Exponential Distribution and the Central Limit Theorem, a first approach

Overview

This project aims to explore the exponential distribution using the functionalities of R, and afterwards verifies its relation with the Central Limit Theorem (CLT). To do so, distributions of averages of 40 exponentials will be produced, and afterwards repeated a thousand times to obtain a large enough sample for the demonstration.

For this simulation, λ, the rate parameter, will be set at 0.2.

Simulations

The simulation consists of generating a sample of 40 random observations respecting the exponential function. The previous simulation will be ran a thousand times to obtain a large enough number of mean observations to test the Central Limit Theorem.

Functions used

Two functions will be used to generate the exponential function observations, and subsequently repeating it a thousand times.

1. **rexp** will generate *n* random observations of the exponential function with λ as rate parameter
2. **replicate** will reproduce the provided code the specified number of times (one thousand in the present scenario).

By definition, the exponential distribution generates the following parameters, with the subsequent results once applied to this case:

- Mean = $\lambda^{-1} = 0.2^{-1} = 5$
- Standard deviation = $\lambda^{-1} = 0.2^{-1} = 5$
- Variance = $(\lambda^{-1})^2 = \lambda^{-2} = 0.2^{-2} = 25$

Sample Mean versus Theoretical Mean

The measured mean provided by the thousand simulations is 5.052105 while the theoretical mean, as stated earlier, is 5. The absolute spread between the results and the theoretical value is therefore of 0.052105.

Graph 1, of **Annex B** shows the density distribution of the experiment. It therefore presents the thousand averages coming from the thousand samples of 40 random observations of the exponential function.

Centred around the theoretical mean (**red line**), a normal distribution (**blue line**) as been overlaid above the distribution to demonstrate the CLT. The actual mean of this distribution (**black line**) is shown as a comparison to expose the proximity to the theoretical one. The standard deviation of this overlaid normal distribution (0.7905694) is the theoretical one, in order to expose the proximity of the data with the theory, thus the normality of this distribution. As a matter of fact, the one measured from the data (0.8148614) proves to be different by 0.024292.

With both observed mean and standard deviation so close to their theoretical values, this experiment tends to validate the Central Limit Theorem.

Sample Variance versus Theoretical Variance

The theoretical variance of the distribution of the mean can be computed using the following formula

- $\sigma^2 = (\frac{\frac{1}{\lambda}}{\sqrt{n}})^2 = 0.625$

The measured variance provided by the thousand simulations is 0.6639991 while the theoretical variance, as provided by the previous calculation, is 0.625. The spread between the results and the theoretical value is therefore of 0.0389991, the experimental value proving to be quite close to the theoretical one.

Distribution

Despite the graphical apparence of normality by the distribution shown in Annex B, further tests should be conducted to validate it is normal, or very close to be.

Two such tests will be done:

1. Is the median similar to the observed and theoretical mean?
2. Are about 95% of the observations within 2 standard deviations (σ) from the mean?

Median tests

The median of the means distribution is 5.018556, for a spread of 0.018556 and 0.033549 with the theoretical and observed means respectively.

The median is very similar to the mean reinforcing the argument for a normal distribution.

Standard Deviation Tests

To be a normally distributed function 95.4% must be between 2σ from the mean.

To obtain this measure with the experimental data, the quantiles located at 2σ (3.576, 6.935) serve as boundaries to count the number of observations between them. Out of the 1000 observations, 954 happen to in this spectrum. This represents 95.4% of the observations within 2σ, respecting the theoretical behaviour of a normal distribution.

Conclusion

In the light of the previous experiment, evidence tends to demonstrate that despite an underlying distribution that is nowhere close to to a normal distribution, the distribution of its metrics, when observed over a large number of samples, will tend towards normality.

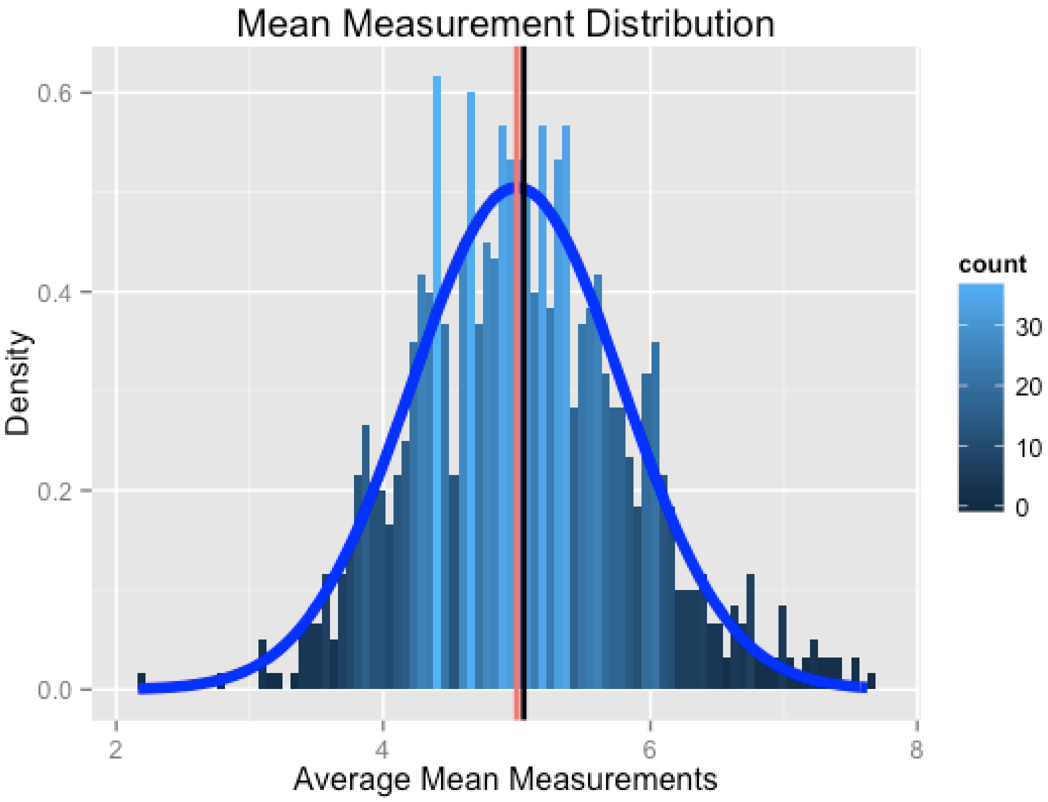
Annex A - The Code

```
# Setting up the initial parameters of the simulation
require(ggplot2)
require(scales)
λ <- 0.2
simul <- 1000
n <- 40

# Running a thousand simulations of 40 observations from the
exponential distribution of λ = 0.2
Sim <- replicate(simul,rexp(n,λ))

## Calculating and storing theoretical values of the exponential
function
# Theoretical mean
μ <- λ^-1
# Theoretical standard deviation
thStdDev <- λ^-1
# Theoretical variance
thVar <- thStdDev^2

# Reformating the object to data frame, because ggplot2 prefers data
frames
SimMeans <- data.frame(colMeans(Sim))
names(SimMeans) <- c("Observations")
mean(SimMeans$Observations)

measuredVariance <- var(SimMeans$Observations)
thSampSD <- (1/λ)/sqrt(n)
thSampVar <- ((1/λ)/sqrt(n))^2

#Storing the percentage of both tails to respect the 2σ rule
tail <- 1-.977249858
#Storing the respective quantiles
quantsMean <-quantile(SimMeans$Observations, c(tail,1-tail))
#Evaluating the percentage of observations between those quantiles
normalityMean <-
nrow(subset(SimMeans,SimMeans$Observations>=quantsMean[1] &
SimMeans$Observations<=quantsMean[2]))/nrow(SimMeans)

## Graph 1 - Distribution of Means
ggplot(SimMeans, aes(x=SimMeans$Observations))+geom_histogram(aes(y =
..density.., fill=..count..),binwidth=.06) + stat_function(fun = dnorm,
colour="blue", arg = list(mean = μ,sd=thSampSD), geom="line", size=2) +
geom_vline(aes(xintercept = μ, colour="red"),size=1) +
geom_vline(aes(xintercept = mean(SimMeans$Observations)),size=1) +
ylab("Density") + xlab("Average Mean Measurements") + labs(title="Mean
Measurement Distribution")
```

Annex B - Graphic

Graph 1 - Distribution of Means

```
## Warning: position_stack requires constant width: output may be
incorrect
```


Mean Measurement Distribution
Density
Average Mean Measurements
count