Statistical Power and Sample Size Computation:

Overview:

We often want to have an idea of statistical power before we initiate a new study.

Suppose we propose a study where our hypothesis is that mice fed a high-fat diet (chow A) have higher glucose levels on average than mice fed a low-fat diet (chow B). We are hoping to use n=15-20 mice per diet group, for a total of 30-40 mice. Before we commit time and resources, we want to estimate power to make sure that this is a large enough sample to detect an effect.

What is statistical power?

Before we define power, let's review **hypothesis testing**.

In the example above, the outcome of interest is glucose level (continuous).

Formally, we can write:

Ho (null hypothesis): There is no difference in average glucose levels between mice fed chow A and mice fed chow B.

Ha (alternative hypothesis): There is a difference in average glucose levels between mice fed chow A and mice fed chow B.

Note that the alternative hypothesis is *two-sided*, meaning we do not pre-specify which direction the difference is in (average glucose lower or higher in chow A group). Two-sided hypotheses are <u>standard</u>, even if you have a sense of which direction the difference will be in.

Now that we have our null and alternative hypotheses formally stated, let's define two more terms – Type I and Type II error.

Type I error is the probability that you falsely reject the null hypothesis. Even though there is no true relationship between glucose levels and diet type, you observed a p-value < 0.05 (alpha). Note that we can use a different alpha level, but 0.05 is standard. If we use an alpha level of 0.01 then we have a higher burden of proof to reject the null hypothesis (i.e.: a more extreme test statistic)

Type II error is the probability that you falsely fail to reject the null hypothesis. Even though there is a relationship between type of chow and glucose levels, you observed a p-value > 0.05. Note that if the alpha level was 0.01, it would be harder to reject the null hypothesis and the probability of Type II error would be higher.

Power is 1 minus the probability of Type II error. It reflects the proportion of times you would get a statistically significant test statistic (i.e.: a p-value < 0.05) given a specific alternative hypothesis is true. When you compute power, *you are specifying an alternative hypothesis and computing the proportion of times you would correctly reject the null hypothesis under this specific scenario.*

How do I actually compute power?

First, think about what statistical test you would perform to analyze your data. A biostatistician can help you with this (See the **Getting Started** section on BERD House). In our example, we know that we can assess if the average glucose level is different between the two groups with a **two-sample t-test** or nonparametric **Mann Whitney test**.

The next step is to think about a *reasonable alternative hypothesis*. Often, you can read the literature to see what the average glucose levels and standard deviation are in mice under standard conditions and then hypothesize what a reasonable change in glucose levels would be for the high-fat diet under consideration. This is usually the most difficult part of a power computation! You want to make sure that whatever differences and variation you are assuming are as realistic as possible. If your particular scenario is unlikely to hold in real world conditions, the computed power based on those assumptions is not going to be informative.

For example, if we know that in previous studies, the average glucose levels in mice fed a standard low-fat diet is normally distributed with a mean level of 115 mg/dL and the standard deviation is 30 mg/dL, we may assume for our power computation that a high-fat diet may cause a spike in blood glucose with a mean of 140 mg/dL and assume that the standard deviation will stay constant at 30 mg/dL. We are thus making 3 important assumptions here:

1) Blood glucose levels are normally distributed within each diet group

2) The difference in average levels is 25 mg/dL.

3) The standard deviation of glucose levels within each diet group is 30 mg/dL.

*Based on assumptions 2) and 3) Cohen's d effect size (mean/SD) is computed to be 25/30 = 0.833

Based on this specific scenario, and a sample size of n=15 per diet group, we compute a power of 60%. **For a sample size of n=20 per diet group, we compute a power of 73%.**

Adequate power is usually 80% or higher. This means that under the specific scenario assumed, we would likely be **underpowered** to detect a difference in average glucose levels. To increase power to 80% or above, we would need to **increase sample size** to n=23 per diet group. Although it is tempting to lower the standard deviation or increase the expected difference in average glucose levels in these computations to compute higher statistical power with a smaller sample size, remember that your assumptions need to align with reality as closely as possible, or the power computation will not be valid. You also should not increase alpha much higher than 0.05 in many circumstances otherwise you run an unacceptably higher risk of a false positive result. It is okay, however, to provide a range of computed power under varying assumptions.

Can I compute power using online tools?

There are many excellent online resources (see **here** and **here**, for example) that will allow you to compute power for different study designs and scenarios. Of course, if your particular design is complex, speak with a biostatistician. We may compute power in these cases via simulation, which provides a lot of flexibility compared to online tools.