Statistical Inference Course Project

Kurt Fitz

April 1, 2019

Overview

The following report is divided into two parts. The first explores the exponential distribution in relation to the Central Limit Theorem, and the second involves some basic inference using the ToothGrowth data in R.

Part 1 - Simulation Exercise

The Exponential distribution is simulated and compared with the Central Limit Theorem. The simulation is performed 1000 times with a sample size n = 40

Sample and theoretical mean

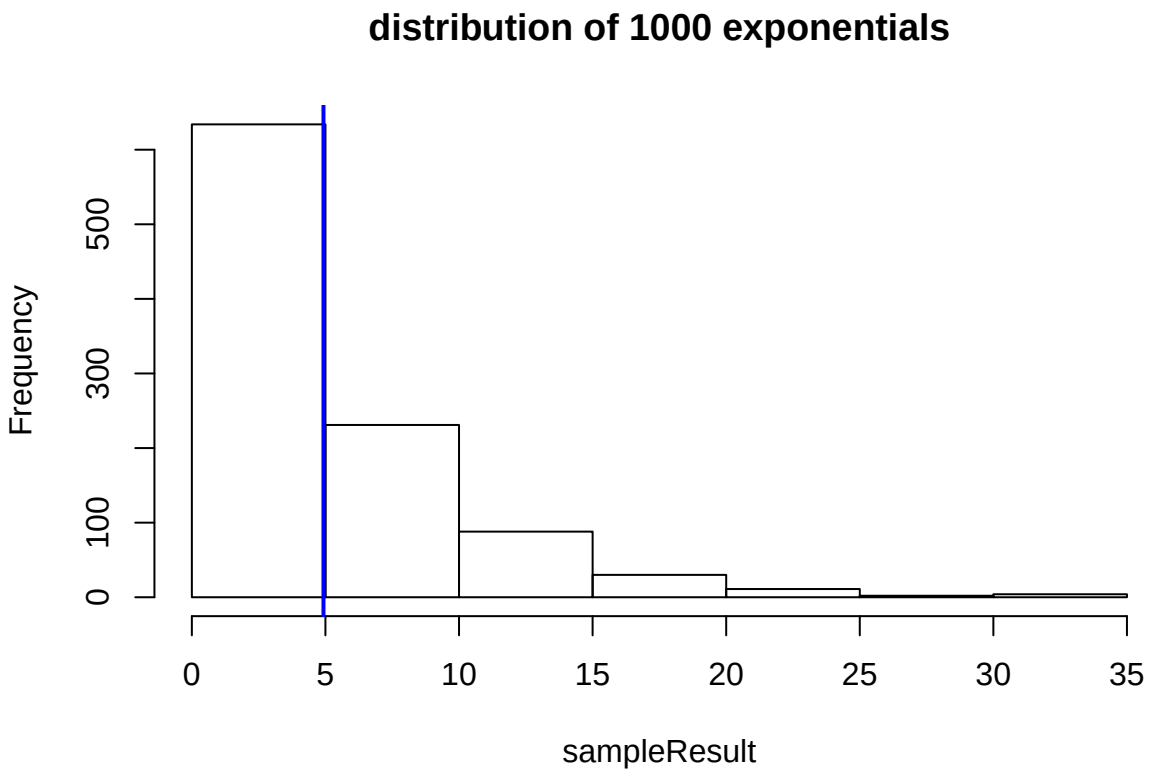
Following we see the distribution of 1000 exponentials:

```
set.seed(33)
lambda <- 0.2 # The rate parameter
n <- 40 # sample size

sampleResult <- (rexp(1000, lambda))


hist(sampleResult, main = "distribution of 1000 exponentials")

sampleMean <- mean(sampleResult)
abline(v = sampleMean, col = "blue", lwd = 2)
```

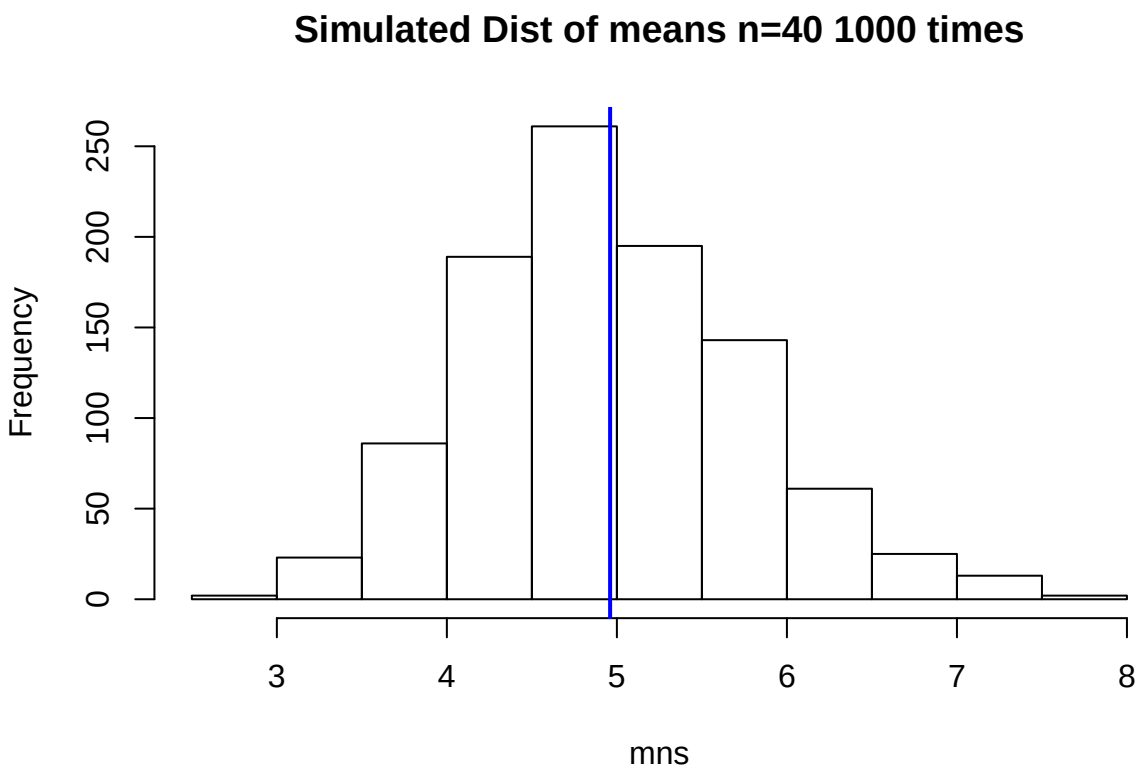


Next, there is the distribution of means of 40 exponentials sampled 1000 times.

```
lambda <- 0.2 # The rate parameter
n <- 40 # sample size
mns = NULL

for (i in 1 : 1000) mns = c(mns, mean(rexp(n, lambda)))
hist(mns, main = "Simulated Dist of means n=40 1000 times")

simMean <- mean(mns)
abline(v = simMean, col = "blue", lwd = 2)
```




```
sampleMean <- mean(sampleResult)
sampleVariance <- var(mns)
sampleStdDev <- (1/lambda)/sqrt(n)

simMean <- mean(mns)
simVariance <- (1/lambda)^2/n
simStdDev <- (1/lambda)/sqrt(n)
```

The sample mean from above is 4.9265823 compared to a simulated theoretical mean of 4.9601793. Notice how the distribution is centered around 5. When we compare the distribution of means from our 40 exponentials simulated 1000 times, we can see that the sample mean from above tends to be centered around the same point.

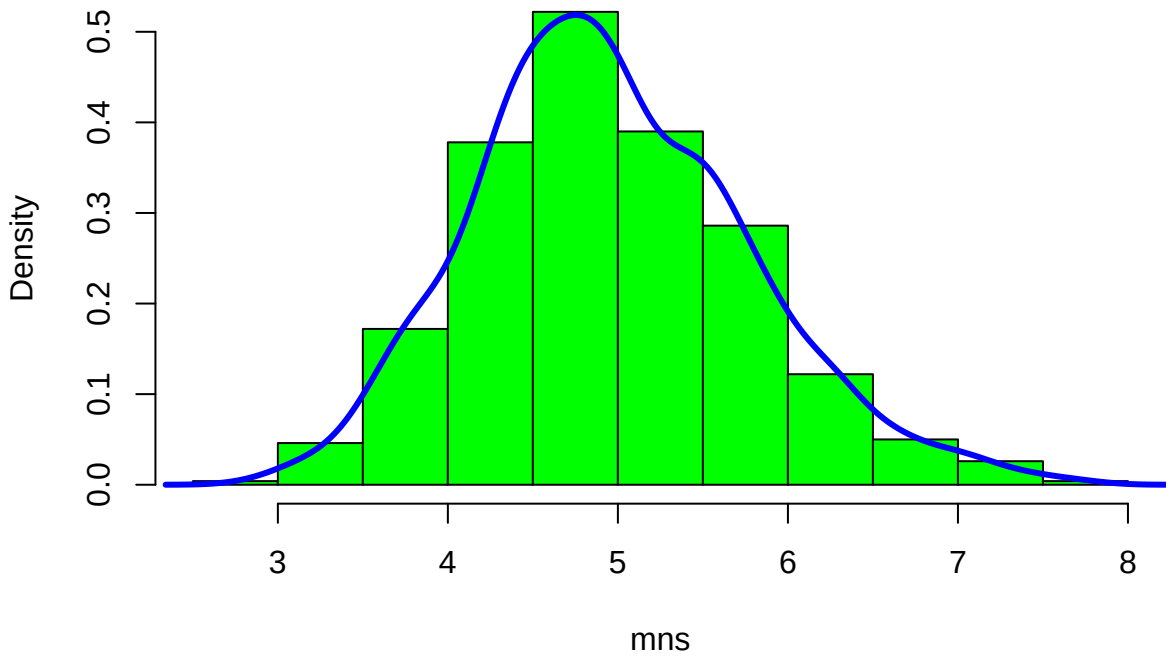
Sample variance and Standard deviations and theoretical variance of distribution

The variance of sample distribution is 0.6513642, while the variance of the means of the simulated samples is 0.625. The sample standard deviation is 0.7905694 and the theoretical standard deviation is 0.7905694.

Normality of distribution

The following graph demonstrates that the simulated distribution of the means of 40 exponentials (1000 times) approximates a normal distribution. This is consistent with the Central Limit Theorem in that the distribution of the averages of samples are normally distributed.

```
hist(mns, prob=TRUE, col="green", main="", breaks=10)
lines(density(mns), lwd=3, col="blue")
```



Part 2 Basic Inferential Data Analysis of ToothGrowth data in R datasets package.

The following basic inferential analysis will compare tooth growth by dose and supplement using confidence intervals.

Below is a basic summary of the tooth growth data:

```
library(datasets)
data("ToothGrowth")

head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

```r
dim(ToothGrowth)
```

```
## [1] 60  3
```

```r
str(ToothGrowth)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```r
summary(ToothGrowth)
```

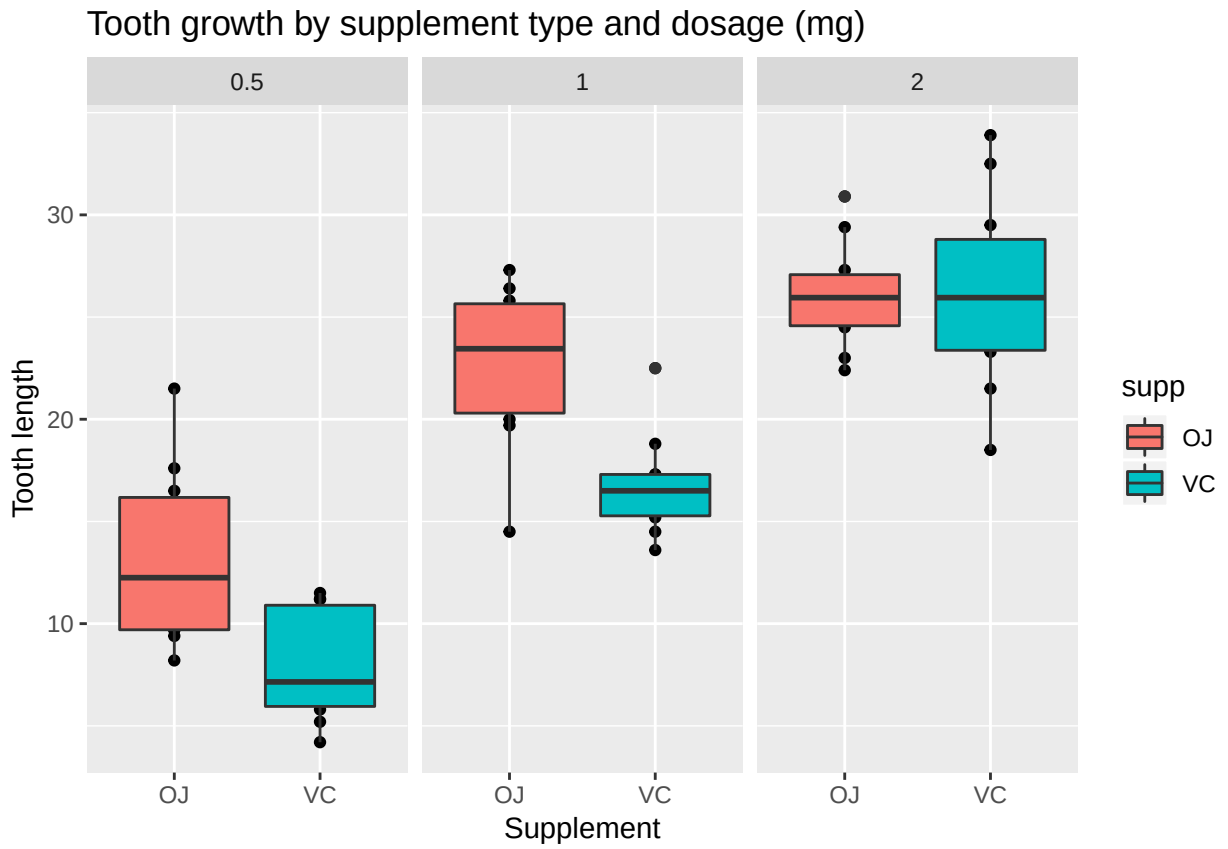
```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

Tooth Growth by Supplement

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```r
qplot(supp, len, data=ToothGrowth, facets=~dose, main="Tooth growth by supplement type and dosage (mg)"
```




```
t.test(len ~ supp, paired = FALSE, var.equal = FALSE, 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
```

We can see from the Student's t-test that the 95% confidence interval contains 0; therefore, we cannot reject the null hypotheses that there is no effect between the two supplements.

Tooth Growth by Dose

Next, we can examine the relationship between tooth growth and dosage.

```
## Subset to dosage 0.5 - 1.0
doseRange1 <- subset(ToothGrowth, dose %in% c(0.5, 1.0))
```

```r
## Subset to dosage 0.5 - 2.0
doseRange2 <- subset(ToothGrowth, dose %in% c(0.5, 2.0))

## Subset to dosage 1.0 - 2.0
doseRange3 <- subset(ToothGrowth, dose %in% c(1.0, 2.0))

t.test(len ~ dose, paired = FALSE, var.equal = FALSE, data = doseRange1)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -6.4766, df = 37.986, p-value = 1.268e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.983781  -6.276219
## sample estimates:
## mean in group 0.5   mean in group 1
##            10.605            19.735
```

```r
t.test(len ~ dose, paired = FALSE, var.equal = FALSE, data = doseRange2)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -11.799, df = 36.883, p-value = 4.398e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -18.15617 -12.83383
## sample estimates:
## mean in group 0.5   mean in group 2
##            10.605            26.100
```

```r
t.test(len ~ dose, paired = FALSE, var.equal = FALSE, data = doseRange3)
```

```
##
##  Welch Two Sample t-test
##
## data:  len by dose
## t = -4.9005, df = 37.101, p-value = 1.906e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.996481 -3.733519
## sample estimates:
## mean in group 1 mean in group 2
##          19.735          26.100
```

Conclusions

From the preceeding t-tests performed on the three dose ranges, we can conclude that there is a relationship between the increase and dosage and increase in growth. The p-value for each of the three ranges was below 0 and the 95% confidence intervals do not contain 0.