Stadistical Inference - Course Project - Part 1 - Simulation Exercise

Jose Maria Vilaverde (jmvilaverde)

Saturday, May 23, 2015

Overview

In this report is investigated the exponential distribution in R and compare it with the Central Limit Theorem.

The CLT states that the distribution of averages of iid variables (properly normalized) becomes that of a stadard normal as the sample size increases.

The result is that $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{sd} = \frac{Estimate - Meanofestimate}{Std.Err.ifestimate}$ has a distribution like that of a standard normal for large n.

Conditions for investigation

For this investigation, the exponential distribution is simulated with R function rexp(n, lambda), where lambda is the rate parameter.

The mean of estimate (μ) of exponential distribution is $\frac{1}{\lambda}$ and the standard deviation (σ) is also $\frac{1}{\lambda}$. Lambda (λ) is set to 0.2 for all simulations.

- $\lambda = 0.2$
- $\mu = \frac{1}{\lambda} = \frac{1}{0.2} = 5$
- $\sigma = \frac{1}{\lambda} = \frac{1}{0.2} = 5$

The distribution of averages of 40 exponentials are investigated in a thousand simulations.

Steps

1. Sample mean and comparation with the theoretical mean of the distribution.

Sample mean:

Is obtained a sample of one thousand exponentials and is calculated the sample mean $\bar{X}_n = \frac{\Sigma_{i=1}^{n} X_i}{n}$

> **Parameters**: $\lambda = 0.2$, n = 1000
>
> **Sample mean**: $\bar{X}_n = \frac{\Sigma_{i=1}^{n} X_i}{n} = 4.992246$ ***(View Code 1./Figure 1.)***

Theoretical mean:

For calculate the theoretical mean of the distribution we use the Central Limit Theorem formula:

CLT Formula: $\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n-\mu)}{sd}$

For parameters: n = 40, $\lambda = 0.2$, $\mu = \frac{1}{\lambda}$, $\sigma = \frac{1}{\lambda}$, $\bar{X}_n = 4.992246$

Is obtained the **theoretical mean** of the distribution: $\frac{\sqrt{n}(\bar{X}_n-\mu)}{sd}$ = -0.0156241 ***(View Code 2./Figure 2.***

Comparation: ***(View Figure 3.)***

Sample mean	Theoretical mean
4.992246	-0.0156241

---

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

Sample variance:

Formula:$\mathrm{Var}(\bar{X}) = \sigma^2$

Sample variance: $\mathrm{Var}(\bar{X}) = \sigma^2/n = 24.8067953$ ***View Code 3.***

Theoretical variance of the distribution:

Formula: $\mathrm{Var}(\bar{X}) = \sigma^2/n$

Parameters: $\sigma = \frac{1}{\lambda}$, $\lambda = 0.2$

Theoretical variance: $\mathrm{Var}(\bar{X}) = \sigma^2/n = (1/\lambda)^2/n = 0.625$ ***View Code 4.***

Comparation: ***(View Figure 3.)***

Sample variance	Theoretical variance
24.8067953	0.625

---

3. Show that the distribution is approximately normal.

In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

(View figure 4.)

Simulations

1. Sample mean and comparation with the theoretical mean of the distribution.

Sample mean:

Code 1:

```
#Calculate the sample
lambda = 0.2
sampleN = 1000 #size of the sample
sample = NULL
for (i in 1:sampleN) sample = c(sample, rexp(sampleN,lambda))

#Mean of the sample
sampleMean = mean(sample)

#Generation of graphic of sample and sample mean
# hist(sample/sampleN, n = sampleN, main="Distribution of sample of 1000 elements", xlab="Exponential v
# abline(v=sampleMean, col="red")
#library(ggplot2)
#g <- ggplot(data.frame(x = 1:sampleN, y = sample), aes(x = x, y = y))
#g <- g + geom_hline(yintercept = 0) + geom_line(size = 2)
#g <- g + labs(title = "Distribution of sample of 1000 elements", x ="Exponential value")
#g
```

Figure 1.

Theoretical mean:

Code 2:

```
#Calculate the distribution of the averages
n = 40 #Number of elements to use each simulation
simulations = 1000 #Number of simulations
lambda = 0.2
mu = 1/lambda
sd = 1/lambda
sampleCLT = NULL
for (i in 1:simulations) sampleCLT = c(sampleCLT, (sqrt(n)*(mean(sample(sample, n))-mu))/(sd))
theoMean = mean(sampleCLT)
```

```r
#Generation of graphic of sample and sample mean
hist(sampleCLT, n=50, main="Distribution of sample of 1000 iterations - 40 elements each iteration", xla
abline(v=theoMean, col="red")
```

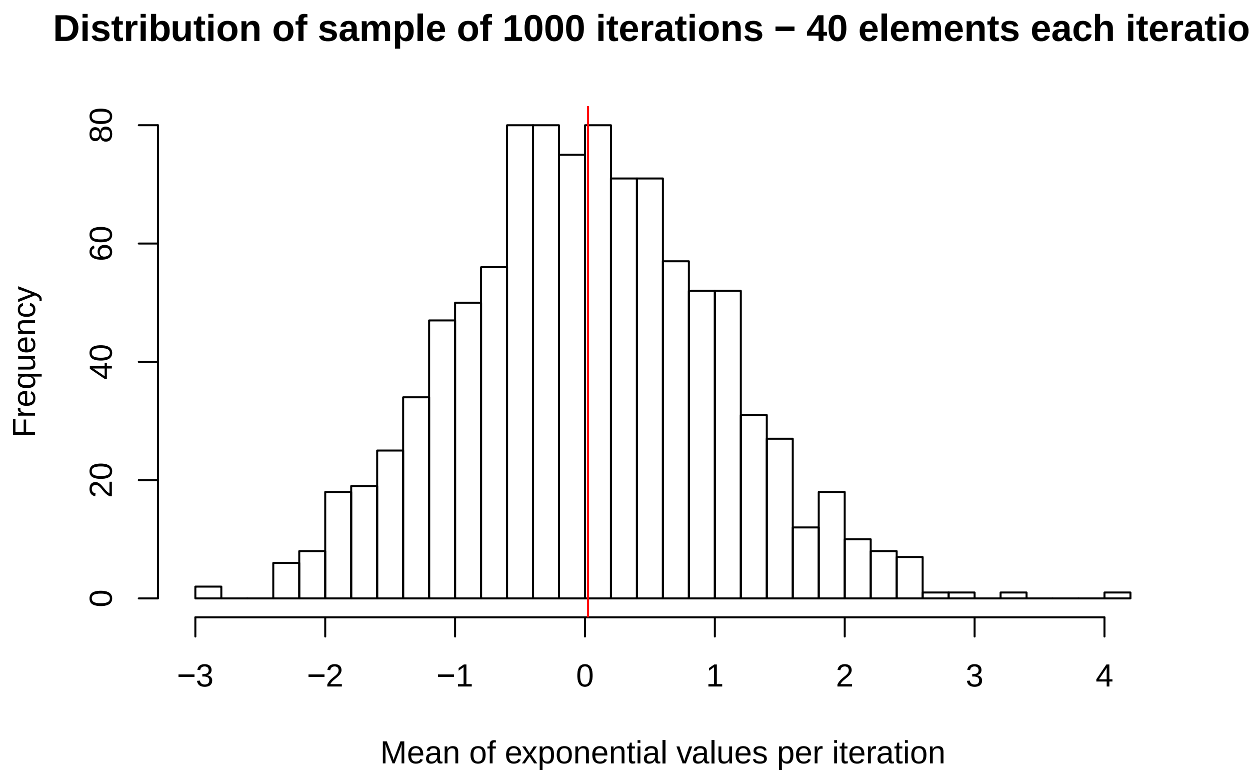


Figure 2.

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

Code 4 - Sample variance:

```r
#Calculate the sample variance:
sampleVar = sd(sample)^2
```

Code 5 - Theoretical variance of the distribution:

```r
n = 40
lambda = 0.2
sigma = (1/lambda)
theoVar = sigma^2/n
```

Code ? - Comparation of variances:

```
plot(c(sampleVar, theoVar), c(0,0), type = "h", xlim=c(theoVar-2, theoVar+2), ylim = c(0,1),
     main= "Variance values", xlab="Variance", ylab=NA)
legend("topright", legend=c("Sample Variance", "Theorical Variance"), col=c("red", "blue"), pch = c(1,1
abline(v= sampleVar, col = "red")
abline(v= theoVar, col = "blue")
```

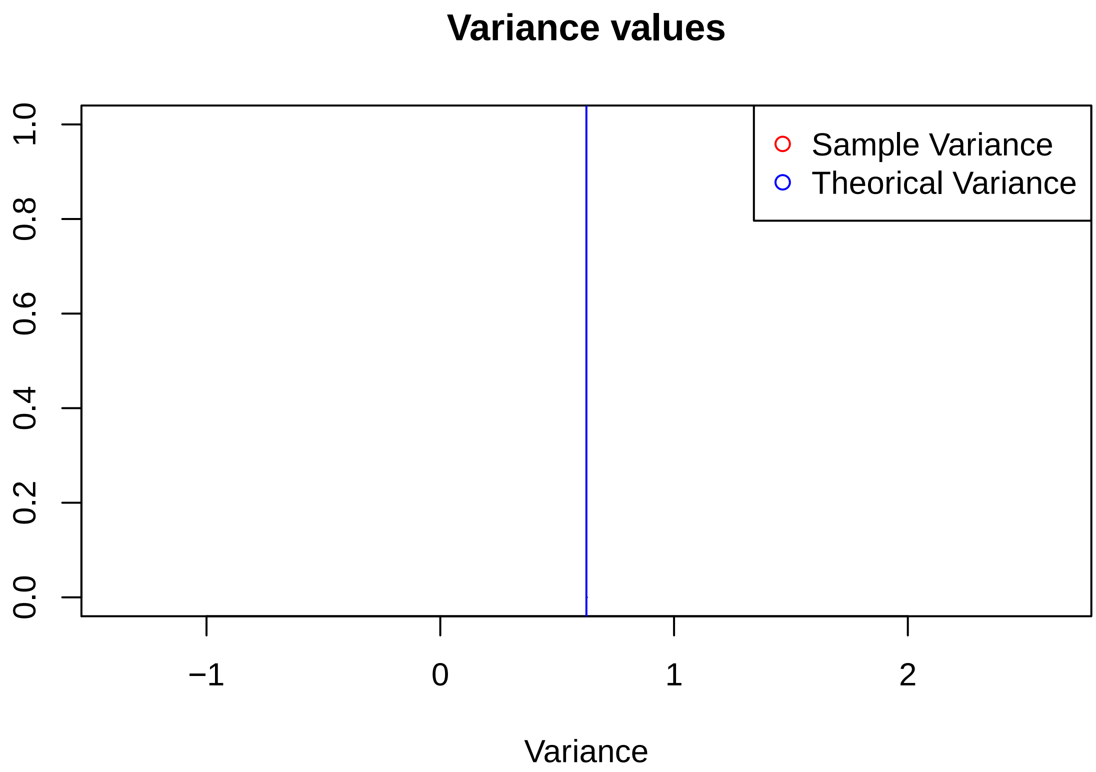


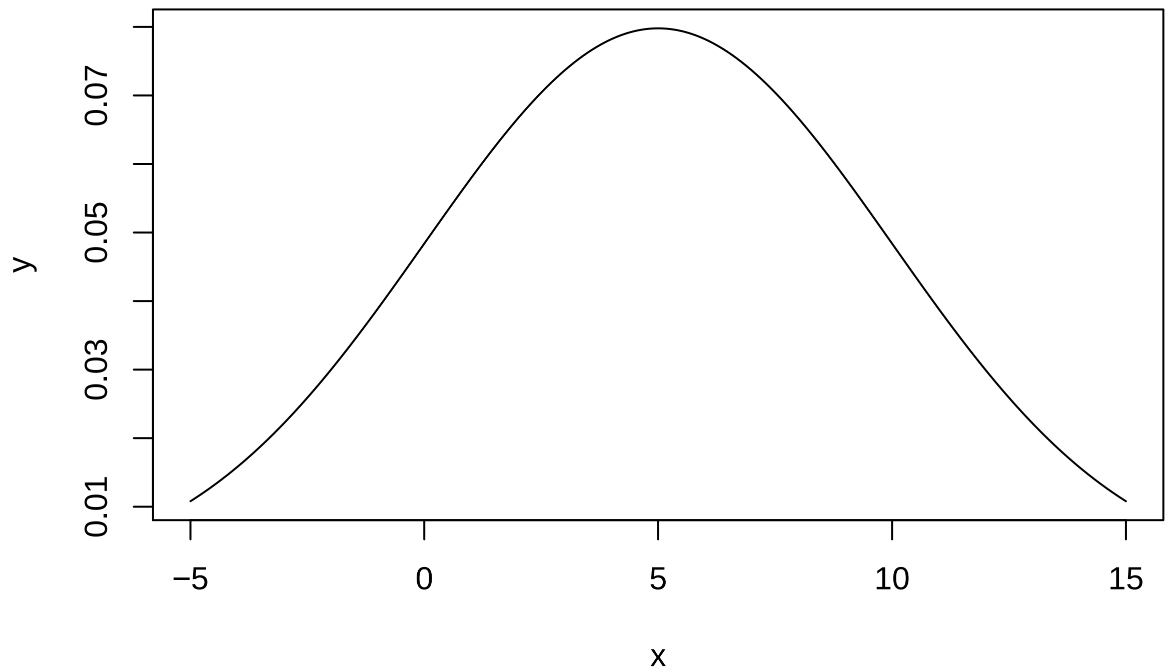
Figure ? - Comparation of variances.

3. Show that the distribution is approximately normal.

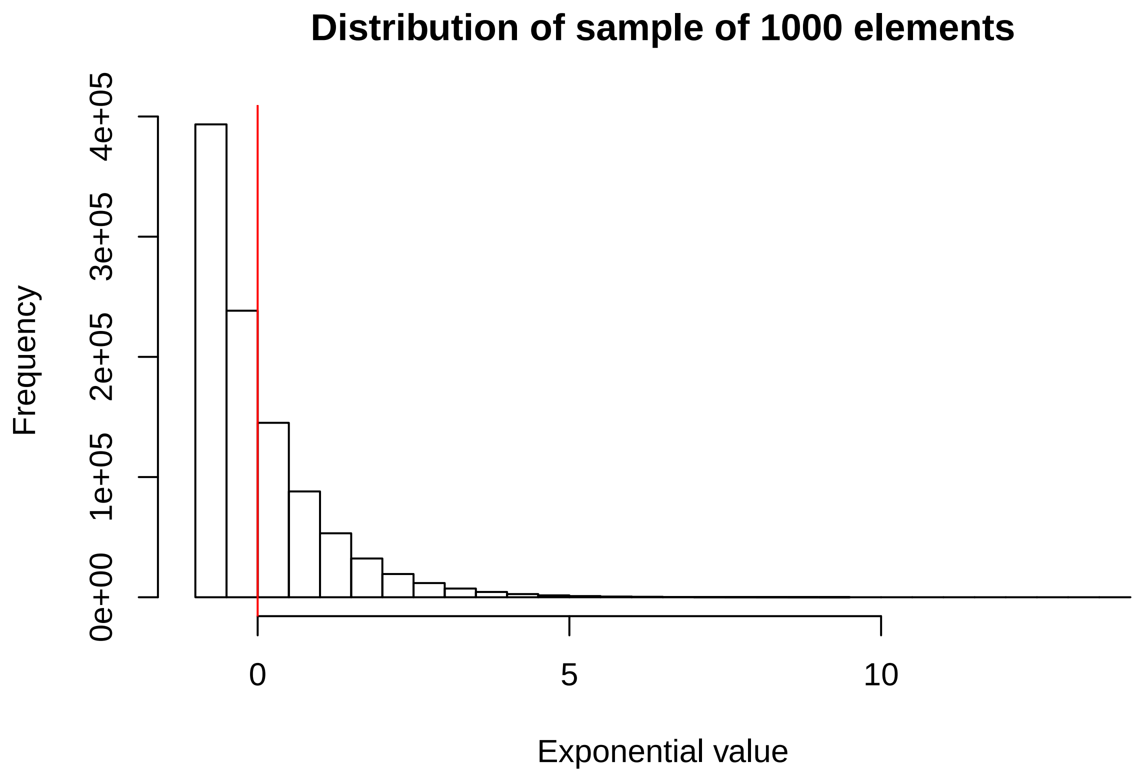
In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

Code 6.

```
x   <- seq(mu-2*sd,mu+2*sd,length=1000)
y   <- dnorm(x,mean=mu, sd=sd)
plot(x,y, type="l", lwd=1)
```

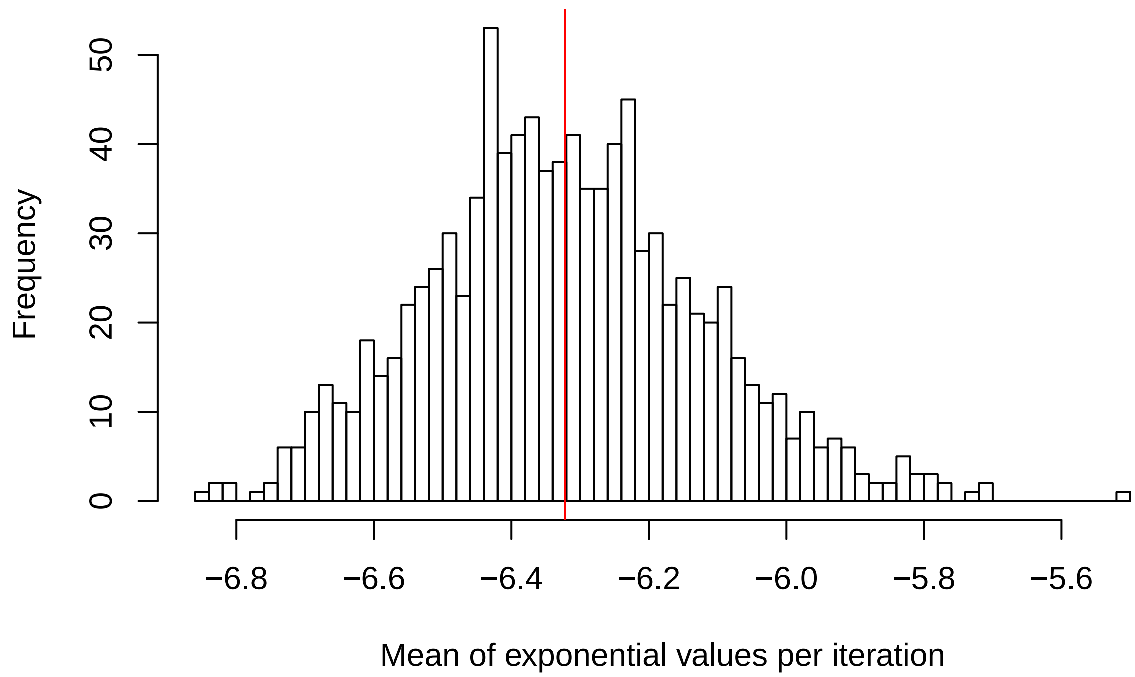



```
#Simple distribution normalized
lambda = 0.2
mu = 1/lambda
sd = 1/lambda
#size of the sample
sampleN = 1000
sample = NULL
for (i in 1:sampleN) sample = c(sample, (rexp(sampleN,lambda)-mu)/sd)
#Mean of the sample
sampleMean = mean(sample)
hist(sample, n = 50, main="Distribution of sample of 1000 elements", xlab="Exponential value")
abline(v=sampleMean, col="red")
```

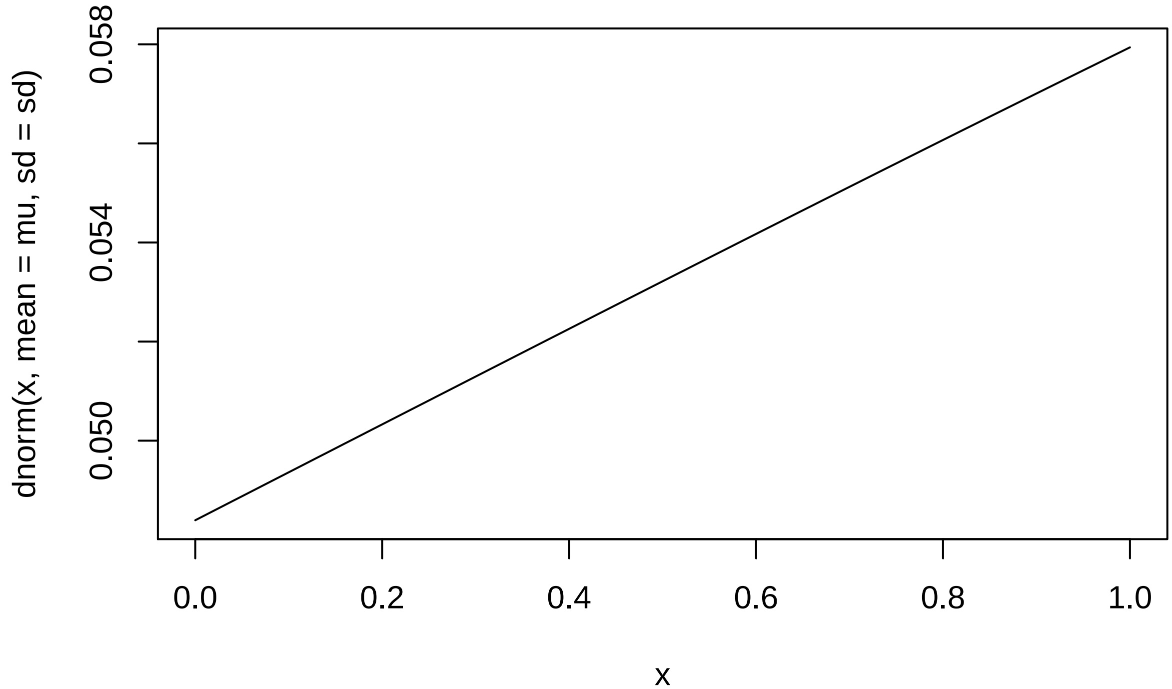



```
##Sample of averages
n = 40
lambda = 0.2
mu = 1/lambda
sd = 1/lambda
sampleCLT = NULL
for (i in 1:sampleN) sampleCLT = c(sampleCLT, (sqrt(n)*(mean(sample(sample, n))-mu))/(sd))
theoMean = mean(sampleCLT)
hist(sampleCLT, n=50, main="Distribution of sample of 1000 iterations - 40 elements each iteration", xla
abline(v=theoMean, col="red")
```

Distribution of sample of 1000 iterations − 40 elements each iteratio




```
curve(dnorm(x,mean=mu, sd=sd))
```


dnorm(x, mean = mu, sd = sd)
0.050
0.054
0.058
0.0
0.2
0.4
0.6
0.8
1.0
x