Sampling Distributions of Exponential Random Variables

Julian Simington

2024-10-25

Overview

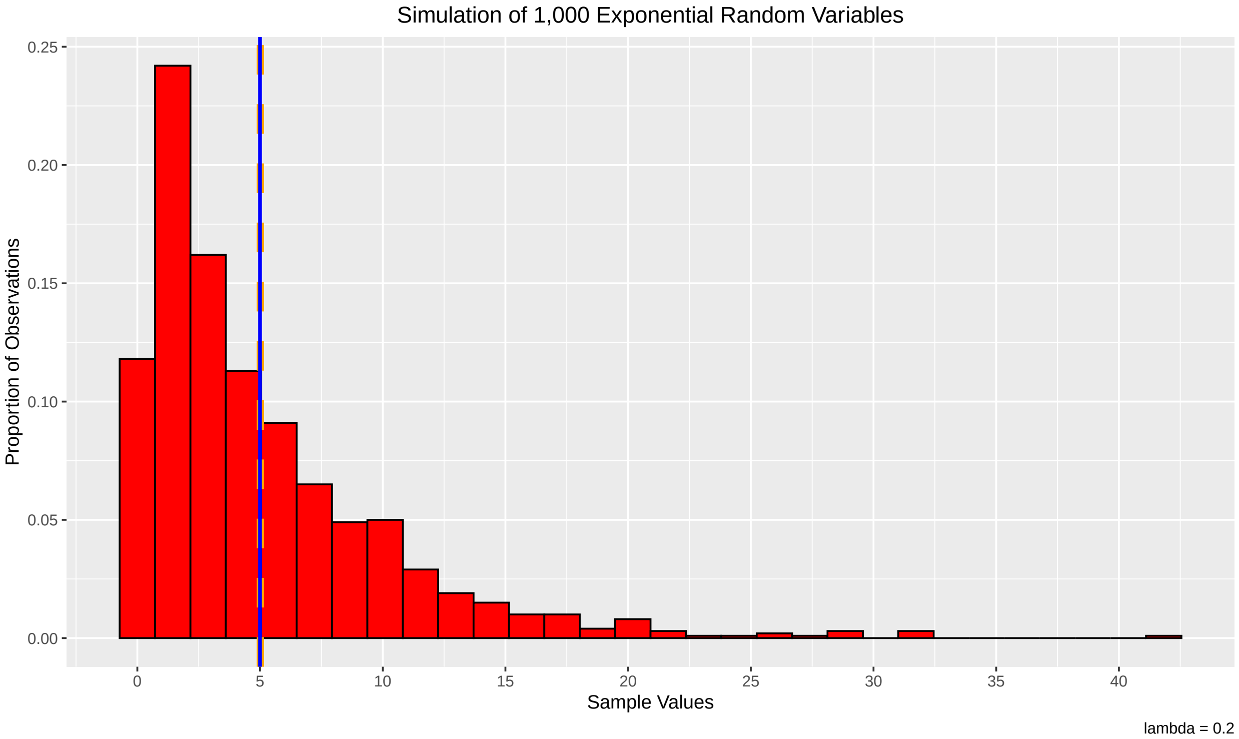
This project examines exponential random variables and their underlying sampling distributions. In particular it uses simulations to compare the sampling distribution of the sample mean to simulations of independent, identically distributed (iid) exponential variables. All simulations were taken from an exponential distribution with a mean and standard deviation of 5.

Package Dependencies

Some plots were created using the `ggplot2` package. See Appendix A.1 for installation code.

Simulating and Plotting Exponential Random Variables

A histogram of 1,000 *simulated* iid exponential random variables is plotted below. The related simulation and plotting code can be found in appendix A.2.

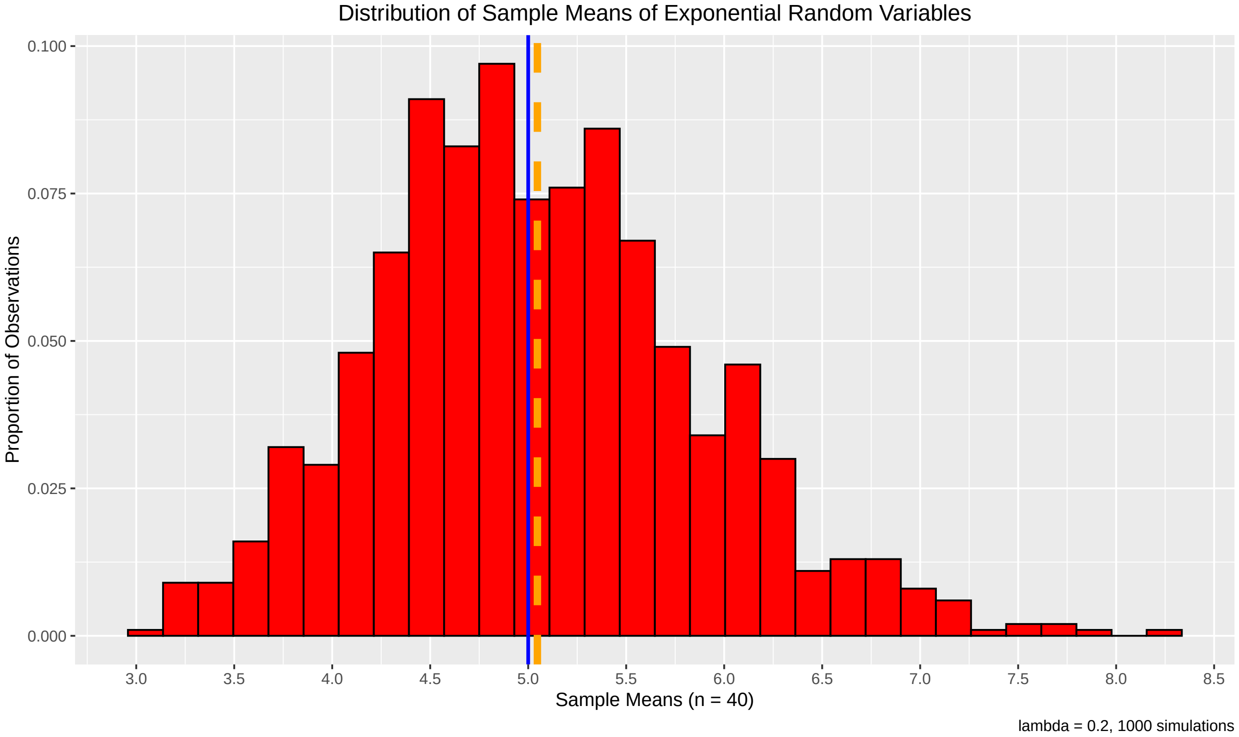


The simulated data resembles an exponential distribution as expected. Its mean of 5.013 is labeled at the dashed-orange line. The blue line labels the theoretical mean of 5. The standard deviation and variance of the simulated distribution are 5.108 and 26.094, respectively. These values are consistent with the respective theoretical values of 5 and 25. Overall, the observed values are close to the theoretical values which is consistent with the law of large numbers.

Summary Statistics of the Sampling Distribution of the Sample Mean

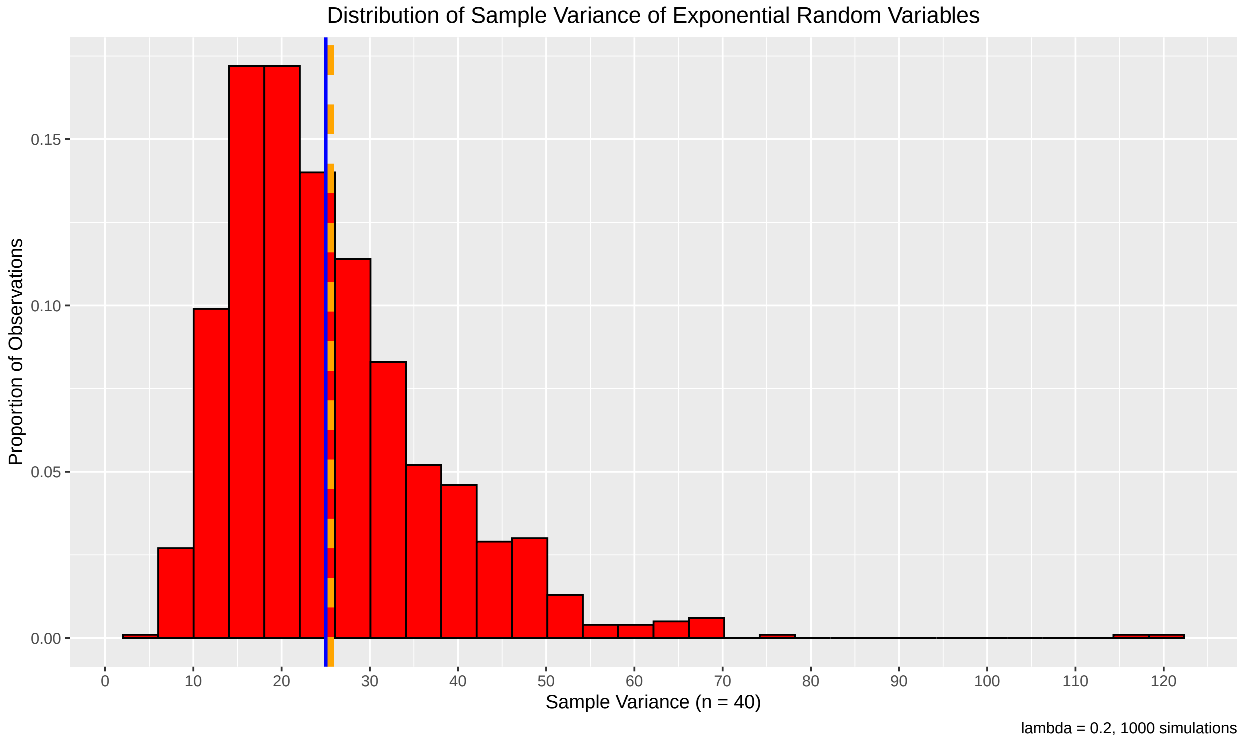
Additionally, 40,000 resulted were simulated from the same population. We partition the 40,000 simulated results into samples of size 40, resulting in 1000 trials total. This constructs the sampling distribution. The mean of the sampling distribution is 5.047, the standard error is 0.823, and the variance is 0.677. These values are close to the respective theoretical values of 5, 5/sqrt(40) = 0.791, and 25/40 = 0.625. The theoretical values are calculated using the probabilistic properties of iid random variables. See Appendix A.3 for the corresponding code that generates the observed statistics.

Plotting the Sampling Distribution of the Sample Mean



The histogram has slight right skew. Overall, it is roughly symmetric which is consistent with a normal distribution. See Appendix A.6 for a QQ plot providing further evidence of normality.

Sample Distribution of Sample Variance



The distribution of sample variances is right skewed with a mean of 25.515, which is close to the theoretical average variance of 25. The center of the sample and the population are labeled via the orange and blue lines, respectively. Note that the limiting distribution of the sample variance is a chi square distribution, which is also right skewed. See Appendix A.5 for related code and plots.

Conclusion

The analysis finds that the distribution of 1,000 simulated iid exponential variables differs considerably from the sampling distribution of the sample mean of 40 iid exponential variables. The corresponding means are quite close and near the theoretical mean of 5. However, the former distribution mirrors the population exponential distribution with regards to its shape and is skewed right with similar variance to the population. In contrast, the sampling distribution of the sample mean behaves in accordance with the CLT and has a normal distribution with the same theoretical mean, but much smaller variance. Again we note that the observed statistics for both distributions are consistent with the corresponding theoretical values.

Appendix

The appendix contains relevant code and figures that have been excluded from the main report for brevity. Note that some code chunks have set `eval = FALSE` so that the code is only **displayed**.

A.1 Package Dependencies Code

The code below installs required packages.

```
install.packages('ggplot2', repos = 'https://cran.r-project.org/')
library(ggplot2)
```

A.2 Simulation and Plotting Code

The code below simulates the 1,000 exponential random variables and plots a histogram indicating the observed and theoretical means.

```
set.seed(5) # set seed for reproducibility
lambda <- 0.2 # set rate parameter
sim_size <- 40 # sample size
num_sims <- 1000 # number of simulations
sims_1000 <- rexp(num_sims, rate = lambda) # 1000 iid sims
sims <- rexp(num_sims*sim_size, rate = lambda) # simulate sampling distr vals
sims_1000_mean <- mean(sims_1000) # sample mean of iid variables
sims_1000_sd <- sd(sims_1000) # sample sd of iid variables
exp_samp <- ggplot(data.frame(vals = sims_1000), aes(x = vals))
exp_sim_plot <- exp_samp +
           geom_histogram(aes(y = after_stat(count)/sum(after_stat(count))),
                  color = 'black', fill = 'red') +
           geom_vline(xintercept = mean(sims_1000), color = 'orange',
                  linetype = 'dashed', linewidth = 2) +
           geom_vline(xintercept = 1/lambda, color = 'blue', linewidth = 1) +
           xlab('Sample Values') +
           ylab('Proportion of Observations') +
           ggtitle('Simulation of 1,000 Exponential Random Variables') +
           scale_x_continuous(breaks = seq(0,75, by = 5)) +
           theme(plot.title = element_text(hjust = 0.5), ) +
           labs(caption = 'lambda = 0.2')
```

```
exp_sim_plot
```

A.3 Sampling Distribution Statistics

The code below generates the sampling distribution of the sample means (sample size 40, 1000 simulations).

```
exp_sims <- matrix(sims, nrow = num_sims, ncol = sim_size) # create matrix
sample_means <- apply(exp_sims, MARGIN = 1, FUN = mean) # form sampling distr
samp_distr_mean <- mean(sample_means) #  mean of sampling distr
samp_distr_sd <- sd(sample_means) # standard error of sampling distr
```

A.4 Sampling Distribution Plotting

The code below plots the sampling distribution of the sample means.

```
exp_means <- data.frame(samp_means = sample_means) # dataframe of simulated means

exp_samp_means <- ggplot(exp_means, aes(x = samp_means)) # create plot object

samp_mean_plot <- exp_samp_means +
              geom_histogram(aes(y = after_stat(count)/sum(after_stat(count))),
                          color = 'black',fill = 'red') +
              geom_vline(xintercept = samp_distr_mean, color ='orange',
                  linetype = 'dashed', linewidth = 2) + # dashed line at  mean
              geom_vline(xintercept =1/lambda, color = 'blue',
                  linewidth = 1 ) + # solid line at population mean
              xlab('Sample Means (n = 40)') +
              ylab('Proportion of Observations') +
              ggtitle('Distribution of Sample Means of
                      Exponential Random Variables') +
              scale_x_continuous(breaks = seq(0,9, by = 0.5)) +
              theme(plot.title = element_text(hjust = 0.5)) +
              labs(caption = 'lambda = 0.2, 1000 simulations')
```

```
samp_mean_plot
```

A.5 Sampling Distribution of Sample Variance Code

The code below generates and plots the sampling distribution of the sample variance.

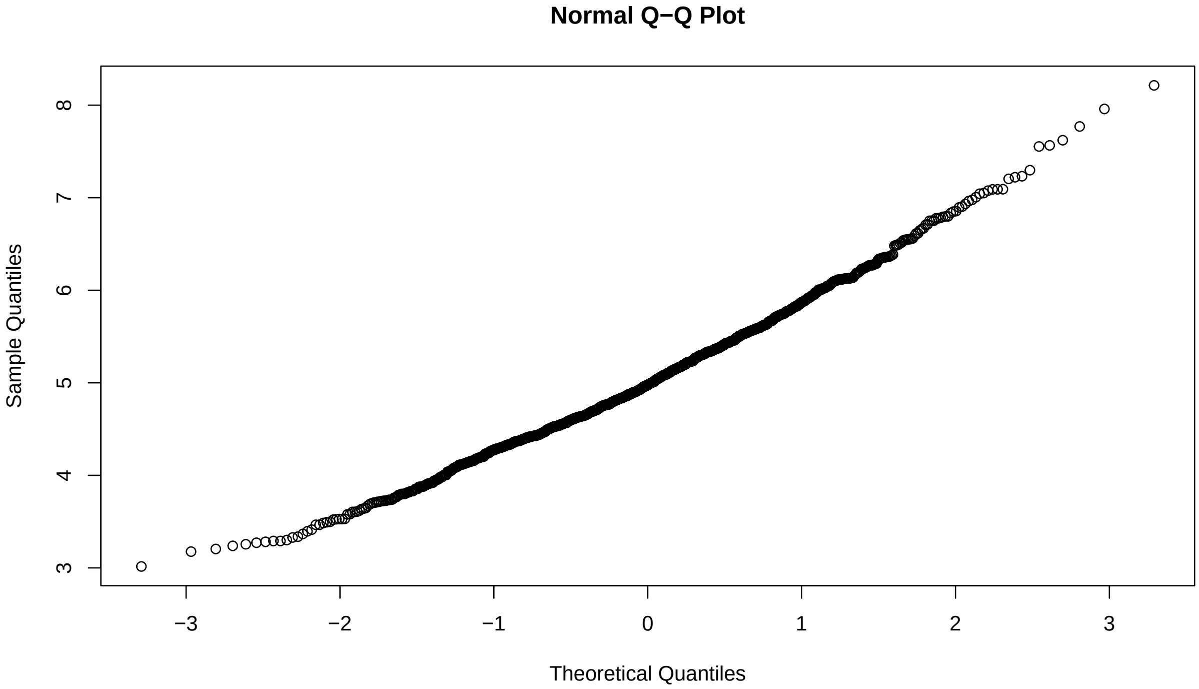
```
sample_vars <- apply(exp_sims, MARGIN = 1, FUN = var) # variance of each sample
mean_sample_var <- mean(sample_vars) # mean of sample variances
exp_vars <- data.frame (samp_vars = sample_vars)
exp_samp_vars <- ggplot(exp_vars, aes(x = samp_vars))
samp_var_plot <- exp_samp_vars +
                geom_histogram(aes(y=after_stat(count)/sum(after_stat(count))),
                           color = 'black', fill = 'red') +
                geom_vline(xintercept = mean_sample_var, color ='orange',
                           linetype = 'dashed',
                           linewidth = 2) + #vertical line through mean sample variance
                geom_vline(xintercept = (1/lambda)^2,
                           color = 'blue', linewidth = 1 ) + # line through pop variance
                xlab('Sample Variance (n = 40)') +
                ylab('Proportion of Observations') +
                ggtitle('Distribution of Sample Variance of Exponential Random Variables') +
                scale_x_continuous(breaks = seq(0,130, by = 10)) +
                theme(plot.title = element_text(hjust = 0.5)) +
                labs(caption = 'lambda = 0.2, 1000 simulations')
```

```
samp_var_plot
```

A.6 QQ Plot

The code below generates a QQ plot to assess normality of the sampling distribution of sample means.

```
qqnorm(sample_means) ## qq norm plot to check normality
```



The QQ plot of normal quantiles demonstrates some departures from linearity in the tails, but is roughly linear. Thus we have additional evidence that the sampling distribution of the sample mean behaves reasonably in line with the CLT.