Statistical Inference course Project - Part 1

Simulation of the distribution of the mean of 40 exponential

Olivia Urbanowski

5/27/2020

Overview

In this report, we 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 λ is the rate parameter.
The mean of exponential distribution is $\mu = 1/\lambda$ and the standard deviation is $\sigma = 1/\lambda$.
We will do our simulations with $\lambda = 0.2$; thus we have $\mu = 5$.

Simulations

The investigation will be done via the distribution of 1000 averages of 40 exponentials.

```
n <- 40
lambda <- 0.2
```

First we create the sample we are going to work with, then we calculate its mean and its standard deviation. For comparison purposes, we also set the theoretical values of the distribution of the mean.

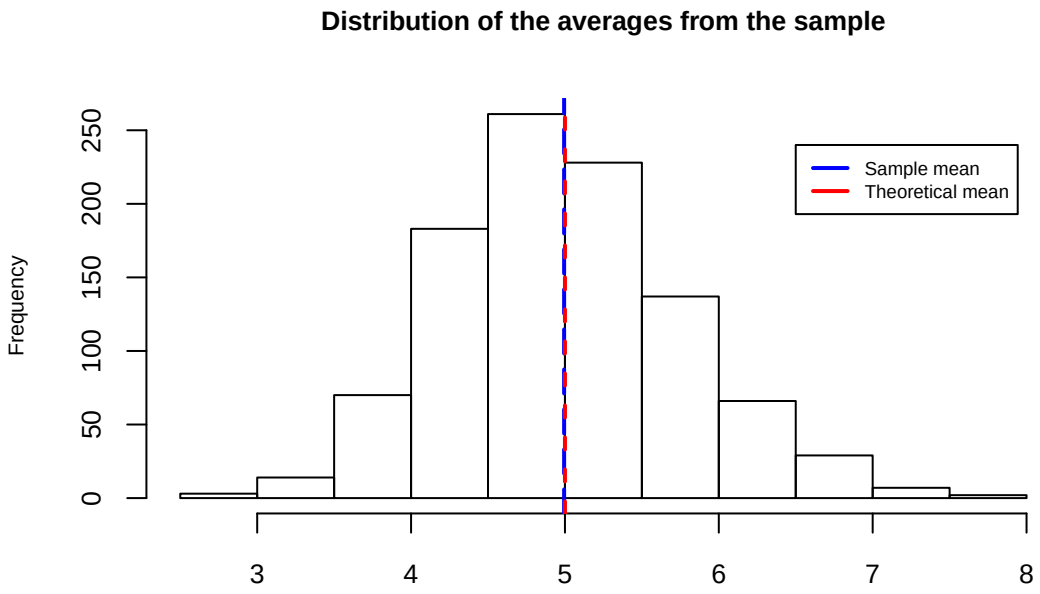
```
expdistr <- NULL
for (i in 1 :1000) {
  sample <- rexp(n,lambda)
  expdistr <- c(expdistr, mean(sample))
}
samplemean <- mean(expdistr)
samplesd <- sd(expdistr)

mu <- 1/lambda
sigma <- 1/lambda
varmean <- sigma^2/n
```

1. Comparing the means

Let us visually compare the mean calculated from our sample 4.99 and the theoretical mean 5.

```r
hist(expdistr, xlab = "", main = "Distribution of the averages from the sample",
     cex.lab=0.7, cex.axis=0.8, cex.main=0.8)
abline(v=samplemean, lwd=2, lty=5, col="blue")
abline(v=mu, lwd=2, lty=2, col="red")
legend(6.5,240, legend = c("Sample mean", "Theoretical mean"),
     col= c("blue","red"), lwd=2:2, cex= 0.6)
```



On this graph we can see clearly that both means are pretty close. Actually they were plotted using different line style so we can differentiate them even if one covers the other.

2. Comparing the variability of the means

To compare the variability of the means from the sample vs. the theory, we will can check the figures we calculated in the "Simulation" section:

```r
mydata <- rbind(c(round(samplemean,2),round(samplesd^2,2)),c(mu,varmean))
colnames(mydata) <- c("Mean","Variance")
rownames(mydata) <- c("Sample", "Theory")
knitr::kable(mydata, caption = "Sample vs. Theory")
```

Table 1: Sample vs. Theory

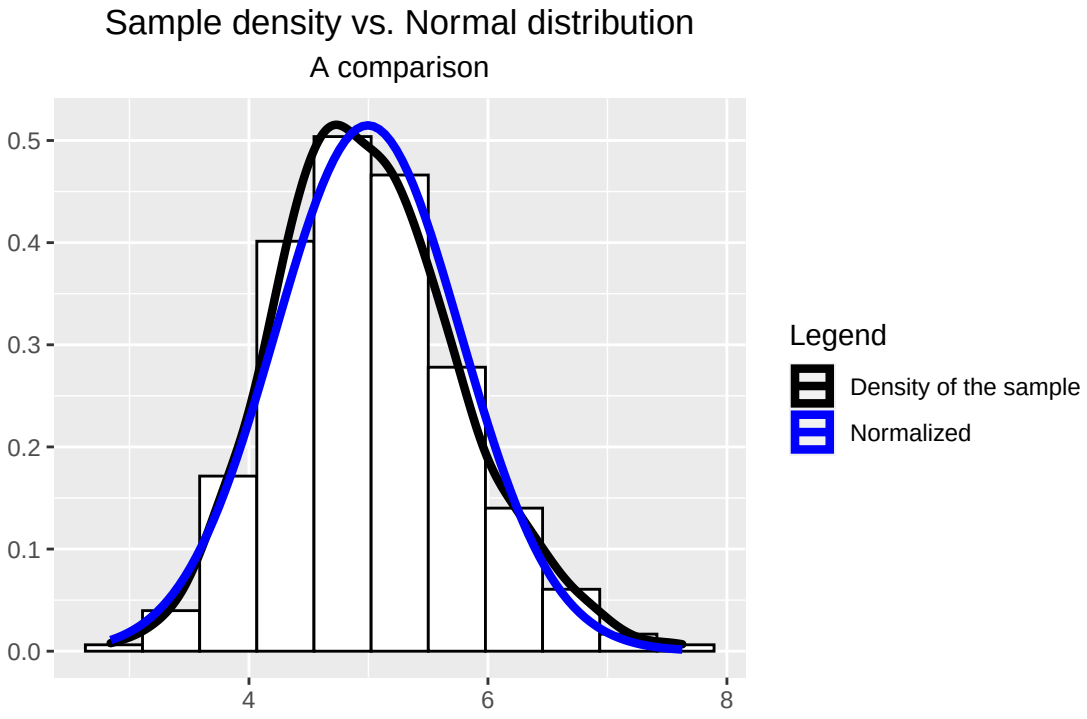
	Mean	Variance
Sample	4.99	0.600
Theory	5.00	0.625

As we can see, both variances are pretty close.

3. Verifying the Central Limit Theorem

The Central Limit Theorem states that the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the sample size increases.
So in our case we are going to check if our distribution is normal, by plotting the density of our samples.

```
library(ggplot2)
g <- ggplot(data = as.data.frame(expdistr), aes(x=expdistr))
g + geom_histogram(aes(y = ..density..), bins=11, fill = "white",
     colour = "black") + geom_density(size = 1.5,
     aes(color = "black")) + stat_function(fun=dnorm, geom = "line",
     args = list(mean = samplemean, sd = samplesd), size = 1.5,
     aes(color = "blue")) + scale_colour_manual(guide = "legend",name =
     'Legend', values =c('black'='black','blue'='blue'), labels =
      c('Density of the sample', 'Normalized')) + labs(title =
     "Sample density vs. Normal distribution", subtitle = "A comparison", x = "",
     y = "") + theme(plot.title = element_text(hjust = 0.5), plot.subtitle =
    element_text(hjust = 0.5))
```



We can clearly see the density approximate that of a normal distribution. If we were to repeat this with a larger sample, it will become even clearer.
The distribution of means of 40 exponentials behaves as predicted by the CLT.

Appendix

For the second question, we could also visualize the difference in variance between the sample and the theory by plotting the curves:

```
library(ggplot2)
g <- ggplot(data = as.data.frame(expdistr), aes(x=expdistr))
g + geom_histogram(aes(y = ..density..), bins=11, fill = "white",
      colour = "black") + stat_function(fun=dnorm, geom = "line", args =
      list(mean = mu, sd = sqrt(varmean)), size = 1.5, aes(color =
      "red")) + stat_function(fun=dnorm, geom = "line", args = list(mean =
      samplemean, sd = samplesd), size = 1.5, aes(color = "blue")) + labs(title =
      "Variance of the mean of 40 exponentials", subtitle =
      "Sample of 1000 vs. theory", x = "", y = "") + theme(plot.title =
      element_text(hjust = 0.5), plot.subtitle = element_text(hjust =
      0.5)) + scale_colour_manual(guide = "legend",name = 'Legend', values =
      c('blue'='blue','red'='red'), labels = c('Sample','Theory'))
```