Fitting a distribution (Discrete)

Collect the dataset and try to approximate a variable with the discrete random variable.

OR

Model a dataset with the discrete random variable and verify it.

Data Set

I have selected the eCommerce customer behaviour dataset.
Source:
https://www.kaggle.com/datasets/uom190346a/e-commerce-customer-behavior-dataset

Basic Analysis:

https://docs.google.com/spreadsheets/d/18_5NHMHLO2W-c66L8k01wJkC5x6xD9cY/edit?gid=1363288999#gid=1363288999

Selection of a Random Variable

I have selected the numerical variable "**Items Purchased**" as Discrete Random Variable. It represents the number of items purchased in a single transaction, and the variability in this number across different customers or transactions makes it suitable for modelling using discrete distributions like the Binomial distribution and Poisson Distribution.

Justification of randomness:

The number of items purchased by a customer in a single transaction is not deterministic—it can vary due to multiple factors such as customer preference, purchasing power, time of the year, promotions, etc. Because of the variability and uncertainty in how many items a customer might purchase in any given transaction, the number of items purchased is justified as a random variable.

The information of Item Purchased is given below in the form of a Pivot Table:

Number of Items Purchased	COUNT of Items Purchased
7	25
8	33
9	34
10	47
11	28
12	33
13	32
14	11
15	24
16	24
17	1
18	9

19	6
20	19
21	24
Grand Total	**350**

Let X be the random variable that denotes the number of items purchased in a transaction.
P(X = x) = $\frac{Count\ of\ x}{Total\ Count}$

P(X=7) = 25/350 = 0.071
P(X=8) = 33/350 = 0.094
P(X=9) = 34/350 = 0.097
P(X=10) = 47/350 = 0.134
P(X=11) = 28/350 = 0.080
P(X=12) = 33/350 = 0.094
P(X=13) = 32/350 = 0.091
P(X=14) = 11/350 = 0.031
P(X=15) = 24/350 = 0.069
P(X=16) = 24/350 = 0.069
P(X=17) = 1/350 = 0.003
P(X=18) = 9/350 = 0.026
P(X=19) = 6/350 = 0.017
P(X=20) = 19/350 = 0.054
P(X=21) = 24/350 = 0.069

P(7≤ x ≤ 21) = 1.00

Now, we can check whether this follows the Binomial distribution.
To check whether present data follows a binomial distribution, we need to compare the observed probabilities with the expected probabilities under the binomial model.

The parameters (n, p) of binomial distribution:

- n represents the total number of trials. In the present case, it is the maximum number of items that can be purchased in a transaction. In the present case n is 21 (the highest observed number of items purchased).
- p can be estimated as the average number of items purchased divided by n.

Now the average (mean) number of items purchased =

∑(Items Purchased×Frequency)/Total Number of Observations

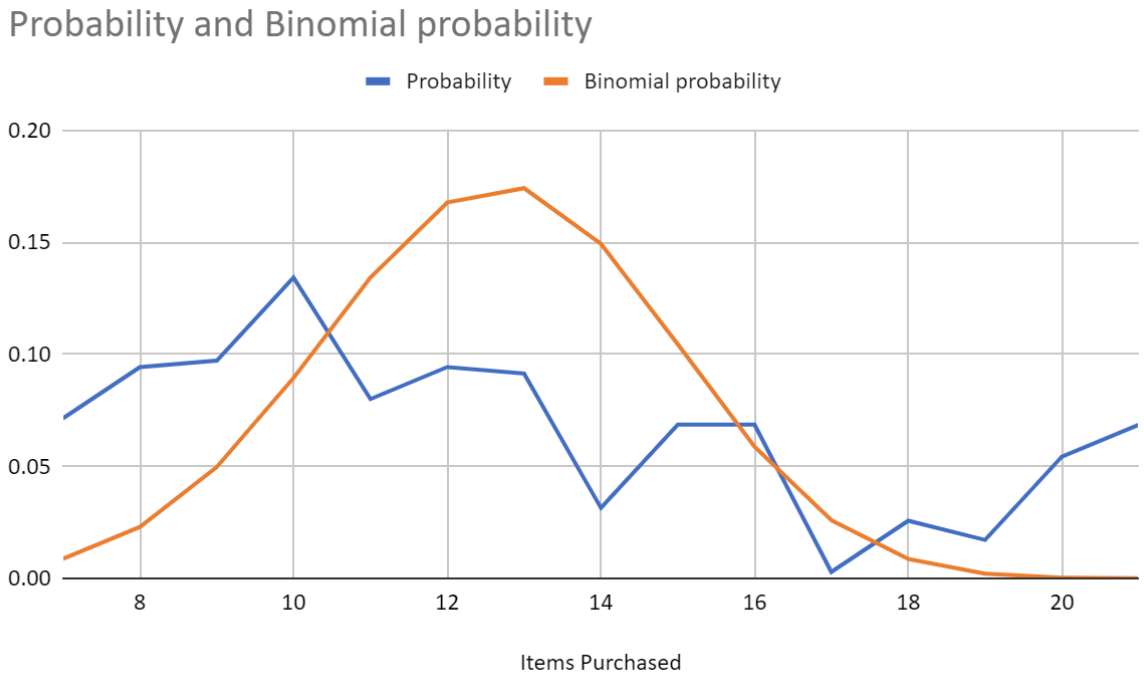
= 12.6

Therefore, p = 12.6/21 = 0.6
X ~ $Bin(21, 0.6)$
The formula in Google Sheet is BINOM.DIST(x, 21, 0.6, false) and the corresponding probabilities are given below:

P(X=7) = 0.00874
P(X=8) = 0.02294
P(X=9) = 0.0497
P(X=10) = 0.08945
P(X=11) = 0.13418
P(X=12) = 0.16773
P(X=13) = 0.17418
P(X=14) = 0.14929
P(X=15) = 0.10451
P(X=16) = 0.05878
P(X=17) = 0.02593
P(X=18) = 0.00864
P(X=19) = 0.00205
P(X=20) = 0.00031
P(X=21) = 0.00002

$P(7 \leq x \leq 21) = 0.99645$
For $x \leq 6$, $P(X=x) = 1 - 0.99645 = 0.00355$.

Given that the sum of the pmf of binomial distribution is nearly 1 and covers almost the entire probability space, it suggests a comprehensive fit to the data. The small discrepancy in the sum indicates that it likely represents the data well, even if it isn't perfect. There is also a notable difference between the observed and binomial probabilities.



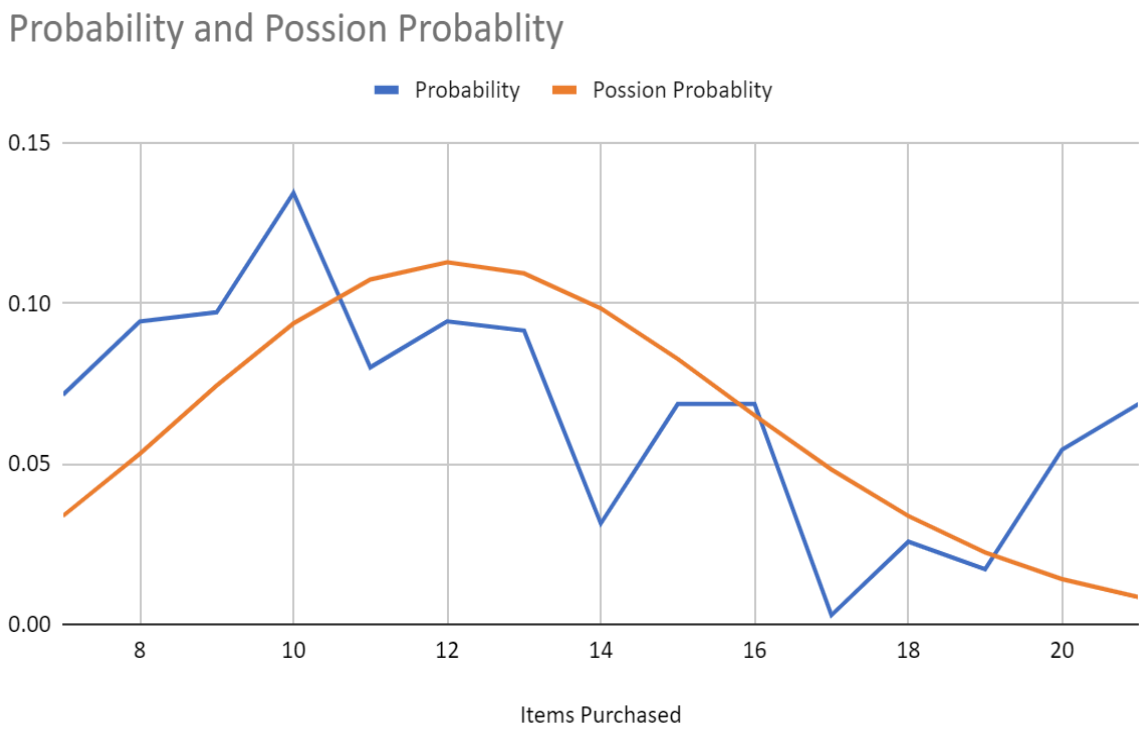
Fitting the data to other distributions like the **Poisson distribution**,

In case of poisson distribution the parameter is λ.
λ is the average rate (mean number of events, or in the present case, the mean number of items purchased) = 12.6
For each observed value of items purchased (x) we can calculate the Poisson probability using the mean λ =12.6.
The formula in Google Sheet would be is POISSON.DIST(x, 12.6, FALSE) and and the corresponding probabilities are given below:

P(X=7)	=	0.0337
P(X=8)	=	0.0531
P(X=9)	=	0.0744
P(X=10)	=	0.0937
P(X=11)	=	0.1074
P(X=12)	=	0.1127
P(X=13)	=	0.1093
P(X=14)	=	0.0983
P(X=15)	=	0.0826
P(X=16)	=	0.0650
P(X=17)	=	0.0482
P(X=18)	=	0.0337
P(X=19)	=	0.0224
P(X=20)	=	0.0141
P(X=21)	=	0.0085

$P(7 \leq x \leq 21) = 0.9571$
For $x<7$ and $x>21$, $P(X=x) = 1 - 0.9571 = 0.0429$..

While the Poisson distribution showed a reasonable fit for certain probabilities, the significant shortfall in the sum of probabilities (0.9571) indicates that it might not be fully appropriate for modelling of the present data across the entire range.



Conclusion:
Based on the analysis, the binomial distribution is likely the better choice for modelling in the present data set. The sum of probabilities being close to 1 suggests that it captures most of the distribution of "Items Purchased" effectively.