Statistical Inference: Course Project Part 1

Anne Strader

28. July 2020

Introduction

In Part 1 of the Coursera Statistical Inference course project, the **exponential distribution** is investigated in R and compared with the Central Limit Theorem. The distribution of **1000 simulated averages of 40 exponentials** is compared to the theoretical mean of the following exponential distribution:

- mean = 1/lambda
- standard deviation = 1/lambda
- lambda = 0.2

The following analysis is performed:

- The sample mean distribution is plotted and compared to the distribution's theoretical mean.
- The variance of the sample mean is calculated and compared to the distribution's theoretical variance.
- The sample mean distribution is shown to be approximately normal.

Sample Mean Distribution Vs. Theoretical Mean

A thousand simulations are performed, where 40 random values from the exponential distribution are selected and averaged:

```
# set seed so that results are reproducible
set.seed(242)
# define lambda value
lambda <- 0.2
# define number of simulations
numSims <- 1000
# define number of random values selected per simulation
sampleSize <- 40
# initialize vector containing sample means
expMeans <- NULL
# calculate sample means
for (i in 1:numSims)
  expMeans <- c(expMeans, mean(rexp(sampleSize, lambda)))
```

The mean of the sample means is compared to the theoretical mean:

```
# average sample mean
avgSampleMean <- mean(expMeans)
avgSampleMean
```

```
## [1] 4.993537
```

```
# theoretical mean
theoreticalMean <- 1.0 / lambda
theoreticalMean
```

```
## [1] 5
```

We observe that the simulated average sample mean is very close to the theoretical mean.

Sample Mean Variance Vs. Theoretical Variance of the Mean

```
# variance of sample mean (through simulation)
varSampleMean <- var(expMeans)
varSampleMean
```

```
## [1] 0.6132664
```

```
# theoretical variance of the sample mean
theoreticalVarSampleMean <- (1.0 / lambda ^ 2.0) / sampleSize
theoreticalVarSampleMean
```

```
## [1] 0.625
```

Again, we observe that the simulated variance of the sample mean is very close to the theoretical sample mean variance.

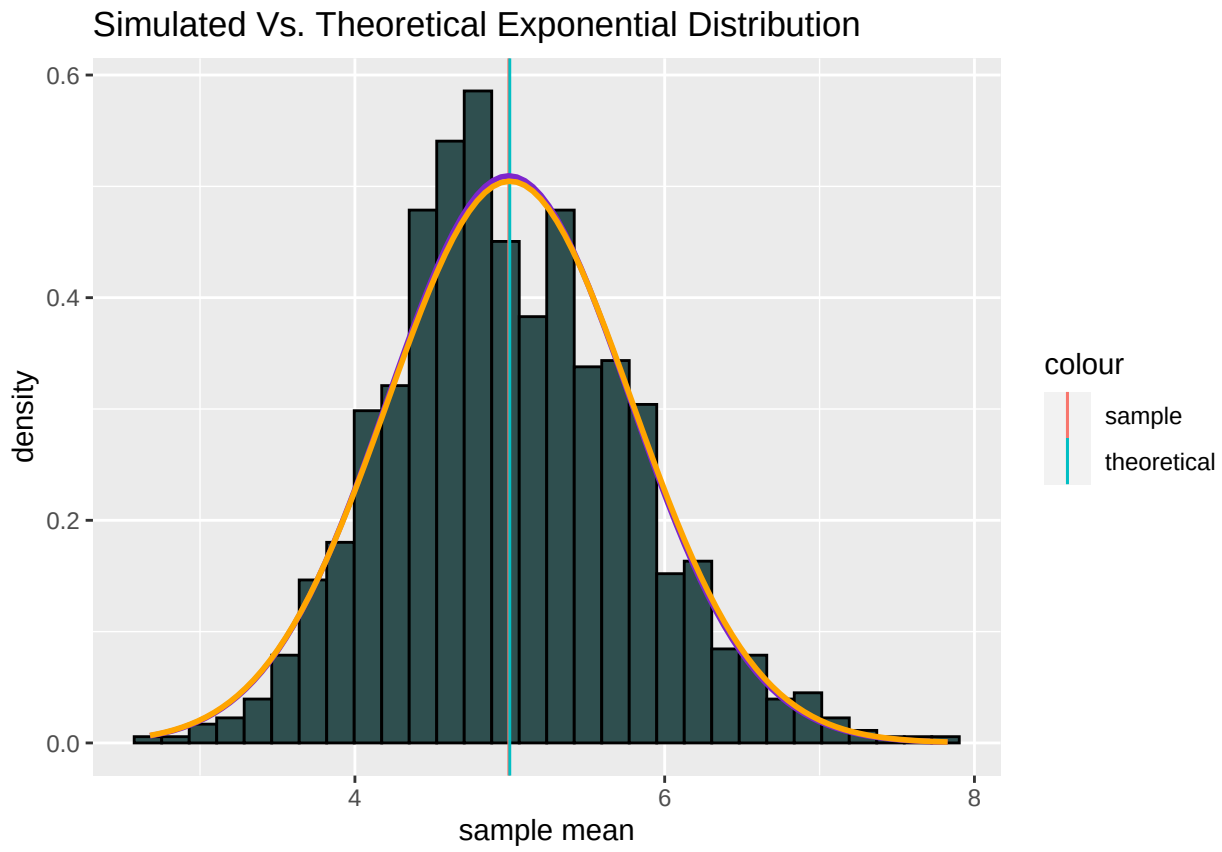
Are the Exponential Distribution Sample Means Normally Distributed?

To qualitatively assess whether the distribution of sample means is approximately normal, the probability density curve generated from the simulated sample mean and variance of the sample mean is compared to the normal density curve generated from the theoretical mean and variance of the mean:

```
library("ggplot2")
# create a dataframe from the sample means
sampleMeanData <- data.frame(expMeans)
# plot distribution of sample means
# initialize plot
g <- ggplot(data=sampleMeanData, aes(x = expMeans))
# generate histogram of sample means showing probability density curve
g <- g + geom_histogram(aes(y = ..density..), color="black", fill = "darkslategrey")
# add vertical lines showing the values of the average sample mean and theoretical mean
g <- g + geom_vline(aes(xintercept = avgSampleMean, color = "sample"))  # average sample mean
g <- g + geom_vline(aes(xintercept = theoreticalMean, color = "theoretical"))  # theoretical mean
# add density curves generated from the simulated and theoretical
# means and variance of the simulated and theoretical means
g <- g + stat_function(fun = dnorm, args = list(mean = avgSampleMean,
                       sd = sqrt(varSampleMean)), color = "purple3", size = 1)
# density curve from sample mean distribution
```

```
g <- g + stat_function(fun = dnorm, args = list(mean = theoreticalMean,
                       sd = sqrt(theoreticalVarSampleMean)), color = "orange", size = 1)
# add axis labels
g <- g + labs(title = "Simulated Vs. Theoretical Exponential Distribution", x = "sample mean",
              y = "density")
# show plot
g
```

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



The red and blue vertical lines show the simulated and theoretical average sample means, respectively. Two normal distributions were generated from the simulated (purple curve) and theoretical (orange curve) means and variance of the means. Both distributions are very similar, and approximately represent the shape of the simulated distribution of sample means.