Statistical Inference: Peer Assessment

Part 1: Simulation Exercise Instructions

Objective: we aimed to investigate the exponential distribution in R and to compare it with the Central Limit Theorem. We investigated the distribution of averages of 40 exponentials with a thousand simulations.

We create a dataframe with 1000 means of 40 exponentials.

```
n <- 40; lambda <- 0.2; nosim <- 1000;
# Create a matrix with a size n*nosim
m <- matrix(data = rexp(nosim * n, lambda), nrow = nosim)
# Calculate mean of each row
df <- data.frame(x = apply(m, 1, mean))
str(df)
```

```
## 'data.frame':    1000 obs. of  1 variable:
##  $ x: num  5.07 4.31 5.17 4.68 3.98 ...
```

```
head(df)
```

```
##          x
## 1 5.071309
## 2 4.309998
## 3 5.166457
## 4 4.681698
## 5 3.984801
## 6 5.162000
```

We understood the true mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Then, we can calculate the theoretical standard deviation,and the theoretical variance, show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

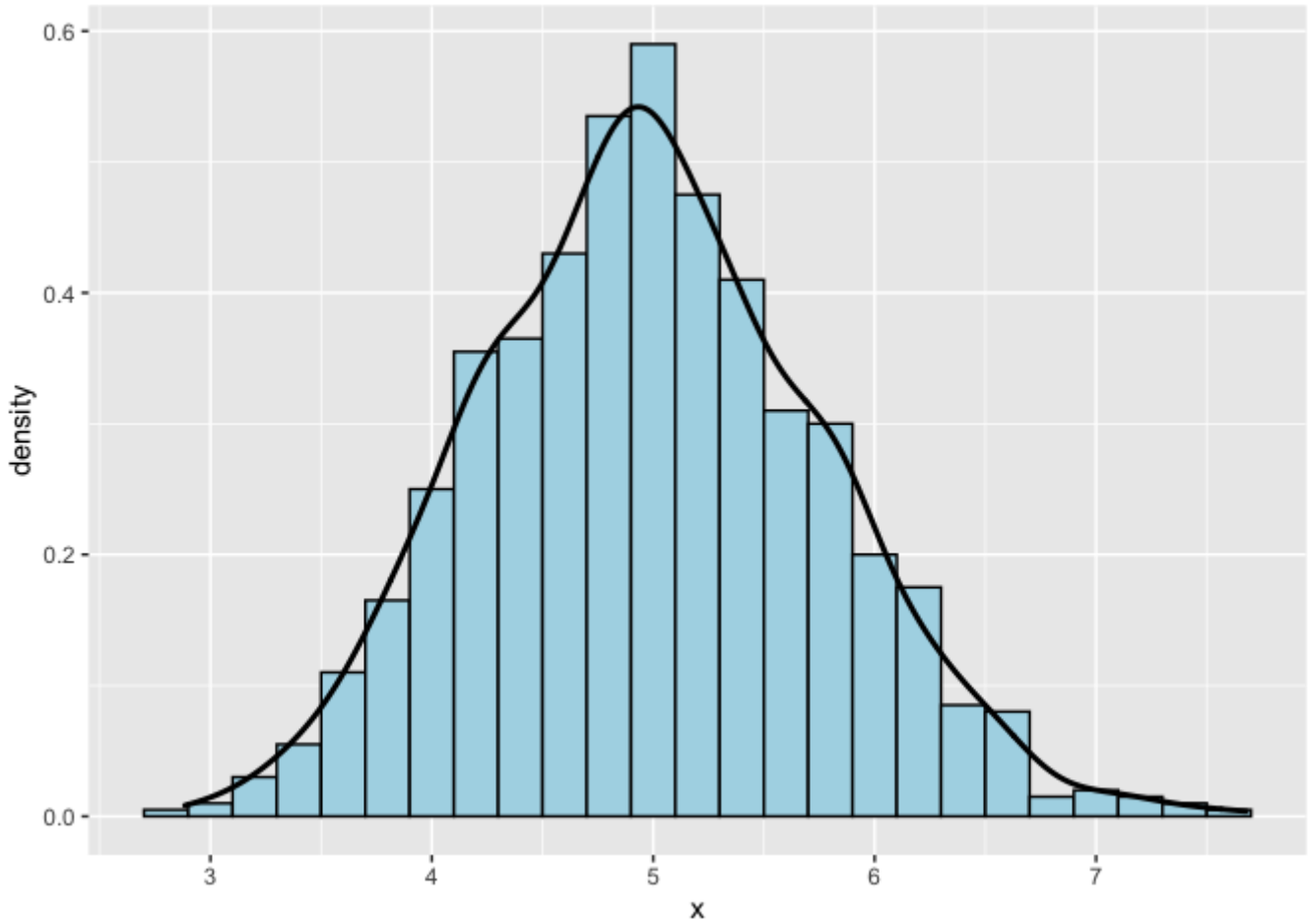
```
a <- round(c(theoretical.mean = 1/lambda, sample.mean = mean(df$x),
             theoretical.sd = (1/lambda)/sqrt(n), sample.sd = sd(df$x),
             theoretical.variance = (1/lambda)/sqrt(n), sample.variance = ((1/lambd
a)/sqrt(n))^2)
           ,3)
a
```

```
##     theoretical.mean          sample.mean       theoretical.sd
##                5.000                4.984                0.791
##            sample.sd theoretical.variance      sample.variance
##                0.766                0.791                0.625
```

The results showed that there are a few difference between theoretical values and sample values.

Finally, we wanted to show that the distribution of averages of 40 exponentials with a thousand simulations is approximately normal.

```
require(ggplot2)
g1 <- ggplot(data = df, aes(x = x))
g1 <- g1 + geom_histogram(aes(y = ..density..), fill = "lightblue", binwidth=0.2, col
our = "black")
g1 <- g1 + geom_density(size = 1, colour = "black")
g1
```



Part 2: Basic Inferential Data Analysis Instructions

Objective: we wanted to explore the ToothGrowth data, and used confidence intervals and hypothesis tests to compare tooth growth by supp and dose.

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

According to the package description, the data is about 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).

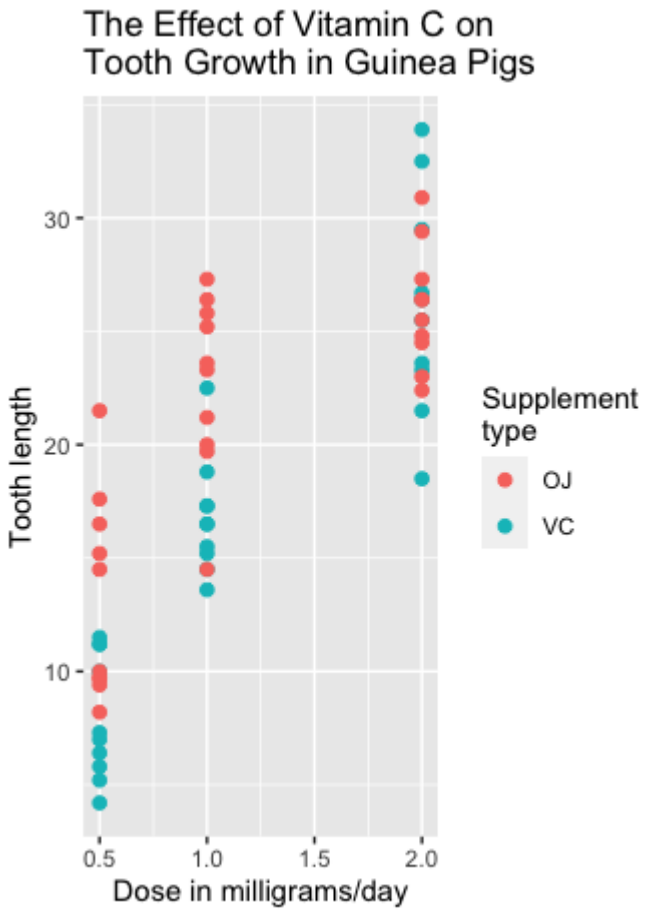
The data contain a data frame with 60 observations on 3 variables.

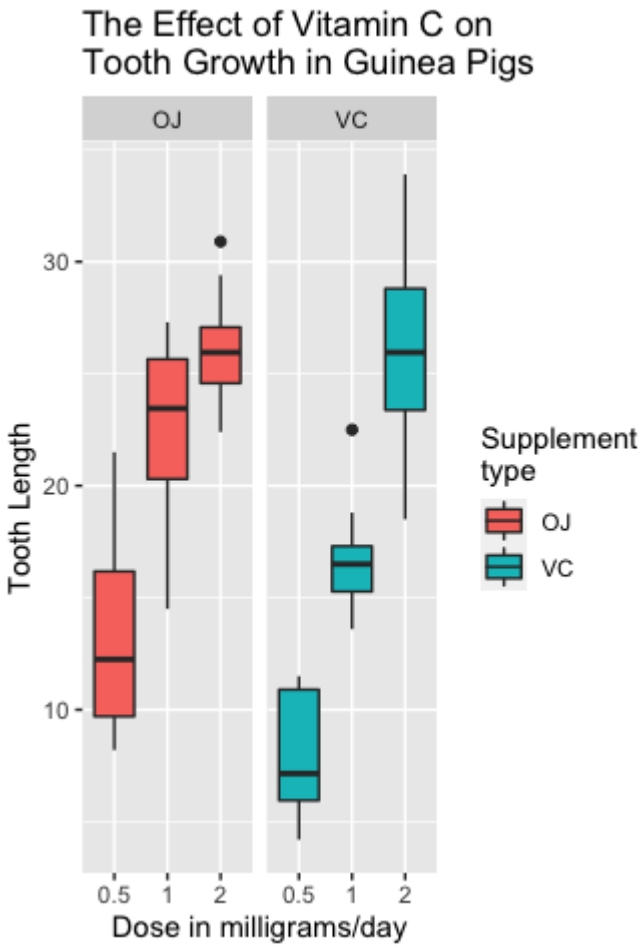
1. [len] numeric Tooth length
2. [supp] factor Supplement type (VC or OJ).
3. [dose] numeric Dose in milligrams/day:

```
# Load data
library(datasets)
data(ToothGrowth)

# Preview data
head(ToothGrowth, n = 5)
```

```
##    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
```

```
# Preview data by a scatterplot
library(ggplot2)
g2 <- ggplot(ToothGrowth, aes(x=dose, y=len, shape=supp, color=supp, size=supp)) +
    geom_point(size=2, shape=19) +
    labs(x = "Dose in milligrams/day",
         y = "Tooth length",
         caption = '') +
    ggtitle("The Effect of Vitamin C on
Tooth Growth in Guinea Pigs") +
    labs(color="Supplement
type")

# Preview data by a boxplot
g3 <- ggplot(ToothGrowth, aes(x = factor(dose), y = len, fill = supp)) +
    geom_boxplot() +
    facet_grid(.~supp) +
    ggtitle("The Effect of Vitamin C on
Tooth Growth in Guinea Pigs") +
    scale_x_discrete("Dose in milligrams/day") +
    scale_y_continuous("Tooth Length") +
    scale_fill_discrete(name="Supplement
type")
require(gridExtra)
grid.arrange(g2, g3, ncol=2)
```





2.2 Provide a basic summary of the data.

```
# Preview any missing values
any (is.na(ToothGrowth))
```

```
## [1] FALSE
```

```
# Preview the dimension
dim(ToothGrowth)
```

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

```
# Preview an internal structure
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 ...
```

```
# Preview a Five-number summary
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
```

```
# Preview data in terms of table
table(ToothGrowth$supp,ToothGrowth$dose, useNA="ifany")
```

```
##
##      0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

2.3 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)

Since the sample size for each groups (OJ, VC) was small (n≤30) and population standard deviations were unknown, the T confidence interval was used.

Question: Is there a statistical difference in the tooth growth by different delievery method?

H0:$\mu 1 - \mu 2 = 0$ (there was no difference between two supplements (OJ, VC) in terms of tooth growth)

Ha:$\mu 1 - \mu 2 > 0$ (there was a significant difference two supplements (OJ, VC) in terms of tooth growth)

Degree of significance, α% = 5%; level of significance α = 0.05

To compute the value of test statistic:

```
# Code: the independent t test with variance equal or unequal assumptions.
equal_t <- t.test(len~supp, paired = FALSE, var.equal = TRUE, alt = "greater", data =
ToothGrowth, conf.level = 0.95)
unequal_t <- t.test(len~supp, paired = FALSE, var.equal = FALSE, alt = "greater", dat
a = ToothGrowth, conf.level = 0.95)
# Note: alternative is a character string describing the alternative hypothesis.

# Code: Permutation tests to test the null hypothesis of no difference between treatm
ent groups.
permuatation <- function(y, group, g1_name, g2_name) {
    # to calculate the mean of two groups, notice here group is a variable.
    testStat <- function(w, g) mean(w[g == g1_name]) - mean(w[g == g2_name])
    # To calculate the difference of mean between group OJ and VC.
    observedStat <- testStat(y, group)
    #Permutation: change the group label of subdata randomly for 10000 times, i.e., i
f the group if irrevelant to the y.
    permutations <- sapply(1 : 10000, function(i) testStat(y, sample(group)))
    permutate_t <- sum(permutations > observedStat)/10000
    permutate_t
}
permutate_t <- permuatation(ToothGrowth$len, as.character(ToothGrowth$supp), "OJ","V
C")

result <- data.frame(
    "p-value" = round(c(equal_t$p.value, unequal_t$p.value, permutate_t),3),
    "t statistic" = round(c(equal_t$statistic, unequal_t$statistic,  NA), 3)
)
row.names(result) <- c("var.equal", "var.unequal", "permutation test")
result
```

```
##                  p.value t.statistic
## var.equal          0.030       1.915
## var.unequal        0.030       1.915
## permutation test   0.033          NA
```

Answer: Assuming that the null hypothesis was true, all p-values were less than 0.05. It meant that there were less than a 5% probability that the null hypothesis was correct. In other words, either the null hypothesis was false or something unlikely had occurred. The result was then said to be statistically significant as it allowed us to reject the null hypothesis, which said no difference between then length and the methods of delivery.

Question (Continue): Different levels of dose by VC and OJ might bring different impacts on the tooth length. So, we continued to investigate them.

H0:$\mu 1 - \mu 2 = 0$ (there was no difference between two supplements (OJ, VC) with various dosages (0.5, 1, 2mg) in terms of tooth growth)

Ha:$\mu 1 - \mu 2 > 0$ (there was a significant difference between two supplements (OJ, VC) with various dosages (0.5, 1, 2mg) in terms of tooth growth)

Degree of significance, α% = 5%; level of significance α = 0.05

To compute the value of test statistic:

```
# Code: the independent t test with variance equal or unequal assumptions.
t0.5 <- ToothGrowth[which (ToothGrowth$dose == 0.5),]
equal_t0.5 <- t.test(len~supp, paired = FALSE, var.equal = TRUE, alt = "greater", dat
a = t0.5, conf.level = 0.95)
unequal_t0.5 <- t.test(len~supp, paired = FALSE, var.equal = FALSE, alt = "greater",
 data = t0.5, conf.level = 0.95)
permutate_t0.5 <- permuatation(t0.5$len, as.character(t0.5$supp), "OJ","VC")


t1 <- ToothGrowth[which (ToothGrowth$dose == 1),]
equal_t1 <- t.test(len~supp, paired = FALSE, var.equal = TRUE, alt = "greater", data
 = t1, conf.level = 0.95)
unequal_t1 <- t.test(len~supp, paired = FALSE, var.equal = FALSE, alt = "greater", da
ta = t1, conf.level = 0.95)
permutate_t1 <- permuatation(t1$len, as.character(t1$supp), "OJ","VC")


t2 <- ToothGrowth[which (ToothGrowth$dose == 2),]
equal_t2 <- t.test(len~supp, paired = FALSE, var.equal = TRUE, alt = "greater", data
 = t2, conf.level = 0.95)
unequal_t2 <- t.test(len~supp, paired = FALSE, var.equal = FALSE, alt = "greater", da
ta = t2, conf.level = 0.95)
permutate_t2 <- permuatation(t2$len, as.character(t2$supp), "OJ","VC")
# Note: alternative is a character string describing the alternative hypothesis.

result <- data.frame(
    "p-value" = round(c(equal_t0.5$p.value, unequal_t0.5$p.value, permutate_t0.5,
                  equal_t1$p.value, unequal_t1$p.value, permutate_t1,
                  equal_t2$p.value, unequal_t2$p.value, permutate_t2), 3),
    "t statistic" = round(c(equal_t0.5$statistic, unequal_t0.5$statistic, NA,
                     equal_t1$statistic, unequal_t2$statistic, NA,
                     equal_t1$statistic, unequal_t2$statistic, NA),3)
    )

row.names(result) <- c("Dosage 0.5mg (var.equal)", "Dosage 0.5mg (var.unequal)", "Dos
age 0.5mg (permutation test)",
                       "Dosage 1mg  (var.equal)", "Dosage 1mg (var.unequal)", "Dosage
1mg (permutation test)",
                       "Dosage 2mg (var.equal)", "Dosage 2mg  (var.unequal)", "Dosage
2mg (permutation test)")
result
```

```
##                                p.value t.statistic
## Dosage 0.5mg (var.equal)         0.003       3.170
## Dosage 0.5mg (var.unequal)       0.003       3.170
## Dosage 0.5mg (permutation test)  0.002          NA
## Dosage 1mg  (var.equal)          0.000       4.033
## Dosage 1mg (var.unequal)         0.001      -0.046
## Dosage 1mg (permutation test)    0.001          NA
## Dosage 2mg (var.equal)           0.518       4.033
## Dosage 2mg  (var.unequal)        0.518      -0.046
## Dosage 2mg (permutation test)    0.514          NA
```

Answer: Assuming that the null hypothesis was true, the p-values of dosage with 0.5mg or 1mg were less than 0.05 (≤ 0.05). It indicated strong support against the null hypothesis, as there were less than a 5% probability. Therefore, we rejected the null hypothesis, which said the (0.5mg or 1mg) dosage made no

difference.

However, the p-values of dosage with 2mg by either VC or OJ were higher than 0.05 (> 0.05). They were not statistically significant and indicated strong support for the null hypothesis. This meant that we retained the null hypothesis, which said the (2mg) dosage make no difference.

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

In our study, we assumed there was no difference between the supplements (VC and OJ) and the tooth length. Our results did not support for our hypothesis. We may conduct further investigation about which supplement was better in terms of the tooth length.

Furthermore, our results did not support that a low-dosage (0.5 and 1mg) had no impact on the tooth length. However, our results supported the hypothesis that (2mg) a high-dosage and the tooth length made no difference,