Statistical Inference Project

Jorge David Guzman

statistical-inference

The project consists of two parts.

A simulation exercise. Basic inferential data analysis.

Part 1 Simulation Exercise Instructions

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.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should

Show the sample mean and compare it to the theoretical mean of the distribution. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution. Show that the distribution is approximately normal. In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

As a motivating example, compare the distribution of 1000 random uniforms

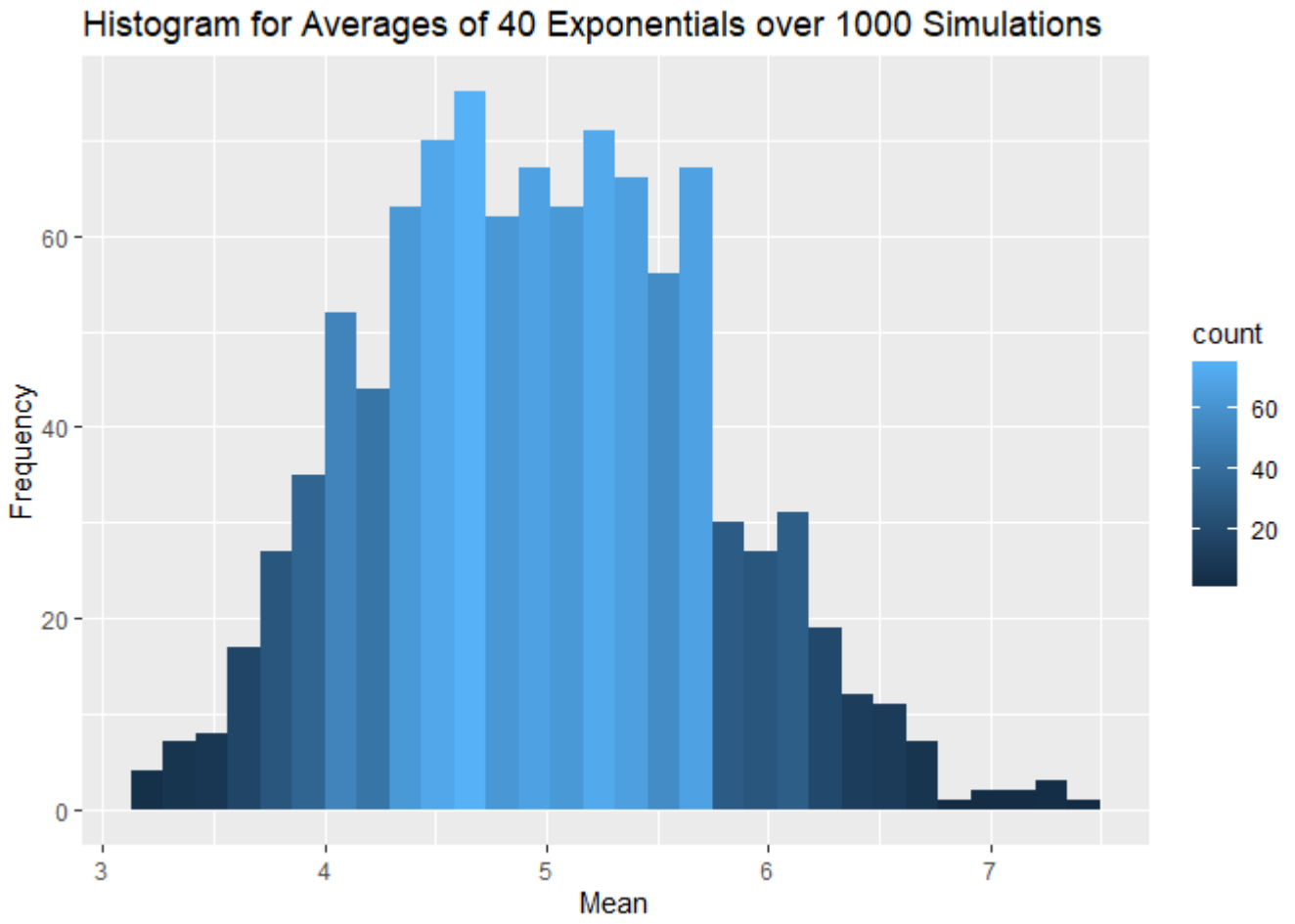
```
library(ggplot2)

lambda <- 0.2
set.seed(1234)
n <- 40

population <- data.frame(x=sapply(1:1000, function(x) {mean(rexp(n, lambda))}))

hist.pop <- ggplot(population, aes(x=x)) +
  geom_histogram(aes(y=..count.., fill=..count..)) +
  labs(title="Histogram for Averages of 40 Exponentials over 1000 Simulations", y="Frequency", x
="Mean")
hist.pop
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Comparison of mean and variance

```
table <- cbind(mean(population$x),  1/lambda, 100*((mean(population$x))-(1/lambda))/(1/lambda) )
colnames(table) <- c("sample", "theoretical", "%dif")
table <- rbind(table, c(var(population$x), ((1/lambda)^2)/n,
            100*(var(population$x)-((1/lambda)^2)/n)/((1/lambda)^2)/n))
rownames(table) <- c("Median", "Variance")
table
```

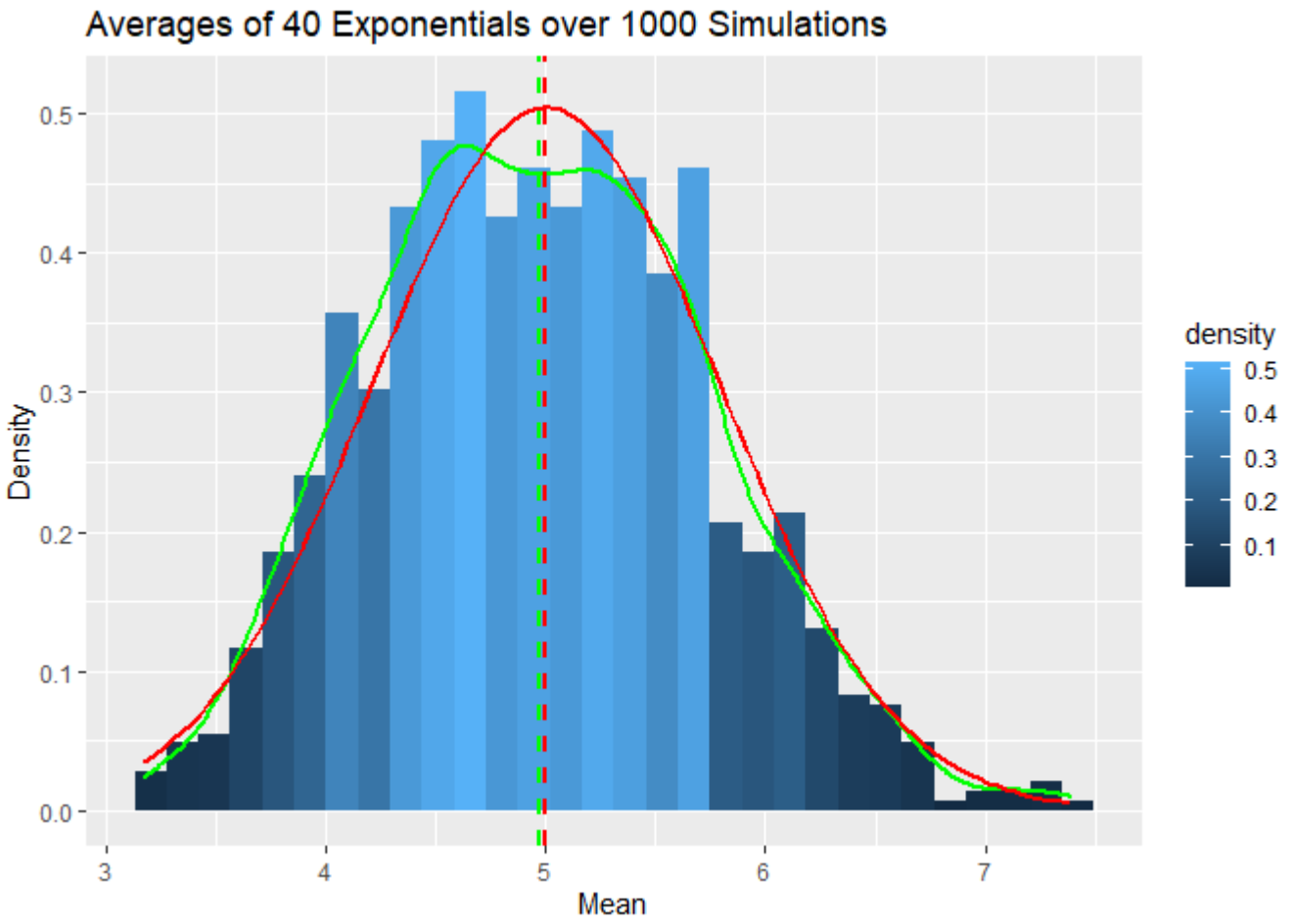
```
##             sample theoretical         %dif
## Median   4.9742388       5.000 -0.515224575
## Variance 0.5706551       0.625 -0.005434495
```

The difference of mean and variance is less than 1 percent between the observed and the theoretical value

Distribution is approximately normal.

```
ggplot(population, aes(x=x)) +
     geom_histogram(aes(y=..density.., fill=..density..)) +
     labs(title="Averages of 40 Exponentials over 1000 Simulations", y="Density", x="Mean") +
     geom_density(colour="green", size = 1) +
     geom_vline(xintercept=mean(population$x), colour="green", linetype="dashed", size = 1) +
     stat_function(fun=dnorm,args=list( mean=1/lambda, sd=sqrt(((1/lambda)^2)/n)),color = "red",
size = 1) +
     geom_vline(xintercept=1/lambda, colour="red", linetype="dashed", size = 1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Part 2 Basic Inferential Data Analysis Instructions

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

- Load the ToothGrowth data and perform some basic exploratory data analyses
- Provide a basic summary of the data.
- 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)

State your conclusions and the assumptions needed for your conclusions.

Load the ToothGrowth data and perform some basic exploratory data analyses

```
library(datasets)
data(ToothGrowth)
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 ...
```

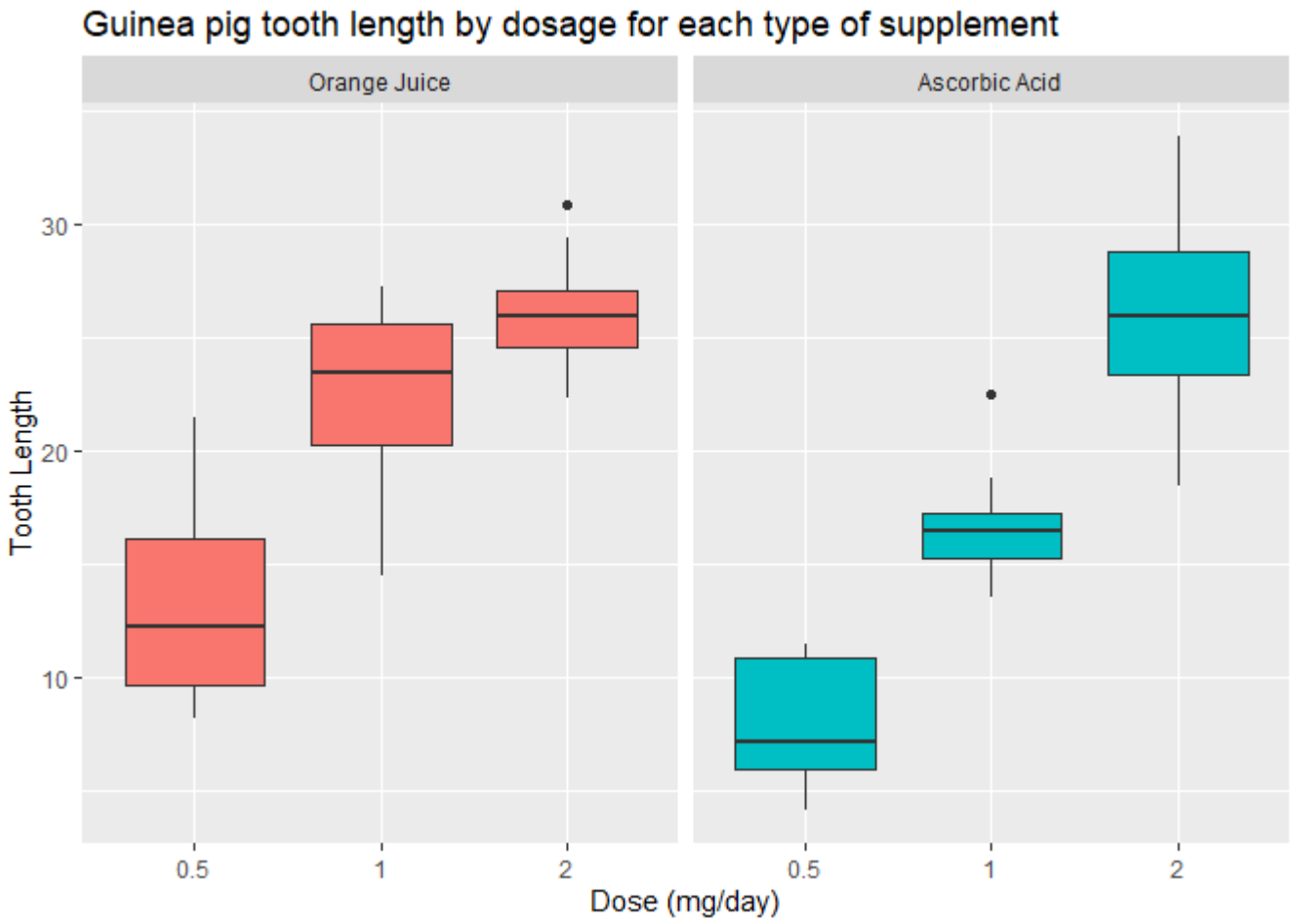
```
head(ToothGrowth)
```

```
##    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
```

```
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
```

```
t = ToothGrowth
levels(t$supp) <- c("Orange Juice", "Ascorbic Acid")
ggplot(t, aes(x=factor(dose), y=len)) +
     facet_grid(.~supp) +
     geom_boxplot(aes(fill = supp), show_guide = FALSE) +
     labs(title="Guinea pig tooth length by dosage for each type of supplement",
          x="Dose (mg/day)",
          y="Tooth Length")
```



The box plots show that increasing the dose increases tooth growth regardless of the supplement. When the dose is 0.5 or 1 mg, orange juice is more effective than ascorbic acid. if the dose is 2.0 milligrams per day, ascorbic acid is slightly more effective despite showing a greater dispersion.

```
table2 <- data.frame(mg_day=numeric(), p.value=numeric(),int1=numeric(),int2=numeric(), response
=character(),      stringsAsFactors = FALSE)
x <- 1
for (n in c(0.5,1.0,2.0)) {
  table2[x,1] <- n
  table2[x,2] <- t.test(len ~ supp, data = subset(t, t$dose ==n))$p.value
  table2[x,3] <- t.test(len ~ supp, data = subset(t, t$dose ==n))$conf.int[1]
  table2[x,4] <- t.test(len ~ supp, data = subset(t, t$dose ==n))$conf.int[2]
  table2[x,5] <- if(table2[x,2] > 0.05 && table2[x,3] < 0 && table2[x,4] >0){"No reject"} else {
"Reject"}
  x<- x+1
}
table2
```

```
##   mg_day     p.value      int1     int2  response
## 1    0.5 0.006358607  1.719057 8.780943    Reject
## 2    1.0 0.001038376  2.802148 9.057852    Reject
## 3    2.0 0.963851589 -3.798070 3.638070 No reject
```

Conclusion

Based on the p-values and the confidence intervals for the doses of 0.5 and 1 mg per day, we can reject the null hypothesis and say that orange juice is more effective than the ascorbic acid, but for the dose of 2 mg per day the null hypothesis cannot be rejected

assumptions

- The length of the teeth has a normal distribution
- There are no other factors that affect tooth growth