18.3: Statistical Power

Remember from the previous chapter that under the Neyman-Pearson hypothesis testing approach, we have to specify our level of tolerance for two kinds of errors: False positives (which they called *Type I error*) and false negatives (which they called *Type II error*). People often focus heavily on Type I error, because making a false positive claim is generally viewed as a very bad thing; for example, the now discredited claims by Wakefield (1999) that autism was associated with vaccination led to anti-vaccine sentiment that has resulted in substantial increases in childhood diseases such as measles. Similarly, we don't want to claim that a drug cures a disease if it really doesn't. That's why the tolerance for Type I errors is generally set fairly low, usually at $\alpha = 0.05$. But what about Type II errors?

The concept of *statistical power* is the complement of Type II error – that is, it is the likelihood of finding a positive result given that it exists:

$$power = 1 - \beta$$

Another important aspect of the Neyman-Pearson model that we didn't discuss above is the fact that in addition to specifying the acceptable levels of Type I and Type II errors, we also have to describe a specific alternative hypothesis – that is, what is the size of the effect that we wish to detect? Otherwise, we can't interpret β – the likelihood of finding a large effect is always going to be higher than finding a small effect, so β will be different depending on the size of effect we are trying to detect.

There are three factors that can affect power:

- Sample size: Larger samples provide greater statistical power
- Effect size: A given design will always have greater power to find a large effect than a small effect (because finding large effects is easier)
- Type I error rate: There is a relationship between Type I error and power such that (all else being equal) decreasing Type I error will also decrease power.

We can see this through simulation. First let's simulate a single experiment, in which we compare the means of two groups using a standard t-test. We will vary the size of the effect (specified in terms of Cohen's d), the Type I error rate, and the sample size, and for each of these we will examine how the proportion of significant results (i.e. power) is affected. Figure 18.4 shows an example of how power changes as a function of these factors.

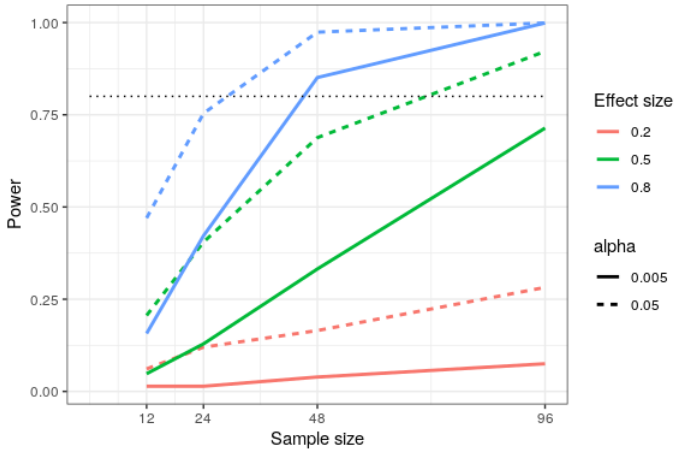

Figure 18.4: Results from power simulation, showing power as a function of sample size, with effect sizes shown as different colors, and alpha shown as line type. The standard criterion of 80 percent power is shown by the dotted black line.

This simulation shows us that even with a sample size of 96, we will have relatively little power to find a small effect ($d = 0.2$) with $\alpha = 0.005$. This means that a study designed to do this would be *futile* – that is, it is almost guaranteed to find nothing even if a true effect of that size exists.

There are at least two important reasons to care about statistical power, one of which we discuss here and the other of which we will return to in Chapter 32. If you are a researcher, you probably don't want to spend your time doing futile experiments. Running an underpowered study is essentially futile, because it means that there is a very low likelihood that one will find an effect, even if it exists.

18.3.1 Power analysis

Fortunately, there are tools available that allow us to determine the statistical power of an experiment. The most common use of these tools is in planning an experiment, when we would like to determine how large our sample needs to be in order to have sufficient power to find our effect of interest.

Let's say that we are interested in running a study of how a particular personality trait differs between users of iOS versus Android devices. Our plan is collect two groups of individuals and measure them on the personality trait, and then compare the two groups using a t-test. In order to determine the necessary sample size, we can use the `pwr.t.test()` function from the `pwr` library:

```
##
##      Two-sample t test power calculation
##
##               n = 64
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

This tells us that we would need at least 64 subjects in each group in order to have sufficient power to find a medium-sized effect. It's always important to run a power analysis before one starts a new study, to make sure that the study won't be futile due to a sample that is too small.

It might have occurred to you that if the effect size is large enough, then the necessary sample will be very small. For example, if we run the same power analysis with an effect size of d=2, then we will see that we only need about 5 subjects in each group to have sufficient power to find the difference.

```
##
##      Two-sample t test power calculation
##
##               n = 5.1
##               d = 2
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

However, it's rare in science to be doing an experiment where we expect to find such a large effect – just as we don't need statistics to tell us that 16-year-olds are taller than than 6-year-olds. When we run a power analysis, we need to specify an effect size that is plausible for our study, which would usually come from previous research. However, in Chapter 32 we will discuss a phenomenon known as the "winner's curse" that likely results in published effect sizes being larger than the true effect size, so this should also be kept in mind.