Statistical Inference Simulation Exercise

Alex Robinson

27 July 2017

Overview

This is an investigation into the exponential distribution and the applicability of the Central Limit Theory (CLT) for the Coursera Statistical Inference module of the Data Science track.

This document records a simulation of repeated sampling and averaging of data from the exponential distribution.

It then discusses how this sampled/averaged data compares to the exponential distribution in line with the expectations of the Central Limit Theory.

Simulations

```r
# set seed to aid reproducibility
set.seed(888)
# set the lambda value to be used throughout
lambda <- 0.2
# set the sample size
s.size <- 40
# set the repetitions
n <- 1000
# generate comparison data n random exp values
non.exp <- rexp(n, lambda)
# quick function to generate the mean data
gen.means <- function(n1, sz, lmb){
        output <- NULL
        # repeat the following n1 times
        for (i in 1 : n1){
                # add a new mean of sz values of rate lmb to the output data
                output <- c(output, mean(rexp(sz, lmb)))
        }
        # return the output
        output
}

avg.exp <- gen.means(n, s.size, lambda)
```

The simulation is performed using R's rexp function.

First an example set of 1000 non-averaged values is generated for the purposes of comparison.

Then a set of a 1000 values is generated by averaging across 40 random values generated with an exponential distribution.

In all cases a rate (lambda) value of 0.2 was used.

Sample Mean vs. Theoretical Mean

```
# Theoretical mean is given by 1/lambda
theoretical.mean <- 1/lambda
# mean of non-averaged example
non.avg.mean <- mean(non.exp)
# mean of averaged values
avg.mean <- mean(avg.exp)
# create a data.frame to display the values
disp.means <- data.frame(theoretical.mean, round(avg.mean, digits = 2),
                         row.names = "Mean")
names(disp.means) <- c("Theoretical", "Avg")

disp.means
```

```
##      Theoretical  Avg
## Mean           5 4.98
```

The theoretical mean of the exponential distribution is 1/lambda so in the case of a lambda of 0.2 the mean would be: 5

As predicted by the Central Limit Theory the observed mean for the averaged data approaches the theoretical value, in this instance it is: 4.98

The CLT states that with a higher number of simulations the sampled data would more closely tend towards the theoretical mean.

Sample Variance vs. Theoretical Variance

```
# Theoretical standard deviation is given by 1/lambda
theoretical.sd <- 1/lambda
theoretical.variance <- (1/lambda)^2
# sd/variance of non-averaged example
non.avg.sd <- sd(non.exp)
non.avg.var <- var(non.exp)
# sd/variance of averaged values
avg.sd <- sd(avg.exp)
avg.var <- var(avg.exp)
# create a data.frame to display the values.
disp.vars <- data.frame(c(theoretical.variance, theoretical.sd),
                        c(round(avg.var, digits = 2),
                          round(avg.sd, digits = 2)),
                         row.names = c("Variance", "SDev"))
names(disp.vars) <- c("Theoretical", "Avg")

disp.vars
```

```
##          Theoretical  Avg
## Variance          25 0.63
## SDev               5 0.80
```

The theoretical variance of the exponential distribution is the standard deviation (1/lambda) squared: 25

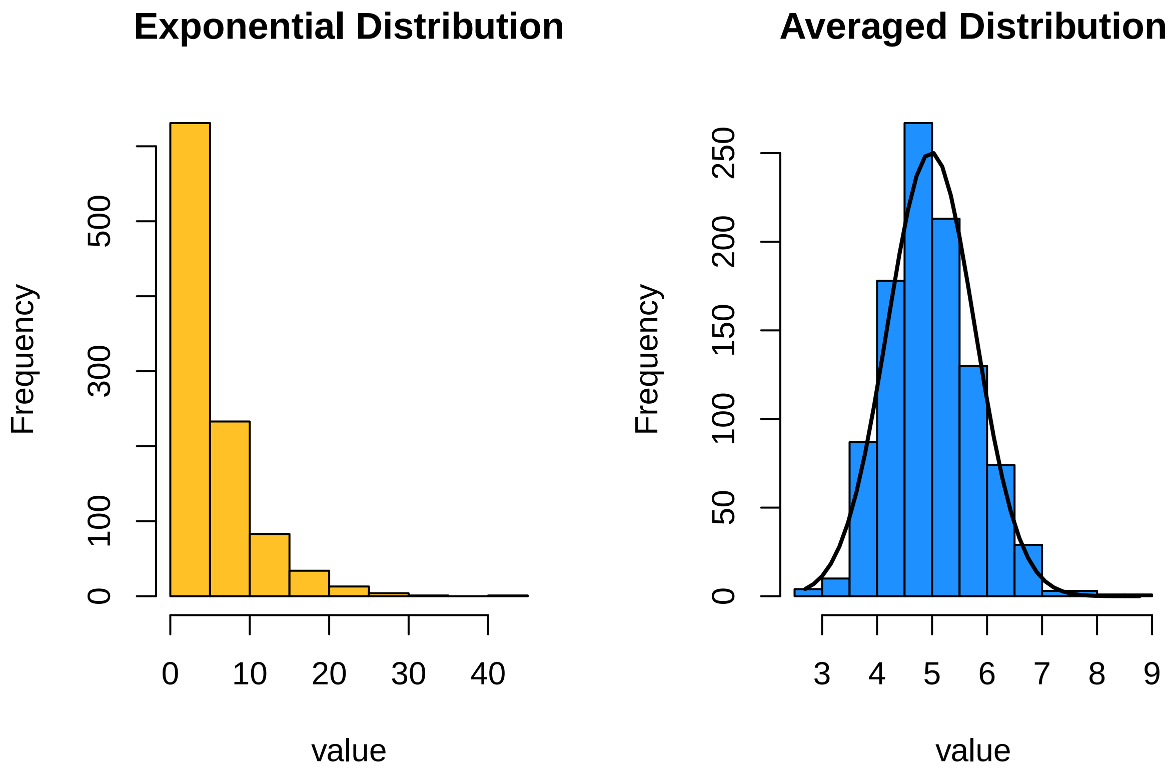
The observed variance of the averaged data is much lower: 0.63

This is consistent with data becoming less variable as the number of samples (n) increases. The variability is essentially removed from the data as a result of taking the mean across a number of observations.

Distribution

```r
# display two plots side by side
par(mfrow = c(1,2))
# Plot the standard exponential distribution
hist(non.exp, main = "Exponential Distribution", xlab = "value",
     col = "goldenrod1")
# Plot the averaged sample information
h <- hist(avg.exp, main = "Averaged Distribution", xlab = "value",
          col = "dodgerblue")
# draw a normal curve on the histogram to demonstrate the point
xfit <- seq(min(avg.exp), max(avg.exp), length.out = 40)
yfit <- dnorm(xfit, mean = avg.mean, sd = avg.sd)
yfit <- yfit * diff(h$mids[1:2]) * length(avg.exp)

lines(xfit, yfit, col = "black", lwd = 2)
```



As can be seen in the comparison plot (above) the averaged data tends much more towards a normal distribution than a random exponentially distributed sample.

This is the **Central Limit Theory** in action!