Exponential Distribution Project and Tooth Growth Analysis

Raymond Dineen

9/21/2020

Part 1:

Overview

In this first section, we will do a bit of exploration of the exponential distribution. We will run simulations of sampling from the exponential distribution and show how the results of these simulations demonstrate the Central Limit Theorem.

Simulations

The first thing we need to do to is to simulate some random data from the exponential distribution. We will create 1000 samples of 40 random exponentials using the rate of $\lambda = 0.2$

```r
B <- 1000
n <- 40
lambda <- 0.2
set.seed(1)
# Creating a matrix of 1000 rows of 40 random exponentials
sim <- matrix(rexp(n * B, rate = lambda), B, n)
```

Sample Mean versus Theoretical Mean

The theoretical mean of the exponential distribution is $\frac{1}{\lambda}$. In our case, our theoretical mean is $\frac{1}{(1/5)} = 5$ and we expect the distribution of our sampled means to be centered around this value. We also expect the mean of our sampled means to be very close to our theoretical mean. Figure 1 shows both of our expectations to be true.

Sample Variance versus Theoretical Variance

The sample variance of our observed means can be calculated with $S^2 = \frac{\sum_{i=1}(X_i - \bar{X})^2}{n-1}$. The `var()` function in R does this for us. Our theoretical variance is σ^2/n which in our case is $5^2/1000 = 0.625$. We expect our sample variance and theoretical variance to be close to each other due to the the high amount of simulations we did. Our expectations are met as shown in the code below.

```r
theoVar <- 0.625
sampleVar <- var(rowMeans)
sampleVar
```

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

Distribution

The Central Limit Theorem tells us that the the distribution of our properly normalized sample means becomes standard normal as the sample size increases. To show this we first must normalize our data (subtract the mean divide by the standard deviation). The `scale()` function in R does this for us. We then plot the the normalized data against a normal distribution curve, as shown in figure 2, and notice how similar the two distributions are. It is also worth noticing how different the distribution of 1000 random exponentials is compared to the distribution of 1000 means of 40 random exponentials, also shown in figure 2. Even though the means are taken the from exponential an exponentially distributed population, they follow a normal distribution due to the Central Limit Theorem.

Part 2:

Overview

The ToothGrowth dataset in R shows data from a test that measured the growth of teeth in guinea pigs after being given varying doses of vitamin C (0,5, 1, and 2 mg/day) by two different delivery methods, orange juice (OJ) and ascorbic acid (VC). We will be looking at whether dosage and delivery method affect tooth growth. We can start with a simple boxplot of data, shown in figure 3, looking at length based on both delivery method and dose. We can see a general trend but not anything strong enough to come to any immediate conclusions.

Does delivery method effect tooth growth?

A good place to start is to see if delivery method effects tooth growth regardless of dose. We can use a t-test to get a p-value. Our null hypothesis is that there is no difference between the two delivery methods. Our alternative hypothesis will be that the OJ results in greater length than VC because of the trend we observed in the boxplot we just made.

```r
data("ToothGrowth")
grpOJ <- ToothGrowth$len[ToothGrowth$supp == "OJ"]
grpVC <- ToothGrowth$len[ToothGrowth$supp == "VC"]
t.test(grpOJ, grpVC, alt = "greater")$p.value
```

```
## [1] 0.03031725
```

Our pvalue is less than 0.05 which leads us to reject the null hypothesis. We can also back up our conclusion with a permutation test which randomly reassigns delivery methods to lengths and see how many end up giving us a more extreme result than the actual result.

```r
lengths <- ToothGrowth$len
supps <- ToothGrowth$supp
testStat <- function(len, method) mean(len[method == "OJ"]) - mean(len[method == "VC"])
observedStat <- testStat(lengths, supps)
set.seed(1)
permutations <- sapply(1:10000, function(i) testStat(lengths, sample(supps)))
mean(permutations > observedStat)
```

```
## [1] 0.0311
```

As we can see from the test, roughly 3% of permutations resulted in a higher difference in means than the observed permutation. This result is very similar to our p-value and further leads us to reject the null hypothesis.

Does dose affect tooth growth?

Our boxplot leads us to believe that higher doses results in greater tooth length. We have more than two groups to compare so we will have to do a few different t-tests to try to get the full picture. We can start with the what seems to have the most obvious conclusion. We test the alternative hypothesis that 2 mg/day results in greater growth than 0.5 mg/day.

```
grp2 <- ToothGrowth$len[ToothGrowth$dose == 2]
grp0.5 <- ToothGrowth$len[ToothGrowth$dose == 0.5]
t.test(grp2, grp0.5, alt = "greater")$p.value
```

```
## [1] 2.198762e-14
```

The p-value for this test is extremely low and we can confidently reject the null hypothesis. Next we want to check if 2 mg/day results in more growth than 1 mg/day.

```
grp1  <- ToothGrowth$len[ToothGrowth$dose == 1]
t.test(grp2, grp1, alt = "greater")$p.value
```

```
## [1] 9.532148e-06
```

Again we see another extremely low p-value which leads us to reject the null hypothesis.

Conclusions

From our analysis we can confidently say that higher doses of vitamin C result in greater tooth growth in guinea pigs tested. We are also fairly confident from our t-test and permutation test that orange juice was a better delivery method than ascorbic acid. We assumed that the data was t-distributed within each unique combination of dose and delivery method.

Appendix

Figure 1:

```
library(dplyr)
library(ggplot2)
data.frame(rowMeans = rowMeans) %>% ggplot(aes(x = rowMeans)) +
    geom_density(size = 2, alpha = 0.2, fill = "blue", size = 2) +
    geom_vline(xintercept = theoMean, color = "red", size = 1.5) +
    geom_vline(xintercept = sampleMean, color = "green", size = 1.5) +
    annotate("text", 6, 0.45, label = "Theoretical Mean = 5", color = "red") +
    annotate("text", 3.8, 0.45, label = "Mean of Sample Means = 4.99", color = "green") +
    theme_bw()
```

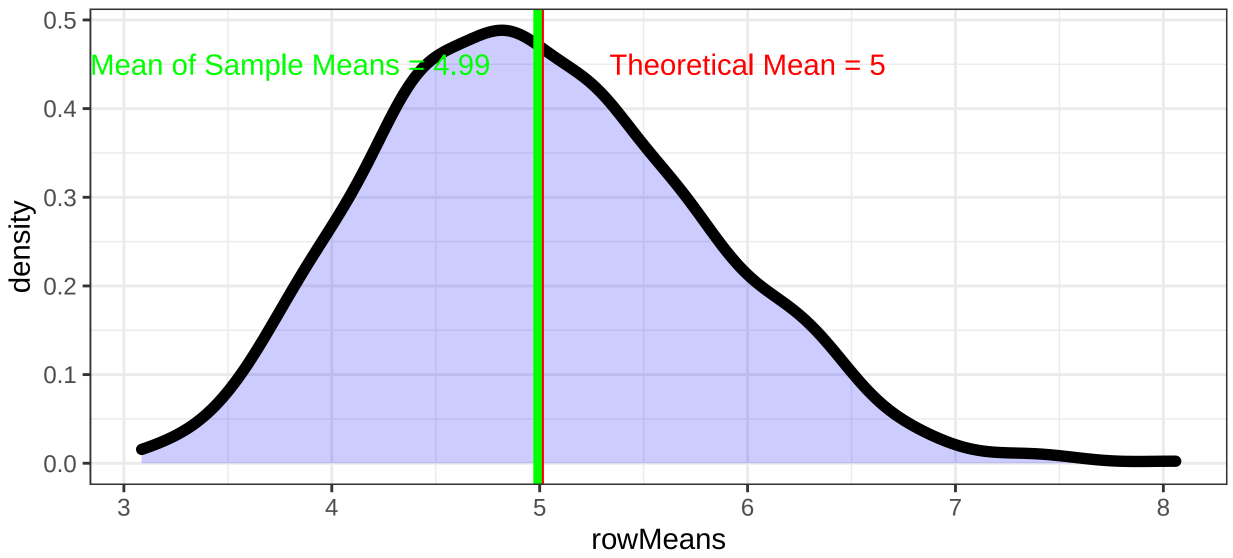


Figure 2:

```
library(ggpubr)
normalized <- scale(rowMeans)
cltdemo <- data.frame(normalized = normalized) %>% ggplot(aes(normalized)) +
    geom_density(size = 2, alpha = 0.2, fill = "blue") +
    stat_function(fun = dnorm, n = 1000, color = "red", size = 1.5) +
    annotate("text", 2.5, 0.35, label = "Normal Distribution", color = "red", size = 3) +
    labs(x = "distribution", title = "Normalized means vs. Normal Distribution") +
    theme_bw() +
    theme(plot.title = element_text(size = 10))

rexpdemo <- data.frame(rexps = rexp(1000, rate = 0.2)) %>% ggplot(aes(rexps)) +
    geom_density(size = 2, fill = "blue", alpha = 0.2) +
    labs(x = "", title = "1000 Random Exponentials") +
    theme_bw()

ggarrange(cltdemo, rexpdemo, nrow = 1)
```

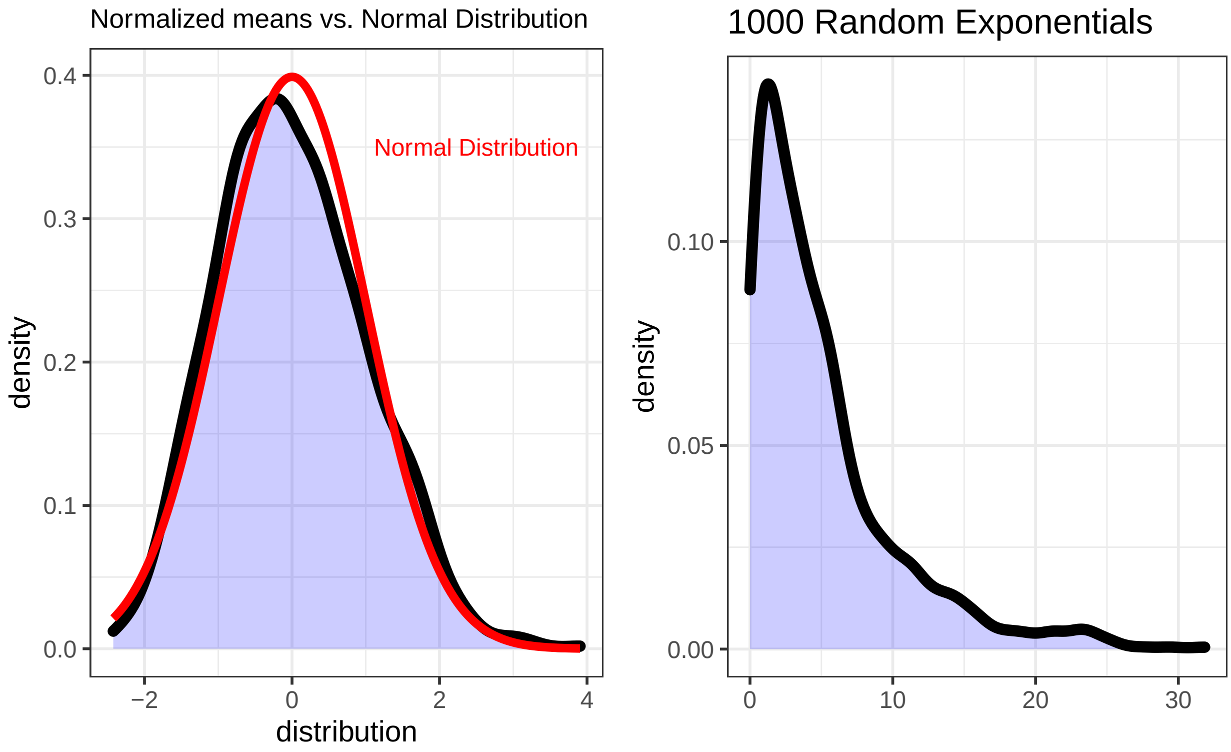
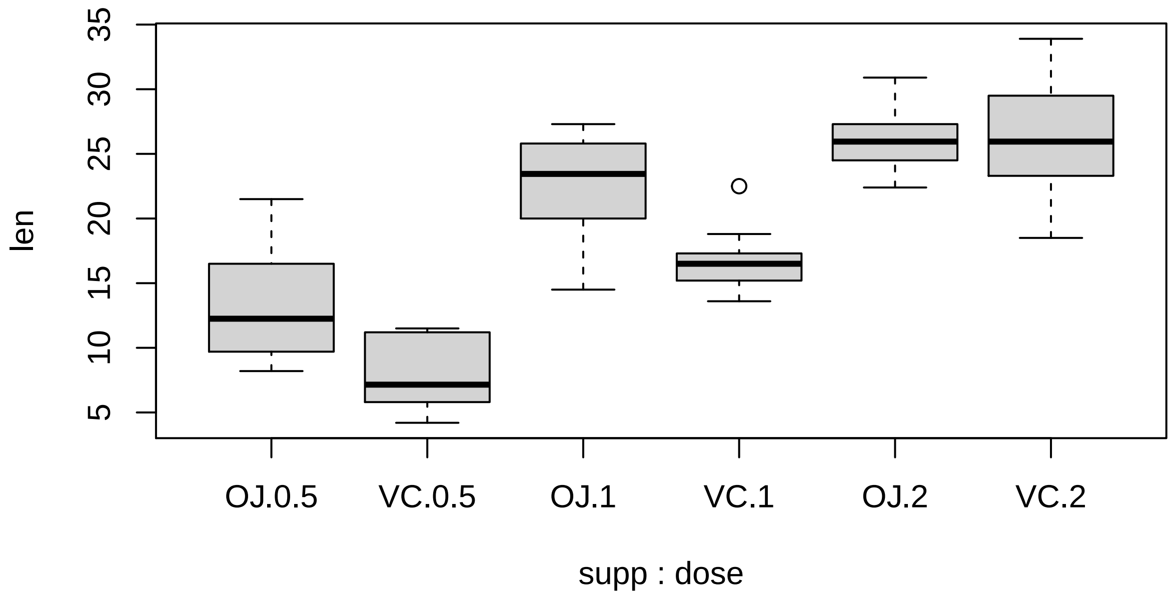


Figure 3:

```
data("ToothGrowth")
boxplot(len ~ supp + dose, data = ToothGrowth)
```


len
OJ.0.5
VC.0.5
OJ.1
VC.1
OJ.2
VC.2
supp : dose