Course Project

Sergio Iván Arroyo Giles

Part 1: Simulation Exercise

Overview

Recall what the **Central Limit Theorem (CLT)** states:

Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample of size n- that is, a sequence of independent and identically distributed random variables drawn from a distribution with mean μ_X and variance σ_X^2. If you take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed. mathematical terms, for n (size of the sample) sufficiently large we have

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu_X, \sigma_X^2/n\right).$$

Here we intend to show how the sample mean of exponential observations satisfies the CLT. That is, how the distribution of sample mean have a very *gaussian* style via simulation of $1,000$ samples of size 40.

Simulations

We denote and use the following variables:

- `n`: Number of simulations to be run
- `S`: Size of each sample
- `lambda`: The rate parameter of our exponential distribution

```
#Parameters of the simulation
n <- 1000
S <- 40
lambda <- 0.2
```

Now, we create a matrix of size `n`×`S` where each of the columns are a simulation of size 40 and then we compute the mean of each of these columns.

```
#Simulations
dataMatrix <- matrix(rexp(S*n, lambda), nrow = S)
sampleMeans <- apply(dataMatrix, 2, mean)
```

Sample Mean versus Theoretical Mean

Therefore, if we perform the sample mean of our observations we get an approximation of the population mean.

```
theoretical.mean <- 1/lambda
sample.mean <- mean(sampleMeans)
cbind(theoretical.mean, sample.mean)
```

```
##      theoretical.mean sample.mean
## [1,]                5    5.024035
```

Sample Variance versus Theoretical Variance

Also, the same relationship works with the sample and population variance

```
theoretical.variance <- ((1/lambda)^2)
sample.variance <- var(sampleMeans)*S
cbind(theoretical.variance, sample.variance)
```

```
##      theoretical.variance sample.variance
## [1,]                   25        25.05991
```

Distribution

To show the normal distribution, we create the probability histogram of the observations and the theoretical distribution (in red). Clearly, the sample mean (marked in blue) is the central point of our distribution.

```
#Plotting the histogram
par(mfrow = c(1,2), mar =c(4.1,3.1,3.1,1.1))
hist(sampleMeans, 30, freq = F, main = "Sample Distribution")
abline(v = 1/0.2, lwd = 3, col = 'blue') #Theoretical Mean

#Theoretical Distribution
pp <- rnorm(2500, mean = 1/lambda, sd = 1/(lambda*sqrt(S)))
lines(sort(pp), dnorm(sort(pp), mean = 1/lambda, sd = 1/(lambda*sqrt(S))), col = "red", lwd = 2)

#Exponential Distribution
xx <- sort(rexp(2500, rate = lambda))
plot(xx, dexp(xx, rate = lambda), type = "l", main = "Exponential Distribution", xlab = "", lwd = 3)
```

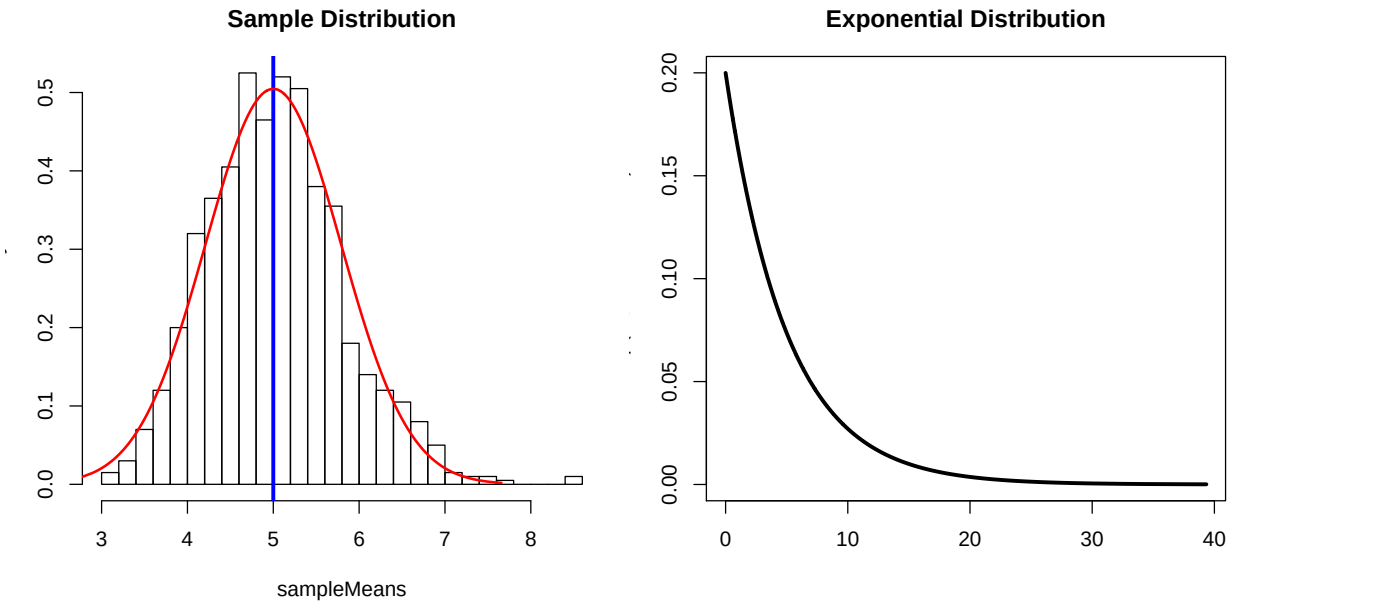

Figure 1: Both distributions of the sample mean and populations, respectively.

Part 2: Tooth Growth. Basic Inferential Data Analysis

We pretend to perform some basic inferential data analysis on the `ToothGrowth` dataset.

```
data(ToothGrowth)
```

From the documentation obtained by `help(ToothGrowth)`:

- The response is 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).

Data Exploration

First we obtain a few statistics of our dataset:

```
library(ggplot2); library(tidyverse)
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
```

```
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 ...
```

```
ToothGrowth %>% group_by(supp, dose) %>% summarise(lenMean = mean(len), count = n())
```

```
## # A tibble: 6 x 4
## # Groups:   supp [2]
##   supp   dose lenMean count
##   <fct> <dbl>   <dbl> <int>
## 1 OJ      0.5   13.2     10
## 2 OJ      1     22.7     10
## 3 OJ      2     26.1     10
## 4 VC      0.5    7.98    10
## 5 VC      1     16.8     10
## 6 VC      2     26.1     10
```

Hypothesis Testing

We are going to perform three analysis comparing the means between the two delivery methods: Oranje Juice (`OJ`) and Ascorbic Acid, i.e, vitamin C (`VC`), via hypothesis Testing:

- without any distinction between doses,
- grouping by dose (3 tests, one for each dose).

Since we don't have enough information about the data, we can perform *t-tests* for this purpose. Considerer the hypothesis test

$$H_0 : \mu_{OJ} = \mu_{VC} \quad \text{vs} \quad H_a : \mu_{OJ} \neq \mu_{VC}$$

where μ_{OJ} and μ_{VC} are the means of the population of the Oranje Juice and Vitamin C methods, respectively.

```
aux <- ToothGrowth %>% mutate(dose = "ALL")
ToothGrowth %>% rbind(aux) %>%
        ggplot(aes(x=supp, y = len, fill = supp)) +
        geom_violin(size=1.5) +
        facet_wrap(~dose)
```

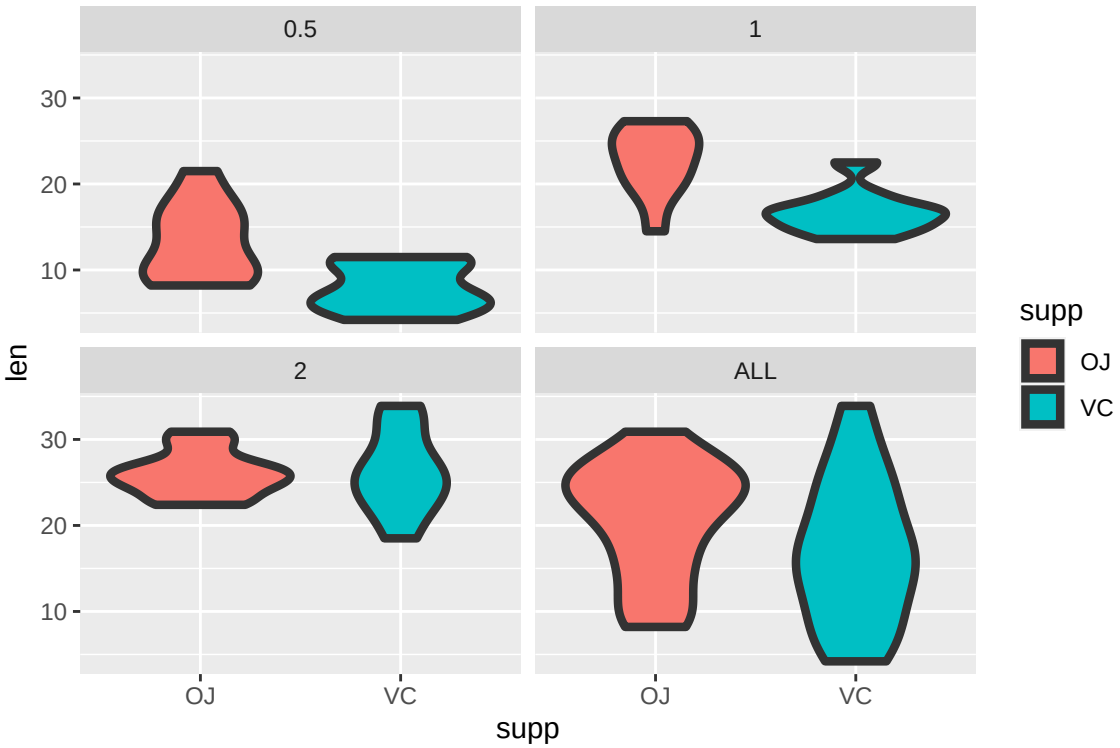

Figure 2: Violin plots for each dosage and without grouping by that.

Regardless of the dosage

First, without taking in consideration the dosage.

```
t.test(len ~ supp, paired = F, 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
```

Since the confidence interval contains the cero and its p-value is $0.0606345 > 0.05$ we failed to reject the null hypothesis. In other words, there is no significance evidence to ensure mean distinction.

Low-dosage

We now filter for the low-dosage.

```
lowdosage <- ToothGrowth %>% filter(dose == 0.5)
t.test(len ~ supp, paired = F, data = lowdosage)
```

```
##
##  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
```

Mid-dosage

Then, we filter for the mid-dosage.

```
middosage <- ToothGrowth %>% filter(dose == 1)
t.test(len ~ supp, paired = F, data = middosage)
```

```
##
##  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
```

High-dosage

Finall, we filter for the high-dosage.

```
highdosage <- ToothGrowth %>% filter(dose == 2)
t.test(len ~ supp, paired = F, data = highdosage)
```

```
##
##  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
```

Conclusions

In the next table, we present all the $p-$values of each hypothesis test

```
##        all.dose    low.dose    mid.dose high.dose
## [1,] 0.06063451 0.006358607 0.001038376 0.9638516
```

From the table, we can conclude the following

- There is no significance evidence to consider the distinct means between the lenght of the cells grouping only by the methods.
- The means between the low (`0.5`) and mid (`1`) dosages are significant distinct.
- In the case of high dosages, we failed to reject the hypothesis of equally means because our $p-$value is nearly 1.