Transcript: Philosophy VIDEO 2.6 – P-Values: What Are They?

This video develops the previous one in the following way. I take that non-black non-ravens puzzle, which probably doesn't really seem real, and try to make it realer. Real scientists really get into trouble in cases like this. And it really seems like part of the problem is that they are doing a thing that works, but it doesn't always work. And, since they aren't totally clear why it works, they aren't totally clear when it works—and doesn't.

Let's begin.

Since the dawn of time, man has wondered: what are p-values?

OK, that's false. Furthermore, you should almost never start your essays with 'since the dawn of time, man has wondered'. It's got a male bias and puts you on the hook for waaay too much anthropology. I know, I kind of did it myself with Kubrick's ape-men, way back in one of those early videos. But I'm a pro. I know how to finesse it.

Let me start over: since the dawn of time, man should have wondered: what are p-values?

That's a little better.

But you're confused. What are p-values?

P for probability. In the sciences, especially the social sciences, there has been a long-standing convention that a p-value of less than .05—5%—is 'statistically significant'. That's key, because statistically significant results are deemed publishable.

In academia, the ultimate rule is 'publish or perish'. So naturally everyone has gone chasing sub-.05 p-values.

This has caused problems because making .05 your 'significant' level is a heuristic. A rule of thumb. Think about it. 6% isn't all that different from 4%. How could it be?

The risk in stipulating a .05 target is that people will try to find ways to game their results to hit the .05 mark.

The problem that has resulted is the so-called replication crisis in the social sciences, particularly psychology. A good experiment needs to be replicable and ideally, should have been replicated. The problem is that a lot of experimental results people thought were very significant and well-confirmed—because of that p-value .05 thing—have recently actually failed.

If you want to know more, Google "replication crisis". I can't spare the time in this video. But it's worth your time. It's interesting. I'll add a personal note. I teach another intro GE module, "Reason and Persuasion". For several years I have used

bits from a popular psychology text, *The Happiness Hypothesis*, by social psychologist Jonathan Haidt. I started to notice some of the neat results he passes along—cool, weird stuff about humans—were turning out not to be, you know, not so true. First it was the priming effects studies. Then it was the willpower depletion stuff. It crept up on me. I was teaching a book that seemed true in 2006 but didn't seem true enough anymore in 2016. It wasn't Haidt's fault. It wasn't my fault for trusting Haidt. But what gave? What has gone wrong?

P-values. Yeah, but it's tricky how so. Let me quote a well-known, very smart statistician Andrew Gelman. He co-wrote, with Erik Loken a good paper on the p-values controversy, entitled, "The Garden of Forking Paths". But I'm just going to quote from a blog post by Gelman, because it reads out clearer:

"Ultimately the problem is not with p-values but with null-hypothesis significance testing, that parody of falsificationism in which straw-man null hypothesis A is rejected and this is taken as evidence in favor of preferred alternative B."

Null-hypothesis A? Holbo, is this your idea of clarity?

OK, you can just think of it in terms of those urns I told you about in the last video. Remember, you've got a big urn. It holds a million balls.

You pick out—black, black, black, black, black, black, black. And on. Hundred black balls right in a row.

Kind of makes you think: looks like all the balls in this urn are black, right? A reasonable induction?

But slow down, cowboy. Hold your horses.

What hypotheses are competing here. Let's take two.

Hypothesis A (null hypothesis): Half of the balls are black, half are white.

Hypothesis B: All the balls are black.

Now, if these are the competitors, then—after 100 black balls—B is the clear winner. The odds of 100 black balls, if Hypothesis A were right—well, that would be exactly like flipping a fair coin and getting 100 heads in a row. Odds against are astronomical. Agreed?

Now you are starting to get it about the null hypothesis, I hope.

In order for the balls you draw to be evidence FOR something—B, in this case—there needs to be something they are evidence AGAINST—A, in this case. The so-called null hypothesis. Null meaning the thing you are aiming to nullify, pretty much.

This is Popper stuff. Gelman says falsificationism. I said disconfirmation, when I explained Popper. Same thing. Science is attempted disconfirmation of someone's general bet. That's the null hypothesis.

Gelman says the problem is when your null hypothesis is a straw man. You know what straw man arguments are. You make yourself seem smart by misrepresenting your opponent's position as dumber than it needs to be.

In this case, B seems smart because A is dumb. But the thing to do is not to believe B, therefore. The thing to do is think of smarter alternatives to B than just A.

I mentioned two of them last time. I'll rename them C and D

Hypothesis C: one of the balls (out of a million) is white.

Hypothesis D: there's a layer of white balls, but that layer is at the bottom of the urn.

It's clear how just drawing black black black black black from the top of the urn is strong evidence against A, in favor of B, but not strong evidence against C or D, in favor of B. It seems like, to support B absolutely, we need evidence that knocks down A and C and D … but where will this end? Just looking at an urn with a million balls in it—that's a lot of possibilities for possible combinations of black and white balls. And why stop there? Maybe the next thing you pull out of the urn will be a rubber chicken. Why not?

We want to know what's real.

About balls in urns or black swans or financial crises. Or whatever.

That seems like it's a straight—what's there?—question.

But really it's probabilistic and it's a 'compared to what?' question. Which brings us to p-values.

Last year the ASA—the American Statistical Association—released a statement, purporting to provide an informal, suitable for non-statistician consumption, statement of what p-values provide, or say, or are.

> Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.

And that's why everyone hates statisticians. I mean, seriously: that was clear as mud.

Even worse, even professionals who work with this stuff, if asked to explain what this informal gloss means—informally—are likely to come out with something wrong (not all of them, but enough that I'm warning you):

'the p-value tells you the likelihood that your result happened just by chance.'

That's the most common, definitely wrong thing you are likely to hear about p-values. People who say this have failed to understand what the question was, that their <.05 answer was an an answer to.

Let's start with a simple case that shows how and why this wrong gloss *has* to be wrong; then, I shall offer my improved informal gloss on the ASA's informal gloss.

What is the simplest case in which we might arrive at a $p < .05$ experimental result in the comfort of our own homes.

Flipping a coin, getting heads 5 times in a row. We know how to calculate the likelihood of that happening: 2×2x2×2x2 = 32.

You have a 1 in 32 chance of flipping a fair coin five heads in a row. So it may take you, I dunno, a couple hours to generate this result. But when you do: science gold!

1 in 32 is < 5% so we publish!

What will our result say, in the pages of *Science* or *Nature*, or *Experimental Numismatics*?

If it were really true that p <.05 says 'the chance of this having happened by chance is less than 5%' then we might run with a headline like this:

Trick coin almost certainly found in local pocket!

No, that's nuts. No one is going to conclude that a coin is a trick coin, just because they flipped five heads. Five heads is unlikely, but the odds that a trick coin snuck into your pocket is, offhand, way more unlikely. Trick coins have got to be 1 in a million in the general coin population. 1 in a 100 million, probably.

In calculating odds concerning a 5-head coin-flip streak, you aren't calculating the chance that your coin is a trick coin that has been weighted somehow to come up heads. There's no way you could be calculating that.

What you are calculating, obviously, is: the odds of a 5-head streak, given that the coin is fair. Your calculation assumes a fair coin. That's what those 2's mean that you multiplied.

Without further ado, Holbo's rewrite of the ASA statement.

1) If this coin is fair, odds are less than 1 in 20 that you could match or beat that 5-head run I just got!

Tying this to the ASA thing (bit loosely):

"under a specified statistical model" = If this coin is fair

"the probability that … a statistical summary of the data … would be equal to or more extreme than" = odds are less than 1 in 20 that you could match or beat
"its observed value" = that 5-heads run I just got!

Whenever you read a scientific paper—be it in psychology or whatever field—and it reports a 'statistically significant' p = <.05 result, mentally rewrite it in terms of my coin case. Scientific results really are a lot like trick coins discovered in our pockets. The world turns out to exhibit some new regularity that is both robust and surprising. All the same, proving you've found a trick coin needs more than just flipping 5 heads in a row.

Putting it another way: the p-value calculation models a possible world in which, by hypothesis, a certain thing would be unlikely to happen. And the conclusion is: we probably don't live in that possible world, since the thing that actually did just happen—5 heads—would be less than 5% likely to happen in that possible world. But that step is suspect, we now see. Even if it weren't, it wouldn't tell us what real world we DO live in.

Incidentally, there are links here to the DN model of scientific explanation. Calculating p-values means relying on deductive-nomological-style thinking. But, here again, the trick is saying where laws come from, hence how this thinking relates to the logic of induction.

Confused? Me, too.

Let's try this. Let's imagine a world in which flipping 5-heads in a row would actually be a good basis for concluding you've found a trick coin.

You worked hard all day and the boss gave you a shiny dollar for your pains. You are running home, clutching the precious coin in your hand. Unfortunately, on the way you collide with Mysterioso the Mysterious, famous magician who lives around

the corner. He is walking along, whistling, flipping his famous trick dollar coin. It always comes up heads.

The two of you collide.

Oh, no, your two coins have gotten mixed up! Whose is whose?

Well, pick one and flip it. If it comes up heads five times in a row there is a less than 5% chance that it's yours. So give that one to the magician.

Note the difference between this case and just flipping a coin from your pocket and happening to get 5-heads.

In the collision case, the competing hypothesis are, independently, equally likely:

Hypothesis A: This is a trick coin.

Hypothesis B: This is a fair coin.

But normally these two hypothesis will not be equally likely.

Let's illustrate with a realistic case concerning which it might be possible to make a mistake, and a lot of people do.

I tell you formula XYZ was administered to 5 cancer patients and they all recovered soon after. Would you say formula XYZ sounds likely to be an effective cancer treatment?

Many would say yes.

But now I add that formula XYZ is water and everyone immediately sees the problem. They were assuming it was independently even-odds that XYZ was curative, or not. But it's obviously not.

It's more likely that we have some kind of cancer recovery fluke here than that plain water cures cancer.

Do you see the connection with the coin case?

If you said it was likely that XYZ is an effective treatment, you think you just collided with Mysterioso. But most days you don't run into that guy, actually.

A cure for cancer is like a trick coin. You don't find one everyday. They're 1 in 10 million. But if you are reasoning as if you just collided with Mysterioso—so a trick coin is equally likely—you may trick yourself into thinking maybe you just cured cancer.

And remember that thing I asked about Plato: whether it was reasonable of him to expect a simple formula for the motions of the wanderers? That's kind of like thinking there might be a neat, trick coin in the heavens. Most coins are wanderers. Heads, heads, tail, heads, tails, tails. Pretty random walk. But some are tricks. But there isn't any simple math that tells you whether you just, as it were, collided with Mysterioso the Mysterious.

What's the point? Well, the big take-away is: friends don't let friends get confused about p-values. Around here it's too easy to answer questions without realizing what the question was.

Also, I hope this has made vivid that weird puzzles about the logic of confirmation have real-world consequences.