Stats_Inference_Course_Project

LI CHENYANG

December 25, 2015

1. Overview: In a few (2-3) sentences explain what is going to be reported on.

- In this section I will investigate the exponential distribution in R and compare it with the Central Limit Theorem

2. Simulations: Include English explanations of the simulations you ran, with the accompanying R code. Your explanations should make clear what the R code accomplishes.

(Majority of the codes referred from Prof Brian Caffo's lecture notes. Therefore all credit goes to him.)

The goal is to demonstrate *Central Limit Therom* with random variables generated by exponential distribution

- First, create a group of 40 random variables x 1000 random variable, with each row represent one observation of the 40 variables.

```
nosim <-  1000
All_observ <- matrix(rexp(40000, rate = 0.2), nosim)
```

- Next creat a function to calculate and normalize all the means from one row of observation

```
cfunc <- function(x, n) sqrt(n) * (mean(x) - 5) / 5
```

- Next create a dataframe by applying function over each row in the All_observ matrix

```
n_size = 40
final_form <- data.frame(samples = apply(X = All_observ, MARGIN = 1, FUN = cfunc, n_size))
```

3. Sample Mean versus Theoretical Mean: Include figures with titles. In the figures, highlight the means you are comparing. Include text that explains the figures and what is shown on them, and provides appropriate numbers.

- According to central limit therom, the theoretical mean for the distribution of sample mean, after proper normalization, should be 0.
- Similarly, the theoretical variance after normalization, should be 1.
- Below is the plot of the distribution of mean of 40 exponential variable.

Distribution of sample mean
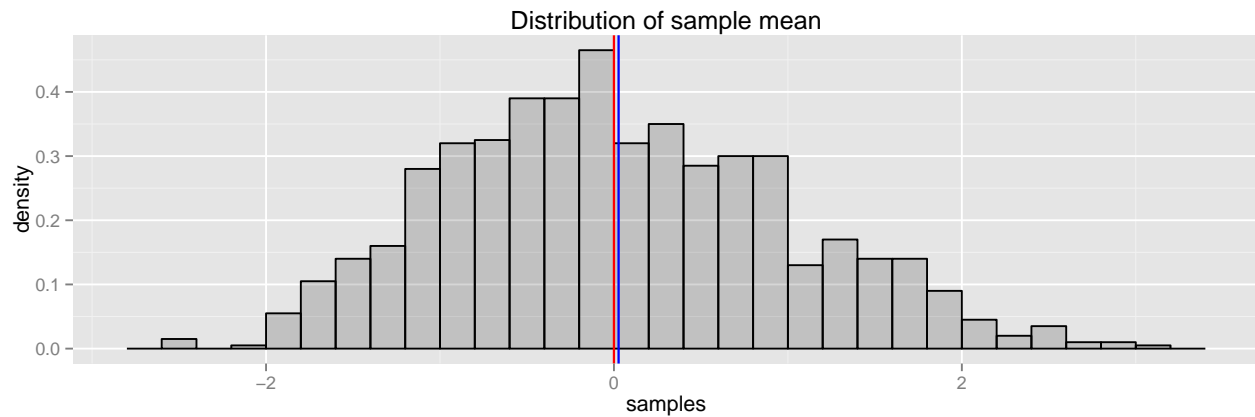
- From the graph, there are two thin lines (my apology it is hard to see, I do not want the lines to overlap).
- One line is red and the other is blue. The red line represents theoretic mean which is 0, and the blue line represent sample mean for the distribution. We can calculate the sample mean and find out that it is in fact

```r
mean(final_form$samples)
```

```
## [1] 0.02668893
```

which is fairly close to the theoretic mean.

- By observation, we can see that the sample distribution is centered comfortably around the population mean.
- Next, we look at sample variance to further validate central limit theorem.

4. Sample Variance versus Theoretical Variance: Include figures (output from R) with titles. Highlight the variances you are comparing. Include text that explains your understanding of the differences of the variances.

- Variability of the graph will affect the width of the graph.
- The higher the variance, the narrower the graph will be.
- Therefore, we can draw in this case a standard normal distribution and try to compare the width with that of the sample distribution

sample distribution & standard normal distribution
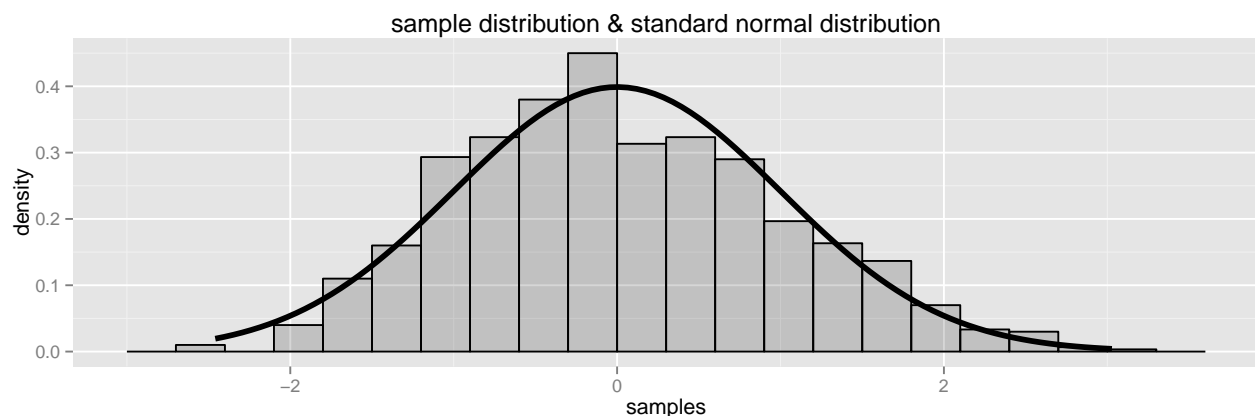
- The black curve represent standard normal distribution.
- As we can see there is almost no difference in width between these two distributions.
- To further prove the point, we calculat the sample variance.

```
var(final_form$samples)
```

```
## [1] 0.9555701
```

- the sample variance is very very close to 1.
- Therefore we can safely conclude that in this case, Central Limit Theorem is validated.

5. Distribution: Via figures and text, explain how one can tell the distribution is approximately normal.

- We can simply tell that the distribution is roughly normal, just by looking at the plot below
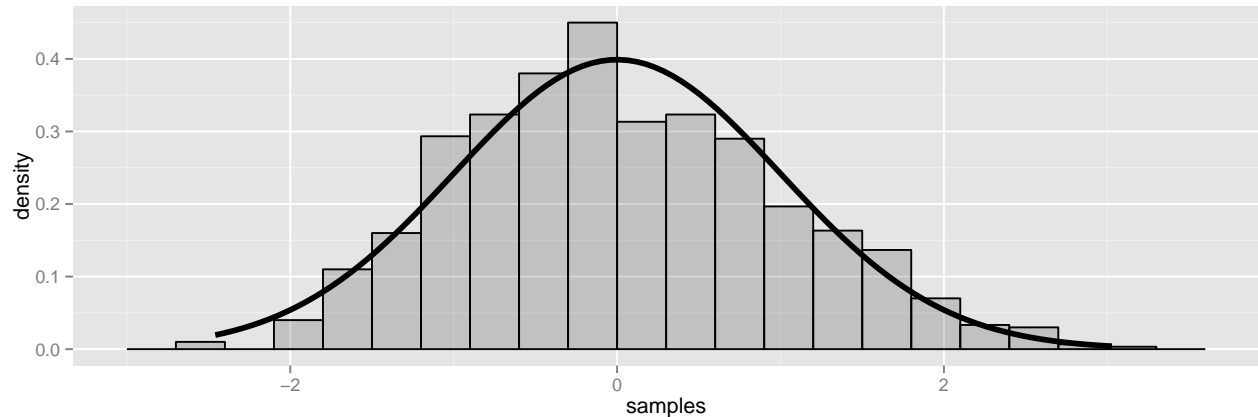
- To show how different it would be if it is just random expontial distribution samples, without taking averages. We construct the following plot:

Histogram of rexp(1000)
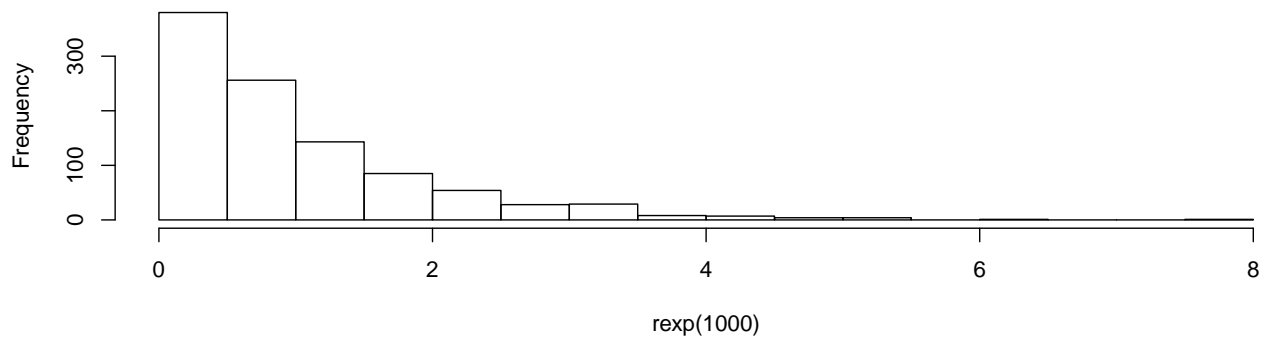
- We can tell by the graph above that the plot is no way gaussian.