Simulation Exercise: Central Limit Theorem

Aslak Eriksen

6 4 2019

Overview

This report has investigated the distribution of means of 40 exponential distributions. The report concludes that the distribution behaves as predicted by the Central Limit Theorem - the distribution of means of 40 exponentials is approximately normal.

Structure of report

The following report will summarize the findings with both figures and text. Please note that the R code can be found in the appendix.

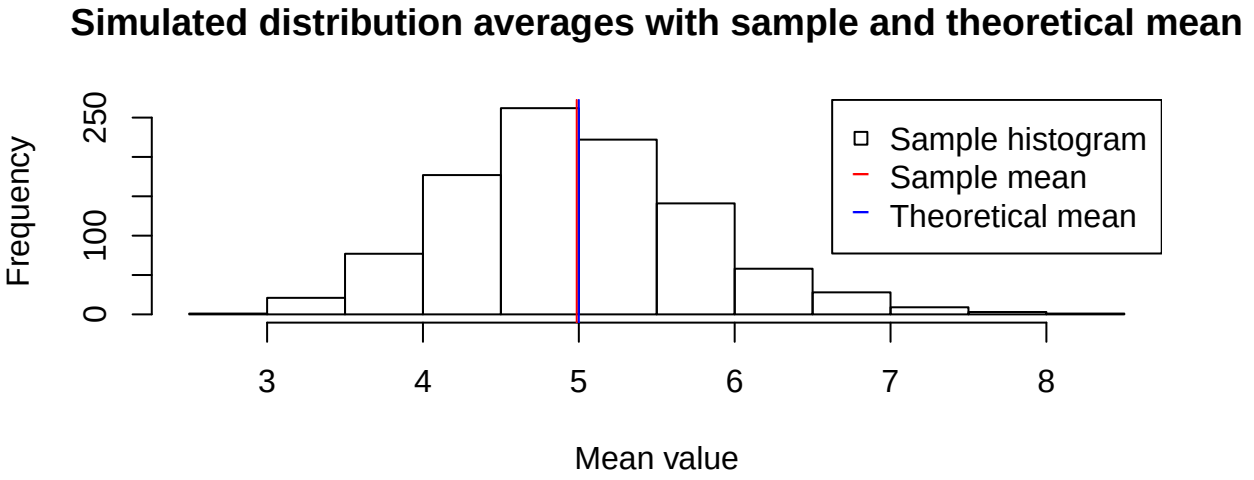
Simulations

The simulations are done for distribution of averages for the **exponential distribution**. Lambda - the rate parameter - is set to 0.2 for all simulations.

Sample Mean versus Theoretical Mean

The simulation data - distribution of averages of 40 exponentials - enable us to compare our sample mean with the theoretical mean.

Sample mean	Theoretical mean
4.986508	5



The table and figure above shows that the simulated mean is close to the theoretical mean. The histogram illustrates the higher frequency of simulated distribution averages around the theoretical mean (5).

Sample Variance versus Theoretical Variance

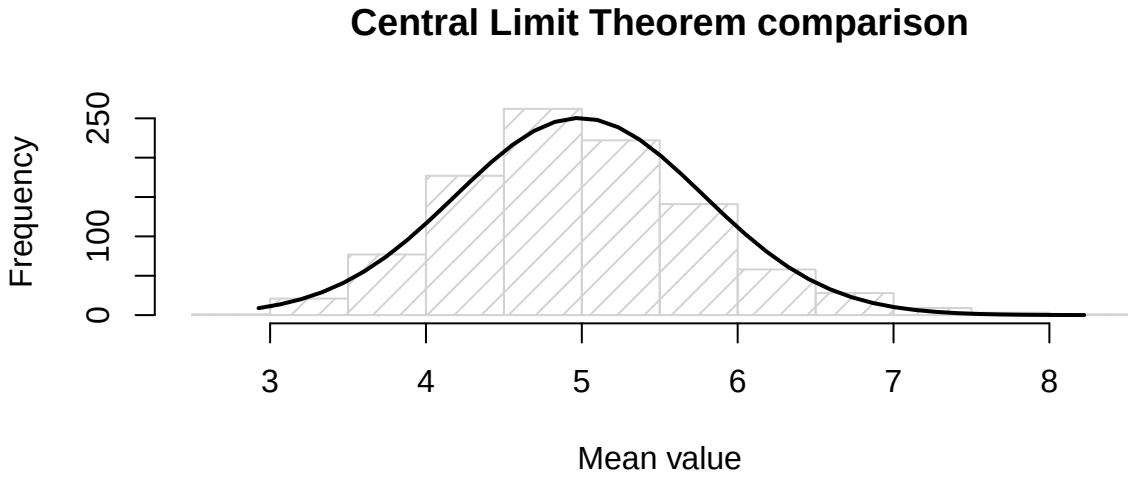
The simulation data - distribution of averages of 40 exponentials - enable us to compare our sample variance with the theoretical variance.

Sample variance	Theoretical variance
0.6344405	0.625

The table above shows that the simulated variance is close to the theoretical variance (standard error) for the distribution. This implies that the histogram from the section above is also similar to the theoretical distribution in regards to variance.

Central Limit Theorem for the simulated distributions

A normal distribution can be added to the histogram for the simulated distribution averages for 40 exponentials. The means and standard deviations are the same for the normal distribution and the simulated averages.



The plot heavily suggests that the simulated averages for 40 exponentials are approximately normal distributed. Hence, it supports the Central Limit Theorem which states that the distribution of averages of iid variables becomes that of a standard normal as the sample size increases.

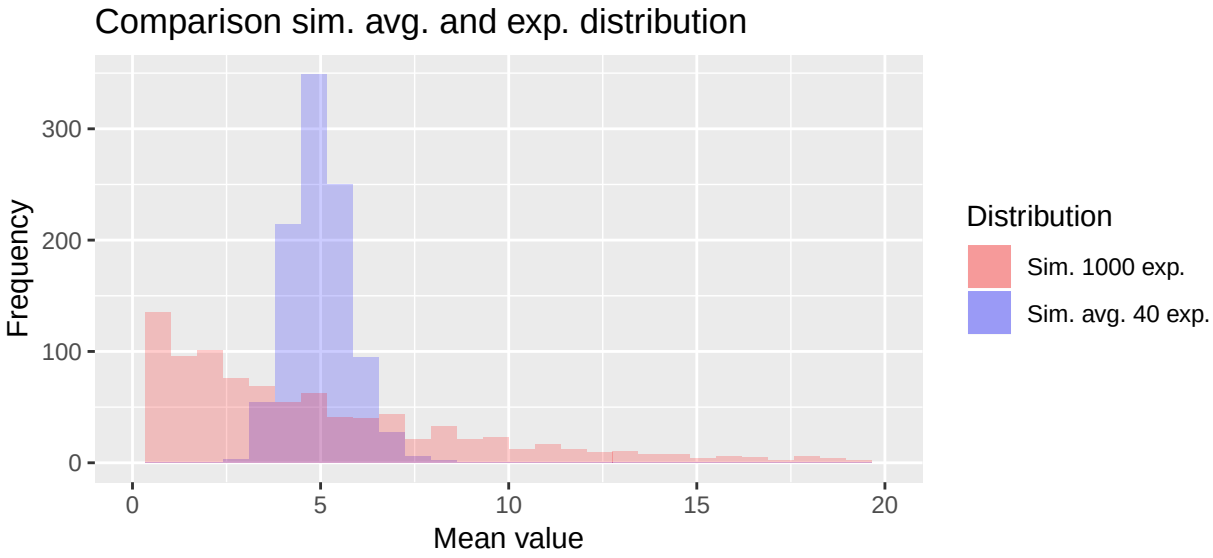
This can be compared to the simulated distribution of 1.000 exponentials. According to the Central Limit Theorem, this distribution should not be normal distributed.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 23 rows containing non-finite values (stat_bin).

## Warning: Removed 2 rows containing missing values (geom_bar).

## Warning: Removed 2 rows containing missing values (geom_bar).
```



The figure above compares the simulated distribution of averages for 40 exponentials with 1000 simulated exponentials. As predicted by the Central Limit Theorem - the 1000 simulated exponentials are does not become standard normal distributed as the sample size increases. The distribution of **averages**, on the other hand, is approximately normal distributed.

Appendix: R code

Simulations

```
## INTRO
# Install required packages if necessary
# install.packages("dplyr")
# install.packages("tidyverse")
# install.packages("reshape2")
# install.packages("tidyr")
# install.packages("kableExtra")

# Load packages
library(dplyr)
library(tidyverse)
library(reshape2)
library(tidyr)
library(kableExtra)

# Set up simulations, start with parameters lambda, mean, standard deviation,
# number of distributions, and number of simulations.
lambda <- 0.2
mean <- 1 / lambda
sd <- 1 / lambda
n <- 40
nsim <- 1000

## SIMULATION
# Set seed
set.seed(42)

# Create empty matrix for averages
mns = NULL

# Run simulation with given parameters
for (i in 1 : nsim) mns = c(mns, mean(rexp(n, lambda)))

# Distribution of 1000 random exponentials
exp_dist <- rexp(nsim, lambda)
```

Sample Mean versus Theoretical Mean

```
# The mean for the simulated data is calculated.
#The theoretical mean is given by the parameters that were initially introduced.
dist_mean <- mean(mns)
theoretical_mean <- mean

# Create a table that compares the sample and theoretical mean.
dt <- matrix(c(dist_mean, theoretical_mean), ncol = 2, byrow = TRUE)
colnames(dt) <- c("Sample mean", "Theoretical mean")
kable(dt)
```

```r
# Create a histogram for the simulation data
hist(
        mns,
        main = "Simulated distribution averages with sample and theoretical mean",
        xlab = "Mean value",
        ylab = "Frequency")

# Create vertical lines for the sample an theoretical means
abline(v = dist_mean,col = "red")
abline(v = theoretical_mean,col = "blue")

# Create a legend for figure.
legend("topright",
  legend = c("Sample histogram", "Sample mean", "Theoretical mean"),
  col = c("black", "red", "blue"),
  text.col = "black",
  pch = c(22, 45, 45))
```

Sample Variance versus Theoretical Variance

```r
# Calculate the variance for the simulation data.
dist_var <- var(mns)
# Standard error is the variance of the population being sampled from
# divided by the number of samples.
theoretical_var <- (sd^2)/n

# Create a table that compares the sample and theoretical variance.
dt_var <- matrix(c(dist_var, theoretical_var), ncol = 2, byrow = TRUE)
colnames(dt_var) <- c("Sample variance", "Theoretical variance")
kable(dt_var)
```

Overlay: normal distribution to the histogram for the simulated distribution averages

```r
## The code belows overlays a normal distribution to the histogram
#for the simulated distribution averages for 40 exponentials.
# Create a histogram for simulation data.
h <- hist(mns, breaks = 10, density = 10,
          col = "lightgray", xlab = "Mean value",
          main = "Central Limit Theorem comparison")

# Overlay a normal distribution over histogram with the same mean and standard deviation.
xfit <- seq(min(mns), max(mns), length = 40)
yfit <- dnorm(xfit, mean = dist_mean, sd = sqrt(dist_var))
yfit <- yfit * diff(h$mids[1:2]) * length(mns)
lines(xfit, yfit, col = "black", lwd = 2)
```

Plot comparison simulated averages with simulated distribution for exponentials

```
# Create data frame with all simulation data.
# Both the simulated averages and the simulated distributions for exponentials.
# Create identifiers for the plotting.
dat <- data.frame(xx = c(mns, exp_dist), yy = rep(letters[1:2], each = 1000))

# Create plot with both histograms shown for each simulation.
ggplot(dat,aes(x=xx)) +
    geom_histogram(
            data=subset(dat,yy == 'a'),
            aes(x = xx, fill = "Sim. avg. 40 exp."),
            alpha = 0.2) +
    geom_histogram(
            data=subset(dat,yy == 'b'),
            aes(x = xx,
                fill = "Sim. 1000 exp."),
            alpha = 0.2) +
        scale_x_continuous(
                    limits = c(0, 20)) +
        labs(
                x = "Mean value",
                y = "Frequency",
                title = "Comparison sim. avg. and exp. distribution") +
        scale_fill_manual(
                        name = "Distribution",
                        values = c("red", "blue"))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 23 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```