A Verbose Explanation of Hypothesis Testing

Owen Pauptit

1 Introduction

With hypothesis testing, the key thing is to realise what is random and what is not. To me, hypothesis testing was always very confusing until I fully understood this distinction, so I've written this explanation in a long-winded way to try to drill in what is random and what is not. After understanding the source of the randomness, the wording should come naturally.

2 Randomness

Let's start with dice, as an initial example. Don't worry if this example doesn't make perfect sense to begin with — just give it a read, see if you get the gist and move on.

Each roll of a fair 6-sided die is random, i.e. it makes sense to say that the probability of the next roll being a 3 is $\frac{1}{6}$.

If I roll the die and see that it was a 4, there is no longer any probability associated with that roll. I rolled a 4, not a 3. We cannot talk about the probability that I rolled a 3 because it didn't happen. Importantly, this is different from having a zero probability of rolling a 3. "I rolled a 3" is simply a false statement, not a probabilistic one.

In statistics, the data we have already collected are no longer random — we know for certain what we recorded*. What is random is the data collection process. Likewise, statements about populations as a whole are generally true or false, but not probabilistic.

3 Hypothesis Testing

"The average height of the MSc cohort is 165 cm" is either a true or false statement. Given enough time and resources, we could accurately find out the average height of the cohort. It does not make sense to say something like "the probability that the average height of the MSc cohort is 165 cm is 95%" because

there is no randomness involved.

Say we took a sample of 5 students and the average height among them was 170cm. This would be possible if the average height of the cohort is 165cm or if it is 170cm. The key thing is that we don't know what it is. The average height of the cohort is fixed but unknown.

This situation is one where we could perform a hypothesis test, with a null hypothesis that the average height of the cohort is 165 cm and an alternative hypothesis that the height of the cohort is not 165 cm. As mentioned earlier, these two hypotheses are statements that can be either true or false, and we cannot say anything probabilistic about them.

Now we are a bit stuck. Our hypotheses are fixed (not random), and so are our data. The only randomness here is the data collection process and our future data.

This is where hypothesis tests get strange and confusing. We start by assuming our null hypothesis is true, regardless of whether it is or not – i.e. assume that the average height of the MSc cohort is 165 cm. Then, we can work out the probability that a (future) random sample of 5 students from the cohort has an average height greater (more extreme) than the data we collected: 170 cm. This probability is known as the p-value.

If the p-value is very small, we can reject the null hypothesis because the likelihood of seeing new data at least as extreme as our data is very small if the null hypothesis is, in fact, true. We might say something like "we have evidence at the x% significance level to reject the null hypothesis" or "there is evidence at the x%significance level that the average height of the cohort is not 165 cm".

If the p-value is large, we cannot reject the null hypothesis because there is a decent chance of seeing data as extreme as what we have already seen, if the null hypothesis is indeed true. Importantly, this does NOT mean that the null hypothesis is true, nor does it mean that the alternative hypothesis is false. We only fail to reject the null hypothesis. We might say something like "there is no evidence at the 5% significance level that the average height is not 165 cm".

You have to be careful with the wording here. The sentence above looks like it has a double-negative, but it does not. Not having evidence that someone is innocent (i.e. not guilty) does not mean that you have evidence that they are guilty.

*In the case of measurement error, we still know what we recorded — we don't, however, know for certain what the real value was (which is now fixed but unknown).