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:
   1. Binomial distribution
   2. Poisson distribution
   3. Uniform distribution
   4. Exponential distribution
2. Sample size: 20, 30, 40, 50
3. Repeat times X = 10, 100, 1000, 10000
4. Steps:
   1. Group A
      1. Take a random sample (with sample size of 20) from a binomial distribution (“rbinom” command)
      2. Calculate the mean of the sample
      3. Repeat the first two steps for X times, store the mean in a vector sample\_mean
      4. Plot the histogram of the sample means
      5. Calculate the mean of the sample means
      6. Plot the distribution of the theoretical means
      7. Compare the histogram and the distribution of the theoretical means
      8. Use Jarque - Bera Normalality Test to test the normality of the sample means
      9. Do the above steps repeatedly with sample size of 30, 40, and 50
   2. Group B
      1. Take a random sample from a Poisson distribution (“rpois” command)
      2. Repeat the same process as Group A
   3. Group C
      1. Take a random sample from a uniform distribution (“runif” command)
      2. Repeat the same process as Group A
   4. Group D
      1. Take a random sample from an exponential distribution (“rexp” command)
      2. Repeat the same process as Group A

# Simulations

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

library(fBasics)

set.seed(3585)

simu = function(sample\_size = 20, repeat\_time = 10){

sample\_mean = c()

lambda = c(0,1,2,3,4)

for (i in 1:repeat\_time){

sample\_mean[i] = mean(rpois(sample\_size, lambda))

}

return (sample\_mean)

}

sample\_size = 200

repeat\_time = 1000

sample\_mean = simu(sample\_size, repeat\_time)

hist(sample\_mean)

jarqueberaTest(sample\_mean)

1. Parameters of each test
2. 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.
3. The testing parameter of the Poisson distribution is the vector [0, 1, 2, 3, 4], which represent the vector of means.
4. The testing parameter of the uniform distribution is default, which means it will generate random number ranging from 0 to 1.
5. 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)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| sample size  repeat times | 20 | 30 | 40 | 50 |
| 10 | 0.8064 | 0.7146 | 0.73 | 0.9837 |
| 100 | 0.2471 | 0.2236 | 0.1192 | 0.1717 |
| 1000 | 0.01836 | 0.1739 | 0.1247 | 0.06067 |
| 10000 | 3.305e-10 | 2.555e-06 | 4.997e-06 | 2.281e-08 |

1. Group B (Poisson distribution)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| sample size  repeat times | 20 | 30 | 40 | 50 |
| 10 | 0.641 | 0.7951 | 0.6233 | 0.655 |
| 100 | 0.5829 | 0.6619 | 0.5933 | 0.5303 |
| 1000 | 0.03308 | 0.1149 | 0.0178 | 0.05361 |
| 10000 | 8.406e-07 | 1.233e-06 | 7.18e-07 | 0.0006701 |

1. Group C (uniform distribution)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| sample size  repeat times | 20 | 30 | 40 | 50 |
| 10 | 0.5447 | 0.7546 | 0.5852 | 0.5644 |
| 100 | 0.5695 | 0.9329 | 0.4924 | 0.008392 |
| 1000 | 0.04496 | 0.1286 | 0.6559 | 0.6989 |
| 10000 | 0.02949 | 0.5717 | 0.9565 | 0.1831 |

1. Group D (exponential distribution)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| sample size  repeat times | 20 | 30 | 40 | 50 |
| 10 | 0.8901 | 0.724 | 0.8778 | 0.768 |
| 100 | 0.03216 | 0.05688 | 0.01675 | 4.011e-05 |
| 1000 | 1.182e-11 | 1.344e-13 | 0.0002431 | 8.641e-05 |
| 10000 | < 2.2e-16 | < 2.2e-16 | < 2.2e-16 | < 2.2e-16 |

# Conclusion

1. It’s surprising that the p-value is getting smaller when the repeat times increase. Theoretically, the results should be opposite.
2. With the sample size increases, the p-value nearly maintains the same, however, theoretically, the distribution of the samples should be closer to the normal distribution, so the p-value should be larger.
3. For different distributions (between groups), the results vary in a great range.
4. Despite of the fact above, the histogram of the sample means in each simulation is very close to normal distribution. When the sample size increases, or/and the repeat times increase, the histogram is closer 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.