DAG C2M3 scripts

Lesson 1 – Confidence intervals

L1V1 – Inferential statistics

Visual

Script

🎙️TH

So far in this course and the previous one, you've been working with descriptive statistics. These describe the behavior of your sample data. Now, with inferential statistics, you'll level up your rigor significantly by using your sample to draw stronger conclusions about your population. Let's get started!

I want to start by asking you some questions about confidence. [CLICK] Say you're trying to figure out the rate of employee satisfaction at a company. [CLICK] There are 10,000 employees. [CLICK] You interview 100 of them, and [CLICK] 82 say they're satisfied. Based on that information, what's your gut feeling about satisfaction among all employees? [pause for thought]

I would say [CLICK] that satisfaction is pretty high, but how confident are you that your sample is representative of the entire company? Would you feel comfortable reporting this result to your boss? [pause for thought]

Okay, maybe you decide that you want to gather some more data, so you expand your survey to [CLICK] 500 randomly selected employees. You find that [CLICK] 455 of them are satisfied, or 91%.

How does [CLICK] this larger sample influence your confidence? If you had to give a range of values to your CEO now, what would you say? [pause for thought]

Last set of questions, say you take a survey of [CLICK] 500 randomly selected employees each month, and you find these percentages of satisfaction:

* [CLICK] April: 91.0%
* [CLICK] May: 92.1%
* [CLICK] June: 88.7%

Based on [CLICK] these three months of data, what would you now estimate the true rate of satisfaction to be? [pause for thought]

This is the idea behind inferential statistics.
* You have some sample data (like your employee survey) and you're trying to make inferences about the population (all 10,000 employees)
* Larger samples provide more reliable estimates than smaller samples – asking 500 employees is more reliable than asking 100.
* Different samples show variability, even when drawn from the same population – month to month the rate of satisfaction changes, even though the underlying population satisfaction might not be changing
* Inferential statistics allows you to quantify your level of confidence in your estimate – it's possible to say mathematically whether a certain estimate is more likely than another

Descriptive statistics state facts about your sample data. Insights like:
* [CLICK] In a sample of 100 employees, 82 said they were satisfied.
* [CLICK] A survey of parents showed that the 71 parents with newborns got on average 6.1 hours of sleep per night while the 89 parents with older children got 8.2 hours of sleep per night.
* [CLICK] In a clinical drug trial, 19% of patients report headaches as a side effect.

[CLICK] Inferential statistics, on the other hand, [CLICK] use the sample data to draw conclusions about the entire population. Insights like:
* [CLICK] Based on a sample of 100 employees, employee satisfaction among all 10,000 employees is likely between 88% and 91%.
* [CLICK] A survey of parents concluded that parents of older children sleep 2 hours more per night compared with parents of newborns.

The first three examples [CLICK] describe the characteristics of a sample, while the last two [CLICK] use the characteristics of the sample to make conclusions about how the population behaves – all employees, all parents.

You previously learned that your sample is a window into the truth. When you peer through that window, you won't necessarily see the truth. Inferential statistics is all about using probability to draw conclusions about the population based on your sample statistics, taking into account the size and variability of the sample, among other factors.

Inferential statistics provide a higher level of analytical rigor compared with descriptive statistics. The specific calculations are more involved–and there are possibilities of errors–but they allow for more robust conclusions about the broader population.

In the employee satisfaction example, you're able to state that 82 out of 100 employees surveyed are satisfied using descriptive statistics, but you can't generalize that to all employees.

Using inferential statistics, you might be able to infer with high confidence that the true proportion of satisfaction is 88-91%. Even though you're less certain about the actual value, you're able to conclude that the true population parameter is likely to fall within those two values.

If you're making a low stakes decision, like understanding an athlete's performance, then descriptive statistics may be completely fine for your use case. However, if you're making a more high stakes decision, like whether to invest millions of dollars into a certain product feature, inferential statistics provides a more robust basis for decision-making. It allows you to quantify the uncertainty in your estimates.

🎙️TH

There are two different types of estimates that you'll find yourself using regularly: point estimates and interval estimates. Follow me to the next video to learn more about the difference!

L1V2 – Point & interval estimates

Visual

Script

🎙️TH

You're 12 minutes away from the restaurant, and your friend texts you to ask how long you'll be. What do you say? 12 minutes? Or.. 10 to 15?

"I'm 10 to 15 minutes away" is a kind of interval estimate – you increase the chance that your estimate is correct by widening its range. Let's see how this works in statistics.

You've already seen two point estimates in the previous modules:[CLICK] x-bar, [CLICK] the sample mean, and [CLICK] s, [CLICK] the sample standard deviation. Each of these is a single value that [CLICK] represents a "best guess" about the population parameters [CLICK] mu and [CLICK] sigma. Another point estimate you'll explore later in this module is [CLICK] p-hat, which represents a [CLICK] sample proportion. [CLICK] For example, p-hat might be that [CLICK] the proportion of employees who are satisfied is 0.82.

Point estimates are useful, but they don't contain any information about the confidence of that estimate. Say you have a random sample of 25 movies from 2013 and the x-bar of their durations is 121 minutes. How certain can you be that the true population mean mu is exactly 121 minutes?

Interval estimates, by contrast, do contain information about how confident you can be. For example, you can be more confident of arriving within 10-15 minutes than you can be if you told your friend exactly 12 minutes.

https://i.imgur.com/AnXY7cO.png

You may have seen a graph like this before, which contains error bars [CLICK]. This is a graph of [CLICK]. three different teams at a company on the x axis and [CLICK]. their employee satisfaction score on the y axis. [CLICK]. Rather than having a point estimate, which would just be a plain bar at [CLICK] 72.7 for Engineering, [CLICK] 60.3 for Design, and [CLICK] so on, [CLICK] this graph shows an interval that [CLICK] represents the range of values where the true proportion is likely to fall.

[CLICK] Based on the intervals, which estimate do you think is the most precise? [pause for thought] That would be the estimate for [CLICK] Sales. It [CLICK] has the narrowest interval. Sales [CLICK] may have a larger sample, or it [CLICK] may have less variability within those samples. By contrast, [CLICK] engineering has the widest interval, indicating that [CLICK] there is a wider range of possible values for the true proportion.

These error bars are just [CLICK] a visual representation of an interval estimate – [CLICK] an inference about the true population proportion based on the sample. If you [CLICK] repeated this sampling process many times, you would [CLICK] expect the true population value to fall within this range most of the time.

🎙️ TH

Interval estimates grapple with a core complexity of statistics: if you take many samples of a population, you'll get different values for your sample statistics. Join me in the next video to see that complexity in action with a simulation!

L1V3 – Sampling distributions and the central limit theorem
Visual
Script
🎙️TH

Like populations, sample statistics also have distributions – ranges of possible values they can take on, and different probabilities associated with each one. Let's walk through an example.

Say you're tasked with [CLICK] estimating the average score on a professional certification exam. [CLICK] Possible scores range from 0 to 100.

If you take [CLICK] one sample, say by [CLICK] asking 50 random people what their scores were, you'll get [CLICK] one x-bar. But what if you take [CLICK] another sample? [CLICK] Will you get the same value? Let's run a quick simulation.
📽️ Screencast – LLM simulation

Link to simulation
Let's generate one new sample. This simulates asking a random sample of 50 people what their score was on the certification exam. So up here you can see some of the values that we sampled already and the sample mean of 76.59. In the next section, you'll see some statistics that summarize all of the sample means that have been generated so far, which is only one. And down below, you'll see a histogram of all the sample means.

For the first sample, the sample mean was 76.5, but if I generate a new sample, I get a different sample mean. This time it is 73. And if I generate another sample, now the sample mean is 74. And the next one is also different, 81.5. As I keep generating more and more samples, you'll start to see more and more values represented. All of these are possible sample means you could get when sampling this distribution.

Now I'll start generating hundreds of samples. A hundred more samples, more and more, and I have over 2,000 samples at this point. What distribution would you guess that this follows? [pause for thought] This curve suggests that the sample means are normally distributed.

Can you guess mu, the true population mean? [pause for thought] Somewhere between 76.5 to 77.5 looks right. In this case, the true population mean is 77.2.

This is a sampling distribution: the range of possible values you can get for x-bar, and the probability associated with each possible value. The idea behind sampling distributions is that you're more likely to get a sample mean that is close to your true population mean, in this case 77.2. As the value gets further and further away from the true population mean, it becomes less likely. If the true average score on this certification exam is 77.2, then if you ask 50 people their scores, an x-bar near the true mean is more likely, while an x-bar that's very low or very high is more rare.

It turns out that these sample means are actually [CLICK] normally distributed! This tendency is explained by the Central Limit Theorem, which states that [CLICK] if you take sufficiently large samples from any distribution and calculate their means, [CLICK] those sample means will be normally distributed. A [CLICK] sufficiently large sample here typically means [CLICK] n greater than 30.

In addition, [CLICK] the mean of that distribution will equal mu, the population mean of the distribution you sampled from. That's great! That means that [CLICK] the central tendency of the sample mean is around the population mean.

The standard deviation of this sampling distribution behaves a bit differently because [CLICK] as you increase your sample size,[CLICK] your estimate becomes more precise. It's similar to improving at archery. [CLICK] As you take more and more shots, the center of your target doesn't change, but your shots will become clustered more tightly around it.

[CLICK] The standard deviation of your sample mean, also called the standard error of the mean, is equal to the [CLICK] true population standard deviation [CLICK] divided by the square root of your sample size [CLICK] (σ / sqrt(n)).

Notice that [CLICK] as n gets larger, sqrt n also gets larger, but at a slower rate.

https://i.imgur.com/JT8UzW6.png

Here’s a graph of that relationship. [CLICK] For small sample sizes, [CLICK] as n gets larger, [CLICK] sqrt n grows quickly, reflecting that [CLICK] there are large gains in precision from adding a few more samples. [CLICK] But, [CLICK] as n gets larger and larger, [CLICK] sqrt n levels off, indicating that [CLICK] there are diminishing returns for the precision of your estimate as your sample size gets larger.

Even if you’re sampling data that isn’t normally distributed, [CLICK] your sample means will be normally distributed as long as your sample size is sufficiently large. Rainfall, for example, might follow a distribution [CLICK] like this, which you saw in the previous lesson. [CLICK] The distribution of the sample means will still be normally distributed.

[CLICK] The central limit theorem also doesn’t just apply to the sample mean x-bar. It also applies to the sample [CLICK] proportion p-hat, and [CLICK] in some cases to the [CLICK] sample variance and [CLICK] sample standard deviation as well. Even [CLICK] if you’re not sure about the underlying distribution of the population, you can still [CLICK] perform inference about these sample statistics based on the normal distribution.

🎙️TH

The central limit theorem is a pretty advanced concept, and if you’re not feeling super comfortable with it, that’s totally okay. The main thread here is that although sample statistics have variability, you can use an interval to estimate the true population mean. Follow me to the next video to see how.

L1V4 – Demo: confidence intervals in action

Visual

Script

Suppose you moved to a new apartment, and the first electricity bill arrived. You are a bit worried because your electricity bill seems a bit higher than what you used to pay, but maybe the appliances in your new house consume more. In order to better plan for costs you decide to estimate what your daily consumption is, so you started recording the daily readings from your electricity meter.

After 30 days, you got a sample of size 30 with sample mean of 8.35 kWh and sample standard deviation of 4.1

* Pop these values on a spreadsheet as you mention them
  * A1: xbar
  * A2: 8.35
  * B1: s
  * B2: 4.1
  * C1: n
  * C2: 30

You could say that the mean consumption is 8.35 kWh. If you want just one value, this is ok, but it doesn't really give you any information about the variability in the consumption. Maybe it's better to get a confidence interval. Let's use 95% confidence.

I will go ahead and do some calculations that might not be super clear right now, but don't worry. You will learn to do all of it in the following videos!

* Choose a cell and compute the margin of error (Say A4)
  * = 1.96 * B2/SQRT(C2)

This quantity represents the uncertainty of the interval that you are developing to estimate the mean electricity consumption. In general, a smaller value means more certainty about the population parameter that you are estimating, and a larger value means more uncertainty.

Finally, let's calculate the lower and upper bounds of the interval
* Choose a cell. Type Lower: and next to it enter =A2 - A4
* Choose a cell. Type Upper: and next to it enter =A2 + A4

So your mean electricity consumption is between 6.87 kWh and 9.81 kWh with 95% confidence. This means that if you were to repeat this experiment of registering your electricity consumption every day for 30 days many many times, then around 95% of the times the interval you found would contain the true mean value.

OK, that's all for this example! Follow me to the next video to learn more about how to calculate confidence intervals like the one I just demonstrated in this video.

L1V5 – Confidence intervals
Visual
Script
🎙️TH
Now that you've seen how confidence intervals can help you develop powerful estimates, let's work through an example calculation and interpretation together.

Say you're working as a data analyst at a bakery that delivers fresh pastries to a local zoo each morning. The items must arrive [CLICK] before 7am, which is when the zoo opens each day. You've been tasked with [CLICK] figuring out how long it takes to deliver the pastries in order to help with [CLICK] scheduling delivery times and [CLICK] creating performance targets.

Now, rather than guessing or using your magic 8 ball, first [CLICK] you'll want to collect some sample data. You [CLICK] monitor the delivery truck for 30 days, [CLICK] each day recording the time it takes to get from the bakery to the local zoo where the pastries are sold.
From your sample, you calculate an [CLICK] x-bar of 43 minutes and an [CLICK] s of 11 minutes. But it's not enough to stop here and say the average delivery time is 43 minutes, because this is just your[CLICK] sample mean. Maybe just [CLICK] due to random chance, your deliveries in those 30 days were [CLICK] quite fast, or maybe they happened to be [CLICK] quite slow.

You only have this one sample, so how can you estimate the true population mean delivery time?

Let's visualize what it would look like if you took 1000s more samples. You might get a sampling distribution of the mean like this [CLICK]. Remember that the [CLICK] sampling distribution is normally distributed, and the [CLICK] sample mean x-bar approximates the true population mean.

Let's add in the [CLICK] standard deviations, so these are all the sample means within one standard deviation of the mean, [CLICK] above or [CLICK] below, [CLICK] 2 standard deviations of the mean and [CLICK] 3 standard deviations of the mean.

[CLICK] Your sample of 30 delivery times, where you calculated x-bar of 43 minutes, is just one of these thousands of samples.

[CLICK] Maybe 43 minutes falls here; it's a longer than average group of times. Remember that one property of the normal distribution is that [CLICK] 50% of the values are above the mean, [CLICK] 50% are below it. So [CLICK] there's a 50% chance of 43 minutes being a longer than average sample mean.

Or maybe 43 minutes is here [CLICK], it's much faster than the true average delivery time. After all, there's a [CLICK] 50% chance of it being a below average time.

Here's one way to think about it. [CLICK] What is the probability that your sample mean 43 minutes is within 2 standard deviations of the true mean? [pause for thought] [CLICK]Based on the three sigma rule you learned in the previous module, that would be [CLICK] 95%. Another way of saying that is: [CLICK] if you choose any random value in this distribution, there's a 95% chance that it's within 2 standard deviations of the true mean.

Taking this a step further, you can quantify the range of values that make up 95% of the possible sample means. Let's quantify that for this particular distribution of delivery times.

* As an approximation of the true mean, you can use your sample mean, which in this case is [CLICK] 43 minutes.
* You also saw in the previous video that the [CLICK] standard error, which is [CLICK] the standard deviation of this sampling distribution, is [CLICK] sigma over square root n. You don't know sigma, since you don't know the true population parameters, however you can use your sample standard deviation as an approximation. In other words, now you have s over the square root of n, which is equal to 11 minutes divided by the square root of 30, which is right around 2.
* To quantify the range of values making up 95% of possible means, you calculate [CLICK] x-bar minus 2 times the standard error as the lower bound and [CLICK] x-bar plus 2 times the standard error as the upper bound.
* Reducing that down, x-bar is [CLICK] 43, minus 2 times 2 equals 39 on the lower bound, and [CLICK] x-bar plus 2 times 2 which is 47 on the upper bound.
* Putting it all together, you can say with 95% confidence, that the true average delivery time is between [CLICK] 39 and [CLICK] 47 minutes.

This is a confidence interval. It's a range of values that is likely to contain an unknown population parameter. It quantifies the uncertainty of an estimate by the relative width of the range. Broad ranges are associated with relatively high uncertainty, whereas narrow ranges offer more precision in the estimate.

You can bring this confidence interval back to your colleagues at the bakery to help with decision-making. [CLICK] Understanding the possible average delivery time can help [CLICK] develop precise scheduling, and can [CLICK] inform performance targets for delivery drivers.

🎙️TH

You just calculated a 95% confidence interval, an extremely powerful statistical technique for estimating population parameters. Even though you can't be certain how unusual your sample is, the confidence interval helps you make an informed estimate about the range of values that is likely to contain the true mean.

Confidence intervals can be tricky to interpret. Follow me to the next video to learn more about what they represent.

L1V6 – Mechanisms of confidence intervals
Visual
Script
🎙️
You've calculated a 95% confidence interval, but what does it mean? In this video, you'll see what the "95%" part of a confidence interval really means and where it comes from.

Let's step back and look at this statement: with 95% confidence, the true average delivery time is between 39 and 47 minutes.

* The 95% confidence reflects the probability that any confidence interval you calculated would contain the true mean. Remember that the population parameter is fixed, but unknown to you. So either it's in this particular interval or it's not.
* Looking at the sampling distribution again, 5% of sample means fall in these two tails of the distribution, 2.5% on the high end, 2.5% on the low end. These are unusual values that happen by chance. It's possible for your sample to have been one of these unusual cases that would lead you to over or underestimate the true mean. In other words, for 5% of confidence intervals you calculate, they won't contain the true population mean.

Screencast

[overlay simulation process on right side of screencast]

🔗LINK TO SCREENCAST (ignore audio)

🔗 Simulation link

Let me show you what I mean. This simulator generates 30 random samples from a normal distribution with a population mean of 45 and a population standard deviation of 10. And then it's going to calculate a confidence interval based on that sample. Let's generate one new sample. On the x axis of this graph, these are all the possible values for the population mean. The red line here is the true population mean of 45.

The confidence interval is the gray area between these two dotted lines, which represents a range of values that are estimated to contain the true mean. In this case, this confidence interval is between 41.4 and 48.6, which does in fact contain the true mean.

Right now I have a success rate of 100%, but I can generate more samples. This one also contains the population mean. And also this one. Now I've generated a few more samples, and you can see that this confidence interval does not contain the true mean. So the upper bound is 44.8, which is just barely lower than the true population mean. There are actually three confidence intervals that do not contain the true mean.

If I generate many more samples, you can see that getting a confidence interval that does not contain the true mean is a relatively rare event. And I'll just generate a bunch more so you can see that now the success rate is down to 95.46%.

What this simulation is showing you is that when you calculate a confidence interval for the mean based on a sample, you have a 95 percent chance that the confidence interval will in fact contain the true mean.

And that's the whole purpose of confidence intervals. You have some uncertainty in your sample; you don't know how far or how close the sample mean you calculated actually is from the true mean. But, using inferential statistics, you can calculate a range of values that is likely to contain the true mean. You're not looking directly at the truth or a singular answer, but a powerful estimate of that truth.

Now that you've developed some intuition for confidence intervals, here's the formula to calculate a confidence interval for the population mean:

X-bar +- z times (s/sqrt(n)). You've seen all these values before: x-bar and s are your sample statistics and n is your sample size.

z represents a z score value from the standard normal distribution, and it controls how “confident” you are that the confidence interval contains the population mean. Recall that the z score is the same as the number of standard deviations from the mean in a standard normal distribution.

In the previous video, 2 standard deviations from the mean to estimate a 95% confidence interval, The actual value that represents this level of confidence is 1.96, and that’s because the three sigma rule is just an estimate. 2 standard deviations above and below the sample mean is actually associated with slightly higher confidence than 95%, so in practice you’ll use a z score of 1.96 to be more precise.

So to put this all together to calculate a 95% confidence interval, you have this formula:

X-bar +- 1.96 times (s/sqrt(n)).

The term on the right [show margin of error] is also referred to as the margin of error, and it quantifies the range within which the population parameter is expected to fall. It’s the part that constructs the confidence interval, and helps you gauge the precision of the sample estimate.

You can calculate confidence intervals with different levels of confidence as well. What confidence levels do you think might be useful? [pause for thought] The most common are a 90% confidence interval, with a z score of 1.645 and a 99% confidence interval with a z score of 2.576.

Note that the higher level of confidence means that you use a higher z score, and therefore generate a wider confidence interval. So in order to increase your confidence for a given sample, you will need to broaden the range of the interval estimate.

Choosing the confidence level for a particular estimate depends on a few factors.
* 95% is the most commonly used confidence level, because it balances confidence with the potential for error. For example, you might estimate the average delivery time for your bakery using a 95% confidence interval between 38 and 49 minutes. It's important to arrive on time to keep your contract with the zoo, so you try to balance certainty about the average delivery time with the precision of your estimate. An

11-minute range means that you don't need to schedule too much buffer time into your scheduled departure.
* A 90% confidence level might be used for preliminary studies or when missing the true value is less important. For example, if you're working with a product research team, you may use a 90% confidence interval to estimate that users on average rate a new feature a 7.2 out of 10. In the early stages of development, a less precise estimate is acceptable to help make initial decisions.
* A 99% confidence level is used when you want to minimize the risk of error. For example, if you're working with a team of scientists estimating pollution in a river, it may be important to have a high confidence level to pass regulatory testing. Calculating that the pollutant concentration in a river is between 5 and 8 parts per million with 99% confidence can help your team reduce the risk of harmful environmental impact.

Keep in mind that for all these estimates, even the 99% confidence interval carries some chance of error.

🎙️TH

Great work going through the simulations of confidence intervals and the general formula! In the past few videos, you've seen that confidence intervals depend on your sample size and confidence level. How do these terms interact? Follow me to the next video to take a look.

L1V7 – Understanding margin of error

Visual

Script

🎙️TH

What makes a confidence interval narrower or wider? Some of the factors depend on your sample data, and one depends on a decision that YOU get to make. Let's take a look.

Let's zoom in on the margin of error. [show formula for margin of error] This term, as you saw in the previous video, determines how wide your confidence interval is. In general, a narrower confidence interval is more desirable because it means you have a more precise estimate. So rather than saying our average delivery time is somewhere between 38 to 49 minutes, wouldn't it be nice to say it's between 40 and 43 minutes? That precision allows for better scheduling and more specific performance targets.

Margin of error depends on three different factors: your desired confidence level, your population standard deviation i.e. the amount of variability in your data, and your sample size. Let's take a look at how changes in each of these factors impacts the margin of error.

For all these confidence intervals I'd like to represent the proportional size of each interval calculated (3 per example)

First up, the desired confidence level. You saw in the previous video that some common confidence levels you'll use are 90, 95, and 99%. And the corresponding z scores are 1.645, 1.96, and 2.576, respectively.

Let's do some quick calculations based on the delivery driver example. I'll hold s and n constant, with s of 10 and n of 100. That means this part of the margin of error term [s/sqrt n] will come out to 10 over 10, or 1.

Starting with a confidence of 90%, that gives you a z score of 1.645. The margin of error would be equal to 1.645, so the width of the full 90% confidence interval would be twice that, or 3.29 minutes wide.

Next up, the 95% confidence interval. z is now 1.96, making the confidence interval 2 times 1.96 or 3.92 minutes wide.

Then finally the 99% confidence interval. z is 2.576, time two gives you a width of 5.152 minutes.

This example illustrates that if you want to have a higher confidence in containing the population mean within your interval, the interval will get wider.

Show numerical comparison first, then show graph

Notice that the width of the interval is not proportional to the relative increase in confidence. A 95% confidence interval is 19% wider than the 90% one, even though you only gain about 5% in relative confidence. Similarly, the 99% interval is 31% wider than the 95% one, and you only get a 4% gain in relative confidence.

This relationship is related to the shape of the normal distribution, which as you know accumulates less probability in the tails. As you move further from the mean, you need to cover an increasingly larger range to capture an additional percentage of confidence.

Let's now zoom in on the impact of variability on the confidence interval. I'll hold n constant at 100 again, and I'll use a z of 1.96 for a 95% confidence interval.

With s=10, you have 10 over 10 which equals 1 times 1.96. So the width of this confidence interval is 2*1.96, or 3.92 minutes wide.

If sigma is 20, the variability in your data is twice as large, so the margin of error is 20 over 10 times 1.96 equals 3.92. That makes your confidence interval 2 * 3.92 or 7.84 minutes. Double the variability makes the confidence interval twice as wide.

Finally for sigma of 30, you get 30 over 10 times 1.96 equals 5.88, times 2 gives you a range of 11.76.

Notice that the relationship between the standard deviation and the margin of error is linear. As s goes up, the margin of error increases proportionally, and the same when s decreases. The main consequence for business applications is that if your data has a lot of variability, you can expect a less precise estimate.

For example, arrival times for a city bus system may be significantly more reliable than for a rural bus system. Which one do you think will be easier to estimate? [pause for thought] That would be the city bus system. The smaller variability makes it easier to estimate the true average arrival time compared with the rural bus system.

Sample size has a bit more of a complicated relationship with margin of error, since you're dividing by its square root. Here's a concrete comparison.

For a constant z of 1.96 and a s of 10, imagine a sample size of 100. Your margin of error works out to 1.96 times 10 / 10 which is 1, so the width of this confidence interval would be 2*1.96 or 3.92 minutes.

Then you could have a sample size of 200, so double the number of delivery times recorded. That works out to 1.96 times 10/sqrt(200). Sqrt(200) is around 14.1, so 10/14.1 equals about 0.71. 0.71 times 1.96 equals 1.39, making the width of your confidence interval 2*1.39 or 2.78 minutes wide.

Finally, with a sample size of 300, you have 10/sqrt(300) which is around 17.3. that gives you 0.58 times 1.96 which is 1.31, times two gives you a range of 2.62 minutes.

What that means is that a larger sample with the same variability allows you to construct a narrower confidence interval with the same chance of containing the true mean. You're dividing the standard deviation by sqrt(n), so a larger n reduces the size of the margin of error.

However, larger sample sizes produce diminishing returns. Increasing your sample size by 100% reduced the size of the confidence interval you could calculate by 29%. Then adding 100 more deliveries only reduced the interval by about 6%.

https://i.imgur.com/5AoHSfk.png

Here's a graph of the relationship between sample size and margin of error. You can see that the relationship is negative, since a larger sample size decreases your margin of error, but there are diminishing returns as the slope flattens out the higher n goes.

Using this graph, can you estimate where the benefit from a larger n seems to level off? [pause for thought] I would say somewhere between 100 and 200 there's an inflection point in the graph.

This relationship is a basic fact of statistics, and it helps explain why you often need a relatively small sample size even to estimate a very large population like the opinions of everyone in the country of France. The consequences in your work will be that, as long as you have n = 30 or higher, getting more and more data may not narrow your confidence interval much.

So to summarize, you can achieve a narrower confidence interval in three ways:

* With a larger sample size, though it has diminishing returns
* By working with data that has less variability

* Or by lowering your confidence level, which can increase your chances of missing the true value

🎙️TH

Great work studying these relationships. So far, you've been working with confidence intervals for means, but you can calculate confidence intervals for other statistics too. Follow me to the next video to learn about confidence intervals for proportions.

L1V8 – Demo: confidence intervals for means

Visual

Script

Let's take a closer look at confidence intervals in action. Let's use a dataset you should already be familiar with: the Forest fires dataset. I've already included the is_small column that you used in module 2.

<open the file and show them the new column>

* As you remember, the area is pretty skewed towards small values, but you would like to know with some confidence what is the true population mean area. What components do we need to calculate the confidence interval? That's right, we need the sample mean, sample standard deviation, number of samples, and the z score
* Create a new sheet, and add a little table:
  * In A1: Mean
  * In B1: SD
  * In C1: n
* Compute these values
  * First, get the center of the interval, a.k.a. the sample mean
    * In A2: = AVERAGE(Data!M:M), it is 1.111
  * Now let's find the sample standard deviation
    * In B2: =STDEV(Data!M:M) = 1.398
  * Find the number of samples
    * In C2: =COUNT(Data!M:M) = 517
* Now we have to make a decision, what confidence do we want? Let's start with 95%.
  * In A3: Confidence:
  * In A4: 95%
  * In B3: z-score
  * In B4: 1.96
  * You can get the margin of error.
    * In C3: Margin of error
    * In C4: B6: = B4 * B2 / SQRT(C2)

* To find the lower bound of the confidence interval,subtract the margin of error to the mean
  * In D3: Lower
  * In D4: A1-C4
* nd to get the upper bound, just add the two
  * In E3: Upper
  * In E4: A1+C4
* That's it! You just calculated your first confidence interval!

* What if you wanted the 99% confidence interval? The sample statistics are the same, you just need to change the z-score to 2.576.
  * In A5: 99%
  * In B5: 2.576
  * And update the Margin of error and bounds of the interval
    * Drag from cells C4:E4
    * Oops, this didn't quite work, I forgot to fix the cells of the statistics
      * Add $ in C4:E4 to fix cells A2, B2, C2
    * Now that is much better!
  * What happened with the length of the interval?
    * The margin of error increased from 0.1205 to 0.1584. That's 30% bigger!

OK, that covers it for computing confidence intervals for the mean in a spreadsheet environment. The key is to calculate the sample statistics that you need, and then make an appropriate selection of the z-score for your desired confidence level.

L1V9 – Confidence intervals for proportions

Visual

Script

🎙️TH

It's useful to estimate the mean of a population, but the mean isn't the end-all, be-all of statistical inference. You're often interested in some other aspects of your population. One of them is the population proportion. Let's define this parameter and cover how to calculate a confidence interval for the proportion.

Let's reconsider the example of analyzing delivery times for the bakery delivery scenario. While estimating the mean delivery time is useful, you might also be interested in investigating the rate of on-time deliveries. You may ask a question like "What proportion of deliveries are on time?" Those are deliveries that made it to the zoo by 7am. You can represent this proportion as p, the true proportion of on-time deliveries.

Suppose you collect a sample of 30 deliveries and record if they were on time or not. The proportion of on-time deliveries in your sample is represented by p-hat, which is your estimate for the true proportion p. Hat is just a funny term for an estimate.

Let's imagine that you measured the proportion of on-time deliveries in your sample as 0.6, which would be 18/30 on-time deliveries and .... Is not great 🙂

Similar to x-bar, the sampling distribution of p-hat is also normally distributed, this time with a mean equal to the true proportion p. So if you took many many samples of 30 deliveries and calculated the proportion of on-time deliveries for each sample, those proportions would be normally distributed. Your p-hat of 0.6 falls somewhere within this distribution. What do you think is the chance of p-hat falling within 2 standard deviations of the true proportion p? [pause for thought] Just like with means, there's a 95% chance that p-hat is within two standard deviations of p.

Now let's talk about the standard error. The standard error for the proportion is written as the square root of p times (1-p) over n. Sometimes, a q is used to represent (1-p), but I am going to minimize introducing new variables! This representation also makes it clear that these two quantities, p and (1-p) are directly related, and are in fact complements of each other.

Let's break this formula down a bit. Just like sigma, the quantity p(1-p) represents the variability in your data. Also similarly, you don't know the true population proportion p, so you use the sample proportion p_hat to estimate this quantity.

Here's a visual example for a scenario with 10 deliveries. Suppose you observed 9 on-time deliveries and 1 late delivery. This sample data would have low variability because only 1 out of the 10 outcomes was different from the others. Contrast that with these 10 deliveries from your sample, where 6 are on time and 4 are late. This data has higher variability because there is a more even mix of success and failures.

In the first scenario, p_hat is 0.9, so p_hat(1 - p_hat) equals 0.09. For the second scenario, p_hat is 0.6, so p_hat(1-p_hat) is 0.24. So mathematically, there is higher variability when p_hat is closer to 0.5.

Next, you'll divide this quantity by your sample size, which again represents the idea that a large sample size gives you a more precise estimate. And lastly, you'll take the square root to make sure this quantity is at the original scale of your data, rather than a squared proportion.

Now, you have your p-hat and standard error, so you can construct a confidence interval for this proportion, similar to how you would construct a confidence interval for the mean.

The interval is defined as p-hat plus or minus z times the standard error, which is the square root of p_hat(1-p_hat) over n. Do you remember the name of this term on the right? [pause for thought] As with the confidence interval for the mean, this term on the right is called the margin of error, and again represents the uncertainty in your estimate.

Here's the 95% confidence interval for the proportion on-time deliveries for your sample scenario. P-hat is 0.6 so (1 - p_hat) is 0.4, z is 1.96 for a 95% confidence interval, and n is 30.

The resulting confidence interval ranges from (0.4247, 0.7753), meaning that with 95% confidence, the true proportion of on time deliveries is somewhere within this range. This wide interval is reflected in the high variability of the data as well as the relatively small sample size. Oftentimes, you need a larger sample size to calculate precise intervals for a proportion.

🎙️TH – WEARING P HAT

So there you have it, you've now calculated confidence intervals for means and proportions! There are a broad range of business questions that you could apply these estimates to. Follow me to the next video to get hands-on with calculating sample means and proportions in spreadsheets. See you there.

L1V10 – Demo: confidence intervals for proportions

Visual

Script

🎙️TH

Let's calculate a confidence interval for the proportion using the forest fires dataset. The steps are very similar to those for the mean!

🖥️ SC

Let’s use the is_small column, and find the confidence interval for the proportion of fires that are small.

* Create a new table
  * In A10: p-hat
  * In B10: 1 - p-hat
  * In C10: n
* Compute this values
  * First, get the value of p-hat from the column is_small
    * In A11: = AVERAGE(Data!N:N),
  * Then get the value for 1 - p-hat. You can do this by simply subtracting p-hat from 1
    * In B11: =1-A11
  * Next find the number of samples
    * In C11: =COUNT(Data!N:N) = 517
  * Ok, you are done with the sample statistics
* Now let’s define the z-score and with that the margin of error and interval bounds
  * Say you want to calculate a 95% confidence interval for p
    * In A13: z-score
    * In B13: 1.96
  * Next, we need to calculate the margin of error. Again, the formula changes a bit from the mean, as the standard error is now the square root of p_hat*(1 - p_hat) / n
    * In C13: =B13*SQRT(A11*B11/C11)
  * For the lower bound, subtract the margin of error from p-hat
    * In D13: =A11 - C13
  * And for the upper bound, you will add the margin of error
    * In E13: =A11 + C13

The resulting confidence interval is (LB, UB), which means that we have 95% confidence that this interval contains the true proportion of fires in small areas.

Ok that’s it for this demo! You just learned how to find confidence intervals for proportions using a spreadsheet.

Lesson 2 – Hypothesis testing for means

L2V1 – Demo: hypothesis testing in action

Visual

Script

🎙️TH

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 82 seconds, while your friend has an average of 80. You friend says “ah 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 an average of 2 seconds fair and square.” Who’s right?

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

The size of the difference is pretty small – you only won by 2 seconds. And you did only solve 20 cubes. Maybe 100 solves would give you a better idea of who’s the best. A hypothesis test can take all this information into account and help you reach a conclusion with high confidence: who’s the better cube-solver, if anyone at all?

🖥️ SC – 🔗 link to demo spreadsheet

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 I just mentioned, 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:A22,0)
* 24.7%! So if you were evenly matched, this kind of result would be pretty common.
* Say you went 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? Now I'll calculate the probability of observing a 3 second difference or larger if you're equally skilled
   * ▶️ In D13, =Z.TEST(C2:C101,0)
* Only 0.24%! So if you were in fact equally skilled, this difference would be rare. There's reason to believe that you are in fact faster than your friend, even if it's only by 3 seconds on average!

L2V2 – Hypothesis testing: means
Visual
	Script
	🎙️TH

	In the previous course, 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 the suspected population mean. Let's see how it works.

	Let's return to the example from the previous module where you were working with a music streaming service. You're working on a new project to figure out if giving users a free trial is effective. You decide to investigate whether users who receive a free trial stay subscribed for longer on average.

Users who don't receive a free trial tend to stay subscribed for about 10 months.

You collect a sample of 100 users who received a free trial and calculate the following descriptive statistics:
* X-bar, the sample mean is 10.4 months
* s, the sample standard deviation, is 2 months

That's close! What do you think, is the difference of 0.4 months large enough to convince you that the free trial is effective? [pause for thought]

You can't be sure just yet. What you don't know based on these descriptive statistics is where this mean of 10.4 falls in your sampling distribution.

You know from studying the Central Limit Theorem that the sampling distribution for means is normally distributed. Here are three scenarios for you to consider:
* The true population mean mu is 10 months, so 10.4 months falls here in the sampling distribution.
* The true population mean mu is 9.6 months, and 10.4 months falls around here in the sampling distribution.
* And lastly, the true population mean mu is 9 months, and 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]

The first outcome is more likely than the other two. Recall that the standard error is s over square root n, so here it's 2 over square root 100, or 0.2. In the first case, x-bar is 2 standard deviations from the mean, in the second, x-bar is 4 standard deviations away, and in the last, 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 extremely unlikely.

The tricky part is, you can never know the true population mean to compare with. But what 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 9 months, it would be 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 10 months, it's much more likely.

Now, in this case, it's so tempting to say 10.4 is higher and call it a day!

But, these numbers are close. It's possible that the true mean is actually 10, and the free trial costs you money while giving you no benefit. Or maybe it's 9.8, and the free trial is slightly worse for retention. This level of precision is important for your conclusions.

What a hypothesis test allows you to do is evaluate whether your sample mean (10.4 months) is significantly different from the suspected population mean (10 months), given the variability in your data and your sample size. It distinguishes between two possible scenarios:

1. The observed difference between your sample mean and the suspected population mean is due to random chance. The values are too close for you to tell if they're actually different.
2. The observed difference reflects a genuine difference between the true population mean and the suspected value.

This distinction is called statistical significance. If the difference is due to random chance, that's not statistically significant. It 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 statistically significant. The difference is likely real and can provide evidence for your hypothesis.

It's very common for you to collect a sample and calculate a mean that's different from the true population mean. For example, when rolling two dice, the population mean of their sum is 7 – that's also the most likely roll. But, you'll roll something other than seven 83% of the time.

So just observing that two values are different isn't enough to conclude that that difference is meaningful, that's just how the cookie crumbles sometimes.

Keep in mind that hypothesis tests only work effectively under certain conditions. It's crucial that your data is a representative sample, ideally a random sample. All statistical tests assume random sampling because if your sample isn't random, you have no way of knowing what biases your sampling method introduced.

The observations in your data must also be independent.

Additionally, your data must meet one of two conditions. Either it is normally distributed, or your sample size must be large. Typically "large" means 30, but 50 or more is ideal. That's because of the Central Limit Theorem, which states that 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 lesson.

Over the next few videos, you'll see how to perform a hypothesis test for means. You'll cover these steps:
* Defining your hypotheses
* Calculating the test statistic
* Defining the significance level
* Calculating the p value
* Making a decision and interpreting the results

You'll also learn how to define errors and 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.

🎙️TH

Hypothesis testing isn't just for business. Doctors use it to determine if a new medication is more effective than the existing one. Environmental scientists use it to check if a conservation effort is improving wildlife populations. And more!

Regardless of the use case, your first step is defining the hypothesis. It's crucial for making sure you construct and interpret your test correctly. Follow me to the next video to learn how.

L2V3 – The hypothesis

Visual

Script

🎙️TH

In statistical hypothesis testing, you'll need to define two related hypotheses: the null hypothesis and the alternative hypothesis. Here's how and why.

Breaking down the music subscription service scenario, you were investigating the mean retention of people who got a free trial. There were two possible outcomes of your test:
* Mu = 10 months
* And mu > 10 months

These are your 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 mu = 10. This hypothesis is called the null hypothesis, and it represents the condition where you aren't able to find evidence that mu was greater than 10. It's associated with finding no effect or no difference. This hypothesis is written as H sub 0, or "H naught".

Alternatively, if you could find evidence that mu is greater than 10 months, that would be great! That evidence would indicate that the free trial was effective in getting users to subscribe for longer. This hypothesis is called the alternative hypothesis – it's the alternative to the null – and it's written H sub 1.

In general, when defining your hypotheses, start with the null hypothesis. Identify the value you would expect if there was no effect. For example, if there was no effect of getting a free trial, you would expect a mean subscription length of 10 months. That gave you H sub 0 of mu = 10. Technically, the null should be the complement of the alternative hypothesis, which would be mu <= 10, but oftentimes it's only written with the equality condition.

For your alternative hypothesis, you'll always compare the population parameter with the value in the null hypothesis, in this case 10. Define the comparison you're interested in: are you hoping to find evidence that the mean is greater than, smaller than, or just different than the expected mean? Your options would be

* H sub 1 mu > 10, is the mean significantly greater than 10?
* H sub 1 mu < 10, is the mean significantly less than 10?
* H sub 1 mu =/= 10, is the mean significantly different than 10?

You can only have 1 alternative hypothesis. So in the example from earlier, you chose mu > 10, 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 evidence for the alternative hypothesis (more on the specifics of how later) – you would say you "reject the null hypothesis". The data suggest that the null hypothesis is likely not true.
* If you don't find evidence for the alternative hypothesis, then you would say that you "fail to reject the null hypothesis". It doesn't mean that the null hypothesis is true, just that we don't have enough evidence to reject it.

The language of statistics is important here. You should avoid phrases like "prove the alternative hypothesis" and "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. This terminology helps avoid overstating your conclusions, and reminds your stakeholders that these tests can never prove with absolute certainty.

As in science, your hypotheses should be based on some theory or observable evidence. In other words, don't just select them at random. For example, when dealing with the music subscription service, it makes sense to choose the null hypothesis that subscription length equals 10, since that's the average for users who didn't get a free trial. It's plausible that the behavior of users who got a free trial is similar.

🎙️TH

Hypotheses are the foundation of your testing strategy, and it's important to get them right. In the next video, you'll get practice identifying hypotheses depending on your business problem. I'll see you there!

L2V4 – 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 testing the pH of a swimming pool, for which the ideal pH is 7.4. Any significant difference from 7.4 is unsafe, whether it's higher or lower. Pause the video for a moment and see if you can jot down the hypotheses you would test. [pause for thought]

Your null hypothesis H sub 0 would be mu equals 7.4. This value represents the status quo.

Your alternative hypothesis would be for unsafe pH levels, in this case 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, 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]

Your null hypothesis would be that mu equals 120 minutes. That value represents no difference between the movies in 2013 and what you expected.

Your alternative hypothesis would be that mu > 120 minutes. You're looking for evidence that movies are longer than two hours. Either you find that evidence, and you reject the null hypothesis, or you don't find that evidence and you fail to reject the null hypothesis.

Earlier in this module, you also constructed confidence intervals for bakery delivery times. You could also use a hypothesis test to investigate 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]

The null hypothesis would be that mu = 45 minutes. That's the status quo, if delivery times weren't different from 45 minutes.

The alternative hypothesis would be mu < 45 minutes. You're interested in finding evidence that the delivery times are less than 45 minutes. You can reject the null hypothesis if you do find that evidence, or fail to reject it otherwise.

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 three types of alternative hypotheses:
* mu > some number
* mu < some number
* And mu =/= some number

These hypotheses are each associated with a different type of test.
* In the first case, you'll reject the null hypothesis if your x-bar is rare on the upper end of the distribution. You should perform a right-tailed test, since you're only interested in the upper tail of values.
* In the second case, you'll reject the null when x-bar is rare on the lower end. You should perform a left-tailed test, since you're interested in rare values only below the mean.

* Both left- and right-tailed tests are considered one-sided 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 reject the null if you find that your x-bar is unusual on either the upper or lower end. You should perform a two-tailed test because you're interested in whether x-bar falls in either the upper or lower tail of values.

Let's see how that shakes out with the three examples you just saw:
* 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 two-tailed test, since either a pH that's too high or too low would be unsuitable.
* Testing movie lengths, recall that your alternative hypothesis was that 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 right-tailed test, since you want to check if x-bar would be unusually high.
* And if you're checking whether the average delivery time is below 45 minutes, you're interested in a left-tailed test.

🎙️ TH

Great work defining your hypotheses and determining your test types! It's nuanced but fun work. Follow me to the next video to see how to complete the next step in conducting your hypothesis test: calculating the test statistic.

L2V5 – 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. Let's see how to perform this calculation and what it tells you.

Say you collected a random sample of 50 movies and calculated a sample mean of 123 minutes and sample standard deviation of 12 minutes. Here's the sampling distribution of the mean for movie duration. Mu is 120 and the standard error is 12/sqrt(50), which equals about 1.7.

You're performing a right-tailed test, with the hypotheses:
* H sub 0, the null hypothesis, states that mu = 120 minutes
* And H sub 1, the alternative hypothesis, states that mu > 120 minutes

Is x-bar different enough from your 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 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 sample size is lower, you may not be certain your results really reflect the population.

The 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, here are the descriptive stats for your sample. [show x-bar, s, and n on screen]

And here were your hypotheses. [show hypotheses on screen]

First, you'll calculate the difference between x-bar and mu, which gives you 3. This step centers your calculations around 0. Then, you'll divide this difference by the standard error, which tells you the number of standard errors between this test statistic and the hypothesized mean.

Since you don't know the population standard deviation, you will use the formula s over square root n, which gives you about 1.7. All together, 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 a z score!

This is the number of standard deviations from the mean on the standard normal distribution. Essentially, you're translating your sample mean into a value on a standardized scale, where the mean is zero, and each step represents one standard deviation.

https://i.imgur.com/zSkWu9B.png

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

Let's visualize this value. Here's the standard normal distribution. Can you spot where your test statistic z = 1.76 falls? [pause for thought] It's right 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. It's not in one of the tails above a z-score of 3, but it's also 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.

L2V6 – 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?

Consider the movie duration example, where you want to 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 "confidence", or the level of certainty you have in your conclusions. You learned that you could construct an interval with 90, 95, or 99% confidence, or some other value as well. Your choice depends on how precise your estimate needs to be. If you needed to be very sure about your results, would you choose 90% of 99% confidence? [pause for thought] You'd choose 99%. Hypothesis tests rely on a similar intuition.

As a data analyst, you will have to make this call to determine what confidence level you are comfortable with. Remember, as with all inferential statistics, you are trying to manage uncertainty. It can never be fully eliminated.

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

As a recap, here are the descriptive stats for your sample. [show x-bar, s, and n on screen]

And here were your hypotheses. [show hypotheses on screen]

Since you've calculated the z-score for your test statistic, you're working with the standard normal distribution.

If you wanted to be 95% confident that you correctly reject the null hypothesis, you would look for test results above the mean that you expect to occur 5% of the time or less. This shaded region represents those values of z that should occur 5% of the time or less. It's called the rejection region, because 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.

When performing a one-sided hypothesis test with a 95% confidence level, your ability to reject the null hypothesis depends on the size of the rejection region, which has an area of 0.05. This value is called the level of significance, and is represented by the Greek letter alpha. An alpha of 0.05 is quite common, and is often used in initial studies for medical research, manufacturing quality control, and the social sciences. As you may realize, confidence is the complement of the level of significance.

https://i.imgur.com/K1B7P4n.png

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 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]

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

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

There's one final nuance you should be aware of. You just saw the rejection region for a right-tailed test. The process is quite similar for a left-tailed test which also has just one rejection region.

However, for a two-tailed test, you're interested in values above and below the mean.

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

For movie lengths, your null hypothesis would remain the same, but this time your alternative hypothesis H sub 1 would be that x-bar =/= 120.

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

Okay, that was a lot of info. To recap the terminology you learned,

* Alpha, your significance level, helps define the strength of evidence required in order for you to reject the null hypothesis. Larger alphas lower the bar for the strength of evidence required, whereas smaller alphas require stronger evidence against the null hypothesis. Common values for alpha include 0.10, 0.05, 0.01, and 0.001.
* The rejection region is the area of the sampling distribution that contains the improbable values that would lead you to reject the null hypothesis. For alpha = 0.05, you saw this rejection region was the area under the curve in the top 5% of the distribution. In the movie theater example, the rejection region represents the range of average movie lengths that are so surprisingly long, you'd conclude movies are indeed longer than 120 minutes on average.

🎙️TH

Alpha, your significance level helps you quantify the amount of uncertainty you're comfortable with. It can be 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.

L2V7 – 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. Let's see how to perform this calculation and why.

https://i.imgur.com/u7CqpJq.png

Earlier, you calculated your test statistic z = 1.76. How rare is this value? Let's visualize this idea by seeing if this value falls within your rejection region.

You can see that z falls within the rejection region, slightly above the boundary. So 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 calculate exactly how rare this value is, using a p-value, the probability of getting a sample mean as rare or rarer than z in the direction of your alternative hypothesis. So in other words, the probability of getting z of 1.76 or higher.

For this step, you'll need one of two things. Either a lookup table, or a spreadsheet or programming language that can calculate it for you.

🖥️ SC

Let's talk through how to calculate the p value for our movie example. Again, you need to calculate the probability of observing a z-score of 1.76 or higher, which will take you back to your coverage of probability and the standard normal distribution. Recall the cumulative distribution function, or CDF, which represents the probability of observing a z-score less than or equal to a specific value. So how might you use the CDF to calculate the probability that you're interested in? [pause for thought] That's right, you need to use the complement rule. The probability of Z > 1.76 equals 1 minus the probability that Z <= 1.76, which is what the CDF tells you.[d]

The Google Sheet function for calculating the CDF value for the standard normal distribution is NORM.S.DIST, which only takes in one value … the z-score for which you want to calculate the CDF value. So =NORM.S.DIST(1.76) tells you the probability that Z <= 1.76, which is about 0.9608. The complement of this probability is then 1 minus this value, or 0.0392. This value is the p-value, the probability of observing a more extreme test statistic than the one you observed, again in the direction of the right-tailed alternative hypothesis.

Suggestion, show where the test statistic appears within the normal distribution, and the area to the right which is the p-value of 0.0384. Compare that to the rejection region defined by the level of significance of 0.05.

Since $p = 0.0392$, if the true average duration of all movies is 120 minutes, you would observe a sample mean of 123 minutes or longer about 3.92% of the time. What do you think, is this rare enough to reject the null hypothesis? [pause for thought]

Compare your p value with your significance level, in this case 0.05. Since your p value is below your significance level, you will reject the null hypothesis. Intuitively what you're doing with this comparison is asking whether this event is expected to happen less than 5% of the time. And that's the case here!

Now, recall that movie schedules might be difficult to adjust, so what if you instead required a 1% level of significance for your hypothesis test? [pause for thought] In this case, you would not reject the null hypothesis because the p-value–which remains unchanged because your test statistic is the same–is greater than your alpha. You would not have observed strong enough evidence that the mean movie length is greater than 120 minutes.

One thing to note with this is that you should think carefully about your level of significance BEFORE you conduct your test. You should not adjust it in order to make the decision you want to make. I am making some of these adjustments for illustration purposes of how the decision could be different.

Let's step back and put everything together. You had a business question about whether movie durations were on average longer than 120 minutes. You collected a sample of 50 movies and found their mean duration was 123 minutes, with a standard deviation of 12 minutes.

You wanted to know, is there sufficient evidence that movie durations are above 120 minutes? Then, via the 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 p value, you found that you would observe a value as extreme or more as 123 minutes about 3.92% of the time. Since you wanted to be 95% confident in your conclusion, you rejected the null hypothesis, concluding that there was 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 several of these calculations in a spreadsheet which will help you develop your intuition for how hypothesis testing will work in practice.

L2V8 – Demo: hypothesis testing for means[e]

Visual

Script

🎙️TH[f]

Now it's time to learn how you can use spreadsheets to get your p-values and make decisions from them. You will see it is pretty similar to the steps you followed for confidence intervals. Let's go back to the forest fires dataset

* Calculate one of every type of distribution & practice with the p value and interpretation of it
* Be very specific with language
* We can come up with examples rather than using real data if necessary

Let's go back to the forest fires dataset. As you can imagine, the bigger the fire size, the larger the number of resources local agencies will need to dispense to put them out. In order to plan accordingly, the agencies want to know if the mean burned area is larger than 1ha, because in that case they would need to plan for more resources.

What should the hypothesis be? [pause for thought]

The null hypothesis, H sub 0, should be that the mean area is 1, while the alternative hypothesis, H sub 1, should be that the mean is greater than 1

As you saw in the lectures, let's begin by finding the test statistic. Remember that you needed the sample mean, or x bar the sample standard deviation, s, and the sample size, n. These are the same statistics you used for the confidence intervals
* Create a new sheet
* In A1: x bar
* In B1: s
* In C1: n
* In A2: =AVERAGE(M:M)
* In B2: =STDEV(M:M)
* In C2: =COUNT(M:M)

Now find the test statistic. Remember you need to subtract mu 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))

Now all that is left is for you 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, that is what rarer means in this case.
* In E1: p-value
You can do this using the NORM.S.DIST function. However, this gives you the probability that a normal variable is less than or equal to some value, but you want it the other way around. By the complement rule you can do
* In E2: 1 - NORM.S.DIST(D2)
This gives a value of 0.035, what would you conclude [pause for thought]

For a 5% significance level you can reject the null hypothesis. However, in this case you might want to be extra careful because a bad planning of resources could end in a tragedy, so you decide to go for a 1% significance level. What happens now? [pause for thought]

Now you can't really say you have enough evidence to reject the null hypothesis that the mean burned area is 1, and should plan accordingly.

Now, let's move to the ISI, or Initial Spread Index, feature. A rating of 10 indicates high rate of spread shortly after ignition, while a rating of 16 or more indicates extremely rapid rate of spread. You are asked to verify that the mean of the ISI stays below 10. What should your hypotheses be? [pause for thought]

Since you want to prove that the mean ISI <10, then that should be the alternative hypothesis, while mean ISI = 10 should be the null hypothesis. Let's repeat the steps.

Again you will need
* 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)
The test statistic will still be the same: You subtract mu=10 from the sample mean and divide by s/square root of n
* In D3: =(A3 - 10) / (B3 /SQRT(C3))
The final step is finding the p-value. In this case, since you have a left-tailed test, you want the probability that a standard normal variable is smaller than or equal to the test statistic. From the value of the test

statistic you should already be able to tell that the p-value will be really small .But how small? This can be done directly
* In E3: =NORM.S.DIST(D3)
You can safely say that you can reject the null hypothesis even at a 1% significance level.

Let's finally look into the temperature. Extreme temperature conditions could make things harder for firefighters, so you are asked to check if the mean temperature differs from 18 degrees Celsius, which, for reference, is equivalent to 64.4 degrees Fahrenheit. What should your hypotheses be? [pause for thought]

Since you want to see if mean temperature deviates, in any direction, from 20 degrees, then you are looking at a bilateral test, where H0 is mu = 20, and H1 is mu =/= 20

First part of the job is exactly the same: Get the feature statistics,
* In A4: =AVERAGE(Data!H:H)
* In B4: =STDEV(Data!H:H)
* In C4: =COUNT(Data!H:H)
And the test statistic
* In D4: =(A4 - 20) / (B4 /SQRT(C4))
Now you need the p-value. Since it is a bilateral test, something "rarer" means something spreading over the two tails of the distribution. In this case, since the test statistic is positive, it means something bigger than the test statistic or something smaller than the negative test statistic. There are two ways in which you can do this:
1. You can sum the two probabilities, each calculated as in the previous two tests, so you get (in cell E4)
   * = 1- NORM.S.DIST(D4) - for the right tail
   * + NORM.S.DIST(-D4) - for the left tail
2. Another way is nothing that the two areas must be the same because the normal distribution is symmetrical, so you can alternatively use (in cell F4)
   * =2*NORM.S.DIST(-D4)
As you can see, you get the same p-value of 0.0004
What can you conclude?
Well, the p-value is smaller that 0.001 you can reject the null hypothesis even at a 1% significance level.

L2V9 – Getting it wrong: false positives and false negatives

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 who's the better solver. In real life, there are only two possible realities:

1. You're equally matched
2. One person is better

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

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

If you're equally matched, and your test didn't find a significant difference, that's great! Your window into the truth saw the correct effect. If you're not equally matched, and your test found one of you to be better, that's great too!

You can see there are two ways to go wrong, though. It's possible that you're equally matched, but your test concluded that one of you is better. That's a false positive. You found a difference, but in reality there isn't one.

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

A table a bit like this:

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

Then there's the true state of the world:
1. Either the null hypothesis is true
2. Or the null hypothesis is false.

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

You have a few mechanisms for controlling these error rates. One of them is your significance level alpha. You saw earlier how your significance level represents the 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 the same as the false positive rate. You can try to minimize false positives by setting alpha to a very low value, but that will increase the likelihood of false negatives. Setting alpha too low means you're being very strict about what counts as significant. This makes it harder to find true effects.

Increasing your sample size, or planning to have a sufficiently large sample, is an important factor in the success of your hypothesis test.
* A large sample provides a more precise estimate and helps reduce both types of errors
* A large sample can also help you detect a very subtle effect. For example, it's possible that 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.

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 manufacturing, you'll likely test for defective products before they're shipped. A false positive would mean identifying a product as defective when it's actually fine, whereas a

false negative would be not detecting a defective product. You may prefer to minimize false positives if customers can easily make returns, to avoid wasting product that's actually fine.
* In medical testing for serious illnesses, however, it's typically more desirable to minimize false negatives. A false negative happens when a patient actually has a particular illness, but the test fails to detect the illness. It's more desirable 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 balance between the two types of errors. For example, if you're working at a bank that approves loan applications, a false positive might mean approving a loan for someone who actually defaults, while a false negative might mean denying a loan to someone who would in fact pay it back fully. The bank will likely conduct 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. Sometimes you're just unlucky.

Now, you're almost to the end of this lesson! There's just one more concept you'll need to learn, and it's a slight nuance with hypothesis testing. Follow me to the next video to learn more.

L2V10 – A big reveal (student's t distribution)

Visual

Script

🎙️TH

So, I temporarily held the full truth from you. 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 carry through with your calculations for hypothesis testing. Let me explain more.

Recall[g][h] the conditions for the Central Limit Theorem, which state that when you are calculating the sample mean based on 30 or more observations, the sampling distribution of the sample mean is normally distributed. So what happens if you calculate a sample mean based on less than 30 observations – do you think the variability will be larger or smaller? [pause for thought] The sample mean for smaller samples is likely to have more variability than for larger samples. This condition introduces more uncertainty into your estimate.

You also saw that the 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 sample standard deviation as an estimate for sigma, you introduce even more uncertainty.

When 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 t-distribution. Let's discuss how the normal and t distributions differ, and how that difference affects your work.

Something like this showing a few t distributions:

The t distribution is very similar to the standard normal distribution. They both have a smooth shape and are symmetric about a mean of zero. However, the t distribution is flatter in the middle and 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 what I just shared about this distribution reflecting more uncertainty about the data that it models.

However, one difference is the t distribution is defined by a parameter known as the degrees of freedom (df), which is a statistical measure that tells you the amount of information in your sample that can be used to form an estimate. It's a bit of an abstract concept, but for this scenario, the degrees of freedom is calculated as the sample size minus 1. Why the 'minus one'? Because you are using the sample standard deviation instead of the population standard deviation in your calculation of the standard error, therefore, the information contained in your estimate is slightly reduced.

As degrees of freedom increases, the t distribution becomes more and more similar to the normal distribution, and therefore your choice of which distribution to use becomes less important.

The fundamental idea of the t distribution is that, in order to calculate a confidence interval or perform a hypothesis test, you generally don't know the population standard deviation. 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.

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

You'll follow a similar process for conducting your hypothesis test as you just learned. You will define your hypotheses the same, and collect the sample statistics needed in order to calculate your test statistic. The calculation for the test statistic is the same as I've described, but actually, when you use s instead of sigma, the test statistic becomes a T, rather than a Z.[i][j][k] Z follows the standard normal distribution, and T follows the t-distribution.

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

Consider the movie duration example, with a hypothesized population mean of 120 and a sample standard deviation of 12 over just 5 samples. Take a look at the difference between the normal and t-distributions. You can see the t distribution has bigger tails. The rejection region for alpha = 0.05 is also different. Notice that the rejection region for the normal distribution starts around 128 minutes, while for the t distribution, it starts around 132 minutes. That means that the t distribution requires stronger evidence to reject the null hypothesis.

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

* In the normal distribution, 130 minutes falls here, and the p value is 0.031, meaning you would 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.

When the t distribution was initially developed, it was much more common to be dealing with small sample sizes of say 10 or 15, and to try to draw conclusions about the population from that small group. Nowadays, you're more likely to be working with larger sample sizes above 30 or 50. You learned in the previous course that you'll often be working with big data. In tech, if you want to survey users, you might be able to get thousands of people. In manufacturing, your systems might record the production time for every product. If you’re working with emergency room data, you might have access to thousands of visits rather than a dozen.

The t distribution is still relevant in many cases, such as in researching rare illnesses or studying endangered species, when you’re working with small samples. When your sample sizes are above say 50, the difference is not likely to impact your decision making.

🎙️TH

And that’s a wrap on hypothesis testing! You’ve learned a ton, from how to formula your hypotheses to identifying your rejection region, calculating your test statistic, and interpreting your p value.

Next up, you’ll complete the practice lab about [topic], which I know you’ll enjoy. When you’ve finished the practice quiz 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!

Lesson 3 – Other hypothesis tests

L3V1 – Hypothesis testing: proportions

Visual

Script

🎙️TH

You’ve practiced creating hypothesis tests for means, and you can actually transfer all that knowledge to other kinds of hypothesis tests! You can test whether a sample has a particular proportion of a certain value, 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:

* What proportion of your coworkers is in favor of the birthday card policy?
* Is the proportion of valid canine DNA test kits really 0.7?
* What proportion of our bakery deliveries to the zoo are on time?

Let's focus on delivery times for this example. Say if more than 5% of deliveries are late, you risk jeopardizing your contract with the zoo, and you would want to move up your delivery driver's start time to ensure they can make it to the zoo on time. So you're interested in the question of 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. Define your hypothesis
2. Define your significance level
3. Calculate your test statistic and p value
4. Make your decision and interpret the results

Suppose you collect a sample of 50 deliveries and record if they were late or not. The proportion of late deliveries in your sample is p-hat, which is an estimate for the true proportion of late deliveries p.

You can define your hypotheses as
* H sub 0 (H0) that p <= 0.05, which corresponds to the status quo, for which the proportion of late deliveries is 0.05 or less. If this hypothesis holds, you would not change your start time because you are confident that your deliveries will consistently arrive on time.
* H sub 1 (H1) 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 alpha of 0.05, which is associated with a 5% false positive rate. The sampling distribution for the sample proportion is the normal distribution, shown here. So you would reject the null hypothesis if your test statistic falls within this rejection region in the top 5% of values.

Let's imagine that you measured the sample proportion of late deliveries as 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 calculate the test statistic z. It's quite similar to the one for means. You use z because the sample proportion is normally distributed.

First, subtract your hypothesized proportion p zero (p0) from your sample proportion p-hat (p̂). This operation has the effect of centering your test statistic at 0. Then you'll divide by the standard error for the distribution, which essentially tells you how many standard errors from the hypothesized proportion your sample proportion is. The standard error looks like the square root of [p zero (p0) times 1 minus p zero (p0) divided by n].

Why not use p-hat to calculate the standard error? In a hypothesis test for means, you hypothesize the true mean, but you often don't know the true standard deviation, so you must use your sample standard deviation to approximate it. However, in the hypothesis test for proportions, you can derive the population standard deviation just from p0 and your sample size. You 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 zero is 0.06 - 0.05 = 0.01. P0 times (1 minus p0) is 0.05 times 0.95, which is 0.0475. This is on the lower end of variability, since you expect most deliveries to be on time.

Divide 0.0475 by 50, your sample size and you get 0.00095. Now take the square root to get 0.0308. So your test statistic z is 0.324. 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, this value is above the mean by less than 1 standard deviation.

How rare would a value of 0.324 or higher be? Take a look at z graphed on this sampling 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 fail to reject the null hypothesis! z is outside the rejection region.

Let's calculate your p value to make sure your intuition is right. You can use a lookup table or software to do so. The p value of z of 0.324 or more extreme, in this case higher, is 0.373. So, if the true proportion of late deliveries is 0.05, 37.3% of sample proportions will be 0.06 or higher.

Next, you'll compare your p value with alpha. 0.373 is greater than alpha, so you fail to reject the null hypothesis.

Your interpretation of this result is that there's not sufficient evidence to conclude that late deliveries happen more than 5% of the time. So, you don't decide to adjust your delivery driver's start time, and 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.

L3V2 – Demo: Hypothesis testing for proportions

Visual

Script

🎙️TH[l]

Let's put what you just learned in action! Let's keep working with the forest fires dataset. This time, let's try to evaluate whether the proportion of small fires constitutes the majority of fires you observe in the data set.

You can see that in this version of the dataset you can find the is_small column, which takes on the value 1 if the burned area was below 0.5ha, and 0 if it was greater.

You might remember from the module 2 graded lab that if you plot a barchart of this feature, the proportion of small and not small fires seem pretty balanced.

Let's create the plot to refresh our memories
* Select is_small → Insert Chart → Bar chart → X axis: is_small → Aggregate → Count

Let's perform a hypothesis test to see if the proportion is in fact 0.5. What should the hypothesis be? [pause for thought]

We can consider the two sided test, where the null hypothesis is p=0.5, and the alternative hypothesis is p=/= 0.5.

Let's get on to it. For the proportions test, just as with the confidence interval, you will need
* [Create a new sheet]
* The sample proportion, p-hat
  * In A1: p-hat
* 1 - P-hat
  * In B1: 1 - p-hat
* And the sample size

  * In C1: n
* 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.
* In D1: 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 E1: p-value

Since this is a two sided test you need to consider both tails of the distribution. In this case, the test statistic is negative so you want values smaller than the test statistic, or bigger than the opposite of the test statistic. Since the normal distribution is symmetric, you can just double the probability of observing a value less than your test statistic.
* In E2: =2*NORM.S.DIST(D2)
This gives a p-value of 0.311

Given this p-value is greater than any reasonable level of significance, you don't have enough evidence to conclude the true proportion is different from 0.5.

L3V3 – 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. Let's see how and why it's done.

You've seen a few cases throughout this course where you might want to compare two samples directly:
* Do movies from 2020 have the same average duration as movies from 1930?
* Is the proportion of people who have a positive opinion of AI the same in rural versus urban communities?

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:
* Define your hypotheses
* Determine the significance level
* Calculate the test statistic
* Calculate the p value
* Interpret the results and make a decision

Here's how you'd set up your hypotheses. First, the null hypothesis:
* Your status quo is that there's no difference between the groups. So you'll write that as H sub 0 is mu sub basic = mu sub premium
* Your alternative hypothesis is that these two means are different, so you can use a two-tailed test with the hypothesis H sub 1 is mu sub basic =/= mu sub 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 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 80 subscribers, with the following descriptive stats:
* Mu of basic = 9.7 months
* Sigma of basic = 3.3 months
* Mu of premium = 10.4 months
* Sigma of premium = 1.8 months

In this case, each of the samples are independent, meaning that the length of the subscription for each basic subscriber is independent of the length of the subscription for each premium subscriber. So we're going to take the approach of treating these samples as independent from each other.

You'll calculate your test statistic a bit differently for two sample tests under different conditions. In this case, you'll typically use the t statistic, which is relevant when your population standard deviation is unknown. You start by subtracting x-bar premium from x-bar basic.

Then you'll divide by the standard error, which represents the estimated variability in this sampling distribution based on the sample size and variability in the sample data. The formula for the standard error depends on a key assumption, which is whether or not you believe–or have evidence–that the variances between the two samples are equal or not.[m][n][o]

Once you've calculated your test statistic, you can move forward to calculate your p value and see whether it's smaller than your alpha.

Calculating this all out, the t statistic is −1.67, with a p value of 0.098. That means you would expect a difference in means of this much or greater about 9.8% of the time. What's your conclusion? Do you reject or fail to reject the null hypothesis? [pause for thought] In this case, you'd fail to reject the null hypothesis, since your p value is not less than alpha.[p][q]

It's important to remember this test assumes the samples are independent. It means that the subscription length for users in the basic plan is not influenced by the subscription length for users in the premium plan. 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 promotional rate to encourage basic subscribers to upgrade, that might shorten the duration for basic subscribers while extending the duration for premium subscribers.

Two sample hypothesis tests are more commonly used in practice than one-sample tests, since you're often interested in comparing two groups.
* In A/B testing, you create two different versions of your product and show those versions to different groups. Then, you calculate whether the groups responded to each version differently.

* In clinical research, you often have experimental and control groups. You give your experimental group the new treatment, such as a new drug, and give the control group a placebo. You then want to compare whether the experimental group showed more benefits than the control group.

🎙️TH

What happens if you're interested in comparing more than two groups, like age brackets? Or pairs of groups like patients' improvement before and after a treatment? There are many types of hypothesis tests beyond the two you've seen. Follow me to the next video to learn more.

L3V4 – Other hypothesis tests

Visual

Script

🎙️TH

You've learned a lot of different statistical tools in just one module! You're almost to the end of it. There are many different hypothesis tests, and they answer different kinds of questions. 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.

Imagine you're not just comparing two groups, like basic and premium subscribers, but three or more groups. For example, you might have a tiered subscription model with basic, premium, and 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 analysis of variance test, also called ANOVA. This test involves calculating the means for each group plus the overall mean. Then, you compare how much the group means differ from the overall mean versus 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. A small p-value (typically less than 0.05) suggests the differences between groups are significant.

Say you're interested in testing whether students' moods improve after drinking a particular energy drink. One option would be to take a random sample of students who had the energy drink, and 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 drink water and assess their mood, then have them drink your energy drink and assess their mood again.

In this case, you actually have 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.

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.

The p-value here indicates how likely you'd see a difference this large by chance if the treatment had no effect. Again, a p-value less than 0.05 is often considered significant.

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 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 creating a table of observed frequencies. Then, you calculate the expected frequencies assuming no relationship. The chi-squared statistic measures how much your observed frequencies deviate from these expected frequencies.

A small p-value suggests that the observed frequencies differ significantly from what you would expect if there is no relationship.

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 testing whether that's actually true. You can use a goodness of fit test.

Say you want to know if customer service times at a call center follow a normal distribution. In a goodness of fit test, you'll calculate the test statistic that measures how much your data of service times deviates from the normal distribution, then 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.

🎙️ TH

So that's a wrap on hypothesis testing! 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 quiz for this lesson, I hope you'll join me in the next and final lesson of this module to learn about how you can use generative AI to construct confidence intervals, perform hypothesis testing, and interpret results from both!

Lesson 4 – Generative AI for inferential statistics

L4V1 – Generative AI for inferential statistics

Visual

Script

Graveyard – rejection region & critical value script

Visual

Script

🎙️ TH

Once you're comfortable with your hypotheses and test type, 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?

Consider the movie duration example, where you want to 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 "confidence", or how the level of certainty you have in your conclusions. You learned that you could construct an interval with 90, 95, or 99% confidence, or some other value as well. Your choice depends on how precise your estimate needs to be. If you needed to be very sure about your results, would you choose 90% of 99% confidence? [pause for thought] You'd choose 99%. Hypothesis tests rely on a similar intuition.

As a data analyst, you will have to make this call to determine what confidence level you are comfortable with. Remember that, as with all inferential statistics, you are trying to manage uncertainty. It can never be fully eliminated.

Images of movie distributions – https://i.imgur.com/d7dBIli.png

Let's see how hypothesis testing can help solve the movie theater with scheduling. Say you collected a random sample of 50 movies and calculated a sample mean of 123 minutes and s of 12 minutes. Here's the sampling distribution of the mean for movie duration. Mu is 120 and the standard error is 12/sqrt(50) equals 12/~7 equals about 1.7.

You're performing a right-tailed test, because you want to know if movies are longer than 120 minutes, not shorter.

If you wanted to be 95% confident that you correctly reject the null hypothesis, you would look for test results above 122.79 minutes. 122.79 is called the critical value here; finding a test result above that value would lead you to reject the null hypothesis, since that value would be very rare if this distribution truly has a mean of 120. This shaded area above the critical value is called the rejection region, since you'll reject the null hypothesis if your test result falls in this region. That conclusion may lead you to adjust the theater's scheduling practices.

When performing a hypothesis test with a 95% confidence level, you're interested in the top 5% of values; or expressed as a proportion, the top 0.05 of the values. This value is called alpha, your significance level. An alpha of 0.05 is quite common, often used in initial studies for medical research, manufacturing quality control, and the social sciences.

One common misconception: you can't just compare the sample mean directly with the critical value; there's actually another step involved, which you'll see in the next video. For now, focus on how the critical value, rejection region, and significance level interact in this distribution.

Here's how to calculate your critical value. The critical value defines the value below which

Imagine that the theater's schedules are expensive to adjust. Choosing to show fewer movies in a day, because movies are longer, may reduce revenue and cause employee schedule changes. You want to 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 alpha, or significance level? [pause for thought]

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

In this case, you start with the same distribution. Do you expect this critical region to be smaller or larger than the previous one? [pause for thought] Your critical region gets smaller. Here's what it looks like. You want to be more precise, so you'll only reject the null for values in the top 1% of the distribution.

Your new critical value is 123.95 minutes. That's about 1.16 minutes difference compared with the critical value for your alpha of 0.05. In the movie theater scenario, this higher critical value means you'd need even stronger evidence of longer movies before changing the scheduling.

There's one final nuance you should be aware of. You just saw the rejection region and critical values for a right-tailed test. The process is quite similar for a left-tailed test which also has just one critical value and one rejection region.

However, for a two-tailed test, you're interested in values above and below the mean. For example, imagine you were interested in whether movie lengths were significantly above or below 120 minutes.

Your null hypothesis would remain the same, but this time your alternative hypothesis H sub 1 would be that x-bar =/= 120.

Here's what that would look like on the distribution. Same distribution, different hypothesis.
* Take a look at the upper rejection region for a moment. The critical value is 123.33. The critical value in your right-tailed test was 122.79. So, is this rejection region for the two-tailed test smaller or larger than just the right-tailed test? [pause for thought]
* It's smaller! In fact, it's half as small. Your rejection region contains 2.5% of the data on either side – upper and lower, totaling 5%. That's because you want to maintain the same precision, with errors only 5% of the time, but you have two rejection regions. If they both contained 5% of the values, that would actually lead to an error rate of 10%, not 5%.

* Notice you also have two critical values, a lower one and an upper one. To reject your null hypothesis, your test statistic can be on either side, as long as it's below the lower one or above the upper one.

Okay, that was a lot of info. To recap the terminology you learned,

* Alpha, your significance level, is your cutoff point; values above that cutoff will lead you to reject the null hypothesis. You saw that common alphas include 0.05, 0.01, and 0.001.
* The critical value is the actual value in the distribution associated with that cutoff point. For example, for alpha = 0.05, you saw the critical value was 122.79 minutes; finding a value above that would lead you to reject the null hypothesis.
* The rejection region is the area of the sampling distribution that contains the improbable values that would lead you to reject the null hypothesis. For alpha = 0.05, you saw this rejection region was the area under the curve above the critical value 122.79. In the movie theater example, the rejection region represents the range of average movie lengths that are so surprisingly long, you'd conclude movies are indeed longer than 120 minutes on average.

🎙️ TH

Alpha, your significance level helps you quantify the amount of uncertainty you're comfortable with. Your rejection region and critical value can then be used to determine whether your sample data leads you to reject or fail to reject the null hypothesis.

Follow me to the next video to see how to calculate your test statistic, which you can then compare with your critical value to perform your hypothesis test.

Lesson 4 – GenAI for inferential statistics

L4V1 – Interpretation with GenAI

Visual

Script

🎙️ TH

You've now seen how to construct confidence intervals and complete hypothesis testing. How can large language models help you with these tasks? They're very complex, there are a lot of different options.

🔗 screencast

One thing I want to show you is using an LLM to help formulate your hypotheses. These can be tricky. For example, 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 H1 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 H0. 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: 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.

You can actually do the same for a confidence interval. Let's start a new chat. So say you constructed a 95 percent confidence interval for the mean rent price of a one bedroom apartment in Budapest, Hungary. I calculated the interval as 550 to 950 euros. Explain what these results mean in business terms.

It means that you can be 95 percent confident that the true mean rent price of a one bedroom apartment in Budapest, Hungary falls within this range. 95% indicates that if we were to repeat the study multiple times, approximately 95 percent of the resulting confidence intervals would contain the true mean rent price.

Then it gives me this business context. In the real estate industry, you could use this information to set rental rates or evaluate potential returns on investment. Again, you can follow up and ask questions like, does this mean that 95 percent of apartments in Budapest are between those two price points?

It says no, it does not mean that. So you can ask follow up questions to clarify your understanding. It's an estimation of the range within which the true mean rent price is likely to fall.

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. 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 [purposely incorrect] 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, 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.

Alright! LLMs can be very useful for helping you formulate hypotheses, set up tests, and interpret results. And you also have to be on the lookout for these potential problems. LLMs are just a statistical representation of language. You have to make sure you're checking their output at every turn.

One helpful tip is, as you tested in this video, you should try to answer every question you ask the LLM. The more you know what to expect, the more you'll be able to debunk any inaccuracies in the LLM's response.

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 and visualize different confidence intervals and hypothesis tests. See you there!

L4V2 – Calculation and visualization with GenAI
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
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 “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, while 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, let's try a confidence interval. Our new app has been downloaded 517 times, with 89 ratings. Of those, 75 users gave it a 5 star rating. Calculate and interpret a 95 percent confidence interval for the true proportion of users who would give the app 5 stars. This is a really interesting sample: we're taking all of the ratings and we're using that sample to try to understand all of the users.

And again, “calculate” means the LLM has to write and run code. If you were just going to use a vanilla LLM, such as the one in Coursera, you wouldn't want it to calculate for you.

Now it's going to give me the steps. Here's the formula for a confidence interval, which you saw already. There's our margin of error for proportions. This is a proportion test: 75 users out of 89. That's not a mean, that is a proportion. First, p-hat is 84%. Then you're going to calculate your standard error and margin of error, which is 0.075, so about 7.5%.

That gives you your confidence interval. which is between 0.767 and 0.919. So we're 95 percent confident that the true proportion of users who would give the app a 5 star rating lies between 76.7% and 91.9%. If we were to repeat this process many times, 95 percent of the calculated intervals would contain the true proportion.

There's a lot of text here. Let's just visualize this confidence interval on a graph of the proportion from 0 to 1, so we can see what that range looks like.

Okay, interesting graph! [tongue in cheek – it’s ugly but functional] The blue dot is p hat, that is the center of our confidence interval, and on either side we have plus and minus one margin of error.

No matter what, this is quite a high proportion. Maybe the true proportion could be below 80%, it could be above 90%.

Now, one more follow up prompt: how would changing the confidence level to 99% change the results and interpretation? Please also visualize this new interval on the same type of graph.

It says the confidence interval will become wider. That's because a higher confidence level requires a larger range to ensure that the true proportion is captured within the interval.

One thing that I'm noticing about this right here is that it doesn't appear to be writing code to calculate this interval. Concerning! If I look at the code, I see something about figures and plots and lines, and I can see the labels for the graph, but I don't see anything here that's actually creating a confidence interval. These are just the confidence levels that it determined.

I'll ask it, please try again and use code to calculate the interval.

Now we got 0.7433 to 0.9421, there we go. So I can see that this time it does, in fact, use code to determine the confidence interval itself. Those values are really close to the original values that it guessed just using its statistical patterns.

I can see that this green confidence interval here below is, in fact, wider than the red one above. So that does reflect my intuition.

Okay, so you've seen how large language models can help you actually calculate confidence intervals and perform hypothesis testing. They're very useful. Again, you have to make sure you're double checking their output at every step.

I hope you go ahead and try these techniques out yourself in the following large language model lab, and then go ahead and follow on to the next graded lab, which is all about constructing confidence intervals and performing hypothesis tests for diamond prices. It's going to be really exciting, I know you'll enjoy it, and I'll see you once you're done in the next module, which is all about regression analysis.