Confidence Intervals: Crash Course Statistics #20

Crash Course: Statistics
https://youtube.com/watch?v=yDEvXB6ApWc
https://nerdfighteria.info/v/yDEvXB6ApWc

===== (00:00) to (02:00) =====

Hi, I'm Adrienne Hill and welcome back to Crash Course Statistics. So, last week I ordered a pair of gold (?) pants with DFTBAQ embroidered on them. The delivery guy said they could come by the next day at exactly 11 AM on the dot! Just kidding. That never happens. Instead of an exact time, the pants guy gave my a range of times. He said they'd be there some time between 8 AM and 2 PM, a lot of anticipation. We've focused a lot on point estimates, like the mean, which are our best guesses but we can give ourselves a little more wiggle room. Let's talk about confidence intervals.

It's useful to give pregnant mothers a due date when they're children will most likely be born, but it might be more accurate to say doctors expect the baby to come around the due date, not exactly on it. And, when pollsters claim that a candidate will get around 30% of the vote plus or minus 2%, we can represent the around part with a confidence interval. You may have seen the term "confidence interval" paired with a percentage like 95%.

A confidence interval is an estimated range of values that seem reasonable based on what we've observed. It's center is still the sample mean, but we've got some room on either side for our uncertainty.

So, when the delivery guy says my pants are coming between eight and two, he's reflecting his uncertainty, the very large frustrating uncertainty about when he'll be here. For example, a dentist thinks the mean number of cavities the average person has in a five year span is greater than one and wants to calculate a 95% confidence interval to see if there's evidence that he's right. He rounds up a random sample of 100 patients from around the country and finds this group has a mean of three cavities with a standard deviation of 0.5 cavities. The way we chose that confidence range is related to the distribution of sample means. The dentist's estimate of the sampling distribution looks like this. And instead of grabbing just the mean the dentist can include a range of the most common 95% of the sample means that we expect from this estimate of the distribution of sample means.

===== (02:00) to (04:00) =====

So now we have a 95% confidence interval from 2.902 to 3.098 cavities. Giving a range of numbers instead of just an estimate for the mean better represents the fact that there's some uncertainty and variation and we estimate population parameters, like the mean proportion or aggression slope from a sample. The interpretation of this confidence interval is a bit more complex. To understand what a confidence interval really is, we have to ask ourselves 'what if'. If the dentist sample was taken again, we wouldn't expect that the mean and standard deviation of cavities would be exactly 3n.5, they'd probably be a little different. Which means that our 95% confidence interval would be different than the one we got before, and if we did it 100 more times with the same sample size, we'd get 100 slightly different confidence intervals.

The 95% in a 95% confidence interval tells us that if we calculated a confidence interval from 100 different samples, about 95 of them would contain the true population mean. Our "confidence" is in the fact that the procedure of calculating this confidence interval will only exclude the population mean 5% of the time. That definition implies that it is possible that the confidence interval that we created doesn't include the true population mean. We have no way of knowing for sure, but the confidence intervals usually contain the true population mean.

Now that we know what a confidence interval is, it might be useful to calculate it. A 95% confidence interval is the range that contains the middle 95% of the values of our estimated sampling distribution, and to get that range, we can use a z-score. A z-score tells us the distance between the mean of a distribution and a data point in standard deviations. Previously we've used z-scores to help us find our percentiles, and we want the middle 95% of the data, so we want our cutoffs to be at the 2.5th percentile and the 97.5th percentile, so that 95% of the values are within our range and 5%, 2.5% on either side, are not.

===== (04:00) to (06:00) =====

To calculate the 95th percent confidence interval for a sample of 49 chocolate cakes with a mean of 3,000 calories and a standard deviation of 500 calories, we can use a z-score of 1.96, which we got from a table to calculate the 97.5th percentile, and a z-score of -1.96 to calculate the 2.5th percentile, but we now need to turn our z-scores back into calorie values. To do so, we multiply by the standard error, 71.4 calories, and add the mean of 3,000 calories to get the 95% confidence interval for our sample. We think it's likely that the real population mean for the number of calories in a chocolate cake is in that range, though we're not sure. What we can have confidence in is that if we're in a situation where we're constantly taking samples like this and we assume that the true mean is inside of every confidence interval, we'll only be wrong 5% of the time.

For example, a gummy worm factory periodically checks whether their bagging machines are calibrated correctly, so each week, they take a sample of 100 bags of gummy worms, measure the mean weight and standard deviation, and calculate a 95% confidence interval. They use the confidence interval to make a decision about whether to pay an expensive repairman to come repair the gummy worm bagging machine. They want their bags of gummy worms to have around 10 oz of gummy treats and decide that as long as the confidence interval contains 10 oz, their ideal weight, they'll assume their machine is fine. Decisions based on their confidence intervals will lead them to call an unnecessary repairman only 5% of the time.

Many researchers use confidence intervals to see if they contain a certain value of interest. A researcher may want to know if a certain number of calories in a cake is plausible. If the sample value were to fall within their confidence interval, it would seem possible, but it's not possible to rule out even if it's outside the interval, because you don't know if you got the 95% of confidence intervals that contain the true mean or the 5% that don't.

===== (06:00) to (08:00) =====

You don't always need to use a confidence interval of 95%. We can calculate other confidence intervals too. You can calculate a 99% confidence interval, or really any percentage confidence interval, but if you try to calculate 100% confidence interval, it will always be negative infinity to positive infinity, which just shows that the larger you want your confidence percentage to be, the wider your interval will be. You can be more hopeful that your confidence interval contains the true population mean, but it's not gonna be that helpful.

So there's a balancing act going on. You want a confidence interval that's narrow enough to be useful but wide enough that the true population mean will usually be inside a confidence interval of that percent. We can't always have large samples. It's often the case there's not enough time or money to collect hundreds of data points to calculate a confidence interval. With small sample sizes, the distribution of sample means isn't always exactly normal, so we often use a T-distribution instead of a z-distribution to find out where the middle 95% of our data is.

The T-distribution, like the z-distribution, is a continuous probability

distribution that's unimodal. It's a useful way to represent sampling distributions. The T-distribution changes its shape according to how much information there is. With small sample sizes, there's less information so the T-distribution has thicker tails to represent that our estimates are more uncertain when there's not much data. However, as we get more and more data, the T-distribution becomes identical to the z-distribution.

Generally, sample sizes that are greater than 30 are considered large enough, because scientists generally believe that sampling distributions where the sample is 30 or more are close enough to normal. The 30 is just an arbitrary cut-off, like .05. However, when we're estimating population proportions, like the proportion of people who are colorblind, the general rule is that your sample size needs to be big enough so that, on average, you'd expect to get at least 10 colorblind and at least 10 non-colorblind people.

===== (08:00) to (10:00) =====

For similar reasons, most people consider that close enough. Since about 8% of males are colorblind, if I only had a sample of 50 males, on average, I'd expect around four males per group to be colorblind. So my sample size wouldn't be quite big enough to assume its normal. Instead, I'd use the almost normal T-distribution. If a drug that's being developed claimed to reduce the proportion of colorblind males born to mothers who took it, we could take a sample of 50 male infants to see if the proportion of colorblindness is different from 8%. Though colorblindness isn't usually life-threatening, it can be inconvenient, so you decide to calculate a confidence interval to see if the drug is likely to be effective. After randomly selecting 50 male infants from mothers who took the drug, you calculate the sample proportion of colorblind infants, which is 6% and calculate the distribution of sample proportions, which has a mean of 6%, the same as the sample mean, and a standard error of .033.

Since our sample size isn't big enough to assume that the distribution of sample proportions is shaped like the z-distribution, we can use the T-distribution to calculate the range of our 95% confidence interval. I mentioned before that the T-distribution shape changes with how much data it has. We'll talk more in detail later as to how to choose the right T-distribution, but for now, we're gonna use this one.

While T-score tables do exist, it's often easier to have a statistical program calculate the T-values that correspond to the 2.5th and 97.5th percentiles, since there are many different T-distributions. Your computer tells you that the T-values corresponding to those percentiles in this case are 2.01 and -2.01, and to convert to a raw score from a T-score, we again use this formula, just with a T-score instead of a z-score. Our confidence interval for proportion of colorblind males is -.6% to 12.63%.

===== (10:00) to (12:00) =====

8% is inside our confidence interval, so it's not too much of a stretch to think that 8% could be the true population proportion, even though we only observed a sample proportion of 6%. Based on this confidence interval, we don't have any evidence to conclude whether this medicine is effective or not, so since the company researching the drug is pretty cautious, they decide not to go ahead with it.

One place you may have seen confidence intervals in the wild is in the news during election season. When newscasters report results from exit polls, they'll usually say something like Candidate A is tracking at 64% with a margin of error of 3%, or you may see a chart like this. The margin of error is usually telling you how far the bounds of the confidence interval are from the mean, and is represented by this part of the confidence interval formula. The margin of error, just like a confidence interval, reflects the uncertainty that surrounds sample estimates of parameters, like the mean or a proportion.

If a poll show that a presidential candidate was tracking at 64% of the vote, +/- 3%, we shouldn't be surprised if it turns out that the true vote was 61%, since that's within the margin of error. You can think of values inside the margin of error, or confidence intervals, as values that might be reasonable estimates of the true population parameter. Confidence intervals quantify our uncertainty.

They also demonstrate the trade-off of accuracy for precision. 100% confidence interval will always contain the true population mean, but it's useless. We have to sacrifice a little bit of accuracy in order to gain more precision. A 99% confidence interval will give us a more useful range since it won't be infinitely long, but it's now possible that our confidence interval won't contain the true mean, and you've probably encountered this trade-off in your daily life.

So you're running a marathon, like everybody does, and you want to load up your iPhone with music, but you don't know how long you're gonna take. You could buy 150 songs on iTunes, which is expensive, or you could buy only 70 and have a chance of running out of music.

===== (12:00) to (13:02) =====

You increase your risk of not having enough music but then again, you're saving yourself from having to buy 80 extra songs, and maybe it's time to consider a streaming service? Confidence intervals demonstrate this delicate balancing act and help us understand how to hit the sweet spot of information vs. accuracy.

Thanks for watching, I'll see you next time in my gold (?~12:22) pants.

CrashCourse Statistics is filmed in the Chad and Stacey Emigholz Studio in Indianapolis, Indiana and it's made with the help of all these nice people. Our animation team is Thought Cafe. If you'd like to keep CrashCourse free for everyone, forever, you can support the series at Patreon, a crowdfunding platform that allows you to support the content you love. Thank you to all our Patrons for your continued support. CrashCourse is a production of Complexly. If you like content designed to get you thinking, check out some of our other channels at complexly.com. Thanks for watching.