Statistical Inference Week4 Programming Assignment (1)

Guang Yang

November 17, 2016

Statistical Inference Week4 Programming Assignment (1)

Overview

This report addresses the questions from Week 4 Assignment of **Statistical Inference**, the Course #6 of the Data Science Specialization series, offered by Coursera.org. The report is mainly consist of 2 parts, aiming to discuss the results of the two topics:

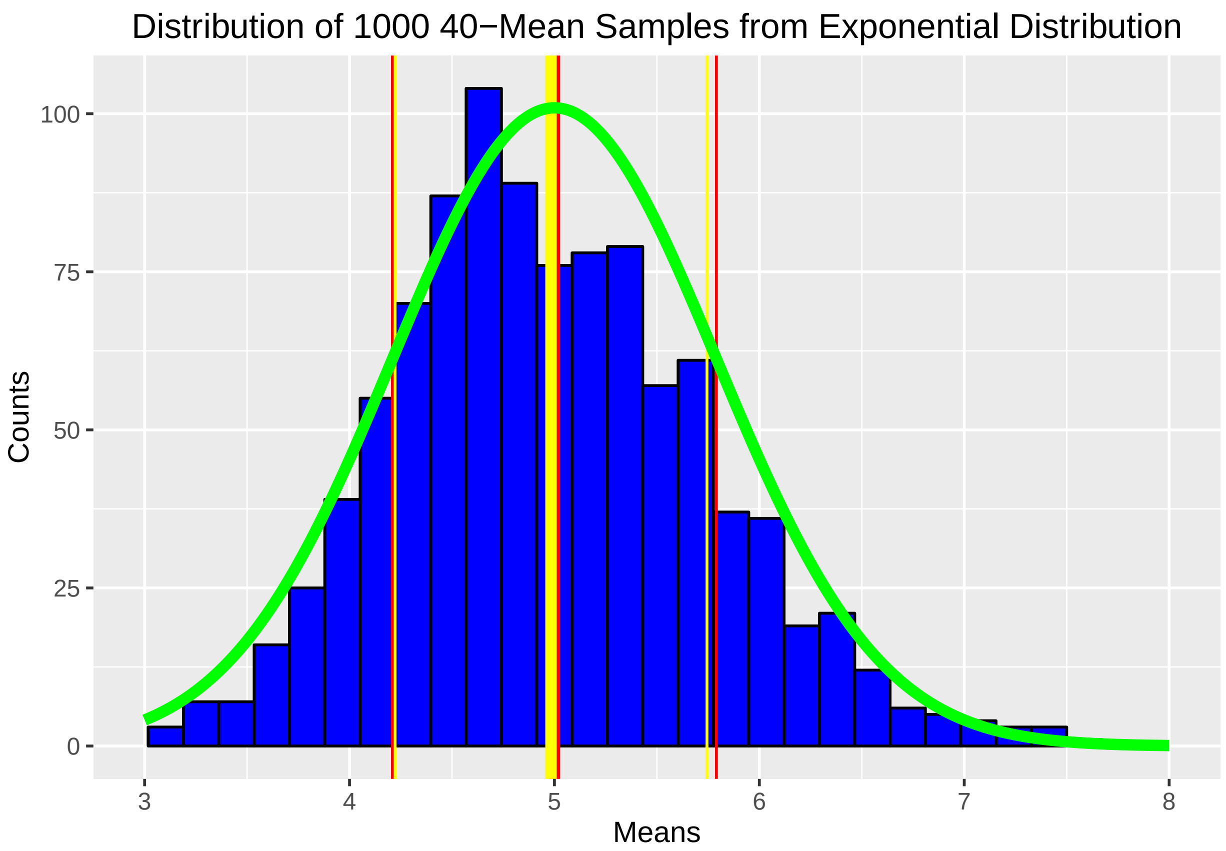
- **Testing CLT with simulation from exponential distribution**;
- Performing inferential analysis on the ToothGrowth dataset.

Case Study 1: Simulation of CLT Based on Exponential Equation

In this project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem (CLT). The exponential distribution can be simulated in R with `rexp(n, lambda)` where `lambda` is the rate parameter. The mean of exponential distribution is `1/lambda` and the standard deviation is also `1/lambda`. Set `lambda = 0.2` for all of the simulations. We will investigate the distribution of averages of 40 exponentials. Note that we will do **1000** simulations.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials.

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```



```
## [1] "Theoretical variance: 0.625"
```

```
## [1] "Sample variance: 0.579"
```

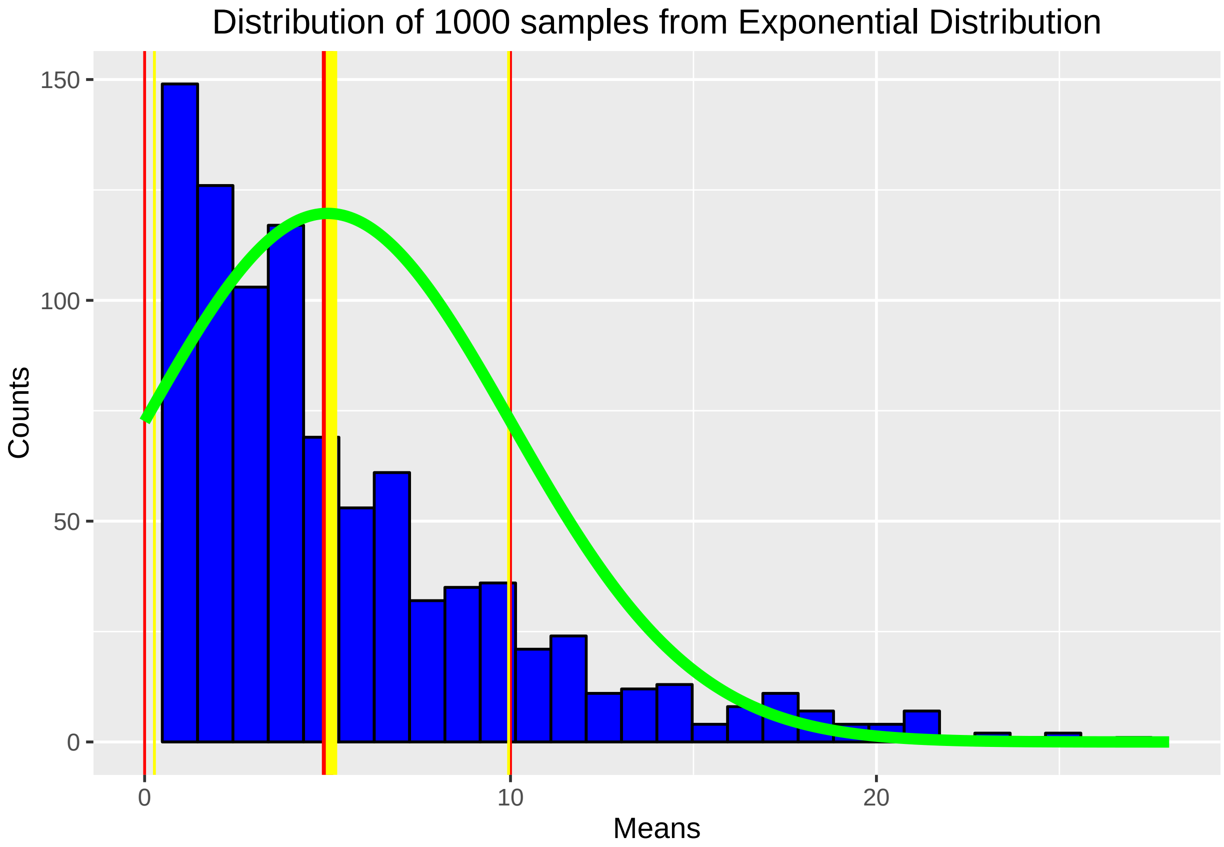
From the above plot we firstly see the **distribution of the sample means** (blue); the top of which stands somewhere between 4 and 5. The above plot:

1. Shows the **sample mean**, `mean(means)` **(yellow)** and compare it to the **theoretical mean**, `1/lambda` **(red)** of the distribution.
2. Shows how variable the sample is and compare it to the theoretical variance of the distribution.

```
+ By variance: the variance of sample means is 0.579, and that of theoretical means is 0.625.
+ By standard deviation: the standard deviations in the plot are marked in orange (sample) and red
```

3. Shows that the distribution is approximately normal by overlaying the **gaussian PDF** (`mean = 1/lambda, sd = 1/(lambda*sqrt(n))`, green curve) with the histogram.

Just to make a comparison, we then quickly plot another set of 1000 samples from distribution. *Note this time there's only 1 sample each, instead of mean of 40.* Let's take a look at it.



```
## [1] "Theoretical variance: 25"
```

```
## [1] "Sample variance: 23.425"
```

Now as we see in the above, the distribution is way much more like exponential, rather than normal, even if the variances are similar (25 of theoretical and 23.425 of sample). In comparison, the 40-mean distribution is very normal-like.

Appendix: Codes

Codes for Plotting "Distribution of 1000 40-Mean Samples from Exponential Distribution"

```
n <- 40                    # Sample Size
times <- 1000              # Repeat Time
lambda <- 0.2              # Rate Constant

# Use set.seed() for reproducible purpose.
set.seed(7)
means <- replicate(times, mean(rexp(n, lambda)))

# Prepare numeric sequence to plot normal curve.
LIM <- c(floor(min(means)), ceiling(max(means)))
x <- seq(floor(min(means)), ceiling(max(means)), length = 1000)
nd <- dnorm(x, mean = 1/lambda, sd = 1/(lambda*sqrt(n)))

# Make plot.
```

```r
library(ggplot2)
ggplot(as.data.frame(means), aes(x = means)) +
        geom_histogram(bins = 30, col = 'black', fill = 'blue') +
        lims(x = LIM) +
        # Theoretical Mean
        geom_vline(xintercept = 1/lambda, col = 'red', lwd = 2) +
        # Sample Mean
        geom_vline(xintercept = mean(means), col = 'yellow', lwd = 2) +
        # Theoretical SD
        geom_vline(xintercept = 1/lambda + c(-1, 1) * 1/(lambda*sqrt(n)), col = 'red') +
        # Sample SD
        geom_vline(xintercept = mean(means) + c(-1, 1) * sd(means), col = 'yellow') +
        # Gaussian PDF
        geom_line(aes(x = x, y = nd*times/5), col = 'green', lwd = 2) +
        labs(x = 'Means', y = 'Counts',
             title = 'Distribution of 1000 40-Mean Samples from Exponential Distribution')

svar <- round(var(means), 3)
tvar <- round(1/lambda^2/n, 3)
print(paste0('Theoretical variance: ', tvar))
print(paste0('Sample variance: ', svar))
```

Codes for Plotting "Distribution of 1000 samples from Exponential Distribution"

```r
times <- 1000           # Repeat Time
lambda <- 0.2           # Rate Constant

# Use set.seed() for reproducible purpose.
set.seed(7)
means2 <- replicate(times, rexp(1, lambda))

# Prepare numeric sequence to plot normal curve.
LIM2 <- c(floor(min(means2)), ceiling(max(means2)))
x2 <- seq(floor(min(means2)), ceiling(max(means2)), length = 1000)
nd2 <- dnorm(x2, mean = 1/lambda, sd = 1/lambda)

# Make plot.
library(ggplot2)
ggplot(as.data.frame(means2), aes(x = means2)) +
        geom_histogram(bins = 30, col = 'black', fill = 'blue') +
        lims(x = LIM2) +
        # Theoretical Mean
        geom_vline(xintercept = 1/lambda, col = 'red', lwd = 2) +
        # Sample Mean
        geom_vline(xintercept = mean(means2), col = 'yellow', lwd = 2) +
        # Theoretical SD
        geom_vline(xintercept = 1/lambda + c(-1, 1) * 1/(lambda), col = 'red') +
        # Sample SD
        geom_vline(xintercept = mean(means2) + c(-1, 1) * sd(means2), col = 'yellow') +
        # Gaussian PDF
        geom_line(aes(x = x2, y = nd2*times*1.5), col = 'green', lwd = 2) +
        labs(x = 'Means', y = 'Counts',
             title = 'Distribution of 1000 samples from Exponential Distribution')

svar2 <- round(var(means2), 3)
tvar2 <- round(1/lambda^2, 3)
print(paste0('Theoretical variance: ', tvar2))
print(paste0('Sample variance: ', svar2))
```