Simulation and Basic Inferential Data Analysis

Kent Mok

September 9, 2015

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, lambda) where lambda is the rate parameter and the mean of exponential distribution is 1/lambda. (Lambda will be set to 0.2 for this simulation.) This experiment will use 1000 simulations of the averages of 40 exponentials to investigate the distribution.

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

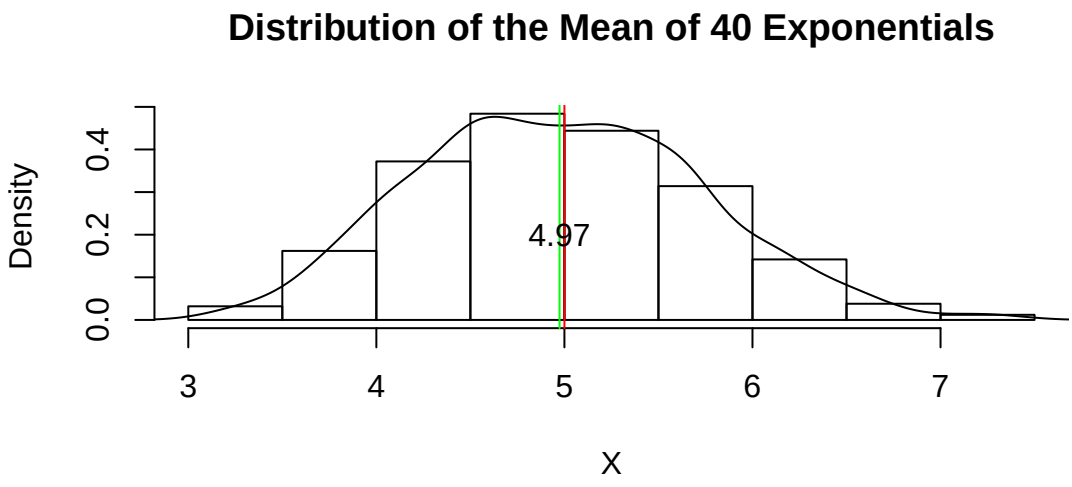
Simulation:

The simulation takes 40 samples of the exponential distribution 1000 times. This creates a data frame with 1000 variables with 40 observations each. The mean and variance of each variable can be taken to analyze the distribution of the mean and variance for the exponential distribution.

```
set.seed(1234)

lambda = 0.2
expdf <- data.frame(replicate(1000, rexp(40, lambda)))
```

Sample Mean versus Theoretical Mean:

```
numofsims = 1000
mns = apply(expdf, 2, 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))
```



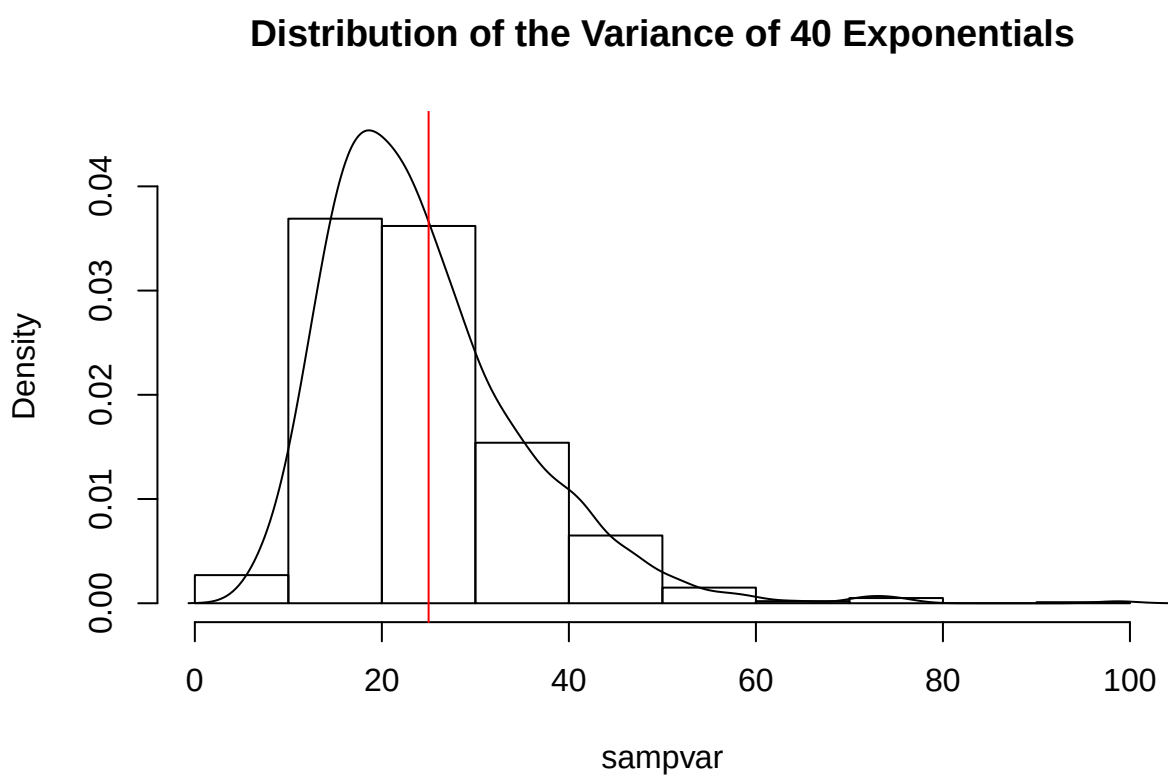
The above figure shows the distribution of the thousand simulated means for 40 exponentials. The sample mean is shown as a green vertical line at the sample mean of 4.974. The theoretical mean is also shown as a red vertical line at $X = 5$. It can be seen that the theoretical mean coincides with the sample mean, at least for a sample of 1000 simulations.

Sample Variance versus Theoretical Variance:

```
sampvar = var(mns)
print(sampvar)
```

```
## [1] 0.5706551
```

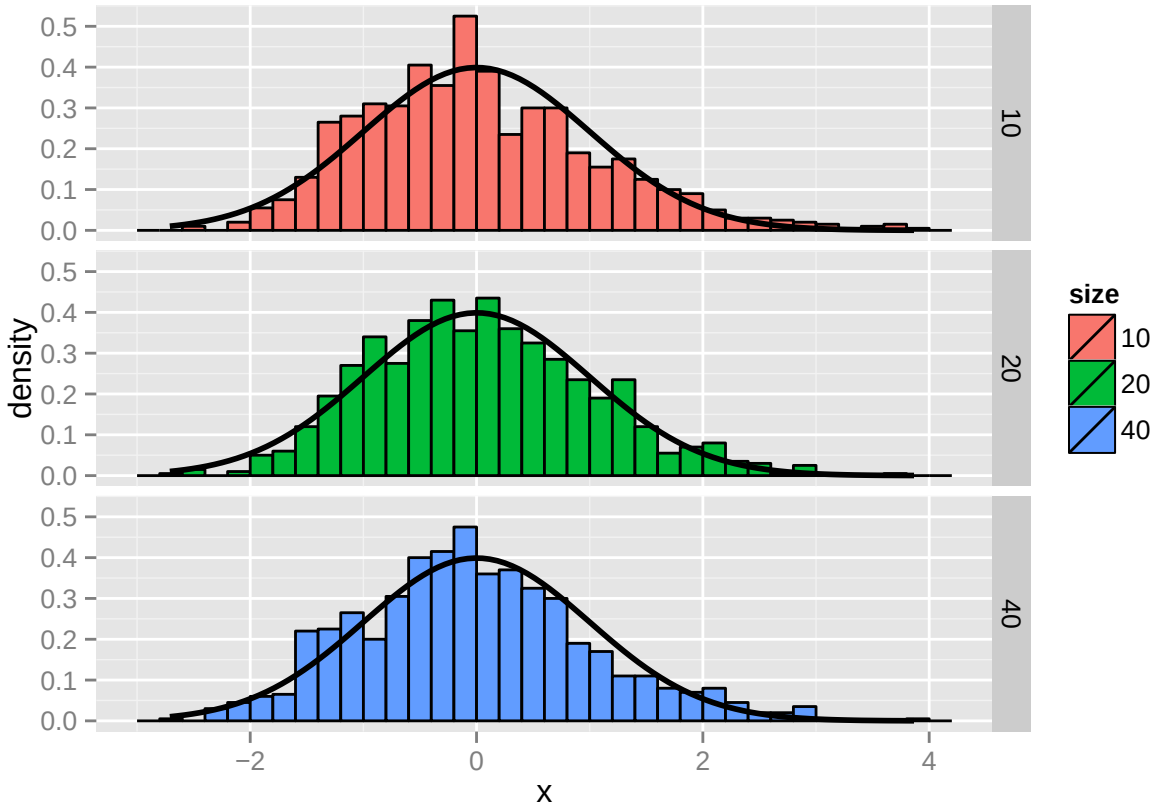
```
sampvar = NULL
for (i in 1:numofsims)
    sampvar = c(sampvar, var(rexp(40, lambda)))

ylimits = range(density(sampvar)$y)
hist(sampvar, freq = FALSE, main = "Distribution of the Variance of 40 Exponentials", ylim = ylimits)
lines(density(sampvar))
abline(v = 1/lambda^2, col = "red")
```



The above figure shows the distribution of one thousand simulations of the variance value for 40 exponentials. The theoretical variance is shown as a red vertical line at $X = 25$. It can be seen that the sample variance coincides with the theoretical variance, at least for a sample of 1000 simulations.

Distribution: Via figures and text, explain how one can tell the distribution is approximately normal.

```
library(ggplot2)
nosim <- 1000
cfunc <- function(x, n) sqrt(n) * (mean(x) - 5) / 5
dat <- data.frame(
  x = c(apply(matrix(rexp(nosim * 10, lambda), nosim), 1, cfunc, 10),
        apply(matrix(rexp(nosim * 20, lambda), nosim), 1, cfunc, 20),
        apply(matrix(rexp(nosim * 40, lambda), nosim), 1, cfunc, 40)
        ),
  size = factor(rep(c(10, 20, 40), rep(nosim, 3))))
g <- ggplot(dat, aes(x = x, fill = size)) + geom_histogram(colour = "black", binwidth = 0.2, aes(y = ..
g <- g + stat_function(fun = dnorm, size = 1)
g + facet_grid(size ~ .)
```


density
0.5
0.4
0.3
0.2
0.1
0.0
10
20
40
−2
0
2
4
x
size
10
20
40