PROBABILITY THEORY AND INTRODUCTORY STATISTICS

ALY6010, FALL 2019

WEEK 3 PROJECT ASSIGNMENT

Electronic Keno

SUBMITTED BY: TWINKLE INAMDAR, SHIVANI ADSAR

NUID: 001050277, 001399374

CRN: 71447

SUBMITTED TO: DR. DEE CHILUIZA REYES

DATE: 10/6/2019

Introduction

The assignment aims at performing probabilities on discrete and continuous probability distributions. In the Keno game, the machine picks up 20 numbers randomly out of a set of 80 numbers and then a player selects 20 numbers. The winning or losing depends on the matches of numbers. We are using the hyper-geometric probability distribution to calculate the mean, variance, standard deviation and simulations to plot the line graph. Using this analysis, we can predict the chances of a player winning or losing the game(1).

Analysis

Part 1

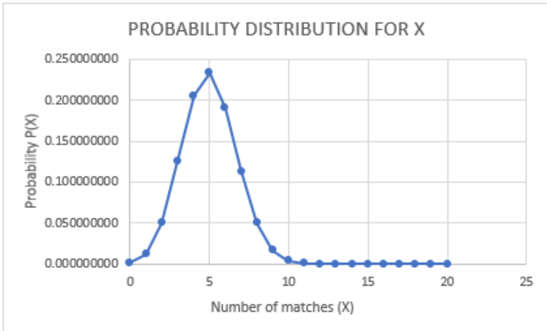
The Keno game to predict the winners using Hyper-geometric probability distribution

The game involves selection of 20 numbers by a player out of set of 1 through 80 numbers. Then computer randomly draws another 20 numbers from the set. The matching numbers amongst the selections of player and the computer decides the winner for the game.

Using the HYPGEOM.DIST function, we are able to display the probability of successes in 20 draws without replacement from a population of 80 numbers.

- **Graph for Probability Distribution**

  The line graph represents the probability distribution and the number of matches. It can be interpreted from the graph that, the highest peak value is 0.23. This indicates that the chances of winning the game by the player in 5 matches.

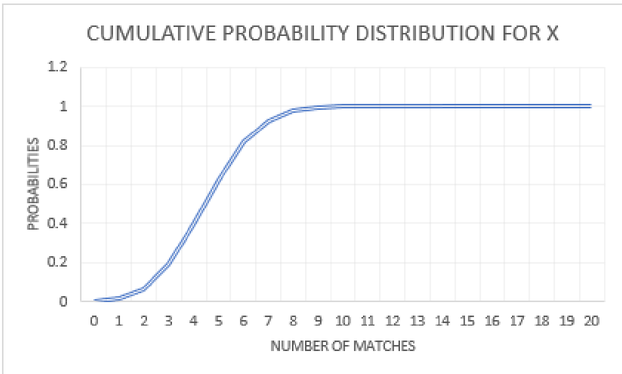


Graph 1: Probability Distribution using Line plot

- **Cumulative Probability Distribution**

  The cumulative probability distribution of a random variable is a distribution function which takes a value less than or equal to x where $P(X \leq x)$.

  It can be interpreted from the line plot that there is a rise from probability 0 to 1 for the first 8 matches, and then the probabilities are constant for the remaining matches.



Graph 2: Cumulative Probability Distribution

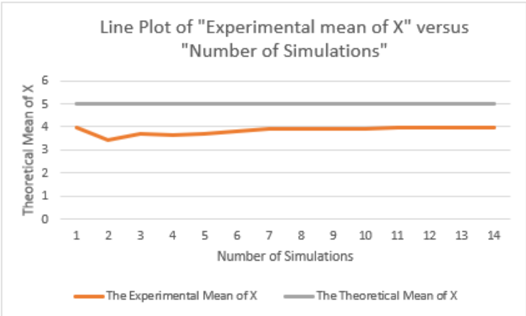
- **Theoretical and Simulated values of X**
  In order to generate simulations for the random values, we have used the RAND() function. The simulated X values have been calculated, as per the VLOOKUP formula which will match the results based on the probabilities. The IFERROR formula is used to eliminate the null values. According to the table, it can be interpreted that the Theoretical value of X is 5 which is greater than the Simulated X value i.e. 4.02. Also, the Variance of X for Theoretical X is 2.84 which is greater than Simulated X value which is 2.66. The Standard Deviation of Theoretical X i.e. 1.68 is more than Simulated X which is 1.63.

	Theoretical X	Simulated X
Expected Value of X	5	4.027
Variance of X	2.848101266	2.66893994
SD of X	1.687631851	1.633689059

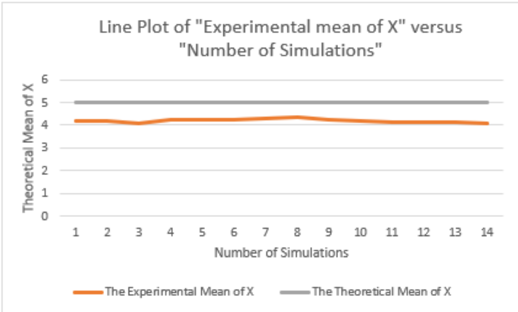
Table 1: Comparison for Theoretical X and Simulated X

- **Line Plot for "Experimental means of X" versus "Number of Simulations"**
  Using the line plot, it can be observed that the Theoretical Mean of X is a straight horizontal line and Experimental Mean of X is a varying based on the random simulations generated.

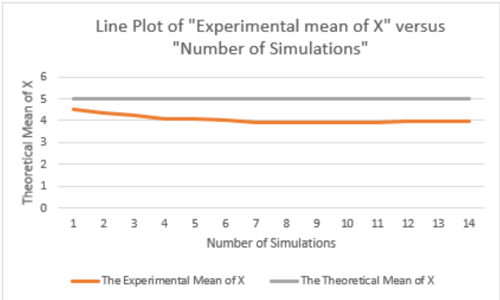
Graphs for comparing the Experimental Mean and Theoretical Mean



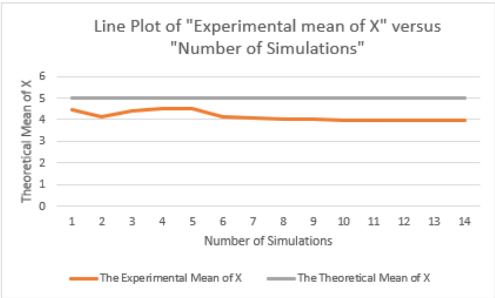
Graph 3.1



Graph 3.2



Graph 3.3



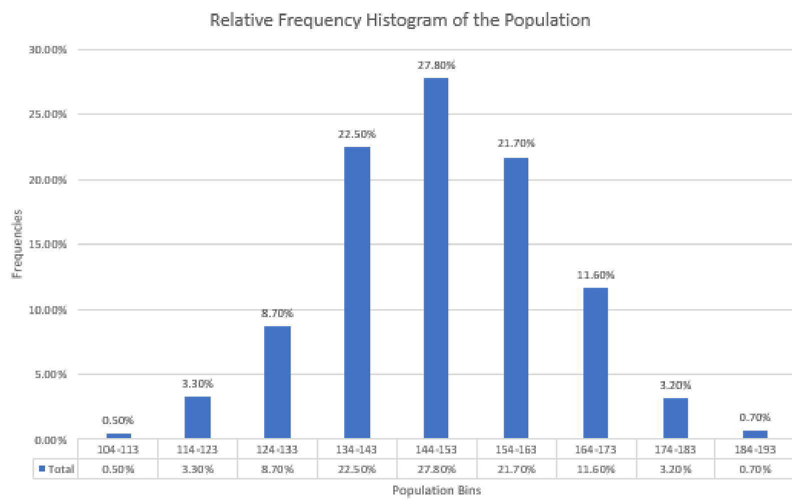
Graph 3.4

According to the Law of Large Numbers, it can be interpreted from the above graphs that, with the increase in the number of trials, the Experimental values of means are approaching the Theoretical mean of X. It can be observed that, the experimental value of mean in Graph 1 was 4, in Graph 2, the experimental value moves closer to theoretical mean with a value of 4.2. Moreover, In Graph 3, the experimental value is 4.5 and in Graph 4, the value reaches 4.7 which is very close to the Theoretical mean value of 5.

Part 2

Random Sampling from a normal population

- On the basis of the normal population given in the table, the relative frequency histogram has been created by generating bins for the population and using the COUNT function to calculate the frequencies from the grand total of the population.



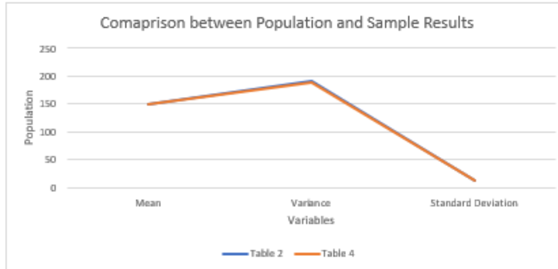
Graph 4: Relative Frequency Histogram of the Population

The relative frequency histogram shows a shape that is equivalent to the symmetric normal distribution of the population. However, since there are some variations in the frequencies, the distribution is not completely normal distribution. The graph focusses on how the number of data in bin relates to the other values in the bin. It can be observed from the graph that, the highest population frequency is 27.8 % and the lowest frequency being 0.50 % from a total of 100 %.

- **Central Limit Theorem**
  On comparing the results of Table 2 and Table 4 based on the Population and Sample results, it can be interpreted that the Mean of Population result i.e. 149.07, Variance i.e. 191.66 and Standard Deviation i.e. 13.84 is more than the Mean i.e. 149.66, Variance i.e. 189.66 and Standard Deviation i.e. 13.69 of the Sample results.

Comparison of Population and Sample Results		
	Table 2	Table 4
Mean	149.077	149.6608333
Variance	191.661	189.663141
Standard Deviation	13.8442	13.69321447

Table 2: Comparison of Results



Graph 5: Line Plot for Population and samples

According to the Central Limit Theorem, it can be observed that the sum of the variables tends towards the normal distribution, even though the frequencies vary from each other(3). As seen in the table, the mean, variance and standard deviation of Population results slightly varies with respect to that of Sample results. The line plot shows the variations between the two tables in terms of Mean, Variance and Standard Deviation.

Conclusion

1. It is inferred from the observations of the Keno game, that by taking the hyper-geometric probability distribution we could calculate the mean, variance and standard deviation of the 20 matches, the random values and simulations to decide the winner for the game.
2. Using the Law of Large Numbers, with the increase in number of trials, the experimental mean for the data was seen to approach the theoretical mean(2).
3. On the basis of the population given, we compared the mean, variance and standard deviation for the normal data with the variables of the 30 samples generated to by using the Central Limit Theorem. It was noted that the sum of samples tends towards a normal distribution even though the original variables are not normally distributed.

References

1. Bruce E. Trumbo California State University, Hayward *Journal of Statistics Education* v.3, n.2 (1995) Retrieved from https://www. http://jse.amstat.org/v3n2/trumbo.html
2. Chibisov, D. M. (2016). Bernoullis Law of Large Numbers and the Strong Law of Large Numbers. *Theory of Probability & Its Applications*, *60*(2), 318–319.
3. Hawkins, D. L., & Han, C.-P. (1996). A central limit theorem for certain nonlinear statistics in repeated sampling of a finite population. Arlington, TX: University of Texas at Arlington, Dept. of Mathematics.