Statiscal Inference Final Project

Maria Cristina Alameda Salas

25/4/2023

Part 1: Simulation Exercise

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:

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. 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.

Simulating

First, we will generate our data. The following code starts defyng some constants used to simulate our 1000 simulations with 40 exponentials for each of them. After that, our simulations are generated and it is calculated the mean of each simulation.

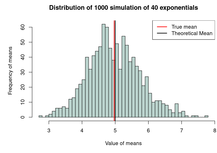
```
lambda <- 0.2
mean <- 1/lambda
std <- 1/lambda
n_similations <- 1000
n_exponentials <- 40

set.seed(69)

# Generate 1000 similations of 40 exponentials
distSims <- matrix(rexp(n_exponentials*n_similations, lambda), nrow=n_similations, ncol=n_exponentials)

# Calculate the mean for each similation (row of the matrix)
# MARGINin indicate that apply similations each row through each row
distMeans <- apply(distSims, MARGIN=1, FUN=mean)

# Calculate the std for each similation (row of the matrix)
# MARGINin indicate that apply similations each row through each row
distVars <- apply(distSims, MARGIN=1, FUN=var)
```

Comparing Sample Mean and Theorical Mean

The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution. One of its properties is the the true mean is equal to the inverse of lambda. With our rate=lambda=0.2 parameter, our true mean should be 5.

To `show` prove this, we will plot an histogram showing the distribution of our 1000 means of 40 exponentials and the theorical mean = 5.

```
hist(x = distMeans,
     main = "Distribution of 1000 similation of 40 exponentials",
     freq = TRUE,
     breaks = 50,
     xlab = "Value of means",
     ylab = "Frequency of means",
     col = "#FFB6C1",
     border = NULL)

abline(v = 1/lambda, lty = 1, lwd = 2, col = "black")

abline(v = mean(distMeans), lty = 1, lwd = 2, col = "red")

legend("topright", lty = 1, lwd =2, col=c("red", "black"), legend=c("True mean", "Theoretical Mean"))
```



As we can see, our two means are very similar, around 5.

Comparing Variance and Theorical Variance

```
trueVariance <- mean(distVars)
theoricalVariance <- (1/lambda)^2

sprintf("True Variance %f", trueVariance)
```

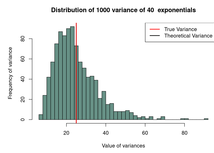
```
## [1] "True Variance 24.986592"
```

```
sprintf("Theorical Variance %f", theoricalVariance)
```

```
## [1] "Theorical Variance 25.000000"
```

The previous code calculate the mean of variance of each row (simulation) of exponents. This value is similar to the theorical variance as CLT says.

```
hist(distVars, breaks = 50, main = "Distribution of 1000 variance of 40  exponentials", xlab = "Value of variance s", ylab = "Frequency of variance", col = "#67918A")

abline(v = theoricalVariance, lty = 1, lwd = 2, col = "black")

abline(v = trueVariance, lty = 1, lwd = 2, col = "red")

legend("topright", lty = 1, lwd =2, col=c("red", "black"), legend=c("True Variance", "Theoretical Variance"))
```



Normal distribution of averages

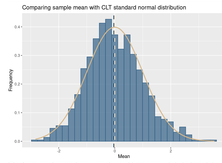
For this task, we need to convert our distribution of average to a 0 mean distribution and compare it to standard normal distribution.

```
standardError <- mean(distMeans)/sqrt(n_exponentials)

# normalization function
standNormalFunc <- function(b){
        round((b - mean(distMeans))/standardError,2)
}

#normalized data of averages
distNormalMeans <- unlist(lapply(distMeans, FUN=standNormalFunc))


library(tibble)
df <- enframe(distNormalMeans)
```

```
# load "tidyverse" plotting package
library(ggplot2)


histogram <- ggplot() +
        geom_histogram(aes(x = df$value, y = ..density..),
                       bins = 30,
                       color = "#315F83",
                       fill="#315F85",
                       alpha = 0.7
        ) +
        labs(
                title = 'Comparing sample mean with CLT standard normal distribution',
                x = 'Mean',
                y = 'Frequency'
        ) +
        geom_vline(
                xintercept = mean(distMeans) - (1/lambda),
                color = "#2B4E5F",
                size = 1,
                linetype="dashed"
        ) +
        geom_vline(
                xintercept = 0,
                color = "#f09d43",
                size = 1,
                linetype="dashed"
        ) +
        stat_function(
                data = df$value),
                size = 2,
                fun = dnorm,
                color = "#2A5B7B8",
                args = list(
                        mean = 0,
                        sd = 1
                )
        )
```

```
## Warning:  `mapping` is not used by stat_function()
```

```
plot(histogram)
```



The two previous graphs looks quite similar. Consequently, it can be said that our data (blue bars) follows a normal distribution (yellow line).

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.

Loading the ToothGrowth data

```
# Load data
data(ToothGrowth)
```

Performing some basic exploratory data analyses

In the following code, we can see that 3 columns compose our dataset: len, supp & dose.

```
# Show first rows
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
```

```
dim(ToothGrowth)
```

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

Provide a basic summary of the data.

This data contains 3 columns. The first column (len) talks about the length of tooth of guinea pigs. It is a continious value in range 4.20 - 33.90 with a mean of 18.81. The columns supp is about the supplement took by the animal. This column is discrete, with just two possible values (OJ & VC).

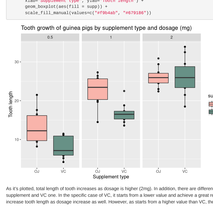
The last column (dose) shows data about dosage of supplement in mg. It goes from 0.5 mg to 2 mg.

```
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
```

Let's plot the previous information by supp group.

```
ggplot(factor(supp), len, data = ToothGrowth, facets=dose,
       main="Tooth growth of guinea pigs by supplement type and dosage (mg)",
       xlab="Supplement type", ylab="Tooth length") +
       geom_boxplot(aes(fill = supp)) +
       scale_fill_manual(values=c("#F39a40", "#67918E"))
```



As it's plotted, total length of tooth increases as dosage is higher (2mg). In addition, there are difference between those animals given OJ supplement and VC one. In the specific case of VC, it starts from a lower value and achieve a great result after increasing dosage. OJ supplement increase tooth length as dosage increase as well. However, as starts from a higher value than VC, the increasing is minor.

Hypothesis Test

Use confidence intervals and/or hypothesis tests to compare tooth growth by supp and dose.

Conclusions

Assumptions

- Supposing the random variable of interest Tooth length has a know mean μ and a variance σ^2. We assume that Tooth length has a normal distribution, that is to say, $X \sim N(\mu, \sigma^2)$.State your conclusions and the assumptions needed for your conclusions.
- Variances of tooth growth are different when using different supplement and dosage.
- The three variables are independent and identically distributed (i.i.d).

Supplement Hypothesis

Null hypothesis: `There is no difference in tooth growth when using supplements (OJ or VC) -> H0: lenOJ = lenVC`

Alternate hypothesis: `There are more tooth growth when using VC than OJ`

```
t.test(ToothGrowth$len[ToothGrowth$supp == 'OJ'], ToothGrowth$len[ToothGrowth$supp == 'VC'], alternative = "great
er", paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$supp == "OJ"] and ToothGrowth$len[ToothGrowth$supp == "VC"]
## t = 1.9153, df = 55.309, p-value = 0.03032
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.4682687       Inf
## sample estimates:
## mean of x mean of y
##  20.66333  16.96333
```

As the p-value (0.03032) is lower than 0.05, then we reject the null hypothesis.

Supplement Hypothesis

Null hypothesis: `There is no difference in tooth growth when using different dosage`

Alternate hypothesis: `There are more tooth growth when the dosage increases`

Comparing 0.5 to 1.0 mg

```
t.test(ToothGrowth$len[ToothGrowth$dose == 0.5], ToothGrowth$len[ToothGrowth$dose == 1], alternative = "less", pa
ired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$dose == 0.5] and ToothGrowth$len[ToothGrowth$dose == 1]
## t = -6.4766, df = 37.986, p-value = 6.342e-08
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -8.753323
## sample estimates:
## mean of x mean of y
##    10.605    19.735
```

Comparing 1.0 to 2.0mg

```
t.test(ToothGrowth$len[ToothGrowth$dose == 1], ToothGrowth$len[ToothGrowth$dose == 2], alternative = "less", pair
ed = FALSE, var.equal = FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  ToothGrowth$len[ToothGrowth$dose == 1] and ToothGrowth$len[ToothGrowth$dose == 2]
## t = -4.9005, df = 37.101, p-value = 9.532e-06
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -4.17387
## sample estimates:
## mean of x mean of y
##    19.735    26.100
```

Both test show a p-values lower than 0.5. Then we reject the null hypothesis. This can be interpreted as, based on these low p-values, it is very likely that dosage has an effect on length, and a major value of dosage, major increase in length.

#Conclusions We conclude that, due to the p-value obtained, there is a difference between administering the OJ supplement and the VC. In addition, there is a greater increase in tooth size when a higher dose of these supplements is administered.