Statistical Inference Project - Simulation Data

ElisaRMA

5/5/2021

Overview

This document describes two statistical analysis and it is divided into two parts: the simulation and the comparision between the simulated data and theoretical parameters of the Exponential Distribution

Simulation

The objective of this first part of the project was to simulate exponential distributions using the `rexp()` function of R and compare it with the Central Limit Theorem.

First a 1000 sets of 40 numbers were generated. These 40 numbers came from the exponential distribution and for each set, the mean was calculated. In other words, the mean was taken from 40 random exponentials, resulting in a 1000 means. These numbers were stored in a matrix.

Then finally, the mean of the 1000 numbers was calculated, therefore, a matrix with a 1000 rows was generated, each row containing one mean.

```
set.seed(1)
lambda <- 0.2
sim<- matrix(rexp(1000*40, lambda), ncol= 40, nrow=1000)

# Now we have a matrix with 40 columns and a 1000 rows.
# Then we take the mean of each row, in  other words, the mean of 40 random exponentials.
# With this, we will have 1000 numbers (a matrix with one column and a 1000 rows)

mn <- matrix(apply(sim, 1, mean))
```

Sample Mean vs Theoretical Mean

As instructed, the 40 numbers generated inittialy were from a exponential distribution with a lambda set as 0.2. For this distribution the average and standard deviation were both 1/lambda.

To compare both means the `apply` function was used, by column. The sample mean was calculated following the theoretical formula and both means were stored in a table, using the package `kableExtra`. It was observed the following:

```
smean <- matrix(apply(mn, 2, mean))

tmean <- 1/lambda
```

```r
library(kableExtra)
comparisonmean <- data.frame(smean, tmean)
colnames(comparisonmean) <- c("Sample Mean", "Theoretical Mean")
kbl(comparisonmean)
```

Sample Mean	Theoretical Mean
4.990025	5

As observed, following the Law of Large Numbers and the Central Limit Theorem both means are very close.

Sample Variance versus Theoretical Variance

As mentioned before, the instructions for this project were to create 40 random exponential, 1000 times, with a lambda of 0.2. Therefore, standard deviation was 1/lambda.

As we know, the standard deviation is sqrt(var), therefore, the variance of this distribution is (1/lambda)^2, and the theoretical variance is this variance divided by the population number, in this case, 40.

To compare both variances the `apply` function was also used, by column, to the means calculated previously. The sample variance was calculated following the theoretical formula and both means were also stored in a table, using the package `kableExtra`. It was observed the following:

```r
svar <- matrix(apply(mn, 2, var))

tvar <- ((1/lambda)^2)/40

# Table comparing both variances
library(kableExtra)
comparisonvar <- data.frame(svar, tvar)
colnames(comparisonvar) <- c("Sample Variance", "Theoretical Variance")

kbl(comparisonvar)
```

Sample Variance	Theoretical Variance
0.6177072	0.625

As observed, following the Law of Large Numbers and the Central Limit Theorem both variances are close.

Distribution

To show how much the simulated data is similar to the Normal Distribution both were plotted within a single histogram. For this, the `ggplot2` package was used.

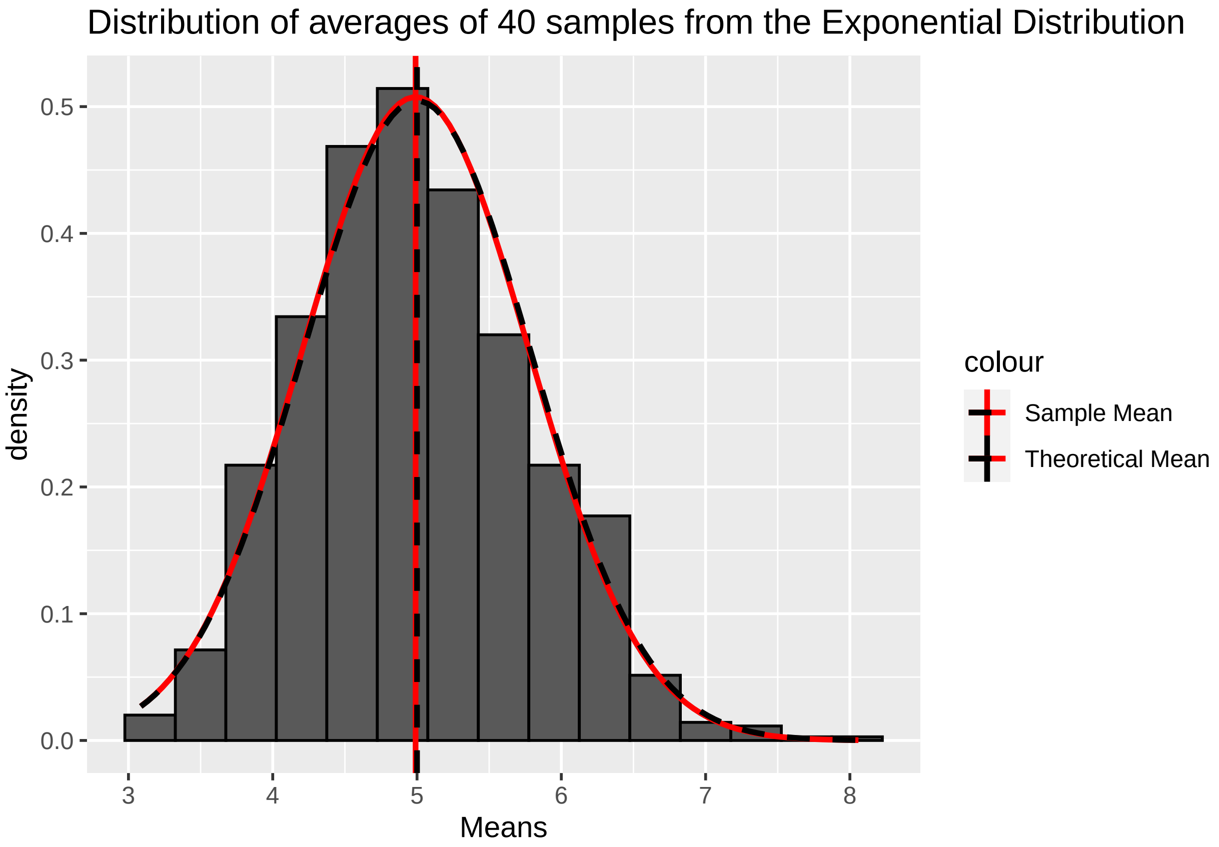
First a ggplot layes was created using the 1000 means in the `aes()` argument. Then, a histogram layers was added, displaying the density (`geom_histogram`) . Then, two vertical lines were added to the plot, to represent both theoretical and sample means (`geom_vline`). The theoretical means, as observed in the legend, was plotted as dashed line.

Finally, both distributions were plotted centered on its respective means and with their respective variances (`stat_function`). The normal distribution were also represented with a dashed line

```
library(ggplot2)

mn <- as.data.frame(mn)

ggplot(mn,aes(V1)) +
geom_histogram(binwidth=.35, colour = "black", aes(y = ..density..))+
geom_vline(aes(xintercept = smean, colour="Sample Mean"),
           size=1) +
geom_vline(aes(xintercept = tmean, colour = "Theoretical Mean"),
           linetype = "dashed", size=1) +
stat_function(fun = dnorm, args= list(mean = smean, sd = sqrt(svar)),
              colour="red", size = 1,show.legend = TRUE )+
stat_function(fun = dnorm, args= list(mean = tmean, sd = sqrt(tvar)),
              colour="black", size = 1, linetype = "dashed", show.legend = TRUE )+
scale_colour_manual(values = c("Sample Mean" = "red", "Theoretical Mean" = "black"))+
labs(title = "Distribution of averages of 40 samples from the Exponential Distribution")+
xlab ('Means')
```



As observed by the plot, both distributions are very similar, as predicted by the Central Limit Theorem and the Law of Large Numbers.