I love the underground map; it's brilliantly complex, and full of interchanges and stations. It got me thinking that there are over 272 stations, and some are more used than others, what are the probabilities of getting off at certain stations, and whether this can be modelled by certain statistical distributions, particularly the Binomial.

Let's begin with Waterloo, the most used Station for 2024 according to TFL data. Waterloo lies on 4 lines, but for now, I will focus on the Jubilee. It is between Westminster and Southwark. The probability of getting to a station and getting off depends on first, which direction one travels on the line. Unless beginning at a terminus, this probability is a half (½), else it is 1, as the Jubilee line is a straight line with no branches.

Suppose X to be the station which one gets off at. The event of getting off at each station is each an independent event, with only two outcomes (stay on/get off) per trial, and a set number of trials, because the Jubilee line only has a set number of stations depending on which station and direction you start at. The probability of success (getting off) can be assumed to be a ½, a fixed uniform distribution, assuming it is influenced by no other factors.

If we use Waterloo as an example, from this, X can be distributed by the Binominal Distribution (9,0.5) if travelling toward Stratford, or X~B (17,0.5) if travelling toward Stanmore.

However, we can then test this distribution with a Chi Squared goodness of fit test, using a 5% significant level, and taking the observed frequencies from the TfL Annual Counts (Entry/Exit) for rail and Tube 2024 data, presuming that all counts for a station are all for the Jubilee Line if there are no interchanges in the station, else the count is divided by the number of lines it serves, and using the distribution for travelling towards Stratford. The observed data is for a week

H0: the data can be modelled by a Binomial Distribution

H1: the data cannot be modelled by a Binomial Distribution

X~B(9,0.5)

But, as there are around 276 million journeys annually, and I have used the weekly observed frequencies, all probabilities calculated using X must then be multiplied by 5 million, as 276,000,000/52 = 5,307,7692

Station number away from Waterloo	Observed Passengers (weekly)	Expected
0 - Waterloo	1,589,208	P(X=0) x 5 x 10^6 = 9750
1 - Southwark	186,862	P(X=1) x 5 x 10^6

		=87500
2 - London Bridge	1,188,796	P(X=2) x 5 x 10^6 =351500
3 - Bermondsey	150,254	P(X=3) x 5 x 10^6 =820000
4 - Canada Water	208,374	P(X=4) x 5 x 10^6 =1230000
5 - Canary Wharf	681,399	P(X=5) x 5 x 10^6 =1230000
6 - North Greenwich	487,281	P(X=6) x 5 x 10^6 =820000
7 - Canning Town	259,119	P(X=7) x 5 x 10^6 =351500
8 - West Ham	112,663	P(X=8) x 5 x 10^6 =87500
9 - Stratford	1,144,201	P(X>=9) x 5 x 10^6 =9750

X^2 = 391,780,000

$$\frac{(1589208-9750)^2}{9750}+\frac{(186862-87500)^2}{87500}+\frac{(1188796-351500)^2}{351500}+\frac{(150254-820000)^2}{820000}+\frac{(208374-1230000)^2}{1230000}$$

$$+\frac{(681399-1230000)^2}{1230000}+\frac{(487{,}281-820000)^2}{820000}+\frac{(259{,}119-351500)^2}{351500}+\frac{(112{,}663-87500)^2}{87500}+\frac{(1144201-9750)^2}{9750}$$

Degrees of Freedom = 10 – 1 = 9

χ^2(9)(0.05) = 16.918

X^2 > χ^2

391,780,000 > 16.918

The result is significant, so there is sufficient evidence to reject the null hypothesis, and so clearly X (the event of the station which one gets off at) cannot be distributed by the Binomial distribution. Also, there are many limitations with this model, due to the assumption of independence as other trials could affect each other, for example, there could be an accident at Canary Wharf, and so more people happen to get off at Canada Water instead, affecting each other's trials. Also, the fixed outcome of success, the probability, does not change and is highly likely to, as most of the observed frequencies deviated hugely from their expected frequencies. In a real-life scenario, it is also highly unlikely that the probability of people getting off at different stations is the same.

So, can we use alternative distributions? For example, the multinomial distribution (multiple stations) or a negative binomial distribution (if modeling the number of trials until a specific destination is reached) could be better alternatives.

Instead of the Binomial, the Multinomial distribution seems more likely. X has a fixed number of stations, and constant probabilities that sum to 1, as the model assigns a fixed probability to each station where you can get off, where the sum of probabilities equals 1 and they are all non-negative. Over a specific time (weekday morning peak), the probability of a passenger getting off at a station can be assumed to be roughly constant. Again, each trial is independent, as the event of getting off at one station is independent of getting off at another station. There are also mutually exclusive categories, as there are a limited number of exit stations (9) and are finite, discrete categories. Therefore, X can be distributed by the Multinomial Distribution, where $n = 10$, $k = 9$

In probability theory, the multinomial distribution is a generalisation of the binomial distribution. This helps as some of the parameters remain the same, such as number of trials, and independent events, but there are differences that make it a better model, such as the different probabilities for different events. Unlike the binomial where every probability for each station was the same, now it is more realistic by having each station have its own individual probability. The Binomial also becomes less accurate for higher numbers of n, and so the Multinomial is better for large sample sizes, especially die to the high usage of Jubilee line, and so makes the multinomial a useful prediction tool and distribution.

However, the Multinomial also has its own limitations. Again, like the Binomial, it assumes independence of passenger choices, but these are likely to be influenced by other external factors, such as if a station is busy or crowded, thus impacting other stations' events and not being fully independent. The constant probabilities don't change between times of the day, from morning peak, to the afternoon and weekends, and so do not accurately represent the usage of the Tube and probabilities of getting off at a certain time. Thus, the multinomial is best used for short-term modelling, instead of long-term. Furthermore, complex routes, such as free walking interchanges (Monument and Bank) create non-linear, complex destination choices that require more advanced nested logic or behavioral models, rather than a simple multinomial set

To conclude, this is only from one line out of 11, and is one of the simplest, with no branches and only two termini at end to end. I also only tested one direction of travel. This becomes more complex when other external factors are added, and so more complex models are required to factor these in, although looking at this section of the tube has been interesting and revealing.