A Simulation of the Exponential Distribution and the CLT

Overview

This report is comprised of two parts. In the first part, I investigated the distribution of averages of 40 exponentials by performing 1000 simulations. I compared the simulated sample mean to the theoretical mean(`1 / lambda`), and the simulated sample variance to the theoretical variance (`1 / lambda`). In part two, I briefly analyzed, summarized, and performed hypothesis tests of a dataset related to tooth growth.

Part 1

Goal: create a distribution of averages of exponentials and compare the sample mean and sample variance to the theoretical mean and theoretical variance; then show that the distribution is approximately normal.

Simulations

For the simulation, I created an empty vector, `v`, and filled the vector with 1000 averages of 40 samples of exponentials with a for loop.

```
set.seed(123)

v <- NULL

for (i in 1:1000) {
  v[i] <- mean(rexp(n = 40, rate = 0.2))
}
```

Sample Mean versus Theoretical Mean

The theoretical mean for an exponential distribution is `1 / lambda`. In the above simulation, the `rate` (lambda) was set equal to `0.2`, so the theoretical mean is `1 / 0.2 = 5`. As you can see from the output below, the sample mean and theoretical mean are very similar. The difference between the values is only about ~0.01.

```
(Sample_Mean = mean(v))
```

```
## [1] 5.011911
```

```
(Theoretical_Mean = 1/0.2)
```

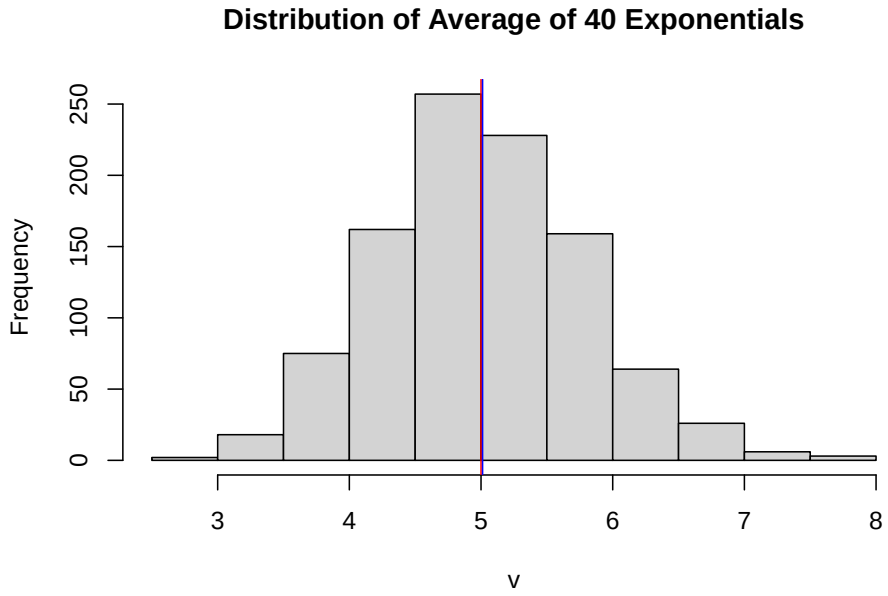
```
## [1] 5
```

```
abs(Sample_Mean - Theoretical_Mean)
```

```
## [1] 0.01191128
```

I plotted a histogram to visually represent the distribution of averages. The theoretical mean is the red vertical line, and the sample mean is the blue vertical line. It's very difficult to distinguish between the lines because of how close the means are.

```r
hist(v, main = "Distribution of Average of 40 Exponentials")
abline(v = 5, col = "red")
abline(v = mean(v), col = "blue")
```



We could eyeball this plot and conclude that the sample mean is not significantly different from the theoretical mean, but aspiring data scientists don't like to just eyeball things. Instead, let's perform a One Sample t-test, with an alpha level set at `p = 0.05`.

```r
t.test(v, mu = 5, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  v
## t = 0.48608, df = 999, p-value = 0.627
## alternative hypothesis: true mean is not equal to 5
## 95 percent confidence interval:
##  4.963824 5.059998
## sample estimates:
## mean of x
##  5.011911
```

The sample mean of 1000 simulations of the average of 40 exponentials is not significantly different from the theoretical mean.

Sample Variance versus Theoretical Variance

The sample variance is slightly smaller that the theoretical variance which is often the case.

```r
(Sample_Variance = var(v))
```

```
## [1] 0.6004928
```

```r
(Theoretical_Variance = (1/0.2^2)/40)
```

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

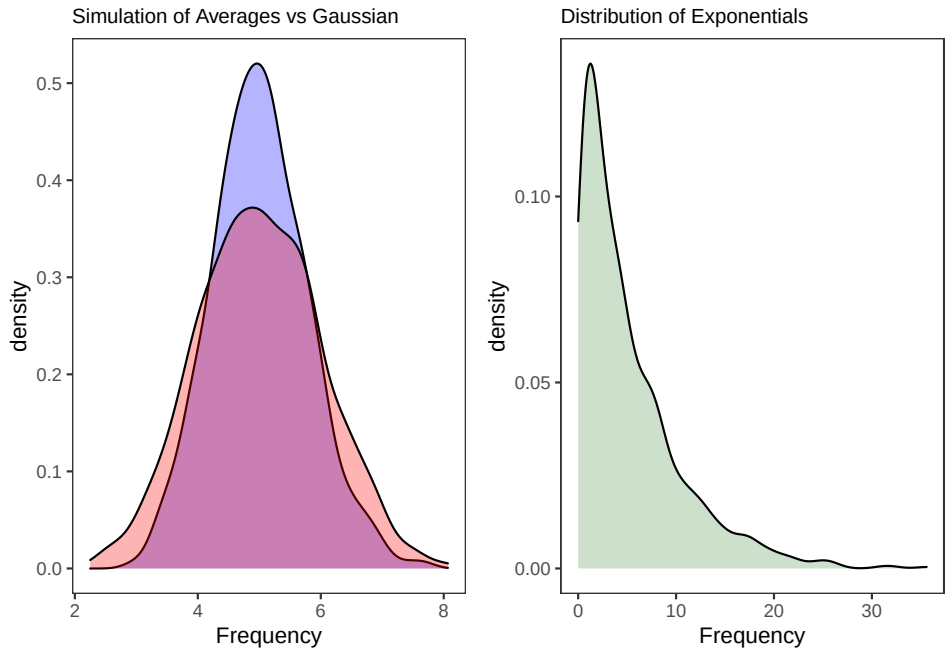
Distributions

```
library(ggplot2)
library(gridExtra)

p1 <-
ggplot(mapping = aes(v)) +
  geom_density(data = as.data.frame(v),
               fill = "blue",
               alpha = .3) +
  geom_density(data = data.frame(v = rnorm(1000, 5)),
               fill = "red",
               alpha = .3) +
  theme_bw() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        plot.title = element_text(size = 10)) +
  labs(title = "Simulation of Averages vs Gaussian") +
  xlab("Frequency")

p2 <-
ggplot(mapping = aes(V1)) +
  geom_density(data = data.frame(V1 = rexp(n = 1000, rate = 0.2)),
               fill = "darkgreen",
               alpha = .2) +
  theme_bw() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        plot.title = element_text(size = 10)) +
  labs(title = "Distribution of Exponentials") +
  xlab("Frequency")

grid.arrange(p1, p2, ncol = 2)
```



Above is a visualization showing that the simulation distribution is relatively normal. The blue histogram on the left is the simulated distribution, and the red (salmon) histogram on the left is a normal distribution (with mean = 5) of 1000 samples created with the `rnorm()` function. The purple shaded area is the overlap between the two distributions. You could perform a test of normality to assess if the simulated distribution is actually normal, but that is beyond the scope of this project. The green histogram on the right is 1000 samples of exponentials; I created this plot to show the stark contrast between samples of exponentials and samples of averages of exponentials. It's very clear that the sample of averages is fairly normal, and the sample of exponentials is not at all normal!