Statistical Inference Course Project

Neil Bungcayao

9/9/2020

Part 1: Simulation Exercise

In this project you 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. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

Parameters

```r
lambda <- 0.2
n <- 40
simulations <- 1000
```

Setting the seed number for reproducibility

```r
set.seed(1337)
```

Simulation

```r
mean = NULL
for (i in 1 : 1000) mean = c(mean, mean(rexp(n, lambda)))
```

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

```r

samplemean<-mean(mean)
```

Sample Mean

The sample mean is `samplemean = ` 5.0559953.

```r
theoreticalmean <- 1/lambda
```

Theoretical Mean

The theoretical mean of the distribution is `theoreticalmean` = 5.

The values of the two means are close to each other.

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

```r
samplevariance<-var(mean)
```

Sample Variance

The sample variance is `samplevariance` = 0.6543703.

```r
theoreticalvariance<- (1/lambda)^2/n
```

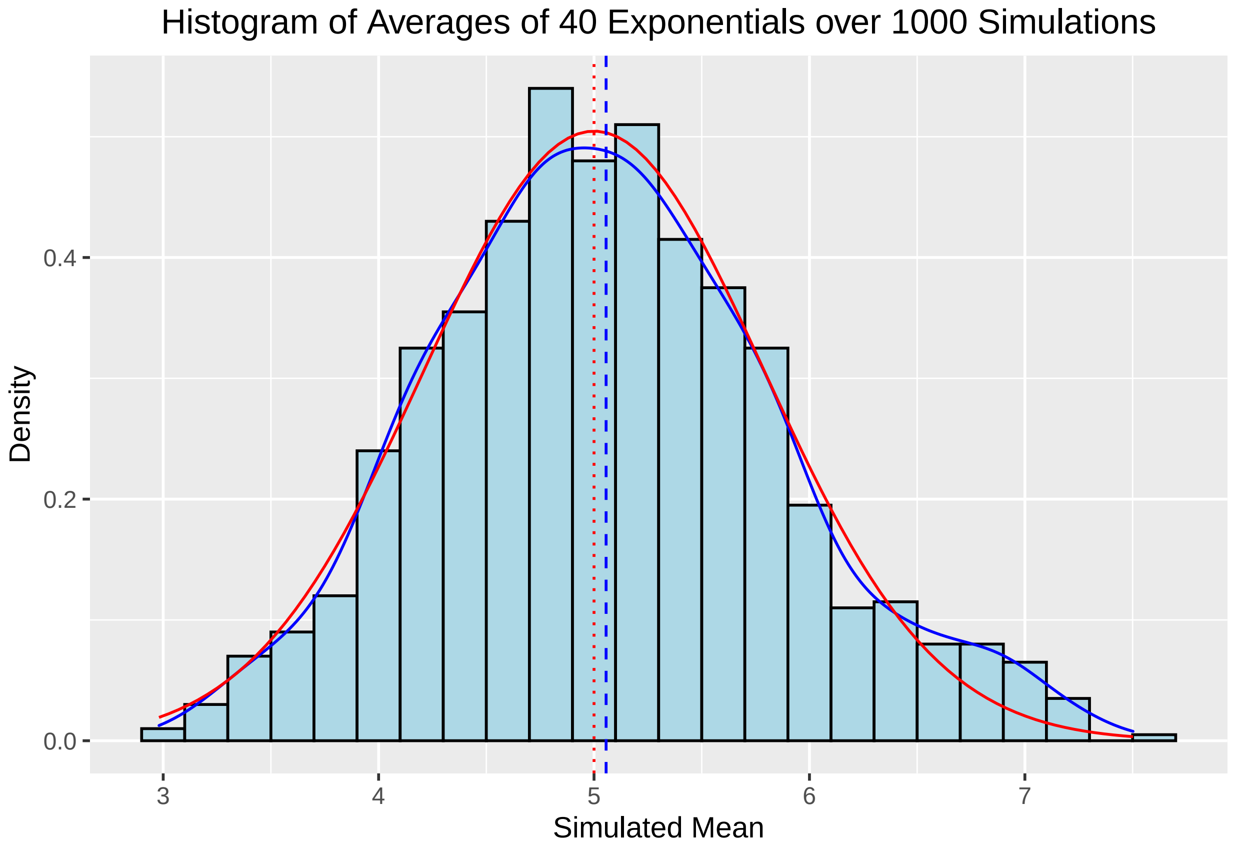
Theoretical Variance

The theoretical variance of the distribution is `theoreticalvariance` = 0.625.

The values of the two variances are close to each other.

3. Show that the distribution is approximately normal.

```r
#Transforming the simulated data
data <- data.frame(mean)
#Plotting the Histogram overlayed by the Density of the Simulated Means and the Normal Distribution
#with Parameters from the Exponential Distribution
library(ggplot2)
ggplot(data, aes(x=mean)) +
        geom_histogram(aes(y = ..density..), binwidth = 0.2,  color ="black", fill = "lightblue") +
        geom_density(color = "blue") +
        stat_function(fun=dnorm,args=list(mean=1/lambda, sd=sqrt(theoreticalvariance)),
                      color = "red") +
        geom_vline(xintercept = theoreticalmean, linetype="dotted", color = "red") +
        geom_vline(xintercept = samplemean, linetype="dashed", color = "blue") +
        labs(title = "Histogram of Averages of 40 Exponentials over 1000 Simulations",
             x = "Simulated Mean", y ="Density") +
        theme(plot.title = element_text(hjust = 0.5))
```



The blue curve is the density of the simulation while the red curve represents the density of the normal distribution with the parameters depending on the lambda parameter of the exponential distribution. It can be seen that the two densities are close and alike. Additionally, the dashed blue line represents the theoretical mean while the dotted red line represents the mean of the sampled 1000 simulations of 40 exponentials.

Part 2: Basic Inferential Data Analysis

Now in the second portion of the project, we're going to analyze the ToothGrowth data in the R datasets package.

ToothGrowth Dataset

Brief Description

The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC).

Dataset

Variable Name	Data Type	Description
len	numeric	Tooth Length
supp	factor	Supplement type (VC or OJ).
dose	numeric	Dose in milligrams/day

1. Load the ToothGrowth data and perform some basic exploratory data analyses

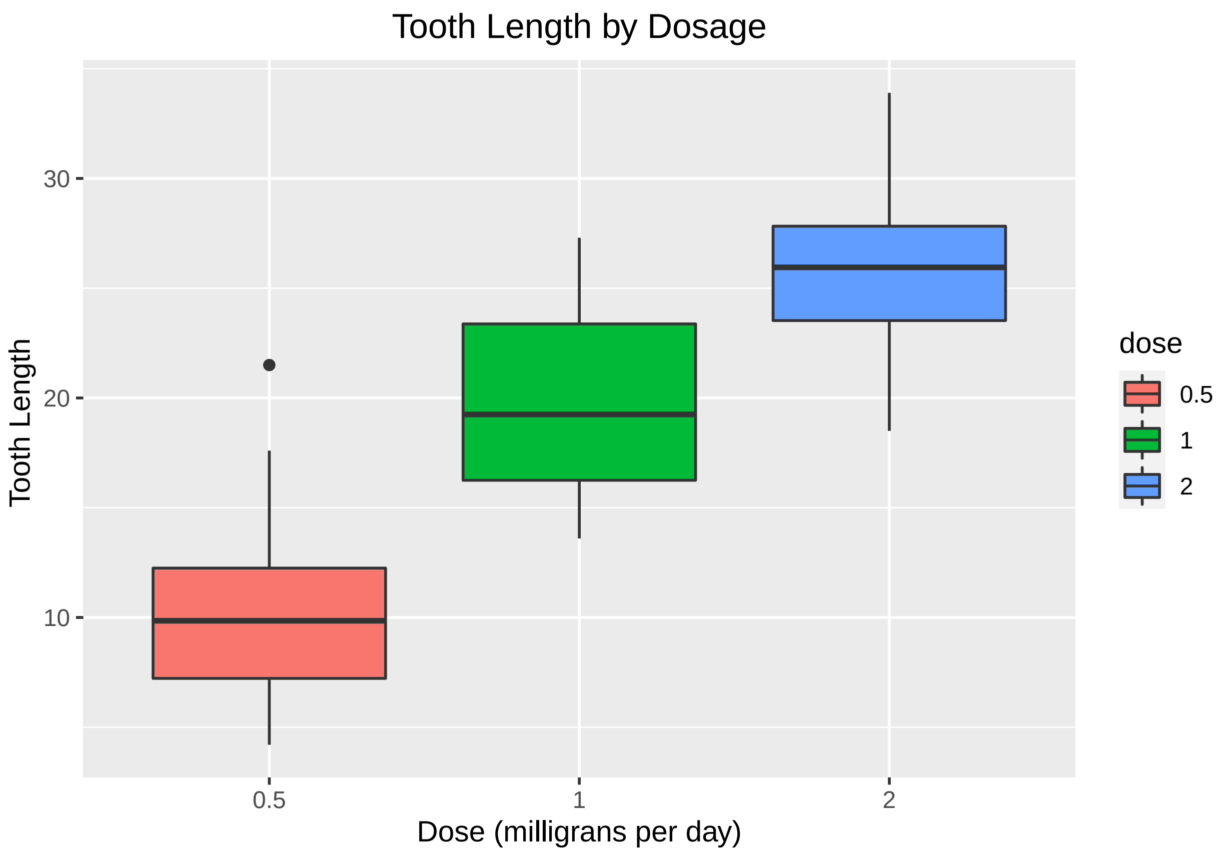
```r
library(datasets)
library(knitr)
data(ToothGrowth)
kable(ToothGrowth[1:10, ])
```

Loading ToothGrowth Dataset

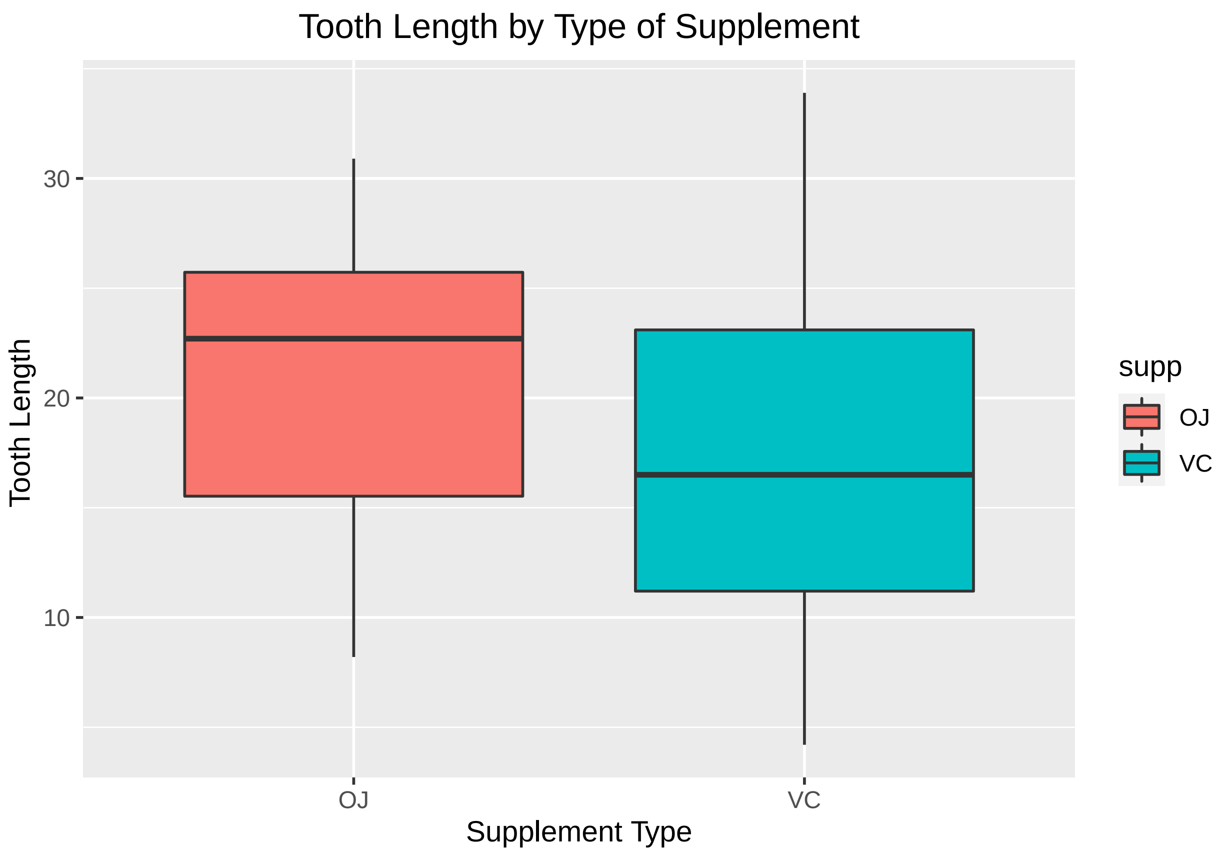
len	supp	dose
4.2	VC	0.5
11.5	VC	0.5
7.3	VC	0.5
5.8	VC	0.5
6.4	VC	0.5
10.0	VC	0.5
11.2	VC	0.5
11.2	VC	0.5
5.2	VC	0.5
7.0	VC	0.5

```r
library(ggplot2)
ToothGrowth$dose<-as.factor(ToothGrowth$dose)
ggplot(data = ToothGrowth, aes(x = dose, y = len)) +
  geom_boxplot(aes(fill=dose)) +
  labs(title="Tooth Length by Dosage",
       x = "Dose (milligrans per day)",
       y = "Tooth Length ") +
  theme(plot.title = element_text(hjust = 0.5))
```

Exploratory Data Analysis



```
library(ggplot2)
ggplot(data = ToothGrowth, aes(x = supp, y = len)) +
  geom_boxplot(aes(fill=supp)) +
  labs(title="Tooth Length by Type of Supplement",
       x = "Supplement Type",
       y = "Tooth Length ") +
  theme(plot.title = element_text(hjust = 0.5))
```



2. Provide a basic summary of the data.

```
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: Factor w/ 3 levels "0.5","1","2": 1 1 1 1 1 1 1 1 1 1 …

```
summary(ToothGrowth$len)
```

Min. 1st Qu. Median Mean 3rd Qu. Max. 4.20 13.07 19.25 18.81 25.27 33.90

```
summary(ToothGrowth$supp)
```

OJ VC 30 30

```
summary(ToothGrowth$dose)
```

0.5 1 2 20 20 20

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

```r
t.test(len ~ supp, data = ToothGrowth, var.equal = FALSE)
```

Type of Supplement: OJ vs. VC

```
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 #### Dosage: 0.5 vs. 1.0

```r
library(dplyr)
ToothGrowth_Pair1 <- ToothGrowth %>% filter(dose %in% c(0.5, 1.0))
t.test(len ~  dose, data = ToothGrowth_Pair1, var.equal = FALSE)
```

```
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
library(dplyr)
ToothGrowth_Pair2 <- ToothGrowth %>% filter(dose %in% c(0.5, 2.0))
t.test(len ~  dose, data = ToothGrowth_Pair2, var.equal = FALSE)
```

Dosage: 0.5 vs 2.0

```
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
library(dplyr)
ToothGrowth_Pair3 <- ToothGrowth %>% filter(dose %in% c(1.0, 2.0))
t.test(len ~  dose, data = ToothGrowth_Pair3, var.equal = FALSE)
```

Dosage: 1.0 vs 2.0

```
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

4. State your conclusions and the assumptions needed for your conclusions.

Assumptions

1. Unequal Variances
2. The distribution of the mean tooth growth is approsimately normal.
3. Randomization mechanism is applied on which guinea pigs will receive a certain type of treatment.
4. Independent and Identcally Distributed for all elements on the study.

Conclusions

1. Based on the Exploratory Data Analysis, the distribution of the tooth growth is higher as the dosage of the supplement is increased. regardless of the type.
2. The concentration of the tooth growth on OC type is a little bit higher on OJ as compared to VC. However, the spread of the tooth growth on VC is much wider than the spread for OJ supplement type.
3. As comfirmed by the t-tests performed on comparing tooth growth by supplement type, at 0.05 level of significance and since p-value = 0.06063, we do no reject the null hypothesis that tooth growth is not affected by type of supplement.
4. However, comparing tooth growth per level of dosage pairwise, all of the p-values are smaller than 0.05. We can reject the null hypothesis that the tooth growth is not affected by level of dosage.
5. Further comparisosn and tests on the interaction of the two variables: supplement type and level of dosage. Additionally, multiple tests can be used to control familywise error rates.