Statistical Inference Course Project Part 1

Hernan Solano

March 05 2018

Overview

In this report we will simulate 1000 samples of the exponential distribution with lambda = 0.2, the size of each sample is 40. Then we will show analyze its theoretical mean and variance with the mean and variance from the simulations. Finally we will show that the distribution of the mean of the simulated data is pretty close to a normal distribution.

We will use the ggplot2 and ggthemes libraries in this report.

Simulations

In this section I first load the libraries necesaries for my analysis (ggplot2 and ggthemes). Then initilized the vectors and constants to be used:

- The means_vector that will hold the mean of each exponential samples
- The mean_dist_vector will be the normalized vector following the Central limit theorem formula
- The theoretical measures (mean, variance and standard error of the mean)

```
library(ggplot2)
library(ggthemes)
##Initializing contants and variables
means_vector = NULL
mean_dist_vector = NULL
theo_mean <- 1/0.2
theo_var <- ((1/0.2)^2)/40
##standard error of the mean
se <- sqrt(theo_var)
```

After this setup I then proceed with the simulations, acumulating each mean in the means_vector variable. Therefore the 1000 values of the vector represent 1000 means, each correspond to an exponential sample of size 40 and lambda 0.2. Then I normalized this vector using the CLT formula and save the values ont he mean_dist_vector. Finally the sumulated summary statistics are calculated (mean, sd and var).

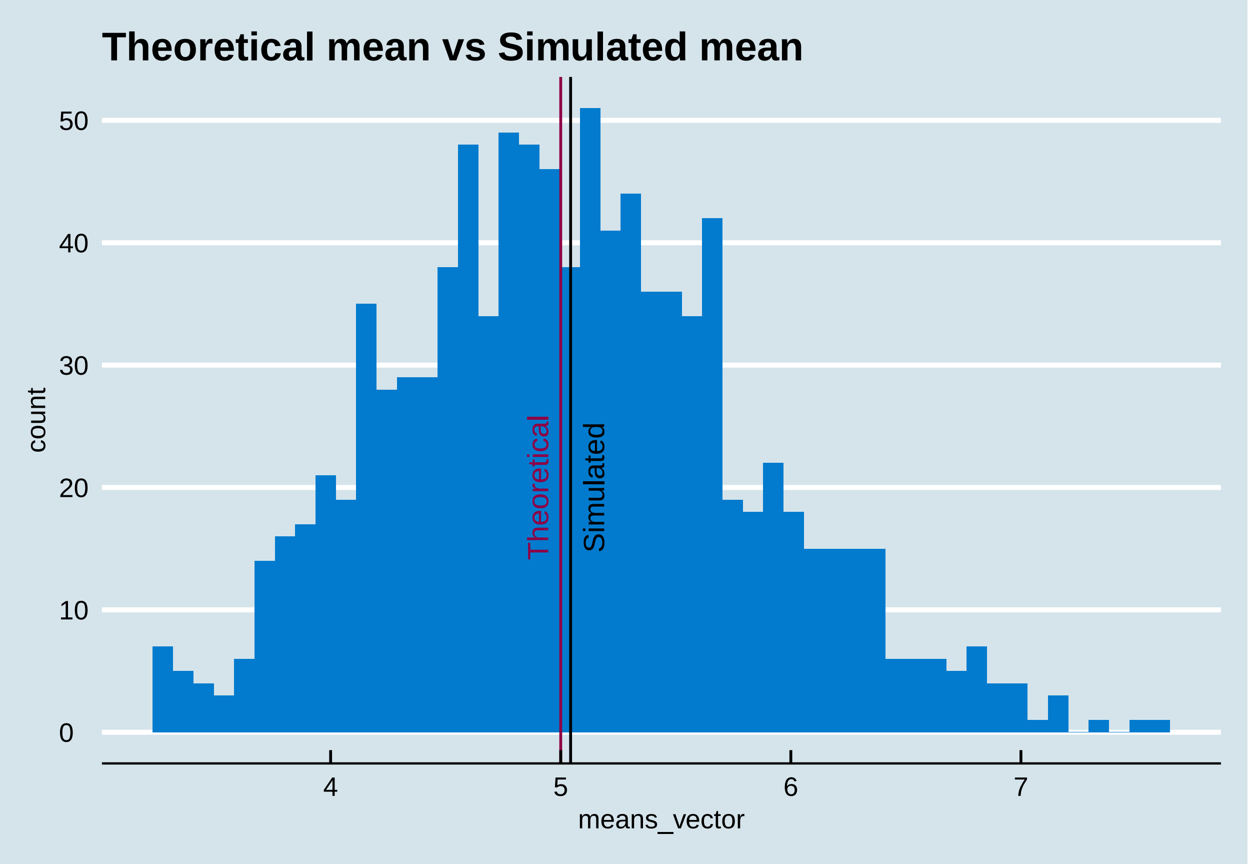
```
##Simulation of a 1000 exponential samples of size 40
set.seed(5)
## I calculated 1.000 averages from 40 ramdom exponentials with lambda equal to 0.2
for (i in 1 : 1000) {means_vector <- c(means_vector, mean(rexp(40,0.2)))}
##Normalized means vector
mean_dist_vector<- (means_vector - theo_mean)/se
##Simulated summary statistics
sim_mean <- mean(means_vector) ## Simulated mean
sim_sd <- sd(means_vector) ## Simulated standart deviation
sim_var <- var(means_vector) ## Simulated variance
```

Sample Mean versus Theoretical Mean

In this section we will see the difference between the theoretical mean (5) and the mean calculated through the simulations (5.043).

- The first thin we can notice is that their difference is: 0.043
- Also the t 95% confidence interval of the calulated mean is: (4.995, 5.091), since this interval includes 5 and the p-value of a the twosided test is 0 which is greater than 0.05, we can say that there is no enough evidence to reject the null hypothesis that the mean is equal to 5. In other words at 5% level of significance we fail to reject the null hypothesis.
- The following figure shows a histogram of the means_vector and shows how close the theoretical and calculated means are.

```
ggplot() + geom_histogram(aes(means_vector),bins = 50,fill = "#027BCE") +
    geom_vline(xintercept =c(theo_mean,sim_mean),colour = c("#8E0045","black")) +
    theme_economist() + geom_text(aes(x=theo_mean-0.1, label="Theoretical", y=20),
                                angle=90,colour = "#8E0045") +
    geom_text(aes(x=sim_mean+0.1, label="Simulated", y=20), angle=90,colour = "Black") +
    labs(title = "Theoretical mean vs Simulated mean")
```



Sample Variance versus Theoretical Variance

In this section we will see the difference between the theoretical variance (0.625) and the variance calculated through the simulations (0.603).

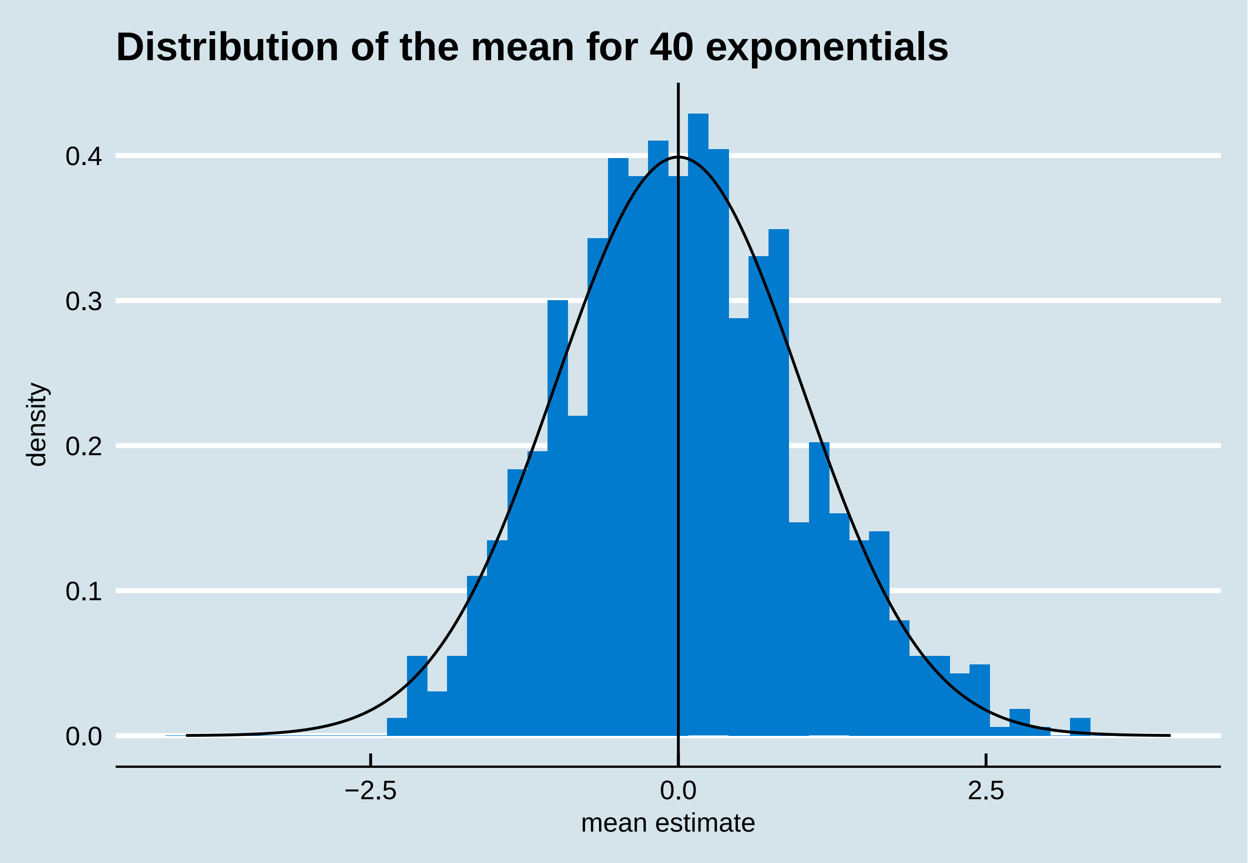
- The difference between the simulated and theoretical variance is: -0.022

- The difference between the standard deviation of the simulated data and the SE of the meam is also minimal: r round(sim_sd- se,3)‘

Distribution

The following graph shows the histogram of the mean_dist_vector which is the normilized verson of the means_vector values. I superposed a standard normal density curve to show that the histogram adjusts to the bell shape of the normal distribution.

```
ggplot() + geom_histogram(aes(x=mean_dist_vector,y=..density..),bins = 50,fill = "#027BCE") +
    geom_line(aes(seq(-4,4,length.out = 1000), y = dnorm(x = seq(-4,4,length.out = 1000))),
              color = "black") +
    labs(title = "Distribution of the mean for 40 exponentials", x = "mean estimate") +
    theme_economist() + geom_vline(xintercept = 0)
```



There is an interesting fact of the hypothesis tests in relation to the CLT. If we run an hypothesis test to see whether the mean of this vector is equal to zero, we will get similar results of those obtained in the sample mean vs theoretical mean section. We also fail to reject the null hypothesis and the p-value is the same.

```
t.test(mean_dist_vector)
```

```
##
##  One Sample t-test
##
## data:  mean_dist_vector
## t = 1.7538, df = 999, p-value = 0.07977
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.006474509  0.115391006
## sample estimates:
##  mean of x
## 0.05445825
```