Assignment 1

CMSC 691 — Intoduction to Data Science

Faisal Rasheed Khan

VB02734

vb02734@umbc.edu

Question 1---

1.

For this question, we will be using binomial distribution. We are using binomial distribution because the problem description satisfies the binomial distribution. For a binomial distribution, there should be fixed trials which should have two possible outcomes, here the two possible outcomes are: purchased and not purchased and the trails are independent. Below is the histogram representation of the data.

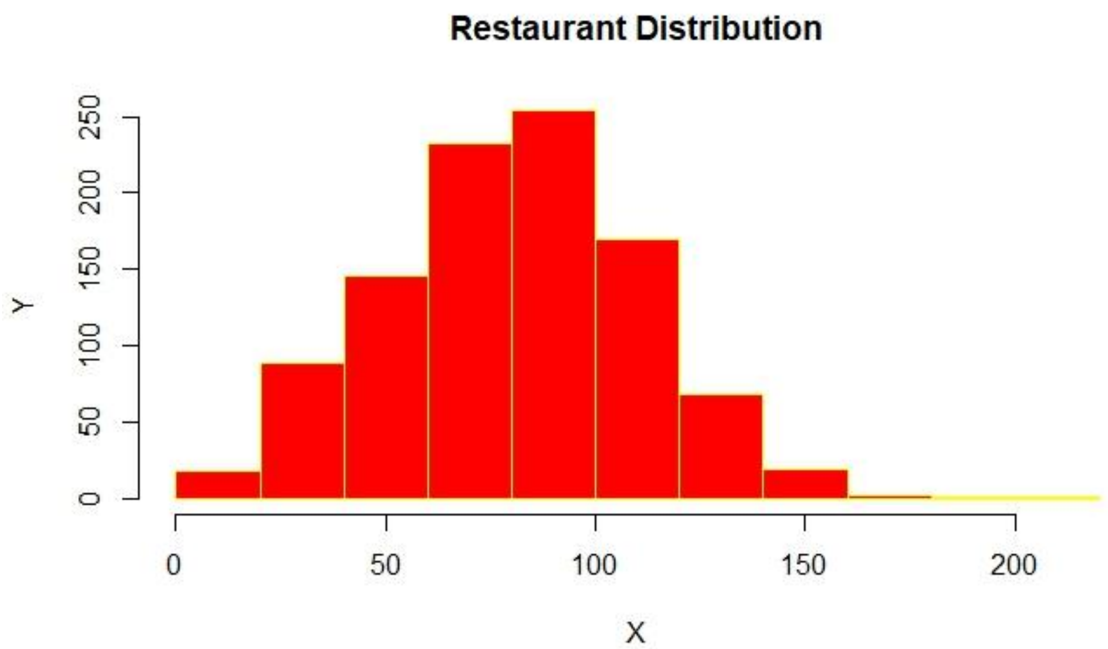
i. The mean of the data is calculated which is 10.12687
The variance is also calculated which is 9.329649
This is probability distribution and is defined as mean = N * Probability
The probability is 0.01012687
the probability that among 1000 customers there will be between 6 and 8 purchases in a day is calculated using dbinom() and is 0.2560151
dbinom() is used to get the probability of exact success(similarly for dpois())

ii. the probability that among 1000 customers there will be exactly 7 purchases in a day is also calculated using dbinom() as its exact probability to calculate and is 0.0865437
iii. the probability that among 1000 customers there will be at most 5 purchases in a day is calculated using pbinom() and is 0.0615141. Here we have used pbinom() because it is used to calculate the cumulative probability of success.(similarly for ppois())
iv. Yes, We can use Poisson's distribution here. Because this data follows the properties of poissons distribution. Poissons distribution is used when the occurrence of event is in a fixed interval of time, if we know the rate of occurrence of that event. It is used when n-> ∞ and p->0, such that n*p is finite. Poissons distribution is used when mean is almost equal to variance. Poissons distribution is used with rare events over fixed interval.
v. In this problem mean is very near to the variance, so we can use poissons distribution here. The mean = 10.12687 and The variance = 9.329649
The probability of poissons distribution is very similar to that of binomial distribution.
the probability that among 1000 customers there will be between 6 and 8 purchases in a day is calculated using dpois() and is 0.2562815
the probability that among 1000 customers there will be exactly 7 purchases in a day is also calculated using dpois() as its exact probability to calculate and is 0.08666683
the probability that among 1000 customers there will be at most 5 purchases in a day is calculated using ppois() and is 0.06243589

R File: IDS_Assign_1.R

2.

The normal distribution is chosen for the restaurant's data because it matches key properties of the data. When visualizing the data via histogram we can observe the bell-shaped plot, which will have the properties of Normal distribution. Moreover, it is of like continuous distribution where the data involves the number of customers visiting, it makes suitable for a continuous distribution like the normal distribution which is bell-shaped.

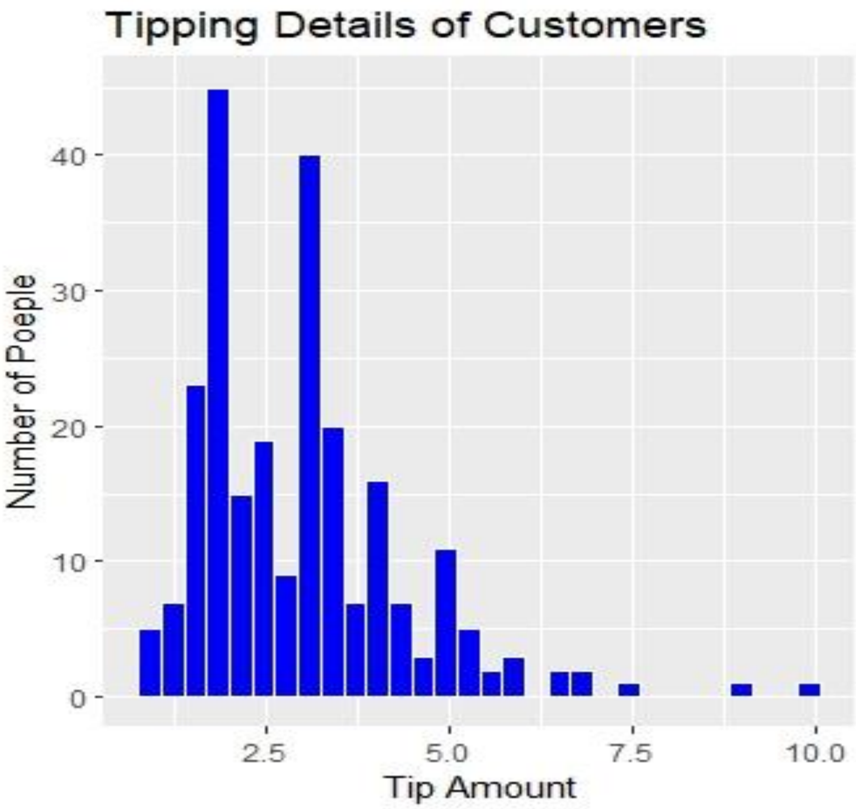


We calculate the mean for the number of customers and is 81.059
The variance is 921.7833 and the standard deviation is 30.36088

The probability distribution is 0.081059
Need to calculate food the restaurant should prepare so that the restaurant won't run out of food 85% of the days, above we have calculated and have the details required for the probability of Normal distribution, and we calculate using qnorm() in R.
The answer is 112.526. 85% of the food won't run out can be consumed around 113 customers

R File: IDS_Assign_2.R

3.



R File: IDS_Assign_3.R

4. The task is to find the point of estimate for both x and y using sample sizes 20 and 75 for five times. I am considering the very first 5 sample sizes from the start. We will be using confidence level of 99%. The standard value for the confidence level 99% is 2.576.
Considering first 20 sample sizes to see how the range of the point estimates are.
The first 20 samples mean of x is 47.24899, mean of y is 9.75, standard deviation of x is 33.32613, standard deviation of y is 2.197487, standard error of x is 7.451949217 and standard error of y is 0.4913730312
Standard error is calculated as standard deviation upon square root of Sample size.
Standard error = standard deviation/ sqrt(Sample Size)
Range of point of estimation is calculated as mean of the sample size plus/minus product of confidence value and standard error of the sample size.
Range of point of estimation for confidence 99% = mean $\pm$ 2.576 * standard error
All the calculations are done in the IDS_Assign_4.R script file
Range for the first 20 samples of x are 28.05276883 to 66.4452117

Range for the first 20 samples of y are 8.484223072 to 11.01577693

Considering first 75 sample sizes to see how the range of the point estimates are.
The first 75 samples
mean of x is 53.04782, mean of y is 10.17333, standard deviation of x is 29.6655, standard deviation of y is 3.2936, standard error of x is 3.425477 and standard error of y is 0.3803144
Standard error is calculated as standard deviation upon square root of Sample size.
Standard error = standard deviation/ sqrt(Sample Size)
Range of point of estimation is calculated as mean of the sample size plus/minus product of confidence value and standard error of the sample size.
Range of point of estimation for confidence 99% = mean $\pm$ 2.576 * standard error
All the calculations are done in the IDS_Assign_4.R script file
Range for the first 75 samples of x are 44.22379 to 61.87185
Range for the first 75 samples of y are 9.193643 to 11.15302

Similarly calculate the ranges for 5 times which are shown in the R script.

```
> print(table_20_x)
  Sample_Size    Mean_x StandardDeviation_x StandardError_x Lower_Range_x Upper_Range_x
1          20 47.24899            33.32613        7.451949      28.05277      66.44521
2          20 53.05808            25.28874        5.654735      38.49148      67.62468
3          20 55.68691            26.82801        5.998924      40.23368      71.14014
4          20 56.96015            33.86909        7.573360      37.45118      76.46913
5          20 57.20367            29.35635        6.564279      40.29408      74.11325
> table_20_y <- data.frame(Sample_Size = 20
```

```
> print(table_20_y)
  Sample_Size Mean_y StandardDeviation_y StandardError_y Lower_Range_y Upper_Range_y
1          20   9.75            2.197487       0.4913729      8.484223      11.01578
2          20  10.95            4.186130       0.9360471      8.538743      13.36126
3          20  10.60            3.152276       0.7048703      8.784254      12.41575
4          20   9.50            3.069373       0.6863327      7.732007      11.26799
5          20  11.10            2.881885       0.6444092      9.440002      12.76000
>
```

```
+ )
> print(table_75_x)
  Sample_Size    Mean_x StandardDeviation_x StandardError_x Lower_Range_x Upper_Range_x
1          75 53.04782            29.66550        3.425477      44.22379      61.87185
2          75 49.09861            31.55560        3.643727      39.71237      58.48485
3          75 49.41138            30.29525        3.498194      40.40003      58.42273
4          75 51.67757            26.49163        3.058990      43.79761      59.55752
5          75 53.41781            27.52430        3.178233      45.23068      61.60494
> table_75_y <- data.frame(Sample_Size = 75
> print(table_75_y)
  Sample_Size    Mean_y StandardDeviation_y StandardError_y Lower_Range_y Upper_Range_y
1          75 10.173333            3.293619       0.3803144      9.193643      11.15302
2          75 10.560000            2.461817       0.2842661      9.827731      11.29227
3          75  9.453333            3.063708       0.3537665      8.542031      10.36464
4          75  9.906667            2.547778       0.2941920      9.148828      10.66451
5          75  9.733333            3.260465       0.3764860      8.763505      10.70316
>
```

Summary and Findings:
We find point of estimation range so that the maximum data lines within that range of the confidence level and the mean lines within the range. It's a way to express the precision of the estimate.
The value of confidence levels are the standard values, for 95% confidence value = 1.96, for 99% confidence value = 2.576
Most of the data distribution falls under this range between lower to upper

Confidence intervals provide a range of values within which we believe the true population (like the mean) is likely to fall with a specified level of confidence.
Not only point estimate provides importance but it also depends on confidence levels provides good understanding of uncertainty.
increasing coincidence level reflects higher degree of confidence level with less precision

R File: IDS_Assign_4.R

References:

https://blackboard.umbc.edu/bbcswebdav/pid-6351621-dt-content-rid-70354645_1/xid-70354645_1

https://blackboard.umbc.edu/bbcswebdav/pid-6362730-dt-content-rid-70461367_1/xid-70461367_1

https://blackboard.umbc.edu/bbcswebdav/pid-6371899-dt-content-rid-70683694_1/xid-70683694_1

https://blackboard.umbc.edu/bbcswebdav/pid-6371900-dt-content-rid-70683695_1/xid-70683695_1

Home - RDocumentation

BINOMIAL distribution in R [dbinom, pbinom, qbinom and rbinom functions] (r-coder.com)

https://blackboard.umbc.edu/bbcswebdav/pid-6367033-dt-content-rid-70554826_1/xid-70554826_1

https://blackboard.umbc.edu/bbcswebdav/pid-6367033-dt-content-rid-70554827_1/xid-70554827_1

https://blackboard.umbc.edu/bbcswebdav/pid-6367033-dt-content-rid-70554828_1/xid-70554828_1

https://blackboard.umbc.edu/bbcswebdav/pid-6367033-dt-content-rid-70554829_1/xid-70554829_1

https://blackboard.umbc.edu/bbcswebdav/pid-6367033-dt-content-rid-70554830_1/xid-70554830_1