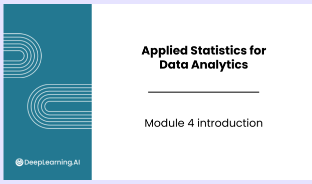
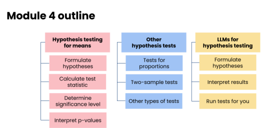
DAG C2M4 scripts

Video title	Isabel	Sean	Slides
L0V1 – Module 4 introduction	✅	✅	✅
L1V1 – Demo: hypothesis testing in action	✅	✅	✅
L1V2 – Hypothesis testing for means	✅	✅	✅
L1V3 – The hypothesis	✅	✅	✅
L1V4 – Identifying the hypothesis and test type	✅	✅	✅
L1V5 – Calculating the test statistic	✅	✅	✅
L1V6 – Determining significance level and rejection region	✅	✅	✅
L1V7 – Calculating the p value	✅	✅	✅
L1V8 – Demo: hypothesis testing for means	✅	✅	✅
L1V9 – Hypothesis testing errors	✅	✅	✅
L1V10 – The t distribution	✅	✅	✅
L2V1 – Hypothesis testing for proportions	✅	✅	✅
L2V2 – Demo: hypothesis testing for proportions	✅	✅	✅
L2V3 – Two sample tests	✅	✅	
L2V4 – Other hypothesis tests	✅	✅	
L3V1 – Interpretation with LLMs	✅	✅	✅
L3V2 – Inference with LLMs	✅	✅	✅
L3V3 – Your next steps	✅	✅	✅

Introduction

L0V1 – Module 4 introduction

Visual	Script
🎙️ TH	Welcome to the final module of this course: Hypothesis Testing! In this module, you'll dive into one of the most powerful and widely-used tools in inferential statistics. Throughout these lessons, you'll see how hypothesis tests can help you make

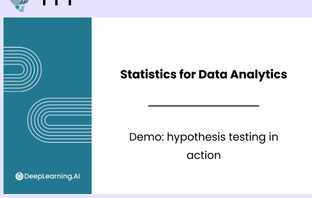
<table>
<tr>
<td>
</td>
<td>data-driven decisions in real-world business scenarios, answering questions like “Does a free trial improve user retention” and “Are delivery times really faster than 45 minutes?”</td>
</tr>
<tr>
<td>
</td>
<td>You'll start by exploring the fundamental concepts of [CLICK] hypothesis testing for means. You'll learn how to [CLICK] formulate your hypothesis, [CLICK] calculate the test statistic, [CLICK] determine your desired significance level, and [CLICK] interpret p-values.

Then, you'll expand your toolkit with [CLICK] hypothesis tests for proportions and [CLICK] two-sample tests. You'll discover how to compare different groups and evaluate whether observed differences are statistically significant. You'll also briefly explore [CLICK] other types of hypothesis tests you might encounter in your work as a data analyst. Throughout the module, you'll get hands-on practice performing various hypothesis tests in spreadsheets, as well as interpreting the results to answer important business questions.

In the final lesson, you'll explore how [CLICK] large language models can assist you in the hypothesis testing process. You'll learn to leverage AI to help [CLICK] formulate hypotheses, [CLICK] interpret results, and even [CLICK] run tests for you.</td>
</tr>
<tr>
<td>🎙️ TH</td>
<td>Are you ready for this final module? Follow me to the next video where you'll jump right into hypothesis testing with a practical example.</td>
</tr>
</table>

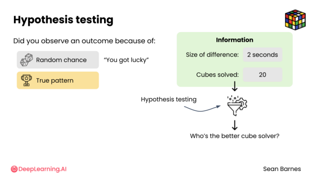
Lesson 1 – Hypothesis testing for means

L1V1 – Demo: hypothesis testing in action

<table>
<tr>
<th>Visual</th>
<th>Script</th>
</tr>
<tr>
<td>🎙️ TH

</td>
<td>Your former college roommate challenges you to see who can solve a Rubik’s cube faster. You think you’re evenly matched, so you agree.

[Sean solves Rubik’s cube in several jump cuts with flourish, sets it down]

You solve the cube in 92 seconds, but he solves it in 71. You lost the first match, but that doesn't make you a loser! You decide to keep going and end up solving 20 cubes each. Your fingers are numb, but you’ve achieved an average time of 80 seconds, while your friend has an average of 82. Your friend says “you just got lucky, that’s nothing, my cubes were more scrambled, I’ll totally beat you next time”. But you’re thinking, “hey, I beat you by 2 seconds on average, fair and square.” Who’s right?</td>
</tr>
</table>

<table>
<tr>
<td>
</td>
<td>Hypothesis testing can help you answer just this type of question. [CLICK] Did you observe a particular outcome because of [CLICK] random chance? That's what your friend is suggesting – that [CLICK] you just got lucky. Or is there some [CLICK] true pattern here, and you are in fact the better cube solver.

The [CLICK] size of the difference is pretty small – you only won by [CLICK] 2 seconds. And you did only [CLICK] solve 20 cubes. A [CLICK] hypothesis test can [CLICK] take all this information into account and help you[CLICK] reach a conclusion with high confidence: [CLICK] who's the better cube-solver?</td>
</tr>
<tr>
<td>🖥️ SC – 🔗 link to demo spreadsheet</td>
<td>Let's see hypothesis testing in action on this example so you can get a feel for how it works.

You have the first 20 solves here in orange, along with the differences in time. Positive values mean that you solved the cube faster than your friend.

▶️ In E3, =average(A3:A22)
This is the average difference that you just saw, that you were 2 seconds faster on average over the first 20 solves.

Now, say you were actually evenly matched. What's the probability that you could observe these results where you were faster than your friend by 2 seconds? I'll calculate it using a hypothesis test, which you'll learn how to do over the next few videos.

▶️ In E6, =Z.TEST(C3:C21,0)

24.7%! So if you were evenly matched, this kind of result would be pretty common. There is a lot of variability in the data, and it's only 20 solves, and it's only a 2 second difference, which is pretty small. So it looks like you're solving faster, but you don't have a ton of evidence.
Now, say after finding out these results, you wanted to settle this once and for all. You go on to do 100 solves and record your results.

▶️ In E10, =average(C2:C)

This time, your average is faster by 3 seconds. Given that this result is based on more tries, could this be evidence that you are actually faster? You can calculate the probability of observing a 3 second difference or larger if you're equally skilled

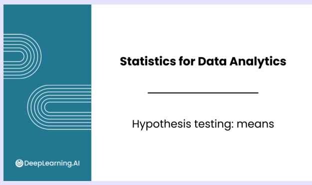
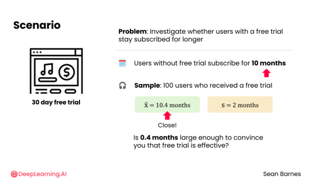
▶️ In D13, =Z.TEST(C2:C101,0)

Only 0.37%! So if you were in fact equally skilled, this difference would be rare. That rarity reflects that you were able to maintain an average of a 3 second advantage over your friend across 100 solves. Let's face it, after 100 chances, if your friend hasn't started solving faster than you, it doesn't seem likely he's truly better than you.
</td>
</tr>
<tr>
<td>🎙️ TH</td>
<td>The hypothesis test you just conducted gives you reason to believe that you are in fact faster than your friend, even if it's only by 3 seconds on average!

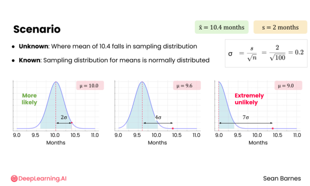
Now that you've seen how hypothesis tests can evaluate the likelihood of certain events, follow me to the next video to see what's happening under the</td>
</tr>
</table>

<table>
<tr><td></td><td>hood.</td></tr>
</table>

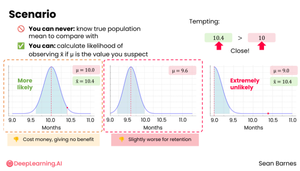
L1V2 – Hypothesis testing for means

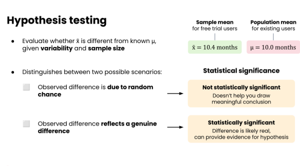
<table>
<tr><th>Visual</th><th>Script</th></tr>
<tr><td>🎙️ TH

</td><td>In Data Analytics Foundations, you saw that data analytics has a lot in common with other investigative fields like science. One powerful investigative tool you have is hypothesis testing for means, which allows you to rigorously evaluate whether your sample mean is significantly different from a particular value.</td></tr>
<tr><td>
</td><td>Let's return to the example from the previous module where you were working with a music streaming service. Your team has rolled out the free trial to more users, and you're working on a new project to figure out if giving users a free trial improves user retention. You decide to investigate [CLICK] whether users who receive a free trial stay subscribed for longer on average.

Previous analytics has shown that [CLICK] users who don't receive a free trial tend to stay subscribed for about 10 months.

You collect a [CLICK] sample of 100 users who received a free trial and calculate the following descriptive statistics:
[CLICK] The sample mean is 10.4 months
[CLICK] The sample standard deviation is 2 months
[CLICK] That's close! What do you think, [CLICK] is the difference of 0.4 months large enough to convince you that the free trial is effective? [pause for thought]</td></tr>
<tr><td>
</td><td>You can't be sure just yet. What you don't know based on these descriptive statistics is [CLICK] where this mean of 10.4 falls in your sampling distribution.

You know from studying the Central Limit Theorem that the [CLICK] sampling distribution for means is normally distributed. Now, there are infinitely many possible scenarios, but here are three for you to consider:
[CLICK] The true population mean mu is 10 months, so [CLICK] 10.4 months falls here in the sampling distribution.
[CLICK] Mu is 9.6 months, and [CLICK] 10.4 months falls around here in the sampling distribution.
[CLICK] And lastly, mu is 9 months, and [CLICK] 10.4 months falls here in the sampling distribution
Based on what you know about the normal distribution, which of these outcomes is most likely? [pause for thought]</td></tr>
</table>

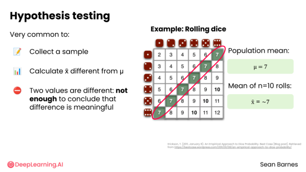
<table>
<tr>
<td></td>
<td>
The first outcome is [CLICK] more likely than the other two. Recall that the [CLICK] standard error is [CLICK] s over square root n, so here it’s [CLICK] 2 over square root 100, which is 2/10 or [CLICK] 0.2. Remember that standard error is just a special name for the standard deviation of the sampling distribution.
In the first case, x-bar is [CLICK] 2 standard deviations from the mean; in the second, [CLICK] x-bar is 4 standard deviations away; and in the last, [CLICK] x-bar is 7 standard deviations away. You know from the three sigma rule that 99.7% of the data falls within 3 standard deviations of the mean, so getting an outcome that is a whopping 7 standard deviations above the mean is [CLICK] extremely unlikely.
</td>
</tr>
<tr>
<td>


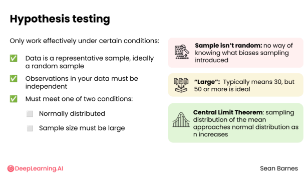
</td>
<td>
The tricky part is, [CLICK] you can never know the true population mean to compare with. But what [CLICK] you can do is calculate how likely it is that you would observe the sample mean you calculated if the population mean is in fact the value you suspect.
The idea here is that if your true population mean was actually [CLICK] 9 months, it would be [CLICK] extremely, extremely unlikely for you to take a sample of 100 users and find that the sample mean was 10.4 months. Whereas, if the true population mean is [CLICK] 10 months, [CLICK] it’s much more likely.
Now, in this case, it’s so [CLICK] tempting to say [CLICK] 10.4 is higher and call it a day!
But, these numbers are [CLICK] close. It is absolutely possible that the true mean is actually [CLICK] 10, and the free trial [CLICK] costs you money while giving you no benefit. Or maybe [CLICK] it’s 9.6, and the free trial is [CLICK] slightly worse for retention. This level of precision is important for your conclusions.
</td>
</tr>
<tr>
<td>


</td>
<td>
What a hypothesis test allows you to do is [CLICK] evaluate whether your [CLICK] sample mean for free trial users (10.4 months) is significantly different from the [CLICK] known population mean for existing users (10 months), given the variability in your data and your sample size. It [CLICK] distinguishes between two possible scenarios:

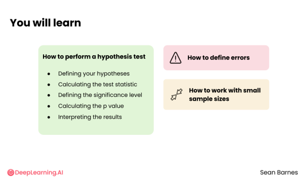
First, that the observed difference between the sample mean of free trial users and the mean of existing users is [CLICK] due to random chance. The values are too close for you to tell if they’re actually different.
Or, that the observed difference [CLICK] reflects a genuine difference between the true population mean of free trial users and the mean of existing users.

</td>
</tr>
</table>

<table>
<tr>
<td></td>
<td>This distinction is called **[CLICK]** **statistical significance**. If the difference is due to random chance, it's **[CLICK]** not statistically significant. It **[CLICK]** doesn't help you draw any meaningful conclusion. On the other hand, if the observed difference reflects a genuine difference between the hypothesized mean and the sampled value, that is **[CLICK]** statistically significant. The **[CLICK]** difference is likely real and can provide evidence for your hypothesis.</td>
</tr>
<tr>
<td>
</td>
<td>It's very common for you to **[CLICK]** collect a sample and **[CLICK]** calculate a mean that's different from the true population mean. For example, when **[CLICK]** rolling two dice multiple times, the **[CLICK]** population mean of their sum is 7, which is also **[CLICK]** the most common roll. But, if you take a sample of say **[CLICK]** 10 rolls and calculate the mean you're very unlikely to get 7 exactly — you **will** get **[CLICK]** values that **vary** around 7.

So just observing that **[CLICK]** two values are different isn't enough to conclude that that difference is meaningful, ~~that's just how the cookie crumbles sometimes.~~</td>
</tr>
<tr>
<td>
</td>
<td>Hypothesis tests only work effectively under certain conditions. It's crucial that **[CLICK]** your data is a representative sample, ideally a random sample. A majority of statistical tests assume random sampling because **[CLICK]** if your sample isn't random, you have no way of knowing what biases your sampling method introduced.

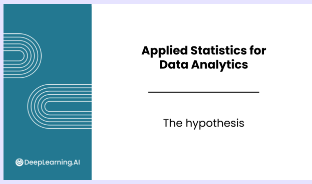
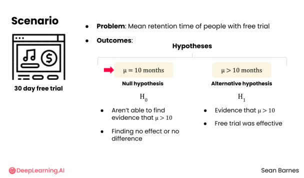
[CLICK] The observations in your data must also be independent.

Additionally, your data must meet **[CLICK]** one of two conditions. Either it is **[CLICK]** normally distributed, or **[CLICK]** your sample size must be large. **[CLICK]** Typically "large" means 30, but 50 or more is ideal. That's because of the Central Limit Theorem, which states that **[CLICK]** the sampling distribution of the mean approaches a normal distribution as the sample size increases. You've already seen this in action during the previous module.</td>
</tr>
<tr>
<td>
</td>
<td>Over the next few videos, you'll see **[CLICK]** how to perform a hypothesis test for means. You'll cover these steps:
● **[CLICK]** Defining your hypotheses
● **[CLICK]** Calculating the test statistic
● **[CLICK]** Defining the significance level
● **[CLICK]** Calculating the p value
● **[CLICK]** And interpreting the results

You'll also learn **[CLICK]** how to define errors and **[CLICK]** work with small sample sizes. Don't worry about all the terminology for now — you'll be a pro at hypothesis testing by the end of this lesson.</td>
</tr>
<tr>
<td>🎙️ TH</td>
<td>Your first step in hypothesis testing is defining the hypothesis. It's crucial for making sure you construct and interpret your test correctly. Follow me to the</td>
</tr>
</table>

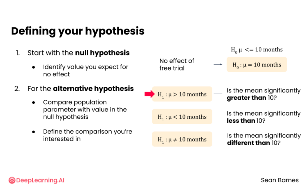
	next video to learn how.

L1V3 – The hypothesis

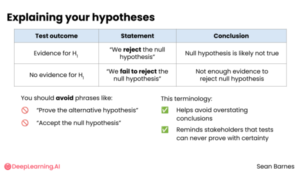
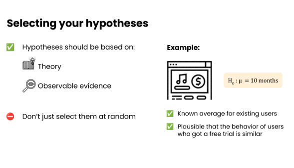
<table>
<tr><th>Visual</th><th>Script</th></tr>
<tr><td>🎙️ TH
</td><td>In statistical hypothesis testing, you’ll need to define two related hypotheses: the null hypothesis and the alternative hypothesis. These two hypotheses complement each other.</td></tr>
<tr><td></td><td>Breaking down the music subscription service scenario, you were investigating [CLICK] the mean retention time of people who got a free trial. There were two possible [CLICK] outcomes of your test:
[CLICK] Mu = 10 months
And [CLICK] mu > 10 months

These are your [CLICK] hypotheses. Now let me ask you, which of these hypotheses corresponds with the idea that getting a free trial has no effect on subscription length? [pause for thought]

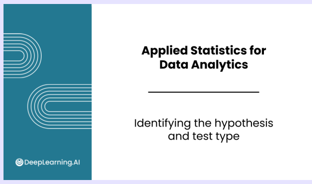
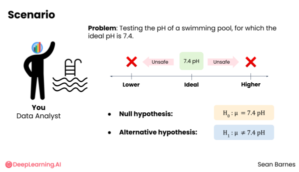
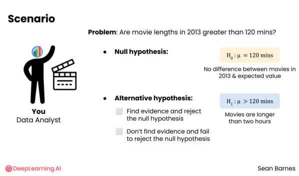
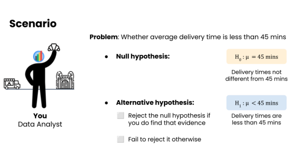
That would be [CLICK] mu = 10. This hypothesis is called the [CLICK] null hypothesis, and it represents the condition where [CLICK] you aren’t able to find evidence that mu was greater than 10. It’s associated with [CLICK] finding no effect or no difference. This hypothesis is written as [CLICK] H sub 0, or “H naught”.

Alternatively, if [CLICK] you could find evidence that mu is greater than 10 months, that would be great! That evidence would indicate that the [CLICK] free trial was effective in getting users to subscribe for longer. This hypothesis is called the [CLICK] alternative hypothesis – it’s the alternative to the null – and it’s written [CLICK] H sub 1.</td></tr>
<tr><td></td><td>In general, when defining your hypotheses, [CLICK] start with the null hypothesis. [CLICK] Identify the value you would expect if there was no effect. For example, [CLICK] if there was no effect of getting a free trial, you would expect a mean subscription length of 10 months, the same as existing subscribers. That gave you [CLICK] H naught of mu = 10.

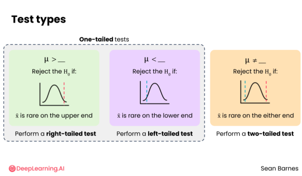
You’ll sometimes see the null hypothesis written as an inequality. In this case, that would look like [CLICK] H naught of mu <= 10. Both approaches are valid, but equality is a more common convention and doesn’t affect the interpretation of results.</td></tr>
</table>

	[CLICK] For your alternative hypothesis, you'll always **[CLICK]** compare the population parameter with the value in the null hypothesis, in this case 10. **[CLICK]** Define the comparison you're interested in: are you looking for evidence that the mean is greater than, smaller than, or just different than the expected mean? Your options would be ● **[CLICK]** H sub 1 mu > 10, **[CLICK]** is the mean significantly greater than 10? ● **[CLICK]** H sub 1 mu < 10, **[CLICK]** is the mean significantly less than 10? ● **[CLICK]** H sub 1 mu =/= 10, **[CLICK]** is the mean significantly different than 10? You can only have 1 alternative hypothesis. So in the example from earlier, **[CLICK]** mu > 10 is most appropriate, since you were hoping to find evidence that the free trial increased subscription length.
	When you're explaining these hypotheses to your business stakeholders, it's important to use exactly the right terms. ● If your test indicates **[CLICK]** evidence for the alternative hypothesis (more on the specifics of how later) – you would say you **[CLICK]** "reject the null hypothesis". The data suggest that the **[CLICK]** null hypothesis is likely not true. ● If you **[CLICK]** don't find evidence for the alternative hypothesis, then you would say that you **[CLICK]** "fail to reject the null hypothesis". It doesn't mean that the null hypothesis is true, just that **[CLICK]** you don't have enough evidence to reject it. The language of statistics is important here. **[CLICK]** You should avoid phrases like **[CLICK]** "prove the alternative hypothesis" and **[CLICK]** "accept the null hypothesis". It may feel like you're being deliberately vague, but remember that inferential statistics is all about managing uncertainty. There's always a chance that your conclusions are incorrect. **[CLICK]** This terminology **[CLICK]** helps avoid overstating your conclusions, and **[CLICK]** reminds your stakeholders that these tests can never prove with absolute certainty.
	As in science, your **[CLICK]** hypotheses should be based on **[CLICK]** some theory or **[CLICK]** observable evidence. In other words, **[CLICK]** don't just select them at random. For example, when dealing with the **[CLICK]** music subscription service, it makes sense to choose **[CLICK]** the null hypothesis that subscription length equals 10, since that's the **[CLICK]** known average for existing users. It's **[CLICK]** plausible that the behavior of users who got a free trial is similar.
🎙️ TH	Hypotheses are the foundation of your testing strategy, and they must be defined first. In the next video, you'll get practice identifying hypotheses depending on your business problem. I'll see you there!

L1V4 – Identifying your hypothesis and test type

Visual	Script
🎙️TH 	Let's take a look back at some of the business problems you've seen in this course so far and formulate hypotheses for them. You'll also see how to determine which type of hypothesis test is appropriate for a given use case.
	First, is the water in a swimming pool safe? Say you were **[CLICK]** testing the pH of a swimming pool, for which the **[CLICK]** ideal pH is 7.4. **Any** significant difference from 7.4 is unsafe, whether it's **[CLICK]** higher **or** **[CLICK]** lower. Pause the video for a moment and see if you can jot down the hypotheses you would test. **[pause for thought]** **[CLICK]** Your null hypothesis H naught would be **[CLICK]** mu equals 7.4. This value represents the status quo. **[CLICK]** Your alternative hypothesis would be for unsafe pH levels, in this case **[CLICK]** H sub 1 of mu =/= 7.4. If the pH differs significantly either above or below 7.4, you would reject the null hypothesis.
	You also investigated business questions related to movie lengths. For example, **[CLICK]** are movie lengths in 2013 greater than 120 minutes? I encourage you to pause the video for a moment and see if you can work out the hypotheses **[pause for thought]** **[CLICK]** Your null hypothesis would be that **[CLICK]** mu equals 120 minutes. That value represents **[CLICK]** no difference between the movies in 2013 and what you expected. Your **[CLICK]** alternative hypothesis would be that **[CLICK]** mu > 120 minutes. You're looking for evidence that **[CLICK]** movies are longer than two hours. Either you **[CLICK]** find that evidence, and you reject the null hypothesis, or you **[CLICK]** don't find that evidence and you fail to reject the null hypothesis.
	In the previous module, you also constructed confidence intervals for bakery delivery times. Recall that you recorded delivery times for 30 days and measured the average time at 43 minutes. You could use a hypothesis test to investigate **[CLICK]** whether the average delivery time is **less than** 45 minutes. Pause the video one more time and try to figure out the hypotheses. **[pause for thought]** **[CLICK]** The null hypothesis would be that **[CLICK]** mu = 45 minutes. That's the status quo, if **[CLICK]** delivery times weren't different from 45 minutes.

<table>
<tr>
<td></td>
<td>The [CLICK] alternative hypothesis would be [CLICK] mu < 45 minutes. You’re interested in finding evidence that the [CLICK]delivery times are less than 45 minutes. [CLICK] You can reject the null hypothesis if you do find that evidence, or [CLICK] fail to reject it otherwise.

Notice that your hypothesized mean delivery time is different from the sample mean you calculated, which was 43 minutes. Your goal is to use that sample mean to understand how likely it is that the population mean is under 45 minutes.</td>
</tr>
<tr>
<td>
</td>
<td>Once you’ve defined your hypothesis, you’ll need to select the appropriate test type. You saw in the previous video that you can select from [CLICK] three types of alternative hypotheses:

[CLICK] mu > some number
[CLICK] mu < some number
[CLICK] And mu =/= some number

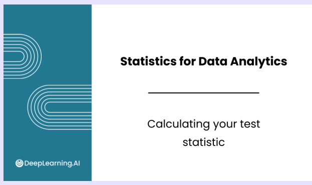
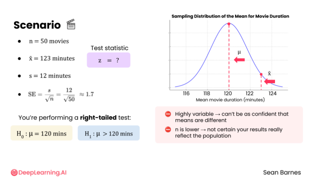
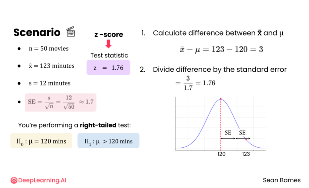
These hypotheses are each associated with a different type of test.

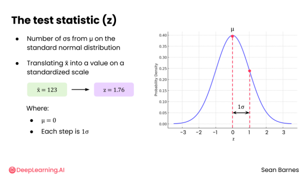
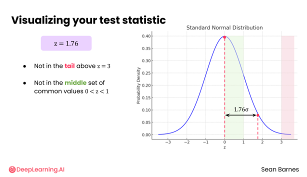
In the first case, you’ll [CLICK] reject the null hypothesis if your sample mean is rare on the upper end of the distribution. You should [CLICK] perform a right-tailed test, since you’re only interested in the upper tail of values.
In the second case, you’ll [CLICK] reject the null when your sample mean is rare on the lower end. You should perform a [CLICK] left-tailed test, since you’re interested in rare values only below the mean.
Both left- and right-tailed tests are considered [CLICK] one-tailed tests, since you’re only checking if a value falls on one side of the distribution.
In the final case, you’re actually interested in both possibilities, so you’ll [CLICK] reject the null if you find that your sample mean is unusual on either the upper or lower end. You should [CLICK] perform a two-tailed test because you’re interested in whether x-bar falls in either the upper or lower tail of values.
</td>
</tr>
<tr>
<td>
</td>
<td>Let’s see how that shakes out with the three examples you just saw:

[CLICK] If you’re testing the water in a pool, with the alternative hypothesis H sub 1 that mu, the average pH, does not equal 7.4. What type of test would you want to perform? [pause for thought] You would want a [CLICK] two-tailed test, since either a pH that’s too high or too low would be unsuitable.
[CLICK] For testing movie lengths, recall that your alternative hypothesis was that mu is greater than 120 minutes, or on average movie durations are above that value. What type of test would that be? [pause for thought] You should perform a [CLICK] right-tailed test, since you want to check if the sample mean would be unusually high.
[CLICK] And if you’re checking whether the average delivery time is below 45 minutes, you’re interested in a [CLICK] left-tailed test.
</td>
</tr>
<tr>
<td>🎙️ TH</td>
<td>Great work defining your hypotheses and determining your test types! It’s</td>
</tr>
</table>

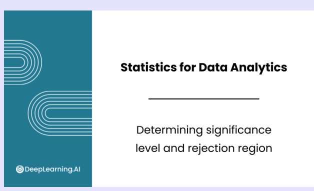
	nuanced but fun work. Follow me to the next video to see how to complete the next step of your hypothesis test: calculating the test statistic.

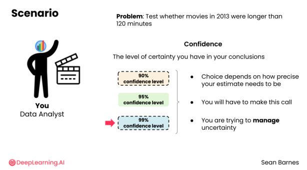
L1V5 – Calculating your test statistic

Visual	Script
🎙️ TH 	How do you make a decision about your hypothesis, given the sample data that you've collected? To answer that question, you'll need to calculate your test statistic.
	Say you're investigating movie durations. You've collected a **[CLICK]** random sample of 50 movies and calculated a **[CLICK]** sample mean of 123 minutes and **[CLICK]** sample standard deviation of 12 minutes. **[CLICK]** You're performing a right-tailed test, with the hypotheses: ● **[CLICK]** H sub 0, the null hypothesis, states that mu = 120 minutes ● And **[CLICK]** H sub 1, the alternative hypothesis, states that mu > 120 minutes **[CLICK]** Here's the sampling distribution of the mean for movie duration, **[CLICK]** centered around your null hypothesis of (mu is 120), and the **[CLICK]** standard error is **[CLICK]** 12/sqrt(50), which equals about **[CLICK]** 1.7. Is **[CLICK]** your sample mean different enough from your **[CLICK]** hypothesized mu to confidently reject the null hypothesis? It's hard to tell. Specifically because both variability and sample size affect your conclusions. ● If your data is **[CLICK]** highly variable, you can't be as confident that these means are truly different. ● The larger your sample size, the more precision your test has, but if your **[CLICK]** sample size is lower, you may not be certain your results really reflect the population. The **[CLICK]** test statistic accounts for the variability and size of your sample data. Once you've calculated it, you can determine how rare of a result it really is.
	~~As a recap, **[CLICK]** here are the descriptive stats for your sample.~~ ~~And **[CLICK]** here were your hypotheses.~~ First, you'll **[CLICK]** calculate the difference between x-bar and mu, **[CLICK]** which gives you 3. This step centers your calculations around 0.Then, you'll **[CLICK]** divide this difference by the standard error, which tells you **[CLICK]**

	the number of standard errors between this test statistic and the hypothesized mean. Since you don't know the population standard deviation, **[CLICK]** you will use the formula s over square root n, which gives you about 1.7. All together, **[CLICK]** the test statistic is 3 / 1.7, which equals 1.76. Did this calculation look familiar to something else you've done with the normal distribution? **[pause for thought]** You just calculated **[CLICK]** a z score!
	Recall that a z score is the **[CLICK]** number of standard deviations from the mean on the standard normal distribution. Essentially, you're **[CLICK]** translating your sample mean into a value on a standardized scale, where the **[CLICK]** mean is zero, and **[CLICK]** each step represents one standard deviation.
 https://i.imgur.com/zSkWu9B.png https://i.imgur.com/UPMCAci.png	Let's visualize this value. **[CLICK]** Here's the standard normal distribution. Can you spot where your test statistic z = 1.76 falls? **[pause for thought]** It's right **[CLICK]** here, 1.76 standard deviations above the mean. How rare would you say this test statistic is? It's tough to say just by looking at the graph. **[CLICK]** It's not in one of the tails above a z-score of 3, but it's also **[CLICK]** not in the middle set of common values where z is between 0 and 1.
🎙️ TH	Is this z score rare enough for you to reject your null hypothesis? Well, that depends. Join me in the next video to see how to answer that question by determining your significance level and rejection region.

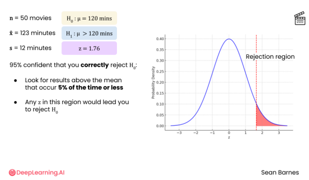
L1V6 – Determining the significance level and rejection region

Visual	Script
🎙️ TH 	In addition to calculating your test statistic, you'll need to determine how precise you want your test to be. What do you consider a sufficiently unlikely value that would lead you to reject your null hypotheses?

<table>
<tr>
<td>


</td>
<td>
Consider the movie duration example, where you want to **[CLICK]** test whether movies in 2013 were longer than 120 minutes on average, to help you schedule movie times in a theater. How high stakes is this decision? Are you okay coming to the wrong conclusion 5% of the time? Only 1% of the time?

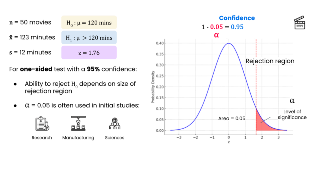
In your previous work with confidence intervals, you encountered the concept of **[CLICK]** “confidence”, or the **[CLICK]** level of **certainty** you have in your conclusions. You learned that you could construct an interval with **[CLICK]** 90, **[CLICK]** 95, or **[CLICK]** 99% confidence — or some other value as well. **[CLICK]** Your choice depends on how precise your estimate needs to be. If you needed to be very sure about your results, would you choose **[CLICK]** 90% or **[CLICK]** 99% confidence? **[pause for thought]** You’d choose **[CLICK]** 99%. Hypothesis tests rely on a similar intuition.

As a data analyst, **[CLICK]** you will have to make this call to determine what confidence level you are comfortable with. Remember, as with all inferential statistics, **[CLICK]** you are trying to **manage** uncertainty. It can never be fully eliminated.
</td>
</tr>
<tr>
<td>


https://i.imgur.com/gRjreRs.png
</td>
<td>
As a recap, **[CLICK]** here are the descriptive stats for your sample.

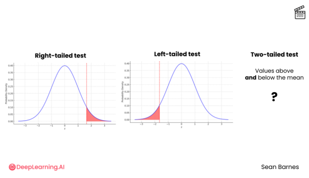
And **[CLICK]** here were your hypotheses.

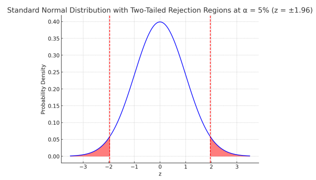
Since you’ve already calculated the **[CLICK]** z-score for your test statistic, you’re working with the standard normal distribution.

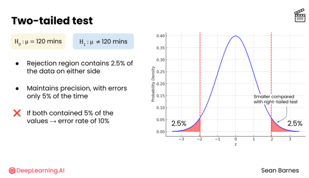
If you wanted to be **[CLICK]** 95% confident that you correctly reject the null hypothesis, you would **[CLICK]** look for test results above the mean that you expect to occur 5% of the time or less. **[CLICK]** This shaded region represents the values of z that should occur 5% of the time or less. It’s called the **[CLICK]** rejection region, because **[CLICK]** any test statistic that falls in this region would lead you to reject the null hypothesis. That conclusion may lead you to adjust the theater’s scheduling practices.
</td>
</tr>
<tr>
<td>


</td>
<td>
[CLICK] When performing a one-tailed hypothesis test with a 95% confidence level, your **[CLICK]** ability to reject the null hypothesis depends on the size of the rejection region, which has an **[CLICK]** area of 0.05. This value is called the **[CLICK]** level of significance, and is represented by the Greek letter alpha. An alpha of 0.05 is quite common, and is **[CLICK]** often used in initial studies for **[CLICK]** medical research, **[CLICK]** manufacturing quality control, and the **[CLICK]** social sciences.

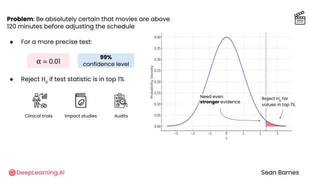
Notice that **[CLICK]** confidence is the complement of the **[CLICK]**level of significance. If you want to be 95% confident, you set your alpha to 0.05, representing a 5% chance of error.
</td>
</tr>
</table>

<table>
<tr>
<td>
</td>
<td>There's one final nuance you should be aware of. [CLICK] You just saw the rejection region for a right-tailed test. [CLICK] The process is quite similar for a left-tailed test which also has just one rejection region.

However, [CLICK] for a two-tailed test, you're interested in values above and below the mean.</td>
</tr>
<tr>
<td>

https://i.imgur.com/cPcYRGK.png

</td>
<td>For movie lengths, your [CLICK] null hypothesis would remain the same, but this time your alternative hypothesis [CLICK] H sub 1 would be that mu =/= 120.

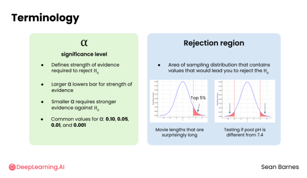
[CLICK] Here's what that would look like on the distribution. Same distribution, different hypothesis and therefore different rejection region.

Take a look at the upper rejection region for a moment. [CLICK] It's smaller for the two-tailed test compared with just the right-tailed test! In fact, it's half as small. [CLICK] Your rejection region contains 2.5% of the data on either side – [CLICK] upper and [CLICK] lower, totaling 5%. That's because you want to [CLICK] maintain the same precision, with errors only 5% of the time, but you have two rejection regions. [CLICK] If they both contained 5% of the values, that would actually lead to an error rate of 10%, not 5%.</td>
</tr>
<tr>
<td>

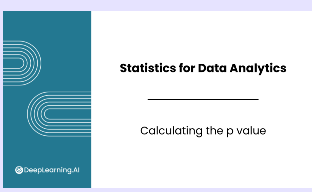
https://i.imgur.com/K1B7P4n.png</td>
<td>Imagine that the theater's schedules are expensive to adjust. Choosing to show fewer movies in a day may reduce revenue and cause employee schedule changes. You want to [CLICK] be absolutely certain that movies on average are indeed above 120 minutes before adjusting the schedule. In that case, would you want a higher or lower level of significance? [pause for thought]

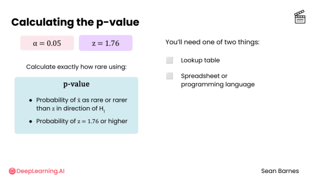
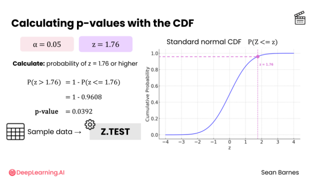
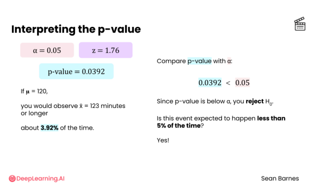
[CLICK] For a more precise test, you can [CLICK] lower your alpha to 0.01, which corresponds to having [CLICK] 99% confidence. So, you would [CLICK] reject the null hypothesis if the test statistic is in the top 1% of all means in this distribution. An alpha of 0.01 is often used in [CLICK] clinical trials, [CLICK] environmental impact studies, and [CLICK] financial audits, when the risk caused by incorrectly rejecting the null hypothesis is higher.

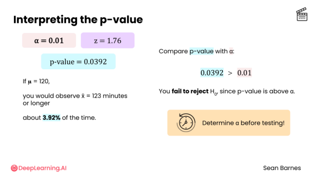
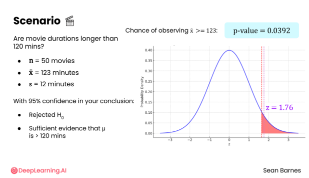
In this case, you start with the [CLICK] same distribution. Do you expect this rejection region for an alpha of 0.01 to be smaller or larger than alpha of 0.05? [pause for thought] Your rejection region gets [CLICK] smaller. Here's what it looks like. You want to be more confident, so you'll only [CLICK] reject the null for values in the top 1% of the distribution. In the movie theater scenario, this smaller rejection region means you'd [CLICK] need even stronger evidence that movies are longer before changing the scheduling.</td>
</tr>
</table>

 	Okay, that was a lot of info. To recap the terminology you learned, ● **[CLICK] Alpha**, your **[CLICK] significance level**, helps **[CLICK]** define the strength of evidence required in order for you to reject the null hypothesis. **[CLICK]** Larger alphas make it easier to reject the null with less evidence, whereas **[CLICK]** smaller alphas require stronger evidence against the null to reject it. **[CLICK]** Common values for alpha include 0.10, 0.05, 0.01. ● **[CLICK]** The **rejection region** is the area of the sampling distribution that contains the improbable values that would lead you to reject the null hypothesis. **[CLICK]** For alpha = 0.05, you saw this rejection region was the area under the curve in the **[CLICK]** top 5% of the distribution. In the movie theater example, the rejection region represents the range of average **[CLICK]** movie lengths that are so surprisingly long, you'd conclude movies are indeed longer than 120 minutes on average. ● You also saw that you can have a **[CLICK]** two-tailed test with **[CLICK]** two rejection regions, for example when **[CLICK]** testing whether a pool's pH is significantly higher or lower than 7.4. For alpha = 0.05, this test will have two rejection regions, each covering 2.5% of the distribution in either tail.
🎙️ TH	Alpha, your significance level helps you quantify the amount of uncertainty you're comfortable with. It's used to determine whether your test statistic is rare enough for you to reject the null hypothesis. Follow me to the next video to see how to calculate that rarity using a p value, which you can then compare with alpha to perform your hypothesis test.

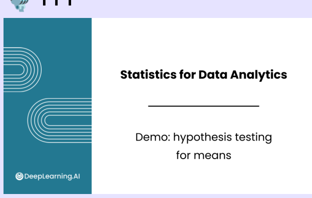
L1V7 – Calculating the p value

Visual	Script
🎙️ TH  	Your next step is to figure out whether your sample mean is rare enough for you to reject your null hypothesis. This is your final step before interpretation.
https://i.imgur.com/u7CqpIq.png  	Suppose the movie theater decided to select an alpha of 0.05. Earlier, you calculated your **[CLICK]** test statistic z = 1.76. **[CLICK]** How rare is this value? **[CLICK]** Let's visualize this idea by seeing if this value falls within your rejection region. You can see that z **[CLICK] does fall** within the **[CLICK]** rejection region, slightly above the boundary. So **[CLICK]** the sample mean of 123 minutes

	would happen less than 5% of the time if the true mean was in fact 120 minutes.
	You can now **[CLICK]** calculate exactly how rare this value is, using a **[CLICK]** p-value, short for probability value. The p-value represents **[CLICK]** the probability of getting a sample mean as rare or rarer than z in the direction of your alternative hypothesis. So in other words, **[CLICK]** the probability of getting z of 1.76 or higher. For this step, **[CLICK]** you'll need one of two things. Either **[CLICK]** a lookup table, or **[CLICK]** a spreadsheet or programming language that can calculate it for you.
	Let's talk through how to calculate the p value for our movie example. Again, you need to **[CLICK]** calculate the probability of observing a z-score of 1.76 or higher on the standard normal distribution. Recall the **[CLICK]** cumulative distribution function, or CDF, which represents the **[CLICK]** probability of observing a z-score less than or equal to a specific value. That means you can ~~easily~~ find the **[CLICK]** probability of getting a z score of **less than or equal to** 1.76 using the CDF. So, how might you use the CDF to calculate the probability that you're interested in? **[pause for thought]** You need to use the complement rule. The probability of z > 1.76 **[CLICK]** equals 1 minus the probability that z <= 1.76, which is **[CLICK]** what the CDF tells you. If you've calculated this z score manually, you can use a spreadsheet function to calculate the p-value. In this case, the probability that z <= 1.76 is about **[CLICK]** 0.9608. The complement of this probability is then 1 minus this value, or **[CLICK]** 0.0392. That gives you the **[CLICK]** p-value, the probability of observing a more extreme test statistic than the one you observed. By the way, you won't typically need to do this step in a spreadsheet. When you're working directly with your **[CLICK]** sample data, as opposed to calculating the z score yourself, you'll use the **[CLICK]** Z.TEST function to do all the steps you previously saw for you – except defining your hypothesis.
	Since p = 0.0392, if the **[CLICK]** true average duration of all movies is 120 minutes, you would observe **[CLICK]** a sample mean of 123 minutes or longer about **[CLICK]** 3.92% of the time. What do you think, is this rare enough to reject the null hypothesis? **[pause for thought]** **[CLICK]** Compare your p value with your significance level, in this case **[CLICK]** 0.05. Since your **[CLICK]** p value is below your significance level, you will reject the null hypothesis. Intuitively what you're doing with this comparison is asking whether **[CLICK]**this event is expected to happen less than 5% of the time. And **[CLICK]** that's the case here!

	Now, consider the case you saw in the previous video where movie schedules are difficult to adjust, so what if you instead required a **[CLICK]** 1% level of significance for your hypothesis test. Is this event rare enough to reject the null hypothesis when alpha equals 0.01? **[pause for thought]** In this case, **[CLICK]** you would not reject the null hypothesis **[CLICK]** because the p-value – which hasn't changed – is greater than alpha. You would not have observed strong enough evidence that the mean movie length is greater than 120 minutes. It's crucial that you **[CLICK]** determine your level of significance BEFORE you conduct your test. Avoid adjusting it in order to make the decision you want to make. That adjustment would introduce analysis bias!
	Let's step back and put everything together. You had a business question about **[CLICK]** whether movie durations were on average longer than 120 minutes. You collected a sample of **[CLICK]** 50 movies and found their **[CLICK]** mean duration was 123 minutes, with a **[CLICK]** standard deviation of 12 minutes. You wanted to know, is there sufficient evidence that movie durations are above 120 minutes? Then, via the **[CLICK]** test statistic, you calculated the chance of observing a sample mean of 123 minutes or above if the true mean is actually 120 minutes. Based on your **[CLICK]** p value, you found that you would **[CLICK]** observe a value as extreme as 123 minutes or more extreme than that about 3.92% of the time. **[CLICK]** Since you wanted to be 95% confident in your conclusion, **[CLICK]** you rejected the null hypothesis, concluding that there was **[CLICK]** sufficient evidence to believe that the true mean is above 120 minutes.
🎙️ TH	Great work calculating and interpreting your first p value. There's a lot going on here, and I don't expect you to remember everything. Follow me to the next video to see this process play out in a spreadsheet, which will help you develop your intuition for how hypothesis testing works in practice.

L1V8 – Demo: hypothesis testing for means

Visual	Script
🎙️ TH 	Now it's time to take it over to a spreadsheet to perform some hypothesis tests yourself using real data. You'll start to see a lot of parallels between confidence intervals and hypothesis tests. Let's get started with the forest fires dataset you've worked with previously.
🔗 [Start] 🔗 [Solution]	As you can imagine, the bigger the fire size, the more resources local agencies will need to deploy to put them out. If the park has a very large average fire size, the agency will need to call in the national fire service as well as allocate

aircraft in case of emergencies.

In order to plan accordingly, the agencies want to know if the mean burned area is larger than 10ha. If so, they'll start working with the national fire service and begin purchasing aircraft.

What should the hypothesis be? **[pause for thought]**

The null hypothesis, H naught, should be that the mean area is 10ha, while the alternative hypothesis, H sub 1, should be that the mean is greater than 10ha.

You'll start by determining the level of significance you're comfortable with. A good starting point is alpha equals 0.05, which has an expected 5% error rate with respect to incorrectly rejecting the null hypothesis.

Then, you'll find the test statistic. Remember that you need the sample mean, x bar, the sample standard deviation, s, and the sample size, n. These are the same statistics you used to compute a confidence interval

- In C2: =AVERAGE(M:M)
- In D2: =STDEV(M:M)
- In E2: =COUNT(M:M)

Now find the test statistic. Remember you need to subtract the hypothesized mean from the sample mean, and then divide that by s divided by square root of n

- In D1: test statistic
- In D2: = (A2 - 1) / (B2 / SQRT(C2))

Next, you need to find the p-value associated with the test statistic. Since this is a right tailed test, you want the probability of seeing a value to the right of your test statistic.

You can do this using the NORM.S.DIST function, the CDF. Remember that the CDF corresponds to the probability that z is less than or equal to some value, but you want the complement: the probability that the test statistic is greater than the value.

- In E2: 1 - NORM.S.DIST(D2)

This gives a value of **[value]**, what would you conclude? **[pause for thought]**

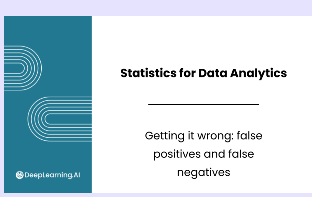
For a 5% significance level, you would fail to reject the null hypothesis and conclude that you do not need to increase your level of resources.

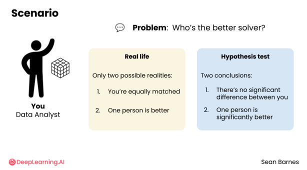
Now, let's move to the ISI feature, or Initial Spread Index. Initial spread index measures how quickly the fire area expands after it starts. A value of 10 indicates a high rate of spread, while a rating of 16 or more indicates extremely rapid rate of spread. The parks department asks you to verify that

	the mean of the ISI stays below 10. If the rate of spread was above 10, they would be required to start a few procedures to help control the rate of spread overall, such as initiating controlled burns or preemptively closing certain trails or campsites. These measures are both very expensive, so the parks department asks you to be extremely confident in your conclusions. You select an alpha of 0.01. What should your hypotheses be for this scenario? **[pause for thought]** Start with your null hypothesis. You're testing whether the ISI is different from 10, so ISI = 10 should be your null hypothesis. You're looking for evidence that ISI is less than 10, so your alternative hypothesis is that the mean ISI is less than 10. You'll need the same statistics to conduct your test: ● The sample mean: In A3: =AVERAGE(Data!H:H) ● The sample standard deviation: In B3: =STDEV(Data!H:H) ● The sample size: In C3: =COUNT(Data!H:H) Calculate the test statistic in the same way: ● In D3: =(A3 - 10) / (B3 /SQRT(C3)) The next step is finding the p-value. In this case, since you have a left-tailed test, you want the probability that a z score is smaller than or equal to the test statistic. You can calculate that probability directly using the CDF: ● In E3: =NORM.S.DIST(D3) You can safely say that you can reject the null hypothesis even at a 1% significance level. One more test. Extreme temperature conditions make firefighting more challenging. If the mean temperature is significantly different from 18 degrees celsius, the parks department will start phasing out older firefighting gear for newer gear with temperature control features. They ask you to check if the mean temperature differs either above or below from 18 degrees Celsius, which is equivalent to about 64 degrees Fahrenheit. They're planning to phase out the equipment anyway, but more extreme temperature conditions would cause them to complete the phase out sooner. So the department is fine with a lower rigor test. What should your hypotheses be? **[pause for thought]** Since you want to see if mean temperature deviates, in any direction, from 18 degrees, then you want to perform a two tailed test, where H naught is mu = 18, and H sub 1 is mu =/= 18 Because of the lower rigor test, you can use a significance level of 0.1.

	First part of the job is exactly the same: Get the feature’s statistics: ● In A4: =AVERAGE(Data!H:H) ● In B4: =STDEV(Data!H:H) ● In C4: =COUNT(Data!H:H) And then calculate the test statistic ● In D4: =(A4 - 18) / (B4 /SQRT(C4)) Then, you need the p-value. Since this is a two-tailed test, the rejection region includes both tails of the distribution, right and left, from the test statistic. In the two-tailed approach, you’re adding the probability that a z score is higher than 3.48 in the positive direction, and lower than -3.48 in the negative direction. Since the distribution is symmetric, you can calculate one of these probabilities and multiply the result by two: ● =2*NORM.S.DIST(-D4) You get a p-value of 0.0004. What can you conclude? The p-value is smaller than 0.1, you can reject the null hypothesis at a 10% significance level. It might seem counterintuitive to perform a two-tailed test when you see that the test statistic is positive. In general, it's important to define your hypotheses in advance based on the question you’re trying to answer. However, if you realize that the answer you’re really looking for is in one direction – say, you’re only interested in seeing if the temperatures are very high – you can reformulate your hypotheses and conduct a different test to get a more applicable result.
🎙️ TH	Great work conducting those hypothesis tests! You’re prepared to determine statistical significance on real world data. Coming up, you’ll see the two possible errors you will encounter when conducting hypothesis tests. I'll see you in the next video.

L1V9 – Hypothesis testing errors

Visual	**Script**
🎙️ TH 	Inferential statistics is all about managing uncertainty, and part of managing uncertainty is understanding how things can go wrong. Hypothesis testing can be wrong in two key ways: you found an effect, but in reality there isn’t one; or, you didn’t find an effect, but in reality there is one. Let’s take a look.

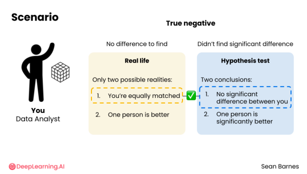


Your hypothesis test uses a sample as a window into the true state of the world. Think back to the example of competing with your friend in solving Rubik's cubes, where they challenged you to see **[CLICK]** who's the better solver. **[CLICK]** In real life, there are **[CLICK]** only two possible realities:

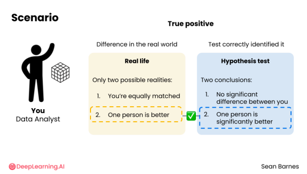
1. **[CLICK]** Either you're equally matched
2. **[CLICK]** Or one person is better

And if you conducted a **[CLICK]** hypothesis test to answer that question, you'd come to one of **[CLICK]** two conclusions:

1. **[CLICK]** There's no significant difference between the two of you
2. **[CLICK]** One person is significantly better

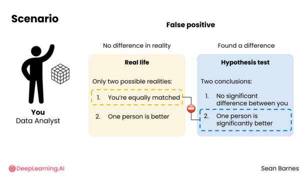


If you're **[CLICK]** equally matched, and **[CLICK]** your test didn't find a significant difference, that's great! Your window into the truth saw the correct effect. This situation is called a **[CLICK]** true negative. You **[CLICK]** didn't find a significant difference, and there **[CLICK]** wasn't one to find.

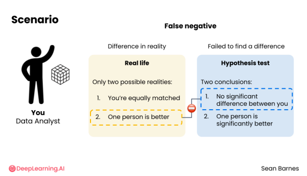


If you're **[CLICK]** not equally matched, and your test found **[CLICK]** one of you to be better, that's great too! This situation is called a **[CLICK]** true positive. There was a **[CLICK]** difference in the real world, and your **[CLICK]** test correctly identified it.

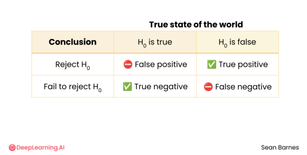
You can see there are two ways to go wrong, though.



It's possible that **[CLICK]** you're equally matched, but your test concluded that **[CLICK]** one of you is better. **[CLICK]** That's a false positive. You **[CLICK]** found a difference, but in reality **[CLICK]** there isn't one.



And it's also possible that your test concluded that **[CLICK]** there's no significant difference, yet **[CLICK]** one of you is in fact better. That's a **[CLICK]** false negative. **[CLICK]** There's a true effect, but you **[CLICK]** failed to find it.



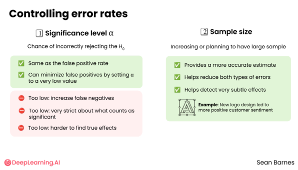
Let's generalize these errors beyond the Rubik's cube example. As you saw in the previous videos, your hypothesis test has **[CLICK]** two possible conclusions: either you **[CLICK]** reject the null hypothesis, or **[CLICK]** you fail to reject the null hypothesis.

Then there's the true state of the world:

1. Either the **[CLICK]** null hypothesis is true
2. Or the **[CLICK]** null hypothesis is false.

~~So you can see~~ if you reject the null hypothesis, but the null hypothesis is true, that's a **[CLICK]** false positive. And conversely, not rejecting the null hypothesis when the null hypothesis is false leads to a **[CLICK]** false negative.

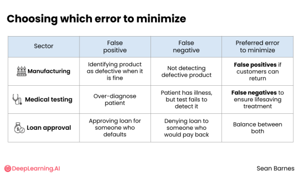
Meanwhile, if you reject the null hypothesis and the null hypothesis is false, that's a **[CLICK]** true positive. And if you fail to reject the null hypothesis and the null hypothesis is true, that's a **[CLICK]** true negative.



You have a few mechanisms for controlling these error rates. One of them is your **[CLICK]** significance level alpha. Your significance level represents the **[CLICK]** chance of incorrectly rejecting the null hypothesis. It's the risk you must take! Based on what you know, does alpha correspond to the rate of false positives or false negatives? **[pause for thought]** It's **[CLICK]** the same as the false positive rate. You can try to **[CLICK]** minimize false positives by setting alpha to a very low value, but that will **[CLICK]** increase the likelihood of false negatives. Setting alpha too low means you're **[CLICK]** being very strict about what counts as significant. This makes it **[CLICK]** harder to find true effects.

[CLICK] Increasing your sample size, or **[CLICK]** planning to have a sufficiently large sample, is an important factor in the success of your hypothesis test.

- A large sample **[CLICK]** provides a more precise estimate and **[CLICK]** helps reduce both types of errors
- A large sample can also **[CLICK]** help you detect a very subtle effect. For example, it's possible that **[CLICK]** a new logo design has led to more positive customer sentiment towards your company. However, that difference may be very small and difficult to detect.



Choosing which error to minimize

Sector	False positive	False negative	Preferred error to minimize
Manufacturing	Identifying product as defective when it is fine	Not detecting defective product	**False positives** if customers can return
Medical testing	Over-diagnose patient	Patient has illness, but test fails to detect it	**False negatives** to ensure lifesaving treatment
Loan approval	Approving loan for someone who defaults	Denying loan to someone who would pay back	Balance between both

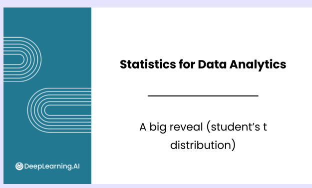
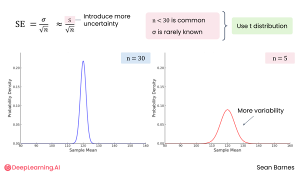
DeepLearning.AI Sean Barnes

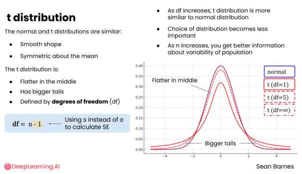
Whether you want to minimize false positives or false negatives depends on the costs and risks associated with each outcome.

- For example, if you're working in **[CLICK]** manufacturing, you'll likely test for defective products before they're shipped. A false positive would mean **[CLICK]** identifying a product as defective when it's actually fine, whereas a false negative would be **[CLICK]** not detecting a defective product. You may prefer to **[CLICK]** minimize false positives if customers can easily make returns, to avoid wasting product that's actually fine.
- In **[CLICK]** medical testing for serious illnesses, however, it's typically more desirable to **[CLICK]** minimize false negatives. A false negative happens when a **[CLICK]** patient actually has a particular illness, but the test fails to detect the illness. While diagnosis can be stressful for patients, it's generally more desirable **[CLICK]** to over-diagnose, since the diagnosis can be corrected with further testing. Meanwhile, a patient whose illness is not detected will not receive potentially lifesaving treatment.
- Most situations call for a **[CLICK]** balance between the two types of errors. For example, if you're working at a bank that **[CLICK]** approves loan applications, a false positive might mean **[CLICK]** approving a loan for someone who actually defaults, while a false negative might mean **[CLICK]** denying a loan to someone who would pay it back in full. The

	bank will likely conduct a significant risk assessment to minimize losses from false positives while maximizing the number of profitable loans.
🎙️ TH	Balancing error types is just another compromise you'll have to make as a data analyst working with uncertainty. There's nothing you can really do to achieve 100% accuracy. Now, you're almost to the end of this lesson! There's just one more nuance to hypothesis testing. Follow me to the next video to learn more.

L1V10 – The t distribution

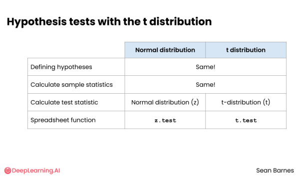
Visual	**Script**
🎙️ TH 	There's one additional element of complexity here, which is that in some cases, the sampling distribution of the sample mean isn't normal, and you'll have to use a different distribution to conduct your hypothesis test.
	Recall the conditions for the Central Limit Theorem, which states that when you are calculating the sample mean based on a **[click]** large sample, typically 30 or more observations, the sampling distribution of the sample mean is **[click]** normally distributed. So what happens if you calculate a sample mean based on **[click]** less than 30 observations – do you think the variability in the sampling distribution of the mean will be larger or smaller? **[pause for thought]** The sample mean for smaller samples is likely to have **[click]** more variability than for larger samples. This condition introduces more uncertainty into your estimate. You also saw that the **[click]** standard error for the sampling distribution for the mean was equal to the population standard deviation (sigma) divided by the square root of the sample size (n). When you use the **[click]** sample standard deviation as an estimate for sigma, you **[click]** introduce even more uncertainty. **[click]** Small sample sizes are common in **many** fields, including healthcare. And, the **[click]** population standard deviation is rarely known. When **[click]** either or both of these conditions are true – a sample size smaller than 30, or an unknown sigma – there is a more appropriate choice for the sampling distribution for the mean, which is known as the **[click]** t-distribution.

<table>
<tr>
<td>


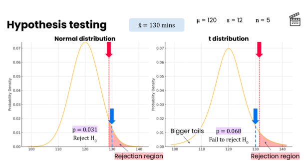
</td>
<td>
The t distribution is very similar to the standard normal distribution. They both have a **[CLICK]** smooth shape and are **[CLICK]** symmetric about a mean of zero. However, **[CLICK]** the t distribution is **[CLICK]** flatter in the middle and **[CLICK]** has bigger tails. Intuitively what that means is that you're more likely to observe values further from the mean. This observation is consistent with the idea that the t distribution reflects more uncertainty about the underlying data.

However, one difference is the t distribution is **[CLICK]** defined by a parameter known as the degrees of freedom (df), which is a statistical measure that tries to quantify the uncertainty in your estimate introduced by estimating population parameters from a limited sample. It's a bit of an abstract concept. But for this scenario, the degrees of freedom is calculated as the **[CLICK]** sample size minus 1. The **[CLICK]** minus one comes from **[CLICK]** using the sample standard deviation instead of the population standard deviation to construct the distribution. Because you're using an estimate, your calculation introduces more uncertainty.

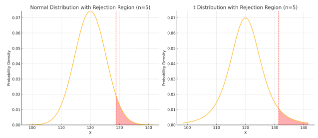
[CLICK] As degrees of freedom increases, the t distribution becomes **[CLICK]** more and **[CLICK]** more similar to the normal distribution, and therefore your **[CLICK]** choice of which distribution to use becomes less important. This convergence on the normal distribution reflects the increased certainty you can have in your conclusions with a larger and larger sample size.

~~The fundamental idea of the t distribution is that, **[CLICK]** in order to calculate a confidence interval or perform a hypothesis test, you generally don't know the population standard deviation.~~ **[CLICK]** As your sample size increases, you get better information about the variability of the population, but with small samples you're working with a less precise estimate.
</td>
</tr>
<tr>
<td>


</td>
<td>
You'll follow a similar process for conducting your hypothesis test as you just learned. You will define your hypotheses the **[CLICK]** same way, and **[CLICK]** collect the sample statistics needed in order to calculate your test statistic. However, you'll use the **[CLICK]** t distribution rather than the normal distribution to figure out whether your sample statistic is sufficiently rare for you to reject the null. Your test statistic will be called t rather than z, and you'll use the **[CLICK]** t.test function rather than the z.test function.

The rejection region, and thus the p-value, are also different, since you're now using the t-distribution to define them.
</td>
</tr>
<tr>
<td>


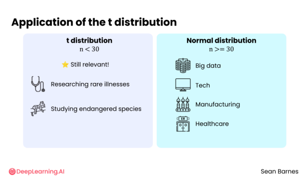
</td>
<td>
Consider the movie duration example, with a hypothesized population mean of 120 and a sample standard deviation of 12 over just 5 samples. **[click]** Take a look at the difference between the normal and t-distributions. You can see the t distribution has **[click]** bigger tails. The **[click]** rejection region for alpha = 0.05 is also different. Notice that the rejection region for the normal distribution starts around **[click]** 128 minutes, while for the **t distribution**, it
</td>
</tr>
</table>

<table>
<tr>
<td>


https://i.imgur.com/svezrJm.png
</td>
<td>
starts around **[click]** 132 minutes. That means that the t distribution requires stronger evidence to reject the null hypothesis. It's more strict.

For example, say the sample mean you observed for the 5 movies was **[click]** 130 minutes.

In the normal distribution, 130 minutes falls [click] here, and the [click] p value is 0.031, meaning you would [click] reject the null hypothesis, since 0.031 is less than your alpha of 0.05.
In the t distribution, 130 falls here. The p value is 0.068. In that case, you would fail to reject the null hypothesis.

So you can see that for small sample sizes, choosing the t distribution over the normal distribution is impactful. It applies a bit more rigor for small sample sizes.
</td>
</tr>
<tr>
<td>


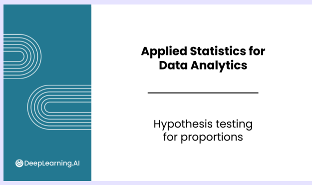
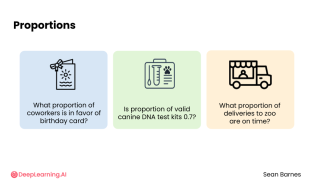
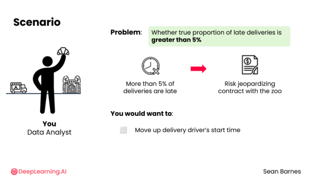
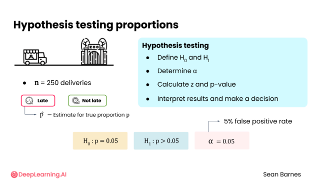
</td>
<td>
When the **[click]** **[click]** t distribution was initially developed, it was much more common to be dealing with **[click]** small sample sizes of say 10 or 15, and to try to draw conclusions about the population from that small group. **[click]** Nowadays, you're more likely to be working with **[click]** larger sample sizes above 30, the general rule of thumb for a "large" sample that will follow the central limit theorem. You learned in Data Analytics Foundations that you'll often be working with **[click]** big data. In **[click]** tech, if you want to survey users, you might be able to get thousands of people. In **[click]** manufacturing, your systems might record the production time for every product. If you're working with **[click]** emergency room data, you might have access to thousands of visits rather than a dozen.

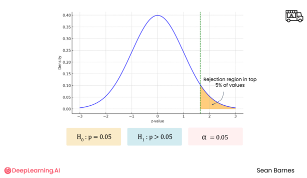
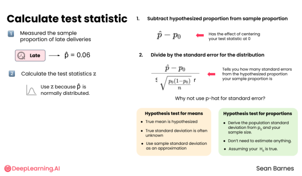
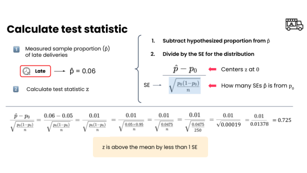
The t distribution is **[click]** still relevant in many cases, such as in **[click]** researching rare illnesses or **[click]** studying endangered species, when you're working with small samples. When your sample sizes are above around 30, the difference is not likely to impact your decision making.
</td>
</tr>
<tr>
<td>🎙️ TH</td>
<td>
That's a wrap on hypothesis testing! You've learned a ton, from how to formulate your hypotheses to identifying your rejection region, calculating your test statistic, and interpreting your p value.

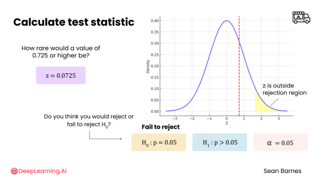
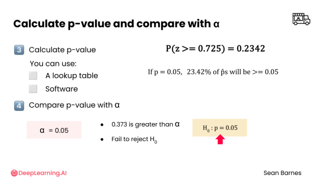
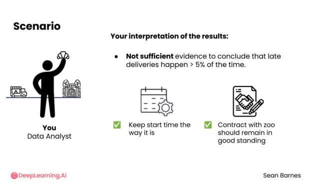
Next up, you'll complete the practice assessment and practice lab, which involves examining human sleep patterns. When you've finished the practice assessment and lab, join me in the next lesson to learn about the wide variety of hypothesis tests you can employ in different business cases. I'll see you there!
</td>
</tr>
</table>

Lesson 2 – Other hypothesis tests

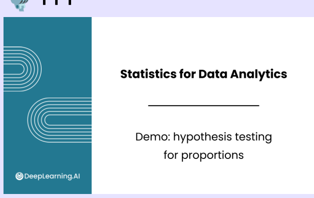
L2V1 – Hypothesis testing for proportions

Visual	Script
🎙️ TH  	Your knowledge about how to conduct hypothesis tests for means also applies to many other kinds of hypothesis tests! For example, you can test a sample proportion, not just the mean; whether two samples have the same mean; and lots more. Let's start with hypothesis testing for proportions.
 	You've seen examples of business questions related to proportions: ● **[CLICK]** What proportion of your coworkers is in favor of the birthday card policy? ● **[CLICK]** Is the proportion of valid canine DNA test kits really 0.7? ● **[CLICK]** What proportion of our bakery deliveries to the zoo are on time?
 	Let's focus on delivery times for this example. Recall the scenario where you were working with a bakery delivering pastries to the local zoo. Say the bakery has just secured a large contract with the zoo to deliver fresh pastries 5 times per day instead of once. According to the contract terms, if **[CLICK]** more than 5% of deliveries are late, the bakery risks **[CLICK]** jeopardizing their contract with the zoo. For that reason, **[CLICK]** you would want to **[CLICK]** move up your delivery driver's start time to ensure they can make it to the zoo. So you're interested in the question of **[CLICK]** whether the true proportion of late deliveries is greater than 5%.
 https://i.imgur.com/vYQq23T.png	You can follow the process for hypothesis testing you learned in the previous videos: 1. **[CLICK]** Define your hypotheses 2. **[CLICK]** Determine the significance level 3. **[CLICK]** Calculate the test statistic and the p value 4. **[CLICK]** Interpret the results and make a decision Suppose you collect a sample of **[CLICK]** 250 deliveries and record if they were **[CLICK]** late or **[CLICK]** not. The proportion of late deliveries in your sample is called **[CLICK]** p-hat, which is an **[CLICK]** estimate for the true proportion of late deliveries p. You can define your hypotheses as ● **[CLICK]** H naught (H_0) that p = 0.05, which corresponds to the status quo, for which the proportion of late deliveries is 0.05. If this

	hypothesis holds, you would not change your start time because you are confident that your deliveries will consistently arrive on time. • **[CLICK]** H sub 1 (H_1) is that $p > 0.05$, which means that the proportion of late deliveries is higher than 0.05 You feel this decision has moderately high stakes, so you decide to select an **[CLICK]** alpha of 0.05, which is associated with a **[CLICK]** 5% false positive rate.
	The sampling distribution for the sample proportion is the normal distribution, shown here. So you would **[CLICK]** reject the null hypothesis if your test statistic falls within **[CLICK]** this rejection region in the top 5% of values. You could also choose the t distribution, but considering your large sample size, the normal distribution is appropriate.
	Let's imagine that you **[CLICK]** measured the sample proportion of late deliveries as **[CLICK]** 0.06. It seems higher than 0.05, but is that difference significant? You can use a hypothesis test for proportions to answer your question. Your next step is to **[CLICK]** calculate the test statistic z. It's quite similar to the one for means. **[CLICK]** First, subtract your hypothesized proportion **[CLICK]** p naught (p_0) from your sample proportion p-hat ($\hat{p}$). This operation has the effect of **[CLICK]** centering your test statistic at 0. Then you'll **[CLICK]** divide by the standard error for the distribution, which essentially tells you **[CLICK]** how many standard errors from the hypothesized proportion your sample proportion is. The standard error looks like the **[CLICK]** square root of [p naught (p_0) times 1 minus p zero (p_0) divided by n]. **[CLICK]** Why not use p-hat to calculate the standard error? In a hypothesis test for **[CLICK]** means, you **[CLICK]** hypothesize the true mean, but you often **[CLICK]** don't know the true standard deviation, so you must **[CLICK]** use your sample standard deviation to approximate it. However, in the hypothesis test for **[CLICK]** proportions, you can **[CLICK]** derive the population standard deviation just from p naught and your sample size. You **[CLICK]** don't have to estimate anything, since you're assuming your null hypothesis to be true.
	Working that out for this example, p hat minus p naught is **[CLICK]** 0.06 - 0.05 = **[CLICK]** 0.01. P naught times (1 minus p naught) is **[CLICK]** 0.05 times 0.95, which is **[CLICK]** 0.0475. This is on the lower end of variability, since you expect most deliveries to be on time. **[CLICK]** Divide 0.0475 by 250, your sample size and you get **[CLICK]** 0.00019. Now take the square root to get **[CLICK]** 0.01378. So your test statistic **[CLICK]** z is 0.725. It corresponds to the number of standard errors away from the

	mean. So in your sampling distribution of the proportion, assuming a mean of 0.05, **[CLICK]** this value is above the mean by less than one standard error.
	How rare would a value of 0.0725 or higher be? Take a look at **[CLICK]** z graphed on the standard normal distribution. Do you think you would reject the null hypothesis in this case or fail to reject it? **[pause for thought]** It looks like you would **[CLICK]** fail to reject the null hypothesis! **[CLICK]** z is not inside the rejection region.
	You should then **[CLICK]**calculate your p value and see if it matches your intution. **[CLICK]** You can use **[CLICK]** a lookup table or **[CLICK]** software to do so. **[CLICK]** The **[CLICK]** probability of getting a test statistic z of 0.725 or more extreme, in this case higher, is 0.2342. So, **[CLICK]** if the true proportion of late deliveries is 0.05, **[CLICK]** 23.42% of sample proportions will be 0.06 or higher. Next, you'll **[CLICK]** compare your p value with alpha of 0.05. **[CLICK]** 0.2342 is greater than alpha, so you **[CLICK]** fail to reject the null hypothesis.
	Your interpretation of this result is that **[CLICK]** there is **not** sufficient evidence to conclude that late deliveries happen more than 5% of the time. So, **[CLICK]** you can keep your delivery driver's start time the way it is, and **[CLICK]** it looks like your contract with the zoo should remain in good standing!
🎙️ TH	Great work on testing proportions! You'll choose a hypothesis test for means or proportions based on the business question you're interested in. Follow me to the next video to see how to conduct a hypothesis test for proportions in a spreadsheet.

L2V2 – Demo: Hypothesis testing for proportions

Visual	**Script**
🎙️ TH 	Let's put what you just learned in action! You'll work with the forest fires dataset to try to evaluate whether the proportion of small fires constitutes the majority of fires you observe in the data set.
🔗 Start 🔗 Solution	Suppose the Portuguese park service has been working to reduce the spread of fires by quickly containing them using a new method. After implementing the new containment procedures, they're interested in investigating whether the proportion of **very small** fires now constitutes the majority of **all** fires. If so, they've met their improvement goals and can be confident that the new measures are working well. This version of the dataset includes the `is_small` column, which takes on the value 1 if the burned area was below 0.5ha, and 0 if it was greater. Recall from

<table>
<tr>
<td></td>
<td>
an earlier lab in this course that the proportion of very small fires and other fires seems pretty balanced.

Select is_small → Insert Chart → Bar chart → X axis: is_small → Aggregate → Count

Your goal is to perform a hypothesis test to see if the proportion is in fact greater than 0.5. What should the hypotheses be? [pause for thought]
You can perform a right tailed test, where the null hypothesis is $p=0.5$, and the alternative hypothesis is $p > 0.5$.
You can start with the default significance level of 0.05.
Let’s get on to it. For the proportions test, just as with the confidence interval, you will need

Now find each of these statistics

In A2: = AVERAGE(Data!N:N)
In B2: =1-A2
In C2: =COUNT(Data!N:N)

Next, as you did with the hypothesis test for the mean you need to find the test statistic. Remember that the test statistic for the proportion looks a little bit different than that for the mean. You still subtract the hypothesized population proportion from the sample proportion, but the standard error that you divide by is equal to the square root of p(1-p) / n:

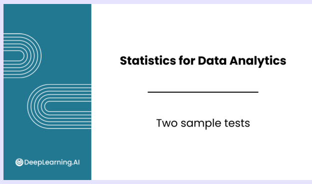
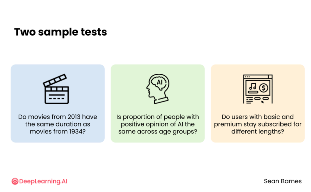
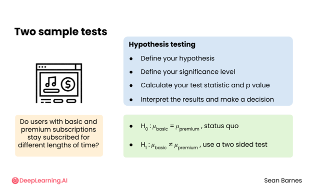
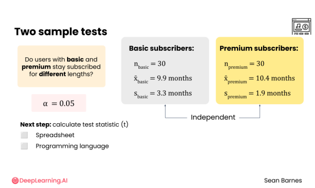
In D2: (A2 - 0.5)/SQRT(A2*B2/C2)

This test statistic follows a normal distribution, so you need to determine the probability of a more extreme value, given your sample data and the structure of your alternative hypothesis

In E2: =1-NORM.S.DIST(D2)

Given this p-value is greater than any reasonable level of significance, you don’t have enough evidence to conclude the true proportion is greater than 0.5. The park service will need to continue their containment measures and gather more data to be certain that the new measures are effective at getting the proportion of small fires above 0.5
</td>
</tr>
<tr>
<td>🎙️ TH</td>
<td>
It’s okay to fail to reject the null! Sometimes non-effects can be just as informative as finding one.
Coming up, you’ll learn how to conduct a two sample test, which compares two means or proportions. See you there!
</td>
</tr>
</table>

L2V3 – Two sample tests

Visual	Script
🎙️ TH 	Oftentimes, you're interested in comparing two samples directly, rather than comparing one sample to a hypothesized value. For example, say you want to compare the mean delivery time on weekends and weekdays.
	You've seen a few cases throughout this course where you might want to compare two samples directly: ● **[CLICK]** Do movies from 2013 have the same average duration as movies from 1934? ● **[CLICK]** Is the proportion of people who have a positive opinion of AI the same in two different age groups? ● **[CLICK]** Do users with basic and premium subscriptions stay subscribed for different lengths of time?
	Let's stick with the music subscription service. Suppose you want to determine whether users with basic and premium subscriptions stay subscribed for different lengths of time. You're not certain what these values are, you just want to see if they're different. You'll follow the same process as usual: ● **[CLICK]** Define your hypotheses ● **[CLICK]** Determine the significance level ● **[CLICK]** Calculate the test statistic and the p value ● **[CLICK]** Interpret the results and make a decision Here's how you'd set up your hypotheses. First, the null hypothesis: ● **[CLICK]** Your status quo is that there's no difference between the groups. So you'll write that as H naught is mu of basic = mu of premium ● Your alternative hypothesis is that **[CLICK]** these two means are different, so you can use a two-tailed test with the hypothesis H sub 1 is mu of basic =/= mu of premium. Or alternatively, if you had a hypothesis that the mean subscription length of one group was higher than the other, you could formulate a one-sided test. Say you're comfortable with an **[CLICK]** alpha of 0.05, with a 5% chance of false positives. Now you are ready to calculate the test statistic.
	Say you have two samples, one of basic subscribers and one of premium. Each sample has 30 subscribers, with the following descriptive stats: ● **[CLICK]** Sample mean of basic = 9.9 months ● **[CLICK]** Sample standard deviation of basic = 3.3 months ● **[CLICK]** Sample mean of premium = 10.4 months ● **[CLICK]** Sample standard deviation of premium = 1.9 months In this case, the two samples are **[CLICK]** independent, meaning that the

<table>
<tr>
<td></td>
<td>length of a basic subscription is independent of the length of a premium subscription. So you're going to take the approach of treating these samples as independent from each other.

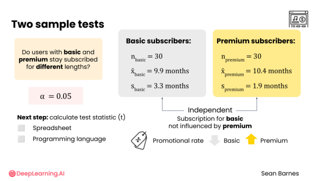
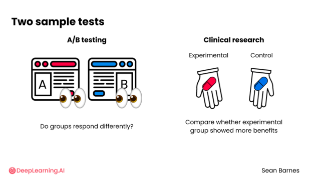
The **[CLICK]** next step would be to calculate your test statistic, which follows the t distribution. You know the theory behind how to conduct this test already, but the math is a bit more involved, both for calculating the standard error and for determining the degrees of freedom. In practice, you'll be using a **[CLICK]** spreadsheet or **[CLICK]** programming language to conduct this test, so let's see how to conduct this test in a spreadsheet</td>
</tr>
<tr>
<td>**[new demo segment here]**
🔗 Start here
🔗 Solution</td>
<td>

- Here's some generated data simulating 30 premium and 30 basic subscribers.
- This time you don't need to calculate any of the statistics like the mean, standard deviation and so on, because the spreadsheet function T.TEST will do that for you. The output of the T.TEST function is just the p value, so you won't need to do any of the intermediary steps to calculate the test statistic.
- T.TEST has several parameters. Remember you can use the help menu to review them.
  - In any cell type: =T.TEST(
    - Help menu
  - You have 4 parameters:
    - Range1: the sample data for the first category
      - Select the column for the T.TEST function
      - ,
    - Range 2: sample data for the second category
      - Select the column for the T.TEST function
      - ,
    - Tails: This tells you if you use 1 tail, or 2 tails of the t-distribution for the p-value. In this case, you want the two tails, because you proposed a two-tailed test (equal vs. different)
      - Type: 2, in the T.TEST function
    - Type: This one is very important because it sets some assumptions on your data.
      - Paired test (1). You'll learn more about paired tests in the next video, but they're good for before and after data, like if you're testing the same person's subscription length before and after getting premium
      - Two sample equal variance (2). This makes the big assumption that the variance is the same across both populations. You probably won't want to use that one often
      - Two sample unequal variance (3). Here you don't make any assumptions on the variance of the

</td>
</tr>
</table>

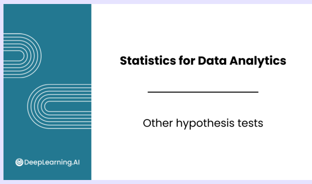
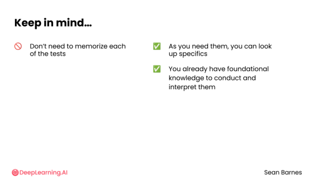
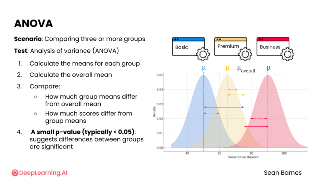
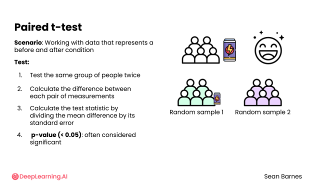
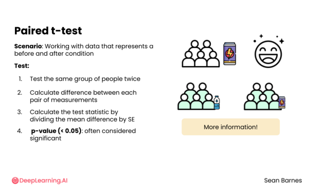
<table>
<tr>
<td></td>
<td>populations.
• Which one do you think is most appropriate here? [pause for thought] You want this last option
○ Type: 3, in the T.TEST function
○) + ENTER

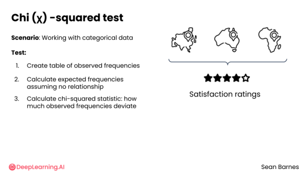
This t test gives you a p-value of 0.604. It’s pretty convenient that you don’t need to calculate any of the sample statistics or the test statistic. How would you interpret the p value? [pause for thought]

This p value is quite large, larger than your significance level of 0.05. So you fail to reject the null hypothesis, concluding that you don’t yet have enough evidence to support the idea that the two different groups of subscribers tend to stay subscribed for different lengths of time.</td>
</tr>
<tr>
<td>
</td>
<td>It’s important to remember this test assumes the samples are independent. It means that the [CLICK] subscription length for basic users is not influenced by the subscription length for premium users, and vice versa. In general it seems fair to conduct this test, but it’s possible that you could inadvertently introduce biases. For example, if your company offered a [CLICK] promotional rate to encourage basic subscribers to upgrade, that might [CLICK] shorten the duration for those subscribers while [CLICK] extending the duration for premium subscribers.</td>
</tr>
<tr>
<td>
</td>
<td>Two sample hypothesis tests are more commonly used in practice than one-sample tests, since you’re often interested in comparing two groups.
• In [CLICK]A/B testing, you create [CLICK]two different versions of your product and [CLICK]show those versions to different groups. Then, you calculate [CLICK] whether the groups responded to each version differently.
• In [CLICK] clinical research, you often have [CLICK] experimental and [CLICK] control groups. You give your experimental group the [CLICK] new treatment, such as a new drug, and give the control group a [CLICK] placebo. You then want to [CLICK] compare whether the experimental group showed more benefits than the control group.</td>
</tr>
<tr>
<td>🎙️ TH</td>
<td>So what happens if you’re interested in comparing more than two samples, like several age groups? Or pairs of samples like patients’ improvement before and after a treatment? There are many types of hypothesis tests beyond the ones you’ve seen so far. Follow me to the next video to learn more.</td>
</tr>
</table>

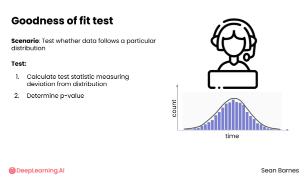
L2V4 – Other hypothesis tests

Visual	Script
🎙️ TH	You’ve seen several kinds of hypothesis tests already. There are so many options out there, and they each answer a different kind of question. You can’t

	learn them all in one day, but in this video you'll see what kinds of questions to answer with which test.
	Keep in mind that **[CLICK]** you don't need to memorize each of the tests in this video, because **[CLICK]** as you need them, you can look up their specifics. **[CLICK]** You already have the foundational knowledge needed to conduct and interpret them.
	Now, imagine you're not just comparing two groups, like basic and premium subscribers, but **[CLICK]** three or more groups. For example, you might have a tiered subscription model with **[CLICK]** basic, **[CLICK]** premium, and **[CLICK]** business. This comparison gets tricky with the tests you saw earlier because the more tests you perform, the more the chance for error accumulates. If you have a scenario like this, you'd want to use the **[CLICK]** analysis of variance test, also called ANOVA. This test involves **[CLICK]** calculating the means for each group plus the **[CLICK]** overall mean. Then, you **[CLICK]** compare **[CLICK]** how much the group means differ from the overall mean versus **[CLICK]** how much individual scores differ from their group means. The p-value in an ANOVA tells you how likely you'd see these differences between groups if there were no real effect. **[CLICK]** A small p-value (typically less than 0.05) suggests the differences between groups are significant.
	It's possible you'll find yourself **[CLICK]** working with data that represents a before and after condition, which can take advantage of a special test. Say you're interested in testing whether **[CLICK]** students' moods improve after **[CLICK]** drinking a particular energy drink. One option would be to take **[CLICK]** a random sample of students who had the energy drink, and **[CLICK]** another random sample of students who had just water, and compare their moods. But, you actually have a more powerful option.
	You can test the same group of people twice – have them **[CLICK]** drink water and assess their mood, then have them **[CLICK]** drink your energy drink and assess their mood again. In this case, you actually have **[CLICK]** more information about the strength of the effect, since you don't have to account for all the possible variability between people. You can perform a paired t test. **[CLICK]** You calculate the difference between each pair of measurements (after minus before). Then, you calculate the test statistic by dividing the mean difference by its standard error.

<table>
<tr><td></td><td>The p-value here indicates how likely you'd see a difference this large by chance if the treatment had no effect. Again, a [CLICK] p-value less than 0.05 is often considered significant.</td></tr>
<tr><td>
</td><td>You may also find yourself working with categorical data. The previous hypothesis tests you’ve learned have all operated on numerical data. Suppose you want to [CLICK] determine if customer satisfaction ratings differ by region.

To test this hypothesis, you can use a chi-squared test. Ah, another good old Greek letter. The procedure involves [CLICK] creating a table of observed frequencies. Then ,[CLICK] you calculate the expected frequencies assuming no relationship. [CLICK] The chi-squared statistic measures how much your observed frequencies deviate from these expected frequencies.

[CLICK] A small p-value suggests that the observed frequencies differ significantly from what you would expect if there is no relationship.</td></tr>
<tr><td>
</td><td>A lot of statistical methods assume that your data follows a normal distribution, or if you know it follows that distribution, then you can often get away with a smaller sample size. For that reason, you may be interested in [CLICK] testing whether that’s actually true. You can use a goodness of fit test.

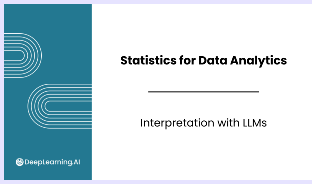
Say you want to know if [CLICK] customer service times at a call center follow a normal distribution. In a goodness of fit test, [CLICK] you’ll calculate the test statistic that measures how much the service times distribution deviates from the normal distribution, then [CLICK] determine the p-value based on that test statistic. Again, a small p value indicates your results are statistically significant, and that your sample data is likely not normally distributed.</td></tr>
<tr><td>🎙️ TH</td><td>Great work learning all of those tests. You’re now equipped to answer a wide variety of business questions about whether an observed effect reflects the truth or is likely due to random chance.

After you finish the practice assessment and practice lab for this lesson, I hope you’ll join me in the next and final lesson of this module to learn how you can use generative AI to perform and interpret hypothesis tests!</td></tr>
</table>

Lesson 3 – LLMs for inferential statistics

L3V1 – Interpretation with LLMs

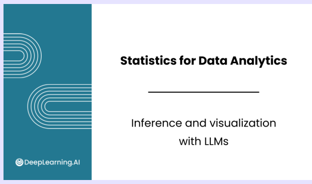
Visual	**Script**
🎙️ TH	How can large language models help you with hypothesis testing? These tests

	are complex operations, there are a lot of different choices you as the data analyst have to make.
🔗 [screencast]()	One thing I want to show you is using an LLM to help formulate your hypotheses. These can be tricky. For example, ***PROMPT:*** *I'm comparing the average delivery times between weekend and weekday orders for a food delivery service. Help me formulate appropriate null and alternative hypotheses for this scenario.* And the model is going to give you a null hypothesis. There's no significant difference in the average delivery times between weekend and weekday orders Or the alternative hypothesis, there is a significant difference in the delivery times. Okay, great. You can follow up on this. For example, it says, a significant difference. Now you can ask, is that H_1 for a one tailed or two tailed? And it says in this case it is for a two tailed test. So let's say that you want a one tailed test to check if the average delivery time on weekends is longer, which is what you're hearing anecdotally from drivers. It will give you an alternative hypothesis: the average delivery time on weekends is longer. Say you really want this to be in notation. Because this is what I just told it! It repeated the same thing back to me. And it gives you u of weekend is greater than u of weekday, and same for H_0. Okay, great. Now, here's another cool thing that you can do: interpreting results. I'm going to start a new chat so I can start the conversation over. Here's an interesting case: ***PROMPT:*** *I conducted a two sample t test comparing mean purchase amounts between two customer segments. The test statistic is 2.45 with a p value of 0.018. Explain what these results mean in business terms.* And it says the test statistic indicates that there is a significant difference in the mean purchase amounts. The p value of 0.018 suggests that there's strong evidence to reject the null hypothesis. And then you can follow up. Let's say that you're curious, can you break down what the p value of 0.018 means? If the null hypothesis is true, i. e. there's actually no difference in mean purchase amounts between the two segments, then there's a 1.8% chance of observing a difference in mean purchase

	amounts as extreme as the one found in this sample. It continues on with the discussion of your alpha level. Okay, a couple other cool things I want to show you. Let's start a new chat and get a clean slate. LLMs can help you double check your work, but you should also be very mindful of their potential for errors. So say you had this scenario. ***PROMPT:*** *Our product lead asked me to investigate if at least 90 percent of production timelines stay within budget. I reviewed 25 budgets and plan to conduct a one sample test for means with a null hypothesis that mu equals 0.9. And an alternative hypothesis that mu is less than 0.9.* Can you spot the problem with this question? I just want to double check my work, so I'll say please evaluate the effectiveness of this approach on a scale of 1 to 10 and provide evidence for your rating. This is a really effective way to get more than just yes or no feedback. It says the effectiveness is 7 out of 10. Now, you may have spotted a moment ago that 90 percent is not, in fact, a mean. This is a proportion. And yet, our LLM goes on to say that the hypothesized population mean is 0.9, which is not correct. This would not be the most effective way to approach this. So I'm not even going to continue reading, but I'm just going to say: are you sure that a test for means is appropriate? You're just nudging the model to think more deeply. It says, apologies for the confusion, you're correct. So you need to watch out for this. When you're using LLMs to help you figure out what your approach should be, don't just take their answer at face value. I can push it further and ask, ***PROMPT:*** *Would a test for proportions lead to a more accurate result?* And it says yes, it would be more accurate. The calculation itself is actually different for standard deviation and the test statistic, so you need to be on the lookout for that.
🎙️ TH	Alright! LLMs can be very useful for helping you formulate hypotheses, set up tests, and interpret results. Remember that LLMs are just a statistical representation of language. You have to make sure you're checking their output at every turn. Follow me to the next video to see how you can get an LLM that can write and

	run code to actually help you conduct hypothesis tests. See you there!

L3V2 – Inference and visualization with LLMs

Visual	Script
🎙️ TH 	How can you get a large language model to actually run a hypothesis test for you? Let me give you a couple of examples.
🔗 [screencast]()	***PROMPT:*** *I'm comparing salaries between two departments. Customer support has a mean salary of $65,000, with a standard deviation of $8,000, n equals 40, while purchasing has a mean of $68,000, standard deviation of $9,000, n equals 35. Perform a two sample hypothesis test to see if purchasing has a higher mean salary.* "Perform!" Perform means the LLM has to write and run code. It has to actually do math. If you're working with a traditional large language model, when you ask it to perform a two sample hypothesis test like this, it will just write out all the steps and you will have to go off and do the calculations yourself. But since we are using ChatGPT 4o, which can actually write and run code with its advanced data it will go ahead and do those steps for you. It's going to use a two sample t test. This is one type of hypothesis test you learned about in the previous videos. And it gives you a fancy name, Welch's t test! Cool. It defines the hypotheses: the mean of purchasing is less than or equal to the mean of customer support is the null hypothesis, and if purchasing is greater than customer support, that is the alternative hypothesis. It gives you the test statistic. Here are the values plugged into the equation, and this is nice to see because you can actually go ahead and double check whether those are in fact the true values. It calculates degrees of freedom. And now it's going to give me the results. I can actually go and look at the code. Now you don't have to look at this code and understand it fully. But the fact that it can write and run code means it does, in fact, have a way to get the true exact values. The test statistic is 1.52, and our p value is 0.067. Based on your own knowledge, if you have a significance level of 0.05, is this result statistically

	significant? **[pause for thought]** It is a relatively low p value, but we're not hitting our significance level yet. So the interpretation is, we do not have enough evidence to reject the null hypothesis. We cannot conclude that the mean salary in the purchasing department is significantly higher. One thing I would like to do is be able to see what this looks like. So I'm going to ask it to simply ***PROMPT:*** *Visualize this test.* We'll see what it comes up with. And here's a graph very similar to the ones you've seen so far in this course. Here is the t distribution. This red area is the rejection region with an alpha of 0.05. This vertical line is the test statistic, 1.52. And you can see that your test statistic does not fall within that rejection region. Visually you can tell this test statistic is a relatively rare value, but not quite as rare as our significance threshold. Finally, one thing that an LLM can be really helpful for is interpreting this for a certain audience. For example, my boss who isn't familiar with statistics. I'll just get a quick opinion here. You could say we compared the average salaries between customer support and purchasing to see if purchasing really pays more based on our analysis, since the purchasing department's average salary is slightly higher. The difference isn't large enough to confidently say that purchasing pays significantly more. The difference could just be due to random variation rather than a real difference. Okay, so you've seen how large language models can help you actually perform hypothesis testing. They're very useful. You have to make sure you're double checking their output at every step.
🎙️ TH	Great work running hypothesis tests using a large language model! That takes you to the end of this module, and you're almost to the end of this course! You've learned so much since histograms, mean, median, and mode. I'm proud of you. Coming up, you'll complete a practice lab with a large language model. Once you're done, you'll also complete the graded assessment and lab for this module, putting your hypothesis testing skills to the test to analyze diamond prices.

	Finally, you'll complete the capstone exercise for this course. You'll act as a data analyst working with a group of cardiologists studying heart disease. Your goal is to help analyze the risk factors associated with heart disease to help with prevention efforts. You'll combine everything you've learned in this module, from descriptive statistics to probability distributions to inferential statistics to put together a comprehensive, rigorous analysis. Once you're done with the graded assessment and lab as well as the capstone exercise, I'll see you in one final video to talk about your next steps as a data analyst. Keep up the great work and I'll see you on the other side of the graded assessments and capstone!

L3V3 – Your next steps

Visual	script
🎙️ TH	Congratulations on completing the capstone and this course! You've made incredible progress in the field of statistics, which is no small feat. You've come a long way since you wondered if you could start a new birthday card tradition among your coworkers, from calculating central tendency variability and skew, to simulating probability distributions in spreadsheets and using LLMs, to calculating confidence intervals and hypothesis tests. You're prepared to run rigorous statistical analyses in data analytics. And there's a lot more to learn! One of my favorite aspects of this work is how much I learn every day, even after years on the job. **[Ending 1]** So I encourage you to keep learning, keep taking courses, and lean into the many opportunities you'll have to learn more on the job. I'll leave you with one piece of advice one of my favorite teachers told me: Stay Classy San Diego! Wait, just one more question... **[MAGIC 8 BALL]** Is success in your data analytics career on the horizon? **[SHAKES 8 BALL]** Without a doubt. **[TOSS 8 BALL OFF SCREEN]** Congratulations again on completing this course and keep learning! **[Ending 2]** So I hope you'll join me in the next course in this series: Data Analysis at Scale with Python. So I hope you'll join me in the next course in this series: Data Analytics at

	Scale with Python. You'll learn the fundamentals of the Python programming language, as well as how to use Python for analysis, data cleaning, and visualization using both pandas and Seaborn. You'll implement all the statistics you've learned and more to create rigorous, scalable, and beautiful analyses. Wait, just one more question... **[MAGIC 8 BALL]** Will I see you again in the next course? **[SHAKES 8 BALL]** Without a doubt. **[TOSS 8 BALL OFF SCREEN]** Congratulations again on completing this course and I'll see you in the next one: Data Analysis at Scale with Python. Congratulations again on completing this course and I'll see you in the next one: Data Analytics at Scale with Python.