Playing with Power: P-Values Pt 3: Crash Course Statistics #23

Crash Course: Statistics
https://youtube.com/watch?v=WWagtGT1zH4
https://nerdfighteria.info/v/WWagtGT1zH4

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

In the last episode we talked about Null Hypothesis Significance testing and p-values and how these two things help us make decisions about things we care about. Like whether babies who drink non-dairy milk are more likely to have allergies, or whether the number of hours you spend watching home makeover shows tends to increase with age.

We don't always come up with the right answer, even if it seemed reasonable. We want to limit our errors as much as possible. Today we'll talk about when and why we might get it wrong.

INTRO In the last episode we briefly touched on "rejecting" the null hypothesis. P-values tell us how "rare" or "extreme" our sample data would be if it really did come from the null distribution. Null means nothing so null hypotheses tend to say that there's no effect, or nothing's going on.

For example, for whether babies who drink non-dairy milk are more likely to have allergies, the null hypothesis (or H0) would be that there is no difference in proportion of babies with allergies between babies who drink non-dairy milk, and those who do not. In the case of home makeover shows, the null hypothesis might be that there's no relationship. So the regression slope--or coefficient--between number of home makeover shows watched and age would be 0: By looking at this slope, we can see it's not exactly flat, but we don't know whether this slope is due to a real relationship, or just random variation.

When we get low p-values, we "reject" the null hypothesis because we've decided that our data would be pretty rare if the null was true since the probability of getting data as or more extreme than ours is below our alpha level. That's option 1. Option 2 is that our p-value is not lower than our pre-selected cutoff which means that we "fail to reject" the null hypothesis.

So, we've narrowed it down to two decisions: we can either reject, or fail to reject the null. The null can either be true, or not true. This means that there are four possible situations: either you correctly reject the null, mistakenly reject the null, correctly fail to reject the null, or mistakenly fail to reject the null.

In two of these situations we make the correct decisions, and in the other two, we'd have made an error. The first error is called a Type I error, which is rejecting the null, even if it's true. It can therefore only happen if the null is true.

Say we've decided that our alpha level is 0.05, so we'll reject the null if our p-value is smaller than 0.05, which means that our sample is in the 5% most extreme values we can expect to get if the null hypothesis were true. So, if the null is true, 5% of the time, we'll still reject it mistakenly, just because we happened to get a rare value. The red shaded region represents all the values from the null distribution that would cause us to decide to "reject" the null, even if it was true.

Since our type 1 error rate is equal to alpha, we get to choose exactly how often we are willing to make Type 1 errors when we choose our alpha. We control our Type I errors by explicitly deciding how often we'll make them. We could also make an error by failing to reject the null hypothesis when it actually is false.

In order for the null hypothesis to be false, some other, alternative, hypothesis must be true. We mentioned in the last episode that we don't actually know any specifics about which hypothesis is correct when we "reject the null", it could be anything. But we can estimate which distribution might be correct, we'll show it outlined in gray, this helps us to compare two distributions instead of just looking at one.

We estimate the alternative distribution based on the mean and standard deviation of our experimental group. The sample mean is our best guess at what he effect size is, so we often use that if we're estimating the alternative after we've collected our data. But sometimes we want to estimate it before we collect data, in which case we use the sample estimates from other, related studies.

We're assuming the Alternative (Ha) distribution looks like this. Our cutoff line is still in the same place; it marks the cutoffs that tell us where the 5% most extreme values are. Any value we get that is to the right of the line causes us to "reject the null" and any value to the left of the line causes us to "fail to reject the null".

The cutoff value doesn't change depending on whether H0 or HA is true. So, if the alternative is true, we still might fail to reject the null if we happen to get a value that is to the left of the cutoff. The blue shaded region shows you the values where we'll make this Type II error.

Just like the rate of Type I errors is equal to alpha, the rate of type II errors is equal to Beta. Since we're only estimating what the alternative distribution looks like, we can't know what Beta is for sure, but again we can estimate it by using our cutoff (alpha) and our best estimates of the shape and position of our alternative distribution to find the approximate area of the shaded region. There's often a trade off between Type I and Type II errors.

Type I errors are essentially False positives: we think we've detected an effect, but there isn't one. And Type II errors are False negatives: there was an effect, we just didn't see it. And while both of these mean we were wrong, there's a lot of times where we may prefer one type of error over the other.

Take smoke alarms. While the sound of the smoke alarm going off is annoying, there's not a lot of cost to having a false positive--or type I error. All you have to do is press a button to reset it.

There is however a huge risk if your smoke alarm does not go off when there really is a fire. For this reason, fire alarms tend to favor having type I errors over type II errors. Which is why sometimes particularly long, hot showers can cause them to go off.

Better safe than sorry. But yeah, turning off an alarm when you're naked and wet. Not fun.

Think about someone in your life who is constantly worried, they operate on the assumption that Type I errors--thinking there'll be an issue when there won't be--are preferable to Type II errors--not preparing for a problem when there really could be one. You can see on this graph that if we assume our null distribution is here, and the alternative is here, then moving the cutoff threshold to the right will cause us--all other things being equal--to have fewer type I errors. But we'll have more type II errors since less of the null distribution is in the "reject" region, and more of the alternative distribution is in the "fail to reject region".

And the opposite happens if we move our cutoff threshold to the left. We'll have more False positives since more of the null is in the "reject" region, but fewer False Negatives because less of the alternative distribution is in the "fail to reject" region. If the error types hard to keep straight, think of the Boy who cried wolf.

In that story the villagers first made a type I error (thinking there was a wolf when there really wasn't), but by the end--and to the detriment of the little boy--they made a type II error: thinking there WASN'T a wolf when there really. Sometimes we do make the right decision and there are two ways to be right: either the null

hypothesis is true and we fail to reject it, or the null hypothesis is false and we do reject it. If the null is true, you'll reject it 1 - alpha of the time.

When alpha is 0.05 that means that when the null is true, we'll correctly fail to reject 0.95 or 95% of the time. If the null is false and the alternative is true, we'll correctly reject the null 1-Beta of the time. If Beta--the proportion of times we will fail to reject the null even though it's false--is 10%, then we'll correctly reject the null 90% of the time.

This proportion (1-Beta) is called our statistical power. As the name suggests, statistical power is really important and something that we want. I mean, it's a power.

I want powers! Statistical power tells us our chance of detecting an effect if there is one. Imagine we design a study to look at whether fish oil makes cat's hair shinier and it has 80% statistical power.

That means we know that if there really is an effect of a certain type of fish oil and if we ran the same experiment multiple times with different samples of cats, the data from 80% of the experiments will lead us to make the correct decision and reject the null hypothesis that fish oil has no effect. This is important because the whole reason that we do experiments is to see whether there's an effect. We don't just test whether fish oil makes cat's hair shiner just for fun, we want shinier cats!

Statistical power tells us about our ability to detect these effects if they exist. It would be a waste of time and money to run an experiment on whether people who play video games have quicker reaction times than those who don't if we only have an estimated 20% power, because that means that even if there gameplay effects reaction time, we often wouldn't be able to tell. Experiments cost money, so if you're going to go through the process of growing cells in a petri dish, or of giving cats fish oil you want to be relatively confident you'll be able to detect an effect if there is one.

Visually we see that statistical power is affected by how much the null and alternative hypothesis distributions overlap. The more they overlap, the less statistical power we'll have, because less of the alternative distribution will be to the right of the cutoff. There are two main ways to get the two distributions to overlap less.

Either you can move them further apart, or you can make them skinnier. The distance between the mean of the two distributions represents something called "effect size". If we're looking at the difference between two groups--like level of neuroticism between cat people and dog people--effect size tells us how big the difference in neuroticism is between the two groups.

If effect size is large, the groups are further away from each other, if it's small, they're pretty close. If two things are really different from each other, it's easier to tell them apart. Say we're researching whether the amount of time people spend in the sun leads to more freckles.

If one group that spent 10 minutes in the sun led to an average of 5 new freckles over the body, it'd be a lot harder to tell than if 10 minutes in the sun led to an average of 500 new freckles. Unfortunately, effect size is largely out of our control. Researchers can't magically change the efficacy of a drug, or the difference in heart rate between people who do kickboxing and people who do Crossfit.

We can also make our distributions overlap less by making them skinnier. And remember, the null and alternative distributions are just sampling distributions. We've seen that as you increase the size of your samples, the distribution of sample means gets thinner. And all other things being the same, they overlap less and we have more power to detect an effect. This shrinking represents the fact that in general, the more data we have, the more information we have. Thankfully we can change sample size.

It might be a pain to sample more people, feed more cats more fish oil, or measure more ocean temperatures, but at least it's within our control, unlike effect size. And that's just what researchers do. We already mentioned that if we're going to take the time to run an experiment or do a study... we want to make sure it has sufficient power to detect any effects out there, and since almost everything else is out of our control, scientists will increase their sample size to get sufficient statistical power to detect these effects.

Across many fields it's considered sufficient to have 80% statistical power or more, and often when researchers are designing studies, they'll decide how many subjects they need based on estimates of effect size and power. So now you're playing with power…and in the next few episodes we'll talk a lot more about exactly when and how you can use p-values, and also some completely different methods for testing ideas. Thanks for watching, I'll see you next time.