Statistical Inference, Coursera Project

Malgorzata MS

January 31, 2017

Simulation Exercise Instructions

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

```
Show the sample mean and compare it to the theoretical mean of the distribution.
Show how variable the sample is (via variance) and compare it to the theoretical variance of the distril
Show that the distribution is approximately normal.
```

Simulation

The code below simulates 1000 exponential distributions of 40 observations. The lambda in exponential distributions is always set to 0.2. The results are stored in a matrix sim_matrix with 1000 rows and 40 columns. Each row is a sample of 40 observation with exponential distribution.

```
lambda=0.2
true_mean <- 1/lambda
true_sd <- 1/lambda
n <- 40
sim_n <- 1000 # number of simulation

sim_matrix <- matrix(NA, nrow=sim_n, ncol=n)
for(i in 1:sim_n){
        set.seed(i)
        sim_matrix[i,] <- rexp(n, lambda)
}
dim(sim_matrix)
```

```
## [1] 1000   40
```

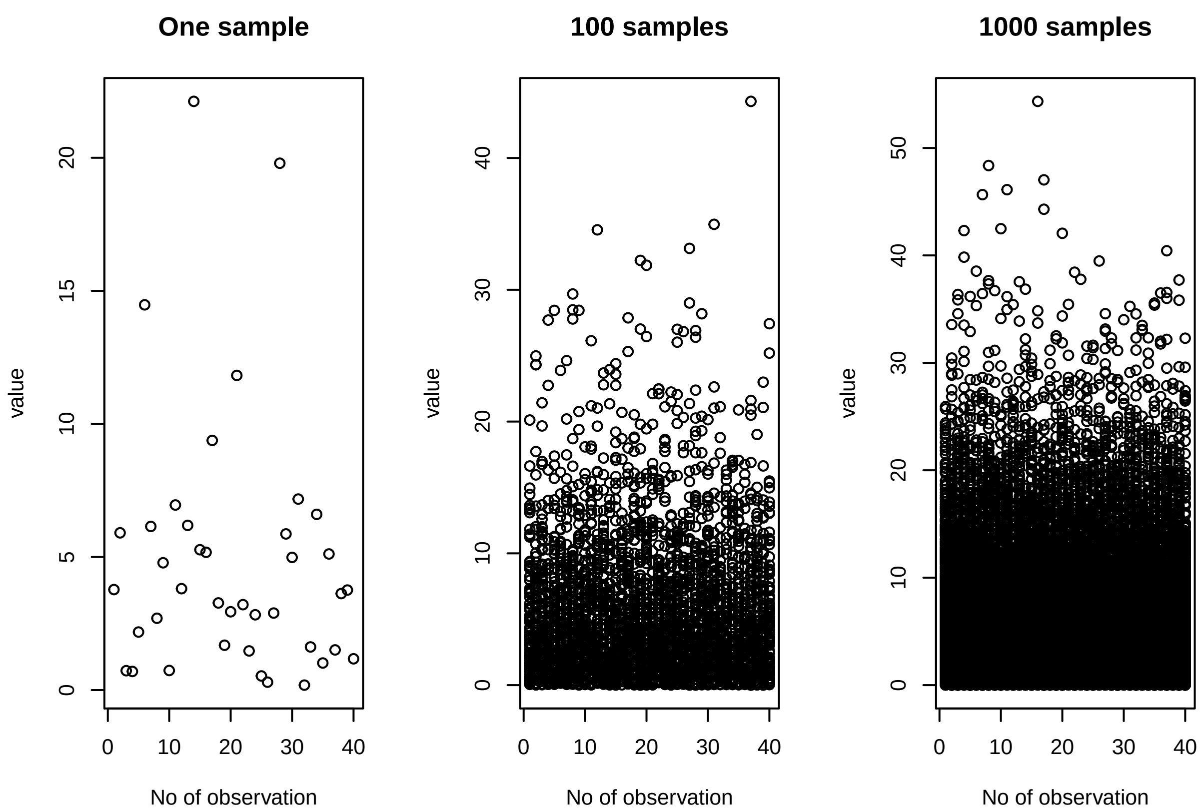
Plot the simulated data

```
par(mfrow=c(1,3))
plot(1:n, sim_matrix[1,], xlab="No of observation", ylab="value",
     main = "One sample")

temp <- rep(1:n, 100)
plot_matrix_100 <- matrix(temp, nrow=100, ncol=n)
plot(plot_matrix_100,sim_matrix[1:100,], xlab="No of observation", ylab="value",
                      main = "100 samples")

temp <- rep(1:n, 1000)
```

```r
plot_matrix_1000 <- matrix(temp, nrow=1000, ncol=n)
plot(plot_matrix_1000,sim_matrix, xlab="No of observation", ylab="value",
                           main = "1000 samples")
mtext("Observation in exponential distribution", outer = TRUE, cex = 1)
```



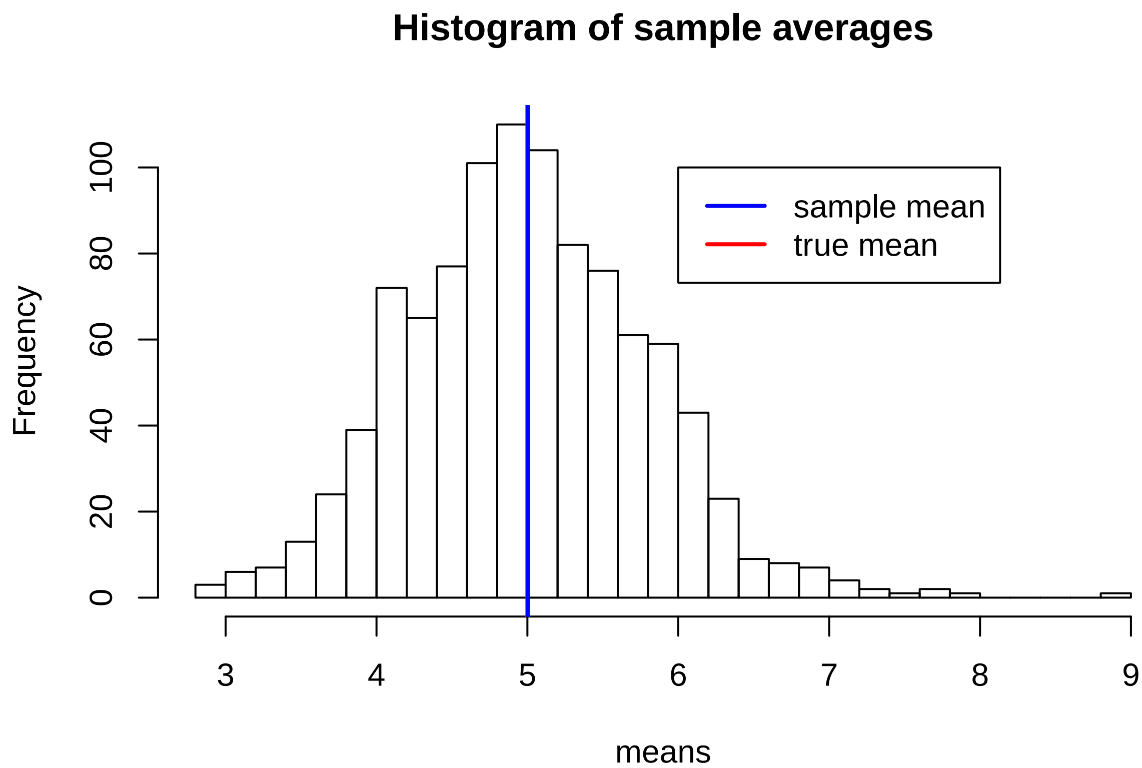

```r
par(mfrow=c(1,1))
```

Sample means versus theoretical mean

```r
means <- apply(sim_matrix, 1, mean)
length(means)
```

```
## [1] 1000
```

```r
hist(means, breaks=40, main="Histogram of sample averages")
legend(6,100, c("sample mean", "true mean") , lwd= c(2,2), lty = c(1,1), col=c("blue", "red"))
abline(v=1/lambda, lwd=2, col="red")
abline(v = mean(means), lwd=2, col="blue")
```




```
paste("Average of sample means:", round(mean(means),3), "vs",
      "theoretical mean:", round(1/lambda,3))
```

```
## [1] "Average of sample means: 5.002 vs theoretical mean: 5"
```

Variance of a sample

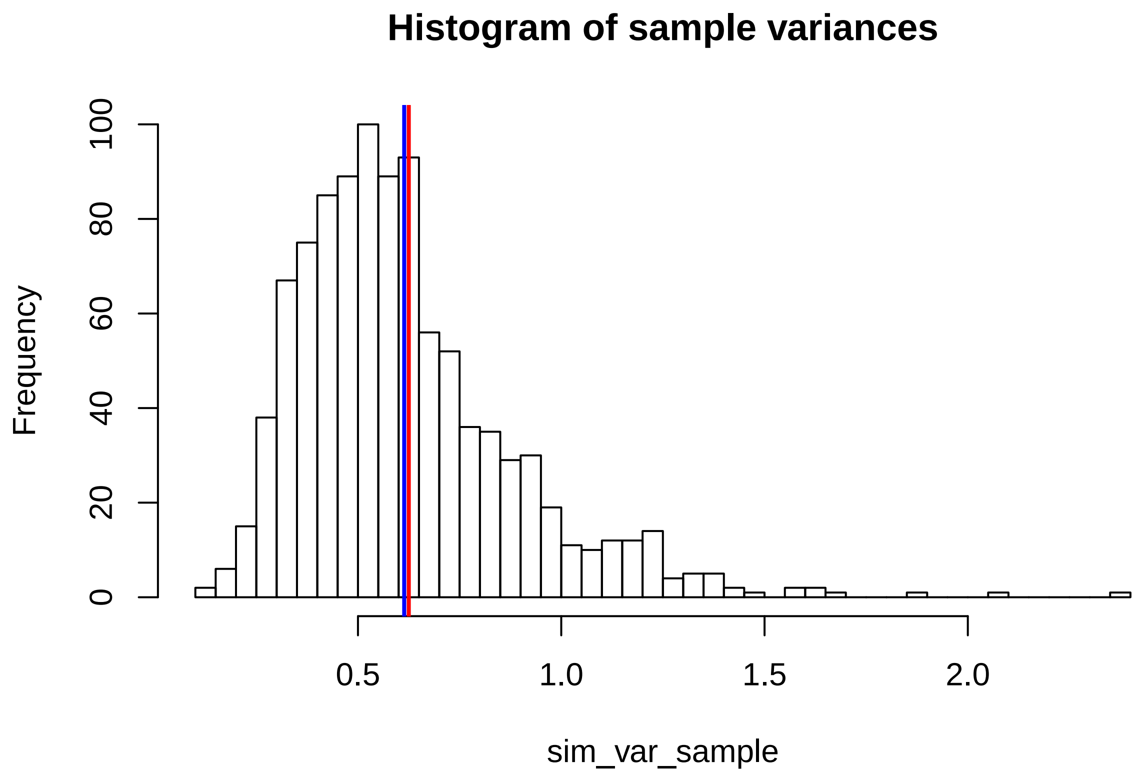
true_var_sample contains theoretical variation of the sample. sim_var_sample is vector with variation for each sample in matrix sim_matrix.The "Variance of sample" is a mean of the variances of each individual sample (composed of 40 observations) and it is compared with the theoretical variance.

```
true_var_sample <- true_sd^2/n

# Calculate the sample variance
sim_var_sample <- vector(mode="numeric", length=sim_n)
for(i in 1:sim_n){
        sim_var_sample[i] <- var(sim_matrix[i,])/n
}
length(sim_var_sample)
```

```
## [1] 1000
```

```
hist(sim_var_sample, breaks=40, main = "Histogram of sample variances")
legend(1,300, c("sample variance", "theoretical variance"),
       lwd= c(2,2), lty = c(1,1), col=c("blue", "red"))
abline(v=true_var_sample , lwd=2, col="red")
abline(v = mean(sim_var_sample), lwd=2, col="blue")
```




```
paste("Variance of sample:", round(mean(sim_var_sample),3), "vs",
      "theoretical variance:", round(true_var_sample,3))
```

```
## [1] "Variance of sample: 0.614 vs theoretical variance: 0.625"
```

The variance of 1000 averages:

```
var(means)/1000
```

```
## [1] 0.0006308244
```

The sample averages (stored in variable "mean") are more concentrated than the individual observations (sim_matrix). Therefore the variance of the averages of samples is smaller than the variance of individual observations.

Distributions

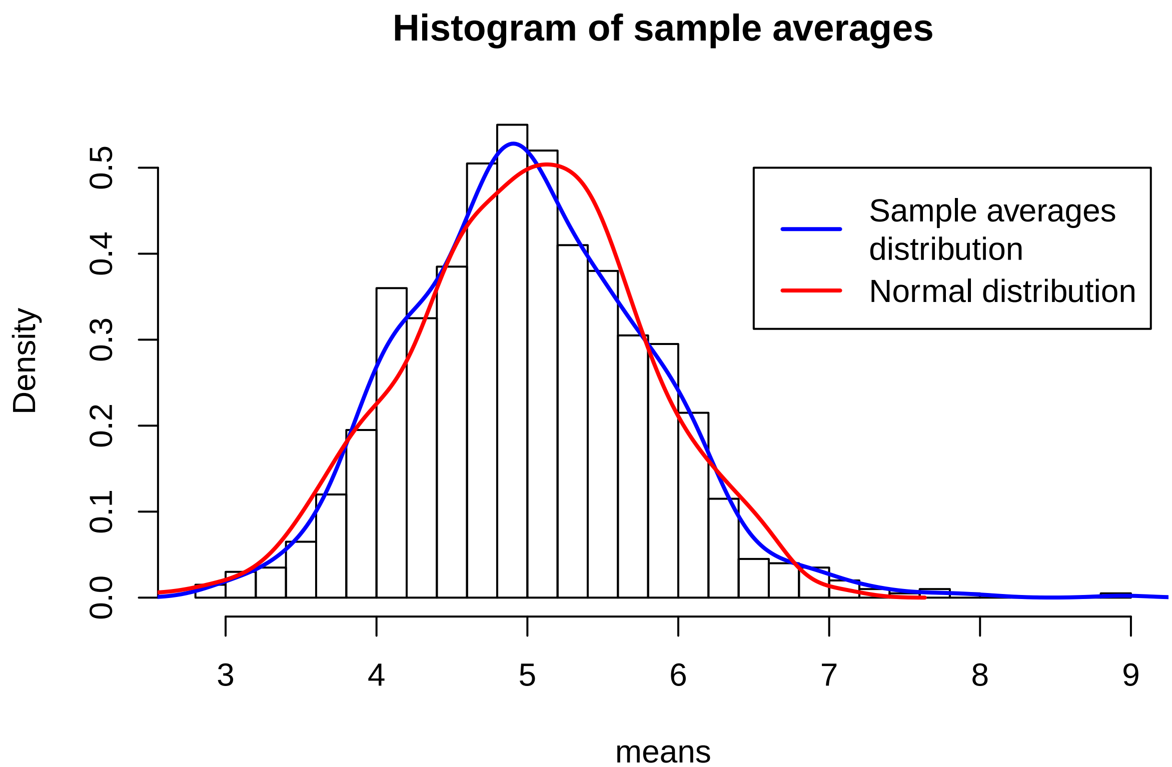
Create 1000 random observation with normal distribution, with mean and standard deviation equal to the analyzed exponential distribution.

```
zvals <- seq(min(means), max(means), length=sim_n)
set.seed(i)
zvals_normal <- rnorm(zvals, mean= 1/lambda, sd=(1/lambda)/sqrt(n))
```

Plot the distribution of averages of exponential distributed observations and normal distributed observations.

```
hist(means, breaks=40, freq = FALSE, main= "Histogram of sample averages")
lines(density(means), col="blue", lwd=2)
lines(density(zvals_normal), col="red", lwd=2)
```

```r
legend(6.5,0.5, c("Sample averages \ndistribution", "Normal distribution"),
       lwd= c(2,2), lty = c(1,1), col=c("blue", "red"))
```



Test if observations in the two distributions are different.

H0 - the difference between two distributions is equal 0. Ha - the difference between two distributions is not equal 0.

Assumptions:

The analyzed variables are independent and identically distributed.

```r
t.test(means, zvals_normal)
```

```
##
##  Welch Two Sample t-test
##
## data:  means and zvals_normal
## t = 0.34713, df = 1996.9, p-value = 0.7285
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05668005  0.08106025
## sample estimates:
## mean of x mean of y
##  5.002327  4.990137
```

The p-value = 0.7285, meaning that for significance level, alfa =0.05, we fail to reject H0. The confidence interval contain 0, meaning that there is no significant difference between the two distributions.