Notebook for Statistical Inference Course Project

Samuel B Baco

PACKAGE LOADING

```
library(tidyverse)
library(ggpubr)
library(datasets)
set.seed(123456)
```

PART 01: SIMULATION EXERCISE

1.1 OVERVIEW

This report is related to course project (week 04) for Johns Hopkins Coursera Statistical Inference classes. The main object is to study the Exponential Distribution e compare it with the Central Limit Theorem.

1.2 SIMULATIONS

Below it is possible to find the code and results to simulate 1000 exponentials, all having sample size if 40 and lambda of 0.2. The result was saved at **exponentials** variable.

```
lambda <- 0.2
n <- 40 # samples for each distribution
N <- 1000 # totals of distribution
exponentials <- replicate(N, rexp(n, lambda))
```

1.3 SAMPLE MEAN VS THEORETICAL MEAN

For $lambda = 0.2$, the exponential distributions has an theoretical mean of 5 (*1/lambda*). To calculate the mean for all 1000 generated distributions, the *apply* function will be used.

```
sMean <- as.data.frame(apply(exponentials, 2, mean))
names(sMean) <- c("mean.exp")
summary(sMean)
```

```
##     mean.exp
##  Min.   :2.527
##  1st Qu.:4.454
##  Median :4.971
##  Mean   :5.023
##  3rd Qu.:5.536
##  Max.   :7.514
```

Comparing the sample mean of theoretical mean, they are pretty close (5.0229151 vs 5).

1.4 SAMPLE VARIANCE VS THEORETICAL VARIANCE

Using the same dataframe created at **exponentials** and the same method (*apply*), the variances for all 1000 distributions were calculated. The theoretical variance for this example is 25 (*[1/lamda]^2*).

```
sVar <- as.data.frame(apply(exponentials, 2, var))
names(sVar) <- c("variance.exp")
summary(sVar)
```

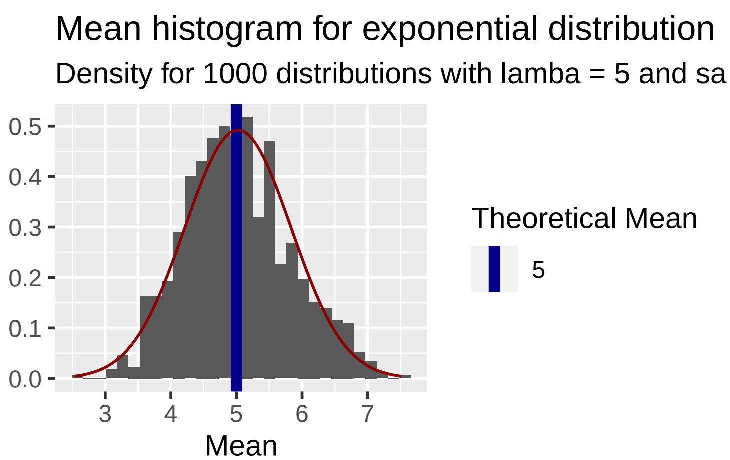
```
##   variance.exp
##  Min.   : 6.518
##  1st Qu.:16.956
##  Median :23.091
##  Mean   :25.242
##  3rd Qu.:30.176
##  Max.   :90.066
```

Comparing the sample variance with theoretical variance, they are pretty close either (25.2424913 vs 25).

1.5 DISTRIBUTION

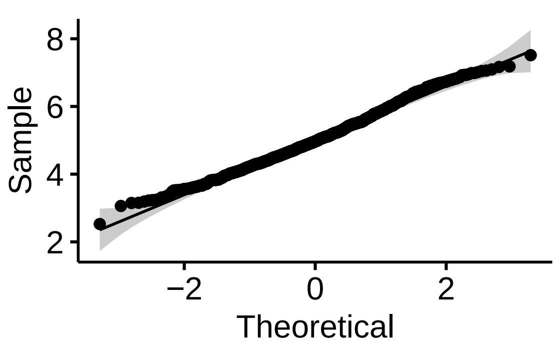
In this section, the normality of the data will be investigated. From Central Limit theorem, it is know that a distribution of means is always normal.

Sample Mean



Doing a graphical analysis, it seems that the distribution of the sample mean is pratically normal. To make sure, it is necessary to run a normality test.

```
ggqqplot(sMean$mean.exp)
```



It is possible to see the data follows a normal distribution.

PART 02: BASIC INFERENTIAL DATA ANALYSIS

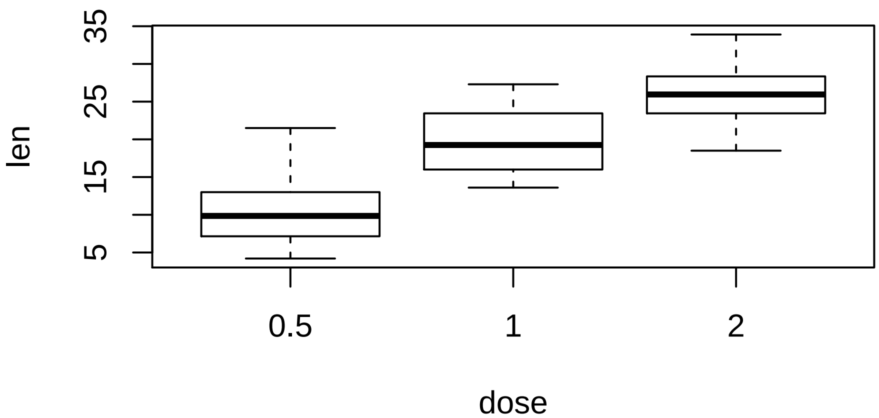
This part consist on analysing the ToothGrowth dataset from *datasets* package. Below there is a summary of the dataset.

```
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
```

It is possible to see that that are 3 possible values from **dose** variable: 0.5, 1, 2.

```
boxplot(len ~ as.factor(dose), data = ToothGrowth, xlab = "dose")
```



There seems to be a considerable difference on variable **len** as **dose** increases. To make sure, it is important to construct the confidence intervals (with 95%) and make sure the extremes does not superimpose themselves.

```
ToothGrowth %>%
        group_by(dose) %>%
        summarise(LCIL = mean(len) - qt(0.975, df = (length(len)-1) * sd(len)/sqrt((length(len)))),
                  UCIL = mean(len) + qt(0.975, df = (length(len)-1)) * sd(len)/sqrt((length(len))))
```

```
## # A tibble: 3 x 3
##    dose  LCIL  UCIL
##   <dbl> <dbl> <dbl>
## 1   0.5  8.51  12.7
## 2   1    17.6  21.8
## 3   2    24.0  27.9
```

As no UCIL (Upper confidence interval limit) is bigger than LCIL (Lower confidence interval limit), it is possible to say that all diferente values of **dose** procuce statiscal different means for **len**, considering a alpha of 5%.

APPENDIX

Code for Exponential Means graphic

```
g <- ggplot(data = sMean, mapping = aes(x = mean.exp))
g + geom_histogram(aes(y = ..density..)) +
        labs (x = "Mean",y = " ", title = "Mean histogram for exponential distribution",
              subtitle = paste("Density for", N, "distributions with lamba =", 1/lambda, "and sample ="
        geom_vline(aes(xintercept = 1/lambda, color = "darkblue"), size = 2) +
        scale_color_identity(guide = "legend", name = "Theoretical Mean", labels = 1/lambda) +
        stat_function(fun = dnorm, args = list(mean = mean(sMean$mean.exp), sd = sd(sMean$mean.exp)), c
print("Don't run me")
```