Statistical Inference Course Project - Part 1

Aaron Cole

8 February 2016

Contents

1 Synopsis

The following text is quoted from the assignment information page on Coursera.

> In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations. Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials.

- Show the sample mean and compare it to the theoretical mean of the distribution.
- Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
- Show that the distribution is approximately normal.

In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

2 Simulations

2.1 Load libraries

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.3
```

2.2 Setup variables

```
# Number of simulations to perform.
simulations <- 1000

# Number of exponentials to generate.
exponentials <- 40

# Lambda value to use for generation.
lambda <- 0.2
```

2.3 Generate the means

```
# Create a vector to hold the means.
means <- as.numeric()

# Iterate for our number of simulations, setting the seed for each iteration.
for (i in 1:simulations) {
  set.seed(i)
  means <- c(means, mean(rexp(exponentials, lambda)))
}

# Convert to a data.frame.
means <- as.data.frame(means)
```

3 Sample mean versus theoretical mean

3.1 Calculate the sample mean to 3 decimals

```
sampleMean <- round(mean(means$means), 3)
print(sampleMean)
```

```
## [1] 5.002
```

3.2 Calculate the theoretical mean

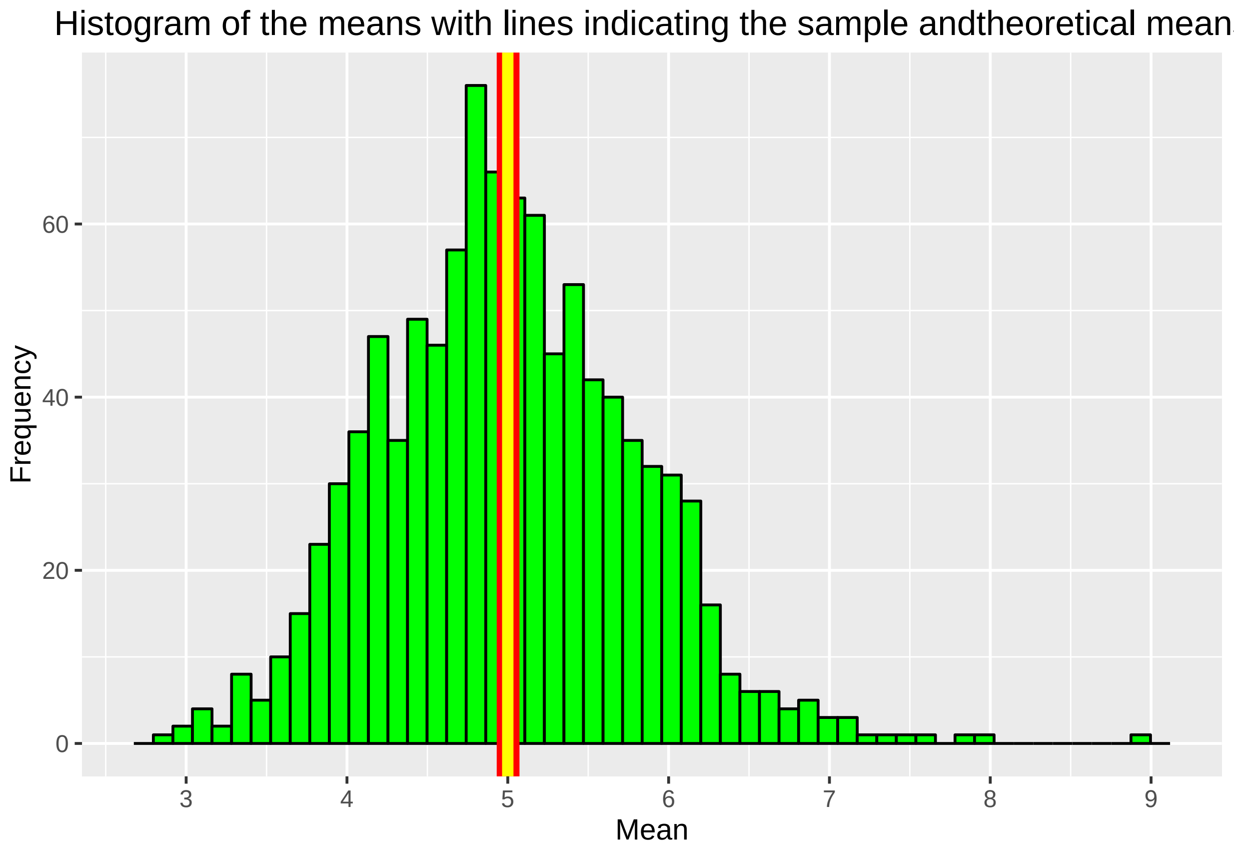
We can calculate the theoretical mean using the formula 1/lambda.

```
theoreticalMean <- 1/lambda
print(theoreticalMean)
```

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

3.3 Plot

```
ggplot(data = means, aes(x = means)) +
  geom_histogram(bins = 50, fill = "green", colour = "black") +
  scale_x_continuous(breaks = 1:10) +
  xlab("Mean") +
  ylab("Frequency") +
  ggtitle(paste("Histogram of the means with lines indicating the sample and",
                "theoretical means", sep = "")) +
  geom_vline(xintercept = sampleMean, colour = "red", size = 4) +
  geom_vline(xintercept = theoreticalMean, colour = "yellow", size = 2)
```



- Red line marks the sample mean value of 5.002.
- Yellow line marks the theoretical mean value of 5.000.

3.4 Summary

- The theoretical and sample mean calcultions both closely coincide, with a difference of only 0.002.

4 Sample variance versus theoretical variance

4.1 Calculate the theoretical variance to 3 decimals

We can calculate the theoretical variance using the formula (1/(lambda^2))/exponentials.

```
theoreticalVariance <- round((1/(lambda^2))/exponentials, 3)
print(theoreticalVariance)
```

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

4.2 Calculate the sample variance to 3 decimals

```r
sampleVariance <- round(var(means$means), 3)
print(sampleVariance)
```

```
## [1] 0.631
```

4.3 Calculate the theoretical standard deviation to 3 decimals

We can calculate the theoretical standard deviation using the formula (1/lambda)/sqrt(exponentials).

```r
theoreticalSD <- round((1/lambda)/sqrt(exponentials), 3)
print(theoreticalSD)
```

```
## [1] 0.791
```

4.4 Calculate the sample standard deviation to 3 decimals

```r
sampleSD <- round(sd(means$means), 3)
print(sampleSD)
```

```
## [1] 0.794
```

4.5 Summary

- The theoretical and sample variance is close, with a difference of only 0.006 (sample 0.631 - theoretical 0.625).
- The theoretical and sample standard deviation is close, with a difference of only 0.003 (sample 0.794 - theoretical 0.791).

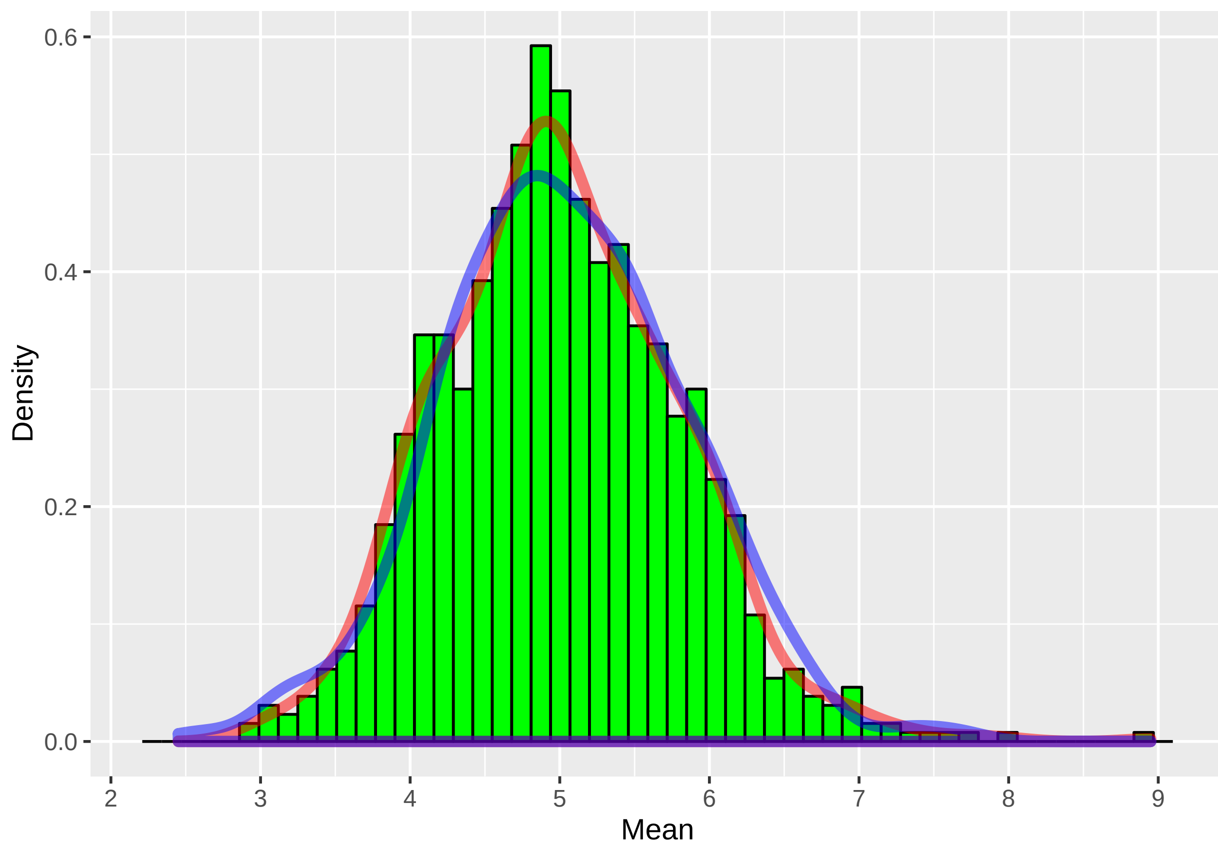
5 Distribution

5.1 Create a normal distribution for comparison

```r
set.seed(1337)
normalDistribution <- data.frame(normalValues = rnorm(simulations,
                                                      mean = sampleMean,
                                                      sd = sampleSD))
```

5.2 Density plot of the means with comparison line for the normal distribution

```r
ggplot(data = means, aes(x = means)) +
  geom_histogram(aes(y=..density..), bins = 50, fill = "green",
                 colour = "black") +
  geom_density(alpha=.5, colour="red", size = 2) +
  geom_density(data = normalDistribution, aes(x = normalValues),
               alpha=.5, colour="blue", size = 2) +
  scale_x_continuous(breaks = 1:10) +
  xlab("Mean") +
  ylab("Density")
```



5.3 Summary

The plot shows that the sample means have a distribution similiar to a normal distribution. This is due to the Central Limit Theorem.