Statistical Inference Course Project - part 1

Andreia Carlos

14 October 2016

Overview

This project consists of two parts:

1. **A simulation exercise**
2. Basic inferential data analysis

This report presents part 1. Statistical Inference (SI) is used to show a simulation of an exponential distribution and calculate it's statistical properties (means and measures of dispersion). Then those statistical properties are compared with with the Central Limit Theorem, as discussed in the SI course slides, and the appropriate statistical tests are used to interpret the results.

Part 1: Simulation Exercise

The intent is to explore the exponential distribution and compare it with the Central Limit Theorem. In this case the exponential distribution is 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`, where `lambda = 0.2` for all of the simulations.

Load required libraries

```
require(dplyr)
require(pander)
require(ggplot2)
```

Question 1: Show the sample mean and compare it to the theoretical mean of the distribution.

Define values to be used

```
set.seed(867564) #setting a random value for reproducibility of the same analysis
numsim <- 1000
num <- 40
lambda <- 0.2
tmean <- 1/lambda
tsd <-  1/lambda
```

The theoretical mean of the distribution is 5. The theoretical variance of the distribution is 25. The following steps describe how the simulation is done to get to the answers to the first and second questions.

Steps for the simulation:

- Firstly create a data frame with 1000 simulations of 40 samples
- add a new column with exponential(0.2) values of n
- prepare data to be grouped by the first id column
- calculate the sample means and standard deviations of the exponential values

- add a new variable with the difference between the theoretical mean and sample means

```
simul <- expand.grid(nsim=1:numsim, id = 1:40) %>%
        mutate(expSim = rexp(n(), lambda)) %>%
        group_by(nsim) %>%
        summarise(smean = mean(expSim)) %>%
        mutate(mu_dif = tmean - smean)
```

A table with a preview of the simulated data is presented in appendix

To compare the sample mean to the theoretical mean of the distribution, first the mean of the sample means is calculetd and then a density distribution function is used, `pnorm`, generating a normal distribution.

```
MuMean <- mean(simul$smean)
MuMean
```

```
## [1] 4.97369
```

The mean of the sample means is equivalent to the the theoretical sample mean, which is 5.

We can also check the mean of the difference between the theoretical mean and the sample mean:

```
mean(simul$mu_dif)
```

```
## [1] 0.02630996
```

The average of this difference is -0.029 which supports the assumption that the means are not so different from each other.

```
pnorm(mean(simul$smean), mean = tmean, sd = tsd)
```

```
## [1] 0.4979008
```

This gives a p-value close to 0.5, which supports the probability of the exponential samples asssuming a normal distribution. This is in line with the Central Limit Theorem (CLT) which states that "the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the sample size increases."

Question 2: Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

The variance of the theoretical distribution is determined by the calculation of the standard error, squared. The se is calculated using the standard deviation `tse` and dividing it by the square root of n `num`.

```
num <- 40
se <- tsd/sqrt(num)
se
```

```
## [1] 0.7905694
```

```
Svar <- se^2
Svar
```

```
## [1] 0.625
```

When comparing it to the sample variance, we use the squared standard deviation to calculate the variance of the sample mean. And we can also determine the standard deviation of the sample mean.

```
MuSd <- sd(simul$smean)
MuSd
```

```
## [1] 0.7650956
```

```
MuVar <- MuSd^2
MuVar
```

```
## [1] 0.5853712
```

The code below gives the same values for the sample mean, variance and standard deviation of the sample mean, but this time using dplyr:

```
simVar <- simul %>% summarise(
                xmean = round(mean(smean),3),
                xvar = round(var(smean),3),
                xsd = round(sd(smean),3)
        )
pander(simVar)
```

xmean	xvar	xsd
4.974	0.585	0.765

The variance of the sample mean **(0.639)** can be compared to the variance of the theoretical distribution **(0.625)**. The standard deviation of the sample mean **(0.799)** can be combared to the standard error of the theoretical distribution **(0.790)**. This supports the assumption of a normal distribution as the sample mean approaches the same asymptothic behavior as of the theoretical distribution.

Question 3: Show that the distribution is approximately normal.

Here are presented some simulations to show the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials. The plots of the respective simulations can be seen in appendix

1000 simulations of 40 exponentials without determining the sample mean

```
# 1000 simulations
simul1 <- expand.grid(nsim=1:1000, id = 1:40) %>%
        mutate(expSim = rexp(n(), lambda))
```

100000 simulations of 40 exponentials without determining the sample mean, just to explore the effect of increasing the number of simulations over the same data.

```r
# 100000 simulations
simul2 <- expand.grid(nsim=1:100000, id = 1:40) %>%
        mutate(expSim = rexp(n(), lambda))
```

1000 simulations of the sample mean of 40 exponentials, to compare the differences without the average of the samples.

```r
simul3 <- expand.grid(nsim=1:1000, id = 1:40) %>%
        mutate(expSim = rexp(n(), lambda)) %>%
        group_by(nsim) %>%
        summarise(smean = mean(expSim))
```

The theoretical distribution of random exponentials has an uniform distribution when compared to the distribution of averages of 40 random uniforms. The latter has a more Gaussian distribution.

Appendix

This table shows a preview of the first simulated data with 1000 simulations of 40 exponentials

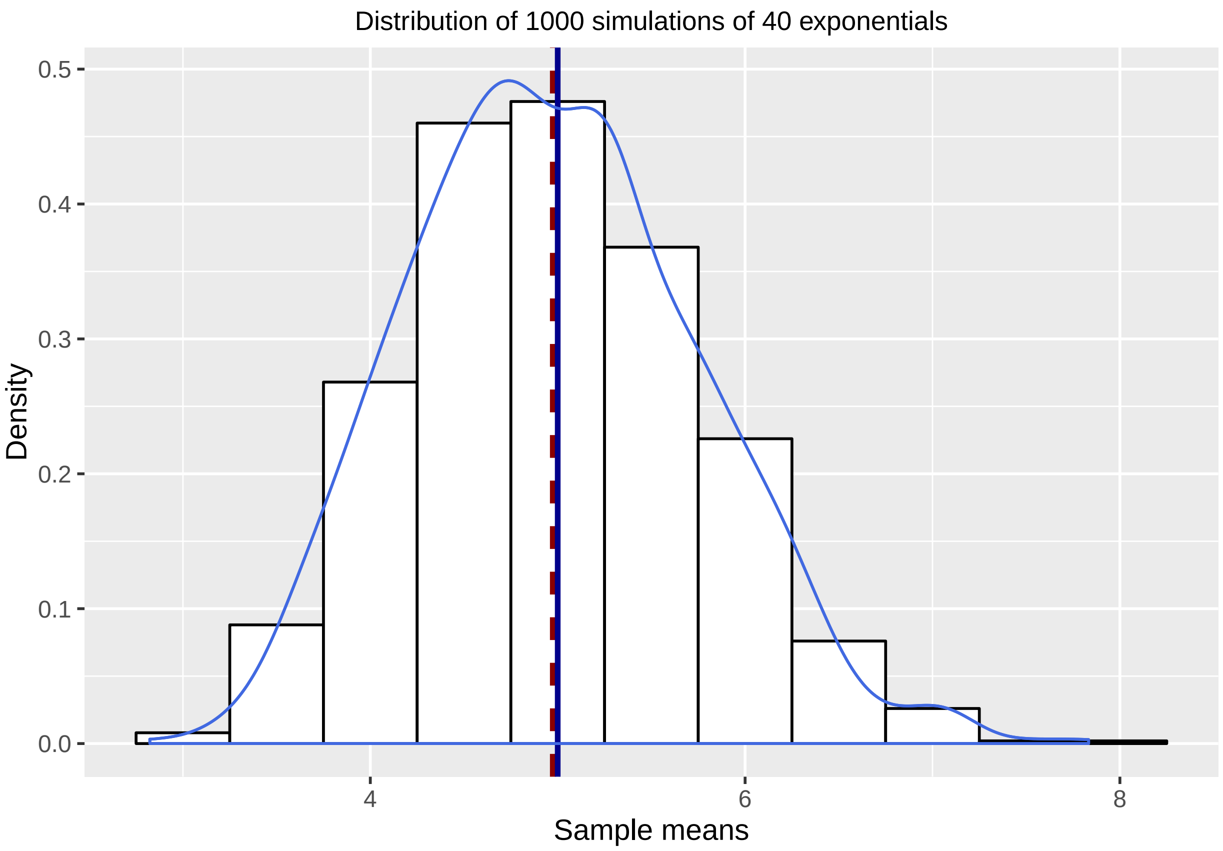
```
pander(head(simul))
```

nsim	smean	mu_dif
1	3.950017	1.0499834
2	4.437274	0.5627265
3	6.188637	-1.1886373
4	5.714509	-0.7145088
5	3.427434	1.5725660
6	4.859178	0.1408222

`nsim` is the id number containing 1000 simulations. `smean` are the estimated means for the 40 exponential(0.2) samples. `mu_dif` is the difference between the theoretical mean `1/lambda` and the `smean`.

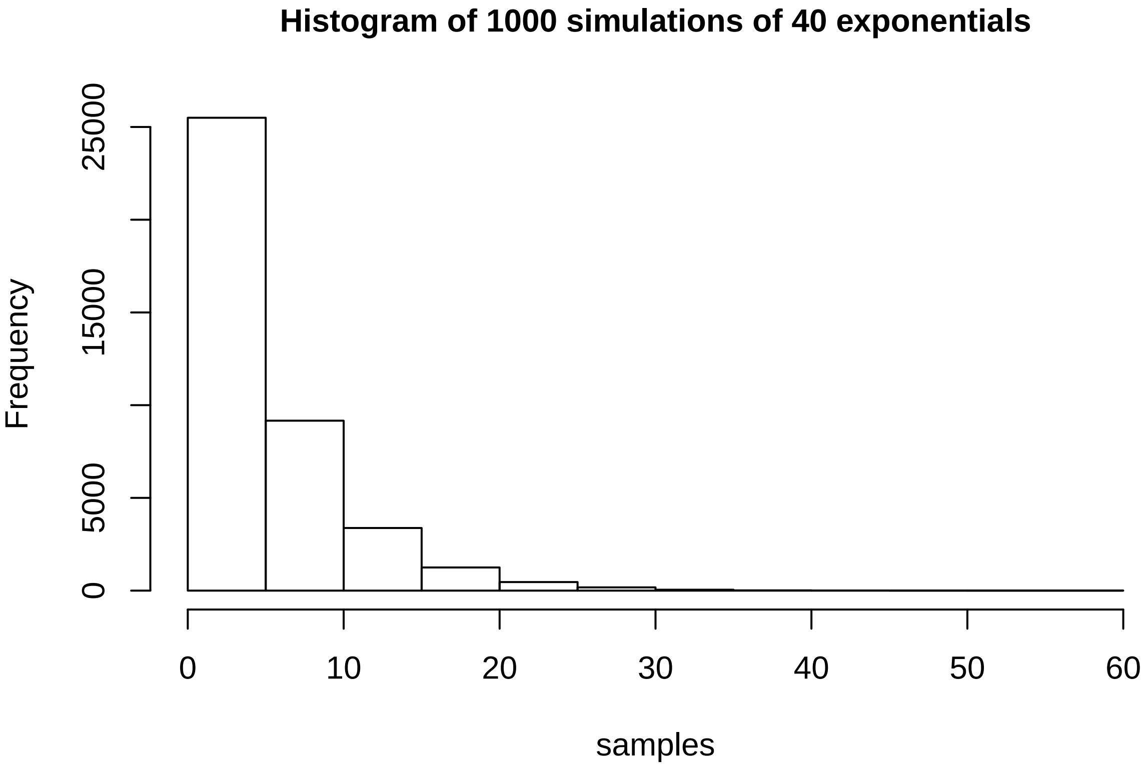
The code below produces a histogram with the distribution of the sample means of 1000 simulations of 40 exponentials, with a dashed line representing the sample mean (the mean of the sample means), which turns out to be equivalent to the theoretical mean of the distribution.

```
ggplot(simul, aes(x = smean)) +
        labs(title = "Distribution of 1000 simulations of 40 exponentials",
             x = "Sample means", y = "Density") + theme(plot.title = element_text(size=10)) +
        geom_histogram(aes(y = ..density..), binwidth = .5, color = "black",
                       fill = "white") +
        geom_vline(xintercept = MuMean, color = "darkred", linetype = "dashed",
                   size = 1) +
        geom_density(fill = NA, colour = "royalblue") +
        geom_vline(xintercept = 5, color = "darkblue", linetype = 1, size = 1) +
        geom_density(fill = NA, colour = "royalblue")
```



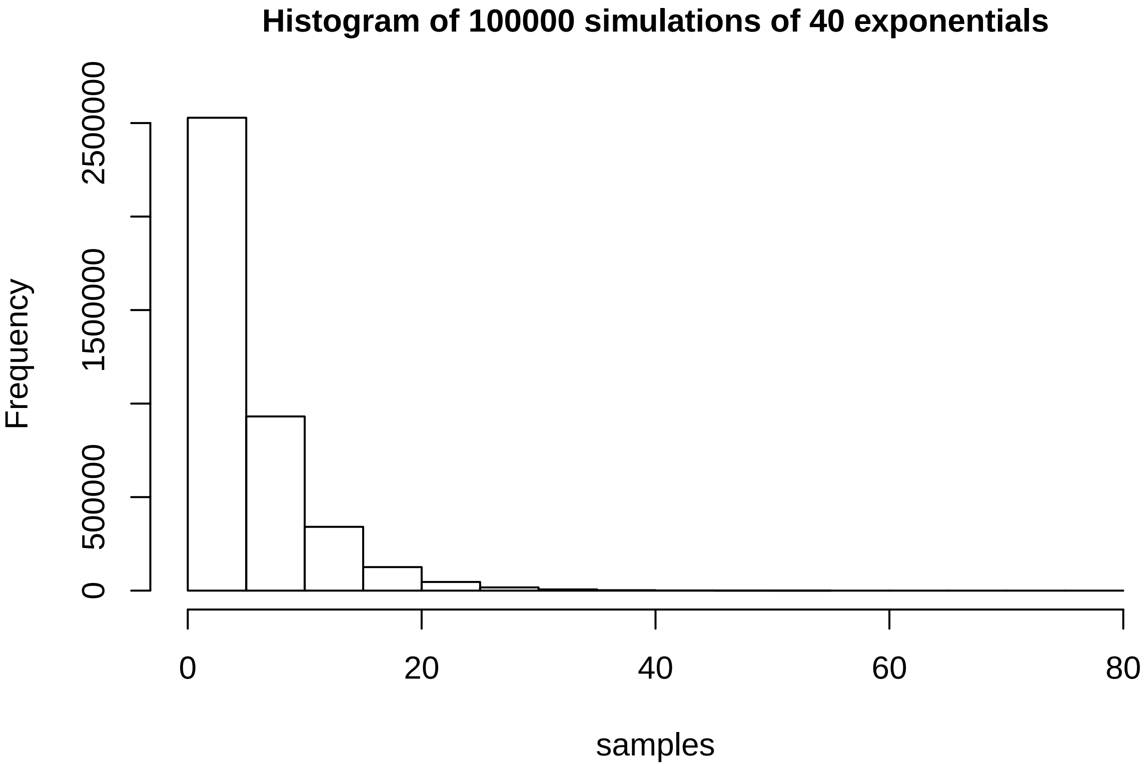
The code below produces a histogram with 1000 simulations of 40 exponentials without determining the sample mean

```
hist(simul1$expSim,
     main = paste("Histogram of 1000 simulations of 40 exponentials"),
     xlab = "samples", ylab = "Frequency", cex.main = 1)
```



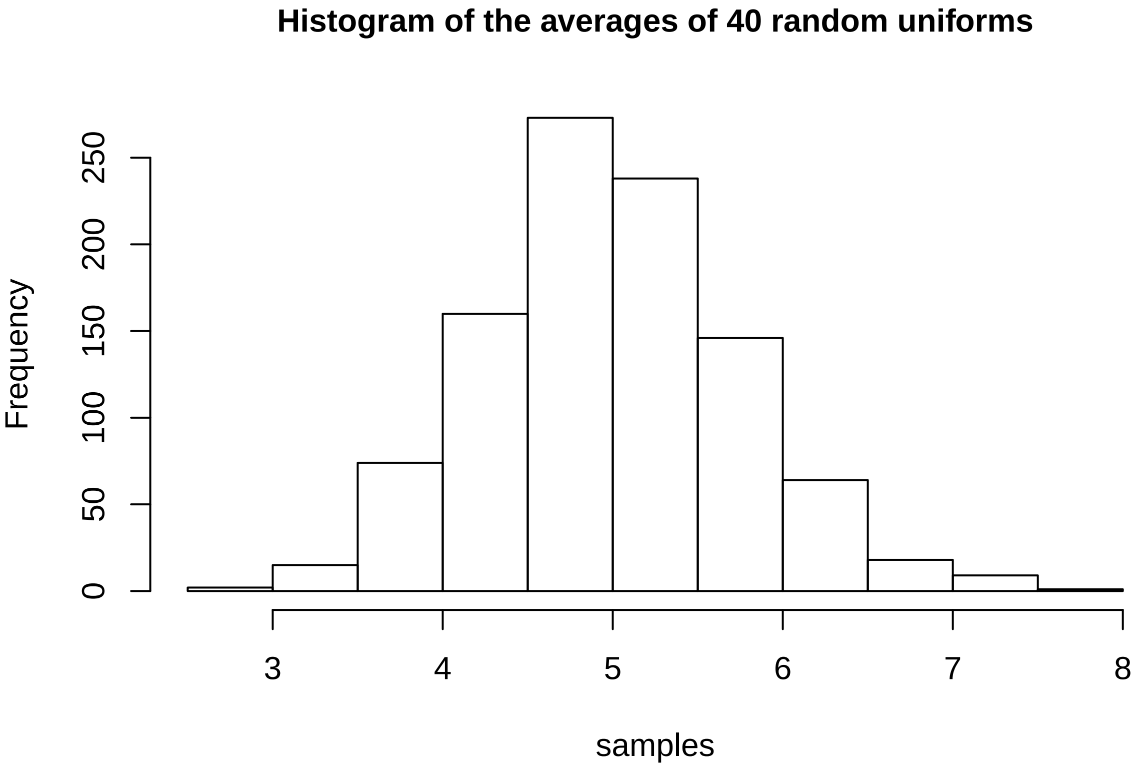
The code below produces a histogram with 100000 simulations of 40 exponentials without determining the sample mean, just to explore the effect of increasing the number of simulations over the same data.

```r
hist(simul2$expSim,
     main = paste("Histogram of 100000 simulations of 40 exponentials"),
     xlab = "samples", ylab = "Frequency", cex.main = 1)
```



The following plot shows 1000 simulations of the sample mean of 40 exponentials, to compare the differences without the average of the samples.

```
hist(simul3$smean,
     main = paste("Histogram of the averages of 40 random uniforms"),
     xlab = "samples", ylab = "Frequency", cex.main = 1)
```



The theoretical distribution of random exponentials has an uniform distribution when compared to the distribution of averages of 40 random uniforms. The latter has a more Gaussian distribution.