Introduction to Statistical Power

Statistical power is, in terms of practical use, kind of the flip side of a p-value.

Remember that a p-value is an attempt to quantify our degree of confidence in not having made a type I error (although it is *not* the probability of having made a type I error), i.e., making a false positive.

Well, statistical power is about avoiding type II errors, that is, making false negatives. Often, it's said that the statistical power of a test 1 - the probability of making a type II error; equivalently, if a study has 80% power, it means that 80% of the time that we ought to be rejecting the null hypothesis, we do so.

From that description, we might think that calculating statistical power isn't very useful after we get a result—and that's generally right. The point of power analysis isn't to say "hey, we got a result, what's the chance of incorrectly rejecting the null." Because power analysis depends on the sample size (among other things, more in a moment), the real point is to figure out what the sample size should be.

Formally speaking, statistical power is a function of the effect size—the extent to which the treatment makes a difference in the null hypothesis (usually standardized into a measure called *Cohen's d* to permit comparison across different studies), the number of observations, and alpha, or the significant level you're planning to use for your study. And it's an increasing function of all these things—that is, larger sample size = more power, larger effect size = more power, and higher significance level (if you're willing to accept p-values of 0.1 rather than 0.05) = more power.

Pause for a moment and think about this. It's important to have an intuition for why this is true. Remember that when we're doing significance testing, we're asking the following question: *Is our data so radically different from what we'd expect under the null hypothesis that it's very unlikely we'd have seen it if the null hypothesis were true?* And we capture that notion of being radically different in terms of standard deviations of the sampling distribution (aka standard errors): if our statistic in the data is lots of standard deviations away from the null hypothesis statistic, then it's a significant result.

The question of power, then, just means "before I do the study, how likely is it that I'll see data lots of standard errors away from the mean in my sample, if there's a difference in the population?" Well, let's reason through this. If we were omnipotent, what levers could we pull in order to change that likelihood?

- We could reduce the standard error. And since the standard error is the population standard deviation (estimated by the standard deviation of the sample) divided by the square root of the sample size, that means we need to increase *the sample size.*
- We could increase the distance between the observed statistic and the null

hypothesis mean, i.e., the *effect size*.

- We could decrease the number of standard errors that count as "lots," a.k.a., the *significance level*.

In Python, these calculations can be found in the statsmodels.stats.power module. I won't give a formal calculation here, although we may talk about power with respect to particular tests and techniques elsewhere in the course. The conceptual framework is more important.

Underpoweredness: pervasive and terrible.

Since messing around with significance levels is usually a bad idea in null hypothesis significance testing, the main way to increase power is to increase the sample size. But by how much? This is where researchers get into trouble.

You see, it's hard to actually do power calculations in real life, because *you don't know what the true effect size out in the world is.* How can you know the true effect size?

Think of a toy study: suppose you were testing the hypothesis that having a garage raises the price of a house. Seems pretty plausible, but by how much? Does having a garage increase the price of the house by 10,000 dollars? By 100,000 dollars? By fifty cents? By 5% of the price? This is effect size. And you don't know it until after you've done the research (and, even then, only if it's correct).

Unfortunately, many researchers in the real world over-estimate the power of their studies because they over-optimistically estimate their effect sizes.

This is particularly relevant to survey and experimental research. Carrying out this research can be super expensive, and so you need to know how many participants you need in order to have a realistic chance of seeing a result. Lots of research in fields like psychology is done with really small groups of participants, and this means that such studies tend to be *underpowered*—this is not a good thing. Statistical power is also relevant to observational research, but the nice thing about observational research using things like census data is that sample sizes tend to be nice and big.

As a consumer of research, underpowered studies are quite bad, because they mean that more published research is likely to be garbage.

To see this, let's think about the problem informally in Bayesian terms.

Suppose there are two disciplines, highpower and lowpower. Highpower customarily conducts studies with 80% power. Lowpower customarily conducts studies with 20% power. Both disciplines customarily find statistical significance with p=0.05.

Now let's suppose the base rate of true findings in both fields is 10%. That is, 10% of the things that researchers test in each field actually has a real effect. And suppose that both fields conduct 10,000 studies. Finally, let's suppose that in both fields there's a very large *publication bias*—research papers with negative findings are never published, and research papers with positive findings are always published. (This is a bit harsh—there is a big publication bias problem, but it's probably not *quite that bad*. But the simplifying assumption will be enough to do the work.)

In highpower, of those 10,000 studies, there will be 1000 true effects, and the researchers will find 800 of them. In addition, they'll find 500 false effects (type I errors).

In lowpower, by contrast, the researchers will only find 200 of the true effects. They'll find the same 500 false effects, because statistical power doesn't affect the false positive rate of an individual study.

But now you be a consumer of research from those two fields. In highpower, there's a $800/1300 = 62\%$ chance of a given research publication being true. In lowpower, there's only a $200/700 = 29\%$ chance of a given publication being true. So you can't really trust the stuff that shows up in the Journal of Lowerpower at all.

Moreover, the studies that do get published are likely to over-estimate the sizes of even true effects. Why? Well, suppose the true effect size out there in the population for what you're studying is ϕ. But random sampling error could mean that, in the sample for any given study, you could see an effect size of $\phi \pm \epsilon$. The problem is that because your study is so underpowered, you're much less likely to detect the effect in samples where it's $\phi - \epsilon$ and much more likely when it's $\phi + \epsilon$. And publication bias kicks in again to mean that in the real research literature we're likely to see over-inflated effect sizes. (And of course ϵ is basically going to be proportional to the sampling variance, so you can expect it to be bigger in the small samples that underpowered studies have anyway.)

With sufficiently small effect sizes, it might even be the case that the sign of the effect could flip, i.e., you think that garages decrease the prices of houses when they really increase it (in our toy equation above suppose $\epsilon > \phi$?!)

Something like this effect is doubtless a major contributor to problems like the replication crisis in psychology research.

Further reading:

- The UCLA Institute for Digital Research and Education has very clear tutorials on using a free program called G*Power to carry out a variety of power analysis tasks.
- Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson & Marcus R. Munafò Power failure: why small sample size undermines the reliability of neuroscience, Nature

Reviews Neuroscience 14: 365–376 (2013) is a detailed description of what can go terribly wrong when power is too low.

- Andrew Gelman & John Carlin, Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors Perspectives on Psychological Science 9(6): 641-651 (2014) describes the authors' efforts at retrospectively analyzing some underpowered studies in order to see how likely it is that they were drastically incorrect. See also this discussion of that paper, particularly the comment by a user named "amoeba" right before the answers.