Statistical Analysis of the Exponential Distribution in R

Arun K Viswanathan

January 19, 2015

Overview

This paper investigates the exponential distribution in R and compares it with the Central Limit Theorem.

Simulations

```
lambda <- 0.2     # Lamdba
n <- 40           # Number of exponentials
numsim <- 100000  # Number of simulations
```

We simulate an exponential distribution using `rexp`, with 40 exponentials and lamda of 0.2. The simulation is performed 100000 times. The code below shows the simulation in action:

```
set.seed(20150119) # Seed to ensure reproducibility
sim <- matrix(rexp(n * numsim, lambda), numsim)
```

Sample Mean versus Theoretical Mean

The code below computes the sample mean from the simulated data.

```
mean.theoretical <- 1/lambda
means.sim <- apply(sim, 1, mean)
mean.sim <- mean(means.sim)
mean.error <- abs(mean.sim - mean.theoretical) / mean.theoretical * 100
```

The sample mean is **4.9996** while the theoretical mean is **5**. This is an error of **0.0079341%**. This clearly shows that the Central Limit Theorem in action as the mean of the simulated means is almost the same as the theoretical mean.

Sample Variance versus Theoretical Variance

The sample variance can be computed using the code show below:

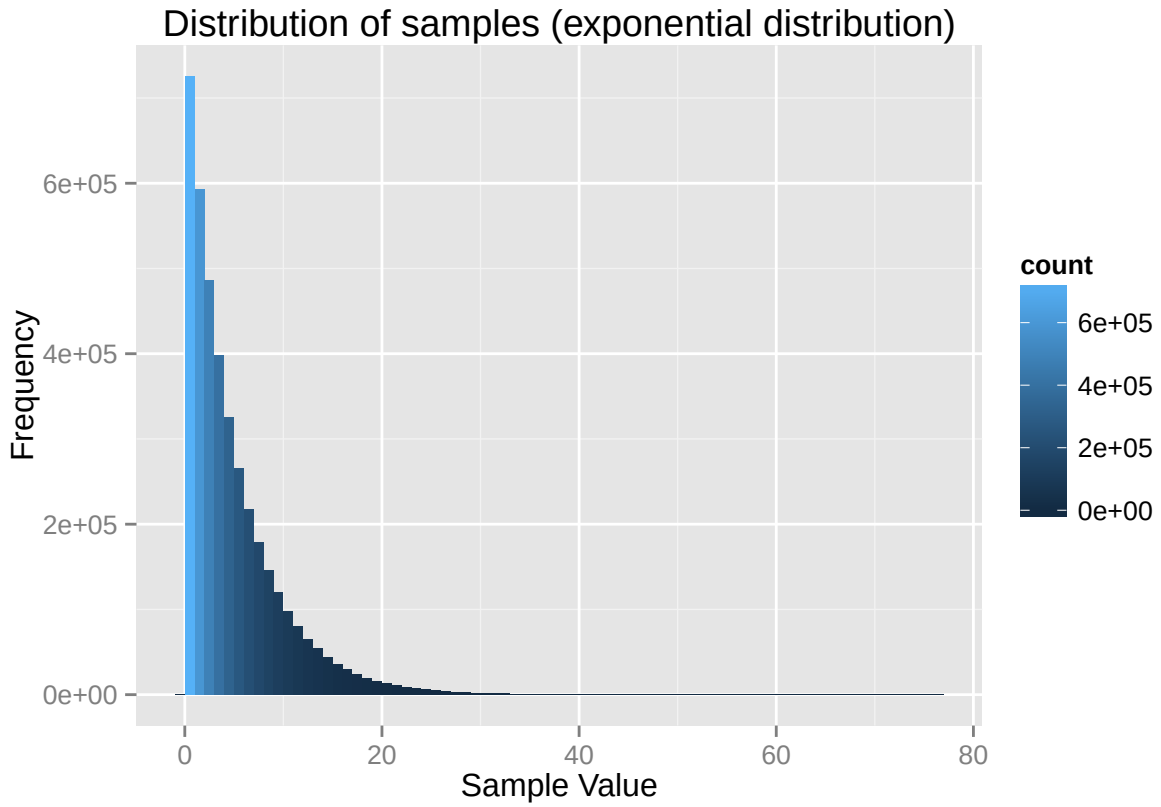
```
sd.theoretical <- 1/lambda
var.theoretical <- sd.theoretical ^ 2
sds.sim <- apply(sim, 1, sd)
sd.sim <- mean(sds.sim)
var.sim <- sd.sim ^ 2
var.error <- abs(var.sim - var.theoretical) / var.theoretical * 100
```

The average sample variance is **23.8824** (standard deviation of 4.88696) while the theoretical variance is **25** (standard deviation of 5). This is an error of **4.47046%** in the variance.

Distribution

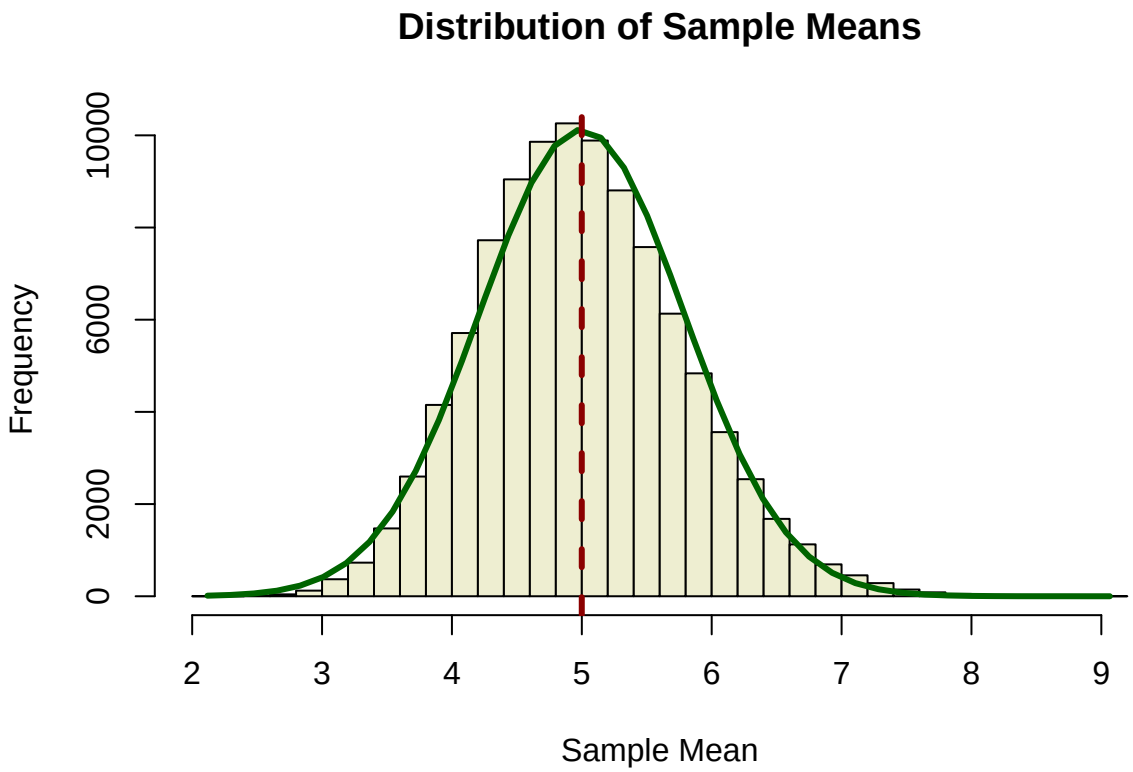
The plot below shows the distribution of the *simulated samples*.

```
library(ggplot2)
g <- ggplot(data.frame(x = as.vector(sim)), aes(x = x)) +
    geom_histogram(aes(fill = ..count..), binwidth = 1) +
    ggtitle("Distribution of samples (exponential distribution)") +
    xlab("Sample Value") +
    ylab("Frequency")
print(g)
```



The histogram below shows the distribution of the *sample means*. The red vertical dashed line is the mean of the sample means and is at 4.9996033. A normal distribution is overlaid over the histogram in dark green.

```
h <- hist(means.sim,
          breaks = 30,
          col = "lightyellow2",
          main = "Distribution of Sample Means",
          xlab = "Sample Mean",
          ylab = "Frequency")
xfit <- seq(min(means.sim), max(means.sim), length = 40)
yfit <- dnorm(xfit, mean = mean.sim, sd = sd(means.sim))
yfit <- yfit * diff(h$mids[1:2]) * length(means.sim)
lines(xfit, yfit, col = "darkgreen", lt = 1, lw = 3)
abline(v = mean.sim, col = "darkred", lt = 2, lw = 3)
```



Looking at the histogram above, the distribution of the sample means appears to be close to a normal distribution.