How P-Values help us test hypothesis: Crash Course Statistics #21

Crash Course: Statistics
https://youtube.com/watch?v=bf3egy7TQ2Q
https://nerdfighteria.info/v/bf3egy7TQ2Q

Hi, I'm Adriene Hill, and Welcome back to Crash Course, Statistics.

We've been talking a lot about how to tell whether two groups are different like whether there's more car accidents on rainy days than snowy days. or whether the IQ of university students is actually different from the population. Today, we're going to start a conversation about statistical inference, which tells us how we can go from describing data we already have to making inferences about data we don't have.

INTRO If you've watched any of the other videos in this series, you've heard a lot about uncertainty. It comes up endlessly in statistics. And uncertainty is at the core of what Inferential Statistics is about: making decisions about ideas, or hypotheses.

I might be interested in whether listening to Mozart while doing calculus homework improves my calculus grades. But I need to test my hypothesis, I can't just have an idea and claim it's correct without any evidence. One thing we need for sure, is data.

So we could randomly sample two groups of 25 people and make half of them listen to Mozart and half to do their homework in silence. We collect their calculus grades and see that those who listened to Mozart scored on average 3 points higher than those who didn't. So Mozart's good.

Problem solved, break out Sonatas, right? Unfortunately, no. We've seen that sample parameters like the mean are just estimates of the mean of the population that they are taken from.

The sample mean score of the Mozart group is higher. But we don't have sufficient evidence that the population mean of Mozart listeners is higher than those who did their work in silence. We may have gotten an especially high sample mean that isn't close to the true population mean.

So we need a way to test our hypothesis while taking into account the random variation of sample means. In theory, one way you could test a hypothesis or model is by how well it predicts the data you got. For example, you and your best friend really love giraffes, and you've spent a lot of time watching them at the zoo and drawing sketches of them.

So you both have a hypothesis about the average number of spots a baby giraffe has, but they're slightly different. You think that baby giraffes have an average of 175 spots, with a standard deviation of 50 spots, and your best friend thinks that baby giraffes have an average of 209 spots with a standard deviation of 45 spots. With the permission of your local zoo, of course, you begin to collect a random sample of baby giraffes and count how many spots they had.

Your sample of 25 baby giraffes had a mean of 200 spots. Now that you have data, you can use it to evaluate which one of you is more likely to be right. Both you and your friend have a model or idea about what the population distribution of baby giraffe spots is.

If you're right, then the sampling distribution of all the possible sample means we could get looks like this: (RED in chart) And the distribution of sample means for your friend's model looks like this: (black in chart) Let's look at where our sample mean of 200 lies on both of these distributions. You can see that you're more likely to see a mean of 200 spots under your friend's hypothesis than yours. If your model were correct, a mean of 200 spots is pretty rare...it's in the top 1.2% most extreme values we'd expect to see, whereas in your friend's model, a mean of 200 spots is only in the top 32%, which means it's pretty common that we'd see sample means around 200 if your friend's model was correct.

But we don't always have predictions that are as specific as you and your friend's predictions about baby giraffe spots. We might have a more general hypothesis, like that the average number of baby giraffe spots is more than 200... but that's all that you really know. In situations like these, one common method of testing ideas is Null Hypothesis Significance Testing (NHST) You have a hypothesis.

That people with a certain gene, we'll call it gene X, eat a different amount of calories than the general population. Null Hypothesis Significance testing asks you to test a different hypothesis--which says there is no difference or effect of this gene. And we'll see how well this null hypothesis predicts the data we've collected.

In this case the null hypothesis--or null model-- is that the population mean caloric intake for people with gene X is actually 2,300, the same as the regular population. If the null hypothesis is found to be infeasible, we can "reject" it. We can represent this hypothesis like this: This might seem like a pretty round about way to test your theory that people with gene X eat differently, and that's because it is.

Null Hypothesis Significance testing is a form of the reductio ad absurdum argument which tries to discredit an idea by assuming the idea is true, and then showing that if you make that assumption, something contradictory happens. For example, you can use reductio ad absurdum to show that there is no largest positive integer. Let's assume there is a largest positive integer.

We'll call it AB for "absurdly big". Now add one to AB. shoot. That would be a larger positive integer...which would be absurd since AB is the largest.

Therefore, by reductio ad absurdum, there is no largest positive integer. By the way, if this kind of argument sounds familiar, it might because reductio ad absurdum is like proof by contradiction. Let's test the null hypothesis for our our gene X case.

First, we assume that the mean number of calories eaten by people with gene X is 2,300, just like the regular population. If we can show that this assumption makes something "absurd" happen, then we can "reject" the idea that it's true. With data from 60 people with gene X, we see that the mean number of calories eaten was 2,400 with a sample standard deviation of 500 calories.

We have to ask how rare or "absurd" it would be to get a sample mean that is this far away from our assumed mean of 2,300. Essentially, we imagine that we take a random sample of 60 people with gene X over and over and over again and calculate the mean. Then we ask how many times out of all those experiments, do we get a sample mean that's as far away from 2,300 as our actual sample mean of 2,400 is.

Even if you haven't heard of the term null hypothesis significance testing, you may have heard of p-values which have been covered everywhere from academic journals, to Buzzfeed articles. A p-value answers the question of how "rare" your data is by telling you the probability of getting data that's as extreme as the data you observed if the null hypothesis was true. If your p-value was 0.10 you could say that your sample is in the top 10% most extreme samples we'd expect to see based on the distribution of sample means.

If we assume that the null hypothesis is true, and the mean caloric intake of people with gene X is 2,300 with a standard deviation of 500 calories, the distribution of sample means will look like this, and tells us which means we expect to see and how often we expect to see each of them. Sample means around 2,300 are most common, but we'll also often see sample means a little bit further away. We

can use this distribution to calculate our p-value.

This is similar to how we compared the likelihood of 200 giraffe spots in you and your friend's models, but with only 1 model this time. Here's our sample mean of 2,400 on this graph. Only about 8.99 percent of the possible sample means are higher than 2,400.

So it's not that unlikely that we'd get a sample mean that's this high if the true population mean was 2,300 calories. This is called a one-sided p-value since it only tells us the probability of getting a sample mean that's higher than 2,400. Often when we ask scientific questions like "Does this medicine have a different level of efficacy than the existing treatment?" we don't know which direction the effect will be in.

The new medicine might be better...or it might be worse. Gene X'ers might eat more, or they might eat less. Because of this--and a few other reasons we'll talk about later in the series--p-values are often two-sided, meaning that we look at how far away a value is from the mean, regardless of if it's higher or lower .

This allows us to reject the null hypothesis if our value is significantly higher than the mean, or if the value is significantly lower than the mean. Because the distribution of sample means is symmetrical, if 9% of the samples of caloric intake are higher than a mean of 2,400, about 18 percent of sample means for calories would be as far away or further from the population mean than 2,400 is in either direction. In other words, a two-sided p-value is a measure of how extreme your sample mean is, because it tells you how often you'll get a value that's as or more extreme than the one you got.

The smaller your p-value is, the more "rare" it would be to get your sample just by random chance alone if the null is true. In our example, we learned that if we assume that there is no effect of gene X on caloric intake, then there would be an 18% chance, about 1 in 5, that we'd see a sample like this just because of the random variation of samples. To finish our attempt at reductio ad absurdum, we have to decide whether this sample is "absurd" or "extreme" enough to lead us to believe that this sample probably isn't from the null distribution.

But that decision isn't always an easy one to make... It's not clear how "rare" or "absurd" a sample needs to be before I decide to "reject" the idea that the sample was taken from a population that has the null distribution. Especially since we don't have another distribution to compare it to, like we did with the giraffes.

Our p-value of 0.18 tells us that if we took a sample like this over and over, about 1 out of every 5 times we'd get a sample with a mean caloric intake that's further from the mean than 2,400 calories is. 1 in 5's not bad...but a 1 in 20 chance might be better. And 1 in 100 better than that. Some statisticians see a p-value as a continuous measure of evidence.

A p-value of 0.18 like ours might be considered pretty weak evidence that our sample isn't taken from the null distribution. But it's better than 0.19, which is in turn better than 0.20 and so on. However, in Null Hypothesis Significance Testing, p-values need a cutoff.

We could set a cut of at 0.05 and say that a p-value that is less than 0.05 is sufficient evidence to allow us to "reject" the idea that the null hypothesis is true. When we can reject the null hypothesis, we consider our result to be "statistically significant", which is basically a phrase that just means "unlikely due to random chance alone". As we'll see later on, it doesn't always mean that it should be "significant" or meaningful to you.

A cutoff of 0.05 means that we want our sample value to be at least in the top 5% of most extreme values in our distribution before we consider the value evidence against that hypothesis. And any p-value less than the 0.05 cutoff counts. 0.049 leads to the same conclusion as 0.0001. Both cause you to reject the null hypothesis.

The current scientific consensus in most fields is that your cutoff--or alpha--should be 0.05. But there's huge disagreement in the field of statistics about whether 0.05 is appropriate, and we're going to dive into later. In the meantime I'm going to get 24 more giraffes so I can compare my model with my friends.

Thanks for watching. I'll see you next time.