The Replication Crisis: Crash Course Statistics #31

Crash Course: Statistics
https://youtube.com/watch?v=vBzEGSm23y8
https://nerdfighteria.info/v/vBzEGSm23y8

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

You might have heard that Power Posing affects how powerful you feel and can change hormone levels. If it does we'd expect to see that effect over and over and over.

Study after study. And it would be pretty disappointing if one study concludes that eating carrots improves your vision, and then after you rushed to sign up for monthly carrot deliveries...5 similar studies find no evidence that munching carrots is good for your eyes. To make sure that an experimental result is sound, we want to replicate the findings.

Results need to be confirmed. Which is why replication--re-running studies to confirm results --and reproducible analysis--the ability for other scientists to repeat the analyses you did on your data--are essential in science. These issues affect basically every field from Artificial Intelligence research to social science.

INTRO A few years ago scientists at a biotech company called Amgen decided to try to replicate more than 50 big-deal cancer treatment studies. These were studies that had been published in respected journals. And the Amgen scientists were only able to replicate the original results 11-percent of the time.

In another reproducibility study...a group of 270 scientists re-ran 100 psychology studies that had been published in 2008 in top-notch journals. Fewer than half of the published results were replicated. Stanford researcher, Dr.

John Ioannidis has claimed that "false findings may be the majority or even the vast majority of published research claims". The journal Nature published a survey a few years back and asked researchers if they thought there was a reproducibility crisis in science. 52% called it a "significant crisis" another 38% called it a "slight crisis". And 90% of researchers thinking they have some size of crisis on their hands is big deal.

The "replicability crisis" has been used in political debates to undermine scientific research. Political activists, especially those that hold opinions that run counter to scientific research, have jumped on the problem of replicability as a way to discredit science more broadly. And when a medical study winds up with invalid conclusions researchers could head down the wrong path people could get misguided treatments based on faulty conclusions they could get sicker even...and a whole lot of money could be wasted researching and providing those treatments.

So, what's causing science's replication problem? There are a lot of answers. Some of them involve unscrupulous researchers--researchers that are more concerned with attention and publishing and splashy headlines than good science.

Here we're talking about fraud. Falsified data. Intentional p-hacking.

Statistical evil doers. But many reasons scientific studies aren't replicable are less nefarious. One issue related to replication--re-doing studies--is reproducibility of the analyses in a paper.

There's not always one prescribed way of analyzing a data set. A researcher named Brian Nosek and his team invited 29 groups of researchers to analyze the same data set--and attempt to answer whether or not soccer referees give more red cards to dark-skinned players than light-skinned ones? Seems simple enough.

These researchers were all working with the SAME data--but they ran different tests. Some used linear regressions. Some went with Bayesian models.

And it's not just the models that the researchers could have differed on. You also have freedom to exclude different outliers, or look at different groups. Twenty of the groups found a statistically significant relationship between skin color and red cards.

Nine groups didn't. The point, says researchers, is that no one analysis is gonna find THE answer, THE singular truth. When researchers aren't clear about how they analyzed their data, from which data points they excluded, to the exact model they ran, it can make it hard for someone to reproduce their results.

Even if they had the same data. Good papers will have detailed descriptions of researchers' methods. When you replicate a study you usually know what model the researcher used or you can ask.

But if scientists aren't clear or consistent about this, it just puts another roadblock in the way of good replication. There are other reasons for the replicability crisis. Some researchers and the folks who report on scientific research don't fully understand p-values.

They make claims that statistical evidence doesn't support. Back in 2016, the American Statistical Association released a statement meant to help researchers understand and use P values better. It was reportedly the first time the 170-plus year old organization made this type of explicit recommendations.

Among the guidelines the Statistical Association published: "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold." And "A p-value, or statistical significance, does not measure the size of an effect or the importance of a result." P-values need to be understood in context. A significant result doesn't mean we ought to all rush out and change what we're doing. But if you like carrots, by all means keep eating them.

Another reason science produces results that can't be reproduced is that published studies have a bias toward overestimating effects--in part because they were published because they had a low p-value. Some studies look promising and then aren't reproducible because they were based on a fluke. When the study is repeated the fluke doesn't repeat itself.

The website FiveThirtyEight offers up this explanation: Say you were looking at the relationship of height and college majors. You gather up your data--including a class of math majors with a few exceptionally tall kids and a class of philosophy majors with an unusually short student. When you compare the averages-- ha ha!

Look at that! Math majors are taller than philosophy majors. You have statistically significant results but when you repeat the study those differences disappear there's regression to the mean which gives you a more accurate picture of pretty similar average heights of each major and nothing all that interesting to write about.

Except a correction to your first paper. Small sample sizes also get blamed. The fewer subjects in a study, the more likely you get skewed and unreplicable results.

DFTBAQ, my friends. Even when results make sense to you--DFTBAQ. So where can researchers start improving the process--to help solve this reproducibility crisis.

For one, researchers argue they need to do a whole lot more replication. Replication allows us to weed out false significant effects: the flukes and the "too good to be true" effects that unfortunately make great headlines. We need to get rid of the idea that one significant test is solid proof of anything.

It isn't. In fact, we need to get rid of the idea that one significant

test is even great evidence of anything. But replication is expensive.

And it's not as sexy as making a new discovery. It doesn't attract the same media attention institutional acclaim or funders. Who wants to say "I found the effect that my colleague found yesterday!"?

So say researchers we've gotta come up with ways to change those incentives. We need to find more funding for replication studies and change the way we all view the value of replication. Some people call for more publication of "null results"--those that DON'T support the hypothesis.

This would allow quality research to be published, even if it didn't show an effect, making p-hacking a little less enticing, since you could still get null results published. Some researchers argue another way to help correct the reproducibility crisis is by reconsidering the standard p-value cut off of .05 for statistical significance. Is it stringent enough?

Or should researchers move it? In 2017, a group of more than 70 researchers co-authored a paper calling for a change in the default P-value threshold from .05 to .005. They wrote: "This simple step would immediately improve the reproducibility of scientific research in many fields." Calling results with a p-value of less than .05 statistically significant they argue results in a high rate of false positives...even when that research is done correctly.

Let's just look at one area of research we've talked about before: Social Priming -- the idea that certain actions or conditions can affect the way you behave. One famous case of social priming is a study where subjects who were exposed to words related to old age--like Florida, bingo, grey, or retired--walked more slowly after exposure than those who were shown neutral words. But recently, many researchers have expressed concerns that some of these social priming results may not hold up.

When they began to see that, many experiments were done with many different priming mechanisms and outcome variables. And we're making this data up here, but let's say that out of 1000 studies done, about 10% or 100 ended up with real effects of social priming. This is a table that displays how often our studies resulted in true positives, false positives, true negatives, and false negatives.

The top row shows the 900 studies where social priming DIDN'T work. Because we used a threshold of 0.05, 5% of those 900 studies will still be statistically significant even though there was no effect. Those 45 are our false positives.

That leaves 855 studies where social priming didn't work and we caught it. Those are our true negatives. The next row contains the 100 studies where social priming DID work.

In those studies, there were actual effects of social priming. There you can see our true positives (60) and false negatives (40). So what does that mean?

Well, remember, statistical power is the ability to detect real effects. Sometimes we can fail to get a significant result, even if an effect of a certain size is real. One estimate suggests that most psychology studies have an average of 60% power.

So, that 60 on our table represents the 60 studies that had a significant effect that was observed. The other 40 weren't caught, giving us false negatives. Using our table, we can look at the percent of significant results that come from studies with no effect.

Our "False Alarms". Our False Discovery Rate is 45 divided by 105, or 42.9%. That means of all the significant effects that were recorded and published in our thought experiment, a bit less than HALF of them are false positives.

Which shows, as we mentioned before, that having a statistically significant effect doesn't make it REAL. All else being equal if we had changed the p-value threshold from .05 to .005, we would have way less false positives. To make the work of reproducibility easier, there are also pushes underway to encourage researchers to share their data more widely.

In the United Kingdom, for example, many research funders expect researchers will make publicly funded research data available--recognizing the data as a public good. Academic journals also play a role in the conversation around reproducibility--many of the most prestigious journals have adopted guidelines and policies that put more emphasis on reproducibility and transparency. In part, to help boost public trust in science and the scientific process.

Let's go back to power posing before we finish today. Really get that blood flowing. Confidence building!

A study on power posing was published in Psychological Science back in 2010 that showed that power posing could change hormone levels and boost confidence. A TED talk about power posing was viewed more than 40 million times. Want a raise?

Respect and awe from your friends and family and enemies? Power pose. Or not.

After power-posing went mainstream other researchers tried to replicate the study--with a larger sample--and didn't come up with the same results. Other researchers found significant problems with the original study and came to the conclusion that quote "the existing evidence is too weak to justify a search for moderators or to advocate for people to engage in power posing to better their lives." Power-posing got labeled pseudo-science. And then in 2018 the original author published a response to some of the critiques about power-posing...with an analysis that suggested the poses could help people feel more confident and powerful.

Now, the newest paper doesn't seem to address all of the critiques about Power Posing study but it comes to the conclusion that researchers shouldn't give up research about the effects of Power Posing quite yet. No. No.

These are not power poses. I'm just trying to find something that indicates confusion. This back and forth of the power posing debate does make it harder to know what's likely to be true.

But it also shows the VALUE of replication and even the reproducibility crisis in research. Science is a push and pull of ideas--researchers are constantly iterating and expanding on ideas that came before. They refine results.

Build on other people's findings. Replication is an essential part of the path to scientific progress and real breakthroughs. The reproducibility crisis means more people are taking the replication step of the process seriously.

Replication has helped us accomplish some pretty important things. Like help change peoples minds about whether smoking caused increases in lung cancer, even though researchers could never do a Randomized Controlled Trial to demonstrate causation. Evidence piled up, and now smoking rates are incredibly low.

No single study is gonna show us the way the world REALLY is but that study and the studies that follow it that do and don't find the same relationships will get us closer and closer. And one day maybe we'll know--with more explicit certainty whether or not we

ought to be putting on our hands on our hips and doing the wonder woman before a big job interview. Thanks for watching, I'll see you next time.