# Test Statistics: Crash Course Statistics #26

Crash Course: Statistics
https://youtube.com/watch?v=QZ7kgmhdIwA
https://nerdfighteria.info/v/QZ7kgmhdIwA

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

Sometimes random variation can make it tricky to tell when there are true differences or if it's just random. Like whether a sample difference of $20 a month represents a real difference between the average rates of two car insurance companies.

Or whether a 1 point increase in your AP Stats grade for every hour you study represents a real relationship between the two. These situations seem pretty different, but when we get down to it, they share a similar pattern. There's actually one idea, which--with a few tweaks--can help us answer ALL of our "is it random...or is it real" questions.

That's what test statistics do. Test statistics allow us to quantify how close things are to our expectations or theories. Something that's not always easy for us to do based on our gut feelings.

And test statistics allow us to add a little more mathematical rigor to the process, so that we can make decisions about these questions. INTRO In previous episodes, z-scores helped us understand the idea that differences are relative. A difference of 1 second is meaningful when you are looking at the differences in the average time it takes two groups of elite Olympic athletes to complete a 100 meter freestyle swim.

It's less meaningful when you're looking at the differences in the average time it takes two groups of recreational swimmers. The amount of variance in a group is really important in judging a difference. Elite Olympic athletes vary only a little.

Their 100 meter times are relatively close together, and a 10th of a second can mean the difference between a gold and a bronze medal. Whereas non professionals have more variation; the fastest swimmers could finish a whole minute before the slower swimmers. A difference of 1 second isn't a big deal between two groups of recreational swimmers because the difference is small compared to the natural variation we'd expect to see.

Two groups of casual swimmers may differ by 10 or more seconds, even if their true underlying times were the same, just because of random variation. That's why test statistics look at the difference between data and what we'd expect to see if the null hypothesis is true. But they also include some very important context: a measure of "average" variation we'd expect to see, like how much novice or pro swimmers differ.

Test statistics help us quantify whether data fits our null hypothesis well. A z-score is a test statistic. Let's look at a simple example.

Say your IQ is 130. You're so smart. And the population mean is 100.

On average we expect someone to be about 15 points from the mean. So the difference we observed, 30, is twice the amount that we'd expect to see on average. Your z score would be 2.

And you can z-score any normal distribution--like a population distribution. But also a sampling distribution which is the distribution of all possible group means for a certain sample size. You might remember we first learned about sampling distribution in episode 19.

We often have questions about groups of people. Finding out that you're two standard deviations above the mean for IQ is pretty ego boosting, but it won't really help further science. We could look at whether children with more than 100 books in their home have a higher than average IQs.

Let's say we take a random sample of 25 children with over 100 books. Then we measure their IQs. The average IQ is 110.

We can calculate a z-score for our particular group mean. The steps are exactly the same, we're just now looking at the sampling distribution of sample means rather than the population distribution. Instead of taking an individual score and subtracting the population mean, we take a group mean and subtract the mean of our sampling distribution under the null hypothesis.

Then we divide by the standard error, which is the standard deviation of the sampling distribution. So, the z-score--also called the z-statistic--tells us how many standard errors away from the sampling distribution mean our group mean is. Z-statistics around 1 or -1 tell us that the sample mean is the typical distance we'd expect a typical sample mean to be from the mean of the null hypothesis.

Z-statistics that are a lot bigger in magnitude than 1 or -1 mean that this sample mean is more extreme. Which matches the general form of a test statistic: The p-value will tell us how rare or extreme our data is so that we can figure out whether we think there's an effect. Like whether children with more than 100 books in their home have a higher than average IQ.

Historically we've done this with tables, but most statistical programs, even Excel, can calculate this. We can use z-tests to do hypothesis tests about means, differences between means, proportions, or even differences between proportions. A researcher may want to know whether people in a certain region who got this year's flu vaccine were less likely to get the flu.

They randomly sample 1000 people and found that 600 people got the flu vaccine, and 400 didn't. Out of the 600 people who got the vaccine, 20% still got the flu. Out of the 400 people who did not get the vaccine, 26% got the flu.

It seems like you're more likely to get the flu if you didn't get a flu shot, but we're not sure if this difference is pretty small compared to random variation, or pretty large. To calculate our z-statistic for this question, we first have to remember our general form: There's a 6% difference between the proportion of the vaccinated and unvaccinated groups, and we want to know how "different" 6% is from 0%. A difference of 0% would mean there's no difference between flu rates between the two groups.

So our observed difference is 6 minus 0 percent, or 6%. For this question, the "average variation" of what percent of people get the flu is the standard error from our sampling distribution. We calculate it using the average proportion of people who got the flu, and didn't get the flu: If our observed difference of 6% is large compared to the standard error--which is the amount of variation we expect by chance--we consider the difference to be "statistically significant".

We've found evidence suggesting the null might not be accurate. There's two main ways of telling whether this z-statistic, which is about 2.2295 in our case, represents a statistically significant result. The first way is to calculate a "critical" value.

A critical value is a value of our test statistic that marks the limits of our "extreme" values. A test statistic that is more extreme than these critical values (that is it's towards the tails) causes us to reject the null . We calculate our critical value by finding out which test-statistic value corresponds to the top 0.5, 1, or 5% most extreme values.

For a z-test with alpha = 0.05, the critical values are 1.96 and -1.96. If your z-statistic is more extreme than the critical value, you call it "statistically significant". So, we found evidence...in this case...that

the flu shot is working.

But sometimes, a z-test won't apply. And when that happens, we can use the t-distribution and corresponding t-statistic to conduct a hypothesis test. The t-test is just like our z-test.

It uses the same general formula for its t-statistic. But we use a t-test if we don't know the true population standard deviation. As you can see, it looks like our z-statistic, except that we're using our sample standard deviation instead of the population standard deviation in the denominator.

The t-distribution looks like the z-distribution, but with thicker tails. The tails are thicker because we're estimating the true population standard deviation. Estimation adds a little more uncertainty ...which means thicker tails, since extreme values are a little more common.

But as we get more and more data, the t-distribution converges to the z-distribution, so with really large samples, the z and t-tests should give us similar p-values. If we're ever in a situation where we had the population standard deviation, a z-test is the way to go. But a t-test is useful when we don't have that information.

For example, we can use a t-test to ask whether the average wait time at a car repair shop across the street is different from the time you'll wait at a larger shop 10 minutes away. We collect data from 50 customers who need to take their cars in for major repairs. 25 are randomly assigned to go to the smaller repair shop, and the other 25 are sent to the larger shop. After measuring the amount of time it took for repairs to be completed, we find that people who went to the smaller shop had an average wait time of 14 days.

People who went to the larger shop had an average wait time of 13.25 days, which means there was a difference of 0.75 days in wait time. But we don't know whether it's likely that this 0.75 day difference is just due to random variation between customers....at least not until we conduct a t-test on the difference between the means of the two groups. Before we do our test, we need to decide on an alpha level.

We set our alpha at 0.01, because we want to be a bit more cautious about rejecting the null hypothesis than we would be if we used the standard of 0.05. Now we can calculate the t-statistic for our two-sample t-test. If the null hypothesis was true, then there would be no real difference between the mean wait times of the two groups.

And the alternative hypothesis is that the two means are not equal. The two sample t-statistic again follows the general form: We observed a 0.75 day difference in wait times between groups. We'd expect to see a difference of 0 if the null were true.

Our measure of average variation is the standard error. The standard error is the typical distance that a sample mean will be from the population mean. This time, we're looking at the sampling distribution of differences between means--all the possible differences between two groups-- which is why the standard error formula may look a little different.

Putting it all together we get a t-statistic of about 2.65. If we plug that into our computer, we can see that this test statistic has a p-value of about .0108. Since we set our alpha at 0.01, a p-value needs to be smaller than 0.01 to reject the null hypothesis.

Ours isn't. Barely, but it isn't. So it might have seemed like the larger repair shop was definitely going to be faster but it's actually not so clear.

And this doesn't mean that there isn't a difference, we just couldn't find any evidence that there was one. So if you're trying to decide which shop to take you car to, maybe consider something other than speed. And we could do similar test experiments for cost or reliability or friendliness.

You might notice that throughout the examples in this episode, we used two methods of deciding whether something was significant: critical values and p-values. These two methods are equivalent. Large test statistics and small p-values both refer to samples that are extreme.

A test statistic that's bigger than our critical value would allow us to reject the null hypothesis. And any test-statistic that's larger than the critical value will have a p-value less than 0.05. So, the two methods will lead us to the same conclusion.

If you have trouble remembering it, this rhyme may help: "Reject H-Oh if the p is too low" These two methods are equivalent. But we often use p-values instead of critical values. This is because each test-statistic, like the z or t statistics, have different critical values, but a p-value of less than 0.05 means that your sample is in the top 5% of extreme samples no matter if you use a z or t test-statistic - or some of the other test-statistic we haven't discussed like F or chi-square.

Test statistics form the basis of how we can test if things are actually different or what we seeing is just normal variation. They help us know how likely it is that our results are normal, or if something interesting is going on. Like whether drinking that water upside down is actually stopping your hiccups faster than doing nothing.

Then you can test drinking pickle juice to stop hiccups. Or really slowly eating a spoonful of creamy peanut butter. Let the testing commence!

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