What is Statistical Power?

When performing hypothesis tests a critical concept to be aware of is power. Statistical power can be thought of as the ability of your test to detect an effect or a difference, between the two values you are testing, when the difference actually exists.

There are four possible outcomes when you carry out a hypothesis test. Two of the possibilities are that the null hypothesis is true and you accept the null hypothesis, or the null hypothesis is true and you reject the null hypothesis. The other outcomes are that the null hypothesis is false and you accept the null hypothesis or the null hypothesis is false and you reject the null hypothesis. Power refers to the last possibility. In reality the null hypothesis truly is false, and your test correctly rejected it. The idea of detecting a difference refers to the fact that when the null hypothesis is false, there is an actual difference in the two values you are testing. So power is more than simply making the correct decision. When the null hypothesis is true and you accept the null hypothesis you have made the correct decision. Yet the null hypothesis being true does not result in a difference, only when the null hypothesis is false is there an indication of a difference. Power specifically refers to detecting this difference, or effect.

So naturally we want whatever test we are using to have a lot of power. If our test doesn’t have very much power we are likely to accept our null hypothesis when in reality the opposite is true. This is known as a Type II error. One factor that affects power is sample size. A larger sample size will allow your test to be better at detecting a difference in your data, and therefore have more power.

Suppose that you want to determine during a card game if the deck of cards is rigged – you have reason to suspect that two cards have been removed and replaced with a two of hearts so there are now a total of three two of hearts in the deck. Your opponent will not let you observe the whole deck but your opponent will let you take a sample of cards from the deck to view. How many cards would you need to sample to detect that there are too many two’s? If you only took three cards and none of them are two’s are you convinced that the deck is fair? What if you were able to take a sample of 20 cards or 30 cards? As your sample size of cards increases and you do not see an unusual amount of two’s showing up, you probably become more confident that the deck is fair. In this case, the null hypothesis is that the deck is fair, and power would be defined as deciding that the deck is rigged when in reality that is the case.

Let’s apply this to another example. You randomly sample 20 people from a population and were hoping for a 75% positive response to a question. The results of your survey show that 13 of 20 people responded yes. This is only 65%, well below what you hoped. At this point you decide that the true population proportion is less than 0.75. Is this a wise decision? Think back to the deck of cards. When you took three cards from the deck and none of them were two’s was that enough to convince you of a fair deck? Probably not. In the same way maybe 20 people should not be enough to convince you that the true proportion is less than 0.75. If your population size is hundreds or thousands of people, you will want to survey many more than 20 people before you begin to feel confident you have enough statistical power.

This is not to say though, that you should keep increasing power or repeating tests until you get the result you want. You must always be willing to accept undesirable results, but having a good understanding of power will go a long way in helping you determine the truth.