Questions Pupils Ask:
Why do p values need to include more extreme possibilities?

By Colin Foster

In learning statistics, p values are notoriously difficult to understand (e.g. Greenland et al., 2016; Humphrey, 2012). For example, people often think that the p value is the probability that the null hypothesis is true, whereas in fact it is the probability, *if* the null hypothesis is true, of getting data at least as extreme as those obtained. The 'at least as extreme as' part of this definition is often confusing to students. Either they overlook this, or they are confused by it, leading to the student's question stated above.

When teaching hypothesis testing, it is not enough just to tell students the right way to do it. We also need to help them see why tempting alternatives that may seem obvious or simpler are *not* right (teaching the negatives as well as the positives). One example of a tempting but incorrect alternative when calculating a p value is not including the more extreme possibilities.

A simple scenario

Many scenarios can be used, but to keep things simple I will consider the issue of determining whether a coin is fair or not (see Note). While this may seem like a 'boring', familiar context, such contexts can be useful when trying to address something that is complicated to think about. The familiarity of the context minimises extraneous cognitive load (Sweller, Ayres, & Kalyuga, 2011), allowing students more headspace to focus on the concepts themselves.

Let's suppose that we throw our coin 10 times and obtain 8 heads. We let X be the number of heads obtained, and we suppose that the throws are independent, with a constant probability of 'success', p, on each trial, so $X \sim \text{Binomial}(10, p)$. We are taking all of this for granted, and the null hypothesis that we wish to test, that the coin is fair, is H_0: $p = 0.5$. We want to know how likely it would be to get data 'like ours' if this null hypothesis is true.

Students may naturally think that this means that they need to calculate

$$P(X = 8) = \binom{10}{8} (0.5)^{10} = 0.044,$$

correct to 3 decimal places, and compare this with the significance level, which we will take as 5%. Since $P(X = 8) < 0.05$, students may conclude that our data would be surprising if the null hypothesis were true, and so they reject the null hypothesis. This is illustrated in Figure 1, where the horizontal line shows a significance level of 0.05.

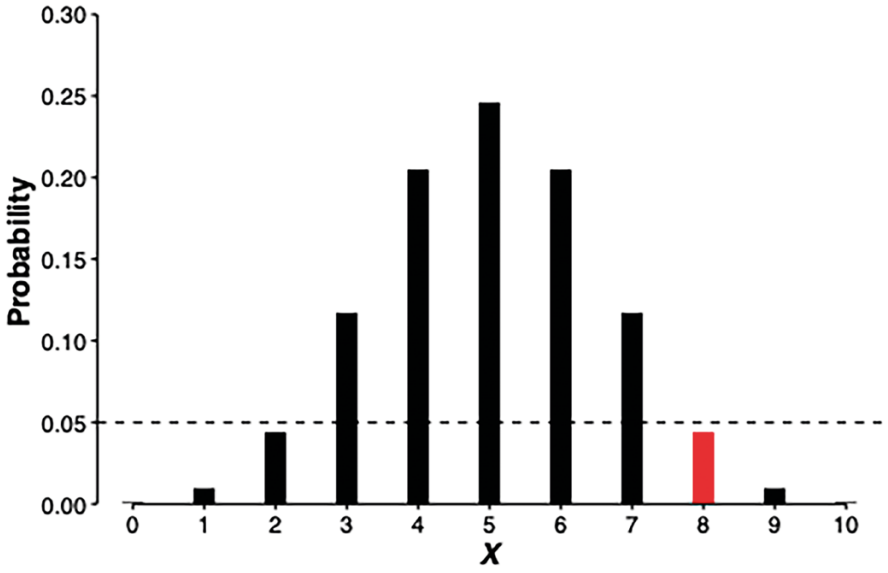

Figure 1. Because $P(X = 8) < 0.05$, a student incorrectly rejects the null hypothesis.

If students do not make this error, I would present this solution to them as though it were from a fictitious student, and ask them what is wrong with it. In my experience, students struggle to critique this kind of solution. They may say that their previous teacher had told them that they should work out $P(X \geq 8)$, rather than $P(X = 8)$, but I rarely find students who can give a convincing explanation of why.

Something that is likely to come up is that we have not yet considered whether a 1-tailed or 2-tailed test is appropriate, and one response students may give here is that $P(X = 2)$ should be included. Since this will be equal to $P(X = 8)$, because our null hypothesis distribution with $p = 0.5$ is symmetrical, including $P(X = 2)$ corresponds to doubling our probability, giving 0.088, which would no longer be significant at the 5% level. It would seem strange to be certain that *if* the coin is biased it can only be biased in favour of heads. Obtaining 8 tails would surely be equally indicative of a biased coin as obtaining 8 heads, so a 2-tailed test seems more appropriate, with the alternative hypothesis H_1: $p \neq 0.5$. And so now we would *not* reject the null hypothesis, because we are sharing out the 5% significance level between the two tails of the distribution. I think that students generally find it difficult to understand why we are doing this. They may say, "But we *didn't* obtain 8 tails, so why are we worrying about something that didn't happen?" How can what didn't happen affect our analysis?

Critical regions

One advantage of thinking in terms of critical regions is that the hard work is done before getting distracted by looking at the data. We want a process which will work *in the long run*, if we are repeatedly doing hypothesis tests many, many times. We want the long-run chance of making a Type 1 error to be less than or equal to our significance level of 5%. From a frequentist point of view, the probability of a Type 1 error only makes sense when imagining doing lots of hypothesis tests, where we wish to make a Type 1 error on average no more than once every 20 tests.

We have to decide, *before knowing that we might get 8 heads*, how many heads would lead us to rejecting the null hypothesis. And we want to decide this in such a way that we will falsely reject the null hypothesis no more than 5% of the time. Now, it is clear that if 8 heads is extreme enough to reject the null hypothesis then 9 or 10 heads will also be. We cannot say that 8 heads would lead us to reject the null hypothesis but that 9 heads wouldn't, because 9 heads is even further from the mean of 5 than 8 is. So, $X = 8$ cannot be in the critical region unless 9 and 10 are as well. This makes it clear why *tails* of the distribution should be the focus, rather than individual values of X (Figure 2).

But now, by including 0, 1, 2, 8, 9 and 10 in our rejection region, we are including far too much probability. If we reject the null hypothesis whenever $X \leq 2$ or $X \geq 8$, then, if we are wrong, and the null hypothesis is true, we will be rejecting it far too often. The sum of these six probabilities is 0.109, so our actual significance level would be around 11%, rather than 5%. The correct 5% critical regions are $X \leq 1$ or $X \geq 9$, and our data with $X = 8$ would *not* be extreme enough to lead to rejection of our null hypothesis.

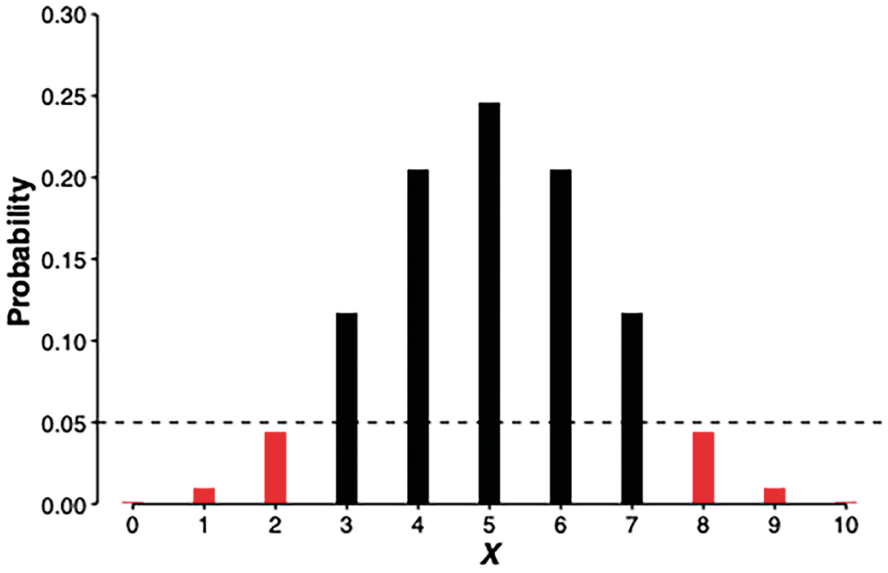

Figure 2. Rejecting the null hypothesis whenever $X \leq 2$ or $X \geq 8$ pushes the significance level up to 11%.

Increasing n

One way to convince students why probabilities for individual X values must be the wrong focus is to consider what happens as the sample size, n, increases. As we see in Figure 3, once we reach $n \geq 255$, $P(X = x) < 0.05$, for *all* values of x.

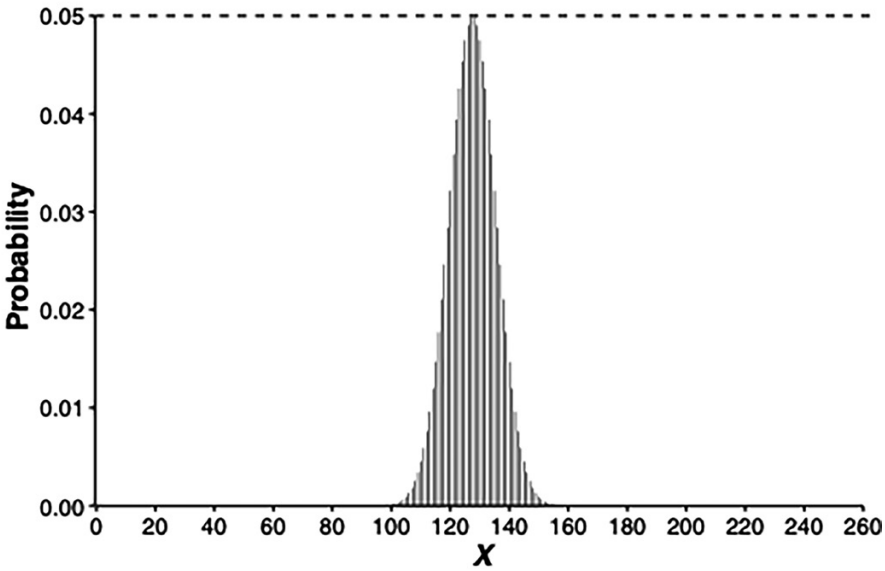

Figure 3. When $n = 255$, we see that $P(X = x) < 0.05$ for *all* values of x.

Using our wrong way of finding a p value, if we threw a coin, say, 256 times, then even if we obtained 128 heads and 128 tails, we would have to reject the null hypothesis that the coin is unbiased. Even though obtaining 128 heads is the *most likely* outcome, there are so many other possible outcomes that its probability does not quite reach the arbitrary 0.05 level. Even if we reduced our significance level below 5%, with enough trials (n large enough), we could always achieve this situation in which any particular arbitrary number of heads will be too unlikely to be 'believable' if the null hypothesis is true.

Thinking about critical regions, we find that for $n = 255$ we have $P(X \geq 144) = 0.022$, whereas $P(X \geq 143) = 0.030$, so the 2-tailed 5% critical regions will be $X \geq 144$ and $X \leq 111$, with a combined total probability of 0.045, after rounding, which is just less than 5%.

Continuous distributions

In many ways, continuous distributions, such as the Normal distribution, are much simpler conceptually than discrete distributions, such as the Binomial distribution. It is usual practice to teach the Binomial distribution first and then move onto the Normal distribution, but I have recently been experimenting with introducing the Normal distribution first. With $X \sim N(170, 6^2)$, for example, based on a population of people with mean height 170 cm and standard deviation 6 cm, it is very clear that the probability that X takes *any* particular single value is going to be zero. Students will accept that $P(X = 170) = 0$, for example, even though 170 cm is the modal height (Figure 4).

Seeing that the area under the probability density curve between two values is equal to the probability of obtaining any value in that interval is quite intuitive. And with continuous distributions we can find tail probabilities that *exactly* add up to the significance level (Figure 5), without having the awkwardness with discrete distributions of trying to find the maximum total that comes to less than 5%.

From a mathematical point of view, the Normal distribution is undoubtedly more complicated than a discrete distribution such as the Binomial distribution, because it involves probability *density*, rather than probability (Foster, 2025), and because the integration needed to find the area under the curve cannot be done analytically. However, from the students' point of view, I think that there are advantages to introducing hypothesis testing first with a continuous distribution, such as the Normal distribution, before moving on to consider the Binomial, Poisson and so on, and then seeing how these can be approximated by the Normal distribution under certain conditions.

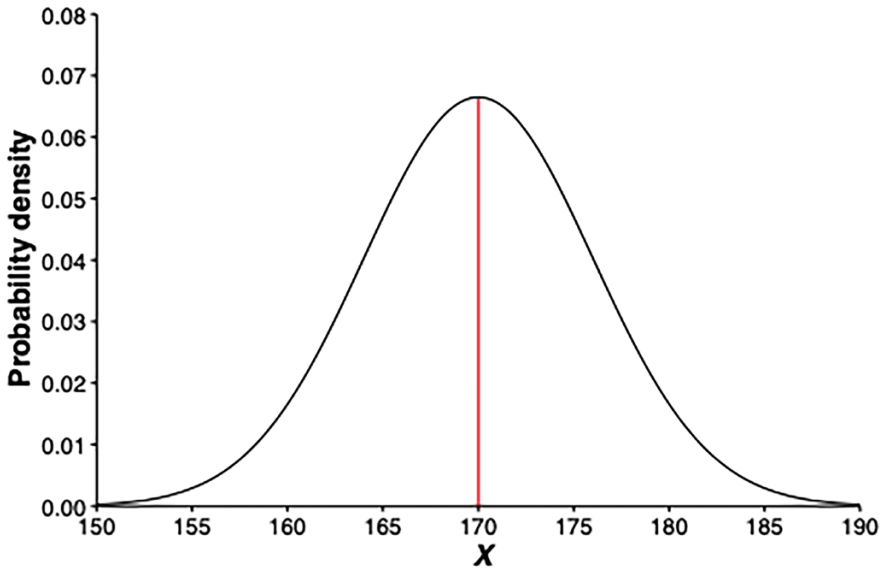

Figure 4. There is no probability associated with any single X value, even if X is the mode.

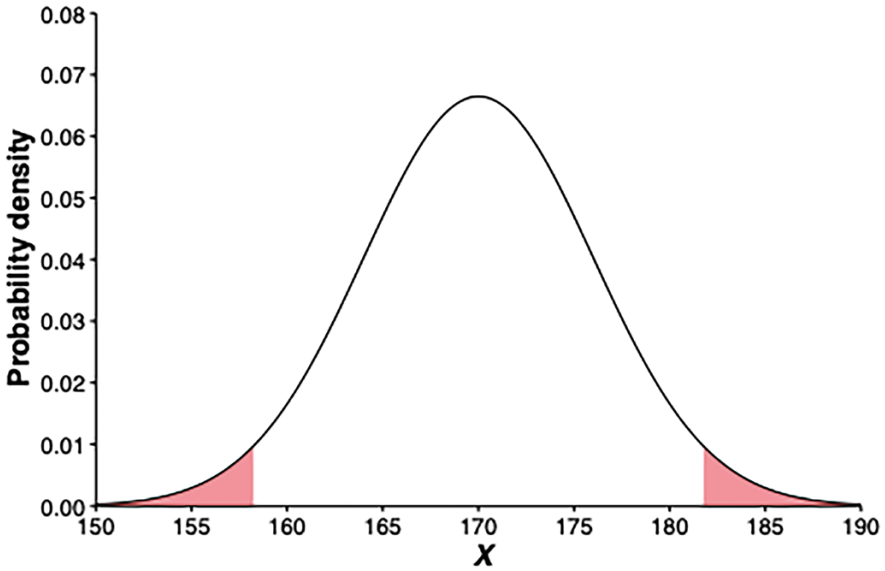

Figure 5. The total shaded area under this Normal curve is exactly 5%.

Note: Although, strictly speaking, there are no biased coins (Gelman & Nolan, 2002)!

References

Foster, C. 2025 'Questions pupils ask: How can probability density be greater than 1?', *Mathematics in School*. 54(5), pp. 2-3.

Gelman, A., & Nolan, D. 2002 'You can load a die, but you can't bias a coin', *The American Statistician*, 56 (4), pp. 308-311. https://sites.stat.columbia.edu/gelman/research/published/diceRev2.pdf

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. 2016 'Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations', *European Journal of Epidemiology*, *31*(4), pp. 337-350. https://doi.org/10.1007/s10654-016-0149-3

Humphrey, P. 2012 'What is a p-value?', *Teaching Statistics*, 34 (1), 18-20. https://doi.org/10.1111/j.1467-9639.2010.00446.x

Sweller, J., Ayres, P., & Kalyuga, S. 2011 *Cognitive load theory*. Springer. http://doi.org/10.1007/978-1-4419-8126-4

Keywords: Hypothesis testing, *p* values, Statistical significance

Author: Prof. Colin Foster, Department of Mathematics Education, Schofield Building, Loughborough University, Loughborough LE11 3TU.

Email: **c@foster77.co.uk**

Website: **www.foster77.co.uk**