Central Limit Theorem Applied to an Exponential Distribution

Stefan Schmager

January 31, 2016

This is the first project for the course Statistical Inference hosted by the Johns Hopkins University on Coursera as part of the Data Science Specialization.

1. Introduction

In this project the **Central Limit Theorem** (CLT), one of the most important theorems in statistics, will be investigated and applied to an **exponential distribution**.

The CLT states that **distributions of averages** of independent and identically distributed (iid) variables from **any distribution (such as the exponential)** becomes that of a **normal distribution** as the sample size increases.

The below mentioned packages are loaded to help manipulate and visualize sample data.

```
library(dplyr)    # Data manipulation

## Warning: package 'dplyr' was built under R version 3.2.5

library(ggplot2)  # Data visualization
```

2. Exponential Distribution

The exponential distribution serves as a CLT application example here although any other distribution (e.g. binomial, uniform, or normal distribution itself) would render similar observations.

```
# Set rate parameter
lambda      <- .2

# Compute theoretical (population) measures
mean        <- 1/lambda
stdev       <- 1/lambda
variance    <- stdev^2
```

Exponential distributions are characterized by *lambda, the rate parameter* which characterizes the distribution steepness and is set to $\lambda = 0.2$. The *theoretical (population) mean of an exponential distribution* μ is $1/\lambda = 5$ and the *theoretical (population) standard deviation* σ is also $1/\lambda = 5$. The *theoretical (population) variance* σ^2 is $1/\lambda^2 = 25$.

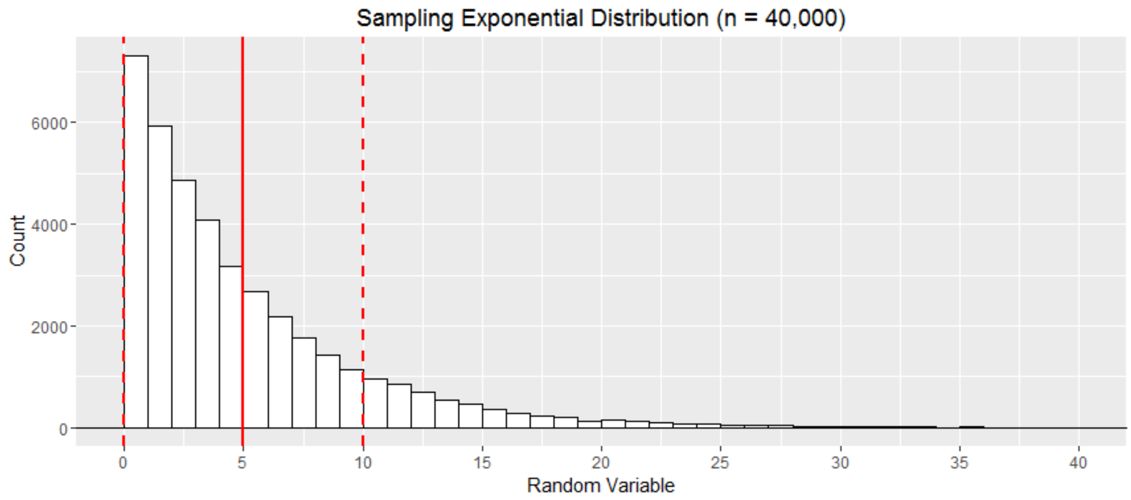
```
# Generate distribution sample
samplesize        <- 40
simreps           <- 1000
set.seed(1)
```

```
exp_distr <- data.frame(rvar = rexp(n = samplesize * simreps, rate = lambda))

# Compute sample measures for exponential distribution
samplemean        <- round(mean(exp_distr$rvar), 2)
samplestdev       <- round(sd(  exp_distr$rvar), 2)
samplevariance    <- round(samplestdev^2, 2)
```

Further, a sample of random variables $X_1, \dots, X_n$ is drawn from the exponential distribution with $\lambda = 0.2$ and a size of $n = 4 \times 10^4$. The sample size is based on the project's simulation requirements: a sample distribution ($n = 1000$) of sample means computed from an exponential distribution ($n = 40$) is going to be simulated.

But before the simulation is further elaborated, let's take a look at the sample exponential distribution

```
# Visualize sample distribution
ggplot(exp_distr, aes(rvar)) +
      labs(title = "Sampling Exponential Distribution (n = 40,000)",
           y     = "Count",
           x     = "Random Variable") +
      geom_histogram(binwidth=1, colour="black", fill="white") +
      # Draw sample measures in red
      geom_vline(aes(xintercept=samplemean),
                 linetype="solid", size=1, colour="red") +
      geom_vline(aes(xintercept=samplemean+samplestdev),
                 linetype="dashed", size=1, colour="red") +
      geom_vline(aes(xintercept=samplemean-samplestdev),
                 linetype="dashed", size=1, colour="red") +
      scale_x_continuous(breaks = seq(0, 40, 5)) +
      coord_cartesian(xlim=c(0, 40))
```



The histogram of the distribution shows a typical exponential shape with exponentially decreasing bins. As mentioned before, the steepness is determined by the rate parameter λ; the higher the parameter value, the steeper and vice versa. The *sample mean* $\bar{x}_n = 4.99$

(red, solid line) and *sample standard deviation* $s_n = 5$ (two red, dashed lines left and right from the mean) are drawn. The *sample variance* is $s_n^2 = 25$. All **sample measure are approximately equal to the theoretical measures**.

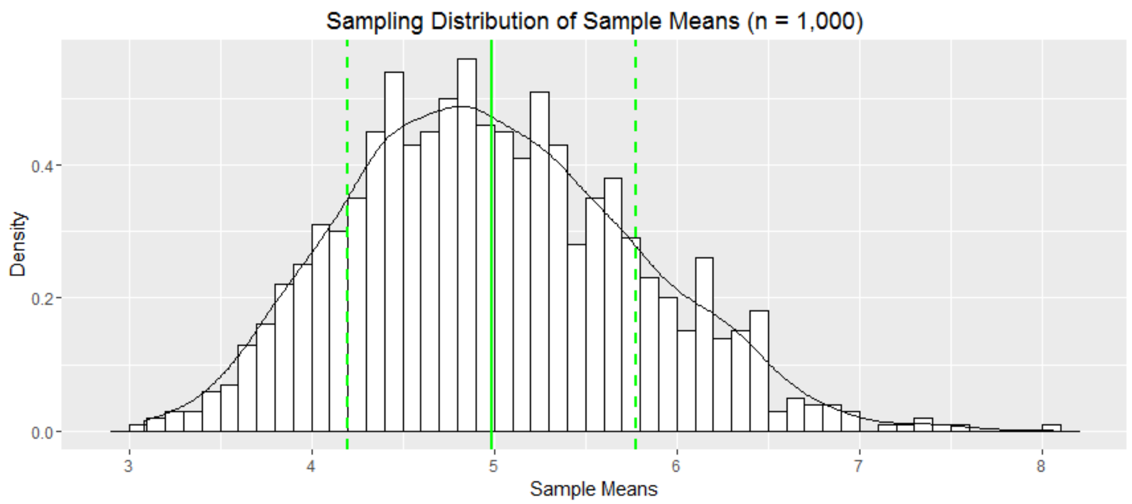
3. Simulation of the Central Limit Theorem

To better understand this simulation, its components are organized in a 40 by 1000 dataset. The simulated sampling distribution of sample averages with 1000 observations and the equal amount of row means of 40 random variables (columns), iid and ramdoly drawn from an exponential distribution, is organized in this data set. The first three rows of the dataset are provided with the first and last three iid random exponential variables in columns V1, V2, V3 and V38, V39, V40 respectively.

```
# Create simulation dataset
simulation <- NULL
simulation <- as.data.frame(matrix(exp_distr$rvar,
                                   nrow    = simreps,
                                   ncol    = samplesize))
simulation <- mutate(simulation, averages = rowMeans(simulation))

# Compute sample measure for distribution of averages means
samplemean2       <- round(mean(simulation$averages), 2)
samplestdev2      <- round(sd(  simulation$averages), 2)
samplevariance2   <- round(var( simulation$averages), 2)

# Display first and last three random variables from exp. distr. and row
mean/average
head(simulation[, c(1:3, 38:41)], 3)

##          V1        V2       V3        V38       V39       V40 averages
## 1 3.7759092 5.0325730  3.18463  0.7586185 0.7747079  2.135148 4.901268
## 2 5.9082139 0.2737037 15.25596 10.3120474 5.2523388  1.945569 5.229248
## 3 0.7285336 4.3983573 17.66216  2.2872371 6.1655369 10.267683 6.401541
```

The subject of CLT is the **sampling distribution of sample averages**, therefore the distribution of the 1000 row means in the column averages. It is visualized below.

```
# Distribution of sample means
ggplot(simulation, aes(averages)) +
      labs(title = "Sampling Distribution of Sample Means (n = 1,000)",
           y = "Density",
           x = "Sample Means") +
      # Histogram with density instead of count on y-axis
      geom_histogram(aes(y=..density..),
                   binwidth=.1,
                   colour="black", fill="white") +
      # Overlay with transparent density plot
      geom_density() +
      geom_vline(aes(xintercept = samplemean2),
                 linetype="solid",  size=1, colour="green")   +
      geom_vline(aes(xintercept = samplemean2 + samplestdev2),
                 linetype="dashed", size=1, colour="green")   +
      geom_vline(aes(xintercept = samplemean2 - samplestdev2),
                 linetype="dashed", size=1, colour="green")
```



The histogram and overlying density function has the shape of a normal distribution, exactly as the CLT states. The proof of that is yet to be delivered. The distribution's *sample mean* $\mu_{\bar{x}} = 4.99$ (green, solid line) and *sample standard deviation* $\sigma_{\bar{x}} = 0.79$ (two green, dashed lines left and right from the mean) are drawn. The *sample variance* is $\sigma_{\bar{x}}^2 = 0.62$. They will now be compared to their theoretical (population) counterparts.

4. Comparison of Theoretical and Sample Measures of the Distributions

Measure	Theoretical Distr.	Exp. Distr. ($n = 4 \times 10^4$)
Mean	$\mu = 5$	$\bar{x}_n = 4.99$
Std. Dev.	$\sigma = 5$	$s_n = 5$
Var.	$\sigma^2 = 25$	$s_n^2 = 25$

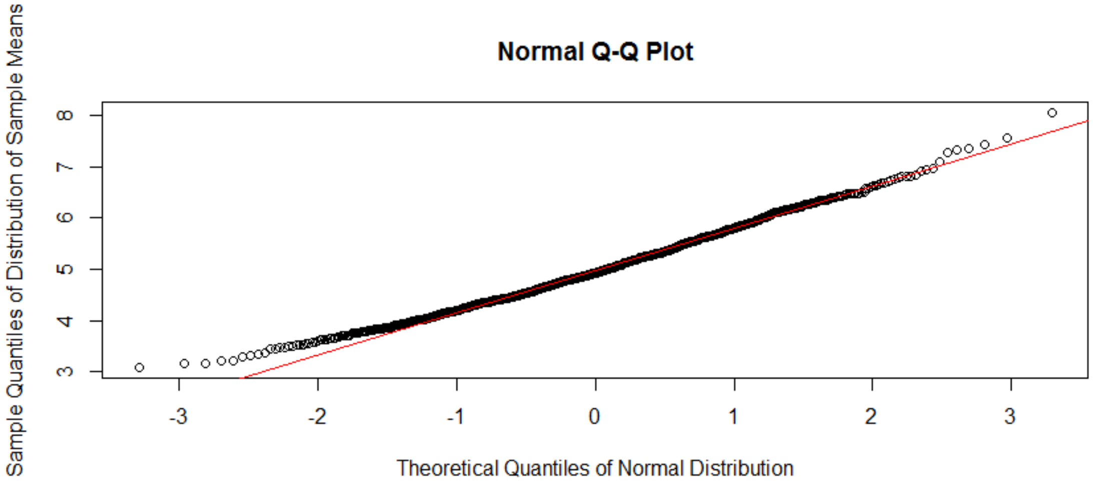
Measure	Theoretical Distr.	Distr. of Sample Means ($n = 1000$)
Mean	$\mu = 5$	$\mu_{\bar{x}} = 4.99$
Std. Dev.	$\sigma = 5$	$\sigma_{\bar{x}} = 0.79$
Var.	$\sigma^2 = 25$	$\sigma_{\bar{x}}^2 = 0.62$

The sample mean $\mu_{\bar{x}}$ of the distribution of sample means $\bar{x}$ is asymptotically equal to that of the theoretical distribution μ; it estimates the population mean. The standard deviation of the sampling distribution of sample means $\sigma_{\bar{x}}$ is also referred to as the **standard error of the mean**.

5. Test for Normality of the Distribution of Sample Means

To reiterate the CLT: a distribution of averages of iid variables (such as exponentials) has a distribution like that of a standard normal for large sample sizes (such as n = 1000). The graph titled "Sampling Distribution of Sample Means" indeed takes the shape of a normal-distribution typical "Gaussian" bell curve centered around the theoretical/sample mean (green, solid line). To test whether that visual impression is in fact true, a Qantile-Quantile (QQ) scatter plot is commonly used in which two sets of quantiles are plotted against one another. If both sets of quantiles came from the same distribution (normal), we should see the points forming a line (an identity line in normalized cases) that's roughly straight.

```r
# Draw a Q-Q plot
qqnorm(simulation$averages,
       xlab = "Theoretical Quantiles of Normal Distribution",
       ylab = "Sample Quantiles of Distribution of Sample Means")
# Draw straight test line
qqline(simulation$averages, col = "red")
```



The quantiles of the sampling distribution of sample means are indeed aligned with the theoretical quantiles along the red line. **The Central Limit Theorem is proven: the sampling distribution of a large collection of sample averages of any distribution is that of a normal distribution.** Although the distribution *where* the sample averages are drawn from *may not be normal* (here: exponential), the distribution *of* sample averages is always *normal*.

6. Conclusion

The CLT was applied to an exponential distribution, thereby explained and proven. The theorem's power and significance is due its application to any other underlying distribution and implications for other statistical concepts. It is indeed one of the most important theorems in statistics because of its implications for confidence intervals and hypothesis testing.

For example: the CLT is very important in poll research for instance where it may be applied to the Bernoulli distribution, e.g. voters vote either for (1) or against (0) a candidate. With the help of the CLT margins of error and certain confidence intervals around an average of "pro" or "contra" voters can be provided.