7.7.2: The p-value of a Test

There are two somewhat different ways of interpreting a p-value ("p" standing for "probability"), one proposed by Sir Ronald Fisher and the other by Jerzy Neyman. Both versions are legitimate, though they reflect very different ways of thinking about hypothesis tests. Most introductory textbooks tend to give Fisher's version only, but that's a bit of a shame. Neyman's version seems cleaner, and actually better reflects the logic of the null hypothesis test. You might disagree though, so both are included . We'll start with Neyman's version…

Neyman: A Softer View of Decision-Making

One problem with the hypothesis testing procedure that has been described so far is that it makes no distinction at all between a result this "barely significant" and those that are "highly significant". For instance, imagine that we tested whether people who say they have extra-sensory perception (ESP) really do by naming the pictures on cards, and found that people were able to correctly identify the picture 62 times out of 100 observations, the calculated value falls just inside the critical region and we get a statistically significant effect. However, it's so close to the "retaining" section of the critical value table that it is pretty nearly nothing. In contrast, suppose a different study found 97 out of the 100 participants got the answer right. This would obviously be significant too, but my a much larger margin. The procedure that we're using makes no distinction between the two. When we adopt the standard convention of allowing α=.05 as the acceptable Type I error rate, then both of these are significant results (62 correct answers and 97 correct answers are both statistically significant).

This is where the p value comes in handy. To understand how it works, let's suppose that we ran lots of hypothesis tests on the same data set: but with a different value of α in each case. When we do that for my original ESP data of 62 correct answers, what we'd get is something like this:

- If 0.05 is the chosen critical value ($\alpha = 0.05$), then the null hypothesis would be rejected.
- If 0.04 is the chosen critical value ($\alpha = 0.04$), then the null hypothesis would be rejected.
- If 0.03 is the chosen critical value ($\alpha = 0.03$), then the null hypothesis would be rejected.
- If 0.02 is the chosen critical value ($\alpha = 0.02$), then the null hypothesis would NOT be rejected.
- If 0.01 is the chosen critical value ($\alpha = 0.01$), then the null hypothesis would NOT be rejected.

hen we test ESP data (X=62 successes out of N=100 observations) using α levels of .03 and above, we'd always find ourselves rejecting the null hypothesis. For α levels of .02 and below, we always end up retaining the null hypothesis. Therefore, somewhere between .02 and .03 there must be a smallest value of α that would allow us to reject the null hypothesis for this data. This is the p value; as it turns out the ESP data has p=.021. In short:

In effect, p is a summary of all the possible hypothesis tests that you could have run, taken across all possible α values. And as a consequence it has the effect of "softening" our decision process. For those tests in which $p \leq \alpha$ you would have rejected the null hypothesis, whereas for those tests in which $p > \alpha$ you would have retained the null. The ESP study obtained X=62, with p=.021. So the error rate I have to tolerate is 2.1%. In contrast, suppose my experiment had yielded X=97. It becomes is a tiny, tiny Type I error rate. For this second case we would reject the null hypothesis with a lot more confidence that the sample really represents the reality of ESP in the population because we only have to be "willing" to tolerate a Type I error rate of about 1 in 10 trillion trillion in order to justify my decision to reject.

> p is defined to be the smallest Type I error rate (α) that you have to be willing to tolerate if you want to reject the null hypothesis.
>
> If it turns out that p describes an error rate that you find intolerable, then you must retain the null. If you're comfortable with an error rate equal to p, then it's okay to reject the null hypothesis in favor of your preferred alternative.

Fisher: The Probability of Extreme Data

The second definition of the p-value comes from Sir Ronald Fisher, and it's actually this one that you tend to see in most introductory statistics textbooks. Notice how the critical region corresponds to the *tails* (i.e., extreme values) of the sampling distribution? That's not a coincidence: almost all "good" tests have this characteristic (good in the sense of minimizing our Type II error rate, β). The reason for that is that a good critical region almost always corresponds to those values of the test statistic that are *least likely* to be observed if the null hypothesis is true. If this rule is true, then we can define the p-value as the probability that we would have observed a test statistic that is at least as extreme as the one we actually did get. In other words, if the data are

extremely implausible according to the null hypothesis, then the null hypothesis is probably wrong. You'll read different versions of this idea throughout this textbook.

A Common Mistake

Okay, so you can see that there are two rather different but legitimate ways to interpret the p value, one based on Neyman's approach to hypothesis testing and the other based on Fisher's. Unfortunately, there is a third explanation that people sometimes give, especially when they're first learning statistics, and it is *absolutely and completely wrong*. This mistaken approach is to refer to the p value as "the probability that the null hypothesis is true". It's an intuitively appealing way to think, but it's wrong. We won't get into why it's wrong because that gets into the philosophy of probability and statistics, but no hypotheses are every proved true or not true. Research hypotheses are support or not, and null hypotheses or retained or rejected. When a null hypothesis is retained, we're saying that our sample suggests that there's no differences between these groups in the population. It's easy fall into this phrasing, though. It might even be in this book somewhere!

Contributors and Attributions

- Danielle Navarro (University of New South Wales)
- Dr. MO (Taft College)