Peer-graded Assignment: Statistical Inference Course Project

Xi Fang

6/11/2020

Part 1: Simulation Exercise

This project will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. I will investigate the distribution of averages of 40 exponentials.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials.

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

```
library(ggplot2)
 l <- 0.2 # the rate parameter
 nsam <- 40 # sample size
 nsim <- 1000 # simulation size
 samplemean <- NULL
 for (i in 1:nsim) {
         samplemean <- c(samplemean, mean(rexp(nsam,l)))
 }
 mean(samplemean)
```

```
## [1] 4.979311
```

```
 mean_theo <- 1/l # theoretical mean
 mean_theo
```

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

The mean is very close to the theoretical mean distribution of 5.

Q2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

```
sigma_theo <- 1/l/sqrt(nsam) # theoretical standard error
var_theo <- sigma_theo^2 # theoretical variance
var_sam <- var(samplemean)
var_theo
```

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

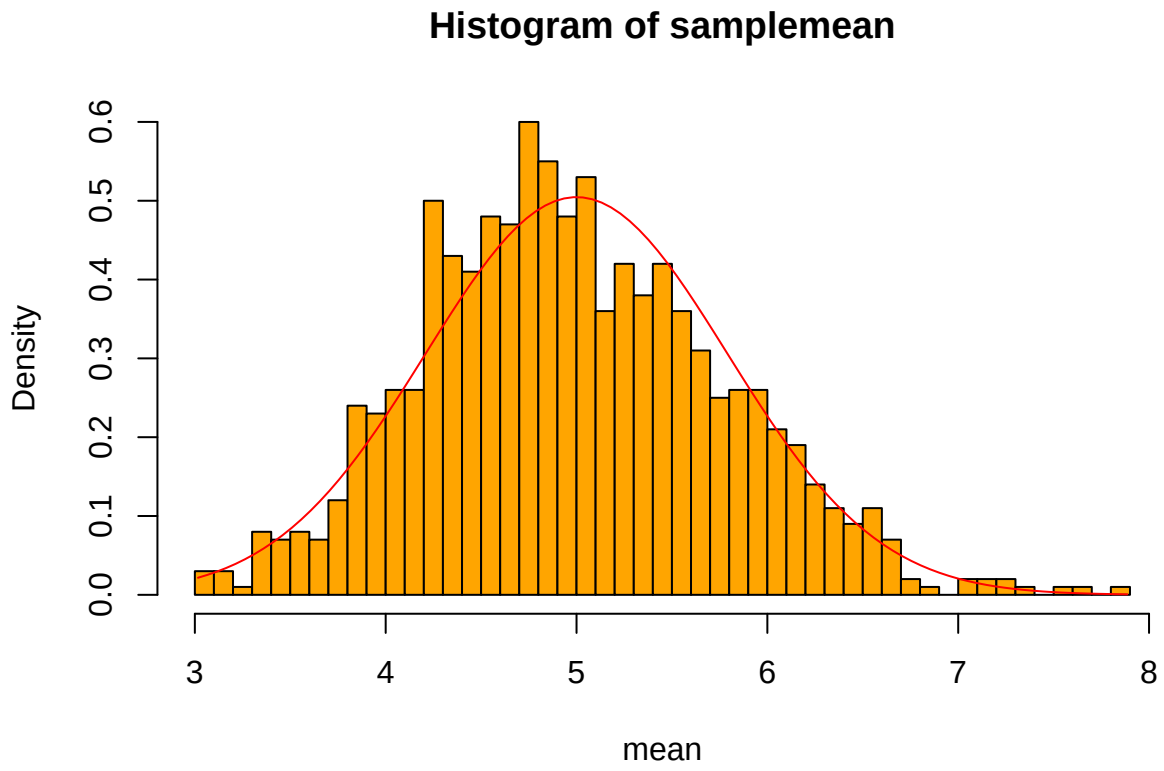
```
var_sam
```

```
## [1] 0.6048647
```

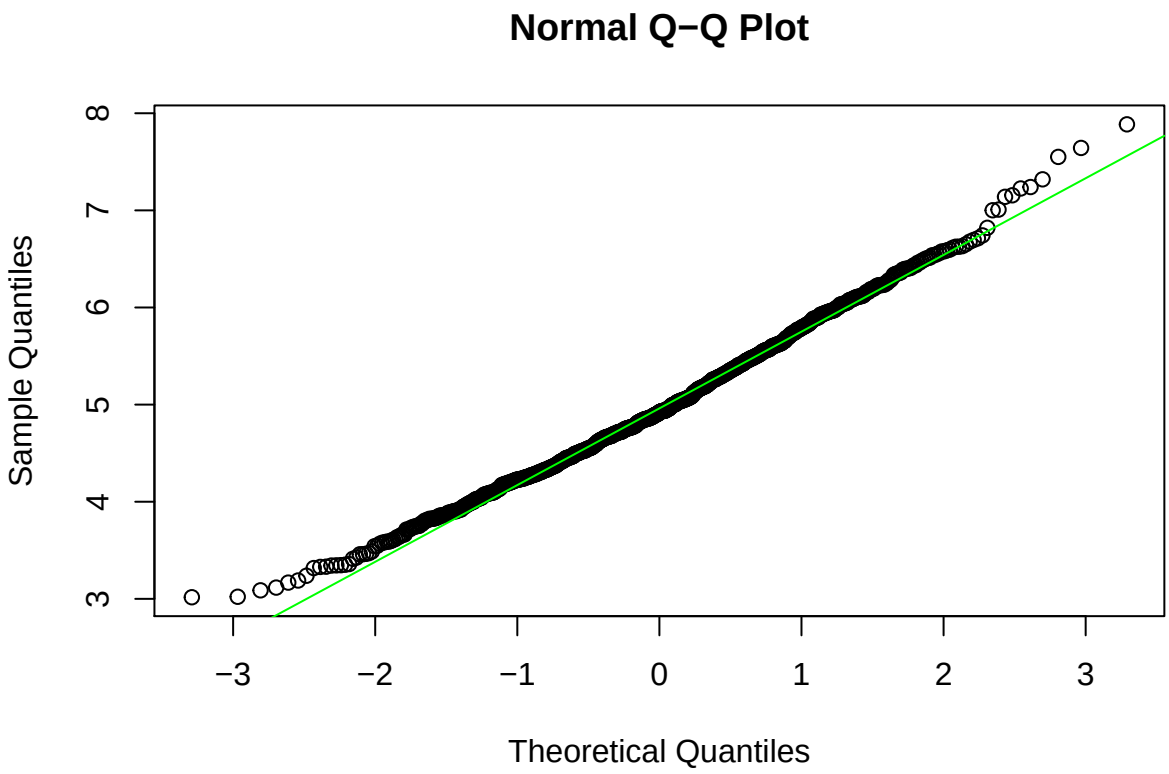
The sample variance is very close to the theoretical variance.

Q3. Show that the distribution is approximately normal.

```
hist(samplemean, breaks = nsam, prob = T, col = 'orange', xlab = "mean")
x <- seq(min(samplemean), max(samplemean), length=100)
lines(x, dnorm(x, mean = 1/l, sd=(1/l/sqrt(nsam))), pch=25, col="red")
```




```
qqnorm(samplemean)
qqline(samplemean, col="green")
```



The sample distribution is very close to a normal distribution.

Part 2: Basic Inferential Data Analysis

This report will analyze the ToothGrowth data in the R datasets package #### Q1. Load the ToothGrowth data and perform some basic exploratory data analyses

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

```
dim(ToothGrowth)
```

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

```
names(ToothGrowth)
```

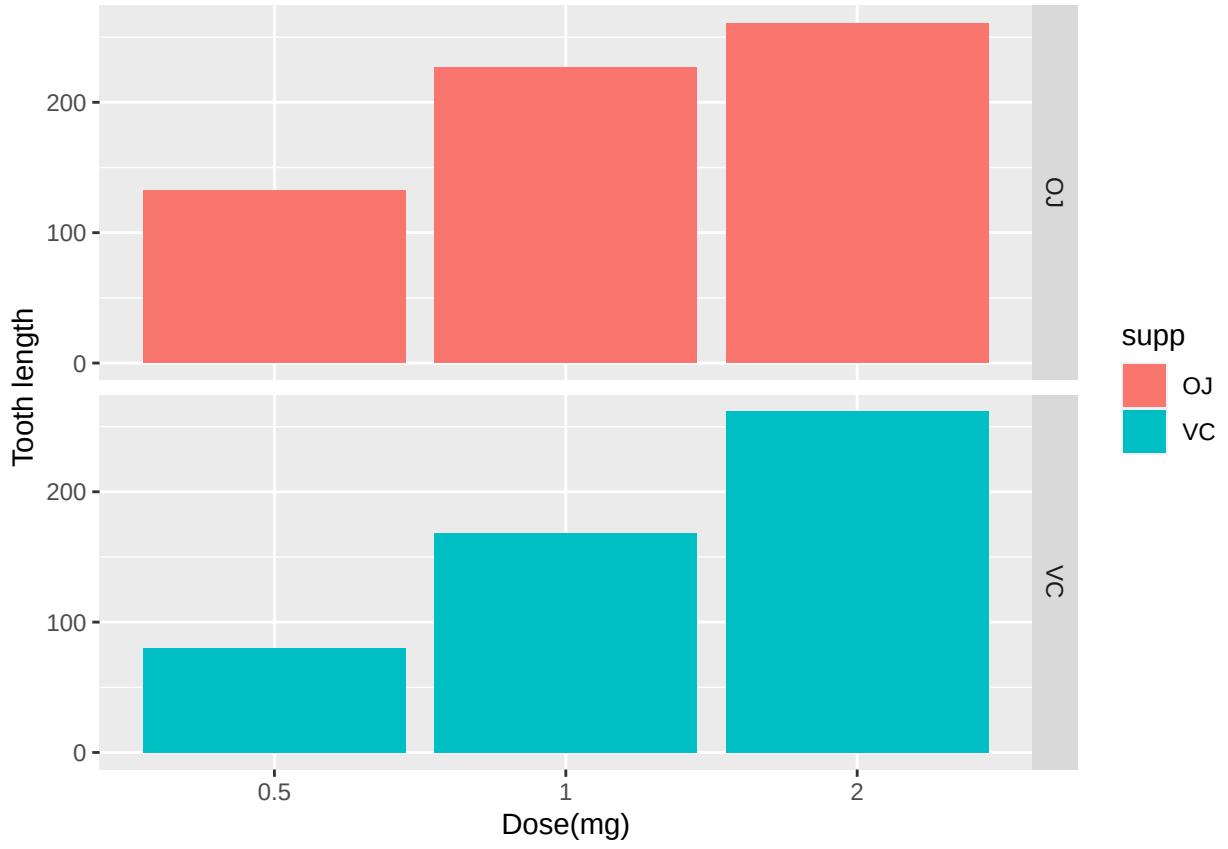
```
## [1] "len"  "supp" "dose"
```

Q2. Provide a basic summary of the data.

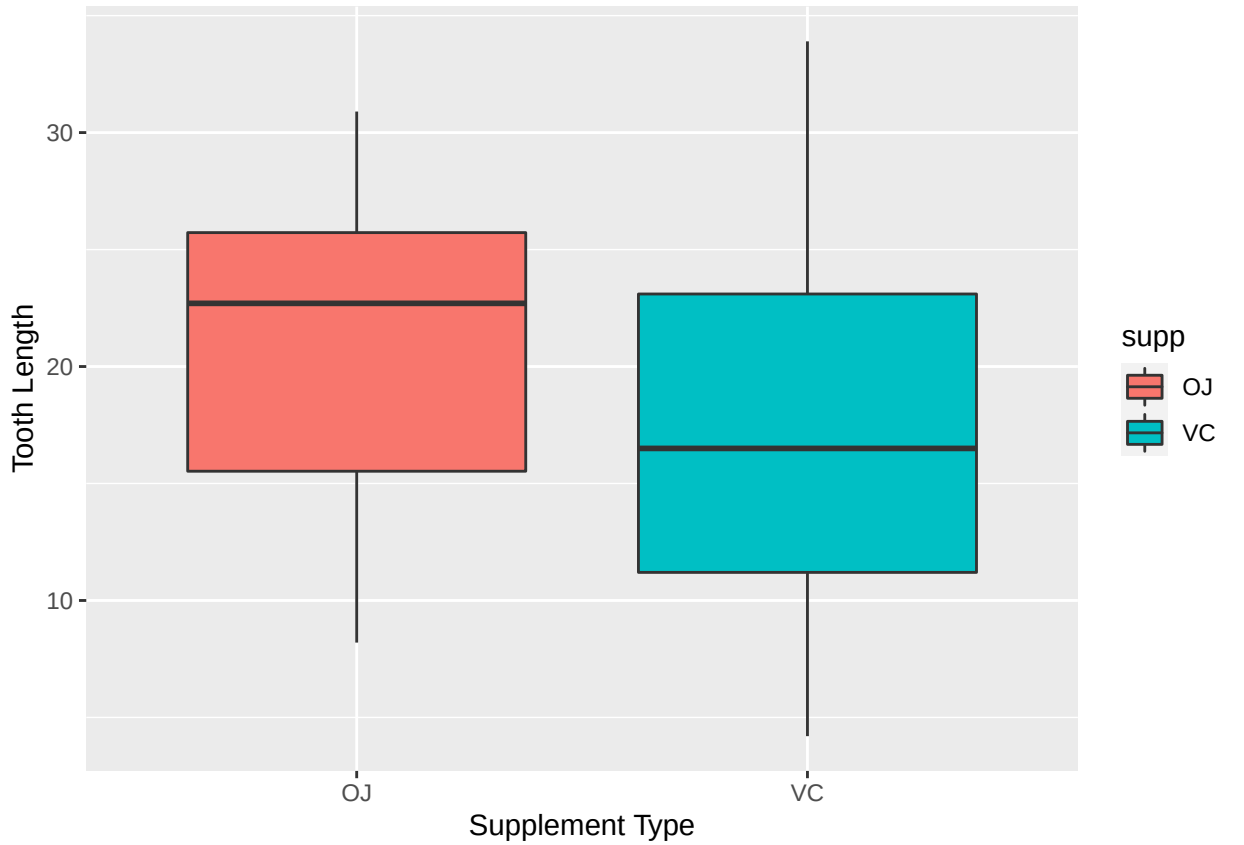
```
library(ggplot2)
summary(ToothGrowth)
```

```
##       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
```

```
ggplot(data=ToothGrowth, aes(x=as.factor(dose), y=len, fill=supp)) +
        geom_bar(stat = "identity") +
        facet_grid(supp~.) +
        xlab("Dose(mg)") +
        ylab("Tooth length")
```




```
ggplot(data = ToothGrowth, aes(x=supp, y=len)) + geom_boxplot(aes(fill=supp)) +
        xlab("Supplement Type") + ylab ("Tooth Length")
```

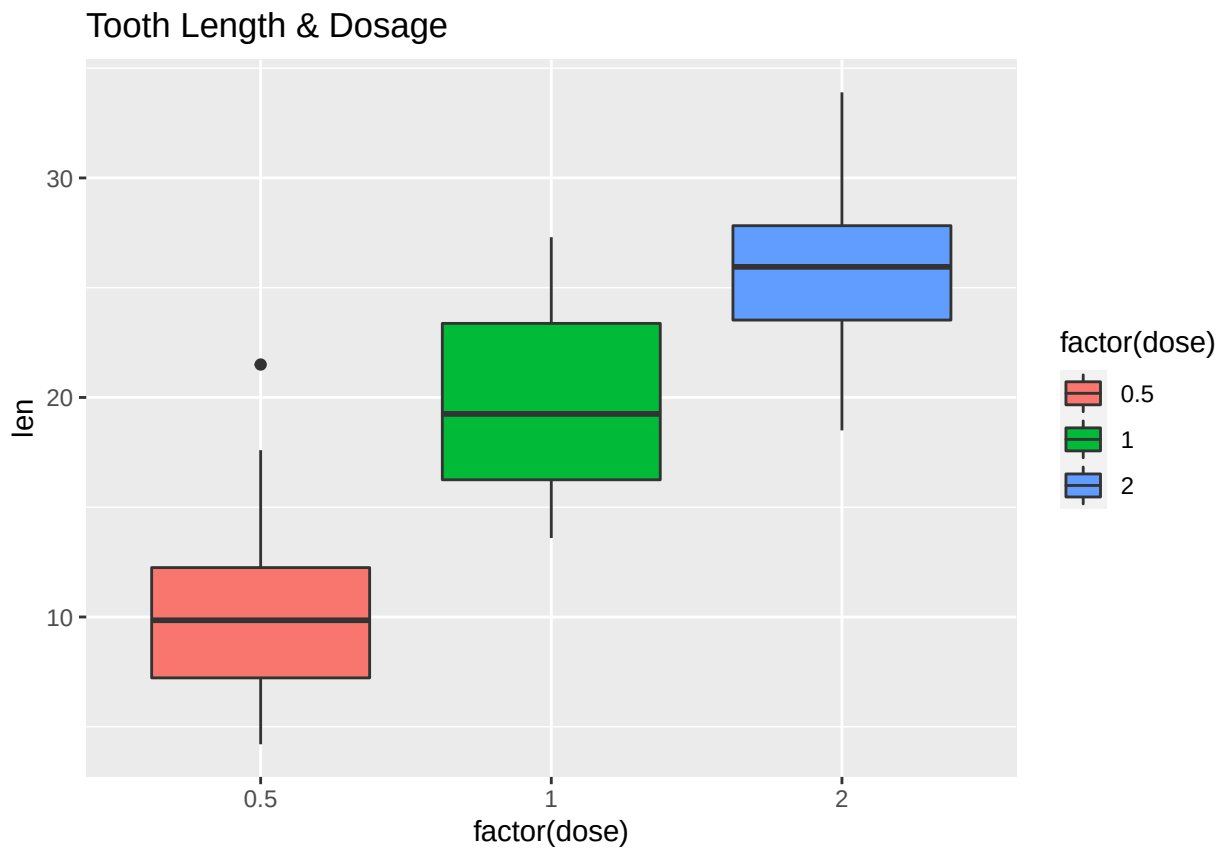


Q3. Use confidence intervals and hypothesis test to compare tooth growth by supp and dose. (Only use the techniques from class, even if there's other approaches worth considering)

```
unique(ToothGrowth$dose) # three dose groups
```

```
## [1] 0.5 1.0 2.0
```

```
ggplot(aes(x=factor(dose), y=len), data= ToothGrowth) +
        geom_boxplot(aes(fill = factor(dose))) +
        labs(title = "Tooth Length & Dosage")
```




```
h1 <- t.test(len ~ supp, data = ToothGrowth)
h1
```

```
##
##  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
```

```
h2 <- t.test(len ~ supp, data = subset (ToothGrowth, dose == 0.5))
h2
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 3.1697, df = 14.969, p-value = 0.006359
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  1.719057 8.780943
## sample estimates:
## mean in group OJ mean in group VC
##            13.23             7.98
```

```
h3 <- t.test(len ~ supp, data = subset (ToothGrowth, dose == 1))
h3
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = 4.0328, df = 15.358, p-value = 0.001038
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.802148 9.057852
## sample estimates:
## mean in group OJ mean in group VC
##            22.70            16.77
```

```
h4 <- t.test(len ~ supp, data = subset (ToothGrowth, dose == 2))
h4
```

```
##
##  Welch Two Sample t-test
##
## data:  len by supp
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean in group OJ mean in group VC
##            26.06            26.14
```

Conclusions

OJ leads to more teeth growth than AC for dosages 0.5 and 1.0. OJ and AC were not different for teeth growth at dose 2.0. However, overall, OJ and AC were not significantly different in terms of the effeciency for teeth growth.