1.2: The Power of p Values

Statistics provides the answer. If we know the *distribution* of typical cold cases – roughly how many patients tend to have short colds, or long colds, or average colds – we can tell how likely it is for a random sample of cold patients to have cold lengths all shorter than average, or longer than average, or exactly average. By performing a statistical test, we can answer the question "If my medication were completely ineffective, what are the chances I'd see data like what I saw?"

That's a bit tricky, so read it again.

Intuitively, we can see how this might work. If I only test the medication on one person, it's unsurprising if he has a shorter cold than average – about half of patients have colds shorter than average. If I test the medication on ten million patients, it's pretty damn unlikely that *all* of them will have shorter colds than average, *unless my medication works.*

The common statistical tests used by scientists produce a number called the p value that quantifies this. Here's how it's defined:

> *The P value is defined as the probability, under the assumption of no effect or no difference (the null hypothesis), of obtaining a result equal to or more extreme than what was actually observed.*[24]

So if I give my medication to 100 patients and find that their colds are a day shorter on average, the p value of this result is the chance that, if my medication didn't do anything at all, my 100 patients would randomly have, on average, day-or-more-shorter colds. Obviously, the p value depends on the size of the effect – colds shorter by four days are less likely than colds shorter by one day – and the number of patients I test the medication on.

That's a tricky concept to wrap your head around. A p value is not a measure of how right you are, or how significant the difference is; it's a measure of *how surprised you should be* if there is no actual difference between the groups, but you got data suggesting there is. A bigger difference, or one backed up by more data, suggests more surprise and a smaller p value.

It's not easy to translate that into an answer to the question "is there really a difference?" Most scientists use a simple rule of thumb: if p is less than 0.05, there's only a 5% chance of obtaining this data unless the medication really works, so we will call the difference between medication and placebo "significant." If p is larger, we'll call the difference insignificant.

But there are limitations. The p value is a measure of surprise, not a measure of the size of the effect. I can get a tiny p value by either measuring a huge effect – "this medicine makes people live four times longer" – or by measuring a tiny effect with great certainty. Statistical significance does not mean your result has any *practical* significance.

Similarly, statistical *in*significance is hard to interpret. I could have a perfectly good medicine, but if I test it on ten people, I'd be hard-pressed to tell the difference between a real improvement in the patients and plain good luck. Alternately, I might test it on thousands of people, but the medication only shortens colds by three minutes, and so I'm simply incapable of detecting the difference. A statistically insignificant difference does not mean there is no difference at all.

There's no mathematical tool to tell you if your hypothesis is true; you can only see whether it is consistent with the data, and if the data is sparse or unclear, your conclusions are uncertain.

But we can't let that stop us.