Project part 1: simulation exercise

Sylvain Tenier

16 novembre 2015

Overview

In this simulation project, we investigate the exponential distribution in R and compare it with the Central Limit Theorem. For this, we compare the distribution of 1000 exponentials with 1000 averages of 40 exponentials. We use 0.2 as the value of lambda for all of the simulations.

This report is written by applying the "Reproducible Research" principle of presenting the code along with the results. We show that although the exponential distribution is not normal, the distribution of its averages is normal.

Simulations

In this section, we present the R code for running 1000 simulations of the mean of 40 values from the exponential distrution. The code is explained through R comments

```r
# Set the number of simulations to 1000
nbsim <- 1000
# Set the experiment values
lambda = 0.2
n = 40
# Create a function that returns the mean of 40 values using the exponential distribution
simfunc <- function(x) mean(rexp(n, lambda))
# Make the experiment reproducible with a seed
set.seed(1979)
# Create a vector with the results
sim <- sapply(1:nbsim, simfunc)
# Display the first and last 5 values of the vector
head(sim)
```

```
## [1] 3.941081 4.484153 5.462776 4.498294 3.400810 4.665680
```

```r
tail(sim)
```

```
## [1] 3.463563 4.689028 3.818944 4.115459 4.210539 7.225896
```

Sample Mean versus Theoretical Mean

The theoretical mean for the exponentional distribution is $1/\lambda$. In our experiment $\lambda = 0.2$ so we expect a value of 5.

The sample mean is calculated using the following R code

```
sm <- mean(sim)
sm
```

```
## [1] 4.993022
```

We notice that 4.9930222 is very close to the expected value of 5

In this section, we highlight the means we are comparing.

Sample Variance versus Theoretical Variance

According to the CLT, the expected variance is $\frac{(1/\lambda)^2}{n}$. In our case $\lambda = 0.2$ and $n = 40$ so we should get $\frac{(1/0.2)^2}{40}$ which gives 0.625.

The sample variance is obtained by squaring the result of the call to R's sd() function

```
sv <- (sd(sim))^2
sv
```

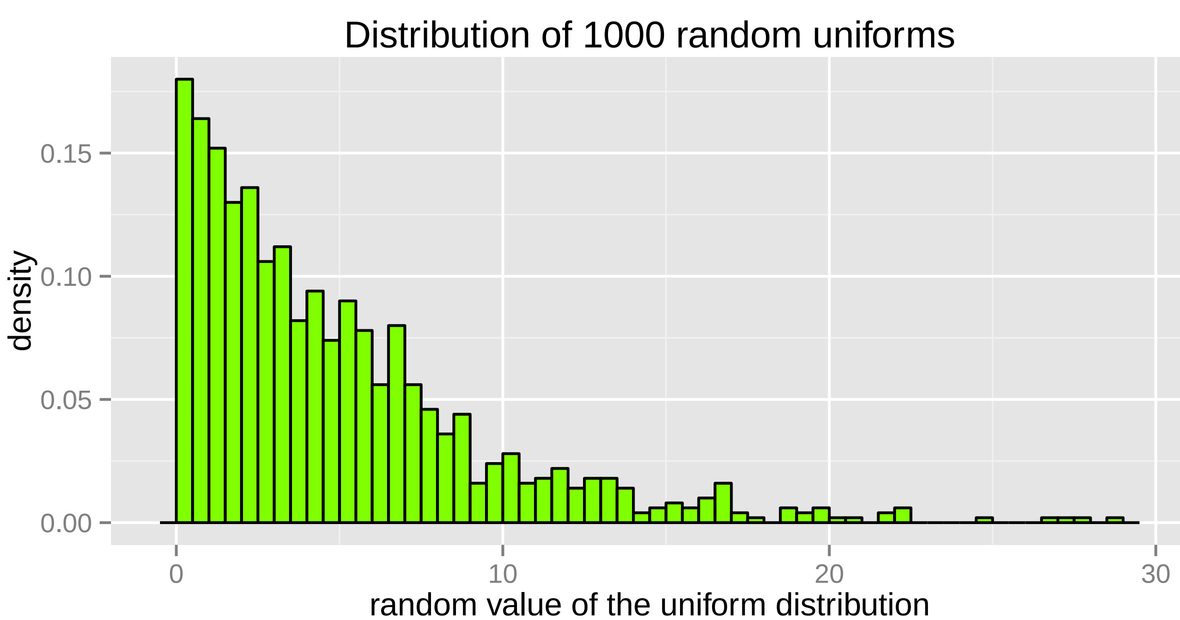
```
## [1] 0.6496487
```

We get a sample variance (0.6496487) that is close to the expected value (0.625) of the variance.

Distribution

First we show that the distribution of 1000 random uniforms is by itself not normal.

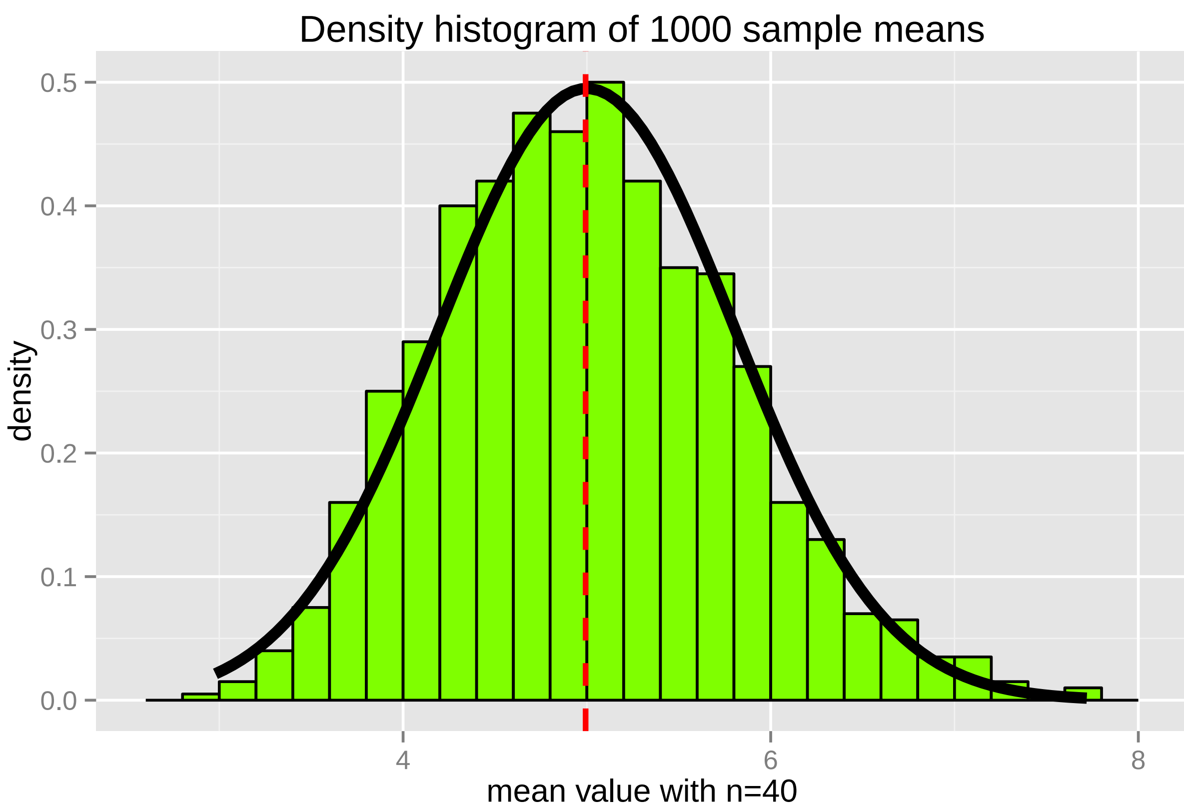
```
library(ggplot2)
df<-data.frame(dist<-rexp(1000, lambda))
ggplot(data=df,aes(x = dist)) +
  geom_histogram(aes(y = ..density..),binwidth = 0.50,colour="black",fill="chartreuse1")+
  labs(title="Distribution of 1000 random uniforms", x="random value of the uniform distribution")
```



However, when we draw

1. a density histogram for the 1000 simulated sample means
2. the normal distribution with the expected mean and standard deviation.

```
library(ggplot2)
df<-data.frame(sim)
ggplot(data=df,aes(x = sim)) +
    geom_histogram(aes(y = ..density..),binwidth = 0.20,colour="black",fill="chartreuse1")+
    stat_function(fun = dnorm,  size=2, arg = list(mean = 5, sd = sd(sim)))+
    geom_vline(aes(xintercept=mean(sim, na.rm=T)),   # Ignore NA values for mean
               color="red", linetype="dashed", size=1)+
    labs(title="Density histogram of 1000 sample means", x="mean value with n=40")
```



This figure demonstrates that the sample distribution of the means approximatively follows the normal distribution. The mean, displayed as a red dotted line, is as expected at a value close to 5.

Conclusion

The large number theory stipulates that with a large-enough number of samples, the distribution of the means of a non-normal distribution is normal. This simulation study demonstrates that this is the case for the exponential distribution with the distribution of the mean of 40 values.