5.1: Differences in Significance

“We compared treatments A and B with a placebo. Treatment A showed a significant benefit over placebo, while treatment B had no statistically significant benefit. Therefore, treatment A is better than treatment B.”

We hear this all the time. It’s an easy way of comparing medications, surgical interventions, therapies, and experimental results. It’s straightforward. It seems to make sense.

However, a difference in significance does not always make a significant difference.[22]

One reason is the arbitrary nature of the $p < 0.05$ cutoff. We could get two very similar results, with $p = 0.04$ and $p = 0.06$, and mistakenly say they’re clearly different from each other simply because they fall on opposite sides of the cutoff. The second reason is that p values are not measures of effect size, so similar p values do not always mean similar effects. Two results with identical statistical significance can nonetheless contradict each other.

Instead, think about statistical power. If we compare our new experimental drugs Fixitol and Solvix to a placebo but we don’t have enough test subjects to give us good statistical power, then we may fail to notice their benefits. If they have identical effects but we have only 50% power, then there’s a good chance we’ll say Fixitol has significant benefits and Solvix does not. Run the trial again, and it’s just as likely that Solvix will appear beneficial and Fixitol will not.

Instead of independently comparing each drug to the placebo, we should compare them against each other. We can test the hypothesis that they are equally effective, or we can construct a confidence interval for the extra benefit of Fixitol over Solvix. If the interval includes zero, then they could be equally effective; if it doesn’t, then one medication is a clear winner. This doesn’t improve our statistical power, but it does prevent the false conclusion that the drugs are different. Our tendency to look for a difference in significance should be replaced by a check for the significance of the difference.

Examples of this error in common literature and news stories abound. A huge proportion of papers in neuroscience, for instance, commit the error.[44] You might also remember a study a few years ago suggesting that men with more biological older brothers are more likely to be homosexual.[9] How did they reach this conclusion? And why older brothers and not older sisters?

The authors explain their conclusion by noting that they ran an analysis of various factors and their effect on homosexuality. Only the number of older brothers had a statistically significant effect; number of older sisters, or number of nonbiological older brothers, had no statistically significant effect.

But as we’ve seen, that doesn’t guarantee that there’s a significant difference between the effects of older brothers and older sisters. In fact, taking a closer look at the data, it appears there’s no statistically significant difference between the effect of older brothers and older sisters. Unfortunately, not enough data was published in the paper to allow a direct calculation.[22]