STAT 3585

Project 1

Wei Liu

201759784

Verification of the Central Limit Theorem (CLT)

Introduction of the Central Limit Theorem

Given any distribution that has a well-defined mean and well-defined standard deviation, it can be either a continuous distribution or a discrete distribution, when we take the samples of the distribution, we calculate the mean of the samples. And we do the process repeatedly (for example, X times), then we can generate X means. The distribution of those means is an approximately normal distribution.

Method

1. Investigate different distributions:
    a. Binomial distribution
    b. Poisson distribution
    c. Uniform distribution
    d. Exponential distribution
2. Sample size from 2 to 50
3. Repeat times X = 100
4. Steps:
    a. Group A
        i. Take a random sample (with sample size of 2 to 50) from a binomial distribution (“rbinom” command)
        ii. Calculate the mean of the sample
        iii. Repeat the first two steps for X = 100 times, store the mean in a vector sample_mean
        iv. Calculate the mean of the sample means
        v. Use Jarque - Bera Normalality Test to test the normality of the sample means, get p-value.
        vi. Repeat the above steps for 100 times to get 100 p-values
        vii. Calculate the proportion of p-value which is greater than 0.05 (which means fail to reject the null hypothesis of the normality test)
        viii. Repeat the above steps for sample size of 2 to 50 to get 49 proportion values
        ix. Plot the 49 proportion values to analyze the trend of the proportion with the change of the sample size
    b. Group B
        i. Take a random sample from a Poisson distribution (“rpois” command)
        ii. Repeat the same process as Group A
    c. Group C

i. Take a random sample from an exponential distribution (“rexp” command)
ii. Repeat the same process as Group A

Simulations

1. Simulation environment: RStudio, install package "tseries"
2. Code in R (for Group C)

```
library(tseries)
set.seed(1)
simu = function(sample_size, repeat_time){
  sample_mean = c()
  lambda = 1
  for (i in 1:repeat_time){
    sample_mean[i] = mean(rexp(sample_size, lambda))
  }
  return (sample_mean)
}

proportion = c()
for (sample_size in 2:50){
  test = c()
  pvalue = c()
  pvalue2 = c()
  for (k in 1:100) {
    repeat_time = 100
    sample_mean = simu(sample_size, repeat_time)
    test = jarque.bera.test(sample_mean)
    test2 = jarqueberaTest(sample_mean)
    pvalue[k] = c(test$p.value)
  }
  proportion[sample_size] = sum(pvalue>0.05)/100
}

plot(proportion)
```

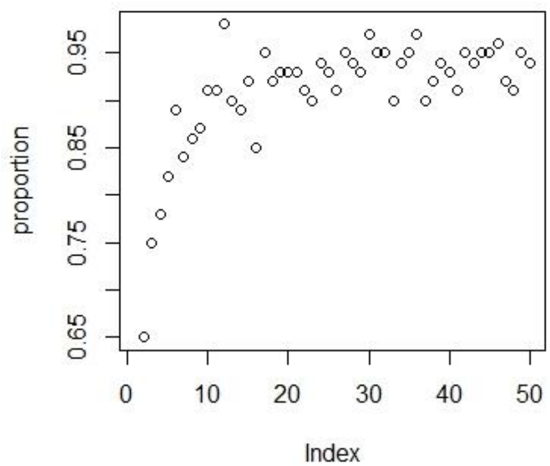
3. Parameters of each test
   a. The testing parameters of the binomial distribution are 10 and 0.1, which represent the number of trials and probability of success on each trial, respectively.

b. The testing parameter of the Poisson distribution is lambda as 1.
c. The testing parameter of the uniform distribution is default, which means it will generate random number ranging from 0 to 1.
d. The testing parameter of the exponential distribution is default, which means the vector of rates is 1.

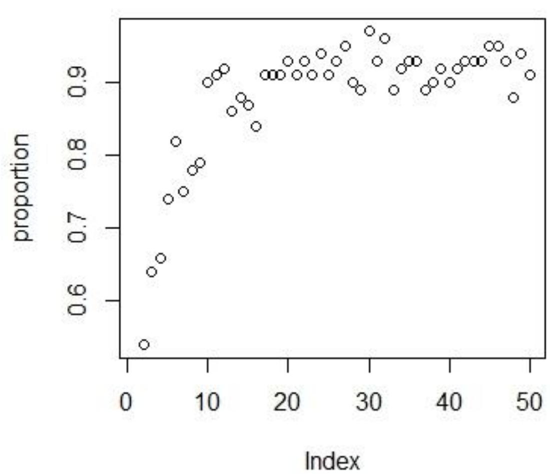
Results

The following tables shows the p value of Jarque - Bera Normalality Test for each simulation.

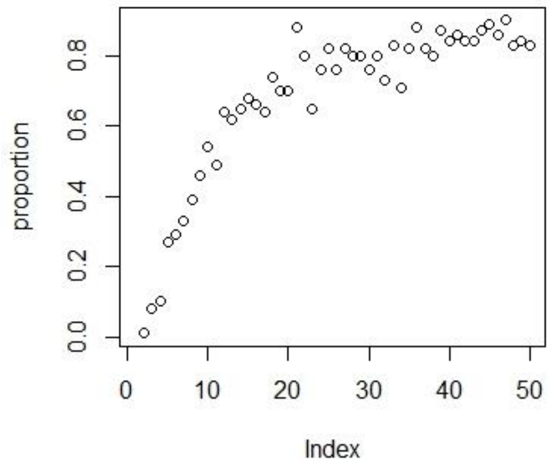
1. Group A (binomial distribution)



2. Group B (Poisson distribution)



3. Group C (exponential distribution)



Conclusion

1. By analyzing the simulation results above, we have verified the correctness of the central limit theorem.
2. When the sample size is greater than 20, the proportions of the p-values for the binomial distribution, Poisson distribution, and exponential distribution are above 90%, 80%, and 60%, respectively. It indicates that in the majority of the normality tests, we fail to reject the null hypothesis, so the distribution of the mean of the samples is approximately a normal distribution.
3. With the sample size increasing, the proportion increases accordingly, which means the larger sample size is, the closer their mean is to the normal distribution.

Further work

1. Different parameters of each distribution should be tested.
2. More distribution should be tested.
3. More methods should be used to test normality.