Simulation of Exponential Distribution

Rachit Kinger

17 August 2017

Overview

This project tests the applicability of the Central Limit Theorem on the exponential distribution in R. The Cental Limit Theorem states that the distribution of averages of iid variables becomes that of a standard normal as the sample size increases. This is true even if the IID variable itself does not follow a normal distribution.

The projects tests the CLT on exponential distribution in R. The properties of the distribution are: * mean of exp distribution is $1/\lambda$
* the standard deviation of the distribution is also $1/\lambda$
* probability density function is $\lambda e^{1/z\lambda}$

Methodology and results

Two simulations were run, one with 40 observations per sample and the other with 200 observations per sample. Each simulation was run a thousand times. Results indicate that
1. observed mean is very close to theoretical mean
2. variance of the sample is very close to the theoretical variance
3. the distribution of sample means is normal and tends to get lower variance as sample size increases from 40 to 200

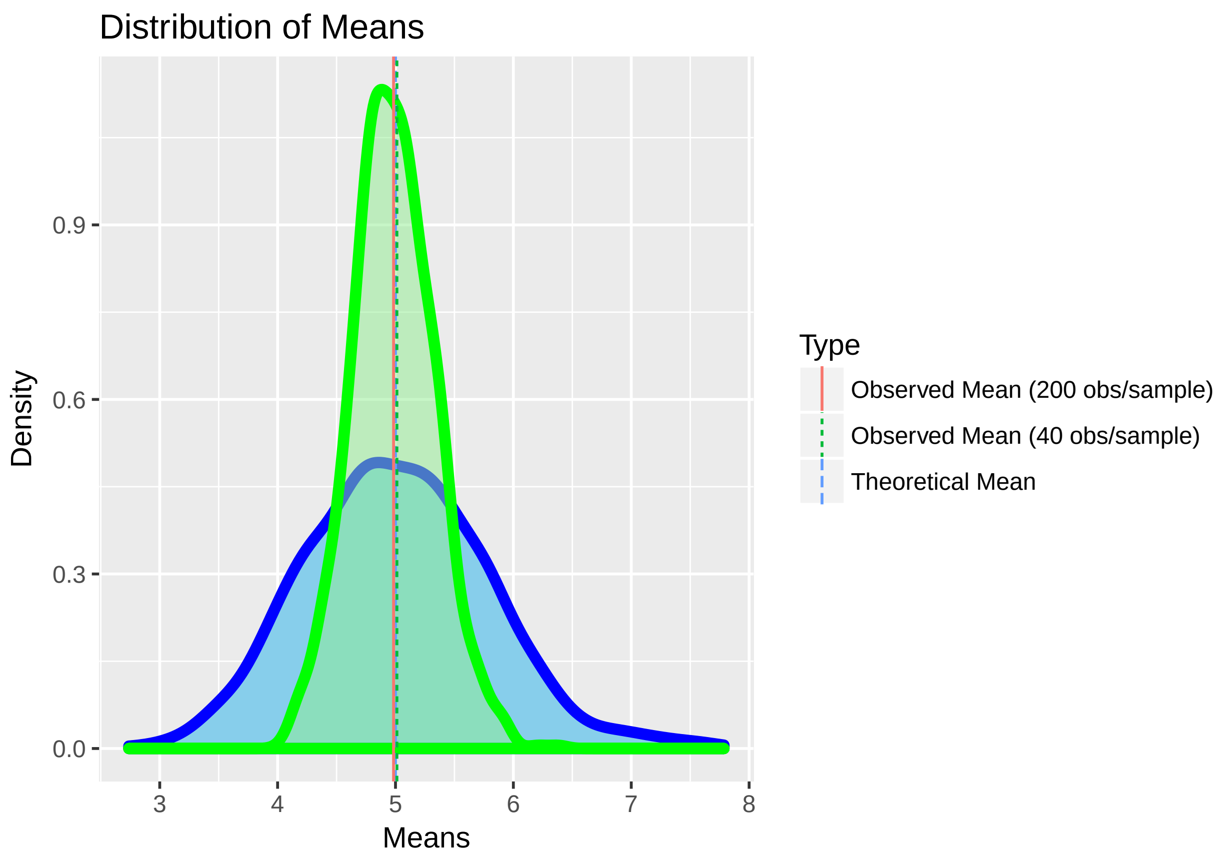
Simulations

We created simulations with λ `= 0.2`. Using rexp 1000 samples were created with 40 observations in each. The code used is as follows:

```
set.seed(123)
nosim <- 1000
samplesize1 <- 40
samplesize2 <- 200
lambda <- 0.2
simulation1 <- rexp(nosim*samplesize1, lambda)
simulation2 <- rexp(nosim*samplesize2, lambda)
simsam1 <- matrix(simulation1, ncol = samplesize1) ##creates the simulated samples with 40 obs each
simsam2 <- matrix(simulation2, ncol = samplesize2) ##creates the simulated samples with 80 obs each
means1 <- apply(simsam1, 1, mean) ##calculates 1000 means of 40 obs each
means2 <- apply(simsam2, 1, mean)
ObservedMean1 <- mean(means1)
ObservedMean2 <- mean(means2)
TheoreticalMean <- 1/lambda
meanCompare <- data.frame(Type = c("Theoretical Mean",
                                   "Observed Mean (40 obs/sample)",
                                   "Observed Mean (200 obs/sample)"),
                          Value = c(TheoreticalMean, ObservedMean1, ObservedMean2))
```

We then plot these means to see what kind of a distribution we have:

```r
library(ggplot2)
g <- ggplot(data.frame(means1), aes(means1))
g + geom_density(lwd = 2, color = "blue", fill = "sky blue") +
    geom_density(aes(data = data.frame(means2), x = means2),
                 lwd = 2, color = "green", fill = "light green", alpha = 0.5) +
    geom_vline(data = meanCompare,
    aes(xintercept = Value , linetype = Type, colour = Type), show.legend = TRUE) +
    ggtitle("Distribution of Means") +
    labs(x = "Means", y = "Density")
```



The blue distribution is of samples with 40 observations each, and green distribution is of samples with 200 observations in each sample. We can see that the distribution of means is Gausian looking, and it gets more centered around its mean as we increase the sample size from 40 to 200. So it looks like the CLT applies well in our scenario. But let's do some further analysis.

Comparing Theoretical & Observed Means

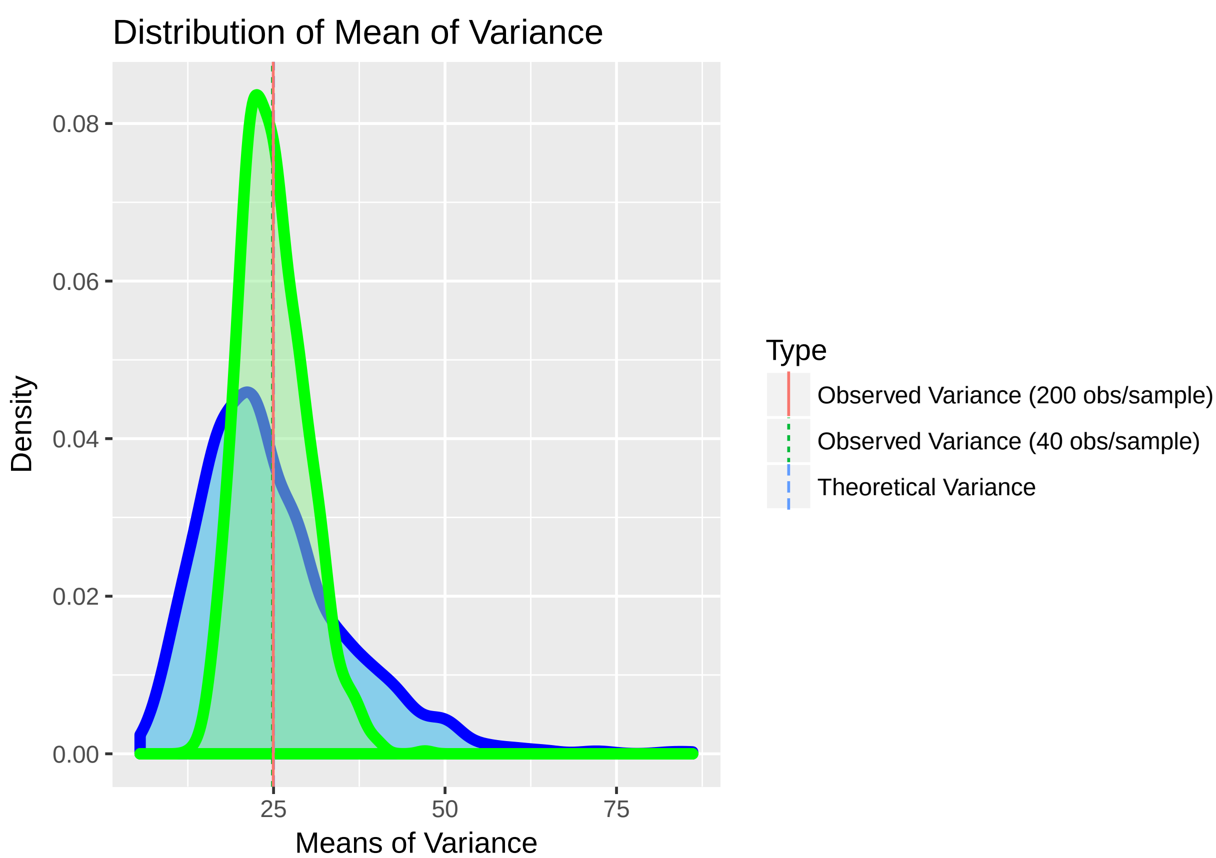
From the three, nearly overlapping vertical lines in the above figure, we can also see that the Observed Means and Theoretical Mean are quite close. We know that the mean for an exponential distribution is $1/\lambda$ so for this simulation since our λ is 0.2 our theoretical mean is 5. Below are the actual values of theoretical and observed means for comparison.

```
##                             Type    Value
## 1               Theoretical Mean 5.000000
## 2  Observed Mean (40 obs/sample) 5.011911
## 3 Observed Mean (200 obs/sample) 4.983538
```

Calculating Simulated Variance

The theoretical variance for an exponential distirbution is $1/\lambda^2$. Since λ is 0.2 in our simulation, the variance $= 25$.

```
vars1 <- apply(simsam1, 1, var) ##calculates 1000 vars of 40 obs each
vars2 <- apply(simsam2, 1, var) ##calculates 1000 vars of 200 obs each
ObservedVar1 <- mean(vars1)
ObservedVar2 <- mean(vars2)
TheoreticalVar <- (1/lambda)^2
varCompare <- data.frame(Type = c("Theoretical Variance",
                                   "Observed Variance (40 obs/sample)",
                                   "Observed Variance (200 obs/sample)"),
                          Value = c(TheoreticalVar, ObservedVar1, ObservedVar2))
q <- ggplot(data.frame(vars1), aes(vars1))
q + geom_density(lwd = 2, color = "blue", fill = "sky blue") +
    geom_density(aes(data = data.frame(vars2), x = vars2),
                 lwd = 2, color = "green", fill = "light green", alpha = 0.5) +
    geom_vline(data = varCompare,
    aes(xintercept = Value , linetype = Type, colour = Type), show.legend = TRUE) +
    ggtitle("Distribution of Mean of Variance") +
    labs(x = "Means of Variance", y = "Density")
```



We can see that the distribution of tending towards Gausian but it isn't completely there yet however as our sample sizes increases from 40 to 200 we can see that the distribution tends towards a normal distribution. So it looks like the CLT applies well in our scenario.

From the two, nearly overlapping vertical lines in the above figure, we can also see that the Observed Variances and Theoretical Variance are quite close. We know that the variance for an exponential distribution is $1/\lambda^2$ so for this simulation since our λ is 0.2 our theoretical mean is 25. Below are the actual values for comparison:

```
##                                 Type    Value
## 1                Theoretical Variance 25.00000
## 2  Observed Variance (40 obs/sample) 24.83462
## 3 Observed Variance (200 obs/sample) 24.95489
```

END OF REPORT