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 repeated a thousand times to obtain a large enough sample for the demonstration.

For this simulation, lambda, 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 run a thousand times in order to extract a large number of mean and variance 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 *lambda* 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 = $\text{lambda}^{-1} = 0.2^{-1} = 5$
- Standard deviation = $\text{lambda}^{-1} = 0.2^{-1} = 5$
- Variance = $(\text{lambda}^{-1})^2 = \text{lambda}^{-2} = 0.2^{-2} = 25$

Sample Mean versus Theoretical Mean

The measured mean provided by the thousand simulations is 5.011803 while the theoretical mean, as stated earlier, is 5. The absolute spread between the results and the theoretical value is therefore of 0.011803.

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 (**green line**) as been overlaid above the distribution to demonstrate the CLT. The standard deviation of this normal distribution is the one measured from the data, and is used in conjunction with the theoretical mean to show the normality of the distribution and thus, validating the Central Limit Theorem.

Sample Variance versus Theoretical Variance

The measured average variance provided by the thousand simulations is 25.1413108 while the theoretical variance, as stated simulation section of this document, is 25. The spread between the results and the theoretical value is therefore of 0.1413108.

Graph 2, of **Annex B** shows the density distribution of the experiment. It therefore presents the thousand variances coming from the thousand samples of 40 random observations of the exponential function.

Centred around the theoretical variance (**red line**), a normal distribution (**green line**) as been overlaid above the distribution to demonstrate the CLT. The standard deviation of this normal distribution is the one measured from the data, and is used in conjunction with the theoretical variance to show the normality of the distribution and validating once again, the Central Limit Theorem.

Distribution

Despite the graphical apparence of normality for both distributions shown in Annex B, further tests on both distributions should be conducted to validate they are normal, or very close to be.

Two tests will be conducted for each distribution:

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

Median tests

The median of the means distribution is 5.0047149, for a spread of 0.0047149 and 0.0070882 with the theoretical and observed means respectively.

The median of the variances distribution is 25.0662249, for a spread of 0.0662249 and 0.075086 with the theoretical and observed means respectively.

For both cases, the medians are very similar to both means hinting for a normal distribution.

Standard Deviation Tests

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

With regards to the means distribution, 95.4% of the observations are within 2 sigmas.

With regards to the variances distribution, 95.4% of the observations are within 2 sigmas.

Again, for both cases, the distributions are very close to a theoretical normal distribution and can thus be assumed to be normal

Conclusion

In the light of the previous two experiments, 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)
lambda <- 0.2
simul <- 1000
n <- 40

# Running a thousand simulations of 40 observations from the
exponential distribution of lambda = 0.2
Sim <- replicate(1000,rexp(simul,lambda))

## Calculating and storing theoretical values
# Theoretical mean
μ <- lambda^-1
# Theoretical standard deviation
thStdDev <- lambda^-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)

#preparing a void object for the for loop
variances <- NULL

# Looping through the thousand simulations to extract the variance of
every sample, and binding them together
for (i in 1:ncol(Sim)){
        variances <- rbind(variances,var(Sim[,i]))
}
# Reformating the object to data frame, because ggplot2 prefers data
frames
variances <- data.frame(variances)
names(variances) <- c("Observations")
#The average value of the observed variances
mean(variances$Observations)

#Storing the percentage of both tails to respect the 2 sigma 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)
```

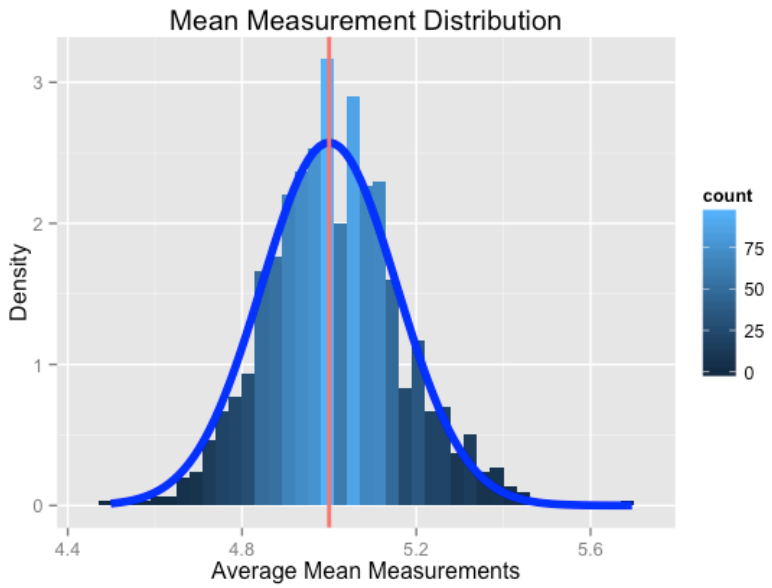
```
#Storing the respective quantiles
quantsVar <-quantile(variances$Observations, c(tail,1-tail))
#Evaluating the percentage of observations between those quantiles
normalityVar <-
nrow(subset(variances,variances$Observations>=quantsVar[1] &
variances$Observations<=quantsVar[2]))/nrow(variances)

## Graph 1 - Distribution of Means
ggplot(SimMeans, aes(x=SimMeans$Observations))+geom_histogram(aes(y =
..density.., fill=..count..),binwidth=.03) + stat_function(fun = dnorm,
colour="blue", arg = list(mean = μ,sd=sd(SimMeans$Observations)),
geom="line", size=2) + geom_vline(aes(xintercept =
mean(μ),colour="red"),size=1)+ ylab("Density") + xlab("Average Mean
Measurements") + labs(title="Mean Measurement Distribution")

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

Annex B - Graphics

Graph 1 - Distribution of Means



Graph 2 - Distribution of Variances