Statistical Inference Course Project 1: Playing with the Central Limit Theorem

Josep Anton Mir Tutusaus

Monday, June 15, 2015

1. Overview

In this report I'm going to play around a bit with the distribution of 1000 exponentials and the distribution of 1000 means of 40 exponentials, with comments on properties of both distributions and on the Central Limit Theorem (CLT from this point onwards).

It is divided as the exercise proposes, and it is 5 pages long. The images are included in the report, so an appendix is not needed.

2. Simulations

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.

In this chunk I'm setting lambda to be 0.2, n to be 40 and nsim to be 1000.

```
lambda <- 0.2
nsim <- 1000
n <- 40
```

Simulation of 1000 random exponentials, using nsim=1000 and lambda=0.2.

```
expsim <- rexp(nsim, lambda)
```

Simulation of 1000 averages of 40 random exponentials, using the code from the example.

```
meanexpsim = NULL
for (i in 1 : nsim) meanexpsim = c(meanexpsim, mean(rexp(n, lambda)))
```

3. Sample Mean versus Theoretical Mean

- The theoretical mean is 1/lambda, that is 5
- The sample mean is 5.073966 when taking 40 simulations and 5.0546147 when taking 1000 simulations.

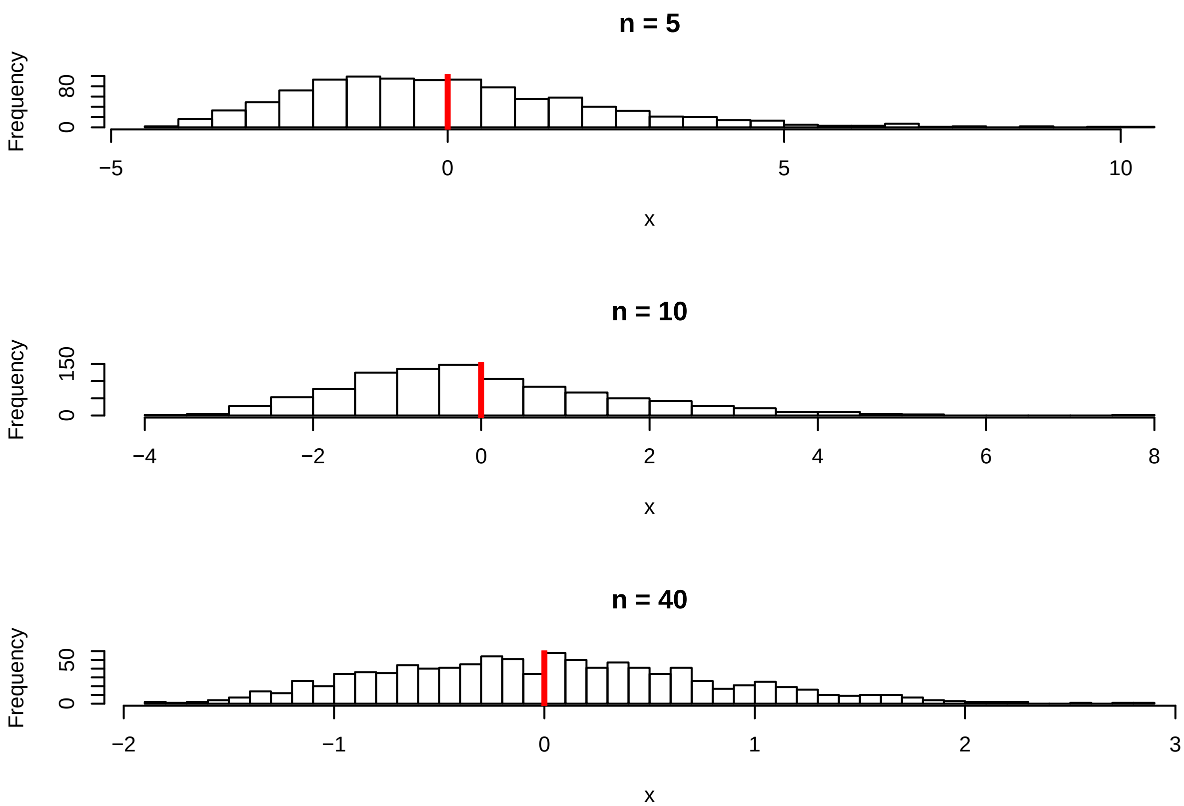
This shows how the higher the sample size, the closer it is his mean to the population average mu that is trying to estimate.

It can be seen as well from these histograms:

```
par(mfrow=c(3,1))
exp = NULL
for (i in 1 : nsim) exp = c(exp, mean(rexp(5, lambda)))
hist(exp-1/lambda, main="n = 5", xlab="x", breaks=40)
abline(v=0, col="red", lwd=3)

exp = NULL
for (i in 1 : nsim) exp = c(exp, mean(rexp(10, lambda)))
hist(exp-1/lambda, main="n = 10", xlab="x", breaks=40)
abline(v=0, col="red", lwd=3)

exp = NULL
for (i in 1 : nsim) exp = c(exp, mean(rexp(40, lambda)))
hist(exp-1/lambda, main="n = 40", xlab="x", breaks=40)
abline(v=0, col="red", lwd=3)
```



As n accumulates more data, the distribution is more centered around 0 (because we substracted mu).

4. Sample Variance versus Theoretical Variance

The variance gives us an idea of how variable a distribution is. - The theoretical variance is 1/lambda^2, that is 25 - The sample variance is 24.7768795 when taking 40 simulations and 22.4534934 when taking 1000 simulations.

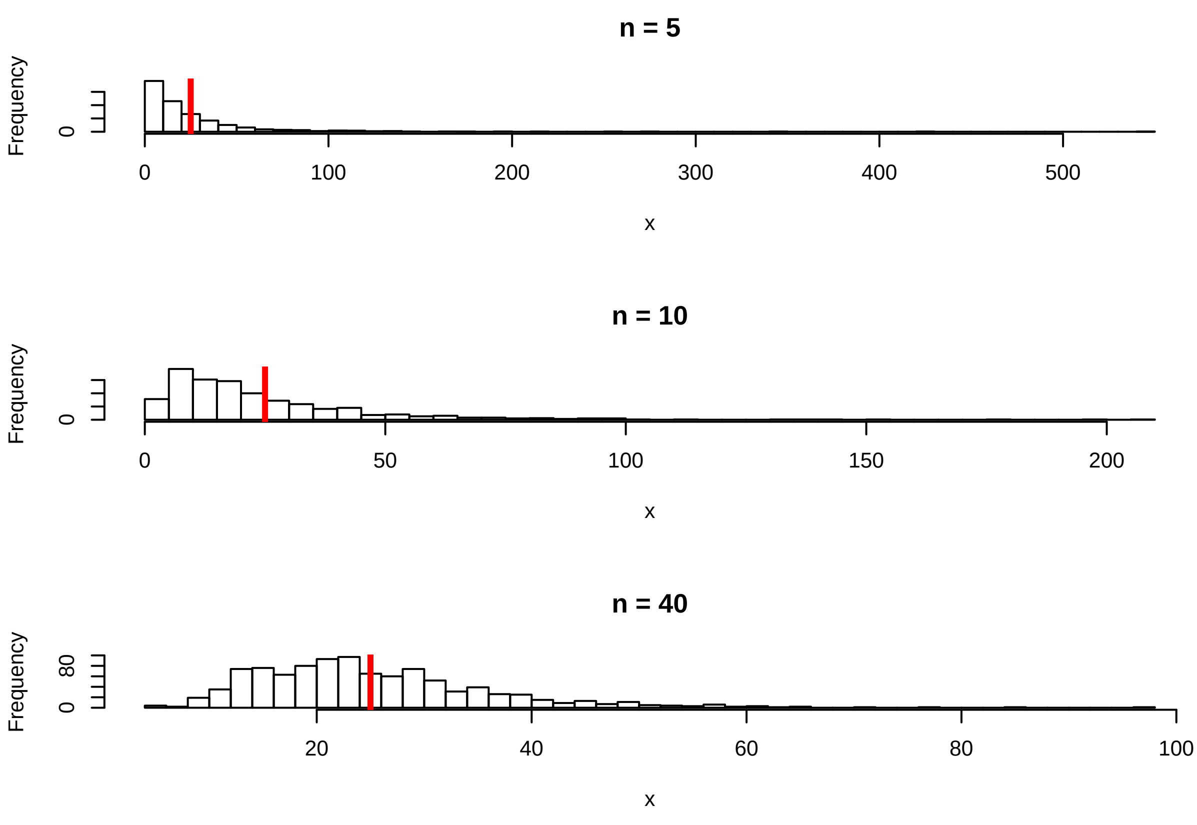
This shows how the higher the sample size, the closer it is the sample variance to the population variance that is trying to estimate.

It can be seen as well from these histograms:

```
par(mfrow=c(3,1))
exp = NULL
for (i in 1 : nsim) exp = c(exp, var(rexp(5, lambda)))
hist(exp, main="n = 5", xlab="x", breaks=40)
abline(v=25, col="red", lwd=3)

exp = NULL
for (i in 1 : nsim) exp = c(exp, var(rexp(10, lambda)))
hist(exp, main="n = 10", xlab="x", breaks=40)
abline(v=25, col="red", lwd=3)

exp = NULL
for (i in 1 : nsim) exp = c(exp, var(rexp(40, lambda)))
hist(exp, main="n = 40", xlab="x", breaks=40)
abline(v=25, col="red", lwd=3)
```

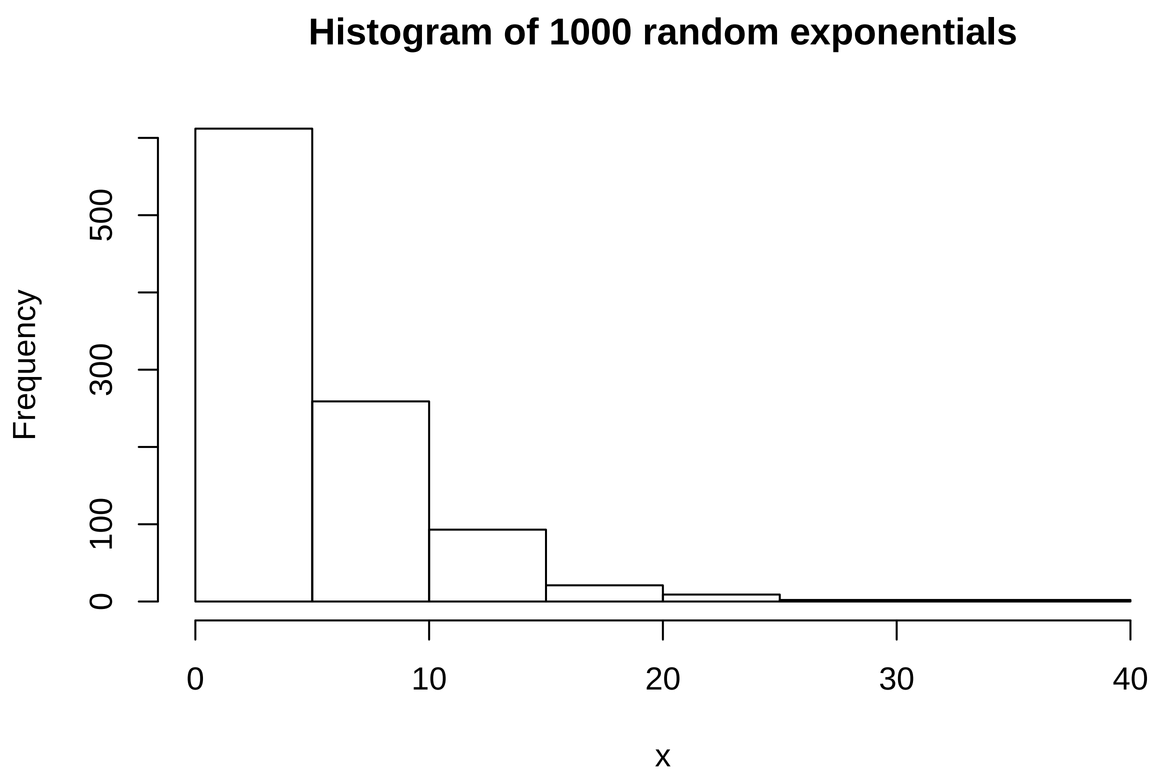


As n accumulates more data, the distribution is more centered around 25, the theoretical population variance.

5. Distribution

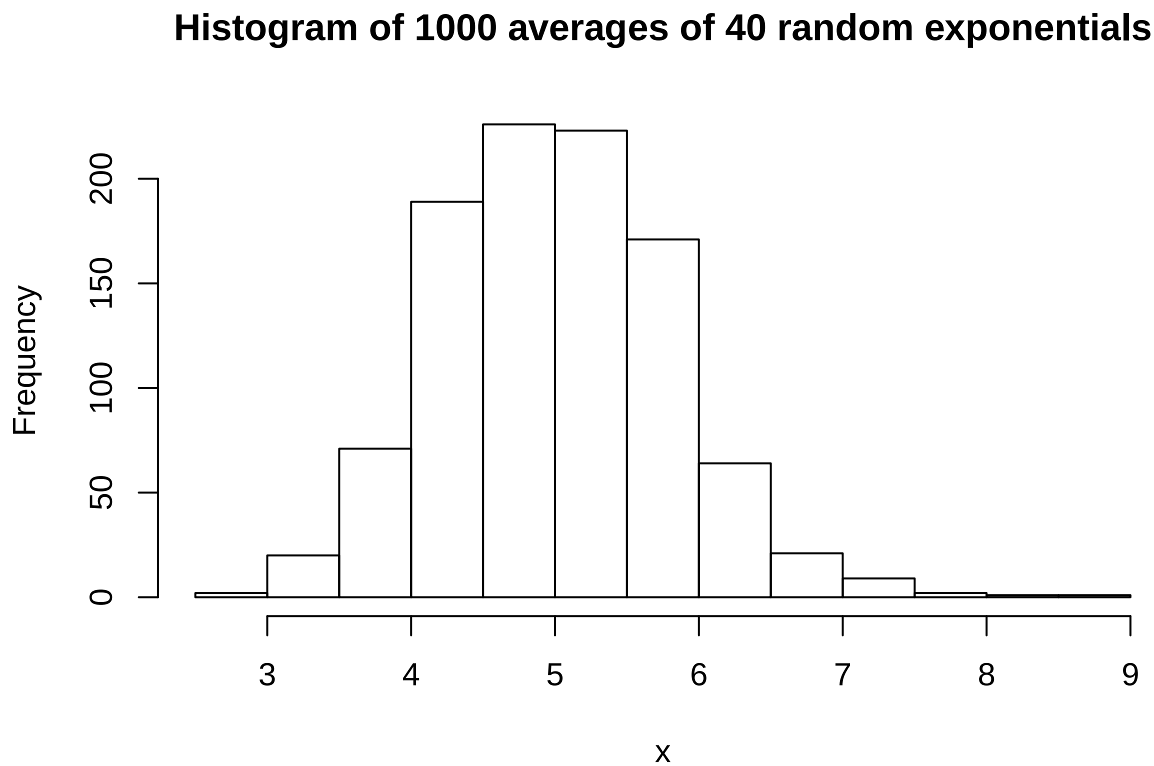
Distribution of 1000 random exponentials

```r
hist(expsim, main="Histogram of 1000 random exponentials", xlab="x")
```



Distribution of 1000 averages of 40 random exponentials

```r
hist(meanexpsim, main="Histogram of 1000 averages of 40 random exponentials", xlab="x")
```



As we can see, the CLT applies perfectly here. A 1000 averages of 40 exponentials are, in fact, a 1000 independent and identically distributed (iid) random variables, and as such, they will be approximately normally distributed, regardless of the underlying distribution.