Sample vs Theoretical Exponential Distribution in R

Benedict Neo

16/12/2020

Overview

In this report, I will be investigating the exponential distribution in R and compare it with the Central Limit Theorem (CLT). I will be investigating the distribution of averages of 40 exponentials, and a total of a thousand simulations.

What is exponential distribution?

From Wikipedia: Exponential distribution describes times between events happening at constant rate lambda with expected value 1/lambda.

In summary, this report will 1. Show the sample mean and compare it to the theoretical mean of the distribution 2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution 3. Show that the distribution is approximately normal

Simulation

The exponential distribution is simulated with `rexp(n, lambda)`, where lambda is the rate parameter. The mean of exponential distribution and standard deviation is 1/lambda. Lambda is set at 0.2 for all simulations.

Sample exponential distribution

```r
set.seed(2021) # for reproducability
nosim <- 1000 # no of simulations
n <- 40 # no of exponentials
lambda <- 0.2 # rate parameter

simdata <- matrix(rexp(nosim * n, rate=lambda), nosim)
sim_mean <- rowMeans(simdata) # row means

# calculate mean, sd and variance of sample exp dist
simdata_mean <- mean(sim_mean)
simdata_sd <- sd(sim_mean)
simdata_var <- var(sim_mean)
```

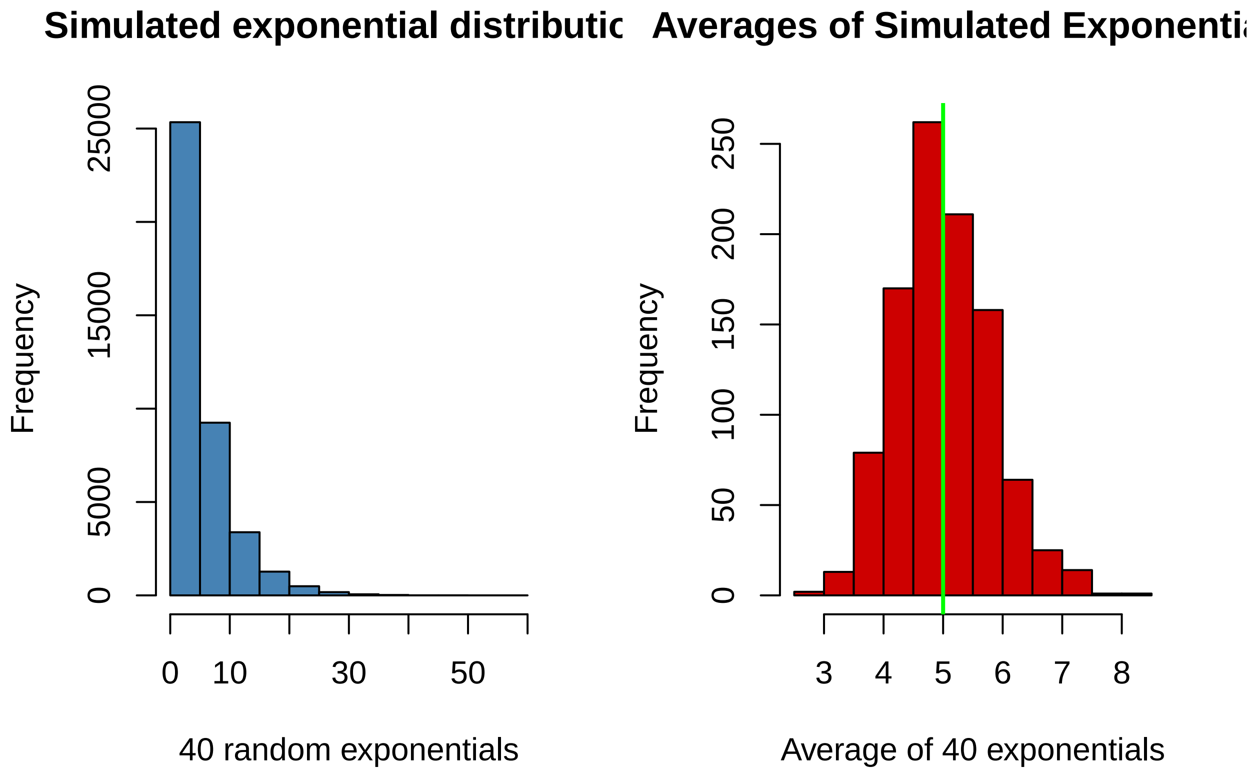
With the no of simulations, no of exponentials, and the rate parameter, we can simulate the exponential distribution by multiplying the exponential by the no of simulations, giving us 1000 simulations of 40 exponentials. We put it in matrix form, and use the apply function to find the mean for each row.

With this, we can then find the sample mean, standard deviation and variance.

Theoretical exponential distribution

```
# calculate mean, df and variance of theoretical exp dist
t_mean = 1/lambda
t_sd = (1/lambda) * (1/sqrt(n))
t_var = t_sd^2
```

Histogram of sample exponential distribution vs Averages of simulated exponentials.



The blue histogram represents the simulated exponential distribution, as you can see most of the data is at the left side of the plot because of the properties of the exponential distribution.

Observing the histogram for the averages of simulated exponentials, we can see it's following the form of a normal distribution. This will be further investigated later on in the report.

The green line represents the theoretical mean of the distribution, and our simulated distribution is centered closely at the line.

Comparison between sample and theoretical statistics

```
##          Sample_stats Theoretical_stats        diff
## Mean        5.0086386         5.0000000 0.008638602
## Std         0.7882570         0.7905694 0.002312423
## Variance    0.6213491         0.6250000 0.003650915
```

Sample Mean versus Theoretical Mean

Observing the table, the sample mean of the exponential distribution is centered at 5.008 whereas the theoretical mean, 1/lambda is 5

The difference between the sample and theoretical mean is 0.0086

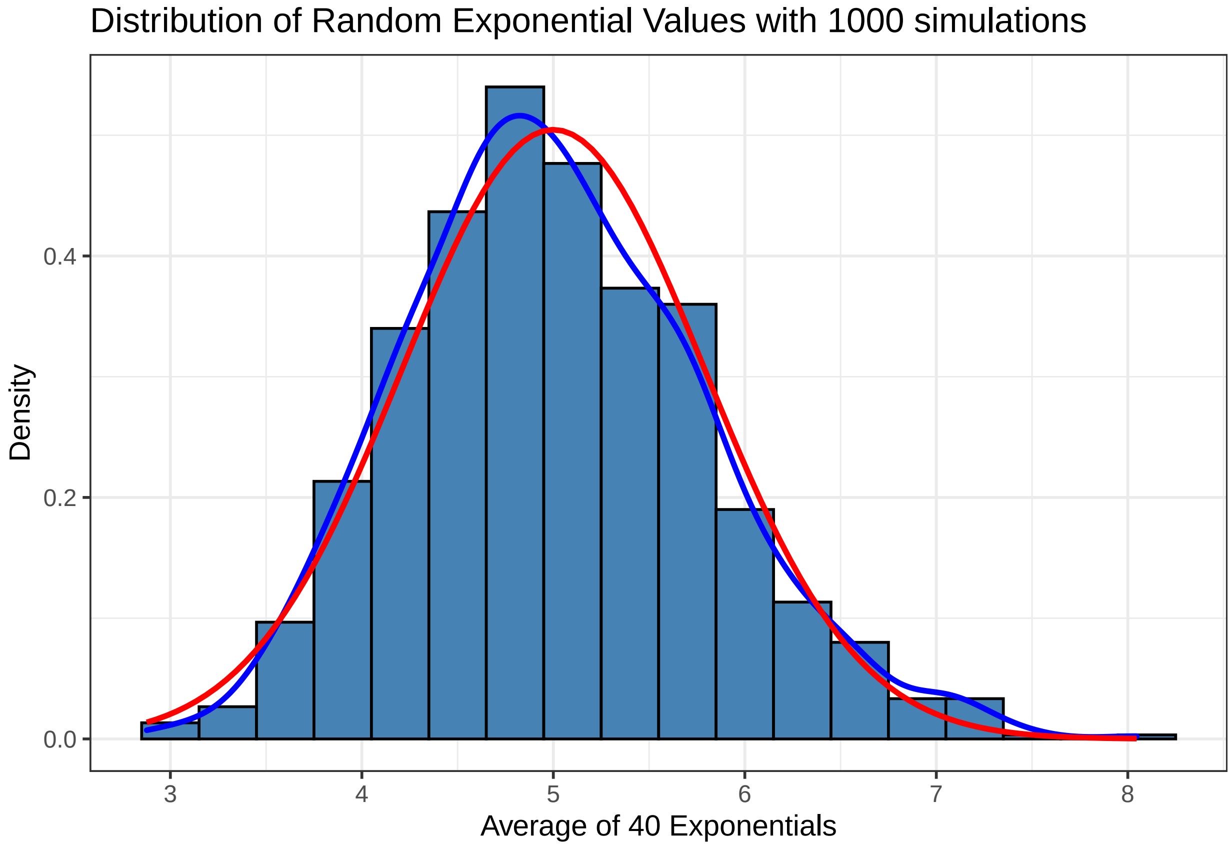
Sample Variance versus Theoretical Variance

The sample Variance is is 0.621, which is very close to the theoretical variance, 0.625.

The difference between them is 0.0037

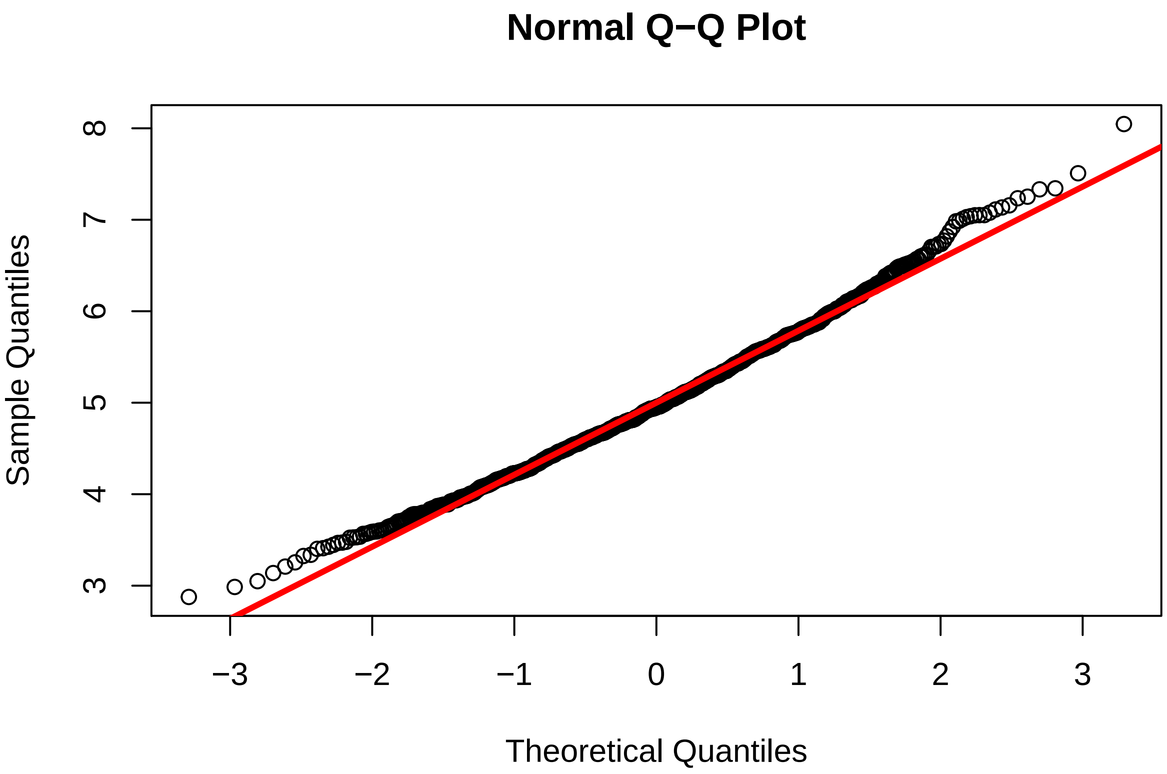
Distribution

Histogram and Density plot



The red line is the theoretical normal distribution density, whereas the blue line is the density of the sample distribution. You can see that the sample distribution is approximately normal.

Q-Q plot



Observing the normal Q-Q plot, we can conclude that the sample distribution approximates the theoretical normal distribution quite closely, with the tails being less normal.

Conclusion

Based on the comparisons and the plots, the simulated sample distribution (as n grows larger) does indeed have similar means and variance with the theoretical distribution. This proves the Central Limit Theorem is in fact true.

An important condition for the central limit theorem is that the random variables are IID, which stands for Independent and Identically Distributed. These conditions are satisfied as we simulated the data using R.

Appendix

R codes for table and plots

Plot 1

```
par(mfrow = c(1, 2))
hist(simdata,
     col = "steelblue",
     main = "Simulated exponential distribution",
     xlab = "40 random exponentials")
hist(sim_mean,
     col = "red3",
     main = "Averages of Simulated Exponentials",
     xlab = "Average of 40 exponentials")
abline(v = t_mean, col = "green", lwd = 2) # theoretical mean
```

Table

```
Sample_stats <- c(simdata_mean, simdata_sd, simdata_var)
Theoretical_stats <- c(t_mean, t_sd, t_var)
diff <-
  c(abs(t_mean - simdata_mean),
    abs(t_sd - simdata_sd),
    t_var - simdata_var)
names <- c("Mean", "Std", "Variance")
data.frame(Sample_stats,
           Theoretical_stats,
           diff,
           row.names =  c("Mean", "Std", "Variance"))
```

Plot 2

```
library(ggplot2)

simdata_mean <- data.frame(sim_mean)
ggplot(simdata_mean, aes(sim_mean)) +
    geom_histogram(
        binwidth = .3,
        fill = "steelblue",
        color = "black",
        aes(y = ..density..)
    ) +
    geom_density(color = "blue", lwd = 1) +
    labs(title = "Distribution of Random Exponential Values with 1000 simulations",
         x = "Average of 40 Exponentials", y = "Density") +
    stat_function(
        fun = dnorm,
        args = list(mean = t_mean, sd = t_sd),
        color = "red",
        lwd = 1
    ) +
    theme_bw()
```

Plot 3

```
qqnorm(sim_mean, col="black") # sample distribution
qqline(sim_mean, col="red", lwd=3) #theoretical
```

Software

```
sessionInfo()
```

```
## R version 4.0.2 (2020-06-22)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS  10.16
## 
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
## 
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## other attached packages:
## [1] ggplot2_3.3.2
## 
## loaded via a namespace (and not attached):
##  [1] knitr_1.30       magrittr_2.0.1   tidyselect_1.1.0 munsell_0.5.0   
##  [5] colorspace_2.0-0 R6_2.5.0         rlang_0.4.8      stringr_1.4.0   
##  [9] dplyr_1.0.2      tools_4.0.2      grid_4.0.2       gtable_0.3.0    
## [13] xfun_0.19        withr_2.3.0      htmltools_0.5.0  ellipsis_0.3.1  
## [17] yaml_2.2.1       digest_0.6.27    tibble_3.0.4     lifecycle_0.2.0 
## [21] crayon_1.3.4     farver_2.0.3     purrr_0.3.4      vctrs_0.3.5     
## [25] glue_1.4.2       evaluate_0.14    rmarkdown_2.5    labeling_0.4.2  
## [29] stringi_1.5.3    compiler_4.0.2   pillar_1.4.7     generics_0.1.0  
## [33] scales_1.1.1     pkgconfig_2.0.3
```