Statistical Inference - Course Project

BMc

Monday, February 16, 2015

Overview:

The purpose of this report is to explore the statistical concepts taught in Coursera: Statistical Inference. To do this, a thousand simulations of 40 random readings each will be taken for an exponential distribution. The average mean (the mean of means) and the average variance will be calculated and compared to the theoretical mean and variance as calculated by the Central Limit Theorem.

The exponential distribution is simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. lambda is set to 0.2 for all of the simulations.

Simulations

The below code simulates a random exponential distribution 1000 times with a sample size of 40 and a lambda value of 0.2. The data is stored for further calculations.

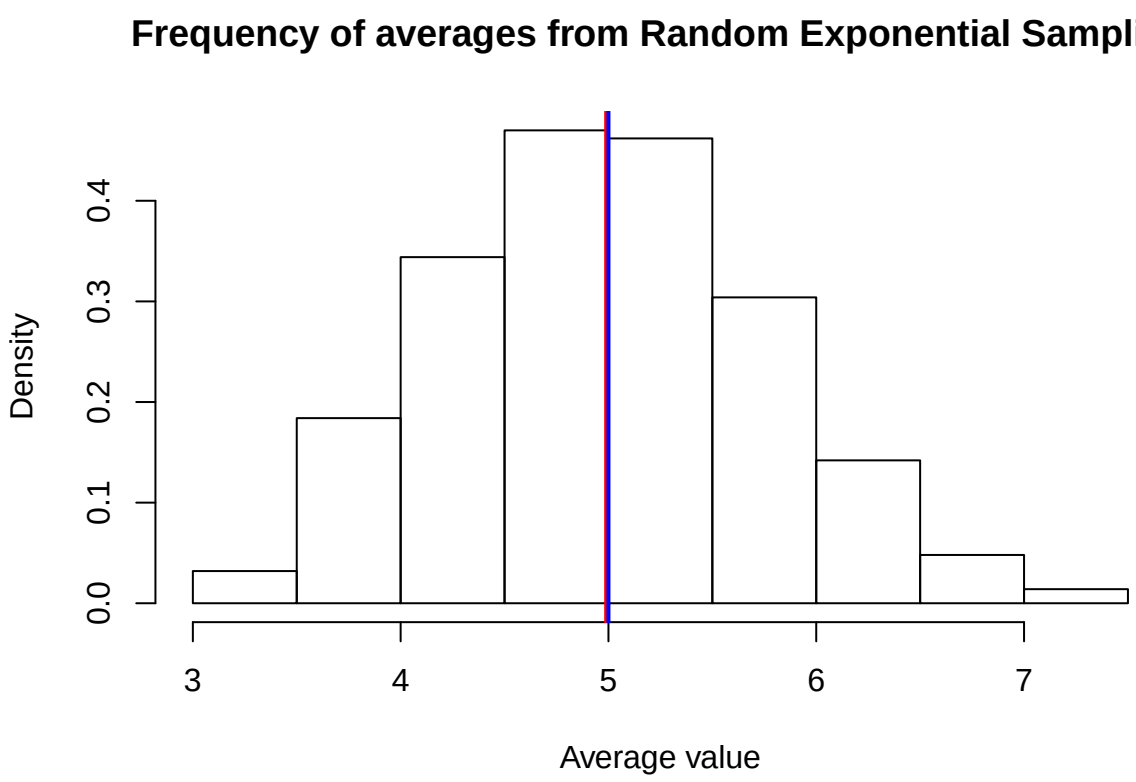
```
mean_sample_data = NULL
lambda <- 0.2
samplesize <- 40
sample_table = NULL

for (i in 1 : 1000)
{
  sample_data <- rexp(samplesize, lambda)
  mean_sample_data = c(mean_sample_data, mean(sample_data))
  # Store Sample data just in case
  sample_table <- rbind(sample_table, sample_data)
}
```

Sample Mean versus Theoretical Mean

The theoretical mean of an exponential distribution is 1/lambda. Given that lambda = 1/5 = 0.2 in this project, the theoretical mean is 5.

The computation of the experimental mean is shown below. It has been mapped against the histogram of sample data. The experimental mean is mapped as a red vertical line and the theoretical mean is mapped as a blue vertical line. As shown in the chart, the two are virtually indistinguishable. Therefore, a calculation is done in order to show the absolute difference between the two.

```
theoretical_mean <- 1 / lambda
experimental_mean <- mean(mean_sample_data)
hist(mean_sample_data, freq=FALSE, xlab="Average value", main="Frequency of averages from Random Exponer
abline(v=experimental_mean, col="red",lwd=2)
abline(v=theoretical_mean, col="blue",lwd=2)
```



The difference between the theoretical and experimental mean is:

```
abs(theoretical_mean - experimental_mean)
```

```
## [1] 0.01087362
```

or a percentage difference from the theoretical mean of:

```
abs(theoretical_mean - experimental_mean) / theoretical_mean
```

```
## [1] 0.002174725
```

Sample Variance versus Theoretical Variance

The theoretical variance of an exponential distribution is 1/lambda. Given that lambda = 1/5 = 0.2 in this project, the theoretical variance is 5.

The computation of the experimental variance is shown below.

```
variance_array <- NULL
for( i in 1:nrow(sample_table))
{
  variance_array <- rbind(variance_array, var(sample_table[i,]))
}
```

```r
sample_variance <- mean(variance_array)
sample_sd <- sd(mean_sample_data)

# Variance is the Standard Deviation squared
# The Standard Deviation of an exponential distribution is 1 / lambda
theoretical_variance <- (1 / lambda) ^ 2

sample_variance
```

```
## [1] 24.35274
```

```r
theoretical_variance
```

```
## [1] 25
```

```r
abs(theoretical_variance - sample_variance)
```

```
## [1] 0.6472626
```

Distribution

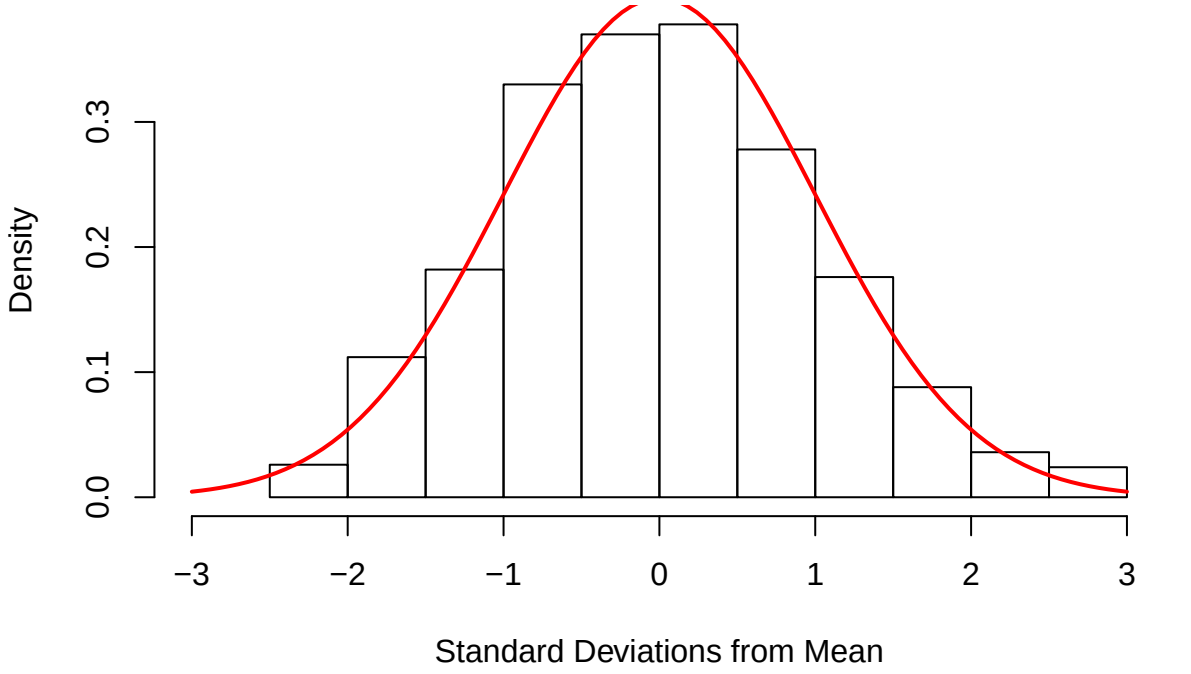
The Central Limit Theorem states that the mean of a large number of samples with a well defined variance will be approximately normally distributed.

To test this, first the mean sample data was normalized by subtracting the mean of the runs from each value and then dividing each value by the standard deviation of the vector of means. The purpose of this is to simply shift the graph and linearly scale the graph so that the x axis shows the number of standard deviations from the absolute mean (represented at x value = 0).

Next, a curve matching a normal standard deviation is plotted on top of the histogram. It is expected that the histogram will closely match this curve (per the CLT).

From below, it can be seen that the simulations run do adhere to the CLT.

```r
adj_data <- (mean_sample_data - mean(mean_sample_data)) / sd(mean_sample_data)
hist(adj_data, freq=FALSE, xlab="Standard Deviations from Mean", main="Frequency of averages from Rando
curve(dnorm(x, mean=mean(adj_data),sd=sd(adj_data)), add=TRUE, col="red", lwd=2)
```


Frequency of averages from Random Exponential Sampling
Density
0.0
0.1
0.2
0.3
−3
−2
−1
0
1
2
3
Standard Deviations from Mean