Statistical Inference Simulation Exercise

Gill Mundin

13 August 2019

1 - Overview

This project investigates the exponential distribution in R and compares it with the Central Limit Theorem (CLT). An exponential distribution is non-normal, and the CLT states that the *distribution of averages* of independent, identically distributed (iid) variables becomes that of a standard normal as the sample size increases. This is demonstrated using simulations and discussions.

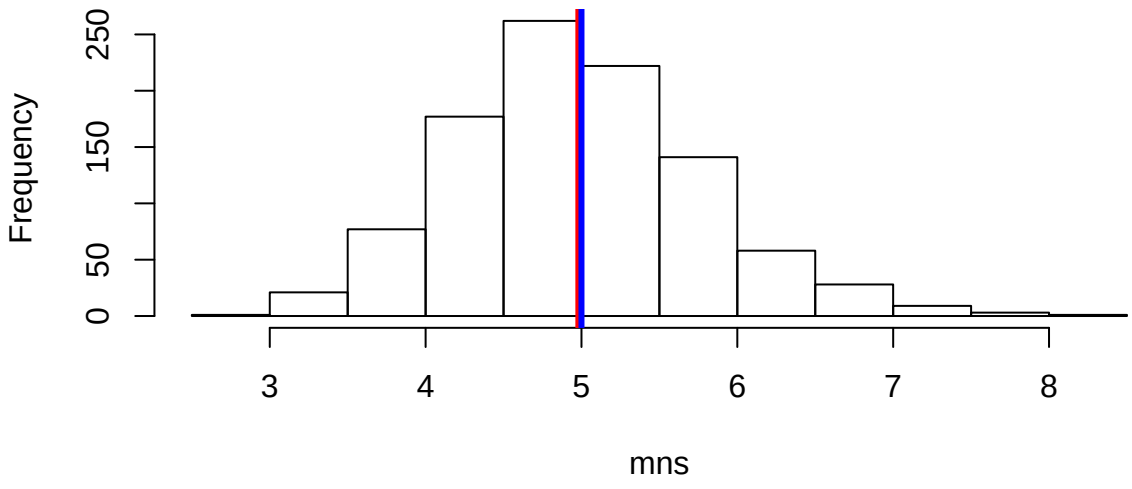
2 - Parameters

The exponential distribution has a rate parameter, lambda. The mean of an exponential distribution is 1/lambda, as is the standard deviation. For the simulations, use the parameters: lambda = 0.2, mean = 1/lambda = 5, stdev = 1/lambda = 5

3 - Simulation - distribution of 1000 averages of 40 exponentials

Define a seed to enable reproducible simulations. Simulate 1000 sets of 40 exponentials. Determine the mean for each set and plot of histogram of the 1000 means of 40 exponentials with lambda = 0.2.

Fig 1 − Hist of 1000 averages of 40 exponentials with lambda = 0.2




```
##        Estimate
## Mean  4.9865083
## Stdev 0.7965177
```

3.1 Explanation

The R code in the appendix simulates 1000 groups of 40 random exponential values with lambda = 0.2 as before. Means of each of the 1000 sets of 40 values are calculated and stored in the variable mns. The mean and standard deviation of mns were calculated, and were 4.99 and 0.80 respectively. The histogram of mns close to a normal distribution and appears symmetrical around the mean. The observed mean (red) and the expected mean (blue) are plotted on the histogram.

4 - Sample Mean versus Theoretical Mean

The theoretical mean of an exponential distribution is equal to 1/lambda, which in this example = 5 (plotted on the histogram with a blue line). Distributions of means of iid variables will be centered around the same spot as the original distribution, ie, the sample mean will be good estimate of the population mean. This is shown by the red line plotted on the histogram being almost exactly in the same place as the population mean in blue. The sample mean is a better estimate of the expected population mean than a single mean of 1000 observations from the population.

5 - Sample Variance versus Theoretical Variance

The variance and the standard deviation are linked, the SD is the square root of the variance. In this example, the expected value of the population SD is 5 (1/lambda).

The summary statistics for the mns variable has a standard deviation of 0.80, which differs considerably from the population SD of 5. This is because the mns SD is the SD that describes how variable the sample means are, not how variable the population observations are.

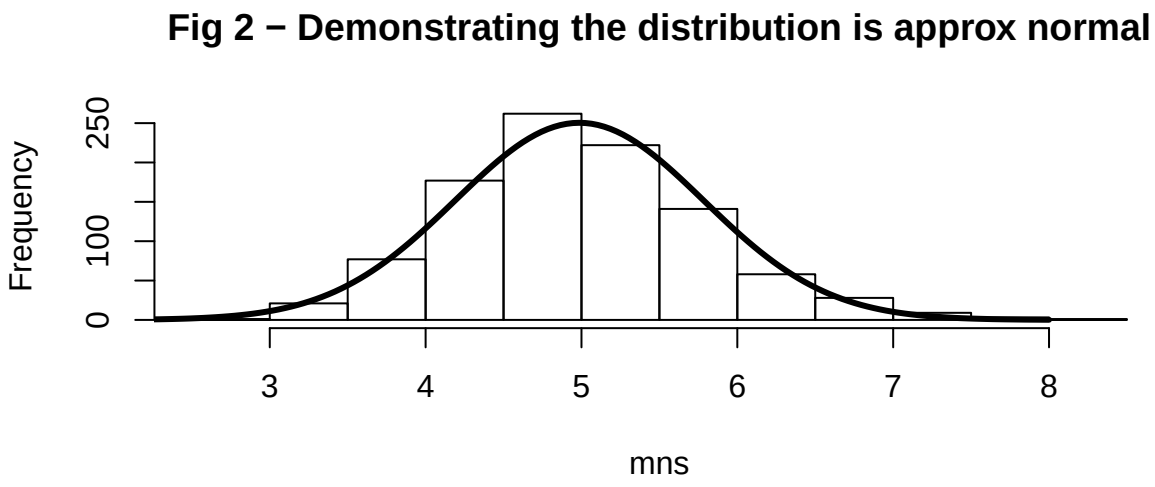
The variance of the sample mean is sigmaˆ2/n. The sample variance, Sˆ2, is an estimator of the population variance, so this can be re-written as Sˆ2/n. The standard error of the mean (the standard deviation, calculated as 0.8), is the square root of Sˆ2/n = S/sqrt(n). Therefore, 0.8 * sqrt(40) should give us a reasonable estimate of what the population standard deviation is:

```
## [1] 5.037621
```

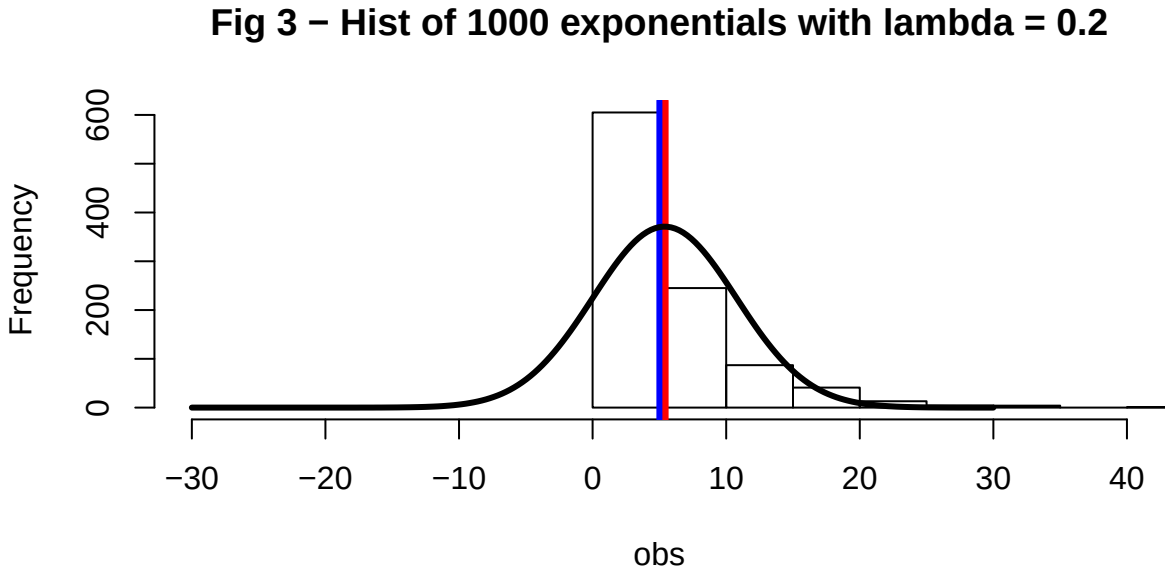
which is very close to the expected population SD of 5. This demonstrates that the variance of the sample mean, specifically the SD of 0.8, can be used to determine the population SD (5) and therefore the variance of the population.

6 - Distribution - approximately normal

The 1000 iid means of n = 40 had a mean = 4.99 and a SD = 0.80. To show that the distribution of iid means is approximately normal, overlay the histogram with a line representing a normal distribution with mean = 4.99 and SD = 0.8. Because generating this density curve gives an total area of 1, there is a mismatch between the y-axis of the density curve and the histogram. This is overcome by using a multiplier of 500 to increase the peak of the density curve from 0.5 to 250. The overlaid curve largely describes the histogram distribution and therefore the distribution is approximately normal.



Compare this with the distribution of 1000 random exponentials, and its associated mean and SD.




```
##       Estimate
## Mean  5.389714
## Stdev 5.376247
```

6.1 - Explanation

Figure 3 with an overlaid normal curve shows that a single sample of 1000 values from an exponential population does not have a normal distribution. However, Figure 2 shows that repeated samples of 40 values from the same population do follow a normal distribution. The distribution of sample means is centred around the original population mean and therefore a better estimate of the population mean, regardless of the original population distribution.

7 - Appendix - R code

Code for Section 3

```
set.seed(42) # ensures reproducible simulations
obs <- rexp(n = 1000, rate = 0.2) #simulates 1000 exponentials with lambda = 0.2
# and stores them in variable obs
hist(obs, # plots obs as a histogram
     main = "Fig 1 - Hist of 1000 exponentials with lambda = 0.2") #chart title
```

```
stats <- as.matrix(c(mean(obs), sd(obs))) #stores mean and sd of obs as a matrix
rownames(stats) <- c("Mean", "Stdev") #assigns row and column names to matrix
colnames(stats) <- "Estimate"
stats #prints stats as output
```

Code for Section 5

```
sd(mns) * sqrt(40) #calculation for estimating population SD
```

Code for Section 6

```
hist(mns,
     main = "Fig 2 - Demonstrating the distribution is approx normal")
x <- seq(0, 8, length = 200) # x vector defining the x axis values
y <- dnorm(x, mean = mean(mns), sd = sd(mns)) #y values for a normal curve with given mean and SD
lines(x, y*500, lwd = 3) #adds the normal density line to the main histogram
```

```
set.seed(42) # ensures reproducible simulations
obs <- rexp(n = 1000, rate = 0.2) #simulates 1000 exponentials with lambda = 0.2
# and stores them in variable obs
hist(obs, # plots obs as a histogram
     main = "Fig 3 - Hist of 1000 exponentials with lambda = 0.2",
     xlim = range(-30, 40)) #Sets the desired limits of the x axis
abline(v = mean(obs), #Adds a vertical line to the histogram, at the mean of obs
       col = "red",
       lwd = 4)
abline(v = 5,  #Adds another vertical line, at 5 on the x axis
       col = "blue",
       lwd = 3)
x2 <- seq(-30, 30, length = 200) #vector for x values for density curve
y2 <- dnorm(x2, mean = mean(obs), sd = sd(obs))#y values for normal curve
lines(x2, y2*5000, lwd = 3) #adds the normal curve to the main histogram
```

```
stats <- as.matrix(c(mean(obs), sd(obs))) #stores mean and sd of obs as a matrix
rownames(stats) <- c("Mean", "Stdev") #assigns row and column names to matrix
colnames(stats) <- "Estimate"
stats #prints stats to the pdf file
```