Exponential Distribution and Central Limit Theorem

Martin Connolly

June 21, 2015

Overview

This report will investigate the exponential distribution in R and compare it to the Central Limit Theorem. This report will detail the investigation of the distribution of averages of 40 exponentials over 1000 simulations.

Simulations

In this investigation we use the `rexp` function within R to generate an exponential sample distribution, The mean of the exponential distribution population is $\mu = 1/\lambda$ and the standard deviation is $\sigma = 1/\lambda$.

```
library(ggplot2, quietly = TRUE)
library(dplyr, quietly = TRUE, warn.conflicts = FALSE)

lambda <- 0.2
n <- 40 #sample size
simul_count <- 1000 #number of simulations
dist_mu <- 1 / lambda #mean of population
dist_sigma <- 1 / lambda #standard deviation of population

set.seed(17) # set seed value for reproducability

# create empty data frame for simulation data
simul_data <- data.frame(matrix(NA, nrow = simul_count, ncol = n + 1))
names(simul_data)[n + 1] <- "sample_mean"

# generate 1000 random exponential distributions of sample size 40
# with each sample in a new row
for(i in 1:simul_count) {
        temp_data <- rexp(n, lambda)
        simul_data[i, ] <- c(temp_data, mean(temp_data))
}
```

Sample Mean versus Theoretical mean

As stated the theoretical mean (mean of the population) is $\mu = 1/\lambda$, as $\lambda = 0.2$ then $\mu = 5$

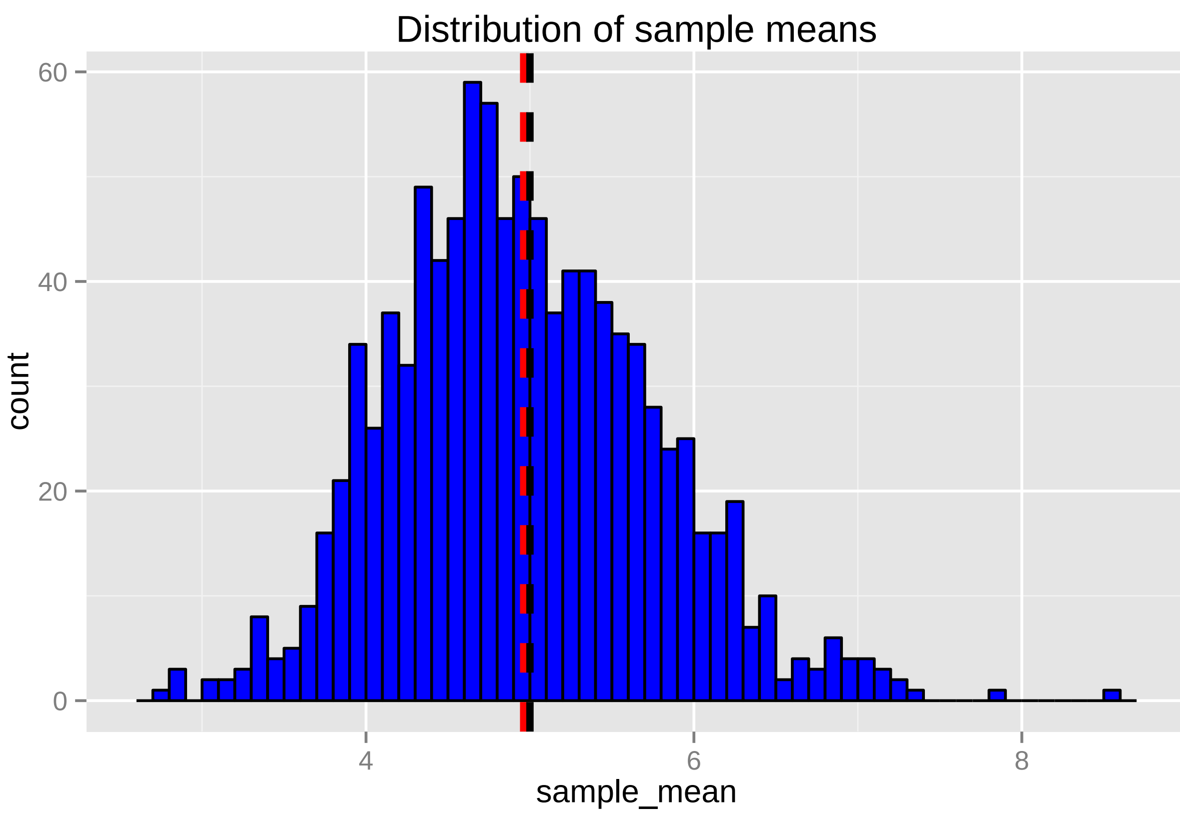
The mean of all the samples can be calculated as follows;

```
mean_of_all_samples <- mean(simul_data$sample_mean)
mean_of_all_samples
```

```
## [1] 4.961683
```

So we can see that mean of the sample distributions 4.9616831 is very near the mean of the population 5. Below you can see a plot of the distribution of sample means, with the mean of all samples marked by a red dotted line and the population mean marked with a black dotted line.

```
ggplot(simul_data, aes(x=sample_mean)) +
        geom_histogram(binwidth = 0.1, colour = "black", fill = "blue") +
        geom_vline(xintercept = mean_of_all_samples, size = 1.3, colour = "red",
                linetype = "dashed", show_guide = TRUE) +
        geom_vline(xintercept = 1 / lambda, size = 1.3, colour = "black",
                linetype = "dashed", show_guide = TRUE) +
        ggtitle("Distribution of sample means")
```



Sample Variance versus Theoretical Variance

The theoretical variance of the sample means is σ^2/n, which is 0.625. The actual variance of the sample means is as follows;

```
sample_var <- var(simul_data$sample_mean)
sample_var
```

```
## [1] 0.6408254
```

So the difference between the theoretical and actual variance can be seen as

```
(dist_sigma^2 / n) - sample_var
```

```
## [1] -0.01582538
```

Distribution

If we look at the distribution of sample means below, we can quickly see that it's approximately normal. The green line is the density function of the sample_mean distribution and the blue line is the normal distribution.

```
ggplot(simul_data, aes(x=sample_mean)) +
        geom_histogram(binwidth = 0.1, colour = "black", fill = "blue",
                       aes(y=..density..)) +
        geom_density(colour = "green", fill = "green", alpha = 0.4, size = 1) +
        stat_function(fun = dnorm, args = list(mean = dist_mu, sd = sqrt(dist_sigma^2 / n)),
                      colour = "blue", size = 1) +
        ggtitle("Distribution of sample means")
```