Class 6 - Course Project: Analysis of Exponential Distributions

Nicholas Ng

Monday, January 26, 2015

Overview

This report was written as part of the course project requirements of class 6 of the Data Science specialisation, Statistical Inference. The aim of this report is to analyse the behaviour of the exponential distribution and asymtoptics. Much of the code has been provided by the lecturer as part of the course.

Simulations

Before the distribution can be analysed, the exponential distribution should randomly generated. As required in the course project, there are 1,000 runs by 40 observations, thus allowing us to generate a data frame of 1,000 observations of sample means and variances. The following code achieves this by first setting the parameters of the simulation (including the RNG seed), followed by the random generation of the exponentially distributed statistics. The code then calculates the mean and variance of the statistics while storing them in a data frame.

```r
set.seed(100)
sim.runs <- 1000
lambda<- 0.2

cfunc <- function(x, n) sqrt(n) * (mean(x) - 1 / lambda) / (1 / lambda)
randvars <- matrix(rexp(sim.runs * 40, lambda), sim.runs)
dat <- data.frame(x = c(apply(randvars, 1, cfunc, 40)),
                  samplemeans = apply(randvars, 1, mean),
                  samplevar = apply(randvars, 1, var),
                  size = factor(rep(c(40), rep(sim.runs, 1))))
```

To present the data in a visual format, the ggplot2 package is required:

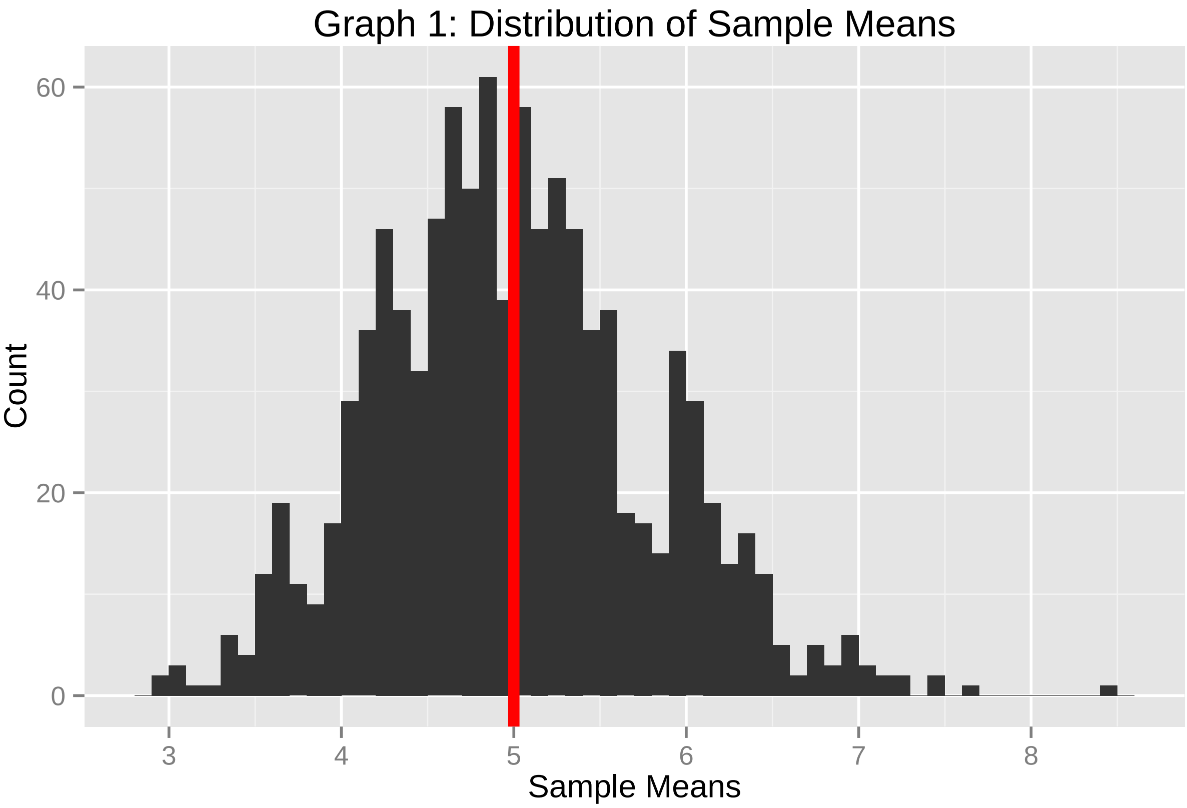
```r
suppressMessages(require("ggplot2"))
```

Sample Mean versus Theoretical Mean

The theoretical mean of the exponential distribution is $1/lambda$, hence given that we have set lambda = 0.2, the means of the sample should theoretically tend to 5 as the number of observations tends to infinity.

To study if this is the case, the following code plots the histogram of the sample means with an annotation on the theoretical mean.

```
g1 <- ggplot(dat, aes(x = samplemeans)) +
        geom_histogram(binwidth = 0.1) +
        labs(title = "Graph 1: Distribution of Sample Means",
             x = "Sample Means",
             y = "Count") +
        geom_vline(xintercept = 1 / lambda, colour = "red", size = 2)
print(g1)
```

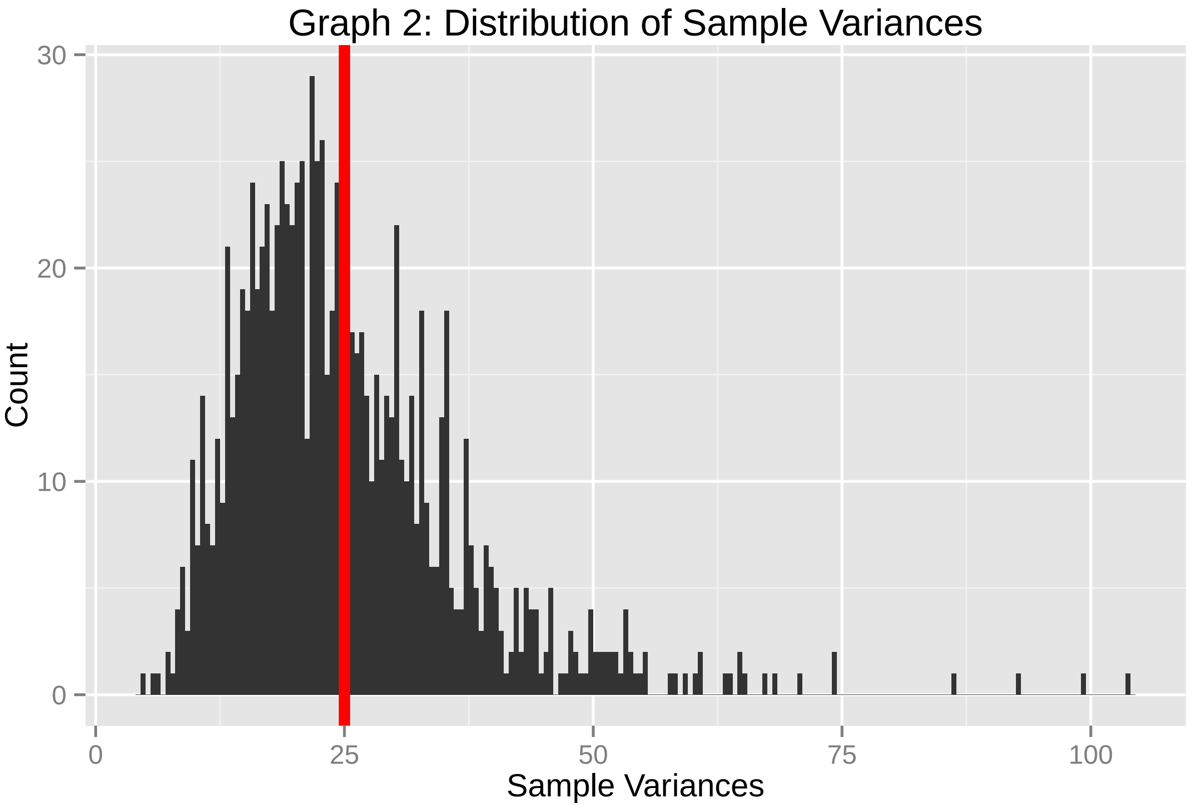


As seen in the above plot, the distribution looks approximately centred around the red line, indicating the theoretical mean of the exponential distribution. If more simulations were run, the result would likely become more obvious, but this plot should be sufficient in indicating that the mean does tend towards the theoretical mean.

Sample Variance versus Theoretical Variance

The theoretical variance of the exponential distribution is $(1/lambda)^2$, hence given that we have set lambda = 0.2, the means of the sample should theoretically tend to 25 as the number of observations tends to infinity.

```
g2 <- ggplot(dat, aes(x = samplevar)) +
        geom_histogram(binwidth = 0.5) +
        labs(title = "Graph 2: Distribution of Sample Variances",
             x = "Sample Variances",
             y = "Count") +
        geom_vline(xintercept = (1 / lambda)^2, color = "red", size = 2)
print(g2)
```



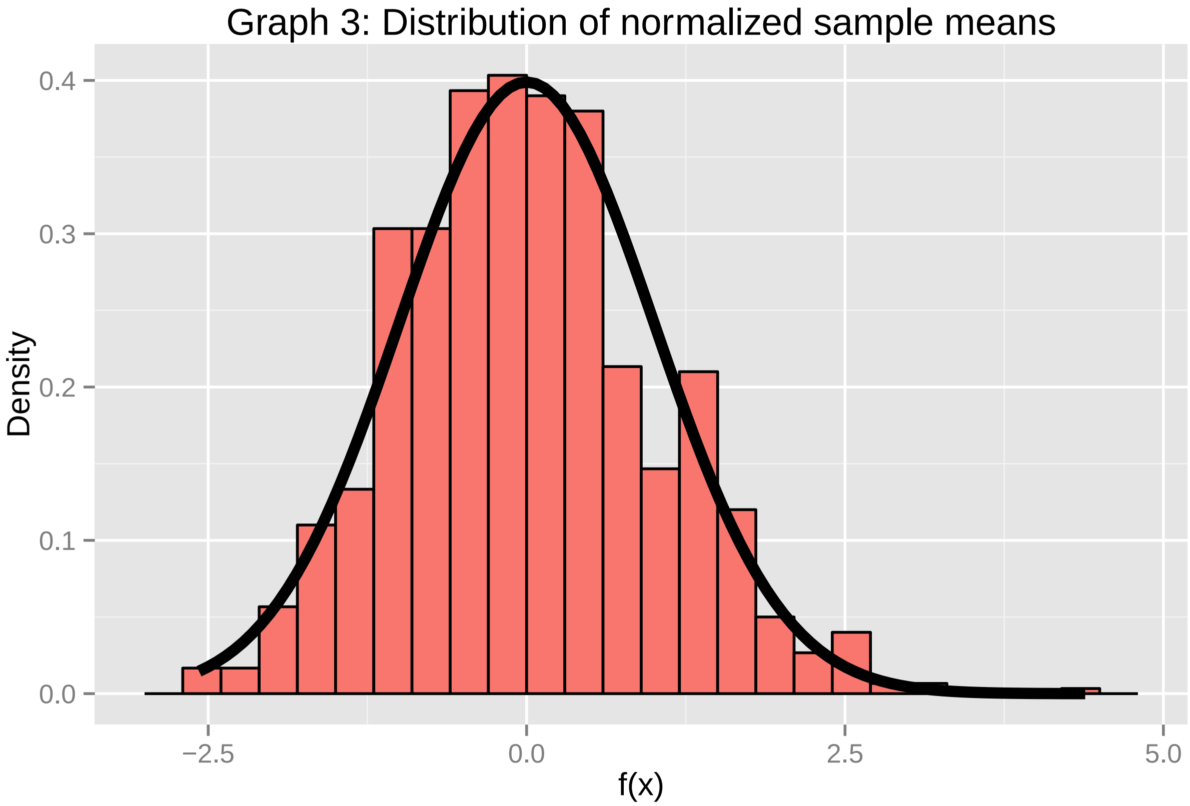
In this particular case, it seems that the distribution of sample variance is not centred around the red line, indicating the theoretical variance of the distribution. This may likely be due to the random number generator seed setting, or an insufficient number of simulations run.

It is likely that with more simulations, one should observe that the distribution will centre more around the red line.

Distribution

Lastly, to study if the exponential distribution tends to the normal distribution as the number of observations tend to infinity, the distribution of the normalised sample means (red bars) is plotted against the standard normal distribution (black line).

```
g3 <- ggplot(dat, aes(x = x, fill = size)) +
        geom_histogram(binwidth = .3, colour = "black", aes(y = ..density..)) +
        stat_function(fun = dnorm, size = 2) +
        labs(title = "Graph 3: Distribution of normalized sample means",
             x = "f(x)",
             y = "Density") +
        theme(legend.position = "none")
print(g3)
```



With this many number of draws, it should be expected that the exponential distribution may be approximated by the standard normal distribution, as per the central limit theorem. As seen, the red bars fit the shape as given by the black line.