we are not including an explanatory variable that measures skill level. Skill level also affects income. This skill level is like a "randomness" that is not be measured or included as an explanatory variable.

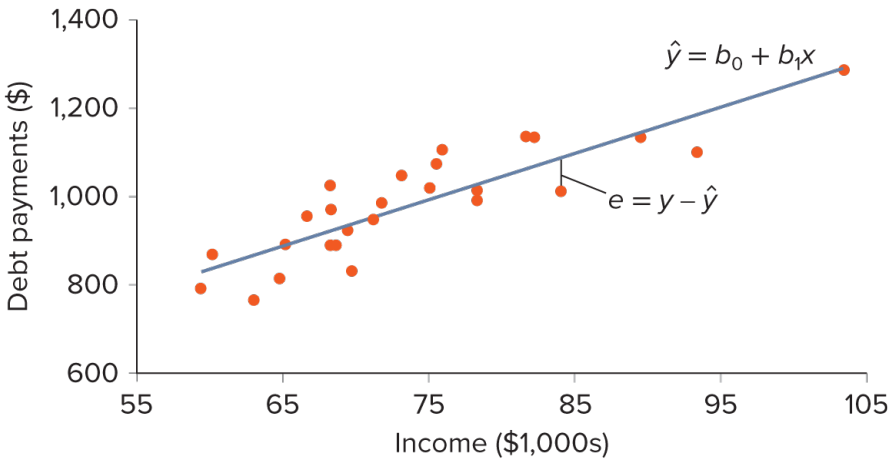
The **Simple Linear Regression Model** is where we have one response variable and one explanatory variable: $y = \beta_0 + \beta_1 x + \epsilon$. This is very similar to $y = mx + b$ where β_0 is the constant and β_1 is the coefficient on the explanatory variable. The ϵ represents the "randomness" in a stochastic model. If this is a deterministic model, $\epsilon = 0$. I'll evaluate an example to give you a better understanding.

If we're given this regression: annual income = 3000 + 2000*(years of education) + ϵ, we can predict someone's annual income. If they have 1 year of education: $5000 = 3000 + 2000 * (1)$. Or, if they have 2 years of education: $7000 = 3000 + 2000 * (2)$. Or, if they have 0 years of education: $3000 = 3000 + 2000 * (0)$. So, every additional year of education adds on to their annual income by 2000.

If $\beta_1 > 0$, there is a positive relationship between the explanatory and response variables. If $\beta_1 < 0$, there is a negative relationship between the explanatory and response variables.

The population parameters β_0 and β_1 used in the simple linear regression model are unknown, and therefore, must be estimated. As always, we use sample data to estimate the population parameters of interest.

One type of linear regression is Ordinary Least Squares (OLS). OLS can be implemented to estimate β_0 and β_1. The big hype of OLS is that it aims to reduce the distance between observed and predicted values. The distance between observed and predicted values is denoted as $e = y - \hat{y}$ where y is the observed value and $\hat{y}$ is the predicted value. In the graph below, the trend line represents our predicted values of debt payments for each income value. The orange dots are the original observed values, or the "real values." Based on these sample observations, we calculated a b_0 and b_1 using OLS to create a formula that predicts debt payments based on income. The reason why b_0 and b_1 are lower-cased is because we are working with sample data, not population data.



The formal equation for calculating b_0 and b_1 is below, but in this class, we'll be using Excel to do all the work for us.

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

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

Let's do an example. Use the Debt Payments data file on canvas to predict Debt based on Income in a simple linear regression model. Interpret the coefficients and predict debt payments if income is $80,000.

Excel Steps for a Linear Regression:

1. First, make sure your Excel is prepared to analyze data. Click on the "Data" tab on the top, then click on "Analysis Tools." If you have a Mac, you need to check the "Analysis ToolPak" box. If you have a PC, you need to "Browse" for the "Analysis ToolPak." Once you have this installed, you're ready to go.

2. Open the Debt Payments data file on Canvas. (Files>Data>ClassData.xlsx>Debt)

3. Choose Data > Data Analysis > Regression from the menu.

4. In the Regression dialog box, click on the box next to Input Y Range, and then select the Debt observations, including its heading. For Input X Range, select the Income Unemployment observations, including the heading. Check Labels. Click OK.

Your results should look just like the table below.

Regression Statistics									
Multiple R	0.86751154								
R Square	0.75257627								
Adjusted R Sq	0.74226695								
Standard Erro	63.2605567								
Observations	26								
ANOVA									
	df	*SS*	*MS*	*F*	*Significance F*				
Regression	1	292136.909	292136.909	72.9995882	9.6603E-09				
Residual	24	96045.5529	4001.89804						
Total	25	388182.462							
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*	
Intercept	210.297683	91.3387356	2.30239319	0.03029326	21.7837983	398.811568	21.7837983	398.811568	
Income	10.4411054	1.2220424	8.54397965	9.6603E-09	7.91893385	12.963277	7.91893385	12.963277	

So, our simple linear regression equation looks like this: $Debt = 210.2977 + 10.4411 * income$. The coefficient on income is 10.4411, which means that for every one unit($1000) increase in someone's income (income is measured in $1,000s) then we predict consumer debt payments to increase by $b1$—that is, by $10.4411. The intercept of 210.2977 suggests that if income equals zero, then predicted debt payments are $210.2977.

To predict debt payments when income is $80,000, we need to plug all of the values into the equation: $Debt = 210.29 + 10.44 * (80)$. We use 80 here because income is measured in thousands. Based on these numbers, we can predict that debt payments will be $1045.59 at an income of $80,000.

The **Multiple Linear Regression Model** is where we have one response variable and multiple explanatory variables: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$. It is just like the simple linear regression model, but we have more than one explanatory variable. There is no limit to the number of explanatory variables. In this case, interpreting the effect of β_1 on y is read as "for every one unit increase in x_1, there is an associated β_1 increase in the predicted value of y, while holding all other variables constant.

12.2 Goodness-of-Fit Measures

So after we've created a linear regression model, how do we measure how "good" it is? In other words, if all of our predicted values are very far from the actual real observed values, it's a pretty bad model. We can quantify how good a model is with these three techniques:

1. Standard Error of the Estimate

   This is a measure of how many times we get a prediction "wrong." The bigger this standard error of the estimate, the more predictions the model has gotten wrong.

2. Coefficient of Determination, R^2

This is a measure of how much of the variation in the response variable can be explained by the regression equation. Let's say we're working with a model that evaluates the effect of education on income. If the model has an R^2 of 0.72, this means that 72% of the variation in income can be explained by education.

3. Adjusted R^2

   This is also a measure of how much of the variation in the response variable can be explained by the regression equation. However, it is adjusted based on the number of explanatory variables and sample size.

I'll dive into detail into these three techniques in this section.

Standard Error of the Estimate

The formula to compute this is:

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n - k - 1}}$$

where:

1. y is each observed value
2. $\hat{y}$ is each predicted value
3. n is the number of samples
4. k is the number of explanatory variables

In this class, we will not compute this standard error manually. We will use Excel to produce a linear regression and identify the standard error of the estimate. Using the regression results from the debt payments example, the standard error of the estimate is reported in the Excel output under "Regression Statistics" as 63.2606. It is the cell highlighted in yellow.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8675115
R Square	0.7525763
Adjusted R Square	0.7422669
Standard Error	63.260557
Observations	26

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	292136.91	292136.91	72.999588	9.66E-09
Residual	24	96045.553	4001.898		
Total	25	388182.46			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	210.29768	91.338736	2.3023932	0.0302933	21.783798	398.81157	21.783798	398.81157
Income	10.441105	1.2220424	8.5439796	9.66E-09	7.9189338	12.963277	7.9189338	12.963277

Coefficient of Determination, R^2

The formula to compute this is:

$$R^2 = 1 - \frac{SSE}{SST}$$

$$SSE = \sum (y_i - \hat{y}_i)^2$$

$$SST = \sum (y_i - \bar{y})^2$$

SSE, the sum of squares estimate, measures the unexplained variation in the response variable. SST, the total sum of squares, measures the total variation in the response variable.

In this class, we will not compute this R^2 manually. We will use Excel to produce a linear regression and identify the R^2. The R^2 is reported in our Excel output under "Regression Statistics" as 0.7526. It is the cell highlighted in orange.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8675115
R Square	0.7525763
Adjusted R Square	0.7422669
Standard Error	63.260557
Observations	26

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	292136.91	292136.91	72.999588	9.66E-09
Residual	24	96045.553	4001.898		
Total	25	388182.46			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	210.29768	91.338736	2.3023932	0.0302933	21.783798	398.81157	21.783798	398.81157
Income	10.441105	1.2220424	8.5439796	9.66E-09	7.9189338	12.963277	7.9189338	12.963277

Adjusted R^2

Because R^2 never decreases as we add more explanatory variables to the linear regression model, it is possible to increase its value unintentionally by including a group of explanatory variables that may have no economic or intuitive foundation in the linear regression model. This is true especially when the number of explanatory variables (k), is large relative to the sample size (n). In order to avoid the possibility of R^2 creating a false impression, virtually all software packages, including Excel and R, include adjusted R^2. Unlike R^2, adjusted R^2 explicitly accounts for the number of explanatory variables (k) and the sample size (n). It is common to use adjusted R^2 for model selection because it imposes a penalty for any additional explanatory variable that is included in the analysis.

The formula to compute this is:

$$Adj.R^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

where:

1. n is the number of samples
2. k is the number of explanatory variables
3. R^2 is the unadjusted R^2 value

In this class, we will not compute this adjusted R^2 manually. We will use Excel to produce a linear regression and identify the adjusted R^2. The adjusted R^2 is reported in our Excel output under "Regression Statistics" as 0.7423. It is the cell highlighted in green.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8675115
R Square	0.7525763
Adjusted R Square	0.7422669
Standard Error	63.260557
Observations	26

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	292136.91	292136.91	72.999588	9.66E-09
Residual	24	96045.553	4001.898		
Total	25	388182.46			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	210.29768	91.338736	2.3023932	0.0302933	21.783798	398.81157	21.783798	398.81157
Income	10.441105	1.2220424	8.5439796	9.66E-09	7.9189338	12.963277	7.9189338	12.963277

13 Week 13: Apr. 17 - Apr. 21

13.1 Regression with Dummy Variables

The explanatory and response variables we have used have mostly been numerical. For example, in Chapter 14, we used income and unemployment to explain variations in consumer debt. However, it is common to include some variables that are categorical. Most categorical variables are binary. They are either a yes or no answer. For example, home-ownership status is a binary variable. The answer is either yes or no.

Let's say we are trying to model professor salary data. (Files>Data>ClassData.xlsx>Professor). After estimating the model in Excel, we get: $\hat{y} = 48.83 + 1.15x$ where y represents salary (in \$1,000s) and x represents experience (in years). The sample regression equation implies that the predicted salary increases by about \$1,150 (1.15×1000) for every year of experience. Arguably, in addition to experience, variations in salary are also associated with a person's sex (male or female) and age (less than 60 years or at least 60 years).

A categorical variable requires special attention in regression analysis because, unlike a numerical variable, the observations of a categorical variable cannot be used in their original form—that is, in a non-numerical format. We need to convert a categorical variable into a dummy variable, which is also called an indicator variable. A dummy variable, d, is defined as a variable that assumes a value of 1 for one of the categories and 0 for the other. For example, when categorizing a person's sex, we can define d as 1 for male and 0 for female. **A dummy variable d is defined as a variable that takes on values of 1 or 0. It is commonly used to describe a categorical variable with two categories.**

Let's interpret the model output below to understand how dummy variables work. Assume that $d_1 = 1$ if the professor is male and $d_2 = 1$ if the professor is at least 60 years old. What is the predicted salary of a 50-year-old male professor with 10 years of experience?

	Coefficients	*Standard Error*	*t Stat*	*p-Value*
Intercept	40.6060	3.6919	10.999	0.000
Experience (x)	1.1279	0.1790	6.300	0.000
Male (d_1)	13.9240	2.8667	4.857	0.000
Older (d_2)	4.3428	4.6436	0.935	0.356

In this case, $d_1 = 1$ because the professor is a male and $d_2 = 0$ because he is not 60 years old. Also, $x = 10$ because he has 10 years of experience. We plug these values into the model to get:

$$\hat{y} = 40.6060 + 1.1279(10) + 13.9240(1) + 4.3428(0)$$

$$\hat{y} = 40.6060 + 11.279 + 13.9240$$

$$\hat{y} = 65.809$$

$$\hat{y} = \$65,809$$

We get a predicted salary of $65,809.

We can use Excel to make dummy variables. Using the Professor data file on canvas, we use the IF function to create a dummy variable for the Sex and Age variables. The formula is: =IF(C2="Male", 1, 0), assuming that cell C2 is the cell we are trying to convert to a dummy variable.

It's possible for categorical variables to have multiple categories. For example, the mode of transportation used to commute to work may be described by three categories: Public Transportation, Driving Alone, and Car Pooling. However, we need to avoid the dummy variable trap. This trap is when we have too many dummy variables. Assuming that the linear regression model includes an intercept, the number of dummy variables representing a categorical variable should be one less than the number of categories of the variable. So if we have three dummy categories of transportation, we can only include 2 in our model. The reason for this is that the dummy variable trap produces an issue called multicollinearity, something that we'll talk about next week.

13.2 Tests of Significance Part. 1

After we construct a linear regression model and get β coefficients, we can conduct hypothesis tests to see if these coefficients are statistically significant. It's important to do these tests because the coefficients are only meaningful if they are statistically significant. Remember, statistical significance is important because that means that the patterns we find in the sample data are not due to random chance. Therefore, this indicates that the patterns we find in the sample will also be present in the population data.

There are three tests we can run:

1. Test of Joint Significance
2. Test of Individual Significance
3. Test for a Nonzero Slope Coefficient

We'll cover the test of joint significance of test of individual significance in this part 1. Then we'll cover the test for a nonzero slope coefficient in part 2, next class.

First, we'll talk about the Test of Joint Significance. This tests if all of the explanatory variables included in a regression are collectively significant in predicting the outcome variable. The null hypothesis is that all of the betas of each coefficient are all 0. The alternative is that at least one $\beta \neq 0$. We can use the P-value approach in the form of a right-tailed F test for this.

The Five Steps of the Test of Joint Significance

1. State the null and alternative hypotheses. The null hypothesis is that all of the $\beta = 0$. The alternative is that at least one $\beta \neq 0$.
2. Choose an α value of .01, .05 or .10. This is the allowed probability of making a Type I error.
3. Compute the test statistic with the following formula:

$$F_{(df_1, df_2)} = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE}$$

4. Compute the p value using the test statistic in Excel. This test is always a right tailed test, therefore the excel command is $= F.DIST.RT(teststatistic, df_1, df_2)$ where $df_1 = k$ and $df_2 = n - k - 1$.
5. Reject or do not reject the null hypothesis.

   If the p value < α, reject the null.

   If the p value > α, do not reject the null.

Let's do an example. Using the data on Canvas, (Files>Canvas>ClassData.xlsx>Baseball), we want to predict a team's winning percentage (Win) on the basis of its batting average (BA) and its earned run average (ERA). BA is a ratio of hits divided by times at bat. ERA is the average number of earned runs given up by a pitcher per nine innings pitched. We expect that a higher BA positively influences a team's winning percentage, while a higher ERA negatively affects a team's winning percentage. We estimate the model: $Win = \beta_0 + \beta_1 BA + \beta_2 ERA + \epsilon$. Conduct a test to determine if BA and ERA are jointly significant in explaining winning percentage at $\alpha = 0.05$. We have 30 observations in our sample. These are the results we get in Excel after running a regression, with wins as the response variable, and BA and ERA as the explanatory variables.

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	0.09578	0.04789	33.966	4.25E-08
Residual	27	0.038068	0.00141		
Total	29	0.133848			

Step one, state the null and alternative hypotheses.

1. $H_0 : \beta_1 = \beta_2 = 0$
2. H_A : at least one $\beta \neq 0$

Step two, choose an α value of .01, .05 or .10. We are using $\alpha = .05$.

Step three, compute the test statistic with the following formula:

$$F_{(df_1, df_2)} = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE}$$

Where:

- n is the sample size
- k is the number of explanatory variables
- SSR and SSE are all going to come from the Excel output

Don't freak out. All of these values are going to come from the regression results in Excel. Therefore, we need to know how to interpret the ANOVA table results. This is the structure of the one-way ANOVA table in Excel that comes with the regression results.

ANOVA	*df*	*SS*	*MS*	*F*	Significance *F*
Regression	k	SSR	$MSR = \frac{SSR}{k}$	$F_{(df_1, df_2)} = \frac{MSR}{MSE}$	$P\left(F_{(df_1, df_2)} \geq \frac{MSR}{MSE}\right)$
Residual	$n-k-1$	SSE	$MSE = \frac{SSE}{n-k-1}$		
Total	$n-1$	SST			

We can compare this to our Excel results.

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	0.09578	0.04789	33.966	4.25E-08
Residual	27	0.038068	0.00141		
Total	29	0.133848			

After looking at this table, we can define all of the values that we need to calculate this test statistic.

- $n - 1 = 29$ therefore $n = 30$
- $k = 2$ (batting average and earned runs average)
- $SSR = 0.0958$
- $SSE = 0.0381$

Now, we can plug all of these into our equation:

$$F_{2,(30-2-1)} = \frac{0.0958/2}{0.0381/(30-2-1)}$$

$$F_{2,27} = \frac{0.0479}{0.0381/27}$$

$$F_{2,27} = \frac{0.0479}{0.0014}$$

$$F_{2,27} = 34.2143$$

Step four, compute the p value with the test statistic in Excel. We enter $= F.DIST.RT(34.2143, 2, 27)$ into Excel and get $p = 3.9605E - 08$ in return, which translates to 0.000000039605.

Step five, reject or do not reject the null hypothesis. We reject the null because p < α, 0.000000039605 < 0.05.

Now we move on to the P Value approach for the Test of Individual Significance. This tests if only one of the explanatory variables (we choose which one we want to evaluate) influences the outcome variable. The null hypothesis is that some $\beta_1 = 0$ (whichever one we are evaluating). The alternative is that $\beta_1 \neq 0$. We can use the P-value approach in the form of a t-test OR the confidence interval approach for this test. We will focus on the P Value Approach.

Five Steps of Test of Individual Significance: P Value Approach

1. State the null and alternative hypotheses. The null hypothesis is that some $\beta_1 = 0$ (whichever one we are evaluating). The alternative is that $\beta_1 \neq 0$.

2. Choose an α value of 0.01, 0.05 or 0.10. This is the allowed probability of making a Type I error.

3. Compute the test statistic with the following formula:

$$t_{df} = \frac{b_j - \beta_{j0}}{se(b_j)} \tag{1}$$

   Let's break down each of these terms:

   - b_j is the estimate of the beta we are evaluating, this is from our model output
   - β_{j0} is the hypothesized estimate of beta in our model (this is from our hypotheses in step 1).
   - $se(b_j)$ is the standard error of the beta we are evaluating, this is from our model output
   - df is the degrees of freedom where $df = n - k - 1$

4. Compute the p value with the test statistic in Excel. These tests are always two-tailed and therefore the command is $= 2{*}T.DIST.RT(ABS(teststatistic), df)$ where $df = n-k-1$, n being the sample size, and k being the number of explanatory variables in the regression.

5. Reject or do not reject the null hypothesis.

   If the p value < α, reject the null.

   If the p value > α, do not reject the null.

Let's do an example. We want to predict a team's winning percentage (Win) on the basis of its batting average (BA) and its earned run average (ERA). BA is a ratio of hits divided by times at bat. ERA is the average number of earned runs given up by a pitcher per nine innings pitched. We expect that a higher BA positively influences a team's winning percentage, while a higher ERA negatively affects a team's winning percentage. We estimate the model: $Win = \beta_0 + \beta_1 BA + \beta_2 ERA + \epsilon$. Conduct a test to determine if BA is independently significant in explaining winning percentage at $\alpha = 0.05$. We have 30 observations in our sample. These are the model output that we get in Excel:

	Coefficients	Standard Error	*t* Stat	*p*-value	Lower 95%	Upper 95%
Intercept	0.1269	0.1822	0.696	0.492	−0.2470	0.5008
BA	3.2754	0.6723	4.872	4.3E-05	1.8960	4.6549
ERA	−0.1153	0.0167	−6.920	1.95E-07	−0.1494	−0.0811

Step one, state the null and alternative hypotheses.

1. $H_0 : \beta_1 = 0$
2. $H_A : \beta_1 \neq 0$

Step two, choose an α value of .01, .05 or .10. We are using $\alpha = .05$.

Step three, compute the test statistic with the following formula:

$$t_{df} = \frac{b_j - \beta_{j0}}{se(b_j)} \tag{2}$$

Let's break down each of these terms:

- b_j is the estimate of the beta we are evaluating, this is from our model output
- β_{j0} is the hypothesized estimate of beta in our model (this is from our hypotheses in step 1).
- $se(b_j)$ is the standard error of the beta we are evaluating, this is from our model output
- df is the degrees of freedom where $df = n - k - 1$

Now we'll define them based on our model output:

- $b_j = 3.2754$
- $\beta_{j0} = 0$
- $se(b_j) = 0.6723$
- $df = 30 - 2 - 1 = 27$

Now we can plug them into our equation:

$$t_{30-2-1} = \frac{3.2754 - 0}{0.6723}$$
$$t_{27} = \frac{3.2754}{0.6723}$$
$$t_{27} = 4.872$$

We can see this matches with the t statistic given to us in the model output for the BA variable.

Step four, compute the p value with the test statistic in Excel. We enter $= 2{*}T.DIST.RT(ABS(4.872), 27)$ into Excel and get $p = 4.29E - 08$ in return, which translates to 0.00000000429.

Step five, reject or do not reject the null hypothesis. We reject the null because p < α because .00000000429 < .05. Therefore, if we've rejected the null, that means we can confidently say that it is impossible that $\beta_1 = 0$ and this would hold true if we collected a greater sample and more observations.

We can also do a test of individual significance using the Confidence Interval approach. However, we are not going to focus on this method in this class. I have the notes below if you'd like to read for yourself.

Four Steps of Test of Individual Significance: Confidence Interval Approach

1. State the null and alternative hypotheses. The null hypothesis is that some $\beta_1 = 0$ (whichever one we are evaluating). The alternative is that $\beta_1 \neq 0$.

2. Choose an α value of .01, .05 or .10. This is the allowed probability of making a Type I error.

3. Compute the confidence interval at 1-α using the following equation, remember that $df = n - k - 1$.

$$b_j \pm t_{\alpha/2,df} * se(b_j)$$

4. Reject or do not reject the null hypothesis.

   If the confidence interval contains the hypothesized value (0), do not reject the null.

   If the confidence interval does not contain the hypothesized value (0), reject the null.

Let's do an example. We want to predict a team's winning percentage (Win) on the basis of its batting average (BA) and its earned run average (ERA). BA is a ratio of hits divided by times at bat. ERA is the average number of earned runs given up by a pitcher per nine innings pitched. We expect that a higher BA positively influences a team's winning percentage, while a higher ERA negatively affects a team's winning percentage. We estimate the model: $Win = \beta_0 + \beta_1 BA + \beta_2 ERA + \epsilon$. Conduct a confidence interval test to determine if BA is independently significant in explaining winning percentage at $\alpha = 0.05$. We have 30 observations in our sample. These are the model output that we get in Excel:

	Coefficients	Standard Error	*t* Stat	*p*-value	Lower 95%	Upper 95%
Intercept	0.1269	0.1822	0.696	0.492	−0.2470	0.5008
BA	3.2754	0.6723	4.872	4.3E-05	1.8960	4.6549
ERA	−0.1153	0.0167	−6.920	1.95E-07	−0.1494	−0.0811

Step one, state the null and alternative hypotheses.

1. $H_0 : \beta_1 = 0$
2. $H_A : \beta_1 \neq 0$

Step two, choose an α value of .01, .05 or .10. We are using $\alpha = .05$.

Step three, compute the confidence interval with the following formula:

$$b_j \pm t_{\alpha/2,df} * se(b_j)$$

The b_j in our case is the coefficient for BA, because we are just interested in the significance of this variable.

Let's define all of these terms:

- $b_j = 3.2754$
- $se(b_j) = 0.6723$
- $df = 30 - 2 - 1 = 27$
- $\alpha/2 = .05/2 = .025$

To find $t_{.025,27}$, we'll use this t value calculator to do this. The calculator finds that $t_{.025,27} = 2.052$ for a two tailed test. When using this calculator, you can enter the significance value as 0.05, the calculator will automatically divide it by 2 for you.

If you would like to be traditional, you can use this t table distribution and manually interpret the table. But the focus of this class is not interpreting these tables, so I am fine with you using online t value calculators. On the test, I will give you the t value in a case like this.

$$t_{30-2-1} = 3.2754 \pm t_{.025,30-2-1} * .6723$$

$$t_{27} = 3.2754 \pm t_{.025,27} * .6723$$

$$t_{27} = 3.2754 \pm 2.052 * .6723$$

$$t_{27} = 3.2754 \pm 1.3796$$

$$t_{27} = 3.2754 + 1.3796 = 4.655$$

$$t_{27} = 3.2754 - 1.3796 = 1.896$$

$$t_{27} = [1.896, 4.655]$$

Step four, reject or do not reject the null hypothesis. The confidence interval does not contain the hypothesized value of 0, therefore we reject the null hypothesis.

14 Week 14: Apr. 24 - Apr. 28

14.1 Tests of Significance Part. 2

Now, we'll move on to a Test for a Nonzero Slope Coefficient. This tests if one of the explanatory variables has a coefficient not equal to 0. The null hypothesis is that some coefficient (whichever one we are evaluating), equals some value: $\beta_1 = 2$. The hypothesized value of $beta_1$ can be any value, just not 0. The alternative is that $\beta_1 \neq 2$. We can use the P-value approach in the form of a t test for this.

Five Steps of Test for a Nonzero Slope Coefficient: P Value Approach

1. State the null and alternative hypotheses. The null hypothesis is that some $\beta_1 > somevalue$ (whichever one we are evaluating). The alternative is that $\beta_1 \leq somevalue$.

2. Choose an α value of .01, .05 or .10. This is the allowed probability of making a Type I error.

3. Compute the test statistic with the following formula:

$$t_{df} = \frac{b_j - \beta_{j0}}{se(b_j)}$$

   Let's break down each of these terms:

   - t_{df} where $df = n-k-1$, n being the sample size, and k being the number of explanatory variables in the regression.
   - b_j is the estimate of the beta in question from our model output
   - β_{j0} is the hypothesized estimate of beta in our model
   - $se(b_j)$ is the standard error of the beta in question from our model output

4. Compute the p value using the test statistic in Excel.

   If the > operator is used in the alternative, it is a right tailed test. The Excel command is $= T.DIST.RT(teststatistic, df)$ where $df = n - k - 1$, n being the sample size, and k being the number of explanatory variables in the regression.

   If the < operator is used in the alternative, it is a left tailed test. The Excel command is $= 1 - T.DIST.RT(teststatistic, df)$ where $df = n - k - 1$, n being the sample size, and k being the number of explanatory variables in the regression.

   If the $\neq$ operator is used in the alternative, it is a two tailed test. The Excel command is $= 2 * T.DIST.RT(ABS(teststatistic), df)$ where $df = n - k - 1$, n being the sample size, and k being the number of explanatory variables in the regression.

5. Reject or do not reject the null hypothesis.

   If the p value < α, reject the null.

   If the p value > α, do not reject the null.

Let's do an example. We want to predict a team's winning percentage (Win) on the basis of its batting average (BA) and its earned run average (ERA). BA is a ratio of hits divided by times at bat. ERA is the average number of earned runs given up by a pitcher per nine innings pitched. We expect that a higher BA positively influences a team's winning percentage, while a higher ERA negatively affects a team's winning percentage. We estimate the model: $Win = \beta_0 + \beta_1 BA + \beta_2 ERA + \epsilon$. Conduct a test to determine if the coefficient estimate of BA is greater than 1 at the 5% significance level. We have 30 observations in our sample. This is the model output that we get in Excel:

	Coefficients	Standard Error	*t* Stat	*p*-value	Lower 95%	Upper 95%
Intercept	0.1269	0.1822	0.696	0.492	−0.2470	0.5008
BA	3.2754	0.6723	4.872	4.3E-05	1.8960	4.6549
ERA	−0.1153	0.0167	−6.920	1.95E-07	−0.1494	−0.0811

Step one, state the null and alternative hypotheses.

1. $H_0 : \beta_1 < 1$
2. $H_A : \beta_1 \geq 1$

Step two, choose an α value of .01, .05 or .10. We are using $\alpha = .05$.

Step three, compute the test statistic with the following formula:

$$t_{df} = \frac{b_j - \beta_{j0}}{se(b_j)}$$

Now we'll define them based on our model output:

- $df = 30 - 2 - 1 = 27$
- $b_j = 3.2754$
- $\beta_{j0} = 1$
- $se(b_j) = 0.6723$

Now we can plug them into our equation:

$$t_{30-2-1} = \frac{3.2754 - 1}{0.6723}$$
$$t_{27} = \frac{2.2754}{0.6723}$$
$$t_{27} = 3.3845$$

Step four, compute the p value with the test statistic in Excel. The greater than (>) sign is used in the alternative, so we enter the command $= T.DIST.RT(3.3845, 27)$ and get a p value of 0.00110 in return.

Step five, reject or do not reject the null. Because the p value is less than the α value of .05 (0.00110 < .05), we reject the null. Therefore, by rejecting the null, we can conclude that there is no way in HELL that $\beta_1 < 1$.

14.2 General Test of Linear Restrictions

In this unit, we'll learn how to test if only a few of all the explanatory variables influence the response variable. This is called a General Test of Linear Restrictions. It also commonly referred to as a "Partial F-test."

Seven Steps for a General Test of Linear Restrictions

1. State the null and alternative hypotheses. The null hypothesis is that some of the $\beta = 0$. The alternative is that at least one $\beta \neq 0$.

   The null hypothesis in this test evaluates "some of the betas." This is because we are only evaluating a few of all the explanatory variables. We can choose which variables to evaluate.

2. Choose an α value of .01, .05 or .10. This is the allowed probability of making a Type I error.

3. Run a regression for the Restricted Model.

   We do this by running a regression in Excel and leaving out the variables that we are testing in the hypothesis. For example, if we are testing β_2 and β_3, then we would run a regression without β_2 and β_3 explanatory variables.

4. Run a regression for the Unrestricted Model.

   The unrestricted model includes all of the explanatory variables in the regression.

5. Compute the test statistic with the following formula:

$$F_{df1,df2} = \frac{(SSE_R - SSE_U)/df_1}{SSE_U/df_2}$$

Let's define each of these terms:

- df_1 = number of variables we are testing (this is the number of variables we leave out in the restricted model)
- $df_2 = n - k - 1$, where n is number of observations and k is total number of variables in the unrestricted model
- SSE_R is the SSE for the restricted model, this shows up in our excel output under the "SS" column in the "Anova" table, for the "residual" row
- SSE_U is the SSE for the unrestricted model, this shows up in our excel output under the "SS" column in the "Anova" table, for the "residual" row

6. Compute the p value with the test statistic in Excel. This kind of test will always be a right tailed test. The command is $= F.DIST.RT(teststatistic, df_1, df_2)$ where df_1 is the number of variables we are testing restrictions on and $df_2 = n - k - 1$, where n is the number of observations and k is the total number of explanatory variables in the unrestricted model.

7. Reject or do not reject the null hypothesis.

   If the p value < α, reject the null.

   If the p value > α, do not reject the null.

Let's do an example. Using the data on Canvas (Files>Data>ClassData.xlsx>Car Wash), we are evaluating a car wash company. A manager at a car wash company in Missouri wants to predict sales based on: 1) price discounts 2) radio advertising expenses and 3) newspaper advertising expenses in 40 counties in Missouri. At the 5% level, determine if the advertisement expenditures on radio and newspapers have a statistically significant influence on sales. This is what the unrestricted model equation looks like: $Sales = \beta_0 + \beta_1 Discount + \beta_2 Radio + \beta_3 Newspaper + \epsilon$.

Step one, state the null and alternative hypotheses.

- $H_0 : \beta_2 = \beta_3 = 0$
- H_A : at least one of the betas $\neq 0$

Step two, choose an α value of .01, .05 or .10. We are using an $\alpha = .05$.

Step three, run a regression for the Restricted Model. The model specification for the restricted model leaves out radio & newspaper variables, so we only include the discount variable: $Sales = \beta_0 + \beta_1 Discount + \epsilon$. Using the "car wash" excel file on canvas, the regression output for our restricted model is:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.59184871							
R Square	0.3502849							
Adjusted R S	0.33318713							
Standard Err	7.57864892							
Observations	40							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	1176.69966	1176.69966	20.4871738	5.76E-05			
Residual	38	2182.56494	57.4359194					
Total	39	3359.2646						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	43.4540802	2.73802386	15.8705995	2.3547E-18	37.9112406	48.9969197	37.9112406	48.9969197
Discount	0.40155567	0.08871657	4.52627593	5.76E-05	0.22195837	0.58115297	0.22195837	0.58115297

Step four, run a regression for the Unrestricted Model. The model specification for the unrestricted model includes all of the variables: $Sales = \beta_0 + \beta_1 Discount + \beta_2 Radio + \beta_3 Newspaper + \epsilon$. The regression output for our restricted model is:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.8002233							
R Square	0.6403574							
Adjusted R Sc	0.6103871							
Standard Err	5.793039							
Observations	40							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	3	2151.1298	717.04326	21.366454	4.019E-08			
Residual	36	1208.1348	33.559301					
Total	39	3359.2646						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	6.7024706	7.1661191	0.9352999	0.3558657	-7.8310925	21.236034	-7.8310925	21.236034
Discount	0.3416662	0.0689231	4.9572063	1.712E-05	0.2018836	0.4814488	0.2018836	0.4814488
Radio	6.0623887	1.6260015	3.7284029	0.0006605	2.7647048	9.3600727	2.7647048	9.3600727
Newspaper	9.3967948	2.2018284	4.2677235	0.000137	4.9312798	13.86231	4.9312798	13.86231

Step five, compute the test statistic with the following formula:

$$F_{df1,df2} = \frac{(SSE_R - SSE_U)/df_1}{SSE_U/df_2}$$

Let's define each of these terms:

- df_1 = number of variables we are testing (this is the number of variables we leave out in the restricted model)
- $df_2 = n - k - 1$, where n is number of observations and k is total number of variables in the unrestricted model
- SSE_R is the SSE for the restricted model, this shows up in our excel output under the "SS" column in the "Anova" table, for the "residual" row
- SSE_U is the SSE for the unrestricted model, this shows up in our excel output under the "SS" column in the "Anova" table, for the "residual" row

Now I'll define these terms in our problem:

- df_1 = 2
- $df_2 = n - k - 1 = 40 - 3 - 1$,
- $SSE_R = 2182.5649$
- $SSE_U = 1208.1348$

We can plug all of these values into our equation:

$$F_{2,40-3-1} = \frac{(2,182.5649 - 1,208.1348)/2}{1,208.1348/(40 - 3 - 1)}$$

$$F_{2,36} = \frac{(974.4301)/2}{1,208.1348/36}$$

$$F_{2,36} = \frac{487.21505}{33.5593}$$

$$F_{2,36} = 14.518$$

Step six, compute the p value with the test statistic in Excel. We enter the command $= F.DIST.RT(14.518, 2, 36)$ and get 2.38054E-05 in return.

Step seven, reject or do not reject the null hypothesis. We reject the null because 0.00002380 < .05. Therefore, at the 5% level, we conclude that the advertisement expenditures on radio and newspapers have a significant influence on sales.

15 Week 15: May 1 - May 5

15.1 Model Assumptions and Common Violations

Linear regressions are wonderful applications for the purpose of addressing public policy questions. However, they are only effective when applied correctly. There are several standards that need to be present or true in order for the regression's estimates to be accurate. If one or more of these standards are not present, our linear regression results are unreliable. The standards act as the foundation to the linear regression (the house). If the foundation is not reliable, the house collapses. Although the house can still exist, it may not be reliable. These standards are known as "assumptions" in formal statistics language. It is because we *assume* that these standards are true and present for our model when presenting linear regression results to others. However, in this class, I will refer to them as standards for simplicity.

Standards of Ordinary Least Squares Regression

1. ***Linear Relationship Between X and Y Variables***

   What does it mean: The relationship between each independent variable and the dependent variable is linear. Therefore, as the explanatory variable changes, the outcome variable consequently changes by a linear factor. This does not allow for exponential, logistical or polynomial relationships, ONLY linear.

   How to check that it's present: Scatter plots for each independent variable; explanatory variable on the x axis and outcome variable on the y axis. By doing that, you can visually identify a linear (or nonlinear) relationship. If the data points are roughly following a straight line, we have a linear regression. However, if you see funky curves, parabolas, exponentials or basically anything other than a straight line, you do not have a linear relationship.

   Bad things that happen if it's not present: The linear regression's beta coefficients will not be applicable to real life data. This will cause a world of problems if you are using your model to predict future values.

   What to do if it's not present: The best solution to fix nonlinearity is to choose better explanatory variables next time :)

2. ***Independence of Observations***

   What does it mean: Observations are not related to each other in any way. Observations need to be authentically unique and independent of each other. An example of nonindependence in an experimental setting would be present if students are cheating on a test together. As a result, the students that cheated will have similar test scores that differ from the rest of the class. Students are NOT working independently in this case, and therefore the test scores they produce are not independent. If each test score was an observation in our study, these observations would NOT be independent.

   How to check that it's present: There's no easy way to check for the independence of observations other than evaluating the data collection process. However, if you are working

with time series data (i.e. GDP data over the years, monthly price data), then we KNOW that observations are correlated. (GDP next year is dependent on GDP this year.) Randomized controlled trials (RCTs) are the easiest way to make sure that observations are independent. In the case of an RCT, participants are randomly assigned to treatment and control groups. Therefore, test subjects are randomly assigned and observations are independent.

Bad things that happen if it's not present: Standard errors for each explanatory variable are inaccurate. This leads to inaccurate tests of individual significance.

What to do if it's not present: Newey-West robust standard errors

3. ***No Hidden or Missing Variables***

What does it mean: This means you have included all relevant explanatory variables in the linear regression. If we are failing to explicitly include necessary variables in the model, then we are missing variables. A "hidden" variable is something that is not explicitly modeled but rather implicitly "hidden" behind the definition of another variable.

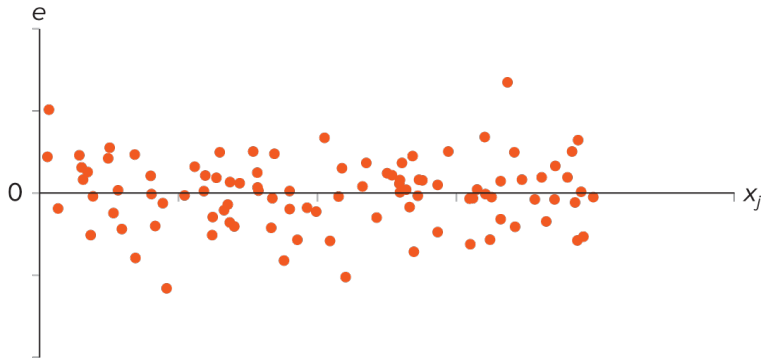
How to check that it's present: If adding a variable to the model makes a significant difference, it means that the model is incorrect and useless without it. That means the model is missing this necessary variable.

Bad things that happen if it's not present: If you do not include all necessary explanatory variables, you will end up with a misspecified model. That means Excel will attempt to produce coefficients for variables that are not included in the model. It also means that the coefficient estimates that it DOES produce will be inaccurate.

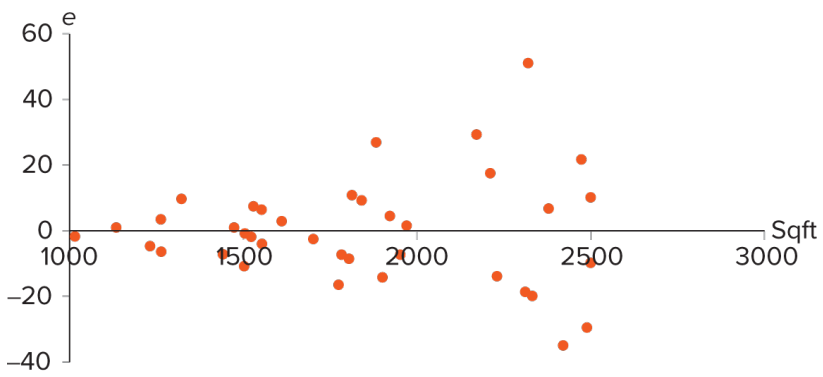
What to do if it's not present: Collect more data and include the missing variables!

4. ***Homoscedasticity***

What does it mean: When homoscedasticity is present, the variance of the error term is constant across all observations. The error term includes anything and everything that is not measured in the model. Remember "variance" is a measure of data dispersion. An example of constant variance is below. The values of the error term are all within the same "range" in distance from the x axis. Constant variance does NOT mean a constant value, but rather a constant range of values.



An example of NON-constant variance is below. The values of the error terms increase as the x values increase.



How to check that it's present: Scatterplot of residuals. Ideally, we want a scatterplot in which the residuals are centered on the horizontal axis and don't "grow" apart from the axis as the x values increase.

Bad things that happen if it's not present: Standard errors for each explanatory variable are inaccurate. This leads to inaccurate tests of individual significance.

What to do if it's not present: White's Robust Standard Errors

5. ***Normality of Residuals***

What does it mean: Normality of residuals simply means that the model's residuals are normally distributed. (Residuals are the difference between each predicted and observed value.)

How to check that it's present: Density plot. This plot should be roughly in the shape of a bell to identify as a normal distribution.

Bad things that happen if it's not present: Standard errors for each explanatory variable are inaccurate. This leads to inaccurate tests of individual significance.

What to do if it's not present: If the residuals are not normally distributed, my initial guess is that this is due to a nonlinear relationship between the outcome and explanatory variables. Therefore, the proposed solution is to choose different variables that would provide a linear relationship.

6. ***Explanatory variables are uncorrelated with each other***

What does it mean: Uncorrelated explanatory variables means we have NO multicollinearity present. This is good. We do NOT want multicollinearity.

How to check that it's present: We can check if explanatory variables are correlated by estimating the correlation coefficient in Excel (=CORREL) between variables. If the correlation value is > 0.5, that means we've got strong correlation between explanatory variables.

Bad things that happen if it's not present: If the explanatory variables ARE correlated with each other, the beta coefficients are biased. This means that if we used our linear regression to predict future values, those predictions would likely be highly inaccurate.

What to do if it's not present: If the explanatory variables ARE correlated with each other, the easiest solution is to remove one of the correlated variables.

7. ***Explanatory variables are uncorrelated with the error term***

What does it mean: This means the error term ϵ is not correlated with any of the explanatory variables; or in statistical terminology, there is no endogeneity. The error term includes anything and everything that is not measured in the model. When things included in the error term are correlated with an explanatory variable, we have endogeneity.

How to check that it's present: Unfortunately, there is no explicit way to test if a model has endogeneity. However, seasoned economists draw from traditional economic theory to determine if an omitted variable is introducing bias in the model.

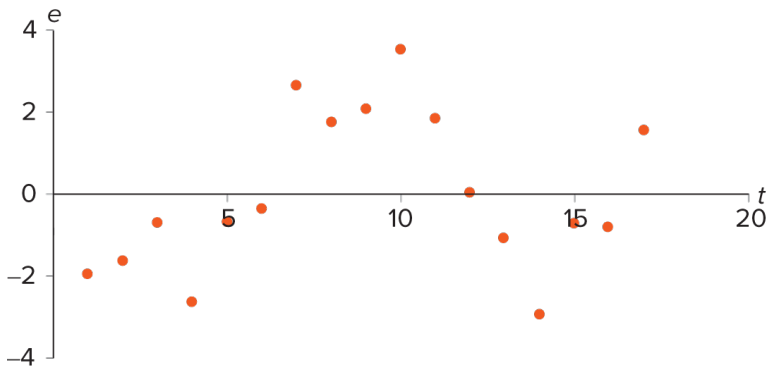
Bad things that happen if it's not present: Omitted variable bias (OVB). The primary consequences of OVB are biased beta coefficients. The linear regression's coefficients are going to be inaccurate if omitted variable bias is happening. This also relates to the "No Hidden or Missing Variables" standard above.

What to do if it's not present: The easiest solution is to explicitly include the omitted variable in the linear regression as one of the explanatory variables.

8. ***Observations of the error term are uncorrelated with each other***

What does it mean: If error term observations are uncorrelated, this means that variables we are not including in the linear regression are NOT correlated with each other. We do not want observations of the error term to be correlated with each other. When they are correlated, that means we have a case of serial correlation and this is NOT good news for our linear regression model. We often see serial correlation in time series data (i.e. GDP over the years).

How to check that it's present: Residual plot (residuals on y axis, observation index on x axis). If the plot reveals some kind of pattern, that means the error terms are correlated across observations. The plot below demonstrates a wave-like pattern for the residuals of time series data. Given this pattern around the horizontal axis, we conclude that the observations of the error terms are correlated.



Bad things that happen if it's not present: Standard errors for each explanatory variable are inaccurate. This leads to inaccurate tests of individual significance.

What to do if it's not present: Newey-West robust standard errors

15.2 Interval Estimates for the Response Variable

Now that we've learned how to make predictions based on a linear regression model, it's important to understand that these results may be off depending on which sample we use. For example, if we are trying to predict house prices in an area based on local crime statistics, our model estimates are going to change depending on which sample of neighborhoods we evaluate. We can compute intervals that will give us a range of expected model beta estimates.

In this section, we will construct two types of intervals:

1. Confidence Interval for the Average Value of Y

   This computes a range of values based on the average value of the response variable. For example, a range of values based on the average house value.

2. Prediction Interval for a Specific Value of Y

   This computes a range of values based on a specific value of the response variable. For example, a range of values based on a specific expected house value.

4 Steps to a Confidence Interval for the Average Value of Y

1. Compute the mean value for each of your explanatory variables.
2. Subtract that mean value from each observation in order to create new "adjusted" explanatory variables.
3. Compute regression results using these newly computed explanatory variables.
4. Compute the confidence interval using the equation below:

$$\hat{y}^0 \pm t_{\alpha/2,df} * se(\hat{y}^0)$$

Let's break down what each of these terms mean:

- $\hat{y}^0$ is the coefficient of the intercept
- $se(\hat{y}^0)$ is the standard error of the intercept
- $t_{\alpha/2,df}$ is the t value at $df = n - k - 1$ where n is the number of observations and k is the total number of variables in the model

Let's do an example. Using the regression model $Win = \beta_0 + \beta_1 * BA + \beta_2 * ERA$, construct the 95% confidence interval for the average winning percentage. We have 30 observations in our sample.

Step one: compute the mean value for each of your explanatory variables. Using the "baseball" file on canvas, compute the average for each of the explanatory variables, that is the averages for BA and ERA. Store these values right below the columns, in cells D32 and E32. The average BA value is 0.257 and the average ERA value is 4.073.

Step two: Subtract that mean value from each observation in order to create new "adjusted" explanatory variables. In cell F2, type the command =D2-$D32$, this will subtract the average BA value from each BA observation. Click and drag down the entire column, so it does this same computation for every other BA value. In cell G2, type the command =E2-$E32$, and click and drag down the entire column. You can label these columns however you like.

Step three: compute regression results using these newly computed explanatory variables. Now, we run a regression with these new variables. That is, we click: Data > Data Analysis > Regression. The Input Y Range will be $C1{:}C31$ and the input X range will be $F1{:}G31$. The regression results are below:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8459241
R Square	0.7155876
Adjusted R S	0.6945201
Standard Err	0.037549
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	0.0957799	0.0478899	33.966292	4.249E-08
Residual	27	0.038068	0.0014099		
Total	29	0.1338479			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.5000667	0.0068555	72.944149	1.501E-32	0.4860004	0.5141329	0.4860004	0.5141329
BA modified	3.2754457	0.6723078	4.8719439	4.297E-05	1.8959841	4.6549073	1.8959841	4.6549073
ERA modified	-0.1152602	0.0166569	-6.9196687	1.952E-07	-0.1494374	-0.0810831	-0.1494374	-0.0810831

Step four: Compute the confidence interval using the equation below:

$$\hat{y}^0 \pm t_{\alpha/2,df} * se(\hat{y}^0)$$

Let's break down what each of these terms mean:

- $\hat{y}^0$ is the coefficient of the intercept
- $se(\hat{y}^0)$ is the standard error of the intercept
- $t_{\alpha/2,df}$ is the t value at $df = n - k - 1$ where n is the number of observations and k is the total number of variables in the model

Let's define each of the terms we need based on the model output:

- $\hat{y}_0 = 0.5000$
- $se(\hat{y}_0) = 0.0069$
- $t_{\alpha/2,df} = t_{.05/2,30-2-1} = t_{.025,27} = 2.052$

  We learned t values back in ECON 261. Instead of looking at a t-table, I recommend using this calculator online, where you can directly enter the value of α without having to divide it by 2. Always select the t-value for a 2-tailed test in these kinds of confidence intervals.

  If you would like to be traditional, you can use this t table distribution and manually interpret the table. But the focus of this class is not interpreting these tables, so I am fine with you using online t value calculators. On the test, I will give you the t value in a case like this.

We can plug all these values into our equation:

$$\hat{y}^0 \pm t_{\alpha/2,df} * se(\hat{y}^0)$$

$$0.50 \pm t_{.025,27} * 0.0069$$

$$0.50 \pm 2.052 * 0.0069$$

$$0.50 + 0.0141 = 0.5142$$

$$0.50 - 0.0141 = 0.4858$$

$$[0.4858, 0.5142]$$

Using this 95% confidence interval, we can state that the mean winning percentage of a team with a batting average(BA) of 0.257 and earned run average(ERA) of 4.073 falls between 0.4858 and 0.5142.

4 steps to a Prediction Interval for a Specific Value of Y

1. Compute the mean value for each of your explanatory variables.

2. Subtract that mean value from each observation in order to create new "adjusted" explanatory variables.

3. Compute regression results using these newly computed explanatory variables.

4. Compute the prediction interval using the equation below:

$$\hat{y}^0 \pm t_{\alpha/2,df} * \sqrt{(se(\hat{y}^0))^2 + s_e^2}$$

Let's break down what each of these terms mean:

- $\hat{y}^0$ is the coefficient of the intercept
- $se(\hat{y}^0)$ is the standard error of the intercept
- s_e^2 is the standard error of the entire regression, this is under the "regression statistics" table in the excel output
- $t_{\alpha/2,df}$ is the t value where $df = n - k - 1$, n is the number of observations, k is the total number of variables in the model

Let's do an example. Using the regression model $Win = \beta_0 + \beta_1 * BA + \beta_2 * ERA$, construct the 95% prediction interval for the average winning percentage. We have 30 observations in our sample.

Step one: compute the mean value for each of your explanatory variables. Using the "baseball" file on canvas, compute the average for each of the explanatory variables, that is the averages for BA and ERA. Store these values right below the columns, in cells D32 and E32. The average BA value is 0.257 and the average ERA value is 4.073.

Step two: Subtract that mean value from each observation in order to create new "adjusted" explanatory variables. In cell F2, type the command =D2-$D32$, this will subtract the average BA value from each BA observation. Click and drag down the entire column, so it does this same computation for every other BA value. In cell G2, type the command =E2-$E32$, and click and drag down the entire column. You can label these columns however you like.

Step three: compute regression results using these newly computed explanatory variables. Now, we run a regression with these new variables. That is, we click: Data > Data Analysis > Regression. The Input Y Range will be $C1$:$C31$ and the input X range will be $F1$:$G31$. The regression results are below:

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.8459241							
R Square	0.7155876							
Adjusted R S	0.6945201							
Standard Err	0.037549							
Observations	30							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	0.0957799	0.0478899	33.966292	4.249E-08			
Residual	27	0.038068	0.0014099					
Total	29	0.1338479						
	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.5000667	0.0068555	72.944149	1.501E-32	0.4860004	0.5141329	0.4860004	0.5141329
BA modified	3.2754457	0.6723078	4.8719439	4.297E-05	1.8959841	4.6549073	1.8959841	4.6549073
ERA modified	-0.1152602	0.0166569	-6.9196687	1.952E-07	-0.1494374	-0.0810831	-0.1494374	-0.0810831

Step four: Compute the prediction interval using the equation below:

$$\hat{y}^0 \pm t_{\alpha/2,df} * \sqrt{(se(\hat{y}^0))^2 + s_e^2}$$

Let's break down what each of these terms mean:

- $\hat{y}^0$ is the coefficient of the intercept
- $se(\hat{y}^0)$ is the standard error of the intercept
- s_e^2 is the standard error of the entire regression, this is under the "regression statistics" table in the excel output
- $t_{\alpha/2,df}$ is the t value where $df = n - k - 1$, n is the number of observations, k is the total number of variables in the model

Let's define each of the terms we need from the Excel output:

- $\hat{y}_0 = 0.5000$
- $se(\hat{y}_0) = 0.0069$
- $s_e^2 = 0.0375^2$
- $t_{\alpha/2,df} = t_{.05/2,30-2-1} = t_{.025,27} = 2.052$

  We learned t values back in ECON 261. Instead of looking at a t-table, I recommend using this calculator online, where you can directly enter the value of α without having to divide it by 2.

If you would like to be traditional, you can use this t table distribution and manually interpret the table. But the focus of this class is not interpreting these tables, so I am fine with you using online t value calculators. On the test, I will give you the t value in a case like this.

We can plug all of these values into our equation:

$$\hat{y}^0 \pm t_{\alpha/2,df} * \sqrt{(se(\hat{y}^0))^2 + s_e^2}$$

$$0.50 \pm t_{0.025,27} * \sqrt{0.0069^2 + 0.0375^2}$$

$$0.50 \pm 2.052 * \sqrt{0.0000 + 0.0014}$$

$$0.50 \pm 2.052 * \sqrt{0.0014}$$

$$0.50 \pm 2.052 * (0.0374)$$

$$0.50 \pm 0.0767$$

$$0.50 + 0.0767 = 0.5767$$

$$0.50 - 0.0767 = 0.4232$$

$$[0.4232, 0.5767]$$

Using this 95% prediction interval, we can state that the winning percentage of a team with a batting average(BA) of 0.257 and earned run average(ERA) of 4.073 falls between 0.4232 and 0.5767. In the previous example, the corresponding 95% confidence interval was between 0.4858 and 0.5142. As expected, the prediction interval is wider because it also accounts for the variability caused by the random error term. The higher variability, captured by the standard error of the estimate s_e, makes it more difficult to predict accurately, thus generating a wider interval.

16 Week 16: May 8 - May 12

16.1 Final Exam Review in class on May 8

16.2 Final Exam on ? via Canvas