Statistical Inference Course Project

Emmanouil Kalaitzakis

2021-07-02 11:54:02

This is a report for the final project of the Statistical Inference course offered by Johns Hopkins University on Coursera. This project is split into two parts:

- The first part involves a simulation exercise intended towards understanding the properties of the distribution of the sum of a number of independent and identically distributed exponential random variables.
- The second part involves the exploration of a data set, some statistical testing, and a discussion of the assumptions supporting the application of the methods.

Part 1: Simulation Exercise

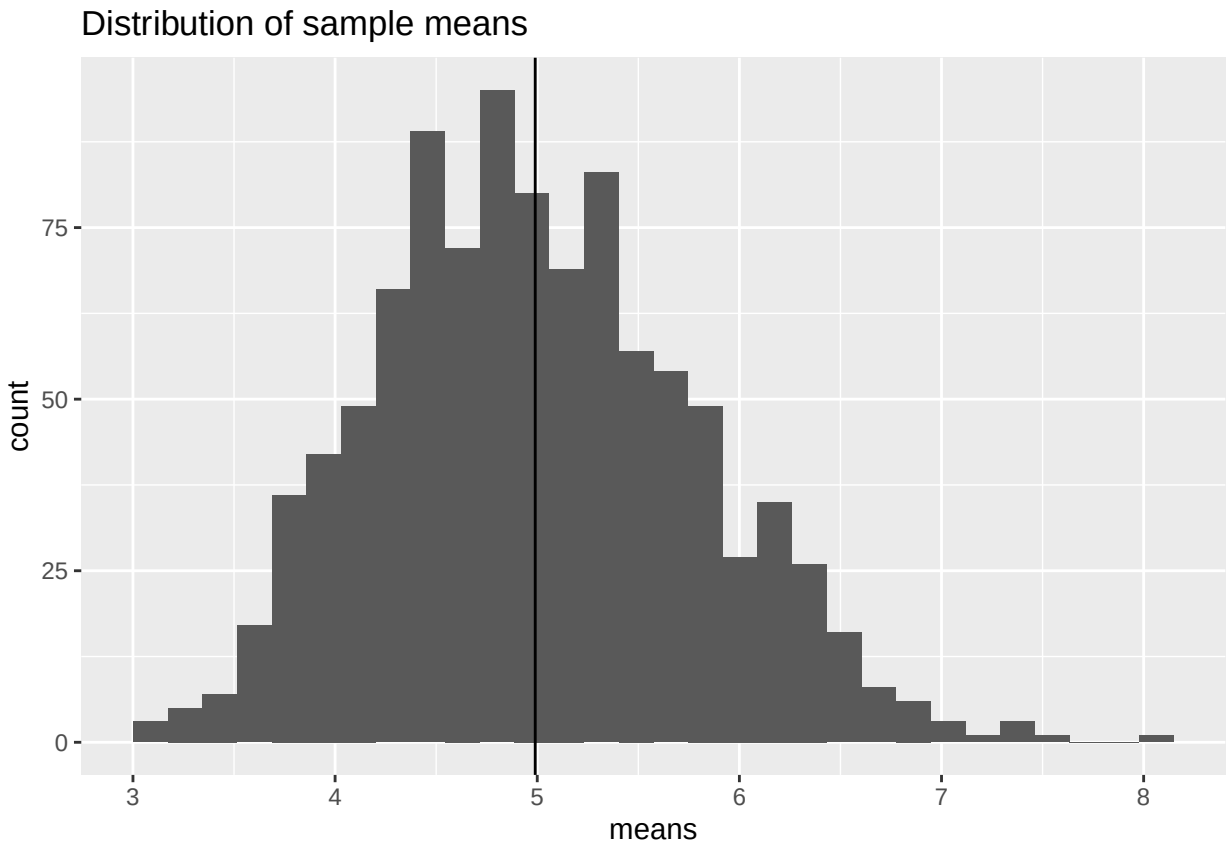
To perform the simulation we create a random matrix where each row is a simulation iteration and each column is one sample.

```
n <- 40
lambda <- 0.2
mu <- 1/lambda
sigma <- 1/lambda

iter <- 1000

set.seed(1)
mc <- matrix(data = rexp(iter*n, lambda), nrow = iter, ncol = n, byrow = FALSE)

mus <- apply(mc, 1, mean)
```

The sample mean of the distribution is 4.9900252 while the theoretical mean of the distribution is 5. The variance of the distribution is 0.6177072 while the theoretical variance of the distribution is 0.625. We can explore the distribution visually with the help of a histogram and a vertical line centered at the sample mean.

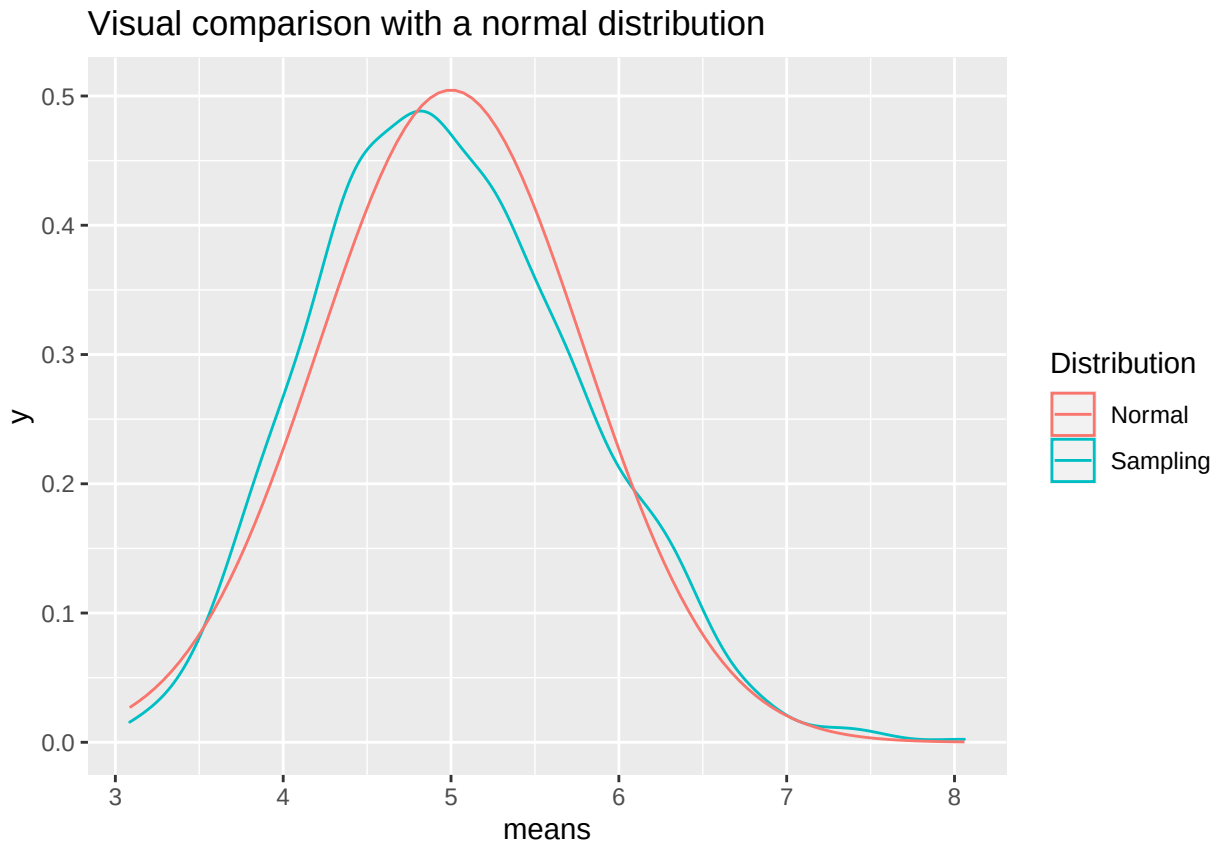
```
library(ggplot2)

ggplot(data.frame(means = mus), aes(means)) +
  geom_histogram() +
  geom_vline(xintercept = mean(mus)) +
  ggtitle("Distribution of sample means")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



To check whether the distribution is approximately normal, we plot it against a normal ditribution with the theoretical mean and variance. We also perform a Shapiro-Wilk normality test to support our finding.

```
ggplot(data.frame(means = mus), aes(means)) +
  geom_density(aes(colour = "Sampling")) +
  stat_function(fun = dnorm, args = list(mean = mu, sd = sigma/sqrt(n)), aes(colour = "Normal")) +
  labs(colour = "Distribution") +
  ggtitle("Visual comparison with a normal distribution")
```




```
shapiro.test(mus)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mus
## W = 0.99157, p-value = 1.759e-05
```

We observe that the sampling distibution is quite similar to the normal. The Shapiro-Wilk test supports this finding further since the p-value is much smaller than 0.1 which is threshold for the distribution to be considered approximately normal.