Inference Course Project - Part A

Joe Nguyen

22 Nov 2015

In this project, we use simulation to explore inference and do some simple inferential data analysis. The project consists of two parts:

- A simulation exercise
- Basic inferential data analysis

A. Simulation for Distribution of Means of Exponentials

We investigate the exponential distribution and the Central Limit Theorem (CLT). The mean of the exponential distribution is $1/\lambda$ and the standard deviation is also $1/\lambda$. We set $\lambda = 0.2$ for all simulations. We investigate the distribution of averages of 40 exponentials using 1000 simulations.

1. Sample Mean

We show the sample mean (from simulations) and compare it with the theoretical mean of the distribution.

```
set.seed(12345)

# Exponential distribution rate parameter
lambda <- 0.2
nSim <- 1000
n <- 40

# Distribution of averages of n (= 40) exponential distributions
datExpMn <- apply(matrix(rexp(nSim * n, lambda), nSim), 1, mean)

## Histogram, sample mean, theoretical mean
mnSample <- mean(datExpMn)
mnTheory <- 1/lambda

cat("Sample mean: ", mnSample, " vs. Theoretical mean: ", mnTheory, sep = "")
```

```
## Sample mean: 4.971972 vs. Theoretical mean: 5
```

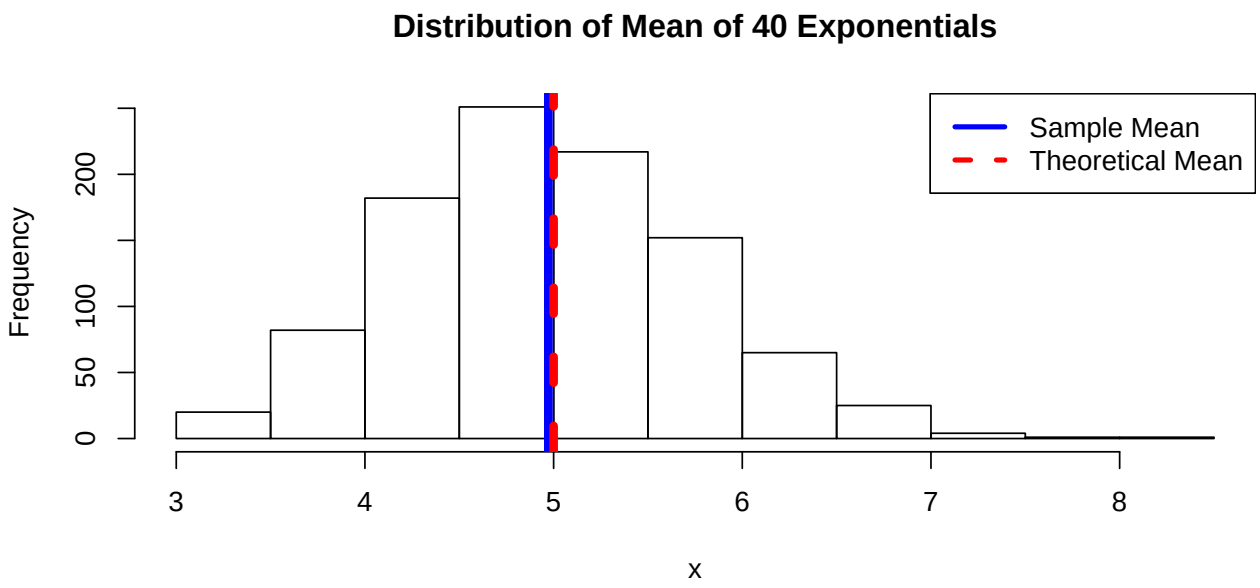

Figure 1. Distribution of mean of 40 exponentials using 1000 simulations.

In Fig. 1, the sample mean (red dashed line) is close to the theoretical mean (blue solid line).

2. Sample Variance

We compare the sample variance to the theoretical variance $(1/\lambda^2)$ by reporting the mean sample variance and theoretical variance, and also plotting the sample variance distribution from the simulations in Fig. 2.

```
## Sample variance: 25.39236 vs. Theoretical variance: 25
```

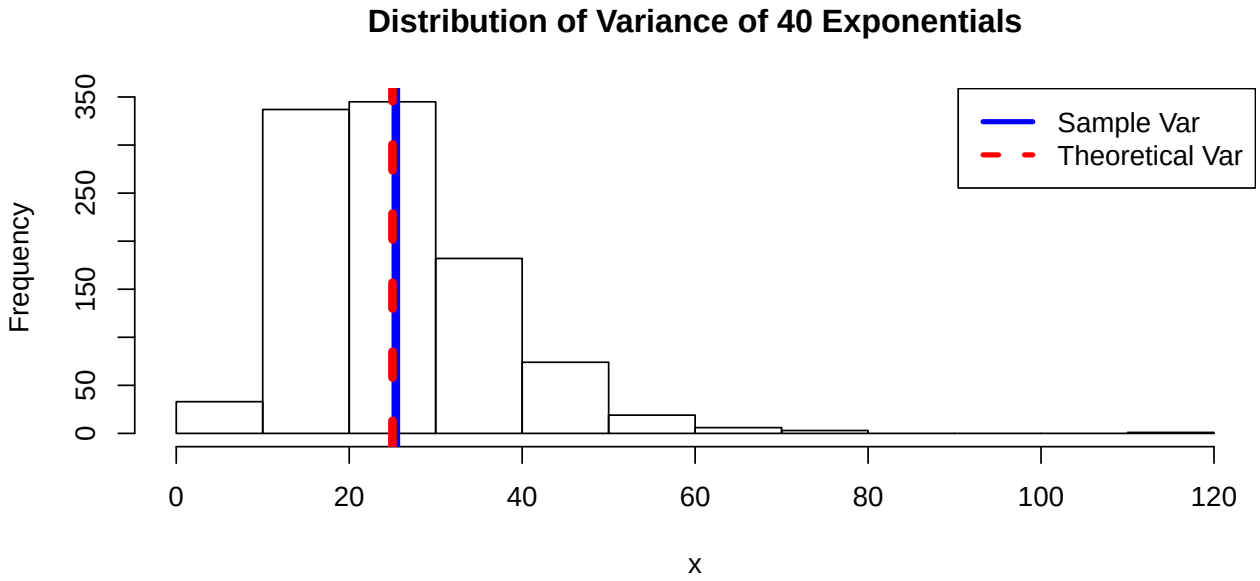

Figure 2. Distribution of variance of 40 exponentials using 1000 simulations.

3. Distribution of Mean of Multiple Draws from Exponential is Approximately Normal

Here, we compare the distribution of 1000 draws of the exponential distribution to the distribution of 1000 "averages of 40 draws from the exponential distribution".

```r
datSample <- data.frame(x = c(rexp(nSim), datExpMn),
                        group = factor(rep(c(1,2), each = nSim),
                        labels = c("Single Draw", "Average of 40 Draws"))
                        )
library(ggplot2)
g <- ggplot(datSample, aes(x)) +
    geom_histogram() +
    facet_wrap(~ group) +
    ggtitle("Exponential Draws Distribution")

# Overlay Normal distribution
x <- mnSample + sd(datExpMn) * seq(-3, 3, length = 1000)
y <- dnorm(x, mnSample, sd(datExpMn))

# Scale density to count of datExpMn
y <- y * max(ggplot_build(g)$data[[1]]$count) / max(y)
datNorm <- data.frame(x = c(x, x),
                      y = c(rep(NaN, length(y)), y),
                      group = factor(rep(c(1,2), each = nSim),
                      labels = c("Single Draw", "Average of 40 Draws"))
                      )
# Plot
g + geom_line(aes(x,y), data = datNorm, col = "red") + facet_wrap(~ group)
```

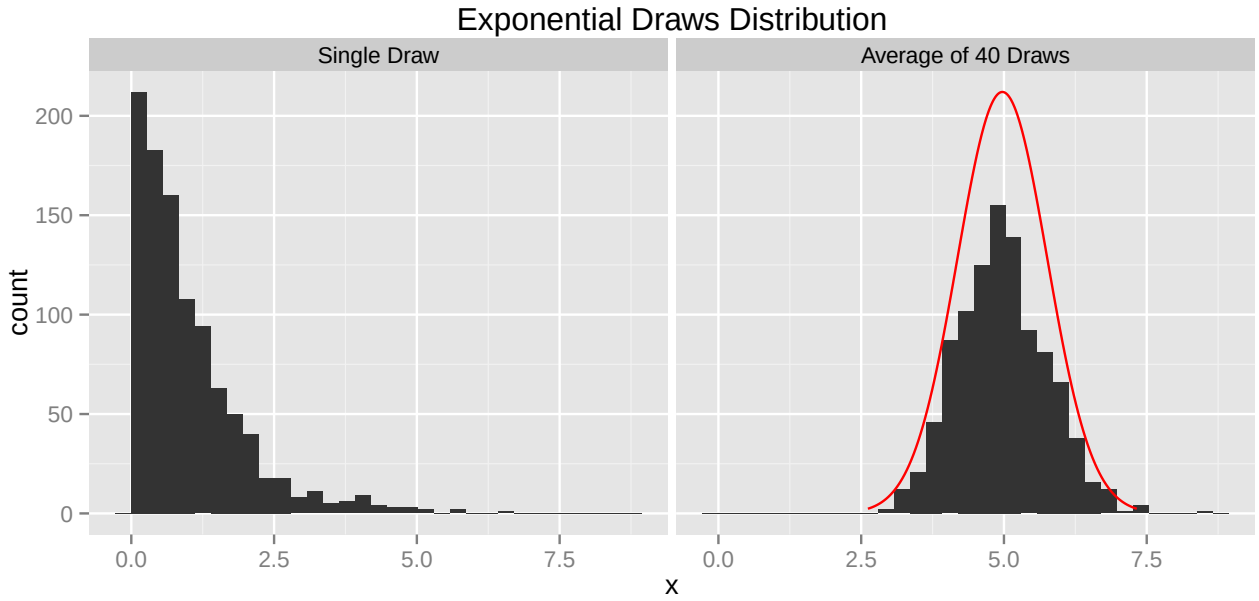

Figure 3. Comparison between sampling distributions with single draws and average of 40 draws from an exponential distribution.

Figure 3 highlights the CLT which states that in the limit as more samples are drawn, the average of these draws follows a Gaussian distribution. The consequence of the CLT is evident in Fig. 3 (right) where the histogram of averages of draws fits in the Gaussian density plot (red line). A beneficial result of the CLT is that in the limit, we can use the sample mean to approximate the mean of the underlying distribution; for the exponential, this is $\frac{1}{\lambda} = \frac{1}{0.2} = 5$ as seen in Fig. 3 (right).