CHAPTER 14

The probability of a dry toothbrush: What is a *p*-value anyway?

I have a party trick: when I tell someone what I do, and they say, "statistician, eh? I took statistics in college," I ask them to define the *p*-value. (I know what you're thinking—not much of a trick. I'm working on some other stuff.) The point is, I have yet to meet anyone who has got anywhere close to the right answer. This is pretty odd because the *p*-value is such a key idea in statistics. Imagine if a literature graduate didn't know whether Shakespeare wrote plays or novels, or someone who'd taken an economics course couldn't describe the relationship between supply and demand. So, if you do nothing else, please try to remember the following sentence: "The *p*-value is the probability that the data would be at least as extreme as those observed, if the null hypothesis were true." Though I'd prefer that you also understood it—about which, teeth brushing.

I have three young children. In the evening, before we get to bedtime stories (bedtime stories being a nice way to end the day), we have to persuade them all to bathe, use the toilet, clean their teeth, change into pajamas, get their clothes ready for the next day and then actually get into bed (the persuading part being a nice way to go crazy). My five-year-old can often be found sitting on his bed, fully dressed, claiming to have clean teeth. The give-away is the bone dry toothbrush: he says that he has brushed his teeth, I tell him that he couldn't have.

My reasoning here goes like this: the toothbrush is dry; it is unlikely that the toothbrush would be dry if my son had cleaned his teeth; therefore he hasn't cleaned his teeth. Or using statistician-speak: here are the data (a dry toothbrush); here is a hypothesis (my son has cleaned his teeth); the data would be unusual if the hypothesis were true, therefore we should reject the hypothesis.

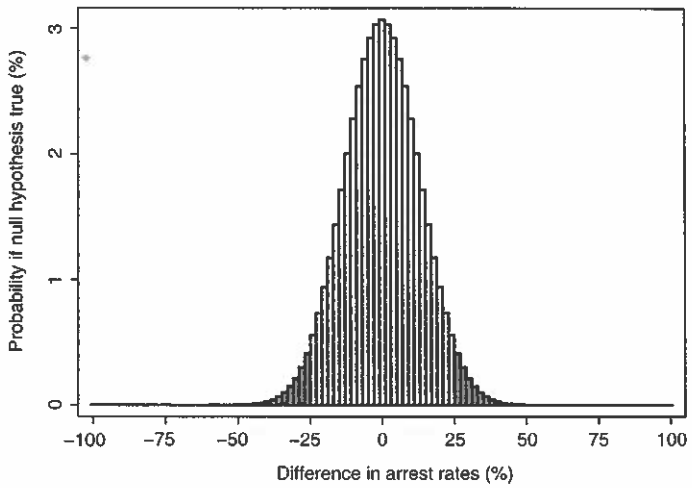
Statistical analysis of a set of data follows a very similar principle—you want to know the probability of the data if the null hypothesis were true. As an example, there was an idea a few years back that young people who had broken the law should be given tours around prisons to show them just how awful their lives would be if they didn't start behaving (that is, they would be "scared straight"). Some social science researchers decided to run an experiment in which young offenders were randomly chosen either to be scared straight or to be treated as usual in the criminal justice system (the control group). Here are some typical data from one of these experiments. 12 of 28 (43%) in the scared straight group committed a new crime compared to 5 of 30 (17%) in the control group.

It looks as though trying to scare teenagers makes things worse, though it could be that the study results were bad luck (in the same way that, by chance, you might beat me handily at Chutes and Ladders). To analyze the data statistically, we first write down a null hypothesis (which roughly speaking is that nothing interesting is going on). So our null hypothesis could be something like: "The chance of committing a crime after a first arrest is the same in teenagers going through scared straight as those going through the usual criminal justice procedures." Next, we conduct a statistical test and get $p = 0.043$. As the p-value is less than 5%, we call our result statistically significant, reject our null hypothesis and conclude that scared straight truly does make things worse.

Now, the p-value of 0.043 isn't quite the probability that, if there was no effect of scared straight, we would see exactly 12 of 28 committing new crimes in the scared straight group and exactly 5 of 30 in controls. This is because we also want to take into account the possibility that the results could have been reversed (that is, more crimes in controls) or might have shown an even bigger difference between groups (e.g., 100% crime rate after scared straight and 0% in controls). In both of these cases we would have rejected our null hypothesis.

Essentially what we do to get the p-value is to write down every possible result of the study (14 of 28 committing crimes in scared straight and 15 of 30 in controls, 1 of 28 crimes in scared straight, 29 of 30 in controls, 24 of 28 and 6 of 30, etc.). We then work out the probability of each result if the null hypothesis were true (e.g., a result like 1 of 28 crimes in scared straight vs. 29 of 30 in controls would be very unlikely if the arrest rate was truly the same in each group). Finally, to get the actual p-value, we add up the probabilities of all results that are at least as unlikely as the one we got.

We can also show this on a histogram. The x-axis gives the difference in arrest rates between the control and scared straight groups and the y-axis gives the probability of each possible result if the null hypothesis were true.



As you'd expect, the most common result if the null hypothesis were true is that there is no difference between groups, though small differences are also pretty common. Very large differences between the groups are extremely rare: the probability of a 100% difference—all controls being arrested and no scared straight arrests—is less than 1 in 10^{16} if the null hypothesis were true (not far from the chance of correctly identifying a randomly chosen grain of sand from all the beaches in the world). The difference we observed in the study was 26% and the shaded areas show results at least as extreme as this. If you add up all the shaded areas, what you get is the probability of getting a difference of 26% or more if the null hypothesis were true. This is the *p*-value: 0.043.

So here is what to parrot when we run into each other at a bar and I still haven't managed to work out any new party tricks: "The *p*-value is the probability that the data would be at least as extreme as those observed, if the null hypothesis were true." When I recover from shock, you can explain it to me in terms of a toothbrush ("The probability of the toothbrush being dry if you've just cleaned your teeth").

• Things to Remember •

1. Inference statistics involves testing a hypothesis, specifically, a null hypothesis.
2. A null hypothesis is a statement suggesting that nothing interesting is going on, for example, that there is no difference between the observed data and what was expected, or no difference between two groups.
3. The *p*-value is the probability that the data would be at least as extreme as those observed if the null hypothesis were true.
4. If the data would be unlikely if the null hypothesis were true, we conclude that the null hypothesis is not true.
5. My son has now worked out my trick and has taken to running his toothbrush under the tap for a second or two before heading to bed.

SEE ALSO: *Choosing a route to cycle home: What p-values do for us*

camera is still in the house." Given that it is easier to pop back inside the house than to unload the car, I decide to test the second hypothesis. A few minutes later, I tell my wife that I have looked in all the normal places inside and couldn't find the camera. We conclude that "it must be in the car somewhere" and head off on our road-trip.

There is something a little odd behind this story: we concluded one thing (that the camera was in the car) because we couldn't find evidence to support something else (the camera was in the house). But as it happens, this is exactly what we do when we conduct a statistical test. First, we propose a null hypothesis, roughly speaking, that nothing interesting is going on (see *The probability of a dry toothbrush: What is a p-value anyway?*). We then run our statistical analyses to obtain a *p*-value. The *p*-value is the probability that the data would be at least as extreme as those observed, if the null hypothesis were true, so if the *p*-value is low (say, less than 0.05) we say, "These data would be unlikely if the null hypothesis were true, therefore it probably isn't." As a result, we declare our result "statistically significant," reject the null hypothesis and conclude that we do indeed have an interesting phenomenon on our hands.

As a simple example, we might have a null hypothesis that girls and boys learn handwriting at the same rate, and a data set of handwriting test scores divided by gender. A statistically significant *p*-value would lead us to reject this null hypothesis and conclude that there are differences in handwriting at an early age.

What we do if *p* is greater than 0.05 is a little more complicated. The other day I shot baskets with Michael Jordan. (Remember that I am a statistician and never make things up.) He shot 7 straight free throws, I hit 3 and missed 4 and then (being a statistician) rushed to the sideline, grabbed my laptop and calculated a *p*-value of 0.07 for the null hypothesis that I shoot baskets as well he does. Now, you wouldn't take this *p*-value to suggest that there is *no* difference between my basketball skills and those of Michael Jordan—you'd probably say something like our experiment hadn't *proved* a difference.

Yet a good number of otherwise smart people come to exactly the opposite conclusion when interpreting the results of statistical tests. Just before I started writing this book, a study was published reporting about a 10% lower rate of breast cancer in women who were advised to eat less fat. If this is indeed the true difference, low fat diets could reduce the incidence of breast cancer by tens of thousands of women each year—astonishing health benefit for something as simple and inexpensive as cutting down on fatty foods. The *p*-value for the difference in cancer rates was 0.07 and here is the key point: this was widely misinterpreted as indicating that low fat diets don't work. For example, the *New York Times* editorial page trumpeted that "low fat diets flub a test" and claimed that the study provided "strong evidence that the war against all fats was mostly in vain." However, failure to prove that a treatment is effective is not the same as proving it ineffective. This is what statisticians call "accepting the null hypothesis" and, unless you accept that a British-born statistician got game with Michael Jordan, it is something you'll want to avoid.