4.6: Controlling the False Discovery Rate

I mentioned earlier that techniques exist to correct for multiple comparisons. The Bonferroni procedure, for instance, says that you can get the right false positive rate by looking for $p < 0.05/n$, where n is the number of statistical tests you're performing. If you perform a study which makes twenty comparisons, you can use a threshold of $p < 0.0025$ to be assured that there is only a 5% chance you will falsely decide a nonexistent effect is statistically significant.

This has drawbacks. By lowering the p threshold required to declare a result statistically significant, you decrease your statistical power greatly, and fail to detect true effects as well as false ones. There are more sophisticated procedures than the Bonferroni correction which take advantage of certain statistical properties of the problem to improve the statistical power, but they are not magic solutions.

Worse, they don't spare you from the base rate fallacy. You can still be misled by your p threshold and falsely claim there's "only a 5% chance I'm wrong" – you just eliminate some of the false positives. A scientist is more interested in the false discovery rate: what fraction of my statistically significant results are false positives? Is there a statistical test that will let me control this fraction?

For many years the answer was simply "no." As you saw in the section on the base rate fallacy, we can compute the false discovery rate if we make an assumption about how many of our tested hypotheses are true – but we'd rather find that out from the data, rather than guessing.

In 1995, Benjamini and Hochberg provided a better answer. They devised an exceptionally simple procedure which tells you which p values to consider statistically significant. I've been saving you from mathematical details so far, but to illustrate just how simple the procedure is, here it is:

1. Perform your statistical tests and get the p value for each. Make a list and sort it in ascending order.
2. Choose a false-discovery rate and call it q. Call the number of statistical tests m.
3. Find the largest p value such that $p \leq iq/m$, where i is the p value's place in the sorted list.
4. Call that p value and all smaller than it statistically significant.

You're done! The procedure guarantees that out of all statistically significant results, no more than q percent will be false positives.[7]

The Benjamini-Hochberg procedure is fast and effective, and it has been widely adopted by statisticians and scientists in certain fields. It usually provides better statistical power than the Bonferroni correction and friends while giving more intuitive results. It can be applied in many different situations, and variations on the procedure provide better statistical power when testing certain kinds of data.

Of course, it's not perfect. In certain strange situations, the Benjamini-Hochberg procedure gives silly results, and it has been mathematically shown that it is always possible to beat it in controlling the false discovery rate. But it's a start, and it's much better than nothing.