Study of sample means for a simulated exponential distribution

Syed Abdullah Hasan

6/20/2021

Overview

This report is prepared as the final project submission for the Statistical Inference course offered by Johns Hopkins University on Coursera. In the first part of this report, a simulation is conducted to evaluate the distribution of the sample mean and variance for an exponential distribution. The simulation includes comparisons between the sample means and theoretical mean, sample variance and theoretical variances, and concludes with a confirmation on the normal distribution of sample means.

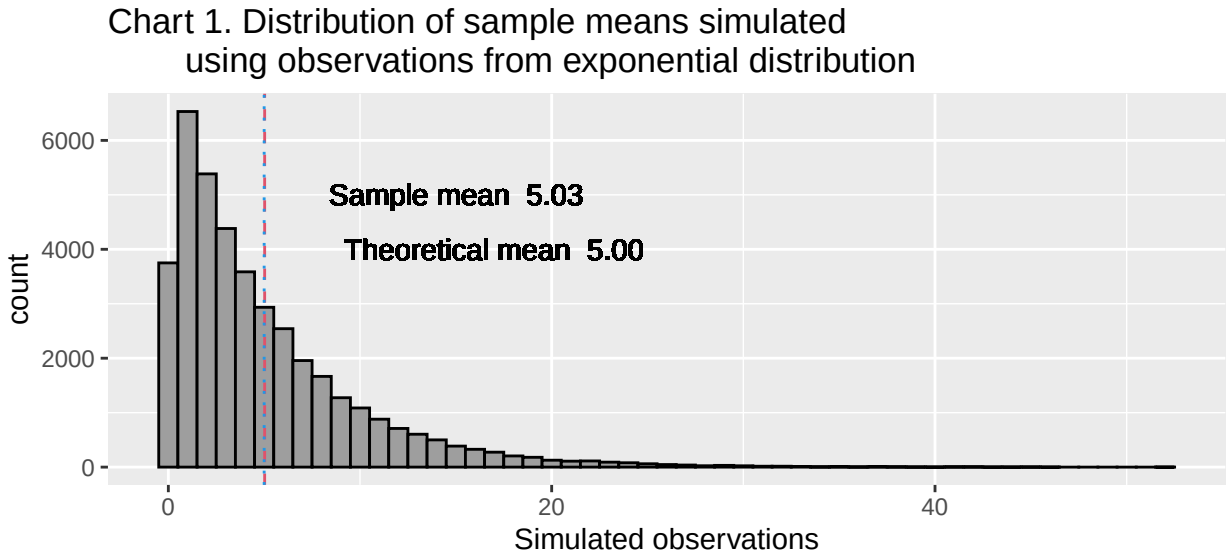
Part 1 - Simulation of an Exponential Distribution

Simulations

For the first section, a simulation of the exponential distribution is conducted by generating 40,000 random observations in a 40 x 1,000 dimension matrix, with lambda set at 0.2. The simulated observations are plotted in Chart 1 below for reference.

```
set.seed = 12345
lambda = 0.2; n = 40
obs_exp =  matrix(rexp(n*1000,lambda),1000,n)
df_obs_exp <- data.frame("obs" = c(obs_exp))
mean_exp = mean(obs_exp)

g1 <- ggplot(df_obs_exp, aes(obs)) +
        geom_histogram(aes(obs), fill=8, col=1, binwidth = 1) +
        labs (title  = "Chart 1. Distribution of sample means simulated
        using observations from exponential distribution",
              x ="Simulated observations") +
        geom_vline (xintercept = mean(df_obs_exp$obs), lty=2, col=2) +
        geom_vline (xintercept = 1/lambda, lty=3, col=4) +
        geom_text (aes(x = mean(obs)+10, y = 5000, label = paste("Sample mean ",
                                    format(mean_exp, digits=2, nsmall=2)))) +
        geom_text (aes(x = 1/ lambda +12, y = 4000, label =
                         paste("Theoretical mean ", format(1/lambda, digits=2,
                                                           nsmall=2))))
g1
```



The sample means for these observations are then calculated row-wise and sample variance and mean determined for all observations. Additionally, the theoretical mean and variance is determined for the simulation for reference.

```
means_t = 1/lambda
sd_means_t = (1/lambda)/sqrt(40)
features_s = data.frame("sample_means" = apply(obs_exp,1,mean),
                        "var_sample" = apply(obs_exp,1,var),
                        "std_sample" = apply(obs_exp,1,sd))
```

Sample Mean versus Theoretical Mean

According to the Central Limit Theorem, the distribution of means and standard deviations of observations generated via the exponential distribution should follow a normal distribution with mean lambda and standard deviation of lambda / sqrt(n) where n is the sample size. Secondly, both the sample means and sample variances should approximate the theoretical or population mean and variance - thereby confirming that these are unbiased indicators.

To demonstrate this, a histogram of the sample means is plotted to examine distribution of means and the sample standard deviation is overlaid for reference.

```
g2 = ggplot(features_s) +
        geom_histogram (aes(sample_means), fill=8, binwidth = 0.2) +
        geom_density (aes(x=sample_means, y=0.22*..count..),
                      col = 2, size = 1,lty=1) +
        geom_vline (xintercept = means_t, col = 2, lty = 1, lwd = 1) +
        geom_vline (xintercept = mean_exp, col = 2, lty = 2, lwd = 1) +
        geom_vline (xintercept = mean_exp + sd(features_s$sample_means), col = 2, lty = 3,
                    lwd = 1.2) +
        geom_vline (xintercept = mean_exp - sd(features_s$sample_means), col = 2,  lty = 3,
                    lwd = 1.2) +
        labs (x = "Sample Means (n = 40)", y = "Frequency",
              title = "Chart 2 - Distribution of sample means for simulation")

g2
```

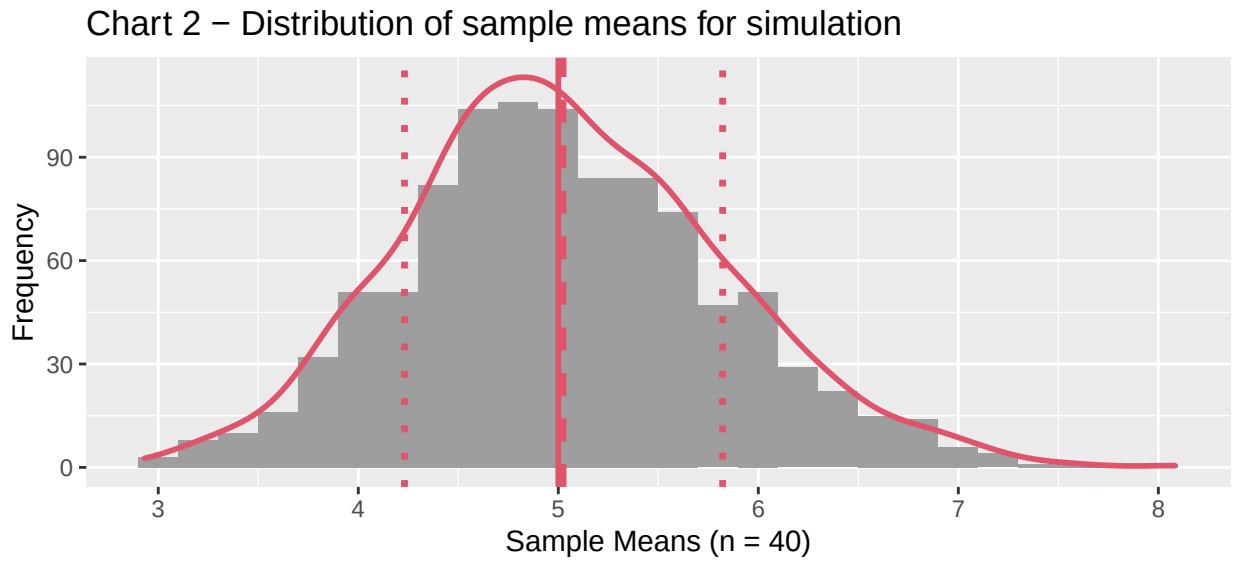


Chart 2 (above) shows the distribution of 1,000 sample means from random observations generated using the exponential distribution with 40 observations per sample and lambda equal to 0.2

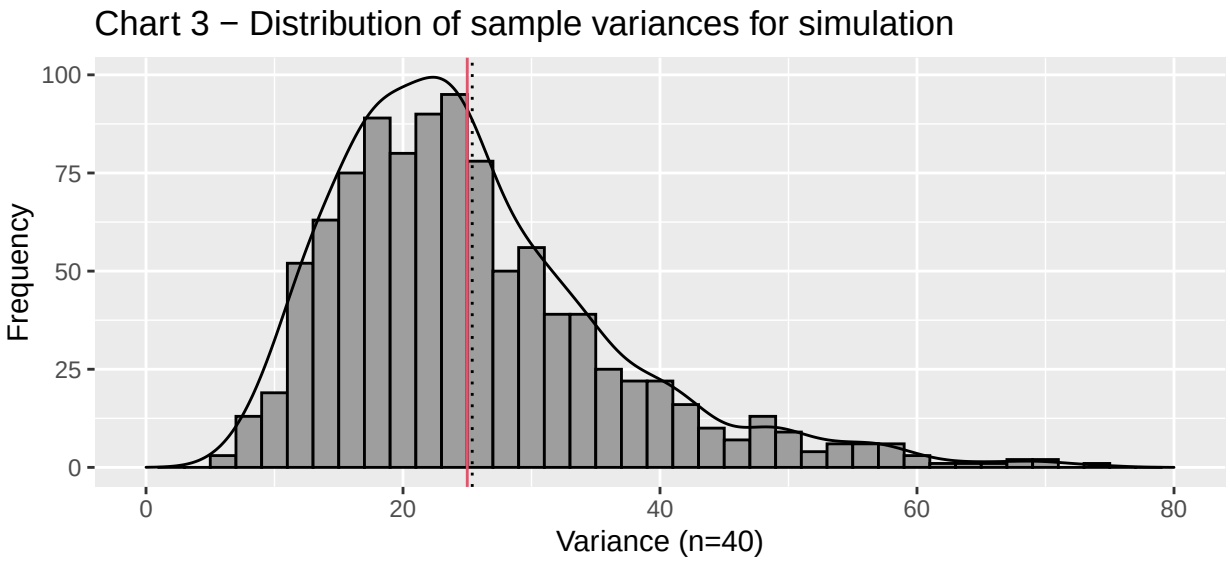
Theoretically, the population mean of this distribution should be 5.00 which is plotted via the line on the chart. As seen on the chart, the histogram of sample means of this distribution is similarly centered at the value 5.03 and plotted via the dashed line.

The variability in sample means is illustrated via dotted lines around the sample mean, representing a standard deviation of 0.7953. This conforms to the Central Limit Theorem which posits a theoretical standard deviation of 0.7906 for the distribution of sample means.

Sample Variance versus Theoretical Variance

```
g3 <- ggplot(features_s, aes(var_sample)) +
        geom_histogram (aes(x=var_sample), col=1, fill = 8, binwidth = 2) +
        geom_density (aes(y=2.25*..count..)) +
        geom_vline (xintercept = mean(features_s$var_sample), lty = 3) +
        geom_vline (xintercept = mean(1/lambda^2), col = 2) +
        xlim (c(0,80)) +
        labs( title = "Chart 3 - Distribution of sample variances for simulation",
              x = "Variance (n=40)",
              y = "Frequency")

g3
```



As seen in Chart 3 above, the distribution of sample variances is centered around the value of 25.3882 (dotted line) which is very close to the theoretical variance of 25.0000 (solid line). However, this distribution is not normal, unlike the distribution of sample means.

Distribution

The distribution of sample means appears to be approximately normal based on the bell curved shape seen in Chart 2. The blue highlighted histogram overlay in Chart 2 shows a simulated distribution of means based on normal variables to demonstrate a very closely approximated fit.

By comparing Chart 1 and Chart 2, we can also clearly conclude that while the distribution of simulated observations is negatively skewed, the distribution of sample means is normal.