Statistical Inference Course Project

Jack Kramer

March 27, 2017

Part 1

Simulation Exercise

In this project, I investigate the Exponential Distribution and compare it with the Central Limit Theorem.

```
lambda = 0.2 ##lambda, the rate parameter, explains both the mean (1/lambda) and the standard deviation
n = 40 ##n is the number of observations we will observe
simulations = 10000 ##simulations is how many simulations we will run on n observations.


simulation.data.frame <- data.frame()

for(i in 1:simulations){
    rexpdata <-rexp(n,lambda)
    simulation.data.frame[i,1] <- mean(rexpdata)
    simulation.data.frame[i,2] <- sd(rexpdata)
    i=i+1
}
```

We should expect to see means and standard deviations around 5 based on our Lambda = 0.2.

```
head(simulation.data.frame)
```

```
##       Mean Standard Deviation
## 1 5.936468           6.500642
## 2 4.112516           4.424011
## 3 4.569992           4.101809
## 4 3.876391           2.975845
## 5 5.213575           5.197311
## 6 4.409024           5.372099
```

Above shows how that appears to be the case.

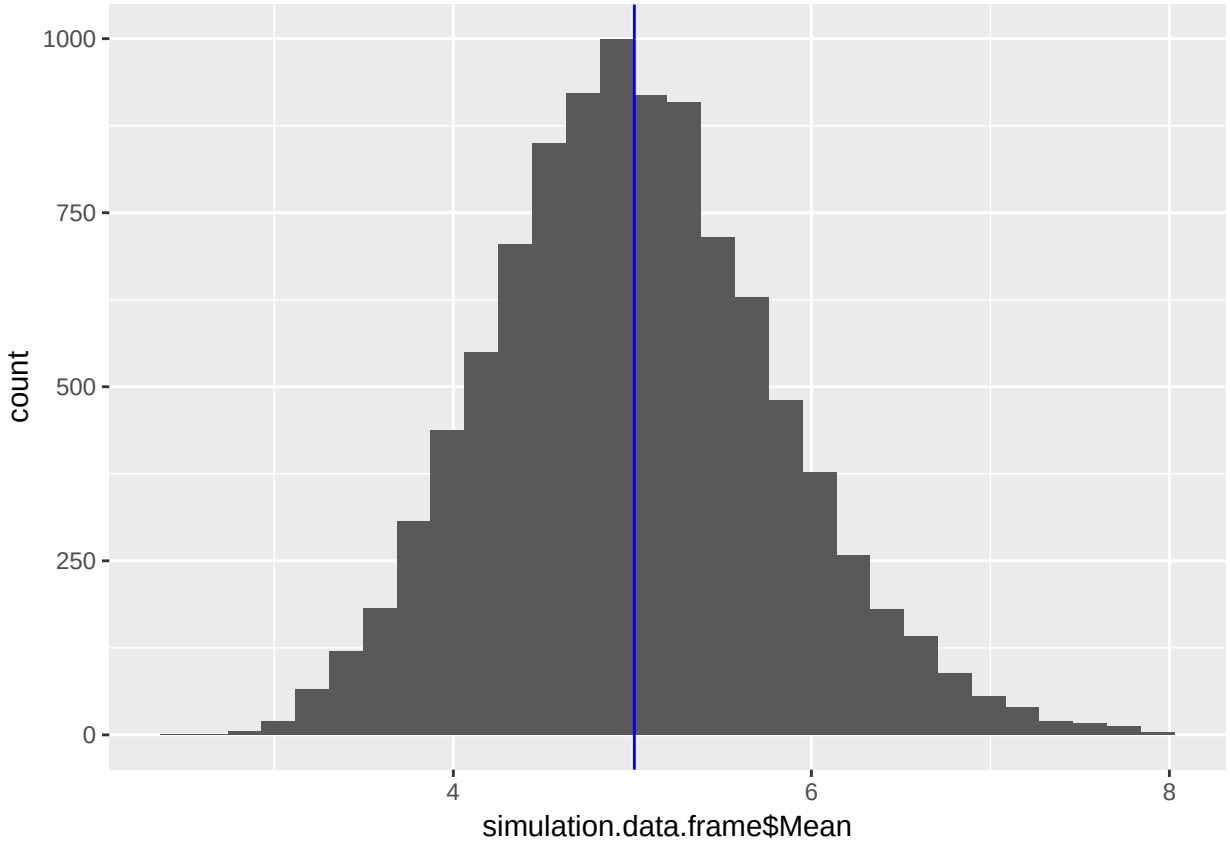
1) Show the sample mean and compare it to the theoretical mean of the distribution.

```
xbar <- mean(simulation.data.frame$Mean) ##sample mean of all our simulations.
mu <- 1/lambda ##Theoretical mean as defined in the Exponential Distribution.

writeLines(paste("Sample mean: \t\t", xbar,"\nTheoretical mean:\t", mu))
```

```
## Sample mean:          5.01109024467169
## Theoretical mean:     5
```

```
ggplot(data = simulation.data.frame, aes(simulation.data.frame$Mean)) +
    geom_histogram(bins = 30) +
    geom_vline(xintercept = xbar, colour = "blue")
```



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

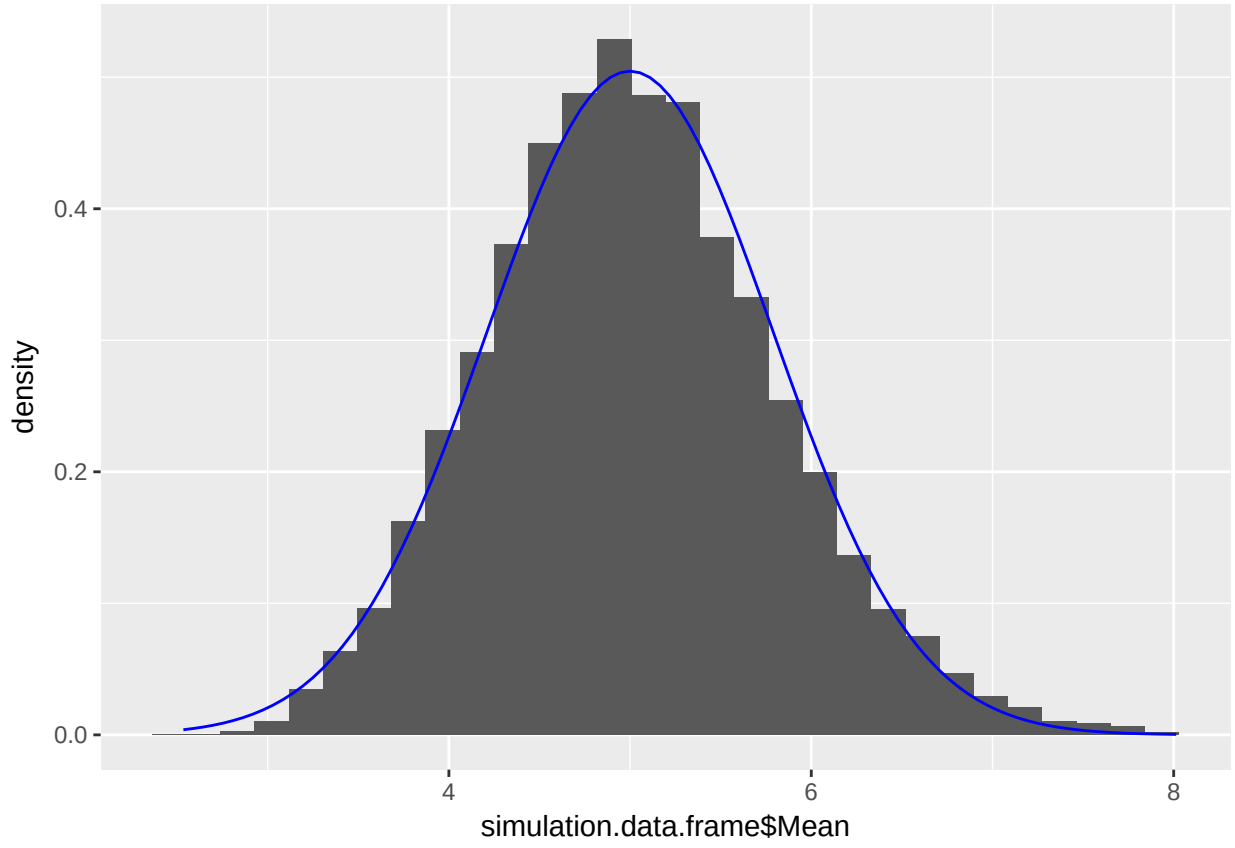
```
s <- sd(simulation.data.frame$Mean) ##sample standard deviation of our means.
sigma <- (1/lambda)/sqrt(n) ##theoretical standard deviation as defined in the Exponential Distribution


writeLines(paste("Sample std. dev:\t\t\t", s,"\nTheoretical std. dev:\t", sigma))
```

```
## Sample std. dev:          0.792361235358986
## Theoretical std. dev:     0.790569415042095
```

3)Show that the distribution is approximately normal

```
ggplot(data = simulation.data.frame, aes(simulation.data.frame$Mean)) +
    geom_histogram(aes(y = ..density..),bins = 30) +
    stat_function(fun = dnorm, colour = "blue", args = list(mean=mu, sd=sigma))
```



Part 2

Basic Inferential Data Analysis

We will be examining the ToothGrowth data set that is included in the R Datasets package

Summary of the Data

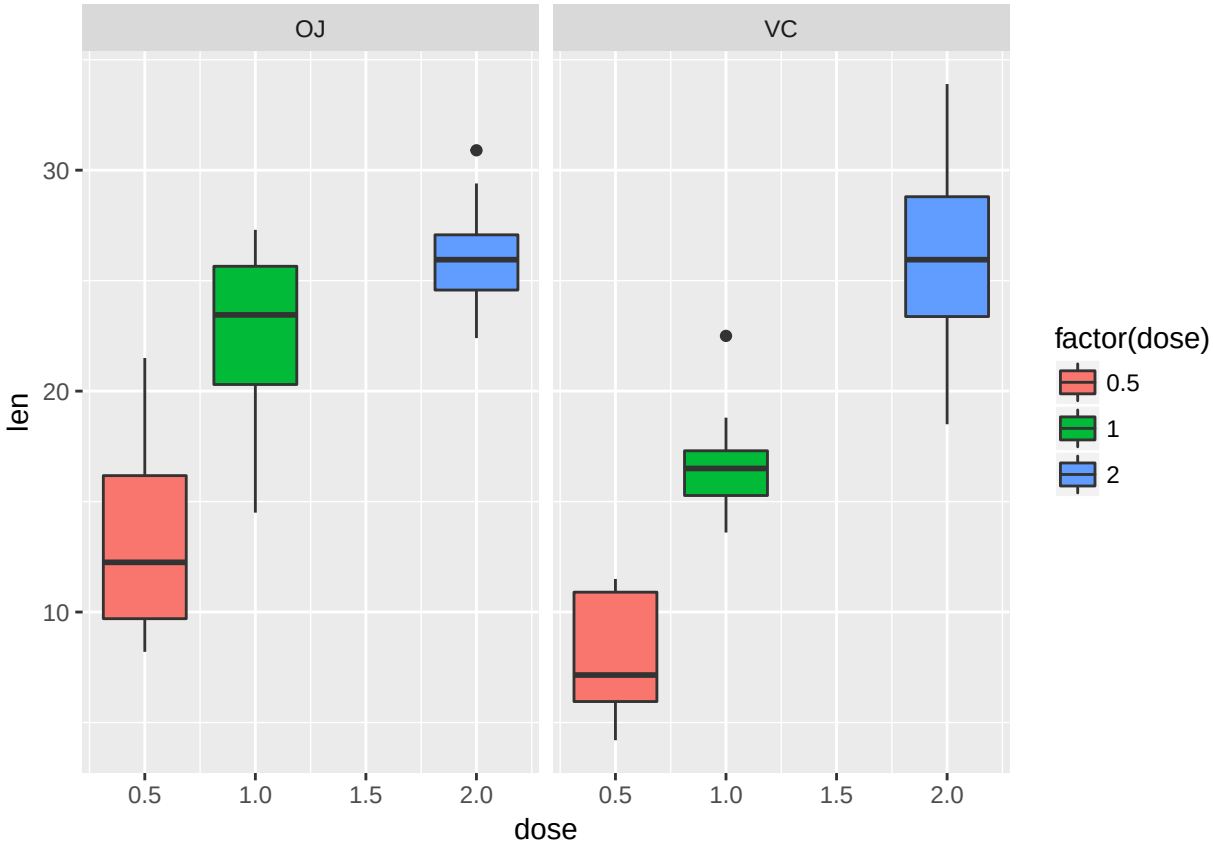
```
library(datasets)
summary(ToothGrowth)
```

```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

The data set examines the tooth length of 60 different guinea pigs after being treated with vitamin C. Tooth length, measured in *len* column, is the length of odontoblasts. The delivery method, denoted in column *supp*, either represents ascorbic acid (VC), or orange juice (OJ). The final column, *dose*, displays the dosage the animal received in mg/day.

Below is a box-plot showing the spread of the data separated by the delivery method *supp*.

```
ggplot(data = ToothGrowth, aes(x = dose, y = len)) +
    geom_boxplot(aes(fill = factor(dose)))+
    facet_grid(~supp)
```



As shown above, there is a clear relationship between dosage and tooth length. The higher the dosage, the longer the tooth. Numericized below:

```
with(ToothGrowth, tapply(len,dose,mean))
```

```
##    0.5      1      2
## 10.605 19.735 26.100
```

Testing Hunches

To confirm this hunch we will test the hypothesis that mean(len|dose=1.0)=mean(len|dose=2.0) without using R's t.test function.

```
xbar.t.test <- mean(ToothGrowth$len[ToothGrowth$dose==2.0]-ToothGrowth$len[ToothGrowth$dose==1.0])
s.t.test <- sd(ToothGrowth$len[ToothGrowth$dose==2.0]-ToothGrowth$len[ToothGrowth$dose==1.0])
n <- length(ToothGrowth$len[ToothGrowth$dose==2.0]-ToothGrowth$len[ToothGrowth$dose==1.0])
Ha <- 0 ##alternative hypothesis

(xbar.t.test-Ha)/(s.t.test/sqrt(n))
```

```
## [1] 4.604647
```

```
qt(0.975,df = n-1) ##0.975 because two sided test
```

```
## [1] 2.093024
```

Since our T-score is greater that our rejection score, we reject the null hypothesis that the difference of the two doses is the same. This leads us to believe that they are indeed different – higher doses do increase tooth length.

Another examination from the box plot above suggests that Orange Juice (OJ) is more effective at increasing tooth length, most notably in lower doses.

```
with(ToothGrowth, tapply(len,supp,mean))
```

```
##       OJ       VC
## 20.66333 16.96333
```

Above shows that in this data set, the mean length for OJ is longer than the mean length of VC. To take this analysis one step further, we will now test the hypothesis that mean(OJ) and mean(VC) are the same using R's t.test function.

```
t.test(len~supp,data = ToothGrowth)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1710156  7.5710156
## sample estimates:
## mean in group OJ mean in group VC
##         20.66333         16.96333
```

There are two things to examine: the p-value and the confidence interval.

Starting with the p-value, since it exceeds 0.05, our level of confidence, we fail to reject the null hypothesis that mean(OJ)-mean(VC)=0.

Similarly, to reject or fail to reject our null with confidence intervals, you examine if the interval contains 0. Since it does contains 0, we fail to reject the null hypothesis that mean(OJ)-mean(VC)=0

Conclusion

We rejected the null that different doses have the same effect on tooth length. This implies that a higher dose leads to a longer tooth.

We fail to reject the null that OJ and VC have the same effect. This implies that the delivery method is irrelevant between these two suppliers.

The above assumes that the mean of the difference between our tested values are normally distributed.