Statistical Inference Course Project

Part 1: Simulation Exercise

In this project I 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. Note that I will need to do a thousand simulations.

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

```
# set seed
set.seed(1234)
# set lambda
lambda <- 0.2
# set number of simulations
n <- 1000
# set number of exponentials
m <- 40
# simulate exponential distribution
sim <- replicate(n, rexp(m, lambda))
# calculate mean of each simulation
means <- apply(sim, 2, mean)
```

```
# calculate mean of means
mean(means)
```

```
## [1] 4.974239
```

```
# calculate theoretical mean
1 / lambda
```

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

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

```
# calculate variance of means
var(means)
```

```
## [1] 0.5706551
```

```
# calculate theoretical variance
(1 / lambda^2) / m
```

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

```
# Comparison
var(means)-(1 / lambda^2) / m
```

```
## [1] -0.05434495
```

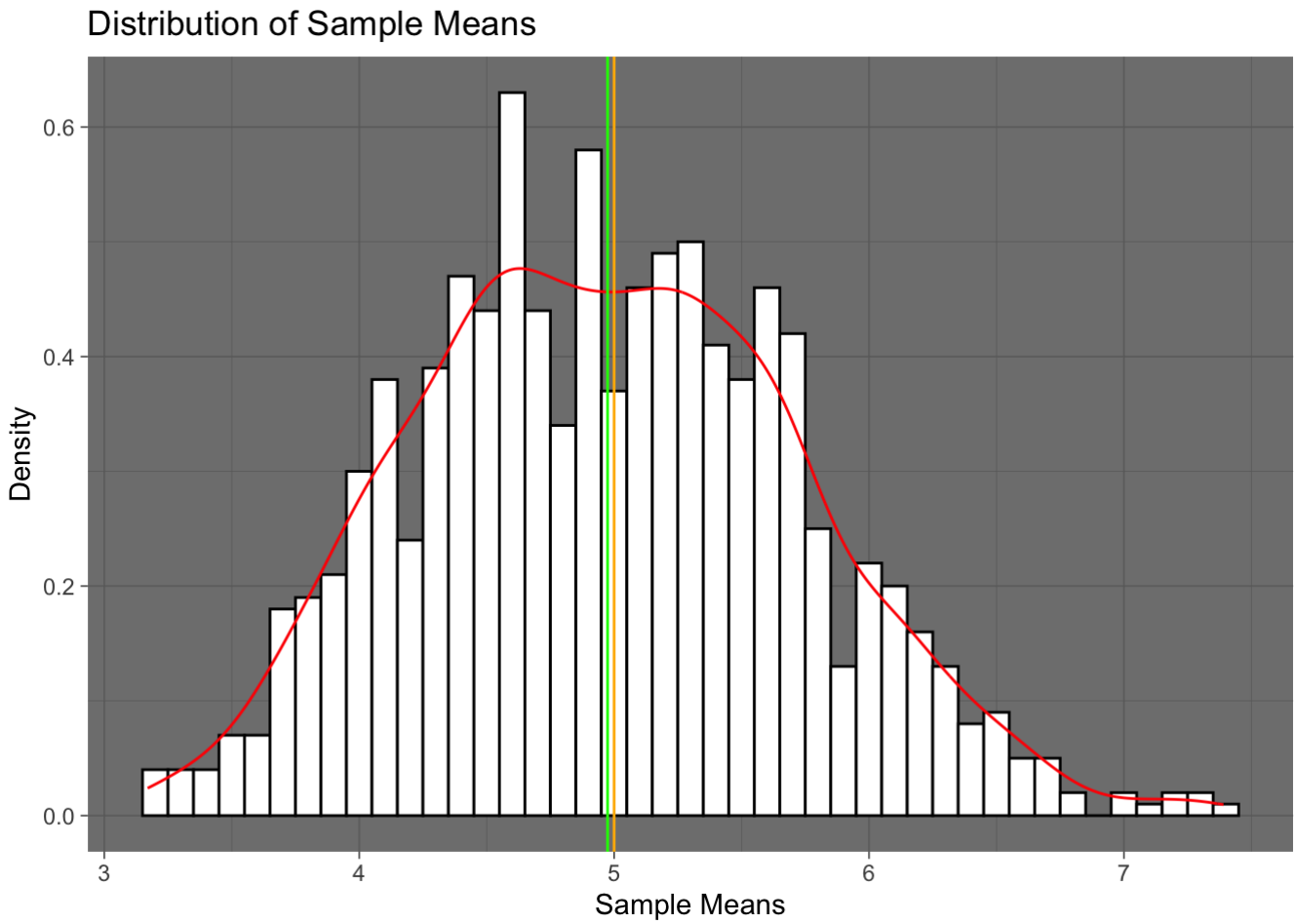
Comparing the two values of the **Variance** we see that the values are **very close**.

3. Show that the distribution is approximately normal.

```
# plot histogram of sample distribution and showing both the mean and theoretical mea
n
library(ggplot2)
# set the ggplot theme
theme_set(theme_dark())
ggplot(data.frame(means), aes(x = means)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.1, colour = "black", fill = "#fff
fff") +
  geom_density(colour = "red") +
  geom_vline(aes(xintercept = mean(means)), colour = "#04ff00", linetype = "solid", s
ize = 0.5) +
  geom_vline(aes(xintercept = 1 / lambda), colour = "#ffb700", linetype = "solid", si
ze = 0.5) +
  ggtitle("Distribution of Sample Means") +
  xlab("Sample Means") +
  ylab("Density")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## ℹ Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



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.

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

```
# load the ToothGrowth data
data(ToothGrowth)

# Change does column to factor data type
ToothGrowth$dose <- as.factor(ToothGrowth$dose)

# check the structure 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 ...
```

```
# check the first 6 rows of the data
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
```

```
# check the last 6 rows of the data
tail(ToothGrowth)
```

```
##     len supp dose
## 55 24.8   OJ    2
## 56 30.9   OJ    2
## 57 26.4   OJ    2
## 58 27.3   OJ    2
## 59 29.4   OJ    2
## 60 23.0   OJ    2
```

```
# check the dimensions of the data
dim(ToothGrowth)
```

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

```
# check the names of the data
names(ToothGrowth)
```

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

```
# check the class of the data
class(ToothGrowth)
```

```
## [1] "data.frame"
```

2. Provide a basic summary of the data.

```
# check the summary of the data
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
```

```
# check the number of observations in each group
aggregate(len ~ supp + dose, data = ToothGrowth, FUN = length)
```

```
##   supp dose len
## 1   OJ  0.5  10
## 2   VC  0.5  10
## 3   OJ    1  10
## 4   VC    1  10
## 5   OJ    2  10
## 6   VC    2  10
```


```
# check the number of observations in each group
aggregate(len ~ supp + dose, data = ToothGrowth, FUN = length)
```

```
##   supp dose len
## 1   OJ  0.5  10
## 2   VC  0.5  10
## 3   OJ    1  10
## 4   VC    1  10
## 5   OJ    2  10
## 6   VC    2  10
```


```
# check the number of observations in each group
table(ToothGrowth$supp, ToothGrowth$dose)
```

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

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)

```
# check for group differences due to different supplement type
# assuming unequal variances between the two groups
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 between group OJ and group VC 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
```

```r
# Create three sub-groups per dose level pairs in order to check for group difference
s.
# assuming unequal variances between the two groups
ToothGrowth.doses_0.5_1.0 <- subset(ToothGrowth, dose %in% c(0.5, 1.0))
ToothGrowth.doses_0.5_2.0 <- subset(ToothGrowth, dose %in% c(0.5, 2.0))
ToothGrowth.doses_1.0_2.0 <- subset(ToothGrowth, dose %in% c(1.0, 2.0))

# Check for group differences due to different dose levels of (0.5, 1.0). Assume uneq
ual variances between the two groups.
t.test(len ~ dose, data = ToothGrowth.doses_0.5_1.0)
```

```
##
##  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 between group 0.5 and group 1 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
# Check for group differences due to different dose levels of (0.5, 2.0). Assume uneq
ual variances between the two groups.
t.test(len ~ dose, data = ToothGrowth.doses_0.5_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 between group 0.5 and group 2 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
# Check for group differences due to different dose levels of (1.0, 2.0). Assume uneq
ual variances between the two groups.
t.test(len ~ dose, data = ToothGrowth.doses_1.0_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 between group 1 and group 2 is no
t 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.

Conclusion:

- ***The p-value is 0.06, and the confidence interval contains zero. This indicates that we can not reject the null hypothesis that the different supplement types have no effect on tooth length.***
- ***For all three of the above t-tests, the resulting p-value is less than 0.5 and the confidence intervals do not contain zero. Thus, we reject the null hypothesis, and establish that increasing the dose level leads to an increase in tooth length.***

Assumptions:

1. **Random Assignment:** The experiment involved the random assignment of guinea pigs to different dose level categories and supplement types. This randomization aimed to control for potential confounders that might influence the outcome.
2. **Representativeness of Sample:** It is assumed that the members of the sample population, comprising the 60 guinea pigs, are representative of the entire population of guinea pigs. This assumption allows for the generalization of the experiment's results to the broader population.
3. **Variances in t-tests:** For the t-tests conducted, it is assumed that the variances are different between the two groups being compared. This assumption is considered less stringent than assuming equal variances and is made to accommodate potential differences in variability between the groups.