Statistical Inference Project Part I - Simulation

William Matthews

January 16, 2018

Set working directory

```
setwd("C:/Users/Bill/Documents/Coursera/JohnsHopkins/Course 6 - Statistical Inference/Week 4")
```

Load packages

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

1. Overview:

In this project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The mean of the exponential distribution is 1/lambda and the standard deviation is also 1/lambda. For this project we will set lambda = .2 for all of the simulations. We will investigate the distribution of averages of 40 exponentials over 1000 simulations.

2. Simulations:

We will simulate the averages of 40 exponential distributions 1000 times to get the sample mean. Set lambda = 0.2. We need to run 1000 simulations of size 40 so set n = 1000 * 40 = 40000.

Set the seed to ensure reproducability of the results. Simulate 40000 exponential random variables using

lambda = .2 and store the results in variable sim.

```
set.seed(1)
lambda <- .2
n <- 40 ## number of exponential variables in sample
sims <- 1000 ## number of simulations
sim <- rexp(n = n * sims, rate = lambda)
```

Collect simulations in a matrix with each row representing a sample of size 40.

```
sim_matrix <- matrix(sim, nrow = sims, ncol = n)
```

3. Exploratory Data Analysis:

View the dimensions of the dataset

```
dim(sim_matrix)
```

```
## [1] 1000   40
```

View the structure of the dataset

```
str(sim_matrix)
```

```
##  num [1:1000, 1:40] 3.776 5.908 0.729 0.699 2.18 ...
```

View range of values

```
range(sim_matrix)
```

```
## [1] 1.838128e-04 5.791790e+01
```

Calculate the mean of each sample (row) and assign to variable mean_sim_matrix.

```
mean_sim_matrix <- apply(sim_matrix, 1, mean)
```

Sample Mean versus Theoretical Mean:

Display the mean of the 1000 exponential samples.

```
mean(mean_sim_matrix)
```

```
## [1] 4.990025
```

Calculate the theoretical mean.

```
tmean <- 1/lambda
tmean
```

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

The estimated mean of the 1000 sample exponential means (4.990025) is very close to the theoretical mean of the exponential distribution (1/lambda = 1/.2 = 5). This distribution is centered near the theoretical center of the distribution.

Sample variance versus theoretical variance:

Display the variance of the 1000 exponential samples.

```
var(mean_sim_matrix)
```

```
## [1] 0.6177072
```

Display the standard deviation of the 1000 exponential samples.

```
sd(mean_sim_matrix)
```

```
## [1] 0.7859435
```

Calculate the theoretical variance

```
tvar <- 1/lambda^2/n
tvar
```

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

Calculate the theoretical standard deviation

```
tSD <- 1/(lambda*sqrt(n))
tSD
```

```
## [1] 0.7905694
```

The estimated variance of 0.6177072 is is very close to the theoretical variance of the exponential distribution (1/lambdaˆ2/n = 1/.2ˆ2/40 = .625).

4. Distribution:

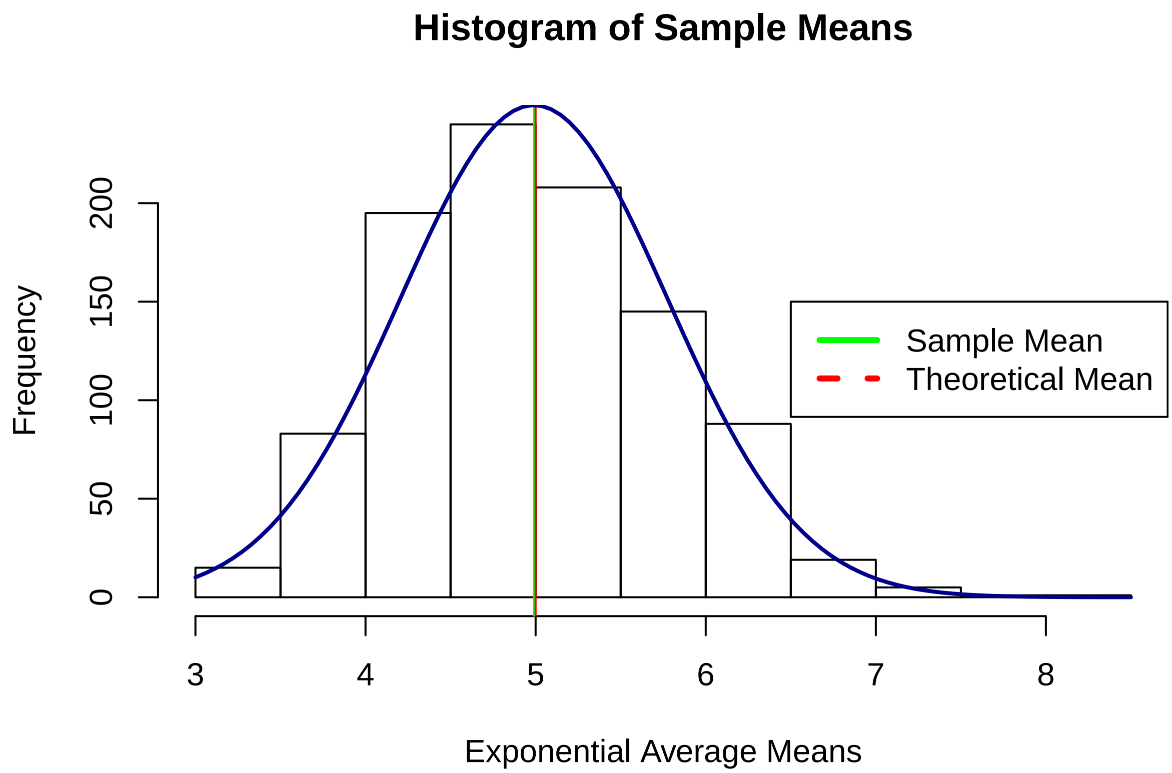
Plot a histogram of the sample average means

```
hist(mean_sim_matrix, xlab = 'Exponential Average Means', main = 'Histogram of Sample Means')

# Add vertical line at the mean of the averages
abline(v = mean(mean_sim_matrix), col = "green")

# Add vertical line at the theoretical mean. Add legend to plot.
abline(v = tmean, col = "red")
legend(6.5,150, c('Sample Mean','Theoretical Mean')
       , lwd = c(3,3), lty = c(1,2), col = c("green","red"))

# Show that the distribution is approximately normal by adding a normal distribution curve using the da
curve(dnorm(x, mean=mean(mean_sim_matrix), sd=sd(mean_sim_matrix))*492, col="darkblue", lwd=2, add=TRUE
```



The sample mean appears to be normally distributed.

Calculate a confidence interval for the simulation.

Calculate the 95% confidence interval for the sample mean.

```
std_err <- sd(mean_sim_matrix)/sqrt(n)
lower <- mean(mean_sim_matrix) - 1.96*std_err
upper <- mean(mean_sim_matrix) + 1.96*std_err
c(lower, upper)
```

```
## [1] 4.746459 5.233592
```

Calculate the 95% confidence interval for the theoretical mean.

```
tstd_err <- tSD/sqrt(n)
tlower <- tmean - 1.96*tstd_err
tupper <- tmean + 1.96*tstd_err
c(tlower, tupper)
```

```
## [1] 4.755 5.245
```

5. Conclusion:

The sample and theoretical confidence intervals are very close.

We can say with 95% confidence that the true mean falls between the lower and upper values of the interval.