Statistical Inference Project

Farhan Choudhary

10 November 2017

Instructions for the Project

This project consists of two parts and in this simulation we will explore inference and perform some simple inferential statistics. The project is divided into two parts

- A Simulation Exercise
- Basic Inferential Data Analysis

Problem Statement

You will create a report to answer the questions. Given the nature of the series, ideally you'll use knitr to create the reports and convert to a pdf. However, feel free to use whatever software that you would like to create your pdf.

Each pdf report should be no more than 3 pages with 3 pages of supporting appendix material if needed (code, figures, etcetera).

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponential(0.2)s. You should

1. Show where the distribution is centered at and compare it to the theoretical center of the distribution.
2. Show how variable it is and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

A. Simulation of 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.

Note: Set working directory first.

1. Show where the distribution is centered at and compare it to the theoretical center of the distribution - Comparing with mean in Figure 1

```r
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```r
require(knitr)
```

```
## Loading required package: knitr
```

```r
set.seed(123)

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: 5.011911 vs. Theoretical mean: 5
```

Plotting the histogram with the following code:

```r
hist(datExpMn, main = "Distribution of Mean of 40 Exponentials",
     xlab = "x", col = 'bisque'
)
abline(v = mnSample, lty = 1, lwd = 5, col = "blue")
abline(v = mnTheory, lty = 2, lwd = 5, col = "red")

legend("topright", legend = c("Sample Mean", "Theoretical Mean"),
       lty = c(1,2), lwd = 3, col = c("blue", "red"))
```

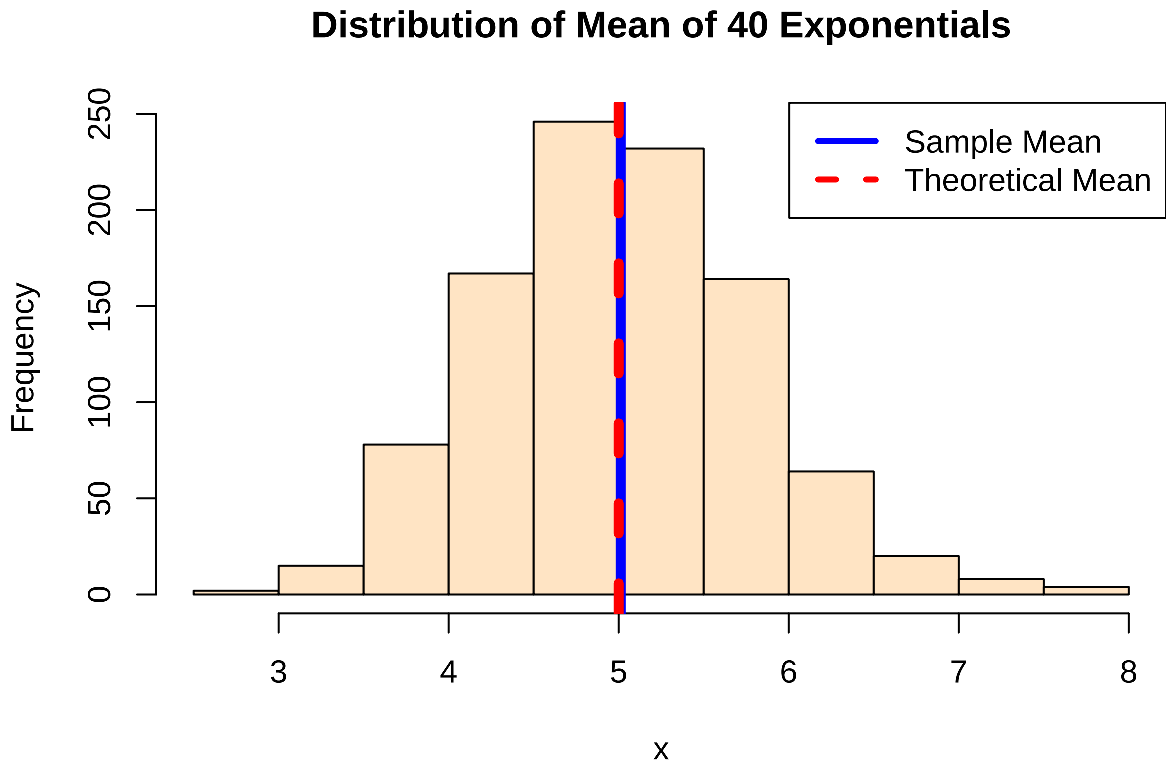

Figure 1: Distribution of mean of 40 exponentials using 1000 simulations.

In the above figure, 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 Figure 2

```
datExpVar <- apply(matrix(rexp(nSim * n, lambda), nSim), 1, var)
varSample <- mean(datExpVar)
varTheory <- lambda ^ -2

cat("Sample variance: ", varSample, " vs. Theoretical variance: ", varTheory, sep = "")
```

```
## Sample variance: 24.79851 vs. Theoretical variance: 25
```

Plotting the Histogram & calculating sample/theoritical variance

```
hist(datExpVar, main = "Distribution of Variance of 40 Exponentials",
     xlab = "x", col = 'darkolivegreen1'
)
abline(v = varSample, lty = 1, lwd = 5, col = "blue")
abline(v = varTheory, lty = 2, lwd = 5, col = "red")

legend("topright", legend = c("Sample Var", "Theoretical Var"),
       lty = c(1,2), lwd = 3, col = c("blue", "red"))
```

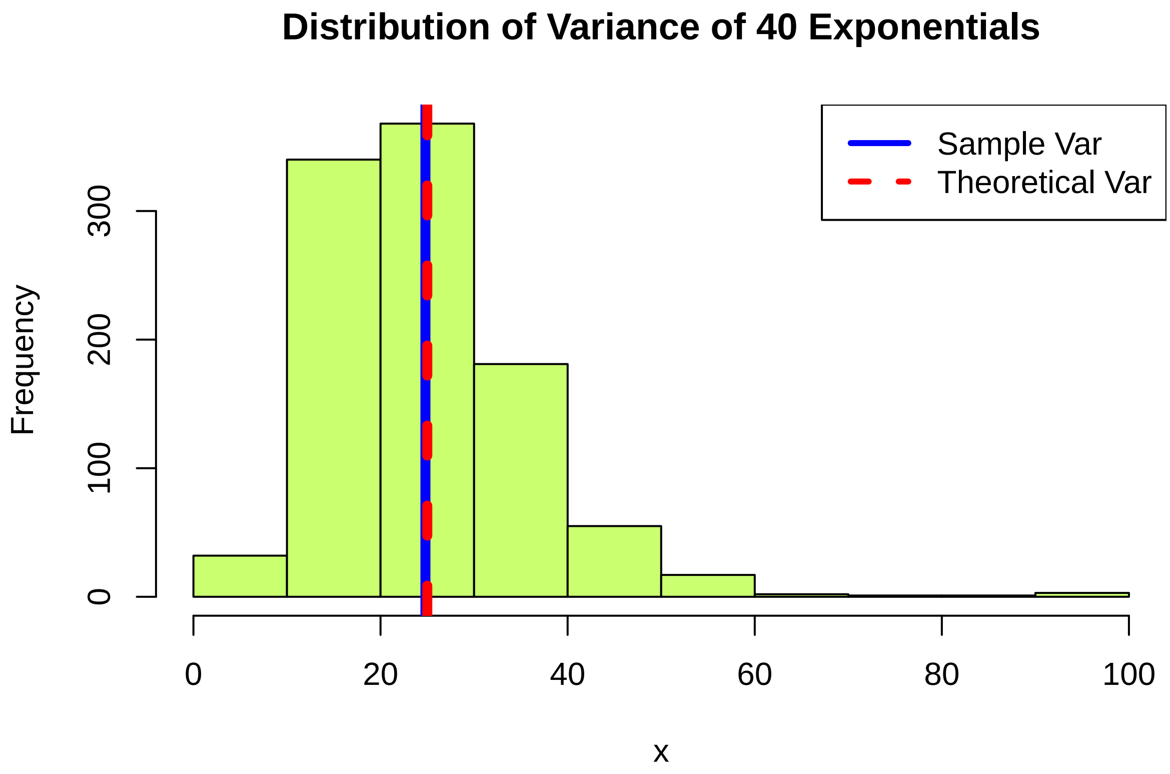

Figure 2: Distribution of 40 exponentials using 1000 simulations

3. Distribution of Mean from multiple exponential draws 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

```
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)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
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)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1000 rows containing missing values (geom_path).
```

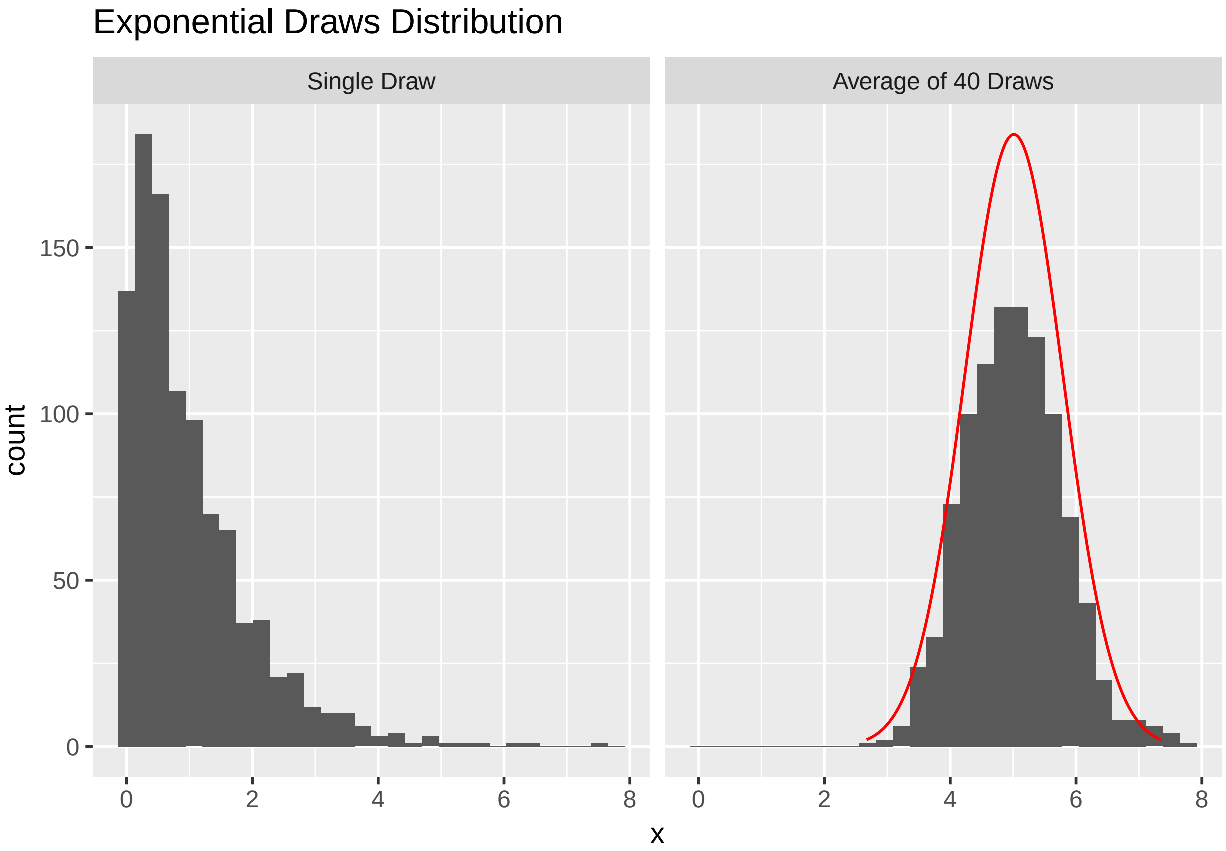

Figure 3: Comparison between sampling distributions with single draws and average of 40 draws from an exponential distribution.

Figure 3 highlights with respect to the Central Limit Theorem (CLT) that as more samples are drawn, the averages follow a Gaussian Distribution. The consequence of the CLT is evident in Figure 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 Figure 3 (right).