Inference Project4: Simulation Exercise and ToothGrowth Data Analysis

Shuai Wang

Dec, 11, 2016

Part 1: Simulation Exercise

Title:

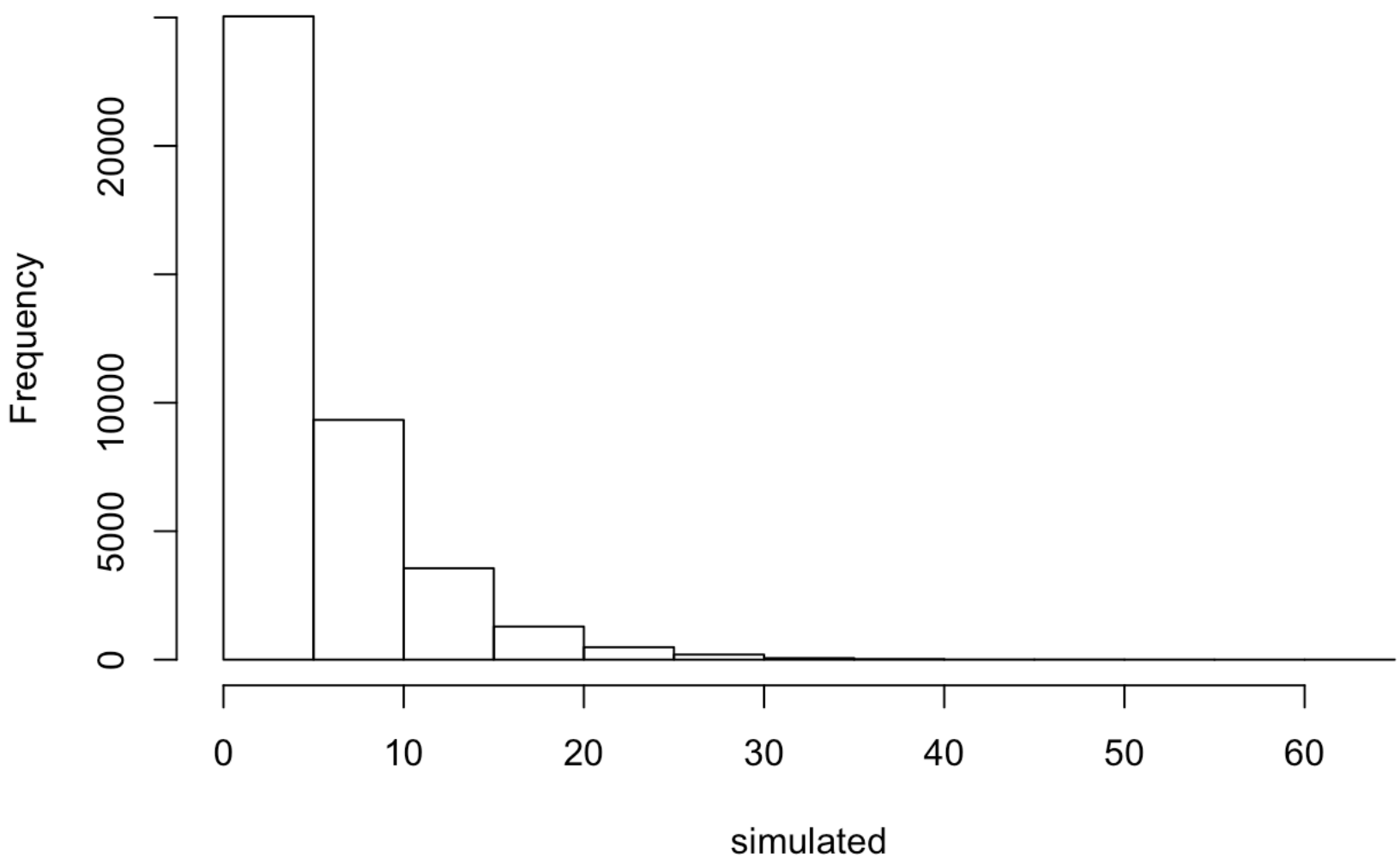
Comparing Simulation Results of exponential distribution with Central Limit Theorem

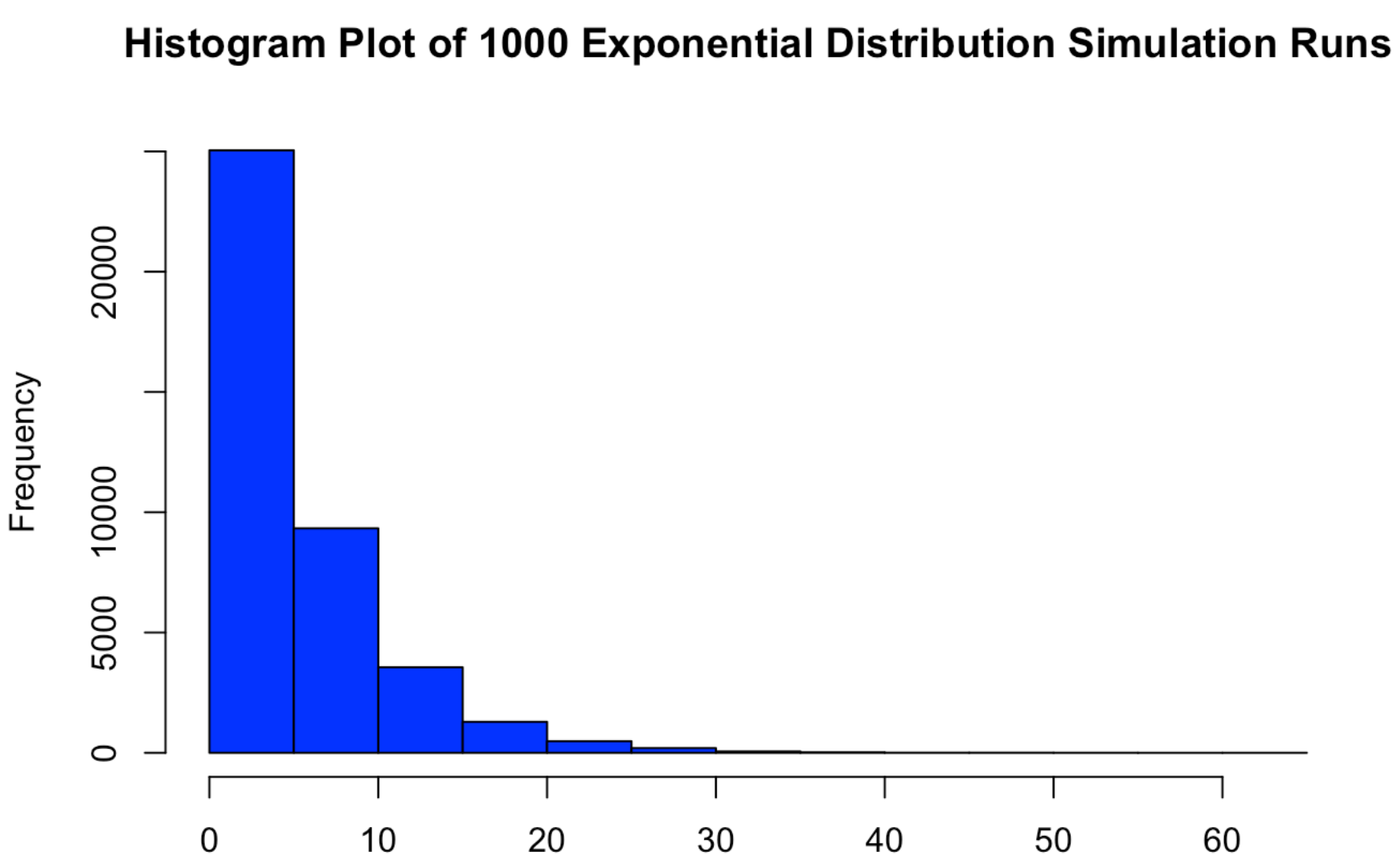
Overview:

In the first part of the project, I will run simulation results of exponential distribution. The exponential distribution number generator is based on rexp() function.

Simulations:

```
# setting up the experiment parameters
nosim <- 1000
n <- 40
lambda <- 0.2
simulated<-matrix(rexp(nosim * n, lambda), nosim)
plot(hist(simulated),xlab="Value",ylab="Frequency",col="blue",main="Histogram Plot of
1000 Exponential Distribution Simulation Runs")
```


Histogram of simulated
Frequency
0
5000
10000
20000
0
10
20
30
40
50
60
simulated


Histogram Plot of 1000 Exponential Distribution Simulation Runs
Frequency
0
5000
10000
20000
0
10
20
30
40
50
60

Value

From the instruction, we know that the the mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda.

Sample Mean versus Theoretical Mean:

Include figures with titles. In the figures, highlight the means you are comparing. Include text that explains the figures and what is shown on them, and provides appropriate numbers.

Here in the figure below. I present the distribution of the normalized means of 1000 simulations in blue histogram plot. I also plot standard normal distribution in the black curve.

```
#plot(hist_plot,xlab="Value",ylab="Frequency",col="blue",main="Histogram Plot of One
Exponential Distribution Simulation Run")
observed.mean<-apply(simulated, 1, mean)
normalizedMean<-observed.mean-1/lambda
normalizedData<-normalizedMean*sqrt(40)/(1/(lambda))
hist(normalizedData,xlab="",ylab="",yaxt='n',xaxt='n',main="",col="blue")
par(new=TRUE)
x<-rnorm(1000, mean = 0, sd = 1)
p<-density(x,main="",xlab="",ylab="")
plot(p,main="",xlab="",ylab="")
title(main="Distribution of Means of Exponential with Normal Distribution Overlap",xl
ab="value",ylab="Frequency",sub="Figure 1. Simulated Exercise")
```

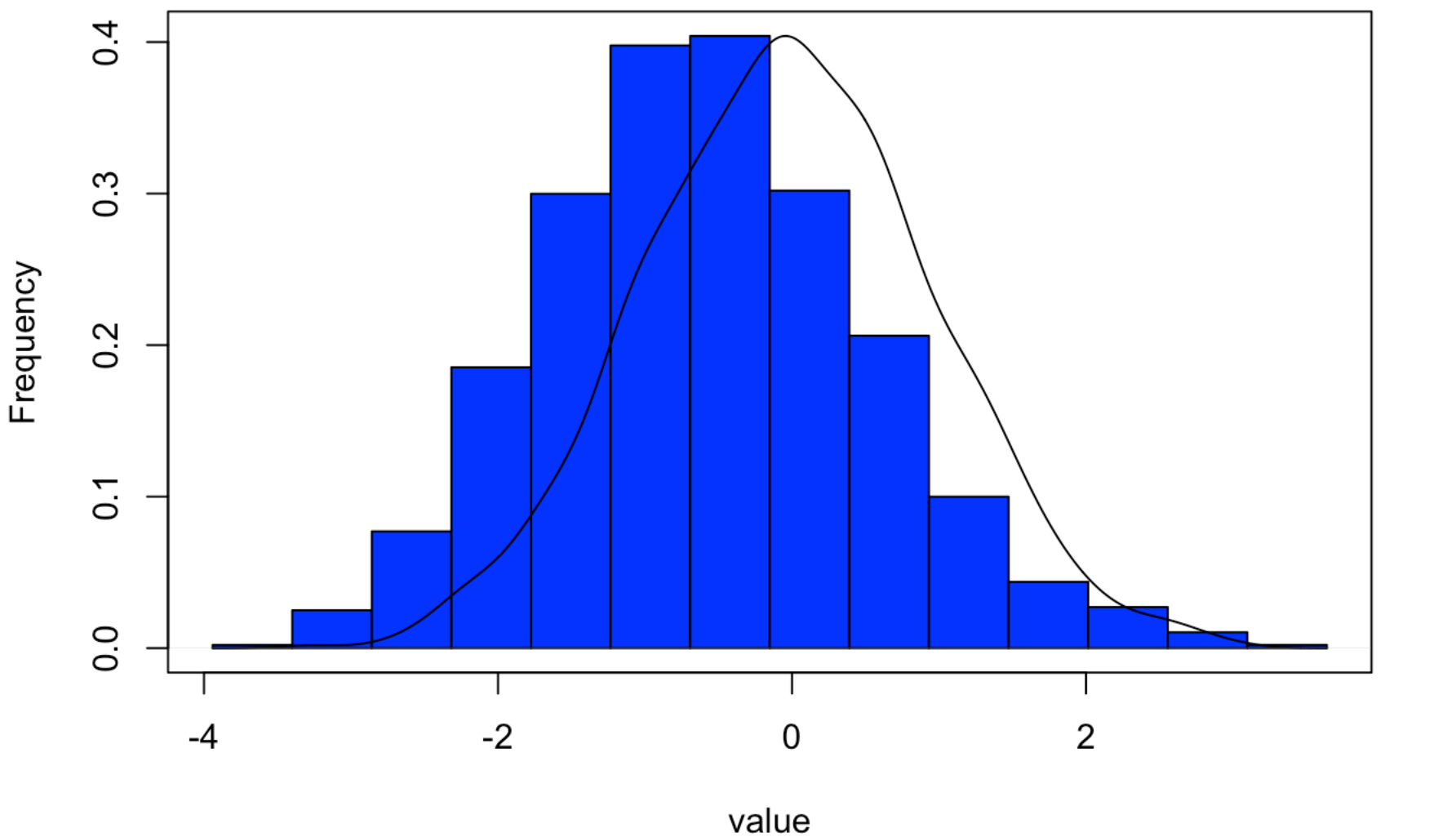


Figure 1. Simulated Exercise

```
#observed.sd
observed.var<-var(apply(simulated,1,mean))
#observed.var
#sqrt(observed.var)
#abline( v = observed.mean, col = "black")
#text(1,0, "Obsevered Mean", col = "white", adj = c(0, -.1))
#abline( v = 1/lambda, col = "red")
#text(3,0, "Theoretical Mean", col = "white", adj = c(0, -.1))
```

Sample Variance versus Theoretical Variance:

Include figures (output from R) with titles. Highlight the variances you are comparing. Include text that explains your understanding of the differences of the variances.

```
observed.var<-var(observed.mean)
sprintf("The observed variance of mean should be:%.2f",observed.var)
```

```
## [1] "The observed variance of mean should be:0.65"
```

```
expected.var<-(1/lambda^2)/n
sprintf("The expected variance of mean should be:%.2f",expected.var)
```

```
## [1] "The expected variance of mean should be:0.62"
```

Distribution:

Via figures and text, explain how one can tell the distribution is approximately normal.

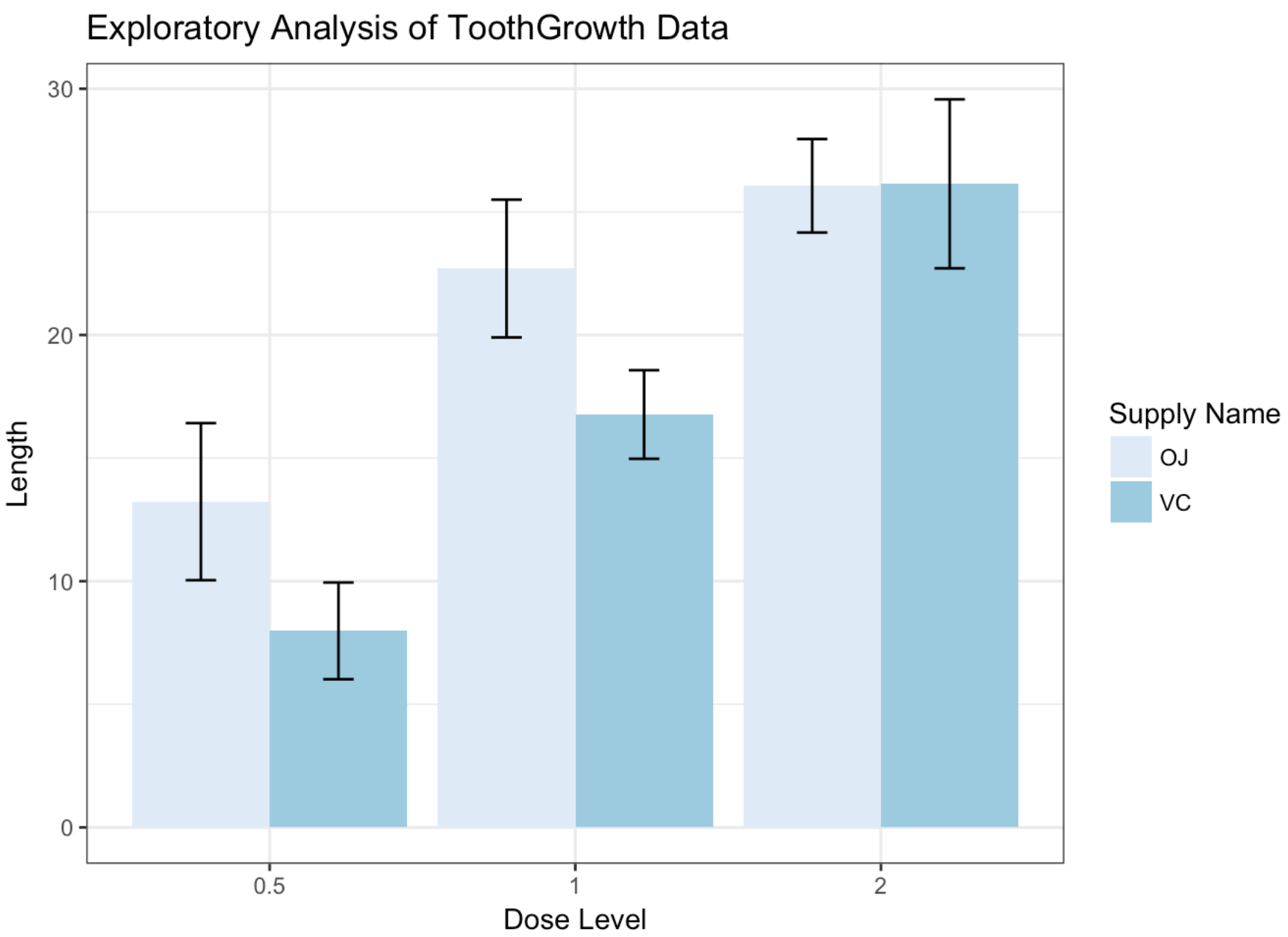
Based on the Central Limit Theorem, we learn that the mean of random distributions follow standard normal distribution when the repetition of simiulations gets larger. From figrue 1, we can see that the distribution of 1000 means of 40 expeontials follow nicely with the standard normal distribution. The observed variance of the mean is also about the same as the expected variance of the mean. From these results, we can tell the distribution is approximately normal.

Part 2: Basic Inferential Data Analysis

load the ToothGrowth data and perform some basic exploratory data analyses

Here I load the ToothGrowth data with basic exploratory data analyses using ggplot2(). I also use summary() function to provide some basic exploratory data analysis.

```
library(datasets)
data(ToothGrowth)
library(ggplot2)
library(reshape2)
ToothGrowth$supp<-as.factor(ToothGrowth$supp)
ToothGrowth$dose<-as.factor(ToothGrowth$dose)
melted_data<-melt(ToothGrowth,id.vars=c("supp","dose"))
#some ggplot examples are from http://www.cookbook-r.com/Graphs/Plotting_means_and_er
ror_bars_(ggplot2)/
library(Rmisc)#get the mean, sd information for plotting
melted_data.Summary <- summarySE(melted_data, measurevar="value", groupvars=c("supp",
"dose"))
  # Use 95% confidence intervals
   p.ci<-ggplot(melted_data.Summary, aes(x=dose, y=value, fill=supp)) +
    geom_bar(position=position_dodge(), stat="identity") +
    geom_errorbar(aes(ymin=value-ci, ymax=value+ci),
              width=.2,                    # Width of the error bars
              position=position_dodge(.9))+
     labs(title="Exploratory Analysis of ToothGrowth Data",x="Dose Level",y="Length")
+
     scale_fill_brewer(name="Supply Name")+
          theme_bw()
plot(p.ci)
```



In the Figure above, The general summary of the data is presented. At x-axis, Mean lengths and 95% confidence Intervals of each supply at each dose level are plotted.

Provide a basic summary of the data.

Explanation: The ToothGrowth dataset has 60 observations with three variables. Variable “length” responses to the length of the cell that reflects to the tooth growth of guinea pigs. The “supply” variable has 2 factors. The “dose” variable, which responses to the dose level of corresponding supplies.

```
summary(ToothGrowth)
```

```
##        len        supp     dose
##  Min.   : 4.20   OJ:30   0.5:20
##  1st Qu.:13.07   VC:30   1  :20
##  Median :19.25           2  :20
##  Mean   :18.81
##  3rd Qu.:25.27
##  Max.   :33.90
```

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)

Below are codes to run t tests at different dose levels. Welch t-test is used.

```
alpha<-0.05
nTestAlpha<-seq(1,3)/3*alpha
Dose.05<-subset(ToothGrowth,dose==0.5)
Dose.1<-subset(ToothGrowth,dose==1)
Dose.2<-subset(ToothGrowth,dose==2)
test.05<-t.test(len~supp,Dose.05)
test.1<-t.test(len~supp,Dose.1)
test.2<-t.test(len~supp,Dose.2)
sprintf("T test Result at Dose 0.5")
```

```
## [1] "T test Result at Dose 0.5"
```

```
test.05
```

```
##
##  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
```

```
sprintf("T test Result at Dose 1")
```

```
## [1] "T test Result at Dose 1"
```

```
test.1
```

```
##
##  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
```

```
sprintf("T test Result at Dose 2")
```

```
## [1] "T test Result at Dose 2"
```

```
test.2
```

```
##
##  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
```

State your conclusions and the assumptions needed for your conclusions.

My conclusions: At Dose level 0.5,1: The supply has a significant effect on the mean length. Lengths supplied by OJ are significantly longer than those supplied by TC. At Dose level 2, the supply doesn't have a significant effect on the mean length.