Peer-graded Assignment: Statistical Inference Course Project

Irina Z

9/22/2020

```
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
```

This report contains two parts of the course project:
1. A simulation exercise
2. Basic inferential data analysis

Part 1: Simulation Exercise

Synopsis

In this report, I investigate the **exponential distribution** in R and compare it with the **Central Limit Theorem**. The exponential distribution will be simulated in R with `rexp(n, lambda)` where lambda is the rate parameter. The mean of exponential distribution is $\frac{1}{\lambda}$, and the standard deviation is also $\frac{1}{\lambda}$. We set $\lambda = 0.2$ for all of the simulations and investigate the distribution of averages of 40 exponentials and do a thousand simulations.

Question 1 : Show the sample mean and compare it to the theoretical mean distribution

First, I run simulation:

```
#  Setting the seed for reproducability
set.seed(1956)
n <- 40
n_sims <- 1000
lambda <- 0.2

## Simulate the sample
SampleMean <- NULL
for(i in 1:n_sims) {
  SampleMean <- c(SampleMean, mean(rexp(n, lambda)))
}
```

The theoretical mean distribution is $E[X] = \frac{1}{\lambda} = \frac{1}{0.2} = 5$, and the sample mean is $\overline{X} = 4.998634$, which is **close to the theoretical mean distribution**.
Difference between the values is:

```
abs((1/lambda) - mean(SampleMean))
```

```
## [1] 0.001365977
```

We can also show how close the sample mean and the theoretical mean are on the plot which is shown in the Appendix (Picture 1).

Question 2 : Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution

The sample variance is:

```
SampleVar <- var(SampleMean)
```

Which is equal to 0.6193824.

The variance of the sample mean is $var(\overline{X}) = \frac{\sigma^2}{n}$, where σ is the standard deviation equal to $\sigma = \frac{1}{\lambda} = \frac{1}{0.2} = 5$. Therefore, the theoretical variance is $var(\overline{X}) = \frac{5^2}{40} = \frac{25}{40} = 0.625$.

Difference between the values is:

```
abs((1/lambda)^2/n - var(SampleMean))
```

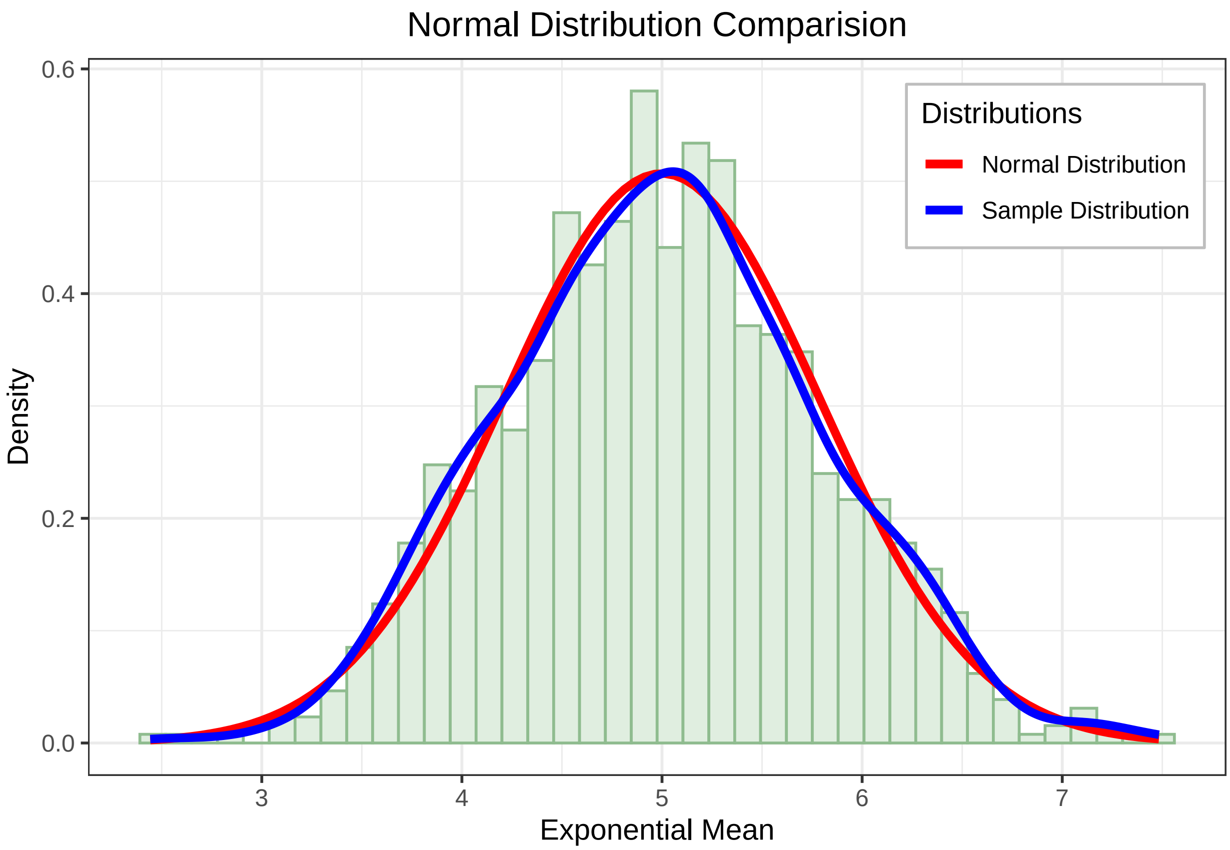
```
## [1] 0.005617598
```

This difference is also small, so **the variance of the sample mean and the theoretical variance are close**.

Question 3 : Show that the distribution is approximately normal

To show that the distribution is approximately normal, I plot histogram with density function:

```
ggplot(data.frame(SampleMean), aes(x = SampleMean)) +
        geom_histogram(bins = 40, color = "darkseagreen", fill = "honeydew2",
                       aes(y = ..density..))+
            stat_function(fun = dnorm, args = list(mean = mean(SampleMean),
                                                    sd = sd(SampleMean)),
                          aes(color = "Normal Distribution"), size = 1.5) +
            stat_density(geom = "line", aes(color = "Sample Distribution"), size = 1.5)  +
            labs(x = "Exponential Mean", y = "Density",
                 title = "Normal Distribution Comparision") +
        scale_color_manual(name = "Distributions",
                             values = c("Normal Distribution" = "red",
                                        "Sample Distribution" = "blue")) +
        theme_bw()+
        theme(legend.position = c(0.85, 0.85),
              legend.background = element_rect(colour = 'grey'),
              plot.title = element_text(hjust = 0.5))
```



This plot indicates that density curve of our sample (*blue line*) is similar to normal distribution curve (*red line*). Therefore, **the distribution is approximately normal**.

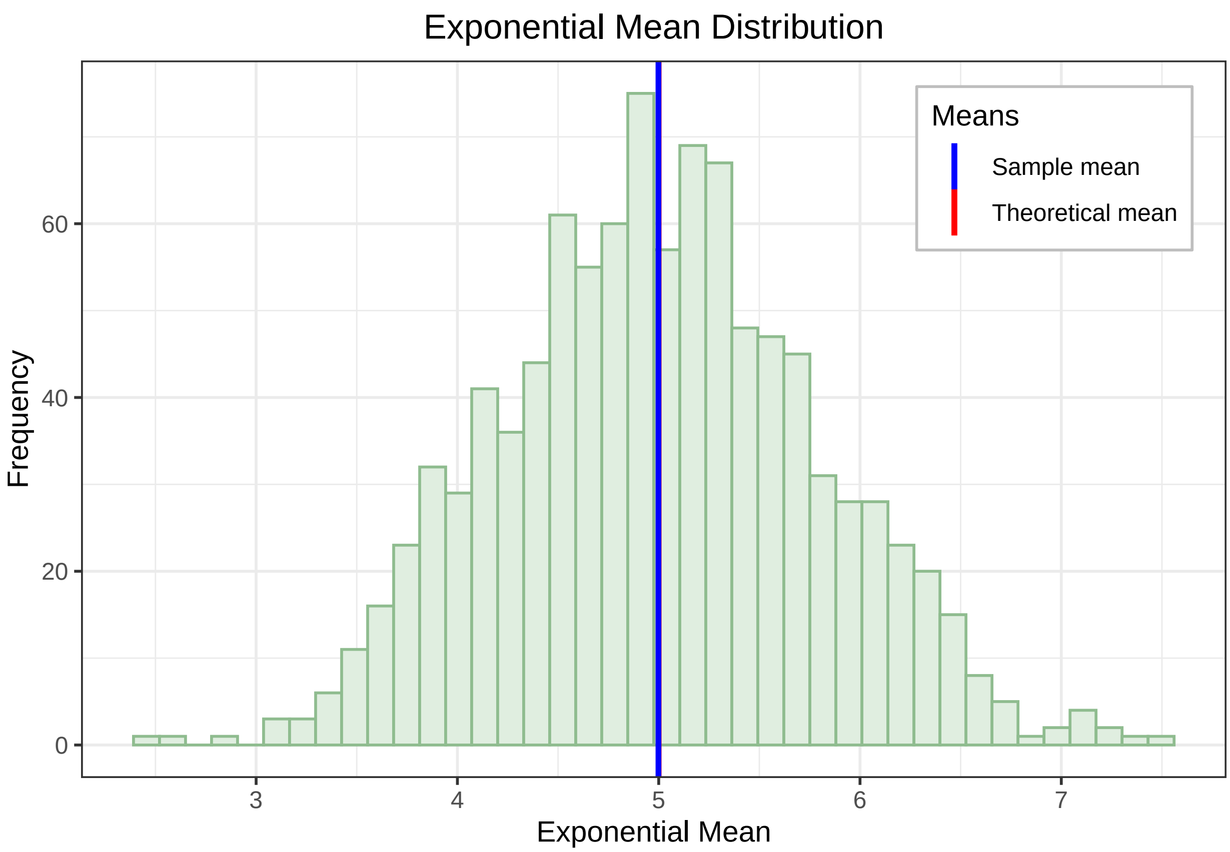
We can also use a **quantile-quantile (Q-Q) plot** of sample mean against a normal distribution. If the data points follow the linearity represented by the normal distribution, then we can say that the data have normal distribution.

The Q-Q Normal Plot is shown in the Appendix (Picture 2). This plot also **indicates the normal distribution of the exponential data**.

Appendix to Part 1

Picture 1. Theoretical and sample means:

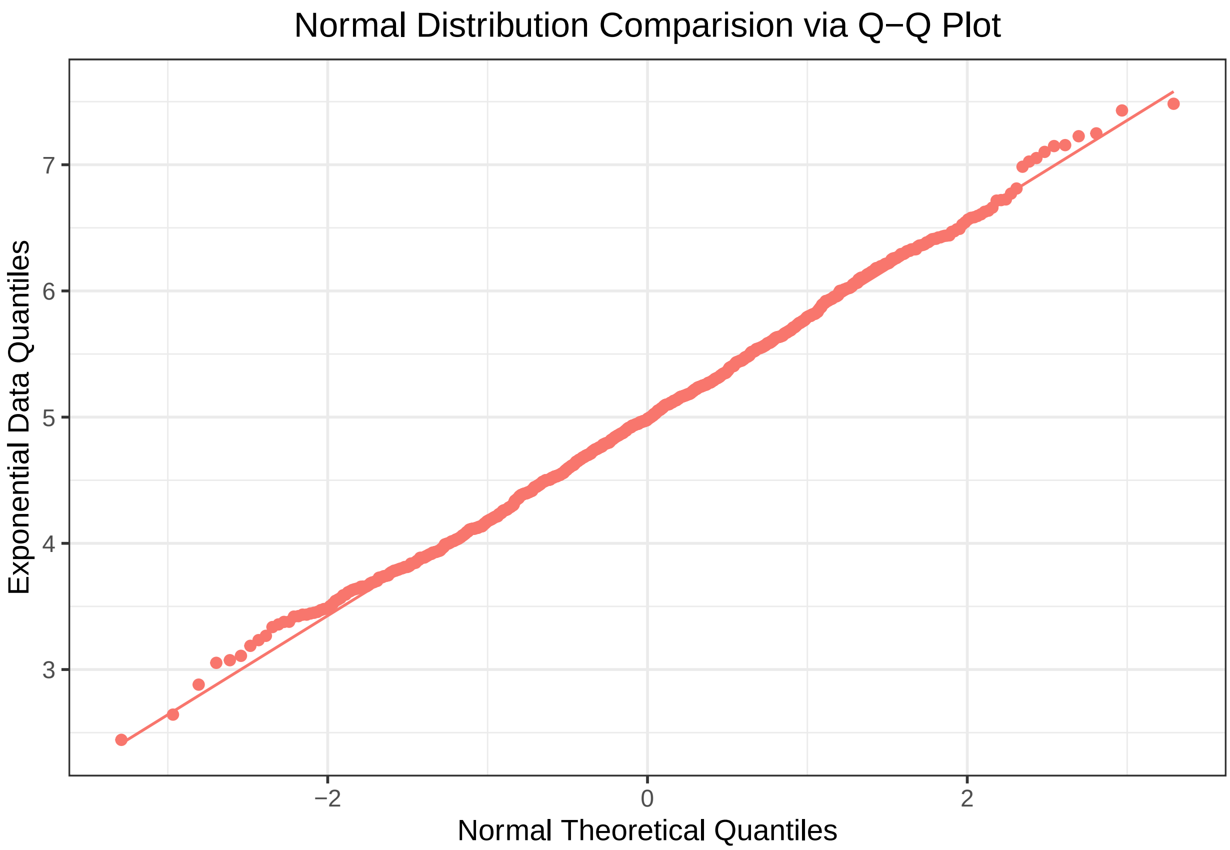
```
ggplot(mapping = aes(SampleMean)) +
        geom_histogram(bins = 40, color = "darkseagreen", fill = "honeydew2") +
        labs(x = "Exponential Mean", y = "Frequency",
             title = "Exponential Mean Distribution") +
        geom_vline(aes(xintercept = 5, color = "Theoretical mean"), size = 1) +
        geom_vline(aes(xintercept = mean(SampleMean), color = "Sample mean"), size = 1) +
        scale_color_manual(name = "Means",
                             values = c("Theoretical mean" = "red",
                                        "Sample mean" = "blue")) +
        theme_bw()+
        theme(legend.position = c(0.85, 0.85),
              legend.background = element_rect(colour = 'grey'),
              plot.title = element_text(hjust = 0.5))
```



Theoretical mean is shown by the *red line*, while the sample mean is shown by the *blue line*. We can see that the two lines are close, therefore, **the values of the sample mean and the theoretical mean are close**.

Picture 2. Comparison to normal distribution via Q-Q plot:

```
ggplot(data.frame(SampleMean), aes(sample = SampleMean, col = "darkseagreen")) +
        geom_qq() +
        stat_qq_line() +
        labs(x = "Normal Theoretical Quantiles", y = "Exponential Data Quantiles",
                 title = "Normal Distribution Comparision via Q-Q Plot") +
        theme_bw() +
        theme(legend.position = "none",
              plot.title = element_text(hjust = 0.5))
```



The data points on the Q-Q plot have outliers in the bottom left and upper right corners, but otherwise the linearity of the points suggests that **the data are approximately normally distributed**.