6.1: Rules of the Game

Medical trials are expensive. Supplying dozens of patients with experimental medications and tracking their symptoms over the course of months takes significant resources, and so many pharmaceutical companies develop “stopping rules,” which allow investigators to end a study early if it’s clear the experimental drug has a substantial effect. For example, if the trial is only half complete but there’s already a statistically significant difference in symptoms with the new medication, the researchers may terminate the study, rather than gathering more data to reinforce the conclusion.

When poorly done, however, this can lead to numerous false positives.

For example, suppose we’re comparing two groups of patients, one with a medication and one with a placebo. We measure the level of some protein in their bloodstreams as a way of seeing if the medication is working. In this case, though, the medication causes no difference whatsoever: patients in both groups have the same average protein levels, although of course individuals have levels which vary slightly.

We start with ten patients in each group, and gradually collect more data from more patients. As we go along, we do a t test to compare the two groups and see if there is a statistically significant difference between average protein levels. We might see a result like this simulation:

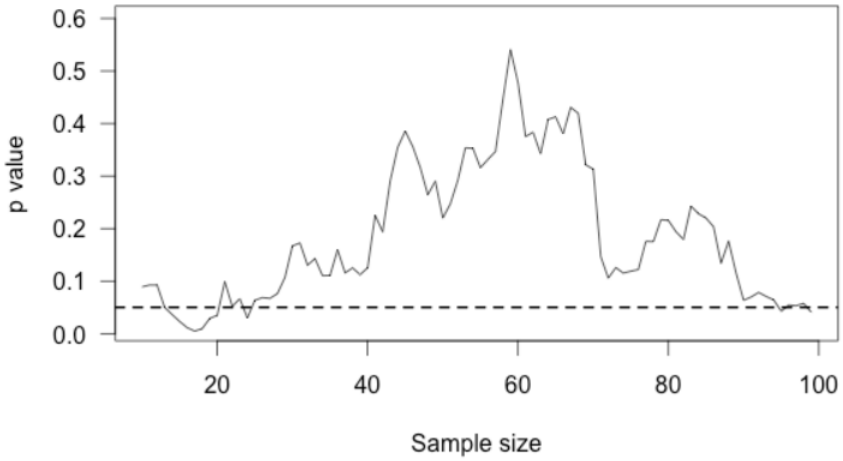

Figure 6.1.1

This plot shows the p value of the difference between groups as we collect more data, with the horizontal line indicating the $p = 0.05$ level of significance. At first, there appears to be no significant difference. Then we collect more data and conclude there is. If we were to stop, we’d be misled: we’d believe there is a significant difference between groups when there is none. As we collect yet more data, we realize we were mistaken – but then a bit of luck leads us back to a false positive.

You’d expect that the p value dip shouldn’t happen, since there’s no real difference between groups. After all, taking more data shouldn’t make our conclusions worse, right? And it’s true that if we run the trial again we might find that the groups start out with no significant difference and stay that way as we collect more data, or start with a huge difference and quickly regress to having none. But if we wait long enough and test after every data point, we will eventually cross *any* arbitrary line of statistical significance, even if there’s no real difference at all. We can’t usually collect infinite samples, so in practice this doesn’t always happen, but poorly implemented stopping rules still increase false positive rates significantly.[53]

Modern clinical trials are often required to register their statistical protocols in advance, and generally pre-select only a few evaluation points at which they test their evidence, rather than testing after every observation. This causes only a small increase in the false positive rate, which can be adjusted for by carefully choosing the required significance levels and using more advanced statistical techniques.[56] But in fields where protocols are not registered and researchers have the freedom to use whatever methods they feel appropriate, there may be false positive demons lurking.