Statistical Inference Course Project

Mark Wan

11/14/2020

Overview

This report consists of 2 parts. The first part explores the CLT theorem with the exponential distribution. The second part conducts an inferential analysis on the ToothGrowth dataset.

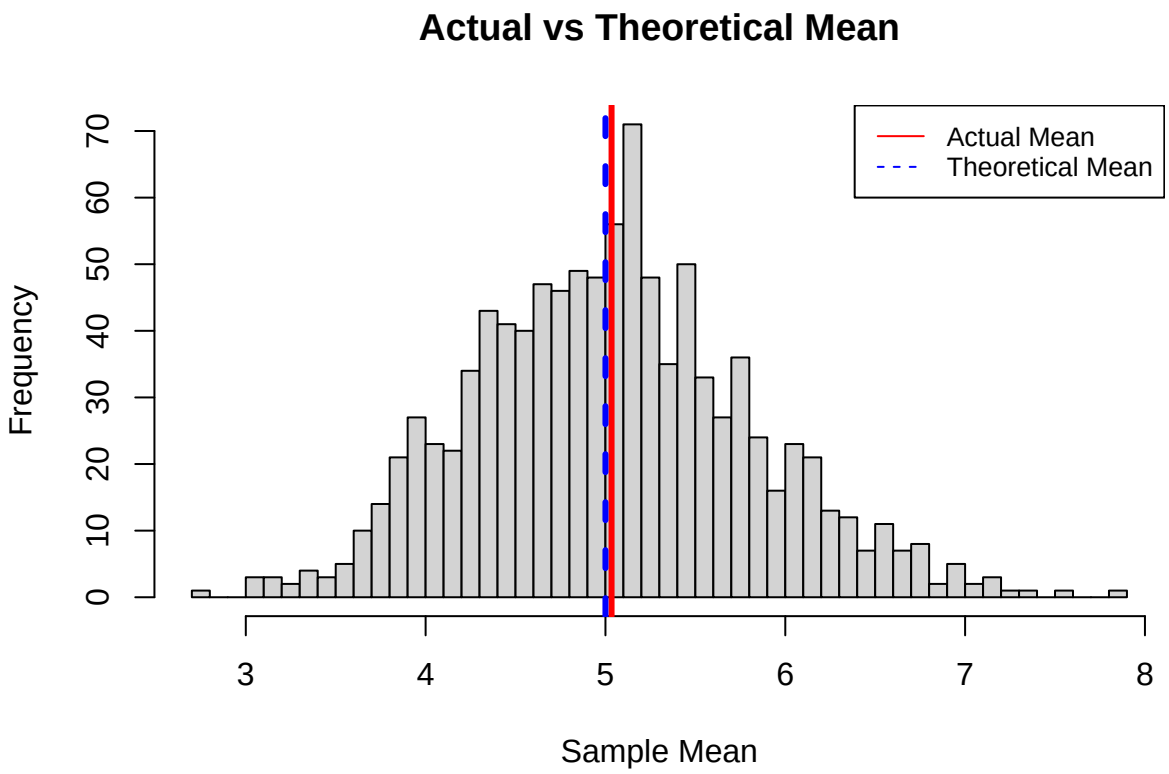
Part 1: Simulations

---

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

The following R code produces 1000 sets of 40 values generated from an exponential distribution with lambda=0.2. The code then generates a histogram of sample means and labels the actual mean against the theoretical mean for a direct comparison.

```
# Set reproducibility
set.seed(2020)
lambda <- 0.2
mu_theo <- 1/lambda
s <- 1/lambda
n <- 40
mns = NULL
for (i in 1 : 1000) mns = c(mns, mean(rexp(n,lambda)))
hist(mns,main="Actual vs Theoretical Mean", xlab="Sample Mean",breaks=40)
abline(v=mean(mns),col="red",lwd=3)
abline(v=mu_theo,col="blue",lwd=3,lty=2)
legend("topright", legend=c("Actual Mean", "Theoretical Mean"),
col=c("red", "blue"), lty=1:2, cex=0.8)
```



The graph above shows that the theoretical mean, 5 is a very good estimate of the actual mean 5.0339482.

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

of the distribution. The code below does a simple calculation to calculate the theoretical variances vs actual variances.

```
var_sample <- var(mns)
var_theo <- s^2/n
```

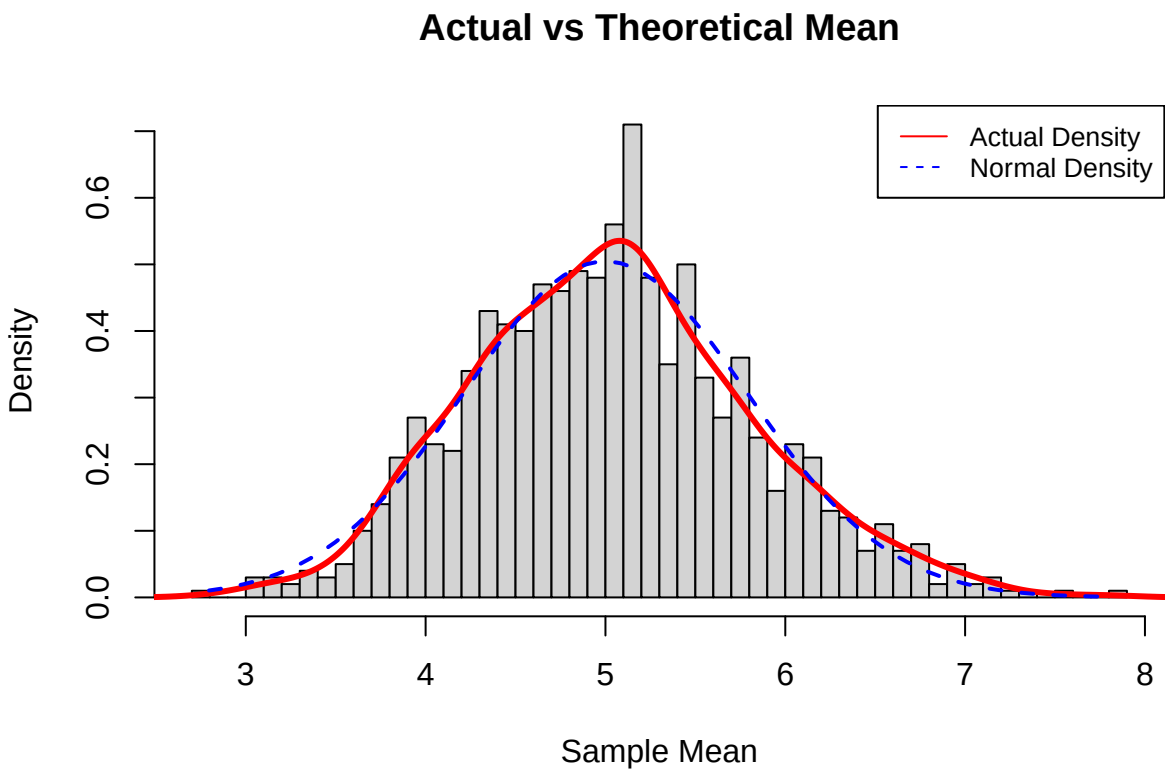
Attribute	Value
Theoretical Variance	0.625
Actual Variance	0.6070127
Theoretical Standard Deviation	0.7905694
Actual Standard Deviation	0.7791102

The table above shows again that the theoretical variance and standard deviation are good estimators of the actual values.

3. Show that the distribution is approximately normal.

The following R code compares the actual sample mean distribution to the normal distribution.

```
hist(mns,prob=TRUE,main="Actual vs Theoretical Mean", xlab="Sample Mean",breaks=40)
lines(density(mns), lwd=3, col="red")
x <- seq(min(mns), max(mns), length=2*n)
y <- dnorm(x, mean=1/lambda, sd=sqrt(((1/lambda)/sqrt(n))^2))
lines(x, y, pch=22, col="blue", lwd=2, lty = 2)
legend("topright", legend=c("Actual Density", "Normal Density"),
col=c("red", "blue"), lty=1:2, cex=0.8)
```



The graph above shows sample mean distribution can be very closely approximated wiht a normal distribution.

Part 2: Basic Inferential Data Analysis

The following R code loads the ToothGrowth dataset and executes a basic exploratory analysis of the data.
1. Load the ToothGrowth data and perform some basic exploratory data analyses

```
data("ToothGrowth")
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
```

```
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
```

```
for (name in names(ToothGrowth)){
print(name)
print(unique(ToothGrowth[,name]))
print("")
}
```

```
## [1] "len"
##  [1]  4.2 11.5  7.3  5.8  6.4 10.0 11.2  5.2  7.0 16.5 15.2 17.3 22.5 13.6 14.5
## [16] 18.8 15.5 23.6 18.5 33.9 25.5 26.4 32.5 26.7 21.5 23.3 29.5 17.6  9.7  8.2
## [31]  9.4 19.7 20.0 25.2 25.8 21.2 27.3 22.4 24.5 24.8 30.9 29.4 23.0
## [1] ""
## [1] "supp"
## [1] VC OJ
## Levels: OJ VC
## [1] ""
## [1] "dose"
## [1] 0.5 1.0 2.0
## [1] ""
```

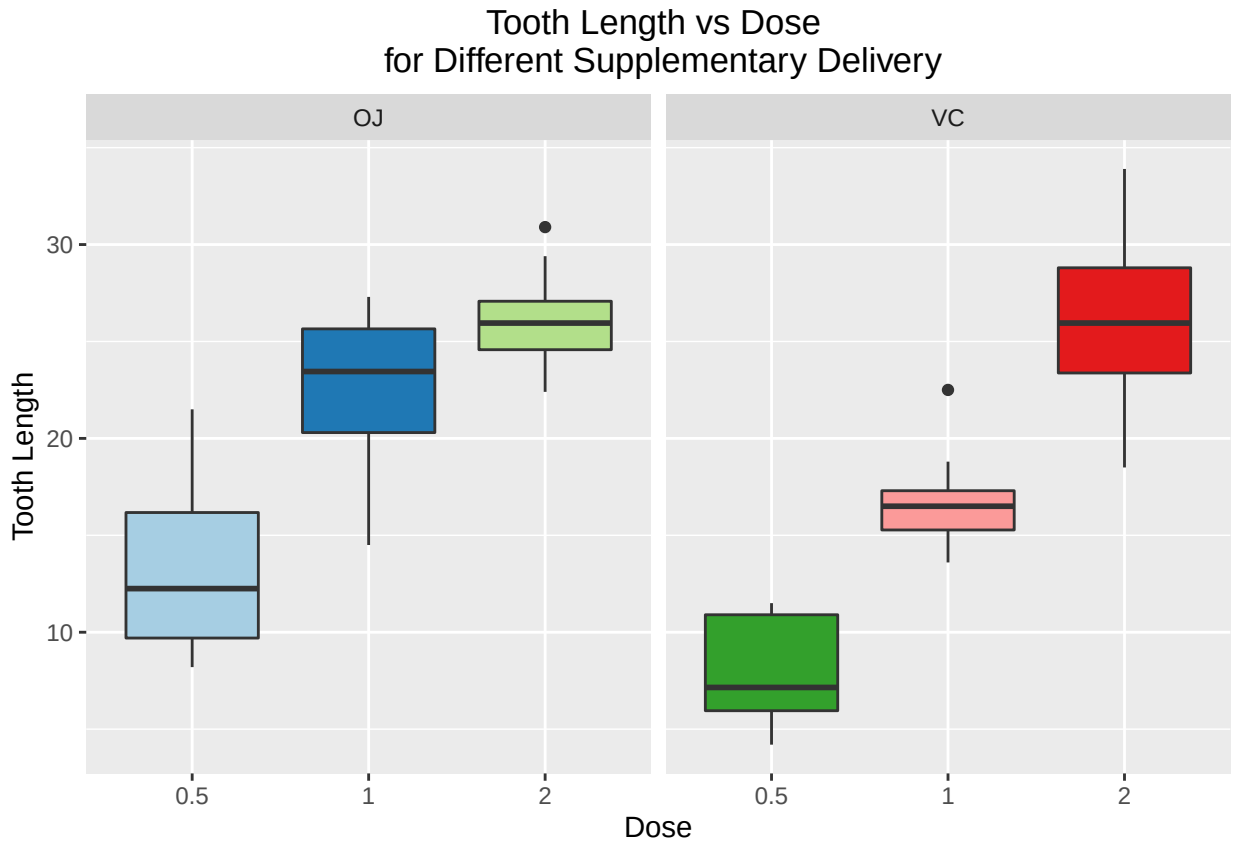
```
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
```

2. Provide a basic summary of the data.

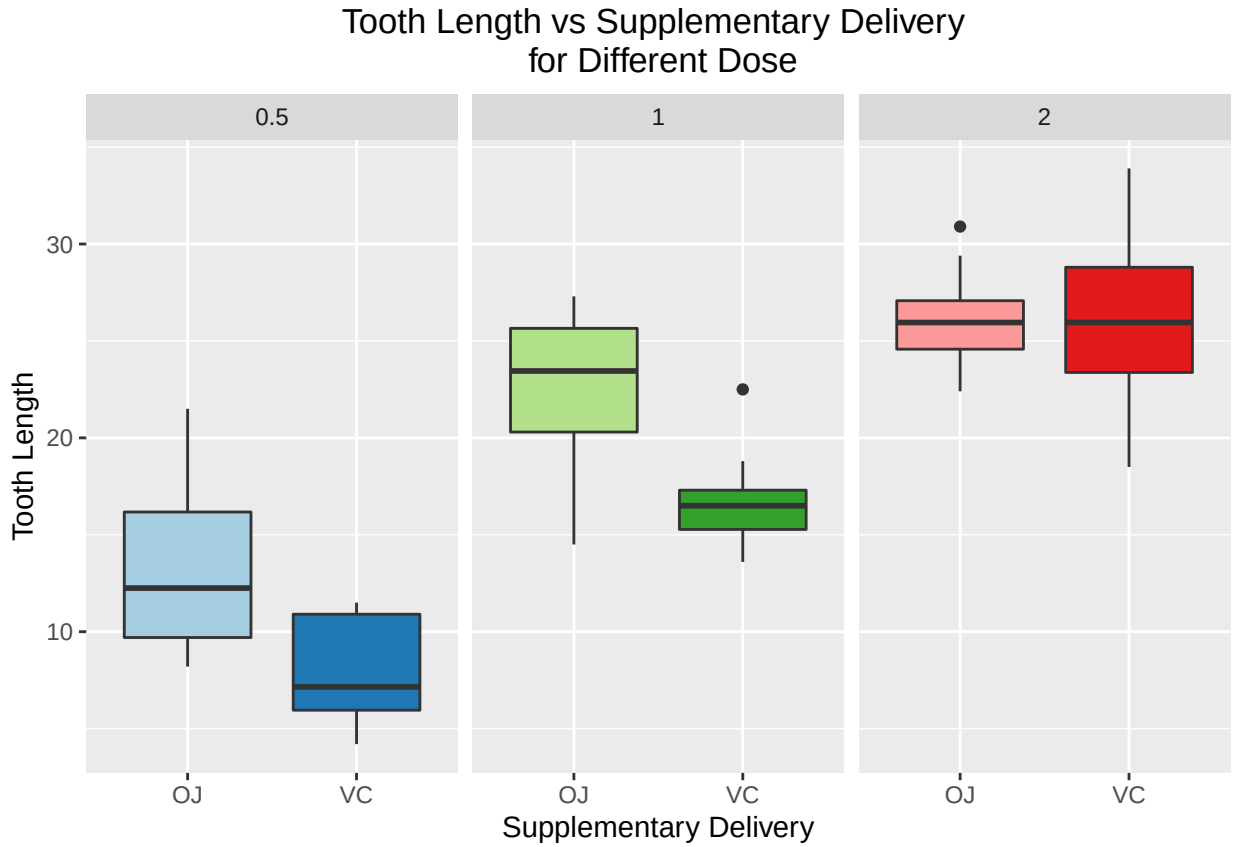
The following code chunk uses ggplot to produce facet plots of tooth length vs dose for each supplementary delivery, and of tooth length vs supplmentary delivery for each dose level.

```
library(ggplot2)
library(RColorBrewer)
g <- ggplot(ToothGrowth,aes(x=dose,y=len)) +
  geom_boxplot(fill=brewer.pal(length(unique(ToothGrowth$dose))*
                               length(unique(ToothGrowth$supp)),"Paired")) +
  facet_grid(.~supp) +
  labs(x="Dose",y="Tooth Length",
```

```
      title="Tooth Length vs Dose \n for Different Supplementary Delivery") +
  theme(plot.title = element_text(hjust = 0.5))
g
```




```
h <- ggplot(ToothGrowth,aes(x=supp,y=len)) +
  geom_boxplot(fill=brewer.pal(6,"Paired")) +
  facet_grid(.~dose) +
  labs(x="Supplementary Delivery",y="Tooth Length",
       title="Tooth Length vs Supplementary Delivery \n for Different Dose") +
  theme(plot.title = element_text(hjust = 0.5))
h
```



The 2 facet plots above show (1) sample averages of tooth length increased with dose for both OJ and VC supplementary deliveries, and (2) sample average of tooth length decreased with a change in supplementary delivery from OJ to VC for dose level 0.5 and 1. However, the sample average between the 2 deliveries did not differ much for dose level 2. These observations are only descriptions of the data. In order to make inferences about the relationship between tooth length and the 2 variables, hypothesis testing is required.

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)

The following 3 code chunks conduct unpaired t-tests for each possible pair of dose levels.

```
# Test Dose 0.5 vs 1
dose_0.5_1 <- subset(ToothGrowth, dose %in% c(0.5,1))
t.test(len~dose,data=dose_0.5_1)
```

```
##
##  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
```

```
# Test Dose 0.5 vs 2
dose_0.5_2 <- subset(ToothGrowth, dose %in% c(0.5,2))
t.test(len~dose,data=dose_0.5_2)
```

```
##
##  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
```

```
# Test Dose 1 vs 2
dose_1_2 <- subset(ToothGrowth, dose %in% c(1,2))
t.test(len~dose,data=dose_1_2)
```

```
##
##  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
```

```
# Test Supplementary Delivery
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
```

For a 5% significance level, the difference in average between VC and OJ was not significant, since the p-value was greater than 5%. That is, the data implies that supplmentary delivery has no impact on tooth length.

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

In conclusion, based on a 5% significance level, the data seems to indicate that tooth length increases with does level. In contrasts, there were no significant association observed between tooth length and supplementary delivery. These observations were based on the assumptions that: 1. The distribution of the sample means follows the CLT theorem, and can be approximated by the t-distribution, 2. The observations are not biased and are a true representation of the underlying population of tooth lengths.