Simulation and Basic Inferential Data Analysis

Kent Mok

Overview:

The purpose of this experiment is to investigate the exponential distribution in R, and to compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, λ) where λ is the rate parameter and the mean of exponential distribution is $\frac{1}{\lambda}$. (λ will be set to 0.2 for this simulation.) This experiment will use 1000 simulations of the average of 40 exponentials to investigate the distribution.

The mean of the exponential distribution is $\beta = \frac{1}{\lambda}$. For $\lambda = 0.2$, $\beta = 5$. The variance of the exponential distribution is given as $1/\lambda^2$. For $\lambda = 0.2$, the variance is $1/\lambda^2 = 1/0.2^2 = 25$.

Simulation:

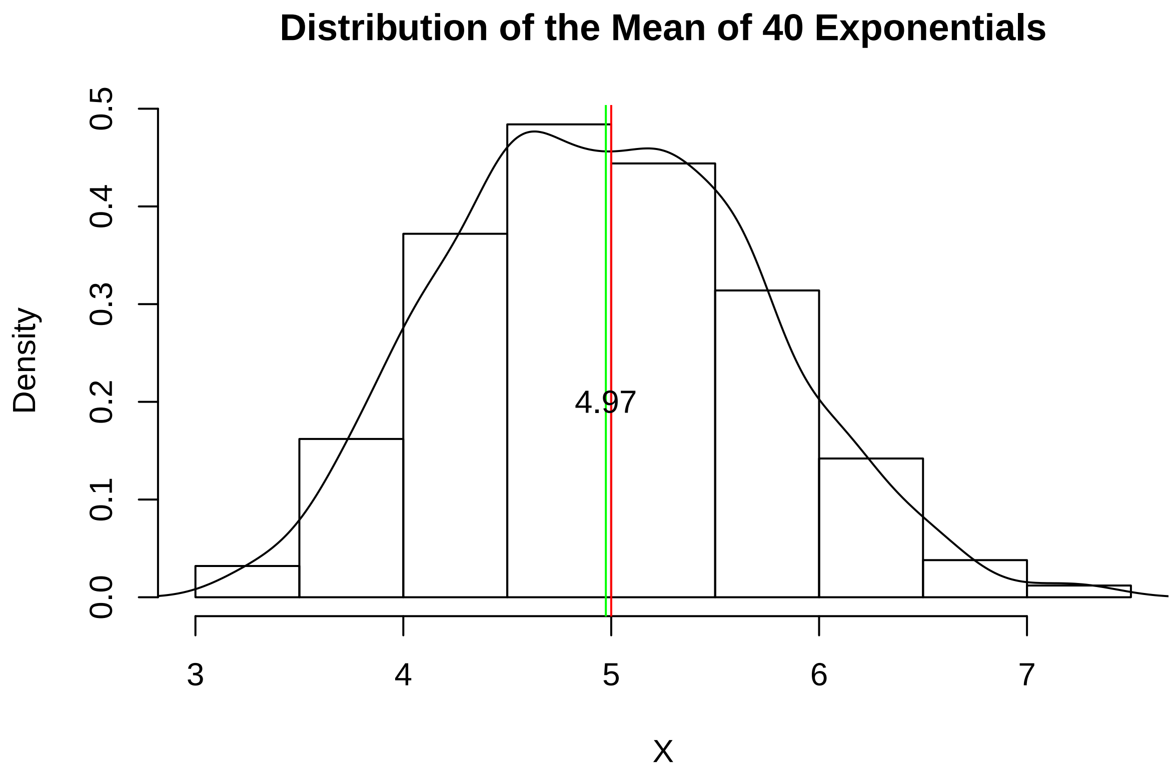
The simulation takes 40 samples of the exponential distribution 1000 times. This creates a data frame with 40 variables and 1000 observations each. The mean each observation is used to create the distribution of the mean of the exponential distribution.

```
set.seed(1234)

lambda = 0.2
numofsims = 1000
# Create data frame of the 1000 simulations of 40 exponentials
expdf <- data.frame(t(replicate(numofsims, rexp(40, lambda))))
```

Sample Mean versus Theoretical Mean:

The Central Limit Theorem states that for large enough sample sizes, the distribution of the mean of independent and identically distributed samples will be a standard normal distribution with mean μ and variance $\frac{\sigma^2}{n}$. Plotting the distribution of the mean of exponentials should yield a normal distribution curve. For the exponential distribution with $n = 40$, $\mu = \beta = 5$ and $\frac{\sigma^2}{n} = \frac{25}{40} = 0.625$.



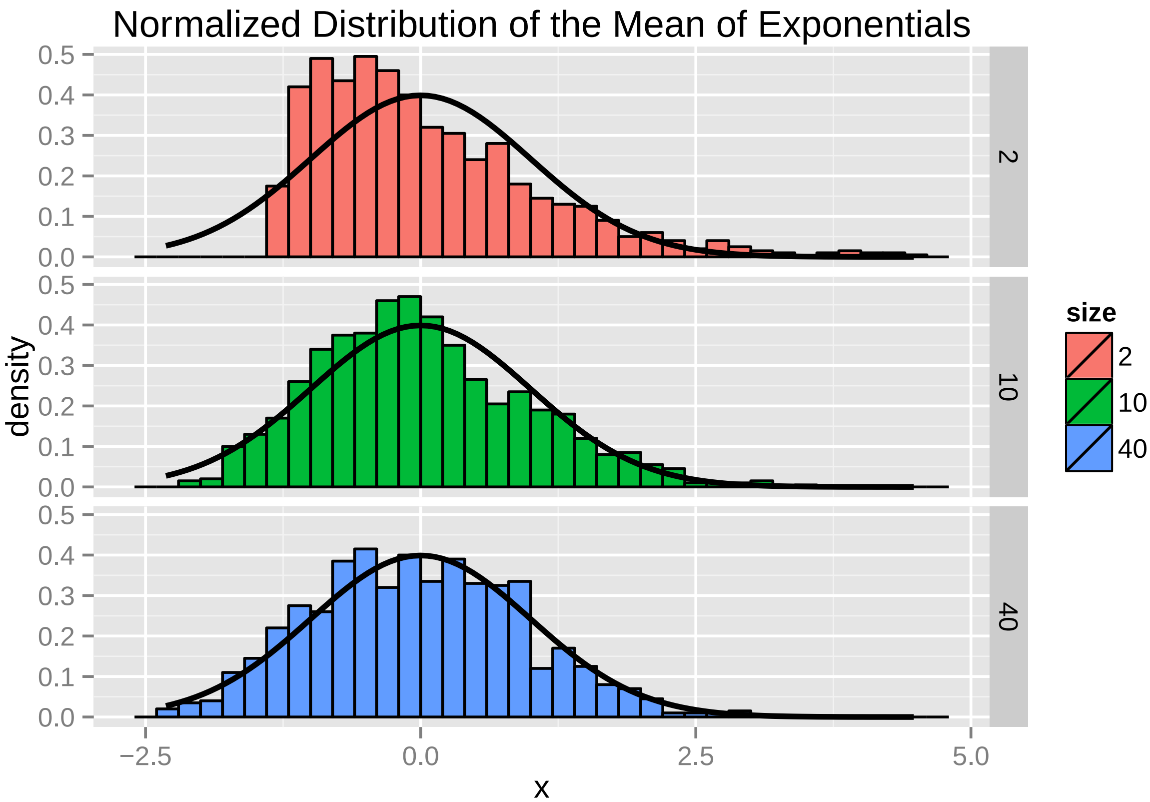
The above figure shows the distribution of the thousand simulated means for 40 exponentials. The mean of the distribution is shown as a green vertical line at **4.974**. The theoretical mean is shown as a red vertical line at $x = 5$. It is apparent that the sample mean for the mean of 40 exponentials matches the theoretical mean of $\mu = 5$.

Sample Variance versus Theoretical Variance:

The variance of these 1000 means of 40 exponentials is 0.571. The theoretical variance of the mean of 40 exponentials is $\frac{\sigma^2}{n} = 0.625$. At $n = 40$, the variance is quite small and would continue to decrease as the sample size increases.

Distribution:

The Central Limit Theorem states that the distribution of averages of independent and identically distributed variables becomes that of a normal distribution as the sample size increases, so $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ would have a distribution like that of a standard normal for large n. From the above simulation, $\mu = 5$ and $\sigma = \sqrt{Var[X]} = 5$. Applying this to the above simulation of the exponential distribution yields the following plots.



The plots above show the normalized mean values of the exponential distribution for 2 samples, 10 samples, and 40 samples. Also, standard normal distribution curves have been superimposed on the plots. It is apparent that the plots are approaching that of the normal distribution curve. Moreover, the curves converge toward the normal distribution curve quite quickly as n is increased towards 40.

Appendix

Sample Mean versus Theoretical Mean Code:

```r
mns = apply(expdf, 1, mean)  # Take mean value of each of the 40 samples
samplemean = mean(mns)  # Average of the 1000 means

hist(mns, freq = FALSE, main = "Distribution of the Mean of 40 Exponentials",
     xlab = "X")
lines(density(mns))  # Add smoothed curve on top of histogram
abline(v = samplemean, col = "green")  # Add vertical line for sample mean
abline(v = 1/lambda, col = "red")  # Add vertical line for theoretical mean
text(x = samplemean, y = 0.2, labels = round(samplemean,2))
```

Sample Variance versus Theoretical Variance Code:

```r
sampvar = var(mns)  # Variance of the 1000 means
```

Distribution Code: