T-Tests: A Matched Pair Made in Heaven: Crash Course Statistics #27

Crash Course: Statistics
https://youtube.com/watch?v=AGh66ZPpOSQ
https://nerdfighteria.info/v/AGh66ZPpOSQ

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

In the last episode we dove into the logic surrounding test statistics and talked about a general formula that allows us to create them for lots different situations. There are so many questions we might want to answer, and it would be rough if we had to memorize a new formula for EVERY Single One.

And sometimes Statistics is taught in a way that makes it seem like there's a different formula you need to know if you want to test whether your bus is late more often than the average bus in your town. Or if burns treated with aloe heal faster than those that are left alone. But!

Hah-zah. We can adapt the general formula...in all sorts of situations. INTRO Let's say that you just moved to a new place, and you're looking for the BEST coffee in town.

Since you've been watching Crash Course Statistics, you decide to do a little impromptu experiment. Word on the street is there are two really popular coffee places near you, Caf-fiend and The Blend Den. So one Sunday after brunch, you grab a random sample of 16 of your new friends, and randomly give half of them an unmarked cup with coffee from Caf-fiend, and the other half an unmarked cup with coffee from The Blend Den.

You made sure to get the same roast--dark--to keep things as even as possible. After delicate sniffs and sips of coffee in a process known as "cupping", the tallies are in. On a scale of 1 to 10, Caf-fiend got a mean score of 7.6 and The Blend Den got a mean score of 7.9 So we observe a difference between the coffee scores.

Coffee from Caf-fiend scored 0.3 points lower than Coffee from The Blend Den. So coffee from The Blend Den is better? Right?

Done and done. Nope not yet. Maybe it's just random chance.

So first we need to define our null. There's no difference between the two coffee shops. And then our alternative hypothesis, that there is a difference.

One is better than the other. In this case, we're interested in whether the mean scores for coffee are different between Caf-fiend and The Blend Den. With a little algebra, we can see that this is the same thing as asking whether the difference between the two means is not zero.

Now that we have our hypotheses, we can do a t-test. Specifically, we'll do a two sample t-test, also called an independent or unpaired t-test. The formula for a two sample t-test follows our general test statistic formula: The difference we observed is 0.3.

If the null hypothesis were true and there's no difference between the coffee shops, we'd expect a difference of 0. So the numerator of our t-test is 0.3. For this kind of t-test, our measure of average variation is the standard error.

For two groups, the standard error is calculated a bit differently since we have to account for the sample variance of two groups. Here, we're squaring the standard deviation to get the variance and n1 and n2 are the sizes of the two groups--both are 8 here. Now that we have our t-value, we can figure out if there's a statistically significant difference between the two coffee shops and there are two ways to do this.

We can calculate the critical t-value and if our t-statistic is GREATER than the critical value we reject the null hypothesis. Or we can calculate the p-value from our t-statistic and we can reject the null hypothesis if the p-value is SMALLER than our chosen alpha level. To do either of these things, we'll need to choose our alpha level.

Again, our alpha is arbitrary. But usually people will use 0.05 since that means that in the long run, only 5% of tests done on groups with no real difference will incorrectly reject the null. So, we'll conform :) and use an alpha of 0.05 here.

To calculate our critical t-value we need to find the t-values which correspond to the top 5% most extreme values in our t-distribution. Usually a computer or a calculator will do this for you, so we won't go into the formula, but here are the cutoffs: The cutoffs for our specific problem are about -2.145 and 2.145. We have two cutoffs because we're doing a two tailed test.

We want to reject the null if coffee from Caf-fiend is better or if coffee from The Blend Den is better. We can already tell that we should fail to reject the null. That there's no clear difference between the quality of the coffee.

Our t-statistic of about 0.44 is isn't close to -2.145 OR 2.145. The critical value and p-value approach will give you identical results, so we don't really need to do both. But for the sake of showing we get the same outcome…our calculated p-value is 0.6684.

We reject the null if the p-value is smaller than alpha, so again we fail to reject since 0.6684 is WAY bigger than 0.05. One thing that's nice about the p-value approach, and the reason we'll mainly rely on it throughout the rest of these examples, is that p-values are easier for us non-computers to interpret. A p-value of 0.6684 means that if there were NO difference in scores between coffee from Caf-fiend and coffee from The Blend Den, we'd still expect to see a difference in our sample means that's 0.3 or greater pretty often... 66.84% of the time.

Since our observed difference of 0.3 or greater is pretty common under the null hypothesis, we haven't found evidence that it's a bad fit. That's why we failed to reject it. So right now we don't have any evidence that one coffee shop is better than the other.

But remember, absence of evidence is not evidence of absence. And while our coffee excursion and experiment were well designed, we can probably improve it. If you look at the scores that your friends gave the coffees, you'll see that there's one person who tried coffee from Caf-fiend and really hated it.

After looking through your scorecards, you realize it's Alex , who has mentioned in the past that she just doesn't love coffee. Which gets you thinking. Even though you randomly assigned your friends to get either coffee from Caf-fiend or coffee from The Blend Den, that design didn't account for the fact that some people just like coffee more than others.

Alex might give the best coffee in the world a measly 6 point rating just because...coffee's not really her thing. Whereas your always caffeinated friend Cameron would probably give that day old coffee in the break room a score of 7 just because he loves coffee. So in addition to any true difference in scores between coffee from Caf-fiend and coffee from The Blend Den, our sample means are also affected by how much the people in each group like coffee.

You randomly assigned your friends to groups, so you don't expect that there's some systematic difference between the average coffee enjoyment of the groups. But random assignment adds variation, which can make it harder to see a true difference between the coffee scores. One solution to this issue is a paired t-test.

You could try to pair up your friends based on how much they like coffee and then randomly assign one to coffee from Caf-fiend and

the other to coffee from The Blend Den, and repeat this over and over until everyone had been assigned. The best match, of course, for a person is themselves. I'm just like me.

So you decide to call another random sample of 16 of your friends. This time you give all of them both Caf-fiend coffee AND The Blend Den coffee and they record their scores. Now that everyone has scored both coffees, you can be sure that the two groups have the exact same level of "coffee affinity" since it's the exact same people.

The mean scores are still affected by variation due to individual coffee preferences, but since the exact same people are in both groups, we can extract that variation and "throw it away" so to speak. One way to do this, is to make a difference score for each person. This will tell you how much more they like coffee from Caf-fiend than coffee from The Blend Den.

Now that we have only one list of values--the difference scores--our matched pairs t-test will look surprisingly similar to the one sample t-test that we've seen before. We observed a mean difference (Caf-fiend - The Blend Den) of -0.18125, which means that on average, people rated coffee The Blend Den 0.18125 points higher than coffee from Caf-fiend. The null hypothesis here is that there's no difference between ratings for coffee from Caf-fiend and coffee The Blend Den, so we'd expect our mean difference to be 0.

And our measure of average variation is just the standard error of the difference scores: Putting it together, we get a t-statistic of about -3.212. Before we get to the corresponding p-value that our computer spit out, let's consider another way to think about what t-statistics are actually telling us. T-statistics tell you how many standard errors away from the mean our observed difference is.

Though the t-distribution isn't EXACTLY normal, it's reasonably close, so we can use our intuition about normal distributions to understand our t-values. Normal distributions have about 68% of their data within one standard deviation from the mean. And about 95% within 2 standard deviations.

That means that t-scores around 3, like ours, are about 3 standard errors away from the mean...only around 0.3% of scores are that far away! So it makes sense that our p-value is very small: 0.00582. Which allows us to reject the null hypothesis that there is no difference between the scores for Coffee from Caf-fiend and coffee from The Blend Den.

Which means that from now on, I'll be buying my coffee from The Blend Den. Except for when I'm meeting up with Alex, then I'll buy` tea. Statistical tests help us wade through the murky waters of variability, and our goal should be to get rid of as MUCH of that variability as possible so that we can see patterns.

We can see whether exercise improves sleep...which your friends might be lacking after all that coffee. Or whether your hearing could be hurt by listening to loud music by Cream or Ice Cube or Vanilla Ice or some other musician that sounds like it belongs in coffee. Like Spoon!

Spoon. Yeah? Brandon Spoon.

But more importantly, we're learning that all those formulas you may have seen floating around, really aren't that different. We're just comparing what we see, to what we think we should see. We're always comparing the way things are to how we expect them to be.

And statistics is no exception. We now have the tools to design experiments and answer a lot of interesting questions and do our own experiments even if we over caffeinate some of our friends in the process. Thanks for watching.

I'll see you next time.