Analysis of the Exponential distribution using simulation

Chris Shaw

3 April 2016

Question

Remove this section in final report

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should

Show the sample mean and compare it to the theoretical mean of the distribution. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution. Show that the distribution is approximately normal. In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

Overview

In a few (2-3) sentences explain what is going to be reported on.

Simulations

The exponential distribution is the probability distribution that describes the time between events in a Poisson process, i.e. a process in which events occur continuously and independently at a constant average rate. This rate is λ. The mean of this distribution is $1/\lambda$ and the standard deviation is also $1/\lambda$.

In this paper we will explore how a sample of 40 random values from this distribution, denoted by X_n, behaves over a large number of simulations. The central limit theorem predicts that distribution of sample means generated by the simulation should converge to the following normal distribution:

$$(X_n)_i \sim N(\mu, \frac{\sigma^2}{n})$$

as the number of simulations $i \to \infty$. The μ and σ^2 are the mean and variance of the underlying population, and n is the sample size.

The simulation code to generate samples 1000 times is shown in the appendix. A matrix called **sim_data** is created with 40 columns and 1000 rows. The mean and standard deviation of each row is collected and the table below shows the average values of these across all the simulations:

Statistic	Value
Population mean	5.000
Mean of Sample means	4.972
Mean of Sample Std dev	4.850

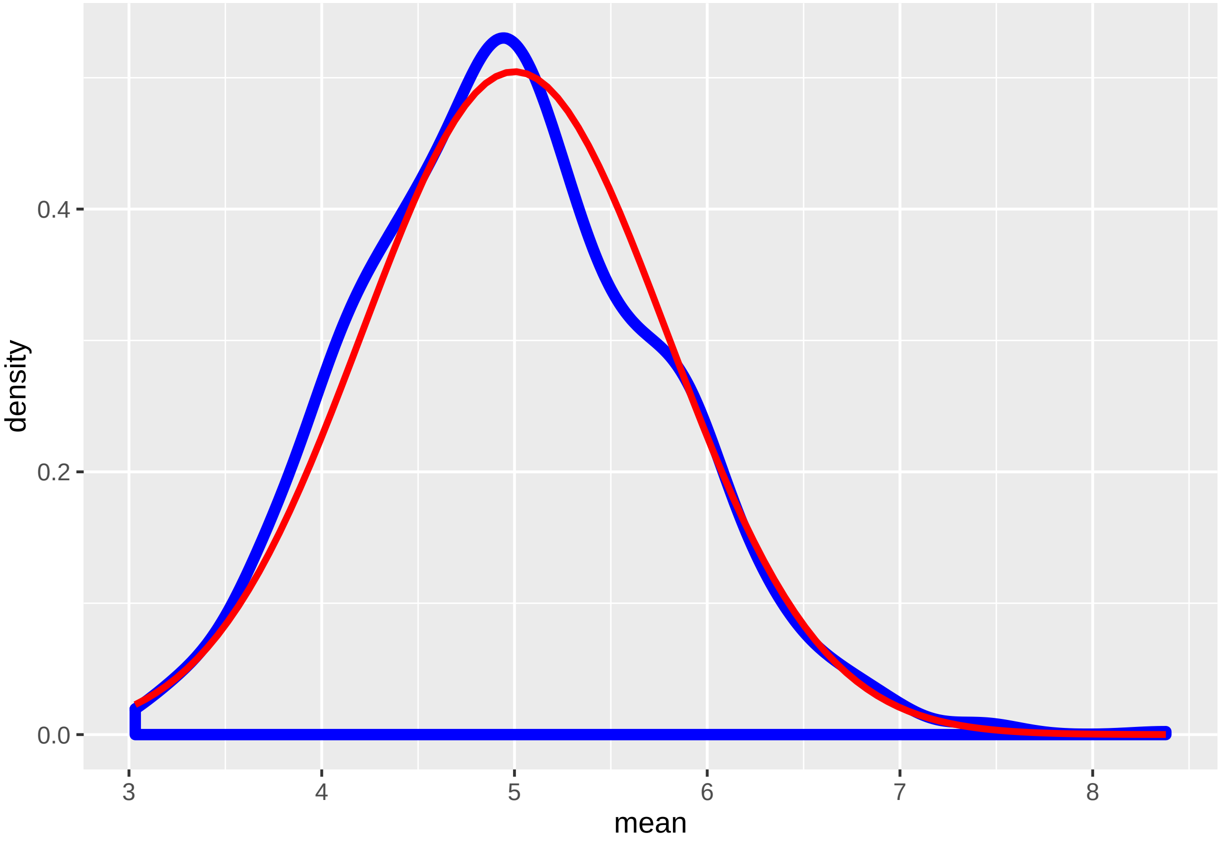
The population standard deviation is also equal to the mean for the exponential distribution, and it can be seen already that the simulated averages are close to the population values.

Distribution

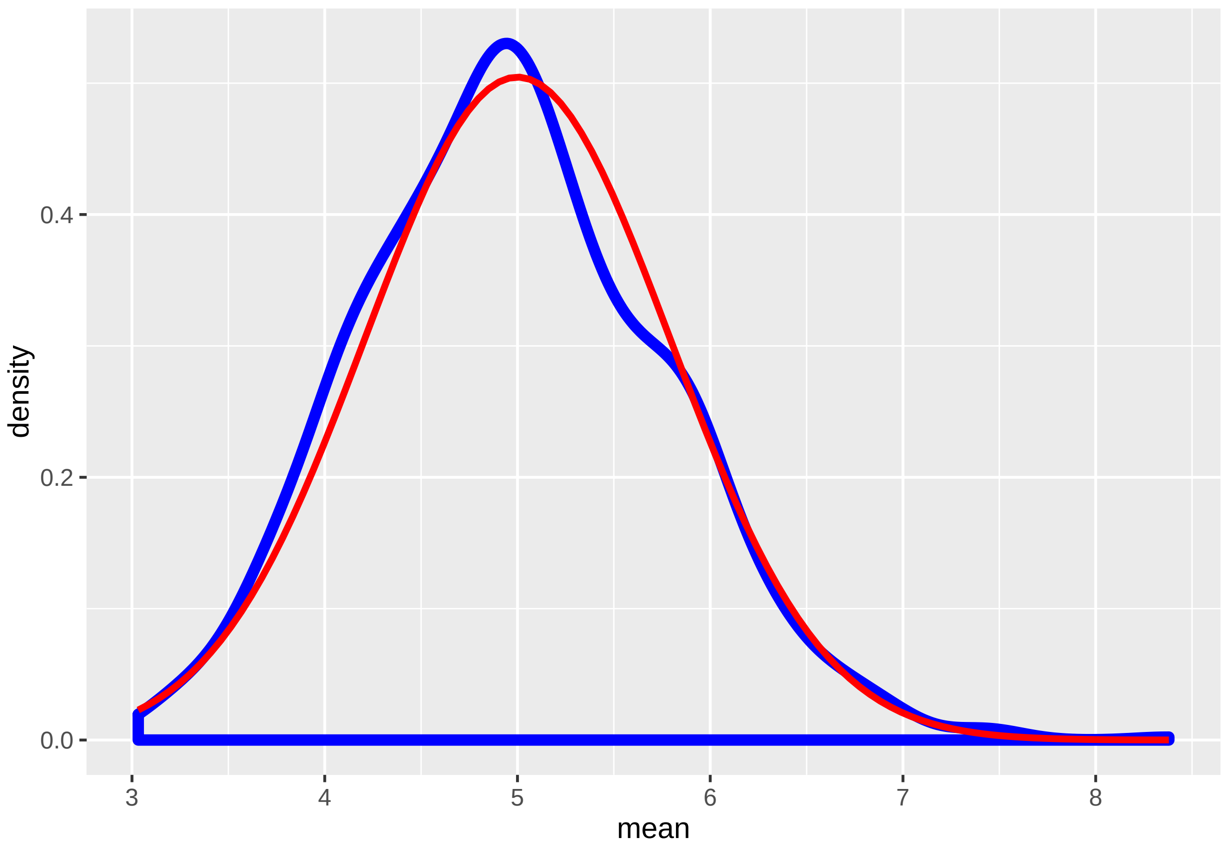
The next task is to compare the distribution density of the sample means to the theorietical distribution under the Central Limit Theorem of:

$$X_n \sim N(\frac{1}{\lambda}, \frac{1}{n\lambda^2})$$

The graph below shows the distribution of sample means in blue and the Normal distribution above in red.



It can be seen that after 1000 simulations, the distribution of the sample means itself forms a distribution which is very close to that predicted by the CLT.



Sample Mean versus Theoretical Mean

Include figures with titles. In the figures, highlight the means you are comparing. Include text that explains the figures and what is shown on them, and provides appropriate numbers.

Sample Variance versus Theoretical Variance

Include figures (output from R) with titles. Highlight the variances you are comparing. Include text that explains your understanding of the differences of the variances.

Appendix

This is the code used to generate the matrix of simulation data, and calculate the sample mean and standard deviation:

```r
# set up parameters of the exponential distribution
lambda <- 0.2
sample_size <- 40
num_sims <- 1000

mean <- 1/lambda
std <- 1/lambda

# Ensure simulations are reproducible
set.seed(12345)

# create a matrix of simulation data (rows = simulations)
sim_data <- matrix(rexp(sample_size*num_sims, lambda), nrow = num_sims)

sample_means <- as.data.frame(apply(sim_data, 1, mean))
names(sample_means)[1]="mean"

sample_sds <- as.data.frame(apply(sim_data, 1, sd))
names(sample_sds)[1]="sd"

sample_mean <- mean(sample_means$mean)
sample_std <- mean(sample_sds$sd)
```

The code below was used to generate the plot to compare the distribution of the sample means with the central limit theorem:

```r
# Plot the density of the simulation sample means against the normal distribution
# with mean of the 1/lambda and standard deviation of 1/lambda divided by
# square root of sample size
ggplot(data=sample_means,aes(x=mean)) +
        geom_density(col="blue", lwd=2) +
        stat_function(fun = dnorm,
                        args = list(mean=mean, sd=std/sqrt(sample_size)),
                        color="red", lwd=1.2, lty=1)
```