Statistical Analysis on Probability Distributions

Xinfei Li

Northeastern University

1. Introduction

The report is divided into two parts. Firstly, based on the distribution characteristics of Keno game endings, it analyzes whether the distribution of the game results conforms to the law of large Numbers, and the results are in line with the law of large Numbers. The second part deals with a group of population data that conforms to normal distribution in advance, calculates the mean, variance and standard deviation of the population, and then expands the sample size to analyze whether there is a central limit theorem. The average of the results showed basically consistent results. But the standard deviation is very close to the population standard deviation, which is not consistent with the central limit theorem. The reasons are then given.

1. Analysis

Part1

The distribution of outcomes of game of Keno conforms to hypergeometric distribution. Keno introduces hypergeometric distribution to demonstrate and deduce the law of large Numbers. First, the hypergeometric distribution formula of R is used to calculate the probability of occurrence of various endings when 20 numbers are selected for each Keno game when the sample space is 20. This probability is used to calculate the expected value as the theoretical average of the Keno game.

The theoretical mean, theoretical mean and theoretical variance of x are calculated as below.

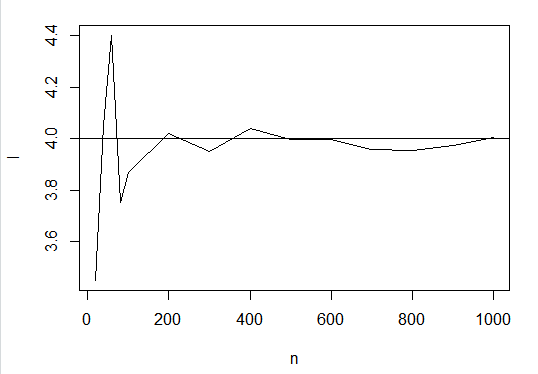
Theoretical Mean:

Theoretical Variance:

Theoretical Standard Deviation:

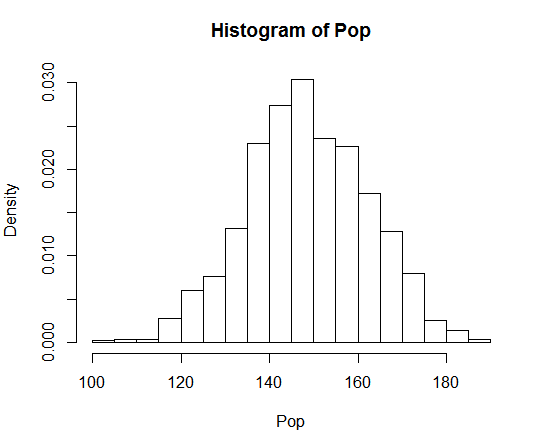
x is the number corresponding to the occurrence of 20 endings in the sample, indicating that the number of player choices in a game is the same as the number of computer choices. P(x) is the probability of happening under the corresponding outcome x.

Then use the random number formula of hypergeometric distribution to expand the number of games, and make experiments with the number of samples of 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000. And calculate the average value of each experiment. Based on this result, a line graph with the number of samples as the horizontal axis and the average as the vertical axis was made. Get the variation of the average as the sample size increases. Finally, the theoretical average value is put into the line graph for comparison.



According to the Law of Large Numbers, as the number of simulations become larger, and the experimental values of the means approach to their theoretical value. The theoretical average value is 4. As shown in the figure, when the number of samples gradually increases from 20 to 1000, the experimental average value and the theoretical average value are gradually consistent and stable around 4. The experimental mean value obtained from sampling is shown in the figure, which is not stable at 4. Especially in 600 to 1000 simulations, has a tendency to deviate from the first, then gradually close to. This is not because there is a problem with the experiment, but because there is a sampling error. Sampling error is inevitable, because with the simulations calculated experimental average cannot always equal to the overall situations. And as the number of simulation continues to increase, the experimental average is still to the theory of convergence of the average.

Part2



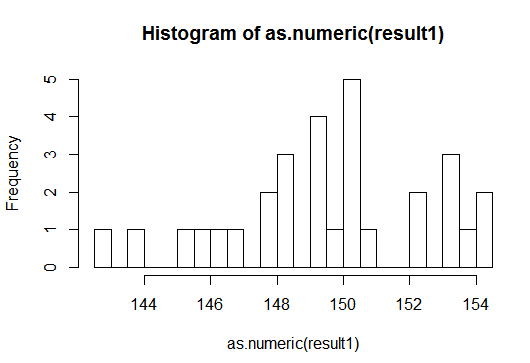
The normal population presents a bell-shaped column chart with symmetrical shape. Theoretically, the explanatory variable corresponding to the highest point is the average of the population, 149.077. As can be seen in the relative frequency histogram, the median and mode are close to the mean. The wider distribution possesses, the larger standard deviation. Since the standard deviation is small, the data shows more concentrated, which can also be observed through this histogram of population.

Then compare the calculated result with the value given by the Central Limit Theorem and analyze the similarities and differences.

|  |  |
| --- | --- |
| Population Mean | 149.077 |
| Sample mean | 149.5733 |
| Population Std. | 13.8511 |
| Sample Std. | 13.77588 |
| Std(Samples)=sd(Pop)/sqrt(30) | 2.528853 |

Because of the existence and irreducibility of the sample deviation, the sample mean is not equal to the population mean, but it is close, which conforms to the central limit theorem. According to the Central Limit Theorem, the sample standard deviation is times less than the population standard deviation, which is shown in the last row of the table in R calculation syntax. But in this demonstration, the sample standard deviation and the population standard deviation are equal.

According to the analysis, it may be that the calculation formula of the population sample difference in R is incorrect, so the function calculating the population sample difference in excel is used for calculation, but the expected result conforming to the central limit theorem is not obtained. So there should be other reasons to think about.



The 30 samples show a roughly bell-shaped shape, which is not large enough to show the characteristics of continuous distribution. With the increase of sample size, it will show a more regular normal distribution. And as the number of samples increases, the data will focus on the average, which means the standard deviation will decrease.

1. Conclusion

In the first part, based on the distribution characteristics of Keno game endings, it analyzes whether the distribution of the game results conforms to the law of large Numbers, and the results are in line with the Law of Large Numbers. As the number of simulation increases, the experimental mean of samples will converge to the real mean value (expected value).

The second part of the calculation did not verify the Central Limit Theorem, the difference is that the population standard deviation is too close to the sample standard deviation. The analysis suggested that part of the reason was that there was an error in the calculation of the standard deviation of the population, and R probably calculated the standard deviation of the population as a sample. But there should be other reasons.