3: Pseudoreplication- Choose Your Data Wisely

Many studies strive to collect more data through replication: by repeating their measurements with additional patients or samples, they can be more certain of their numbers and discover subtle relationships that aren't obvious at first glance. We've seen the value of additional data for improving statistical power and detecting small differences. But what exactly counts as a replication?

Let's return to a medical example. I have two groups of 100 patients taking different medications, and I seek to establish which medication lowers blood pressure more. I have each group take the medication for a month to allow it to take effect, and then I follow each group for ten days, each day testing their blood pressure. I now have ten data points per patient and $1,000$ data points per group.

Brilliant! $1,000$ data points is quite a lot, and I can fairly easily establish whether one group has lower blood pressure than the other. When I do calculations for statistical significance I find significant results very easily.

But wait: we expect that taking a patient's blood pressure ten times will yield ten very similar results. If one patient is genetically predisposed to low blood pressure, I have counted his genetics ten times. Had I collected data from $1,000$ independent patients instead of repeatedly testing 100, I would be more confident that differences between groups came from the medicines and not from genetics and luck. I claimed a large sample size, giving me statistically significant results and high statistical power, but my claim is unjustified.

This problem is known as pseudoreplication, and it is quite common.[38] After testing cells from a culture, a biologist might "replicate" his results by testing more cells from the same culture. Neuroscientists will test multiple neurons from the same animal, incorrectly claiming they have a large sample size because they tested hundreds of neurons from just two rats.

In statistical terms, pseudoreplication occurs when individual observations are heavily dependent on each other. Your measurement of a patient's blood pressure will be highly related to his blood pressure yesterday, and your measurement of soil composition here will be highly correlated with your measurement five feet away. There are several ways to account for this dependence while performing your statistical analysis:

1. Average the dependent data points. For example, average all the blood pressure measurements taken from a single person. This isn't perfect, though; if you measured some patients more frequently than others, this won't be reflected in the averaged number. You want a method that somehow counts measurements as more reliable as more are taken.
2. Analyze each dependent data point separately. You could perform an analysis of every patient's blood pressure on day 5, giving you only one data point per person. But be careful, because if you do this for every day, you'll have problems with multiple comparisons, which we will discuss in the next chapter.
3. Use a statistical model which accounts for the dependence, like a hierarchical model or random effects model.

It's important to consider each approach before analyzing your data, as each method is suited to different situations. Pseudoreplication makes it easy to achieve significance, even though it gives you little additional information on the test subjects. Researchers must be careful not to artificially inflate their sample sizes when they retest samples.