2.2: The Power of Being Underpowered

After hearing all this, you might think calculations of statistical power are essential to medical trials. A scientist might want to know how many patients are needed to test if a new medication improves survival by more than 10%, and a quick calculation of statistical power would provide the answer. Scientists are usually satisfied when the statistical power is 0.8 or higher, corresponding to an 80% chance of concluding there's a real effect.

However, few scientists ever perform this calculation, and few journal articles ever mention the statistical power of their tests.

Consider a trial testing two different treatments for the same condition. You might want to know which medicine is safer, but unfortunately, side effects are rare. You can test each medicine on a hundred patients, but only a few in each group suffer serious side effects.

Obviously, you won't have terribly much data to compare side effect rates. If four people have serious side effects in one group, and three in the other, you can't tell if that's the medication's fault.

Unfortunately, many trials conclude with "There was no statistically significant difference in adverse effects between groups" without noting that there was insufficient data to detect any but the largest differences.[57] And so doctors erroneously think the medications are equally safe, when one could well be much more dangerous than the other.

You might think this is only a problem when the medication only has a weak effect. But no: in one sample of studies published between 1975 and 1990 in prestigious medical journals, 27% of randomized controlled trials gave negative results, but 64% of these didn't collect enough data to detect a 50% difference in *primary outcome* between treatment groups. Fifty percent! Even if one medication decreases symptoms by 50% more than the other medication, there's insufficient data to conclude it's more effective. And 84% of the negative trials didn't have the power to detect a 25% difference.[17, 4, 11, 16]

In neuroscience the problem is even worse. Suppose we aggregate the data collected by numerous neuroscience papers investigating one particular effect and arrive at a strong estimate of the effect's size. The median study has only a 20% chance of being able to detect that effect. Only after many studies were aggregated could the effect be discerned. Similar problems arise in neuroscience studies using animal models – which raises a significant ethical concern. If each individual study is underpowered, the true effect will only likely be discovered after many studies using many animals have been completed and analyzed, using far more animal subjects than if the study had been done properly the first time.[12]

That's not to say scientists are lying when they state they detected no significant difference between groups. You're just misleading yourself when you assume this means there is no *real* difference. There may be a difference, but the study was too small to notice it.

Let's consider an example we see every day.