Statistical Inference Course Project: Part 1

Bruno

2023-08-04

Overview

In this project, we investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of an exponential distribution with rate λ is $\mu = 1/\lambda$, and the standard deviation is also $\sigma = 1/\lambda$.

Let us consider samples of $n = 40$ independent exponential random variables with $\lambda = 0.2$. In this work, we concluded the following tasks:

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

Simulations

In the following, we obtain a set of $N = 10000$ samples, where each one of them consists of $n = 40$ independent realizations of exponential random variables with rate function $\lambda = 0.2$.

```
## Number of samples
n <- 40

## Number of iterations
N <- 10000

## Rate parameter
lambda <- 0.2

## Simulation
sam <- matrix(rexp(n = n * N, rate = lambda), nrow = N, ncol = n)
```

Sample mean versus the theoretical mean

Let us consider the sample mean of the realizations of $n = 40$ exponential random variables and build a histogram in order to compare its relationship with the theoretical mean $\mu = 1/\lambda = 5$.

```
library(ggplot2)

mn <- apply(X = sam, MARGIN = 1, FUN = mean)

ggplot(data = data.frame(mn), aes(x = mn)) +
```

```r
        labs(title = "Histogram of population mean for n = 40",
             x = "Sample mean",
             y = "Count") +
        geom_histogram(color = "white",
                       fill = "blue") +
        geom_vline(xintercept = 5, color = "red")
```

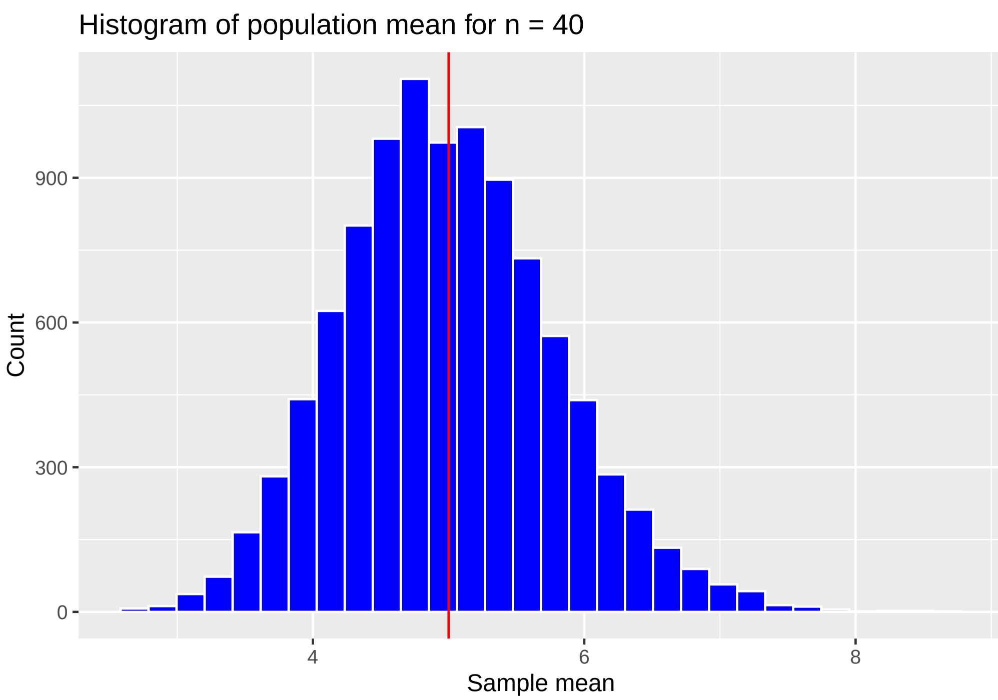


Figure 1: Histogram of sample mean

As we can see in the histogram from Figure 1, the values for the sample mean distributes around the population mean $\mu = 5$, which is consistent with the fact that the sample mean is a unbiased estimator for the population mean.

Sample variance versus theoretical variance

Similarly, let us consider the histogram of the sample variances and see how it compares with the theoretical variance $\sigma^2 = 1/\lambda^2 = 25$.

```r
var <- apply(X = sam, MARGIN = 1, FUN = var)

ggplot(data = data.frame(var), aes(x = var)) +
        labs(title = "Histogram of sample variances for n = 40",
             x = "Sample variance",
             y = "Count") +
        geom_histogram(color = "white",
                       fill = "red") +
        geom_vline(xintercept = 25)
```

Note that the histogram from Figure 2 distributes around the theoretical variance $\sigma^2 = 25$, being consistent with the fact that the sample variance is an unbiased estimator for the population variance.

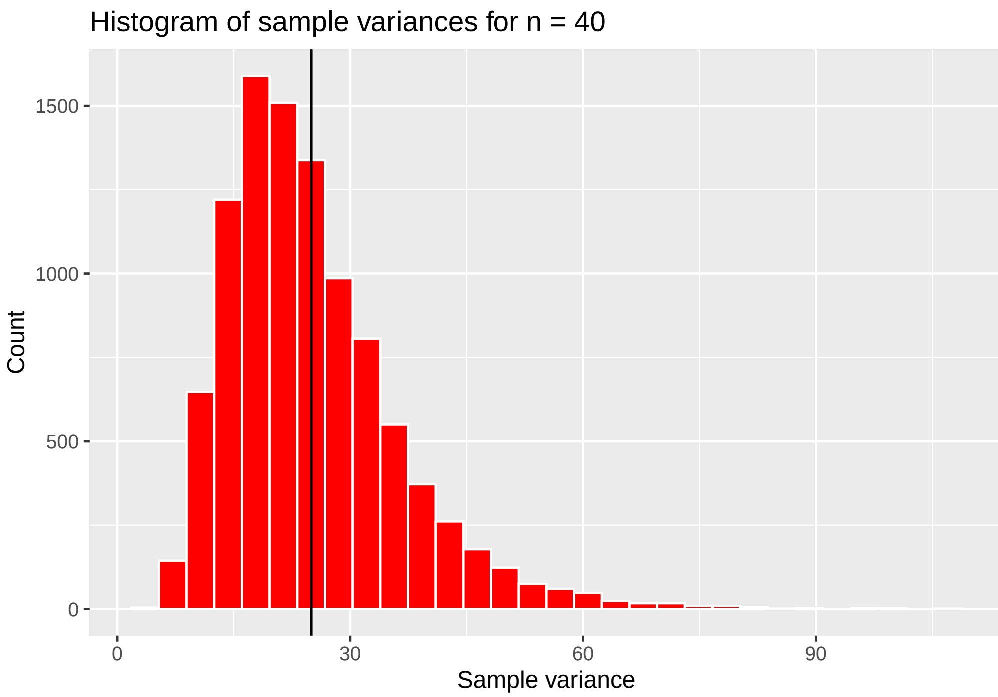


Figure 2: Histogram of sample variance

Central Limit Theorem

According to the Central Limit Theorem (CLT), given n identically distributed random variables $X_1, X_2, \ldots, X_n$ with mean μ and standard deviation σ, the quantity

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right),$$

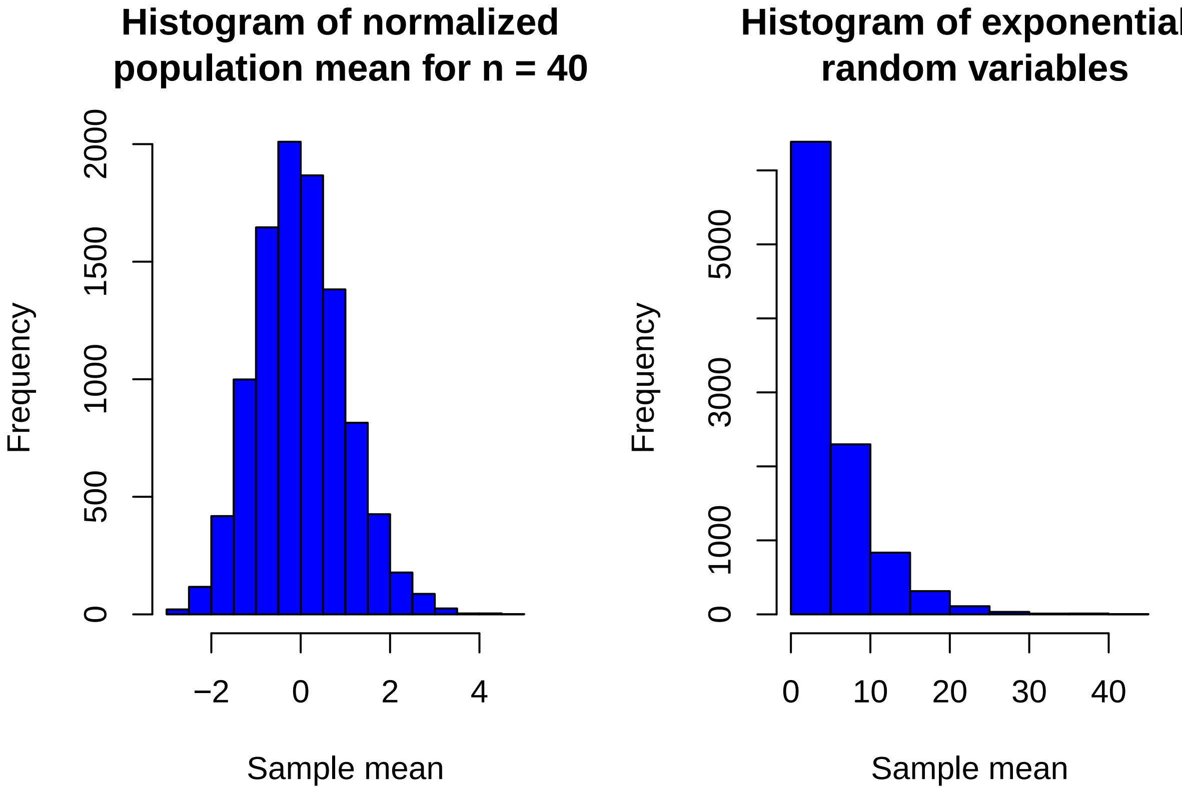
where $\bar{X}_n$ stands for the sample average $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, is approximately normally distributed as n gets large.

In order to check whether the sample mean is approximately normal, we rely on the central limit theorem and consider the quantity $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$.

```
nor <- (mn - 1/lambda) * sqrt(n) / (1 / lambda)

par(mfrow = c(1,2))
hist(
        x = nor,
        xlab = "Sample mean",
        main = "Histogram of normalized \n population mean for n = 40",
        col = "blue"
)

hist(
        x = rexp(n = N, rate = lambda),
        xlab = "Sample mean",
        main = "Histogram of exponential \n random variables",
        col = "blue"
)
```



Histogram of normalized population mean for n = 40
Frequency
0
500
1000
1500
2000
−2
0
2
4
Sample mean
Histogram of exponential random variables
Frequency
0
1000
3000
5000
0
10
20
30
40
Sample mean