P-Value Problems: Crash Course Statistics #22

Crash Course: Statistics
https://youtube.com/watch?v=PPD8lER8ju4
https://nerdfighteria.info/v/PPD8lER8ju4

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

To recap from last time, P-values tell us how "rare" something is. So far, we've been using that information to decide whether or not our hypotheses are reasonable, and using P-values to reject or fail to reject an idea.

Today, we're going to explore p-values a little more and talk about the logic of p-values and some of the problems that come up. INTRO Remember, to calculate a p-value, we first assume that the null distribution is the true distribution our sample was taken from. Then we calculate how often we'd see a value that is at least as extreme as our observed value.

So in probability terms, the p-value is the probability of getting a sample as or more extreme than ours, given that the null hypothesis is true: So all the values that we see in the sampling distribution are means we could actually get if the null hypothesis was true. For example, let's say the average cat weigh 10lbs (or 4.5 kg). We might want to calculate the probability of getting a group of 30 randomly selected calico cats who have an average weight of 11 lbs (or 5 kg) if calico cats have the same average weight as the whole population of cats.

The first issue is if, in real life, there is no connection between two things like fur color and weight --we still might get samples of calicos, mackerel tabbies, or tortoise shells that are different enough to cause us to "reject" the null hypothesis that there is no difference. Our alpha tells us how often this will happen. Let's say our hypothesis is that the reaction time of older professional chess players is different from the reaction time of the general population of professional chess players.

Even if older chess players are the same as their colleagues, if we ran this study over and over, we'd expect that 5% of the time, we'd mistakenly reject the null if it were true. This is one reason why p-values are pretty controversial in the statistical community right now. Not everyone agrees that a p-value less than 0.05 is sufficient evidence to reject the null hypothesis.

In fact, some studies that look at incredibly important things like new medications, have already decided that an alpha of 0.05 isn't low enough. They want p-values lower than 0.01 so that if the null hypothesis is true, they'll only mistakenly reject it 1% of the time. Still others argue that 0.005 is the better cutoff.

As you can see, the standard cutoff is arbitrary. Null Hypothesis Significance Testing requires that we draw a line in the sand somewhere, but it isn't clear where. Arguments have been made that we can have different p-value cutoffs--our alphas--depending on the situation, and that scientists should be allowed to justify their reasons for picking a certain cutoff.

But on the whole, many fields that regularly use p-values have some sort of "official" cutoff that they use. The second, related issue is that a p-value tells you how "extreme" your data would be if you assume the null hypothesis is true. But when you really think about it...that's not what we want to know.

We want to know whether the null is correct, or at least probably correct. In other words, the probability of the null, given that we've seen our data. A p-value of 0.02 in a study on cancer rates in mice tells you that if your new drug didn't work and there was no difference between the cancer rates of mice on and off the drug, then you'd only expect 2% of identically run studies to produce a difference in cancer rates that's as or more extreme than the one you just observed.

But we can't use these p-values alone to tell us about the probability of the null being true or false, even though it can be tempting to think we can. One common misinterpretation of a p-value is that it can tell you the probability that the null hypothesis is true. For example, if a random sample of tuna has a 10% higher mercury content than a random sample of mahi-mahi, it would be incorrect to say that a p-value of 0.02 in this case means there's only a 2% chance that the null hypothesis is true.

This is an especially tempting misinterpretation because it feels like it maybe should be true, but again, when we calculate our p-value, we've already assumed for a moment that the null hypothesis is true and that any sample differences we see are actually due to just random sampling variation. If our p-value for the chess study was 0.01, that means that we already assumed older chess players were the same as the general population of chess players, so 0.01 can't tell us much about the probability that older chess players are the same as their colleagues. That would be like saying "assuming that grass is green, what's the probability that grass is green?" It just doesn't make much sense.

Similarly, p-values can't tell you the probability that you've made an error, given that you rejected the null. Again, this is because p-values don't tell you about the probability of the null being true or false. If you've rejected the null hypothesis--like that drinking orange juice is not associated with higher levels of cavities than drinking coffee--either you did so correctly, because there really is a difference between cavities in OJ and coffee drinkers, or you did so mistakenly because there really is no discernible difference.

But p-values--since they assume the null is true--don't tell you how likely either of these options is. Ronald Fisher--one of the first proponents of Null Hypothesis Significance Testing wrote that: " In general tests of significance are based on hypothetical probabilities calculated from their null hypotheses. They do not generally lead to any probability statements about the real world, but to a rational and well-defined measure of reluctance to the acceptance of the hypotheses they test." In other words, getting a p-value of 0.04 doesn't mean that there's a 4% chance that the null hypothesis is true.

The probability we want to know is the opposite conditional probability from what a p-value gives you. We want to know the probability of the null hypothesis given that we got this data. But that's not what we get.

From the p-value we get the Probability of the data given the null. For example, we calculate P(data older chess players are the same as population of chess players) but we wish we could calculate P(older chess players are the same as population of chess players

data). And while all the same pieces are there, they're not the same. This is made even more clear when you realize the probability of being a child, given that you're at Chuck E Cheese is NOT the same as the probability of being at Chuck E Cheese, given that you're a child.

This is one reason why p-values are so perplexing. They don't give us the probability that we truly want. There are some statistical methods that will give you the probability of a hypothesis given the data, and we'll talk about those later.

A third issue is that if you reject the null, you still don't have much information about the alternative. When the data is pretty improbable under the null hypothesis, we reject the null and accept the hypothesis that the data came from another distribution that is not the null distribution. We call this the alternative distribution, and the hypothesis that goes with it, the alternative hypothesis.

If we reject the null that Mrs. Smith and Mr. Kennedy give the same

amount of homework each week, then the alternative is that they don't give the same amount each week.

But, we don't know whether the difference is by 30 minutes, 25 minutes...45 minutes. Or, for example,we might want to know whether people who were primed with the words "Elderly, Florida, and Retired" walked more slowly than the average person who takes 10 minutes to go around our office building, with a standard deviation of 1 minute. We think they will.

We take a sample of 50 people, primed them, and set them off. Their mean time is 10.5 minutes, which corresponds to a p-value of 0.00036. We already decided beforehand to make our alpha (or predetermined cutoff) 0.005.

So our p-value which is less than 0.005 allows us to reject the null hypothesis...in this case that the people primed with words about being old take a mean of 10 minutes to walk around the building. But what now? While we've rejected the null hypothesis that the primed subjects take a mean of 10 minutes.

The alternative hypothesis is just that their mean isn't 10. Our p-values can't tell us anything else. A fourth common issue for p-values is more about how we interpret "non-significant" p-values.

If our p-value isn't lower than our predetermined cutoff, our alpha, we "fail to reject" the null hypothesis. Notice that we say fail to reject, not accept. Null hypothesis testing doesn't allow us to "accept" or provide evidence that the null is true, instead we've only failed to provide evidence that it's false.

Consider this: Your best friend makes the statement, "there are no black swans in China". You think she's wrong, so you go to China and you look at a bunch of swans, and none of them are black. You may, at a certain point, decide that you've seen SO many swans that if there were black swans in China, it's unlikely that you wouldn't have seen one yet.

But you can't PROVE there are no black swans until you've seen EVERY. SINGLE. SWAN.

Just like you can't prove the null is true--that there's no relationship between two variables, you can only show that you didn't find any evidence it's false. The absence of evidence is not the evidence of absence. "failing to reject" the null hypothesis doesn't mean that there isn't an effect or relationship, it just means we didn't get enough evidence to say there definitely is one. If we looked whether bees produce more honey when it's warm than when it's cold, we could look at some data and calculate a p-value of 0.25.

Since we decided beforehand that our alpha would be 0.01, we fail to reject the null hypothesis that bees produce the same amount of honey in hot and cold seasons. But we can't conclude that there is no difference or even that it's unlikely that there's a difference. We can only conclude that we didn't find any evidence of one.

Since null hypothesis significance testing is often the first type of statistical inference that people learn, it can seem pretty limiting to know that you can't provide good evidence for the null hypothesis being true. In some cases the null hypothesis might be what you actually want to demonstrate. For example, say there are two groups: people who play a souped up, bells and whistles version of a cognitive training game and those who plan a less fancy version of the game.

If these two groups have the same amount of improvement in cognitive abilities (which is our null hypothesis says) that's really interesting. It means that researchers could feel comfortable using whichever version of the game that they want. If playing the fancier, more aesthetically pleasing game made people with strokes, or children with learning differences more likely to play it, researchers would know that's fine.

They wouldn't have any concerns that the bells and whistles would detract from the cognitive benefits. P-values can be perplexing. But they give us insight into how to make decisions about data.

They also remind us that people's perception of evidence can be arbitrary. What you consider sufficient evidence might not be enough to convince someone else. When you read about the results of scientific studies, you can see the alpha they used and decide if you think it's a stringent enough criteria.

More than that, though, we now know what p-values are and how to interpret them. This helps us compare the logic of null hypothesis significance testing with how we normally reason about the world. Thanks for watching, I'll see you next time.