Statistical Inference Project (Part 1)

Eric Erkela

11/16/2020

Overview

In this section of the two-part statistical inference project, we will investigate the Central Limit Theorem (CLT) as it relates to estimates of population means taken from a simulated data set. From a randomly-generated, exponential data set, we will show that the distribution of sample averages is approximately normal, with mean and variance identical to the exponential distribution from which they were taken.

Dependencies

Below are all the packages used in this analysis, produced here for convenience:

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

Part 1: Simulations

In the first part of this assignment, we will be applying the Central Limit Theorem (CLT) to show the Gaussian convergence of a set of sample means, each taken from 40 random observations of an exponential distribution(lambda = 0.2). Before we do, however, we must generate the data itself:

```
set.seed(123)        # for reproducibility
n <- 1000; b <- 40  # number of averages, number of observations per average
lambda <- 0.2        # exponential rate, determined in project instructions
sim_data <- matrix(rexp(n * b, rate = lambda), nrow = n, ncol = b)
sim_means <- apply(sim_data, 1, mean)
head(sim_means)
```

```
## [1] 4.438396 5.698263 6.963634 4.702773 4.106572 4.665833
```

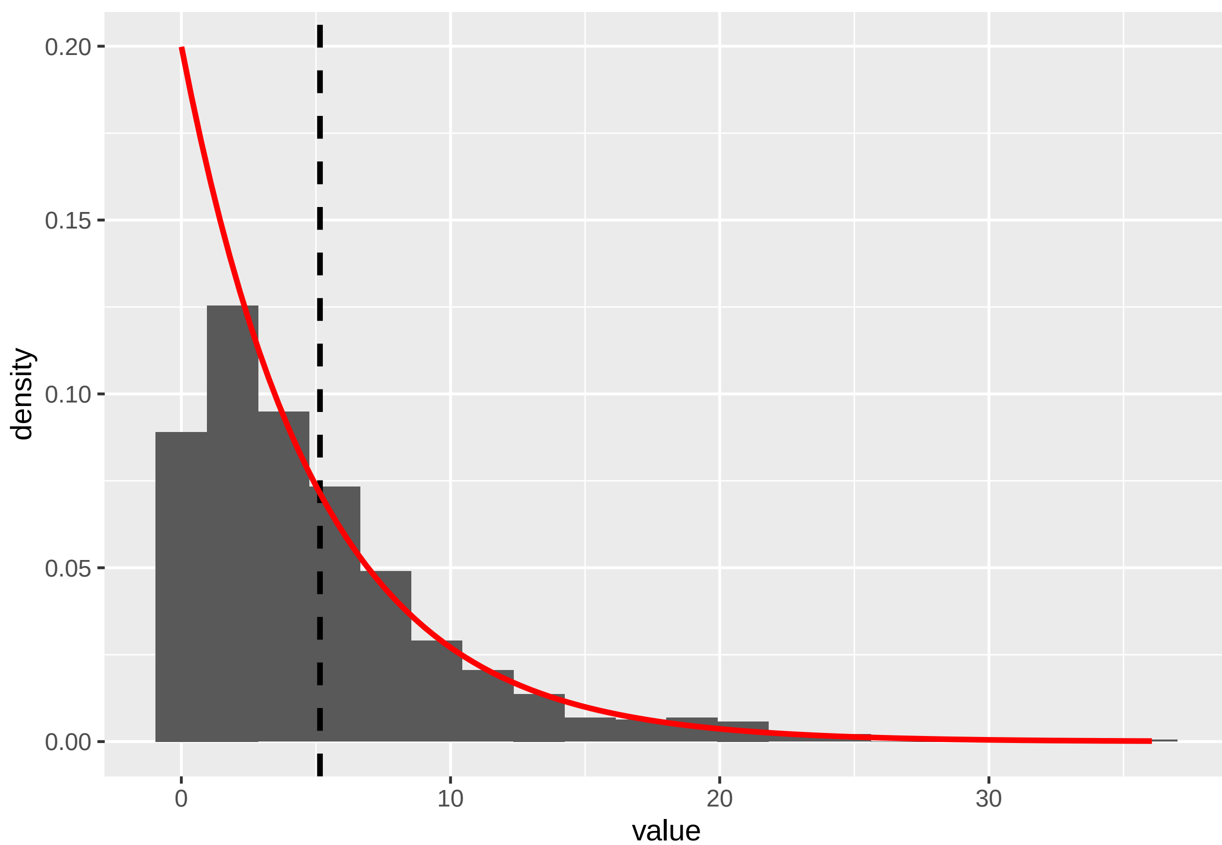
Et voilà! Since our data comes from an exponential distribution for which we know lambda, we can exactly calculate its expected mean/variance. How do these compare to what we've observed?

```
data.frame(Mean = c(mean(sim_means), 1 / lambda),
           Variance = c(var(sim_means), (1 / lambda)^2 / b),
           row.names = c("Observed", "Expected"))
```

```
##              Mean  Variance
## Observed 5.011911 0.6088292
## Expected 5.000000 0.6250000
```

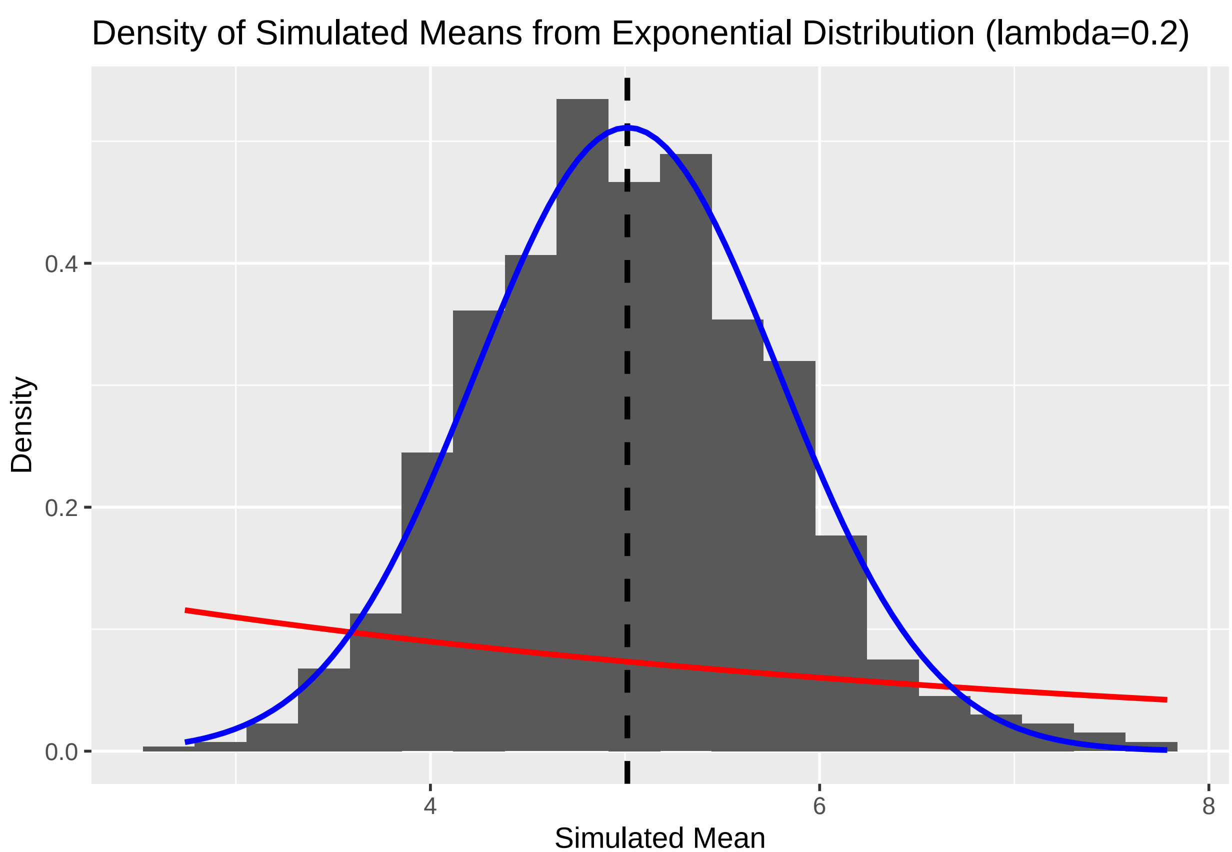
Pretty close! This shows that our simulated data does in fact estimate the theoretical population mean/variance of the exponential distribution it is based on, a basic tenet of statistical inference. If we were to plot the probability density of our sample means by their value, what distribution would we get? A naïve guess might be that they match the distribution from which they were taken, an exponential with lambda = 0.2. In this scenario, our data would look something like this, where we've substituted our sample means for the first column of our original, simulated data set.

```r
g <- ggplot(data = tibble(value = sim_data[, 1]), aes(x = value))
g +
  geom_histogram(aes(y = ..density..),
                 bins = 20) +
  geom_vline(xintercept = mean(sim_data[, 1]),
             linetype = "dashed",
             color = "black",
             size = 1) +
  stat_function(fun = dexp,
                n = 101,
                args = list(rate = lambda),
                color = "red",
                size = 1)
```



Makes sense, right? But what do we actually get?

```r
g <- tibble(value = sim_means) %>% ggplot(., aes(x = value))
g +
  geom_histogram(aes(y = ..density..),
                 bins = 20) +
  geom_vline(xintercept = mean(sim_means),
             linetype = "dashed",
             color = "black",
             size = 1) +
  stat_function(fun = dexp,
                n = 101,
                args = list(rate = lambda),
                color = "red",
                size = 1) +
  stat_function(fun = dnorm,
                n = 101,
                args = list(mean = mean(sim_means), sd = sd(sim_means)),
                color = "blue",
                size = 1) +
  labs(title = "Density of Simulated Means from Exponential Distribution (lambda=0.2)",
       x = "Simulated Mean",
       y = "Density")
```



As we can see, our sample means are clearly not exponentially distributed (the red line). Instead, we find that they follow a normal distribution (blue line) with mu = mean(sim_means), which matches the dashed black line in the figure above, and sd = sd(sim_means). This matches what we'd expect from the CLT.