Statistical Inference Course Project Part 1

Semin (Sammie) Bae

University of Waterloo

Overview:

This report will explore the Central Limit Theorem which states that the distribution of averages of variables becomes that of a standard normal as the sample size increases. This theorem will be tested by comparing running mathematical models on sample data.

Simulations

Code: #-comment/explanation

```
#create simulation data with 4000 data sets
simulationData = rexp(n = 4000, rate = .2)
hist(simulationData)
#make into matrix  with data sets grouped of 40 exponentials
matrixdata =  matrix(simulationData, nrow = 100, ncol = 40)
dim(matrixdata)
resultCLT <- apply(matrixdata, 1, mean)
#mean and variance
mean(resultCLT)
var(resultCLT)
#create the histogram of Central Limit Theorem applied grouped data set
hist(resultCLT)
```

The above code first creates simulation data of 4000 data sets with 0.2 lambda, creates and histogram as shown in figure 1. Then new data set of matrix data is created by grouping 40 exponentials into 100 groups and its histogram is created as shown in figure 2. This R program also shows both mean and variance of the distribution.

Mean Analysis

Sample Mean

Code:

```
> mean(resultCLT)
[1] 5.035011
```

Theoretical Mean (*μ*)

“mean is 1/lambda”:

Therfore: Theoretical Mean (μ) = 1/0.2 = 5

Conclusion (Mean):

The sample mean is very close to the actual value of 1000 sample data. This means that the distribution is centered around there and is very close to the thereotical value of 5. As data size increases, the sample mean will approach closer to 5

Variance Analysis

Sample Variance

Code:

```
> var(resultCLT)
[1] 0.6685216
```

Theoretical Variance

Variance = √Standard Deviation = sqrt(5) = 2.2360679775

Theoretical Standard Deviation(σ) = 1/0.2 = 5

Theoretical Standard Deviation(σ) = 1/0.2 = 5

Conclusion (Variance):

The sample variance is considerably lower than the theoretical variance (it is only 30% of the theoretical). This is because as central of Central Limit Theorem; as averages of the random samples were taken, it became more clustered around mean and therefore variance was reduced.

Distributions

Figure 1: Raw 4000 exponents:

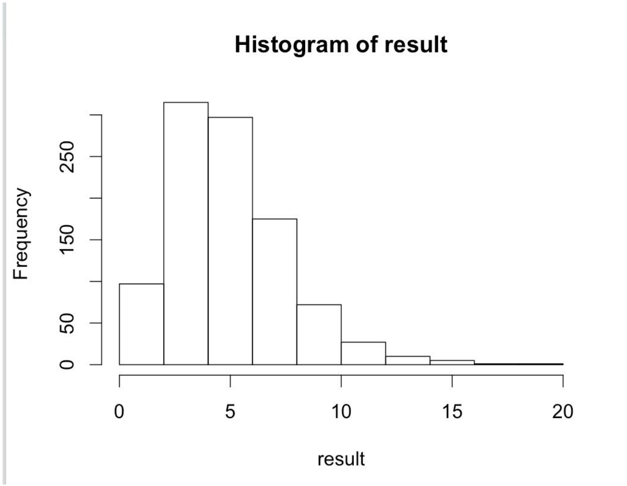
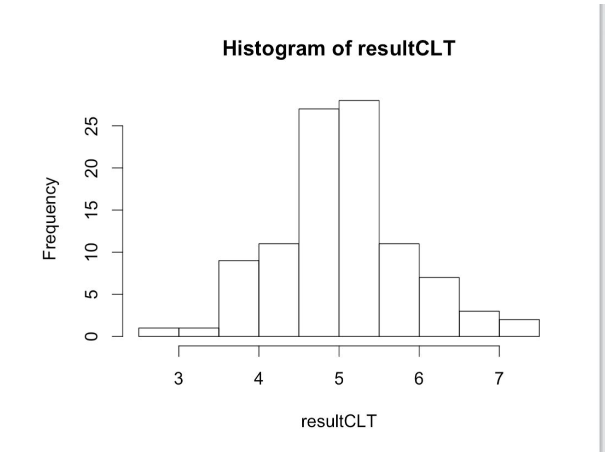


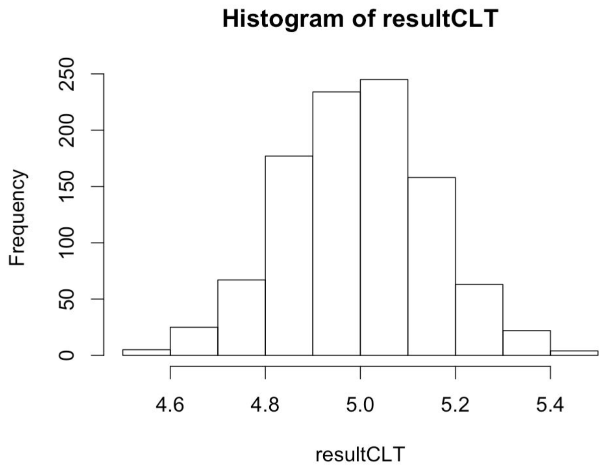
Figure 2: 100 groups of 40 exponents:



Conclusion (Distribution):

As shown in both Figure 1 and Figure 2, both distributions seem to normally distributed with the mean (5) at the center. One key thing to note, however, is that Figure 2 is has less variance and more clustered around the middle. As data size and grouping size increases, it will become more of a perfect normal distribution as exemplified here:

- 1000 by 1000 data set



-