Part 1: Simulation Exercise Instructions

Tomislav Kralj

26/08/2022

Overview

In this paper, we simulate samples from the exponential distribution and calculate their sample means. The analysis focuses on the distribution of these sample means and seeing the Central Limit Theorem at work.

Simulations

We simulate the samples from the $\text{Exp}(\lambda = 0.2)$ distribution. Each simulated sample has size $n = 40$ and we generate $m = 1000$ samples. Finally, for every sample, we calculate the corresponding sample mean. The code for the simulation is given below.

```
#Set the parameters
lambda = 0.2
n = 40
m = 1000

#Generate the samples
set.seed(26082022)
samples <- matrix(rexp(n*m, lambda), ncol = n, nrow = m)
means <- apply(samples, 1, mean)
```

Sample Mean versus Theoretical Mean

In this heading, we compare the sample mean versus the theoretical mean. If $X \sim \text{Exp}(\lambda = 0.2)$, we have that the theoretical mean of X is $E(X) = 1/\lambda$. The sample mean of a sample of $n = 40$ IID variables $X_{i1}, \ldots, X_{in}$ is equal to

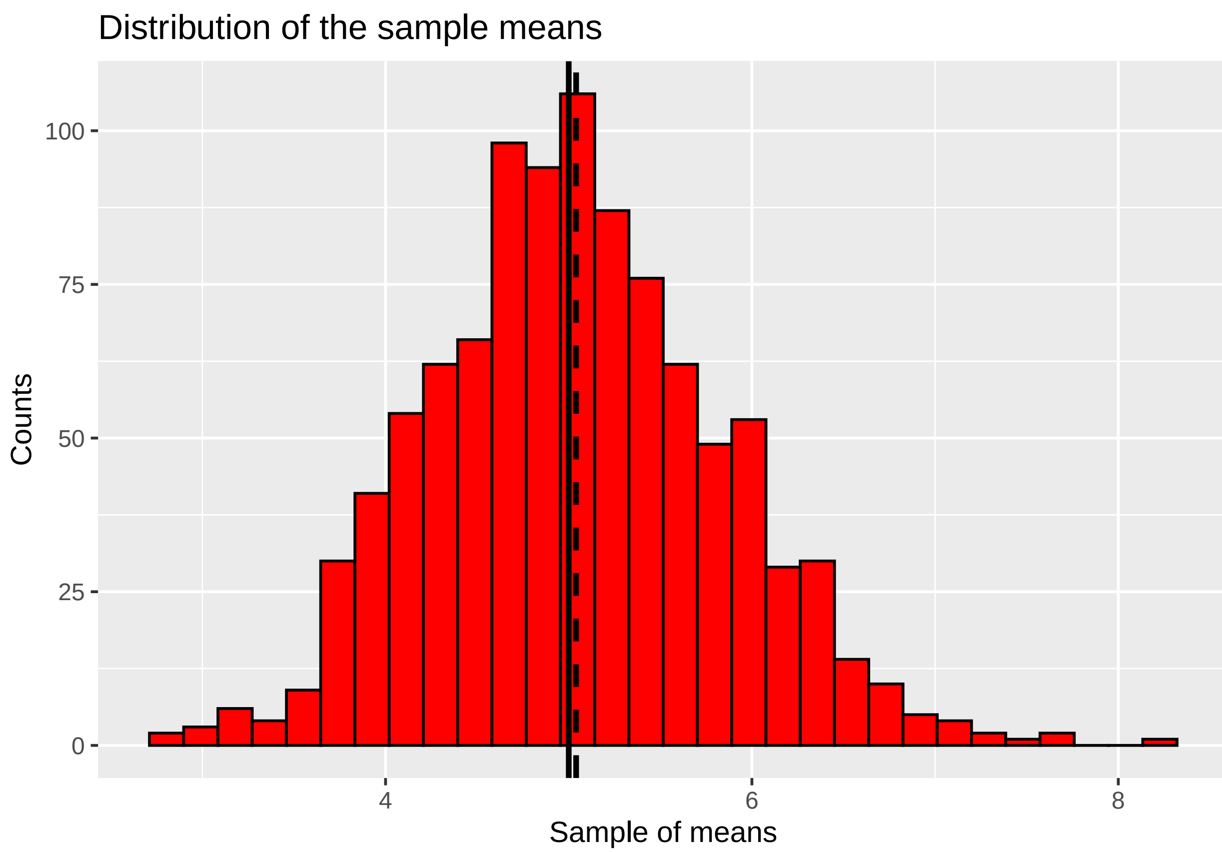
$$\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$$

which implies that the theoretical mean of the sample mean is $E(\bar{X}_i) = 1/\lambda$. Finally, since we have $m = 1000$ sample means, the grand sample mean is equal to

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{n} \bar{X}_i$$

where one can show once again that $E(\bar{X}) = 1/\lambda$. Let's compare the theoretical mean $1/\lambda = 5$ with the simulated value $\bar{X}$.

```r
ggplot(data.frame(means = means), aes(x = means)) +
  geom_histogram(color = "black", fill = "red") +
  xlab("Sample of means") +
  ylab("Counts") +
  ggtitle("Distribution of the sample means") +
  geom_vline(xintercept = 5, size = 1) +
  geom_vline(xintercept = mean(means), size = 1, linetype = "dashed")
```



The full line represents the theoretical mean (which is $1/\lambda = 5$), while the dashed line represents the sample mean. For the clearence, the simulated sample mean is equal to

```
## [1] 5.040496
```

These values are very close to each other, which was undoubtely expected. In the next heading, we're gonna see why.

Sample Variance versus Theoretical Variance

It is known that if $X \sim \text{Exp}(\lambda = 0.2)$, we have that the theoretical variance of X is $\text{Var}(X) = 1/\lambda^2$. The sample mean of a sample of $n = 40$ IID variables $X_{i1}, \ldots, X_{in}$ is equal to

$$\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$$

which implies that the theoretical variance of the sample mean is $\text{Var}(\bar{X}_i) = 1/(n\lambda^2)$. Finally, since we have $m = 1000$ sample means, the grand sample variance is equal to

$$s^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\bar{X}_i - \bar{X})^2$$

where one can show once again that $E(s^2) = 1/(n\lambda^2)$. Let's compare the theoretical variance $1/(n\lambda^2) = 1/(40 \cdot 0.2^2) = 0.625$ with the simulated value s^2.

```
#Theoretical variance
1/(n*lambda^2)
```

```
## [1] 0.625
```

```
#Observed sample variance
var(means)
```

```
## [1] 0.6226884
```

It can be seen that the theoretical variance is very close to the simulated one. Also, the variance is relatively small, so that's the reason why the sample mean is close to its theoretical value.

Distribution

Because of the Central Limit Theorem, we have that

$$\frac{\bar{X}_i - E(\bar{X}_i)}{\text{sd}(\bar{X}_i)}$$

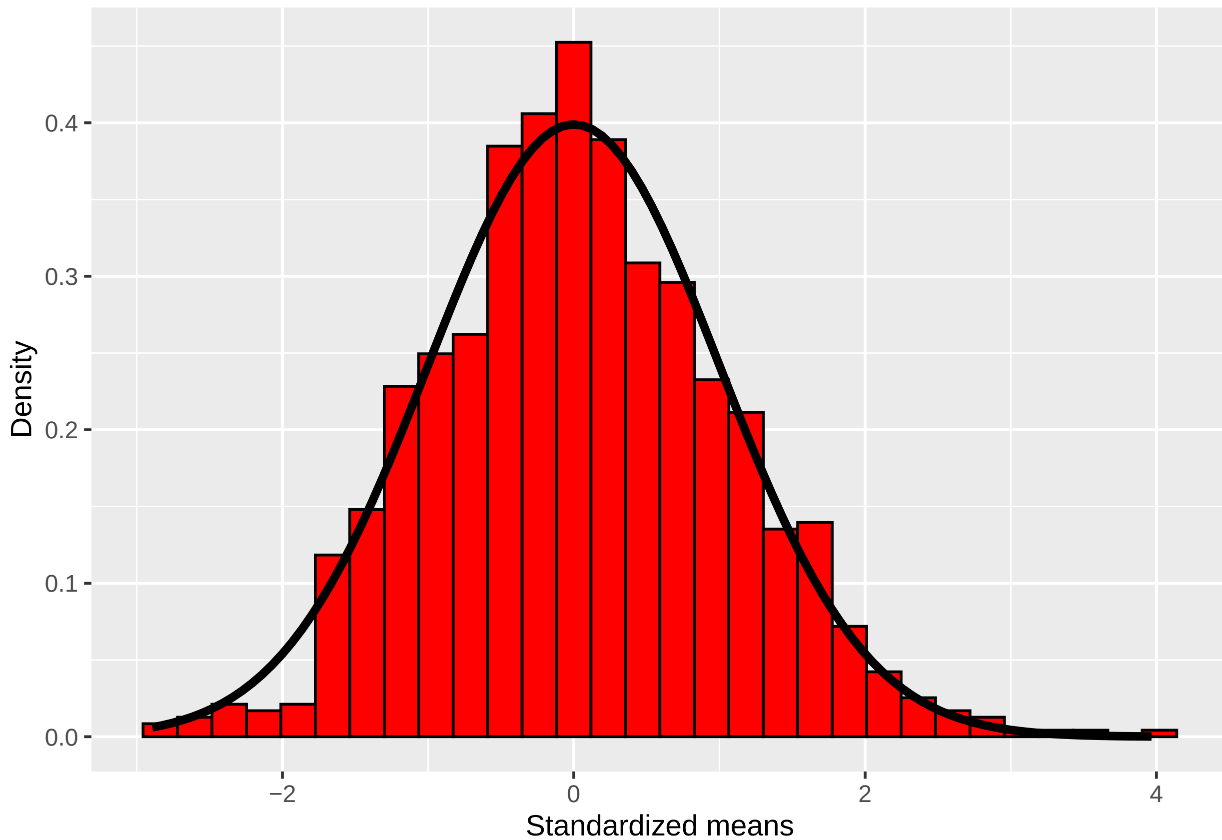
has approximately normal distribution. Since we have already showed that

$$E(\bar{X}_i) = 1/\lambda = 5, \quad \text{sd}(\bar{X}_i) = \frac{1}{\sqrt{n\lambda^2}} \approx 0.7906.$$

Let's show that graphically! We plot the histogram of the standardized variables $\frac{\bar{X}_i - E(\bar{X}_i)}{\text{sd}(\bar{X}_i)}$ and add the density function of the standard normal random variable.

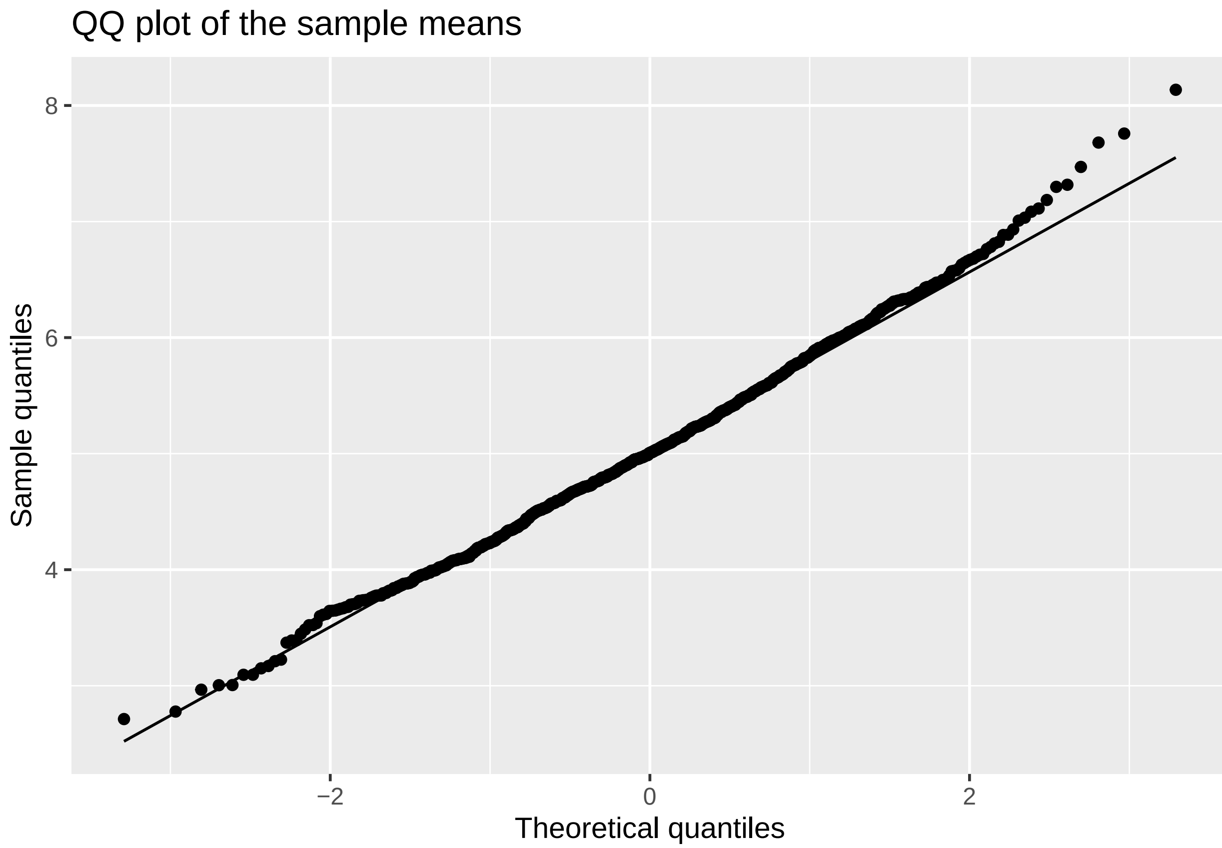
```
ggplot(data = data.frame(means = (means-5)/sqrt(0.625)), aes(x = means)) +
  geom_histogram(aes(y = ..density..), color = "black", fill = "red") +
  stat_function(fun = dnorm, size = 1.5) +
  xlab("Standardized means") +
  ylab("Density")
```

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



One can see that the density function of the standard normal random variable fits very well on the histogram. Alternatively, we could make a QQ plot for the sample means to check the normality. It is ploted below.

```
ggplot(data= data.frame(means)) +
  geom_qq(mapping = aes(sample = means)) +
  xlab("Theoretical quantiles") +
  ylab("Sample quantiles") +
  labs(title = "QQ plot of the sample means") +
  stat_qq_line(mapping = aes(sample = means))
```



We can see that the dots lie approximately on the line, so the data is approximately normal.