Statistical Inference Course Project 1

K. Divis

Overview

This report uses simulation to investigate the exponential distribution in R and compare it with the Central Limit Theorem. I will compare the mean and variance of the distribution of the mean of 40 exponentials with the theoretical mean and variance, and show that the distribution is apporoximately normal.

This report was completed as part of (and under the guidelines of) the Johns Hopkins Data Science specialization Statistical Inference class on Coursera.

Simulations

First I simulate 1000 sets of expoential distributions with lambda=.2 and n=40.

```
# Set constants
lambda = 0.2
n = 40
numSim = 1000

# Create 'mns' variable to hold each of 1000 simulations of the mean of the
# exponential distribution with n=40, lambda=.2
mns = NULL
for (i in 1:numSim) {
    mns = c(mns, mean(rexp(40, lambda)))
}
```

Now there is simulated data stored in the "mns" variable. I can now use that data to compare the distribution of the mean of the exponential distribution with the Central Limit Theorem.

---

1. Sample Mean versus Theoretical Mean

We know that the theoretical mean of the distribution is 1/lambda. We can use our simulated data to calculate the sample mean.

```
theoryMean = 1/lambda  #Calculate theoretical mean
sampleMean = mean(mns)  #Calculate sample mean from simulated data
diffMean = sampleMean - theoryMean
```

The theoretical mean is **5** and the sample mean is **5.004**. The two means are very close–there is only a difference of **0.004** units between the two means.

See Figure 1 for a histogram and frequency polygon of the simulated exponential distribution of means. The sample mean is plotted as the vertical red bar and the theoretical mean is plotted as the vertical blue bar. As this figure demonstrates, the two bars are very close to one another. This demonstrates that sample mean appears to be an unbiased estimator of the theoretical mean.

---

2. Sample Variance versus Theoretical Variance

We know that the theoretical standard deviation of the distribution is (1/lambda)/sqrt(n). This is the familiar equation for the standard error of the mean. We can use our simulated data to calculate the sample standard deviation. The theoretical and sample variances are just their respective standard deviations squared.

```
meanSD = sd(mns)  #Calculate sample standard deviation from simulated data
theoryMeanSD = (1/lambda)/sqrt(n)  #Calculate theoretical standard deviation
# Convert to variance:
meanVar = meanSD^2
theoryMeanVar = theoryMeanSD^2
sdDiff = meanSD - theoryMeanSD
```

The theoretical standard deviation is **0.791 (variance: 0.625)** and the sample standard deviation is **0.768 (variance: 0.589)**. The two standard deviations are very close–there is only a difference of **-0.023** units between the two standard deviations.

See Figure 2 for a histogram and frequency polygon of the simulated exponential distribution of means. Overlayed in red is a Gaussian distribution using the sample mean and sample standard deviation. Overlayed in blue is a Gaussian distribution using the theoretical mean and theoretical standard deviation. The Gaussian distribution gives a good sense of the "spread" (or variance) of the sample distribution. As you can see, the red and blue lines are very similar, showing the similarity in variance and mean between the sample and theoretical distributions.

Distribution

The Central Limit Theorem tells us that the distribution of averages of iid variables becomes that of a standard normal as sample size increases. That means that the distribution of the averages of the exponential distributions should be normal at a large enough sample size.

See Figure 3. First let's look at the exponential distribution (the first panel). It does not follow a classic "bell-shaped curve" like a normal distribution. Next we can look at the distribution of averages of the exponential distributions and see how it becomes more like a normal distribution with increased sample size (here we used n=5, n=10, and n=40 in the remaining panels).

Figure 1

Code to create Figure 1, followed by Figure 1:

```
library(ggplot2)
library(gridExtra)

h <- ggplot(data = NULL, aes(x = mns)) + geom_histogram(alpha = 0.2, binwidth = 0.3,
    colour = "black", aes(y = ..density..))
h <- h + geom_vline(xintercept = sampleMean, color = "red")
h <- h + geom_vline(xintercept = theoryMean, color = "blue")
h <- h + xlab("Exponential Distribution Means") + ylab("Density") + ggtitle("Histogram of Exp. Dist. Mea
```

```
d <- ggplot(data = NULL, aes(x = mns)) + geom_freqpoly(aes(y = ..density..))
d <- d + geom_vline(xintercept = sampleMean, color = "red")
d <- d + geom_vline(xintercept = theoryMean, color = "blue")
d <- d + xlab("Exponential Distribution Means") + ylab("Density") + ggtitle("Freq. Polygon of Exp. Dist

grid.arrange(h, d, nrow = 1)
```

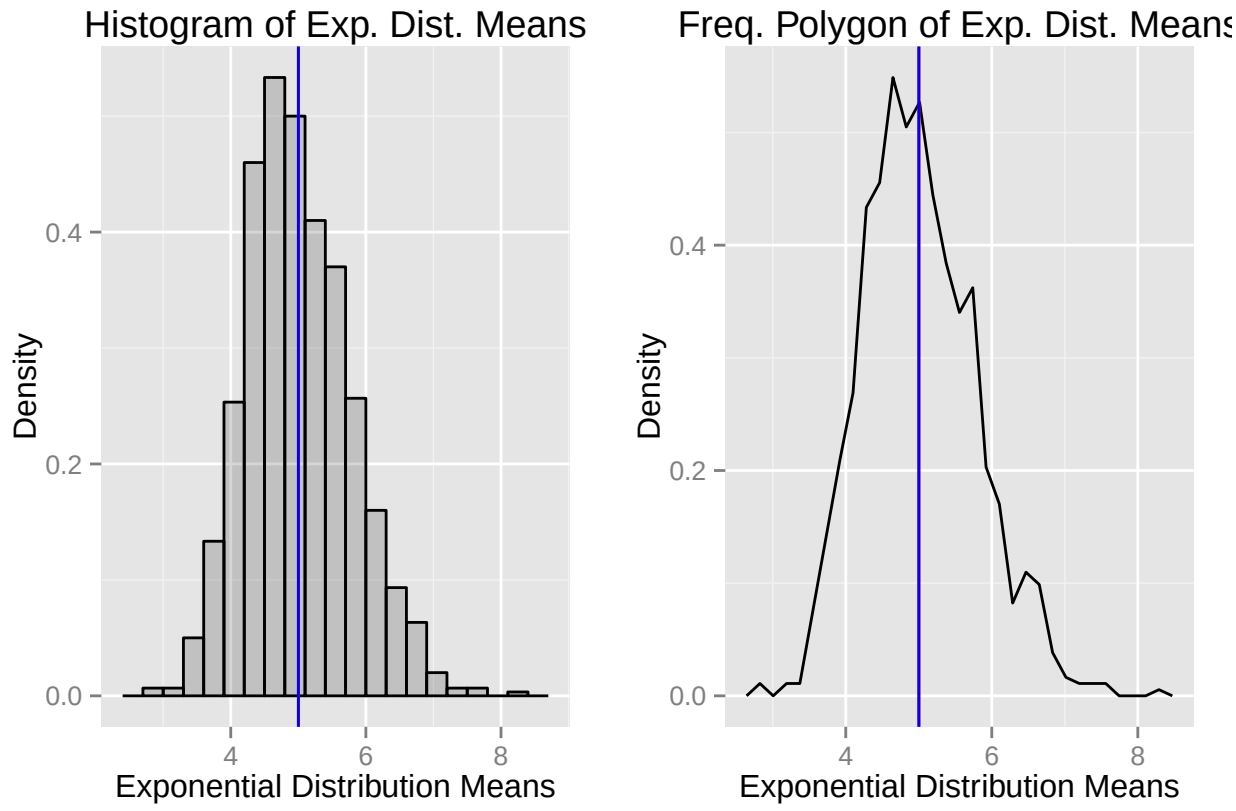


Figure 2

Code to create Figure 2, followed by Figure 2:

```
h2 <- ggplot(data = NULL, aes(x = mns)) + geom_histogram(alpha = 0.2, binwidth = 0.3,
    colour = "black", aes(y = ..density..))
h2 <- h2 + stat_function(fun = dnorm, color = "red", args = list(mean = sampleMean,
    sd = meanSD))
h2 <- h2 + stat_function(fun = dnorm, color = "blue", args = list(mean = theoryMean,
    sd = theoryMeanSD))
h2 <- h2 + xlab("Exponential Distribution Means") + ylab("Density") + ggtitle("Histogram of Exp. Dist.

d2 <- ggplot(data = NULL, aes(x = mns)) + geom_freqpoly(aes(y = ..density..))
d2 <- d2 + stat_function(fun = dnorm, color = "red", args = list(mean = sampleMean,
    sd = meanSD))
```

```
d2 <- d2 + stat_function(fun = dnorm, color = "blue", args = list(mean = theoryMean,
    sd = theoryMeanSD))
d2 <- d2 + xlab("Exponential Distribution Means") + ylab("Density") + ggtitle("Freq. Polygon of Exp. Dis

grid.arrange(h2, d2, nrow = 1)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

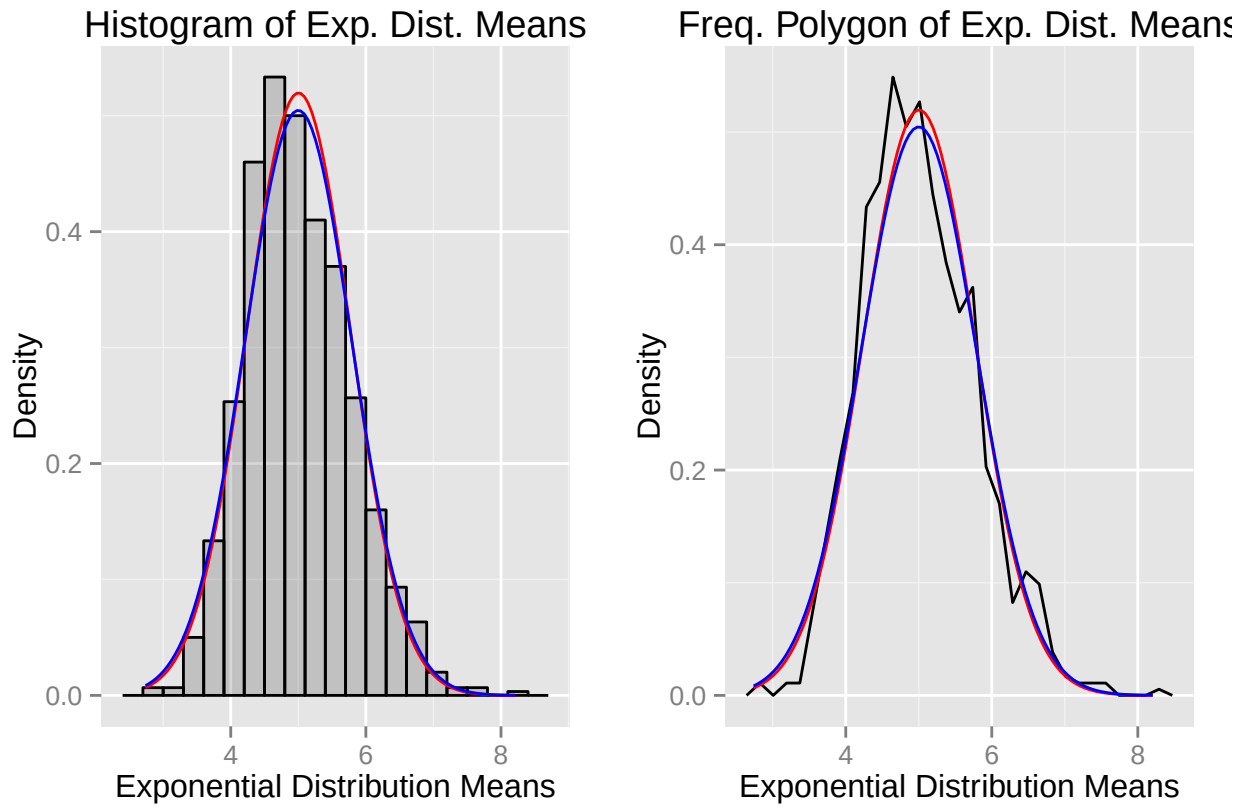


Figure 3

Code to create Figure 3, followed by Figure 3:

```
# Plot exponential function (Equivalent to 'averaging' 1 function)
exp1000 = rexp(1000, lambda)
e <- ggplot(data = NULL, aes(x = exp1000)) + geom_histogram(alpha = 0.2, colour = "black",
    aes(y = ..density..))
e <- e + stat_function(fun = dnorm, args = list(mean = mean(exp1000), sd = sd(exp1000)))
e <- e + xlab("Exponential Distribution") + ylab("Density") + ggtitle("n=1")

# Average 5
mns5 = NULL
```

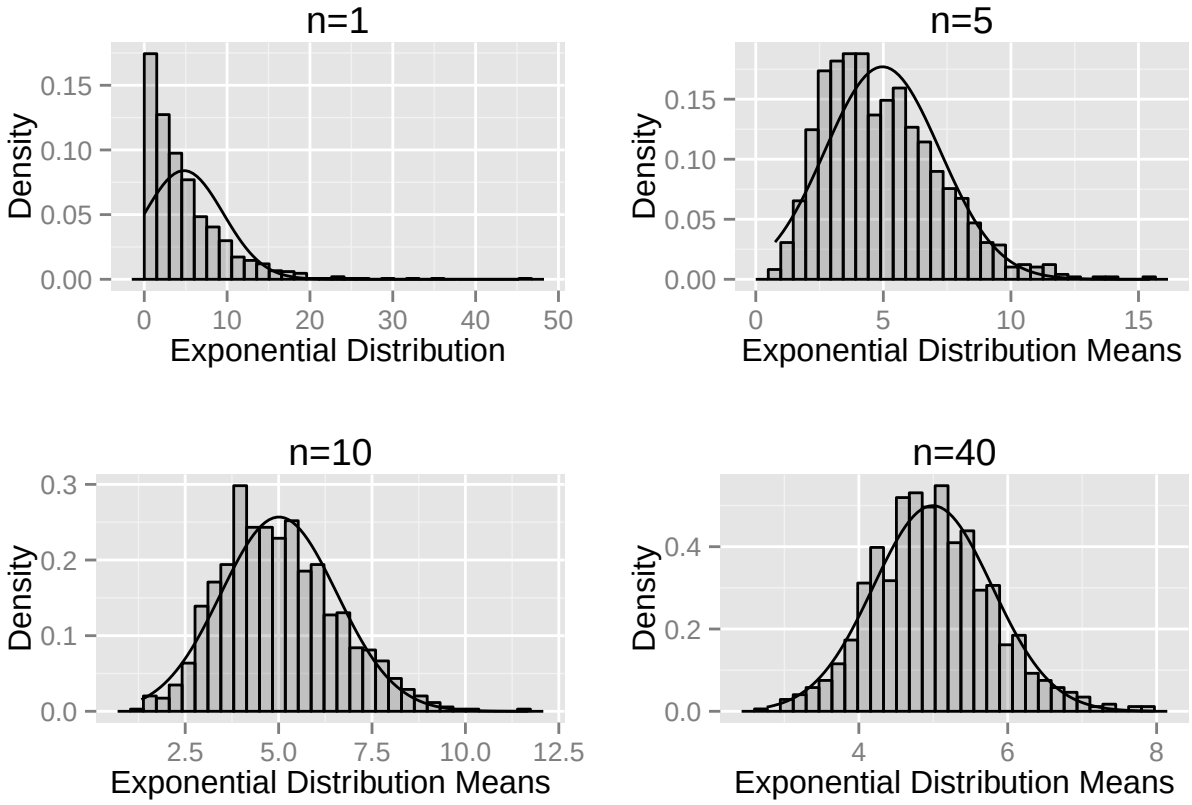
```r
for (i in 1:numSim) {
    mns5 = c(mns5, mean(rexp(5, lambda)))
}
h5 <- ggplot(data = NULL, aes(x = mns5)) + geom_histogram(alpha = 0.2, colour = "black",
    aes(y = ..density..))
h5 <- h5 + stat_function(fun = dnorm, args = list(mean = mean(mns5), sd = sd(mns5)))
h5 <- h5 + xlab("Exponential Distribution Means") + ylab("Density") + ggtitle("n=5")

# Average 10
mns10 = NULL
for (i in 1:numSim) {
    mns10 = c(mns10, mean(rexp(10, lambda)))
}
h10 <- ggplot(data = NULL, aes(x = mns10)) + geom_histogram(alpha = 0.2, colour = "black",
    aes(y = ..density..))
h10 <- h10 + stat_function(fun = dnorm, args = list(mean = mean(mns10), sd = sd(mns10)))
h10 <- h10 + xlab("Exponential Distribution Means") + ylab("Density") + ggtitle("n=10")

# Average 40
mns40 = NULL
for (i in 1:numSim) {
    mns40 = c(mns40, mean(rexp(40, lambda)))
}
h40 <- ggplot(data = NULL, aes(x = mns40)) + geom_histogram(alpha = 0.2, colour = "black",
    aes(y = ..density..))
h40 <- h40 + stat_function(fun = dnorm, args = list(mean = mean(mns40), sd = sd(mns40)))
h40 <- h40 + xlab("Exponential Distribution Means") + ylab("Density") + ggtitle("n=40")

grid.arrange(e, h5, h10, h40, nrow = 2)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```


n=1
Density
0.15
0.10
0.05
0.00
0
10
20
30
40
50
Exponential Distribution
n=5
Density
0.15
0.10
0.05
0.00
0
5
10
15
Exponential Distribution Means
n=10
Density
0.3
0.2
0.1
0.0
2.5
5.0
7.5
10.0
12.5
Exponential Distribution Means
n=40
Density
0.4
0.2
0.0
4
6
8
Exponential Distribution Means