Hypothesis Testing: Conceptual Introduction (draft)

Now that we understand distributions and the central limit theorem, we're in a good position to make sense of the notion of a hypothesis test. It's actually very simple.

Suppose you do an experiment. Let's say you want to find out whether a company is engaging in racial discrimination in interviewing, so you manufacture a bunch of resumes and then send them out with randomly changed names, street addresses (in a racially segregated city), and other markers of racial identification. This is a kind of study that economists actually do.

Now suppose you find out that, say, white-appearing resumes get interviews 10% of the time, while black-appearing resumes get interviews 8% of the time. Is this fact alone enough to show that the company engages in discrimination?

Well, one worry you might have is that this difference is just the result of bad luck. Maybe, just by sheer coincidence, more of the black-appearing resumes landed on harsher recruiters' desks, or showed up in the mail just before lunch, when people were hungry and impatient and crabby. You need statistics to quantify the likelihood that the difference you saw was just the result of chance.

So here's one way you might think of the problem. Let's assume what we saw, i.e., that the real-world (population) average (that is, mean) number of interviews that white-appearing applicants get per 100 resumes is 10. Suppose as a hypothesis (this is the famous **null hypothesis**) that the real-world mean number of interviews that black-appearing resumes get is *also* 10. How likely is it that we would have actually seen black-appearing resumes only yield 8 interviews in a sample?

The central limit theorem gets us most of the way to the answer: recall that it says that the sample mean will be normally distributed around the mean of the actual population. So *if* the population mean of interviews for black-appearing resumes is actually 10 (that is, if there's no discrimination), how likely is it that we'll actually see 8 in our sample? Well, that depends on the standard deviation of the sample mean.

The calculation of that quantity is a topic for another lesson. (This is just a conceptual introduction, remember?) The important thing to know right now is that it will get smaller as sample size increases. (Think about that in light of our lesson on the central limit theorem again: with a larger sample size, it's more likely that the sample mean will correspond to the population mean; so a population full of large sizes will on the whole have less dispersion around the sample mean than a population full of small sample sizes will. This is just the law of large numbers.)

So let's suppose that the standard deviation of the sample mean is 0.5. Then

our observation is four standard deviations out from the mean. We know from the properties of the normal distribution that if the mean of interviews for black-appearing resumes really is 10 and the standard deviation of the sample means is .5, only a very small percentage of possible samples would be as low as 8. So we have some good evidence that discrimination is happening.

By contrast, if the standard deviation is 2, well, then it's only one standard deviation away. We know that in a normal distribution upwards of 30% of observations will fall one standard deviation or more away–so our result shouldn't give us much reason to think that discrimination is happening.

This is the essence of hypothesis testing. In its classical form, we pick a threshold value—95% is a common one—and we figure out the distribution that applies to the statistic (like mean, or variance, or whatever) that we're trying to figure out (it isn't always normal, because we aren't always just trying to figure out the mean, and actually we usually use a slightly different distribution called the t distribution anyway—more on this later).

Then we pick a null hypothesis—a description of the quantity we would expect to see if there's nothing interesting going on (no discrimination, the drug has no effect, etc), and we look at our data.

If the statistic in question is far enough away from what we would expect the statistic to be under the null hypothesis, then we count that as evidence that the null hypothesis isn't true—that there is something going on. "Far enough" is defined by our threshold value: if we pick a value of 95%, it means that we'll only count the data as evidence of something if the statistic of interest is so extreme that it's outside the range that we would expect 95% of samples to fall in.

The infamous p-value

That 95% example we just gave is also the source of the "p-value." Once we see how many standard deviations our sample is from the null hypothesis statistic, we know how big a chunk of the distribution of possible observations the observation we got is out of. And we get our p-value from that.

Here's an example. If we see a mean that is two standard deviations away from the null hypothesis mean, we know that we would expect 95.4% of observations to be closer to the null mean. To get a p-value, we just turn that to a decimal and subtract it from 1: we would say that we have a p value of 0.046. With a standard significance threshold of 0.05 (the equivalent of our 95% threshold value above), we'd say that the p-value is low enough for statistical significance—for us to think we've detected a real effect.

The story I just told is a very conventional, classical, approach to hypothesis testing. Unfortunately, it tends to be misinterpreted and misused a lot. We'll

talk more down the line about how this goes wrong, but here is one quick warning to start with

Watch out for p-hacking.

What happens if you take a bunch of samples? For example, suppose you have 20 different ways to formulate the null hypothesis or to analyze the data, and you try them all. 19 of them don't turn out to be significant at an 0.05 level, but one of them does. Does this mean you have a real result? You should be able to see that it probably doesn't: if an 0.05 significance level means that you should see a result this extreme 5% of the time even if the null hypothesis is true, well, dude, you just did 20 tests, and you saw it once... that is 5% of the time! You have effectively no evidence here.

People intentionally or unintentionally cheat like this all the time—it's called **p-hacking**. Another way of p-hacking is to not have a "stopping rule"—to analyze the data while you're collecting it, and then to stop collecting data when you see a significant result. Many scientists "preregister" their studies by publicly committing to the data they will collect and the analyses they will run in advance. This is a good practice to head off p-hacking. This blog post has a good explanation of p-hacking and preregistration (And actually, even if you don't cheat like this, the chances that you incorrectly rejected the null hypothesis even with a p value of 0.05 is still substantially higher than 5%, because of Bayes Rule: see this blog post for an ex