Exploring The Central Limit Theorem With Exponential Random Variables

Author: DRC

Part I

Overview

The following is a simulation to explore the Central Limit Theorem using exponential random variables. The mean of the exponential distribution is 1/ lambda, the standard deviation is also 1/ lambda, and lambda is set equal to 0.2 for 1000 simulations conducted. In each simulation, 40 random variables from the exponential distribution were generated for each experiment.

Simulations

A matrix of random variables from the exponential distribution was created where each row contained 40 random variables. There are 1000 rows representing the number of simulations.

```
lambda <- 0.2
num_of_sims <- 1000
trials_mat <- matrix(rexp(40 * num_of_sims, lambda), num_of_sims, 40)
```

Then I created a data frame with 3 columns. The first column contains the average cumulative means for the 1000 simulations. This means the first number in this column is the average mean of the first experiment (the first 40 random variables from the first row of the trials_mat matrix), the second number in this column is the average of the means from the first two experiments, the third number is the average of the first 3 experiments and so on. Similarly, the second column contains the average cumulative variances of all the simulations.

The third column is the means of each experiment, after being standardized to the standard normal distribution. To standardize the means a function called stnd_normalizer was created to subtract the theoretical mean and divide he difference by the standard normal error (which is the standard deviation divided by the square root of the number of variables in each experiment).

```
mns <- apply(trials_mat, 1, mean)
vrs <- apply(trials_mat, 1, var)

stnd_normalizer <- function(x) {
  mu <- 1 / lambda
  stddev <- 1 / lambda
  sqrt(40) * (x - mu) / stddev
}

mns_standardized <- lapply(list(mns), stnd_normalizer)

df <- data.frame(
  cumsum(mns)/(1:1000),
  cumsum(vrs)/(1:1000),
  mns_standardized
)
names(df) <- c("cum_means", "cum_variances", "stnd_means")
```

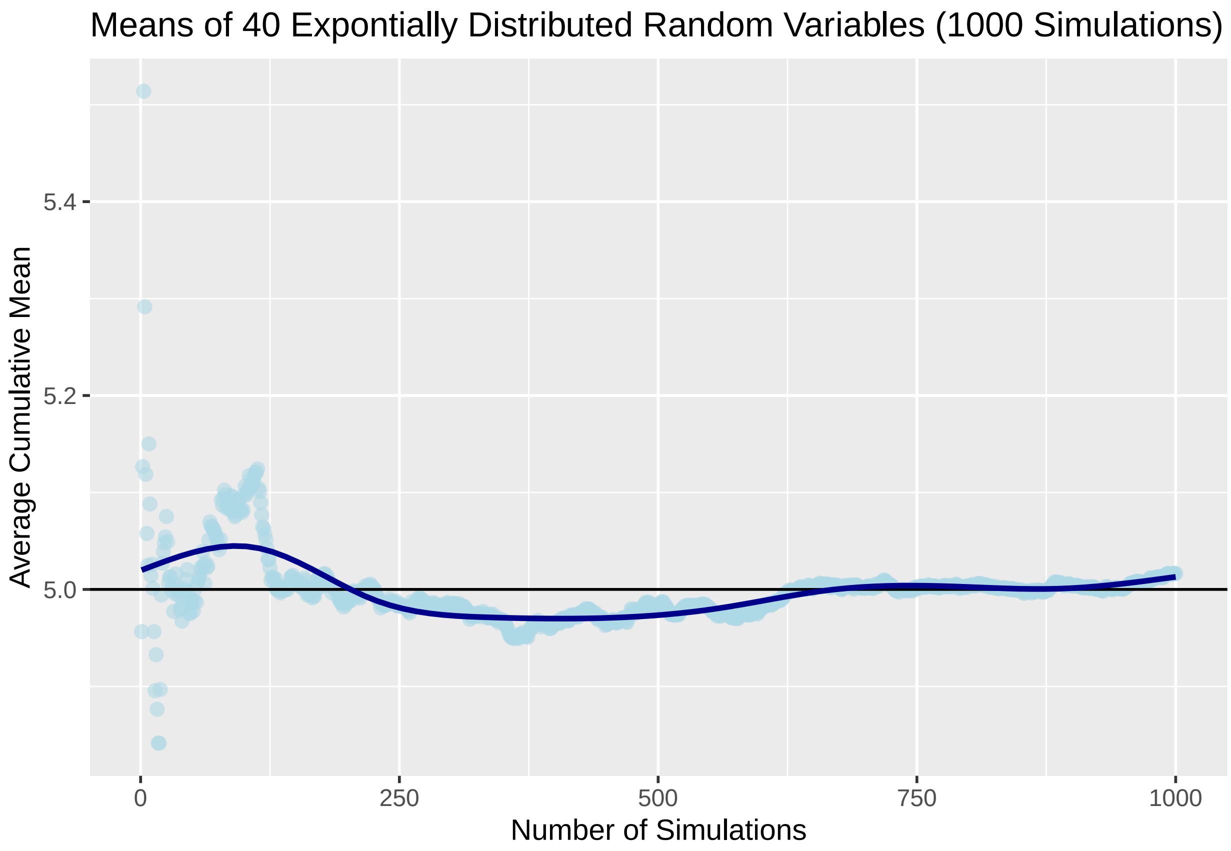
Sample Mean versus Theoretical Mean

The theoretical mean of the exponential function is 1/ lambda. For these experiments, the theoretical mean = 1/ 0.2 = 5.

Because of the Law of Large Numbers (LNN), we know that the average of the samples collected limit to the theoretical mean. To demonstrate this, a scatter plot was created with the number of simulations along the x-axis and the average cumulative means from those numbers of simulations along the y-axis.

```
library(ggplot2)

ggplot(df, aes(x=(1:1000), y=cum_means)) +
  geom_point(color = "lightblue", alpha = .58, size = 2) +
  geom_hline(yintercept=1/lambda) +
  geom_smooth(color = "darkblue", se=FALSE) +
  xlab("Number of Simulations") +
  ylab("Average Cumulative Mean") +
  ggtitle("Means of 40 Expontially Distributed Random Variables (1000 Simulations)")
```



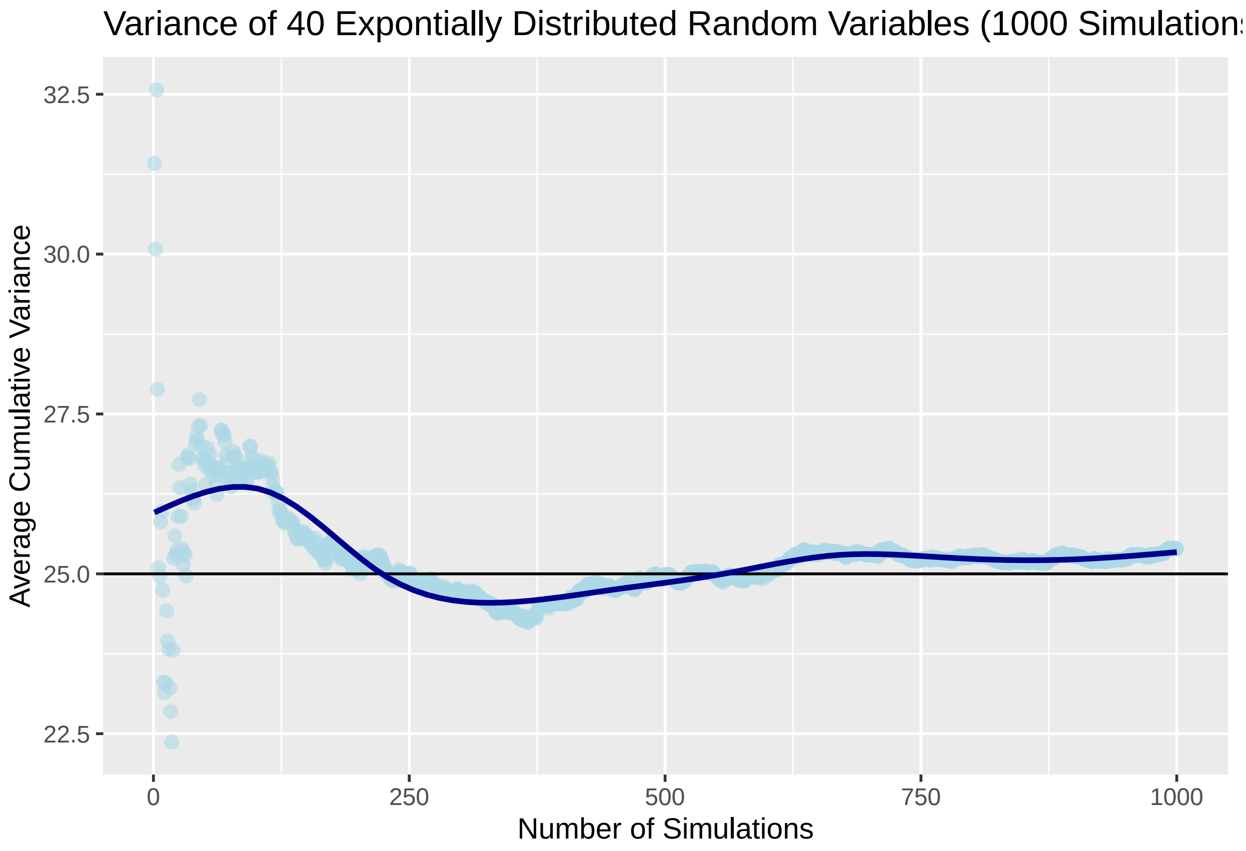
A smoothing curve was added to the scatter plot above, as well as the theoretical mean when lambda = 0.2. As can be seen, the average cumulative means begin to get closer to the theoretical mean as the number of simulations increases. This is to be expected because as the number of trials increases, so does the sample size from which the average mean is calculated. Though the average mean may not be closest to the theoretical mean at the 1000th simulation when compared to the previous simulations, for a sufficiently large number of simulations it will.

Sample Variance versus Theoretical Variance

The theoretical standard deviation of the exponential function is 1/ lambda. For these experiments, the theoretical standard deviation = 1/ 0.2 = 5. This means the theoretical variance is 5^2 = 25.

Because of the LLN, we know that the average of the samples collected limit to the theoretical standard deviation and variance. To demonstrate this, a scatter plot was created with the number of simulations along the x-axis and the average cumulative variances from those numbers of simulations along the y-axis.

```
ggplot(df, aes(x=(1:1000), y=cum_variances)) +
  geom_point(color = "lightblue", alpha = .58, size = 2) +
  geom_hline(yintercept=(1/lambda)^2) +
  geom_smooth(color = "darkblue", se=FALSE) +
  xlab("Number of Simulations") +
  ylab("Average Cumulative Variance") +
  ggtitle("Variance of 40 Expontially Distributed Random Variables (1000 Simulations)")
```

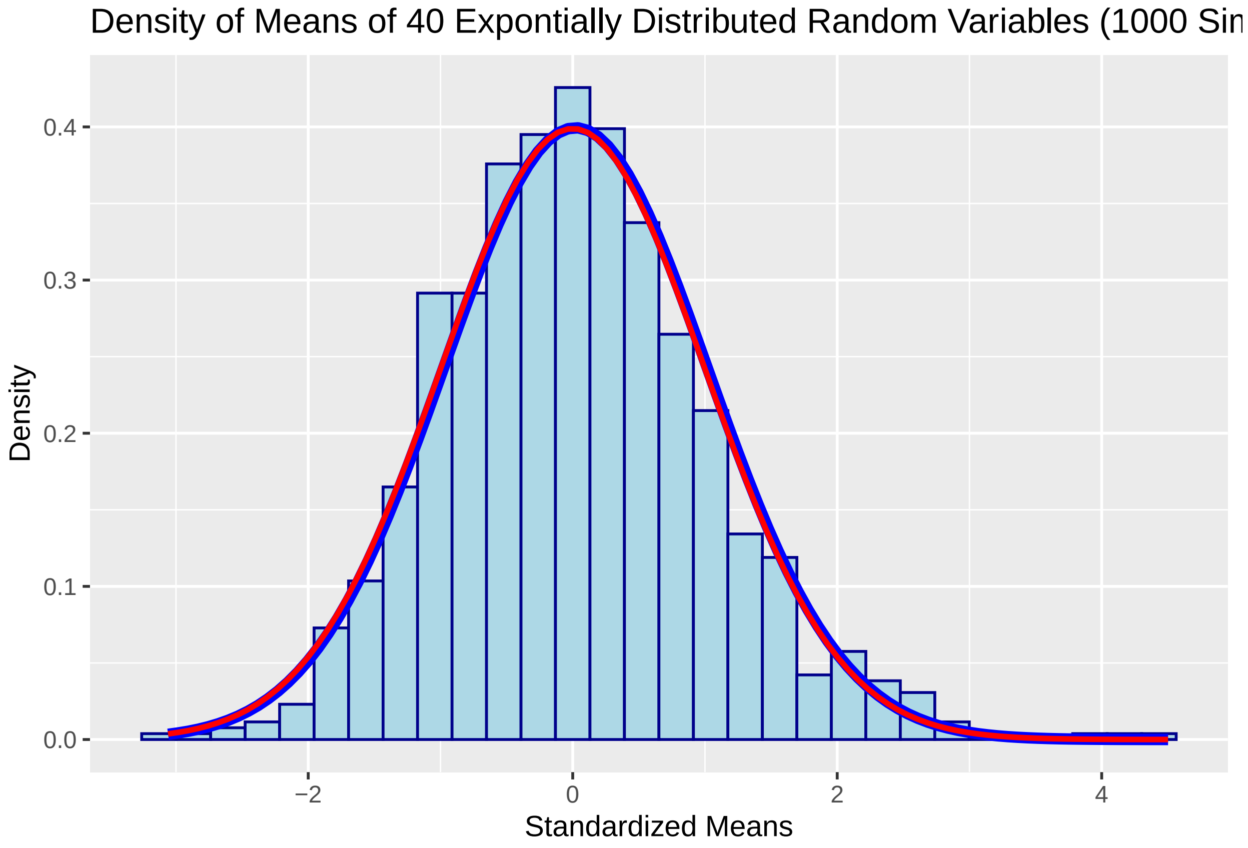


A smoothing curve was added to the scatter plot above, as well as the theoretical variance when lambda = 0.2. As can be seen, the average cumulative variances begin to get closer to the theoretical variance as the number of simulations increases. This is to be expected because as the number of trials increases, so does the sample size from which the average variance is calculated. Though the average variance may not be closest to the theoretical mean at the 1000th simulation when compared to the previous simulations, for a sufficiently large number of simulations it will.

Comparing Distributions

The Central Limit Theorem (CLT) states that the distribution of averages (once properly normalized) becomes that of a standard normal distribution as the sample size increases. To demonstrate this, a histogram was made of the averages from each experiment containing 40 random variables from the exponential distribution. The means were standardized by subtracting the theoretical mean, and dividing that difference by the standard error.

```
ggplot(df, aes(stnd_means)) +
  geom_histogram(aes(y=..density..), color="darkblue", fill="lightblue") +
  stat_function(fun = dnorm,
                args = list(mean=mean(df$stnd_means),
                            sd=sd(df$stnd_means)),
                col = "blue",
                size = 2) +
  stat_function(fun = dnorm,
                args = list(mean=0,
                            sd=1),
                col = "red",
                size = 1) +
  xlab("Standardized Means") +
  ylab("Density") +
  ggtitle("Density of Means of 40 Expontially Distributed Random Variables (1000 Simulations)")
```



As can be seen, the histogram representing our experimental data appears to be normally distributed. The blue line is the normal density function centered at average of means from each experiment with the standard deviation from the means from each experiment. The red line is the standard normal density function. These two lines nearly perfectly overlap each other, showing that the means from our experimental data are normally distributed. Because of CLT, for a sufficiently large number of simulations these two lines would perfectly overlap.