A statistical trip to Tesco

Introduction

Picture this: it was my best friend Angela's birthday tomorrow, and I had still not gotten her a present. In my defence, sixth form is busy; I had Economics essays to write, Maths homework to submit, and a ludicrous number of Physics practical write-ups to complete. But none of that changed the fact that Angela, who had been my best friend since year 7, deserved better than a last-minute petrol station card.

Thankfully, I went to a school with a saving grace - during free periods, students were allowed to leave the site to go to the nearby town centre, with just a tap of their lanyards. The problem was that I had already left 3 times that week, and as I reached for my lanyard for the fourth time, I created this beautiful question in my head (the kind of question only someone doing too much statistics revision lately would come up with). *At what point do I become a mathematical anomaly?*

Making the model

Let's look at the school as a whole. My first order of business was to make a model, a type of mathematical framework to describe how students in my school leave during the day. There were 500 students that attended my sixth-form. Naturally, not everyone wanted to go out during their free periods; some would be doing their overdue homework, some would be playing card games. So, let's assume that the chance of any student leaving during their free period was 40%, or 0.4.

Could I use the binomial distribution here? Of course, this model rested on some assumptions. The four key conditions required for a probability distribution to be binomially distributed are that there is a fixed number of trials, each trial is independent, there are only two possible outcomes, and that the value of p stays constant. Of these, the first and third conditions are straightforward: I could say the first one was true as there were 500 students and each counted as a trial. I could also confirm the third one to be correct: a student would either leave school or not leave school during the day.

However, the second and fourth were trickier. In reality, whether one student left might have influenced another. For example, one person going out to Tesco might have triggered the rest of their friend group to come with them. For the sake of simplicity I assumed that each trial was independent. Regarding the value of p staying constant, did every student really have the same probability of leaving? A year

13 with lots of frees probably left school more often than a busy Year 12 with a packed timetable would. This means the p = 0.4 was really an average across everyone.

I could introduce the discrete random variable X, where

$$X = \textit{the number of students leaving the school during a free period.}$$

This gave

$$X \sim B(500,\ 0.4)$$

meaning, for any given free period, the average number of students leaving the school would be

$$np = 500 \times 0.4 = 200$$

To calculate the probability that (for example) 210 students leave the school during any given free period, I used the binomial probability mass function:

$$P(X = r) = \binom{n}{r} p^r (1 - p)^{n-r}$$

When the correct values are substituted in, the equation gives us:

$$P(X = 210) = \binom{500}{210} 0.4^{210} (1 - 0.4)^{500-210}$$
$$= 0.02387.. = 2.39\% \ (3\text{ sf})$$

Which told me that the probability 210 students leave the school during any given free period is 2.39%. But what a faff that was..! Typing this into my calculator took way too long, time I could have spent doing that overdue Maths homework that I forgot about (this is true). I wanted to simplify this even more.

Poisson

Watch what happens when I let n grow large while p shrinks small (keeping np = 200 the same). Let $np = \lambda$, therefore $p = \frac{\lambda}{n}$. Rewriting the binomial probability mass function using the definition of the binomial coefficient and substituting p for $\frac{\lambda}{n}$:

$$P(X = r) = \frac{n!}{r!(n-r)!} \left(\frac{\lambda}{n}\right)^r \left(1 - \frac{\lambda}{n}\right)^{n-r}$$

$$= P(X = r) = \frac{n!}{r!(n-r)!} \times \frac{1}{n^r} \times \lambda^r \left(1 - \frac{\lambda}{n}\right)^n \times \left(1 - \frac{\lambda}{n}\right)^{-r}$$

$$= P(X = r) = \frac{n!}{r!(n-r)!} \times \frac{1}{n^r} \times \lambda^r \left(1 - \frac{\lambda}{n}\right)^n \times \left(1 - \frac{\lambda}{n}\right)^{-r}$$

As n → ∞, the red box approaches $\frac{1}{r!}$, and the blue box approaches 1. However the pink box:

$$\lim_{n \to \infty} (1 - \frac{\lambda}{n})^n = e^{-\lambda}$$

Which makes the whole equation:

$$P(X = r) = \frac{\lambda^r e^{-\lambda}}{r!}$$

The Poisson distribution!

To test whether the Poisson actually does approximate the Binomial distribution, I let

$$X = \textit{the number of students leaving the school during a free period under the Poisson approximation}$$

$$X \sim Po(200)$$

$$P(X = 210) = \frac{200^{210} e^{-200}}{210!}$$
$$= 0.02151.. = 2.15\%$$

Very close to the binomial answer!
But let's not get side tracked - we have to focus on Angela's present. If 200 out of 500 students leave per day on average, then the probability of me specifically being one of them on any given day is:

$$\frac{200}{500} = 0.4$$

Meaning I leave on 40% of school days on average. Since there are 5 days, and I would leave each day independently with a probability of 0.4, my weekly exits would follow:

$$Y \sim B(5, 0.4)$$

where

$$Y = \textit{number of times I leave school in a week}$$

Note that Poisson isn't used here because n=5 is very small, and p=0.4 is relatively large. The Poisson approximation would work best when n is large and p is small which isn't really the case.

I could now calculate whether or not me leaving would be anomalous, and to do this properly, I decided to run a hypothesis test.

Hypothesis testing

The question was simple: was I actually leaving more than the typical student, or would four exits in that week be nothing more than normal variation? Recall that we established p = 0.4 as the probability of any student leaving on a given day, so under the null hypothesis this would apply to me too. My null hypothesis was therefore:

$$H_0: p = 0.4$$

My alternate hypothesis stated that I left school **more** than the average student:

$$H_1: p > 0.4$$

I used a 5% significance level for this.

$$P(Y \geq 4) = P(Y = 4) + P(Y = 5)$$

$$= (\binom{5}{4} 0.4^4 \times 0.6^1) + (\binom{5}{5} 0.4^5 \times 0.6^0)$$

$$= 0.0768 + 0.01024$$
$$= 0.08704 = 8.70\% \text{ (3sf)}$$

Since 8.70% > 5%, I could conclude that I thankfully was not an anomaly for leaving school four times that week!

Of course, a truly rigorous statistician would have verified this model using a chi squared goodness of fit test, but with Angela's birthday the next day I decided this was a problem for another day.

Conclusion

Armed with the knowledge that I was no statistical anomaly for leaving school so many times, I tapped my lanyard and went out to the shops to buy Angela a teddy bear. It wasn't much (but then again neither was the 8.7% probability). I've done more

than justify a shopping trip though, because even just leaving school for thirty minutes contains, if you look closely enough, a remarkable amount of mathematics.