P-Hacking: Crash Course Statistics #30

Crash Course: Statistics
https://youtube.com/watch?v=Gx0fAjNHb1M
https://nerdfighteria.info/v/Gx0fAjNHb1M

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

Lies. Damn lies.

And statistics Stats gets a bad rap. And sometimes it makes sense why. We've talked a lot about how p-values let us know something significant in our data--but those p-values and the data behind them can be manipulated.

Hacked. P hacked. P-hacking is manipulating data or analyses to artificially get significant p-values.

Today we're going to take a break from learning new statistical models, and instead look at some statistics gone wrong. And maybe also some props gone wrong. INTRO To recap to calculate a p-value, we look at the Null Hypothesis--which is the idea that there's no effect.

This can be no effect of shoe color on the number of steps you walked today, or no effect of grams of fat in your diet on energy levels. Whatever it is, we set this hypothesis up just so that we can try to shoot it down. In the NHST framework we either reject, or fail to reject the null.

This binary decision process leads us to 4 possible scenarios: The null is true and we correctly fail to reject it The null is true but we incorrectly reject it. The null is false and we correctly reject it. The null is false and we incorrectly fail to reject it.

Out of these four options, scientists who expect to see a relationship are usually hoping for this one. In NHST, failing to reject the null is a lack of any evidence, not evidence that nothing happened. So scientists and researchers are incentivised to find something significant.

Academic journals don't want to publish a result saying: "We don't have convincing evidence that chocolate cures cancer but also we don't have convincing evidence that it doesn't". Popular websites don't want that either. That's like anti-clickbait.

In science, being able to publish your results is your ticket to job stability, a higher salary, and prestige. In this quest to achieve positive results, sometimes things can go wrong. P-hacking is when analyses are being chosen based on what makes the p-value significant, not what's the best analysis plan.

Statistical tests that look normal on the surface may have been p-hacked. And we should be careful when consuming or doing research so that we're not misled by p-hacked analyses. "P-hacking" isn't always malicious. It could come from a gap in a researcher's statistical knowledge, a well-intentioned belief in a specific scientific theory, or just an honest mistake.

Regardless of what's behind p-hacking, it's a problem. Much of scientific theory is based on p-values. Ideally, we should choose which analyses we're going to do before we see the data.

And even then, we accept that sometimes we'll get a significant result even if there's no real effect, just by chance. It's a risk we take when we use Null Hypothesis Significance Testing. But we don't want researchers to intentionally create effects that look significant, even when they're not.

When scientists p-hack, they're often putting out research results that just aren't real. And the ramifications of these incorrect studies can be small--like convincing people that eating chocolate will cause weight loss--to very, very serious--like contributing to a study that convinced many people to stop vaccinating their kids. Analyses can be complicated.

For example, x-k-c-d had a comic associating jelly beans and acne. So you grab a box of jelly beans and get experimenting. It turns out that you get a p-value that's greater than 0.05.

Since your alpha cutoff is 0.05, you fail to reject the null that jelly beans are not associated with breakouts. But the comic goes on there are different COLORS of jelly beans. Maybe it's only one color that's linked with acne!

So you go off to the lab to test the twenty different colors. And the green ones produce a significant p-value! But before you run off to the newspapers to tell everyone to stop eating green jelly beans, let's think about what happened.

We know that there's a 5% chance of getting a p-value less than 0.05, even if no color of jelly bean is actually linked to acne. That's a 1 in 20 chance. And we just did 20 separate tests.

So what's the likelihood here that we'd incorrectly reject the null? Turns out with 20 tests--it's way higher than 5%. If jelly beans are not linked with acne, then each individual test has a 5% chance of being significant, and a 95% chance of not being significant.

So the probability of having NONE of our 20 tests come up significant is 0.95 to the twentieth power, or about 36%. That means that about 64% of the time, 1 or more of these test will be significant, just by chance, even though jelly beans have no effect on acne. And 64% is a lot higher than the 5% chance you may have been expecting.

This inflated Type I error rate is called the Family Wise Error rate. When doing multiple related tests, or even multiple follow up comparisons on a significant ANOVA test, Family Wise Error rates can go up quite a lot. Which means that if the null is true, we're going to get a LOT more significant results than our prescribed Type I error rate of 5% implies.

If you're a researcher who put a lot of heart, time, and effort into doing a study similar to our jelly bean one, and you found a non-significant overall effect, that's pretty rough. Dissapointing. No one is likely to publish your non-results.

But we don't want to just keep running tests until we find something significant. A Cornell food science lab was studying the effects of the price of a buffet on the amount people ate at that buffet. They set up a buffet and charged half the people full price, and gave the other half a 50% discount.

The experiment tracked what people ate, how much they ate, and who they ate it with, and had them fill out a long questionnaire. The original hypothesis was that there is an effect of buffet price on the amount that people eat. But after running their planned analysis, it turned out that there wasn't a statistically significant difference.

So, according to emails published by Buzzfeed, the head of the lab encouraged another lab member to do some digging and look at all sorts of smaller groups. "males, females, lunch goers, dinner goers, people sitting alone, people eating in groups of 2, people eating in groups of 2+, people who order alcohol, people who order soft drinks, people who sit close to buffet, people who sit far away…" According to those same emails, they also tested these groups on several different variables like "[number of] pieces of pizza, [number of] trips, fill level of plate, did they get dessert, did they order a drink..." Results from this study were eventually published in 4 different papers. And got media attention. But one was later retracted and 3 of the papers had corrections issued because of accusations of p-hacking and other unethical data practices.

The fact that there were a few, out of many, statistical tests

conducted by this team that were statistically significant is no surprise. Many researchers have criticized these results. Just like in our fake jelly bean experiment, they created a huge number of possible tests.

And even if buffet price had no effect on the eating habits of buffet goers, we know that some, if not many, of these tests were likely to be significant just by chance. And the more analyses that were conducted, the more likely finding those fluke results becomes. By the time you do 14 separate tests, it's more likely than not that you'll get at LEAST one statistically significant result, even if there's nothing there.

The main problem arises when those few significant results are reported without the context of all the non-significant ones. Let's pretend that you make firecrackers. And you're new to making fire crackers.

You're not great at it. And sometimes make mistakes that cause the crackers to fizzle when they should go "BOOM". You make one batch of 100 firecrackers and only 5 of them work.

You take those 5 exploded firecrackers (with video proof that they really went off) to a business meeting to try to convince some Venture Capitalists to give you some money to grow your business. Conveniently, they don't ask whether you made any other failed firecrackers. They think you're showing them everything you made.

And you start to feel a little bad about taking their million dollars. Instead, you do the right thing, and tell them that you actually made 100 firecrackers and these are just the ones that turned out okay. Once they know that 95 of the firecrackers that you made failed, they're not going to give you money.

Multiple statistical tests on the same data are similar. Significant results usually indicate to us that something interesting could be happening. That's why we use significance tests.

But if you see only 5 out of 100 tests are significant you're probably going to be a bit more suspicious that those significant results are false positives. Those 5 good firecrackers may have just been good luck. When researchers conduct many statistical tests, but only report the significant ones, it's misleading.

Depending on how transparent they are, it can even seem like they only ran 5 tests, of which 5 were significant. There is a way to account for Family Wise Errors. The world is complex, and sometimes so are the experiments that we use to explore it.

While it's important for people doing research to define the hypotheses they're going to test before they look at any data, it's understandable that during the course of the experiment they may get new ideas. One simple way around this is to correct for the inflation in your Family Wise Error rate. If you want the overall Type I error rate for all your tests to be 5%, then you can adjust your p-values accordingly.

One very simple way to do this is to apply a Bonferroni correction. Instead of setting a usual threshold--like 0.05--to decide when a p-value is significant or non-significant, you take the usual threshold and divide it by the number of tests you're doing. If we wanted to test the effect of 5 different health measures on risk of stroke, we would take our original threshold--0.05--and divide by 5.

That leaves us with a new cutoff of 0.01. So in order to determine if the effect of hours of exercise--or any of our other 4 measures--has a significant effect on your risk of stroke, you would need to have a p-value of below 0.01 instead of 0.05. This may seem like a lot of hoopla over a few extra statistical tests, but making sure that we limit the likelihood of putting out false research is really important.

We always want to put out good research, and as much as possible, we want the results we publish to be correct. If you don't do research yourself, these problems can seem far removed from your everyday life, but they still affect you. These results might affect the amount of chemicals that are allowed in your food and water, or laws that politicians are writing.

And spotting questionable science means you not have to avoid those green jelly beans. Cause green jelly beans are clearly the best. Thanks for watching, I'll see you next time.