Statistical Inference Course Project

Pradeep Babburi

10/25/2016

Part 1: Analysis of Exponential Distribution in R

Introduction

This report is an illustration of the analysis on exponential distribution in R in comparision to the Central Limit Theorem. Analysis is made by comparing the theoretical mean and variance to the sample mean and its variance for a sample of 40 exponents. As suggested in the guidelines, the rate parameter for the exponential distribution *lambda* is set to a constant value 0.2 for all the simulations and the number of simulations is set to 1000.

According to the definition of exponential function, the theoretical mean and standard deviation are calculated as

$$\mu = \sigma = \frac{1}{\lambda} = \frac{1}{0.2} = 5$$

Analysis of Mean

The R function **rexp** is used to generate an exponential distribution of 40 samples and is simulated for *sim* = 1000 times. A data matrix is created with 1000 rows of 40 columns each. The means of all 1000 rows are calculated and stored in the variable *mns*.

```
set.seed(222)
lambda <- 0.2                  # rate parameter
sim <- 1000                    # number of simulations
tmn <- tsd <- 1/lambda         # theoretical mean and standard deviation
dt <- rexp(40*sim, lambda)
m.dt <- matrix(dt, nrow = 1000, byrow = T) # matrix of 1000 rows with 40
exponents each.
dim(m.dt)

## [1] 1000   40

mns <- apply(m.dt, 1, mean)       # row wise calculation of mean
```

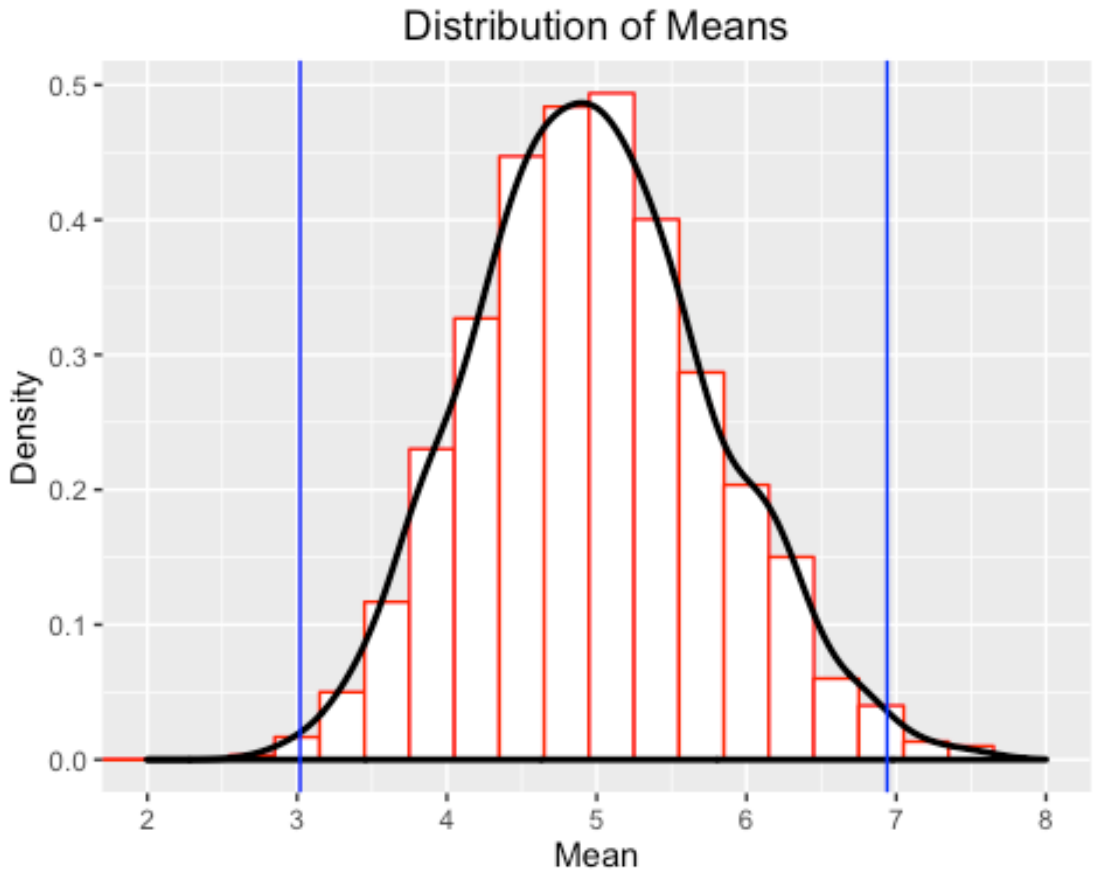
Let us see how a random mean from the 1000 means will compare to the theoretical mean.

```
cat('The value of sample  mean is ', round(mns[sample(1:sim, 1)],2))

## The value of sample  mean is  5.3
```

From the output we see that a randomly picked sample mean seems like a good estimate to the actual or theoretical mean. Next, let's see how the distribution of these 1000 means looks like.

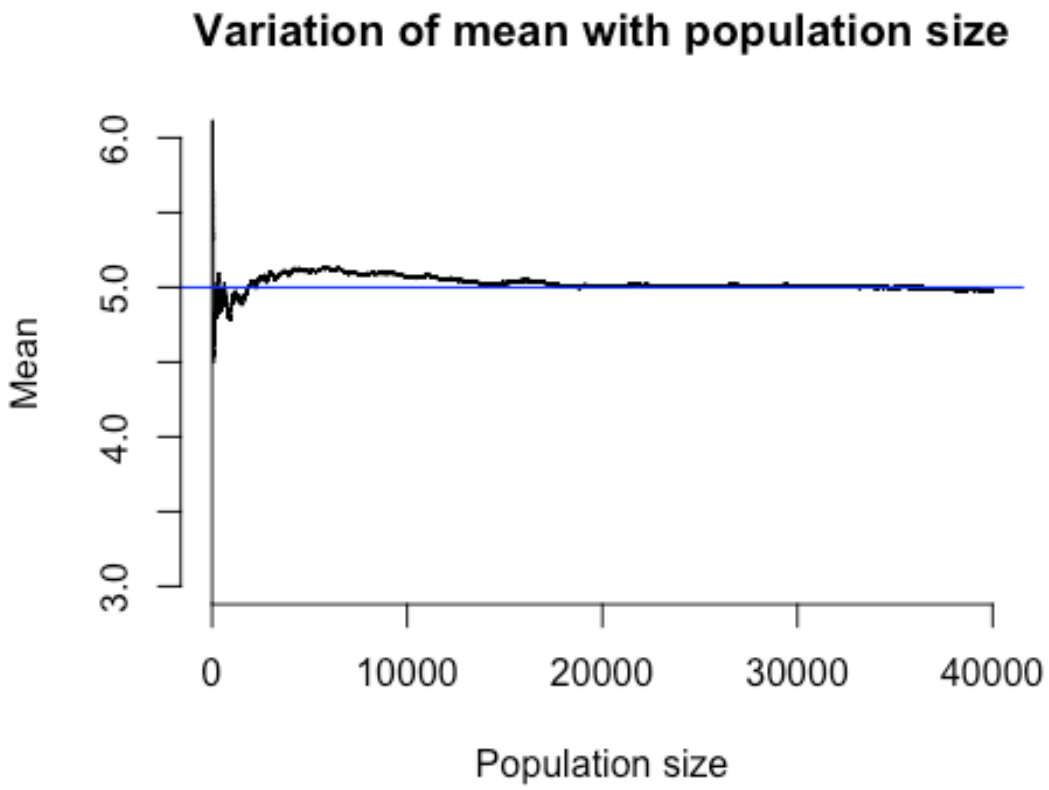
```
mns <- data.frame(m = mns)
g <- ggplot(mns, aes(x = m))
g <- g + geom_histogram(binwidth = 0.3, fill="white", colour = "red", aes(y =
..density..)) +  geom_density(size = 1)
g <- g + geom_vline(xintercept = qnorm(p = c(0.025, 0.975), mean =
mean(mns$m)), colour = "blue")
g + scale_x_continuous(limits = c(2,8), breaks = 2:8, na.value = 0) +
ggtitle("Distribution of Means") + labs(x = "Mean", y = "Density")
```



As we can see the distribution is centered around 5 which is the theoretical mean of the function and the density looks almost normal with peak located close to 5. It can also be inferred that 95% of the area under the curve falls within 2 standard deviations from the mean on either side. The vertical blue lines indicate the 0.025 and .975 quantile values which are close to 3 and 7 ($\mu \pm z_{1-\alpha/2}$) respectively. This tells us that there is 95% chance that any random draw from the 1000 means of 40 exponents estimates the mean to be within 2 standard deviations of the theoretical mean.

In addition, if we were to observe the value of mean with the population size, it can be said from the following graph that the mean converges to the theoretical mean as population size increases conforming to the Law of Large Numbers.

```
plot(1:length(dt), cumsum(dt)/1:length(dt), type = "l", xlab = "Population
size", ylab = "Mean", main = "Variation of mean with population size", ylim =
c(3,6), frame.plot = FALSE)
abline(h = tmn, col = "blue")
```



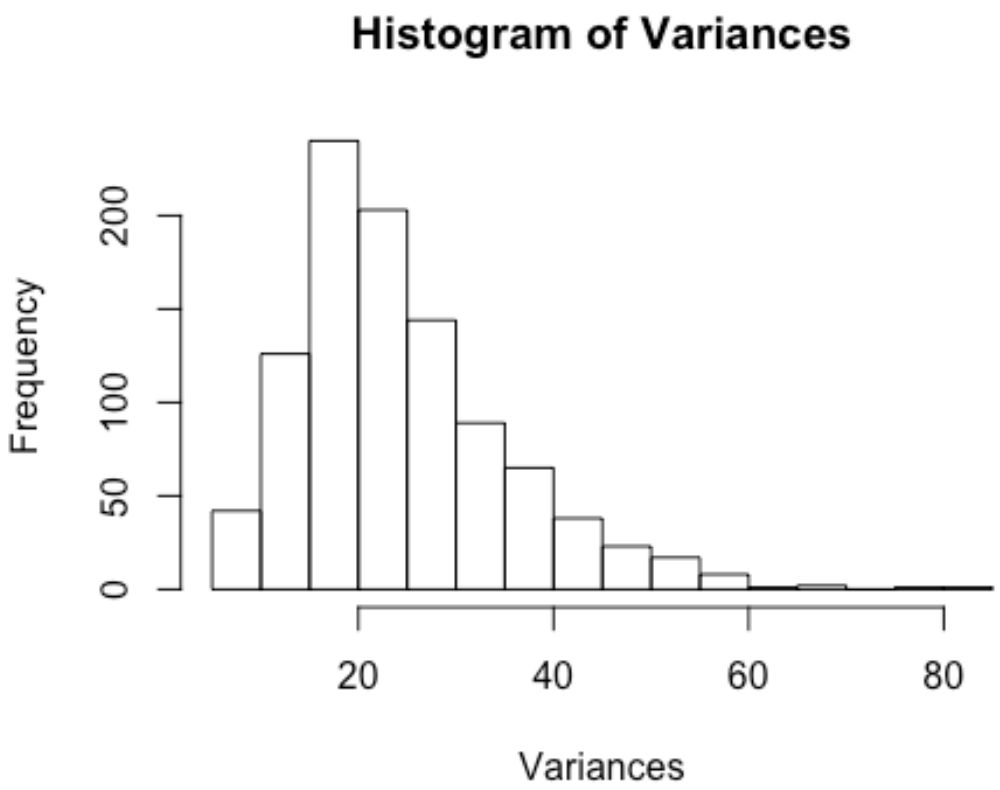
Analysis of Variance

Given a theoretical standard deviation of $\sigma = 5$, the theoretical variance can be calculated as square of the standard deviation i.e., $\sigma^2 = 25$. Now, let's see how the estimated variance from the simulated population compares to the actual one.

```
vrs <- apply(m.dt, 1, var)
cat('The value of sample variance is ', round(vrs[sample(1:sim, 1)],2))

## The value of sample variance is  40.9

hist(vrs, xlab = "Variances", main = "Histogram of Variances")
```



As can be seen in the histogram, the distribution looks a bit different than for the means but it can be said that the bulk of the variance lies around a value of 20. If we were to pick a random variance of 40 exponents from the population, the probability of estimating the variance close to the actual variance is comparatively less than that for the mean. However, the average of all 1000 variances yields a value of 24.95 which is pretty close to the actual variance.