Hypothesis Testing

In these notes, we discuss a very important but often misunderstood topic: hypothesis testing.

The logic of all hypothesis tests is exactly the same, but at the same time quite subtle. Many students try to memorise the steps for each test in a robotic manner, without full understanding; but these students often then struggle with more difficult problems, and have to memorise all the steps again when they learn a new kind of test. It can also be quite tricky to remember which way around all the tests go if you don't really know what you are doing.

Learning fixed procedures certainly has its place, but rather than taking a purely formulaic approach we will aim to also understand the underlying logic fully. For most students, this requires a lot of repetition until the key ideas 'sink in'. It is also helpful to try to *care deeply* about the result, as far as possible. Hypothesis testing can be quite hard to get a grasp of it if remains merely abstract, so imagine you have a genuine practical interest in the outcome; that something important is at stake, or there is a burning question you really need to get to the bottom of.

Flipping Coins

We introduce the main idea of hypothesis testing in a gentle, non-mathematical way, aiming to bring the key principle to light with a concrete example. Consider the following story.

Imagine that you regularly play tennis, and your tennis club have matched you with a new partner. At the start of each match, your partner flips a coin to see who will serve first. However, on first impressions your partner seems a bit shifty and not very trustworthy. Moreover, you notice they always use exactly the same coin, and are keen to impose the rule that if it shows 'Heads' then it will be they who serve first.

After a while, you realise they seem to be serving first every time! You start to suspect your opponent of foul play; perhaps they are using a weighted coin – or one with 'Heads' on both sides, like that used by the Batman character, Harvey 'Two Face' Dent.

After some careful deliberation, you pluck up the courage to confront your opponent about the coin. In response, they make the following claim:

> 'This is a fair coin. The probability it comes out heads is 50%; same as for tails.'

Unsatisfied with merely verbal assurance, you get them to agree that together you will *test* the claim they've made. How might you go about this?

A natural approach is to flip the coin a bunch of times and see what happens. Suppose you do so, and this is the result:

How would you react to this outcome? Pause here and think about what your conclusion would be. If you are now convinced the coin was unfair, try to explain *exactly* why you believe this. Be as clear and precise as you are able, since this is the central point. I'll wait! Here's a scary picture of Two Face in the meantime.

Following the logic of hypothesis testing, we would in this case ordinarily conclude that the coin is *not* fair. The line of reasoning supporting this decision can be summarised in a single short sentence, but it is important to get the logic exactly right. Suppose you respond to your partner:

> 'If the coin had been fair, then this outcome would be very unlikely;
> so, the coin can't be fair after all.'

You may have thought something similar earlier – and if not, hopefully this response to the partner seems fairly natural. Yet the exact logical principle that we need to pick up on here is extremely specific, and deceptively hard to grasp fully. So, let's unpack what we have said a little further.

Our response to the partner is about a hypothetical possibility; what would

be the case *if the coin were fair*. This point cannot be emphasised enough. But I'll try to do so, using lots of italics. Think *hypothe*sis testing = *hypothe*tical reasoning. We are first imagining that the coin *is fair*, and then considering *how likely* the specific results we actually got would be *if this were the case*.

If we then find that our results *would be* very unusual in the event of a fair coin, then – exiting our hypothetical daydream where it *is* fair and returning to the real world – we conclude that actually, the coin *cannot be fair after all.* Let's write this out explicitly:

- **Premiss 1**: *If* the coin were fair, getting 10 heads in a row *would be* very unlikely
- **Premiss 2**: When we did the test, we *actually did get* 10 heads in a row
- **Conclusion**: Therefore, the coin *cannot be fair after all.*

To spoil the story slightly by bringing some maths into it, there is a reason the logic of testing the partner's claim needs to be framed *precisely* like this. This is because we need to *assume* something about the probability of getting heads on the coin in order to *do any calculations*. We now explore these calculations in detail.

Let's call the probability of getting heads on the coin p. Then according to our partner's claim, $p = \frac{1}{2}$. Of course, this may or may not actually be true. But we can say, hypothetically, that *if* it were true, the probability of getting Heads 10 times in a row would be $\frac{1}{2} \times \frac{1}{2} \times ... \times \frac{1}{2} = \left(\frac{1}{2}\right)^{10}$. Numerically, this is about 0.000977. So, *if* the coin were fair, we would have essentially just witnessed a miracle!

Suppose now you present these calculations to your partner. After some rapid blinking, they respond as follows:

> 'Sure, that's a low number. But *any* sequence of H and T is equally unlikely. For instance, the sequence $H, T, H, T, H, T, H, T, H, T$ also occurs with probability 0.000977. So, the test doesn't prove *anything*.'

How would you reply? It seems intuitively that there is a big difference between the sequences $H, H, H, H, H, H, H, H, H, H$ and $H, T, H, T, H, T, H, T, H, T$, although they are indeed equally likely given that $p = \frac{1}{2}$.

A next thought that might occur to us here is that it is not the actual *sequence* that matters, but the *number of* heads obtained. After all, in the latter sequence, it is easy to show that 5 would be the *expected number* of heads if $p = \frac{1}{2}$; but getting 10 heads seems a little extreme in this case. Of course, if instead $p = 1$, getting 10 heads is not an extreme outcome at all!

In response to this development of our case, the partner makes a final protest:

> 'Okay, this result was extreme; but there are other extreme results it could have been too.'

We may concede that it was a little unfair to focus only on the exact outcome we got. Rather, we should perhaps focus on the chances of getting results *as extreme as this one, or even more extreme*, again in the case that $p = \frac{1}{2}$. Yet the only other outcome that would be equally extreme would be getting 0 heads. This would also be very unlikely *if* the coin were fair.

More explicitly: letting X be the number of heads, we may calculate the 'fairer' probability we suggested as follows:

$$\mathbb{P}(X \geq 10 \text{ or } X \leq 0 \mid p = \tfrac{1}{2}) = \left(\tfrac{1}{2}\right)^{10} + \left(\tfrac{1}{2}\right)^{10} \cong 0.00195$$

Note the conditioning on $p = 1/2$ in the first expression here, the claim we're testing.

Now, we may retort that this number is still low enough to cast a strong doubt on our partner's claim. Or, instead of making this concession at all, we may instead choose to persist with our original calculation, saying that an extreme outcome in the *positive* direction is precisely the kind of possibility that we had been suspicious about in the first place; so, in fact it was fine to just look at $\mathbb{P}(X \geq 10 \mid p = 1/2)$, and this is just the same as $\mathbb{P}(X = 10 \mid p = 1/2)$ after all. So, way the test should be conducted depends on our initial intentions in conducting it.

These disputes about how the test should operate can be tricky – but actually, what really matters for drawing valid conclusions is that the testing procedure is fixed in advance. We shouldn't make up the rules once we've already done the test and looked at the results! Unfortunately, it's too late for that now, since we've already done the test. So, we just have to hope that our partner agrees that looking at how 'extreme' the results are (either in one direction, or in both) is the natural thing to do. But to avoid further potential arguments, in future we should aim to clarify in advance how our testing procedure will work.

Terminology

We close by introducing some terminology to summarise what we've done here. The partner's claim that we're testing, that the coin is fair, is called the 'null hypothesis'. We write it as follows:

$$H_0 : p = \tfrac{1}{2}$$

This is the thing that does all the work; that is, we have to *assume* that it's true to do any calculations. The rival to this claim is the 'alternative hypothesis', which we write as follows:

$$H_1 : p \neq \tfrac{1}{2}$$

We don't ever do much with the alternative hypothesis directly. The reason should now be clear: if we only assume that $p \neq \frac{1}{2}$, we can't do any actual

calculations, as this requires a *specific* value for p. That is, if $p \neq \frac{1}{2}$ we have no idea how likely our data is, since this depends on the exact value of p; just knowing that it is not $\frac{1}{2}$ is too vague to be helpful.[1]

By taking the alternative as $H_1 : p \neq \frac{1}{2}$, we are also assuming that, prior to the test, we are open to the possibility that the coin is actually biased towards tails ($p < \frac{1}{2}$). Essentially, this would mean that, when we look at results 'as extreme or more extreme' than the result we got, i.e. getting 10 heads, we should indeed look at both $\mathbb{P}(X \geq 10)$ and $\mathbb{P}(X \leq 10)$, given the null. But if the alternative was specified in advance as only $H_1 : p > \frac{1}{2}$, it would be okay to only look at $\mathbb{P}(X \geq 10 \mid p = \frac{1}{2})$.

Summary

The procedure for testing a null hypothesis is always the same: we *assume* that the null hypothesis is true, do some kind of numerical *test*, and then see how likely the *result* we got would be *if the null were true*. If the answer is 'not very likely at all', then this casts doubt on whether the null is true after all. We then exclaim that we 'Reject the null hypothesis' – a phrase you hear a lot in statistics.

[1] Technically, the null can also be 'composite'; i.e. encompass a range of values of p. Then we would focus on the particular case that's most plausible given the data, and proceed as above. But this isn't really a case that ever comes up in econometrics.