Data Science Specialism

Module 6 - Statistical Inference (June 2015)

Investigating a sample population of 40 random exponential values and their relationship to the natural exponential density function using the Central Limit Theorem and Law of Large Numbers.

Author: Fiona Young

Submission date: 14 June 2015

SUMMARY

The Central Limit Theorem states that the sampling distribution of any statistic will be normal or nearly normal if the sample size is large enough. The theoretical (exponential) density function (EDF) and sample distribution means were calculated as 0.975 and 0.9943 respectively. The theoretical (exponential) density function and sample distribution variances 1000 means of 40 exponential variables were 0.994 and 0.005 respectively. To accurately represent the EDF a sample population should have all its values centred around the mean value with a small variance and that is what we observe. However a better reflection of the exponential density function is to remove the first 150 of the 1000 calculated means as they vary considerably. What we discover is that the sample population mean 0.9871 is almost the same as the EDF mean 0.975 and the variance is almost zero at 0.00002. Using 40 exponential random variables is indeed enough to represent a EDF and satisfy the Central Limit Theorem.

INTRODUCTION

The Central Limit Theorem (CLT) definition on StatTrek (2015) states “that the sampling distribution of any statistic will be NORMAL or NEARLY NORMAL if the sample size is large enough”.

How do we know that the sample is large enough? We draw on the idea of the Law of Large Numbers (LLN). As we take a LLN (running mean) of the samples eventually the probability will be equivalent to the mean (mu) of a population density function. It is at this point of equivalence onwards that we can use the number of samples that accurately reflects the population density function.

According to some statistians a sample size of 30 to 40 values is large enough to be representative of a population density function (StatTrek 2015). Is this the case? Our sample size for this analysis has been set at 40 randomly generated exponential values.

We will attempt to answer three questions as they relate to an exponential density function and they are:

1. How does the theoretical mean and sample mean compare?
2. How does the theoretical variance and sample variance compare?
3. How does the theoretical density function and sample distribution curve differ?

What is an exponential density function?

Before proceeding to answer our three questions we ought to determine what is a exponential density function?

The equation for an exponential curve is given as:

```
y = e^x
```

where -

y = y axis of a graph x = x axis of a graph e is known as Euler’s number (Wikipedia 2015a) as described by Leonhard Euler (Wikipedia 2015b).

e is an irrational number and the base of Natural Logarithms which were invented by John Napier (Wikipedia 2015c).

e is calculated:

```
e = 1/n!
```

The value of e is known to 1.250 billion digits of accuracy (Gourdon 1999). Here are the first 20 digits to get you started.

```
e = 2.7182818284590452353...
```

METHODOLOGY / DATA PROCESSING

For this analysis I will generate five graphs:

Two histograms demonstrating a normal population density function (figure 1) and an exponential density function (figure 2) both using 1000 random variables.

A line graph (figure 3) demonstrating the Law of Large Numbers (LLN) and two histogram of the means of 40 random exponential variables

generated in 1000 simulations (figure 4) and with the first 150 mean values removed (figure 5).

In order to compare the distribution curves in figures 2, 4 and 5 requires the generation of mean and variance values (table 1).

RESULTS

The calculated mean and variance for an exponential density function, a sample population of means of 40 exponentials simulated 1000 times and a sample population of means of 40 exponentials simulated 1000 times with the first 150 samples removed.

```
options(scipen=999)
title <- "Table 1 Comparison of mean and variance"
EDF <- cbind("Exponential Density Function", round(mean(exponentials), digits = 4), round(var(exponentials), di
gits = 4))
SDC <- cbind("Sample distribution", round(mean(samplemeans), digits = 4), round(var(samplemeans), digits = 4))
RDC <- cbind("Sample distribution (first 150 variables removed)", round(mean(removed), digits = 4), round(var(r
emoved), digits = 4))
comparison <- rbind(EDF, SDC, RDC)
colnames(comparison) <- c("analysis?", "Mean", "Variance")

row.names(comparison) <- NULL
print(title)
```

```
## [1] "Table 1 Comparison of mean and variance"
```

```
comparison
```

```
##      analysis?                                              Mean     Variance
## [1,] "Exponential Density Function"                         "0.975"  "0.8372"
## [2,] "Sample distribution"                                  "0.9943" "0.0054"
## [3,] "Sample distribution (first 150 variables removed)" "0.9781" "0.0002"
```

The five graphs:

```
par(mfrow = c(1,2))

##1 - plotting the PDF graph
hist(population, col = "yellow", main = "Figure 1.  Normal population density function (PDF)\n (Gaussian densit
y distribution)", xlab = "Population variance\ndemonstrating the +/- 1st, +/- 2nd and +/- 3rd standard deviatio
ns", probability = TRUE)

##1.2 - plotting the density curve
lines(density(population), col = "black", lwd = 3)

    ##2 - graphic EDF
    hist(exponentials, prob = TRUE, col = "red", xlab = "Number of randomly exponential generated values", main
 = "Figure 2. Exponential density function (EDF)")

    ##2.1 - plotting the natural exponential line
    x.est <- fitdistr(exponentials, "exponential")$estimate

    curve(dexp(x, rate = x.est), add = TRUE, col = "black", lwd = 3)
```

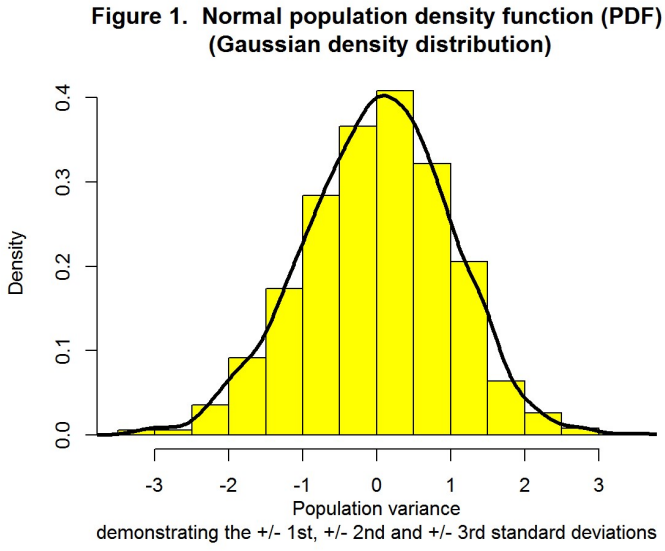


Figure 1. Normal population density function (PDF) (Gaussian density distribution)

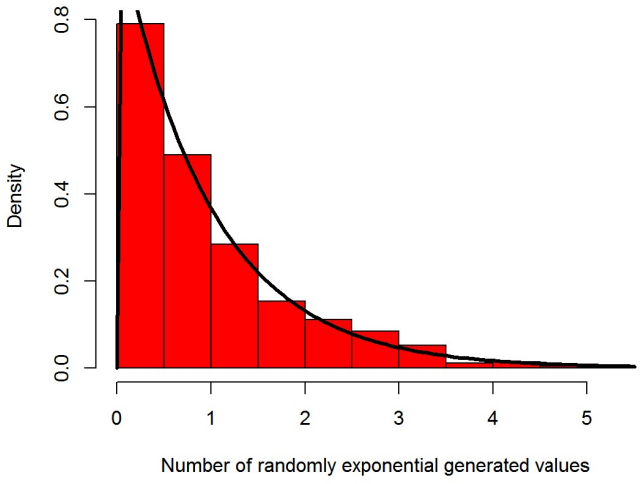
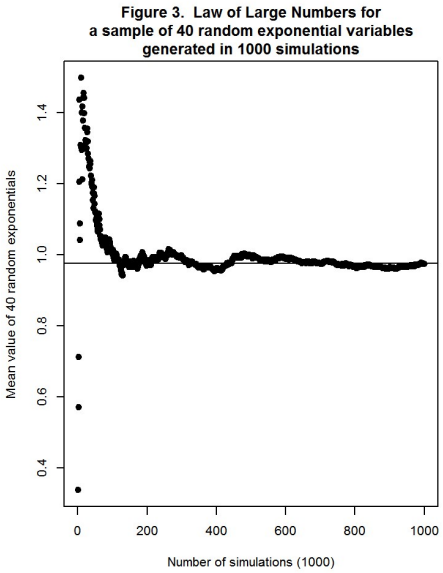
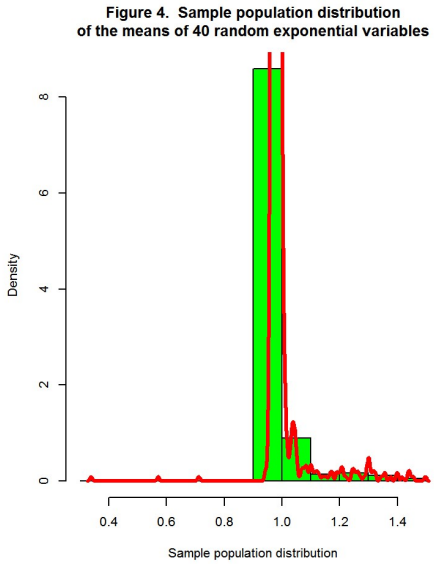


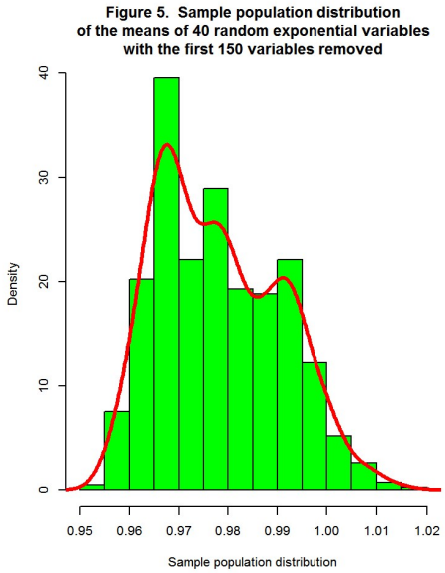
Figure 2. Exponential density function (EDF)

```
par(mfrow = c(1,3))

##3 - plotting the graph of Law of Large Numbers
```

```
plot(samplemeans, xlab = "Number of simulations (1000)", ylab = "Mean value of 40 random exponentials", pch = 1
9, main = "Figure 3.  Law of Large Numbers for\na sample of 40 random exponential variables\ngenerated in 1000
simulations")

##3.1 - plotting a line to represent the mean value of the exponential density function
abline(h = mean(exponentials))

    ##4 - plotting a histogram of mean value of 40 exponentials
    hist(samplemeans, xlab = "Sample population distribution", col = "green", main = "Figure 4.  Sample populat
ion distribution\nof the means of 40 random exponential variables", prob = TRUE)

    ##4.1 - plotting the sample population curve
    lines(density(samplemeans), col = "red", lwd = 3)

        ##5 - re-plotting of means without first 150 samples
        hist(removed, xlab = "Sample population distribution", col = "green", main = "Figure 5.  Sample populat
ion distribution\nof the means of 40 random exponential variables\nwith the first 150 variables removed", prob
= TRUE)

    ##4.1 - plotting the sample population curve
    lines(density(removed), col = "red", lwd = 3)
```







CONCLUSION

This analysis addresses three questions relating to an exponential density function and a sample population of 40 exponentials values.

Question 1 - How does the theoretical mean and sample mean compare?

Table 1 demonstrates that the mean value for the theoretical or exponential density function and a sample population of 40 exponentials is 0.975 and 0.994 respectively. The mean values are both approximately one which is what we would expect as our sample population is meant to accurately represent a normal population.

Question 2 - How does the theoretical variance and sample variance compare?

Referring back to table 1 the variances of the theoretical or exponential density function and sample population of 40 exponentials is 0.837 and 0.054 respectively. The variance is remarkably different the sample variance is closer to the mean value.

Question 3 - How does the theoretical density function and sample distribution curves differ?

Going back to the Central Limit Theorem where “the sampling distribution of any statistic will be NORMAL or NEARLY NORMAL if the sample size is large enough”. According to some statistians 30 to 40 samples is enough to satisfy the Central Limit Theorem is this the case?

For this question we need to refer to figures 1, 2, 3 and 4. Figures 1 and 2 show how theoretical density functions of a normal population and an exponential population differ. The exponential distribution is skewed to the left side of its graph but the population variance is with three standard deviations.

Comparing the theoretical (exponential) density function (figure 2) and sample means distribution (figure 4) shows a peak density of 0.8 and 8 respectively. If figure 4 is to accurately represent the EDF it should have all its values centred around the mean value with a smaller variance and it does.

If we remove the first 150 widely fluctuating samples shown in figure 3 and replot the samples (figure 5) we see that the distribution curve is approximately normally distributed (unlike figure 4) and that this sample mean (0.9781) almost matches the exponential density function mean of 0.9750. The variance becomes even tighter from 0.0054 to 0.0002. I can now confidentially say that 40 samples is enough to satisfy the Central Limit Theorem.

APPENDICES

R packages required for this analysis.

```
##1 - loading R packages
library(MASS)
library(stringr)
library(knitr)
library(rmarkdown)

    ##NOTE - for knitr/rmarkdown to work in RCONSOLE you are required to download the PANDOC package available online at: http://pandoc.org/installing.html
```

What hardware/software combination I am using for this analysis?

```
    ##2 - hardware/software?
    sessionInfo()
```

```
## R version 3.2.0 (2015-04-16)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 8 x64 (build 9200)
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.1252
## [2] LC_CTYPE=English_United Kingdom.1252
## [3] LC_MONETARY=English_United Kingdom.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] rmarkdown_0.6.1 knitr_1.10.5    stringr_1.0.0   MASS_7.3-40
##
## loaded via a namespace (and not attached):
## [1] magrittr_1.5    formatR_1.2     htmltools_0.2.6 tools_3.2.0
## [5] stringi_0.4-1   digest_0.6.8    evaluate_0.7
```

R SCRIPT - generating variables

creating a normal density function

```
set.seed(67)
population <- rnorm(1000)
```

creating an exponential density function

```
set.seed(53)
exponentials <- rexp(1000)
```

40 exponential samples

```
set.seed(43)
sample <- rexp(40)
```

generating 1000 means from the exponential samples

```
set.seed(63)
n <- 1000
samplemeans <- cumsum(sample(exponentials, n, replace = TRUE))/(1:n)
```

Removing the first 150 samples

```
removed <- samplemeans[-c(1:150)]
```

REFERENCES

Gourdon X (1999) Plouffe's Inverter: e to 1.250 billion digits, accessed at 17:35 on 9 June 2015 from http://www.plouffe.fr/simon/constants/expof1.txt

StatTrek.com (2015) Central Theorem Limit accessed at 11:30 on 9 June 2015 from http://stattrek.com/statistics/dictionary.aspx?definition=central_limit_theorem

Wikipedia (2015a) Euler's number accessed at 17:20 on 9 June 2015 from http://en.wikipedia.org/wiki/E_%28mathematical_constant%29

Wikipedia (2015b) Leonhard Euler accessed at 17:20 on 9 June 2015 from http://en.wikipedia.org/wiki/Leonhard_Euler

Wikipedia (2015c) John Napier accessed at 17:20 on 9 June 2015 from http://en.wikipedia.org/wiki/John_Napier

generating an html document

render(“project1.Rmd”, html_document())