Statistical Inference Course Project Part 1

Scott Lin

Description

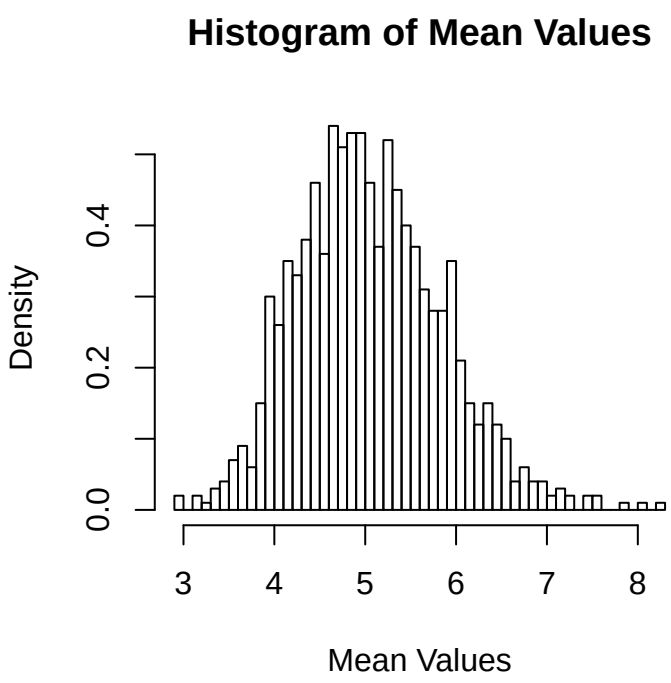
This project is to analyze the exponential distribution and see if random samplings of it adheres to the Central Limit Theorem. The CLT states that the distribution of mean values will approach a normal distribution as the sample size reaches infinity.

Q1: Show the sample mean and compare it to the theoretical mean of the distribution

First, the theoretical mean for an exponential distribution is 1/lambda, as stated in the project description. When lambda is 0.2, the mean is 5.

We'll set up a simulation with a lambda of 0.2 and the sample size of 40. We'll run it 1000

```
nosim = 1000; n = 40; lambda = 0.2; mean_values = NULL
for(i in 1:nosim) mean_values = c(mean_values, mean(rexp(n,lambda)))
hist(mean_values, breaks = 40, main = "Histogram of Mean Values", xlab = "Mean Values", prob=TRUE)
```



We do see that the sample mean distribution is centered around 5 and in fact, the mean of those values is very close to the theoretical mean and will get closer with larger sample sizes.

```
mean(mean_values)
```

```
## [1] 5.060348
```

---

2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution

First, the theoretical variance should be 25 (sd = 1/lambda, lambda = 0.2, sd = 5, var = sd^2, var = 25).

CLT states that the variance of the distribution of averages is the variance of the population divided by sample size. In this case, the sample size is 40, so 25/40 is 0.625.

Let's check the variance of the mean_values from question 1:

```
var(mean_values)
```

```
## [1] 0.633915
```

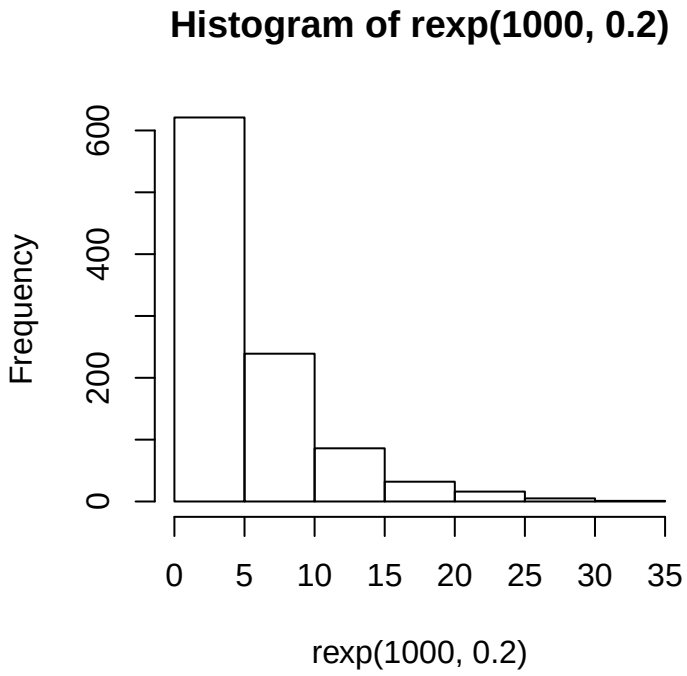
There may be differences in the values and should get closer with larger sample sizes.

---

3. Show that the distribution is approximately normal

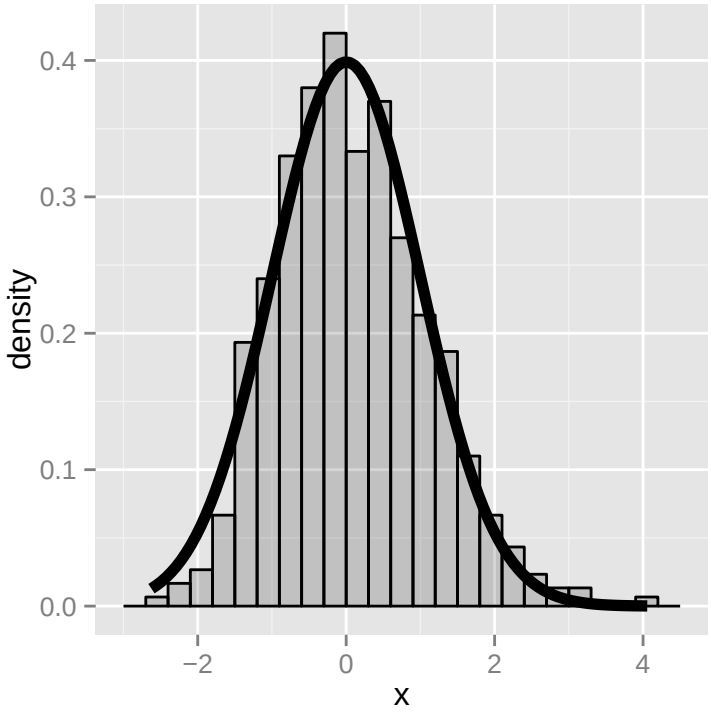
This question wants us to prove the central tenant of the CLT, which is that the distribution of averages is normally distributed as the sample size gets bigger, even if the underlying data is not. Let's check out the underlying exponential distribution.

```
hist(rexp(1000,0.2))
```



Clearly not normally distributed. Let's look at the distribution of the averages again, this time using the formula for the CLT, we converted it the standard normal and overlaid the normal distribution curve over it:

```
library(ggplot2)
normal_mean_values = NULL
for (i in 1:length(mean_values)) normal_mean_values[i] <- sqrt(40) * (mean_values[i]-5)/5
dat <- data.frame(x=normal_mean_values)
g <- ggplot(dat, aes(x = x)) + geom_histogram(alpha = .20, binwidth=.3, colour = "black", aes(y = ..dens
g <- g + stat_function(fun = dnorm, size = 2)
g
```



The eyeball test says that the data looks pretty normally distributed against the normal distribution density curve.

Appendix

Just for fun, let's create one with a sample size of 200 and recheck the mean (should be closer to 5), the variance (with the new sampel size should be close to 25/200 which is 0.125) redraw the normal histogram.

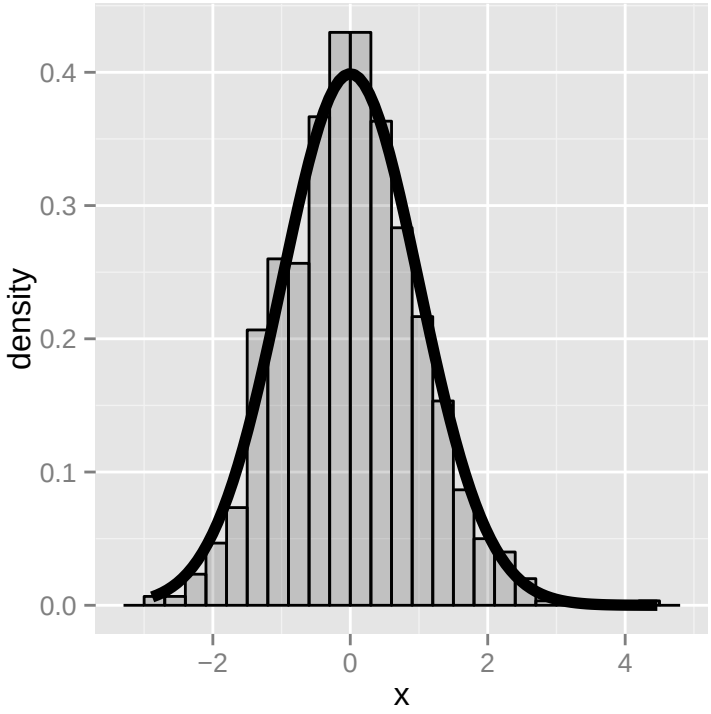
```
library(ggplot2)
new_mean = NULL
nosim = 1000; n = 200; lambda = 0.2
for (i in 1:nosim) new_mean = c(new_mean, mean(rexp(n,lambda)))
mean(new_mean)
```

```
## [1] 5.002315
```

```
var(new_mean)
```

```
## [1] 0.1184851
```

```
normal_mean_values = NULL
for (i in 1:length(new_mean)) normal_mean_values[i] <- (new_mean[i]-5)/(5/sqrt(n))
dat <- data.frame(x=normal_mean_values)
g <- ggplot(dat, aes(x = x)) + geom_histogram(alpha = .20, binwidth=.3, colour = "black", aes(y = ..dens
g <- g + stat_function(fun = dnorm, size = 2)
g
```



Looks pretty good!