# Simulation exercise

*kwatanwa17*

*4/9/2018*

## Overview

This course project consists of two parts. The first part corresponds to the comparison of the simulation sample mean to the theorical mean of exponential distribution. In the second part, we conducted t test using ToothGroth data set.

## Part 1: Simulation Exercise

### Simulation

We simulated 1000 times 40 samples of exponential distribution with lambda = 40 and averaged it by each simulation. Likewise, we took 40 samples from the theorical distribution 1000 times.

Note that the mean of exponential distribution is 1/lambda and its standard deviation is also 1/lambda. So, in our case, the mean and standard deviation is 1/0.2 = 5.

```
n <- 40
lambda <- 0.2
nosim <- 1000

set.seed(1)
Simulation <- replicate(nosim, rexp(n, lambda))
Simulation_means <- apply(Simulation, 2, mean)
Simulation_vars <- apply(Simulation, 2, var)

set.seed(1)
Theorical <- replicate(nosim, rnorm(40, mean = 1/0.2, sd = 1/0.2))
Theorical_means <- apply(Theorical, 2, mean)
Theorical_vars <- apply(Theorical, 2, var)

dat <- data.frame(
      Mean = c(Simulation_means, Theorical_means),
      Variance = c(Simulation_vars, Theorical_vars),
      Type = rep(c("Simulation", "Theorical"), each = nosim)
    )
```

### Sample Mean versus Theoretical Mean

Both distributions are very closed

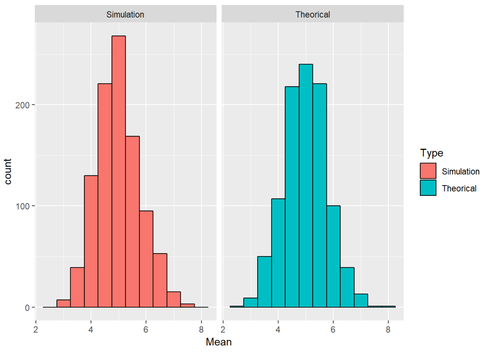
```
summary(Simulation_means)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.108   4.445   4.950   4.990   5.492   7.491
```

```
summary(Theorical_means)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.637   4.456   4.993   4.989   5.532   7.823
```

```
library(ggplot2)
g <- ggplot(dat, aes(x = Mean, fill = Type)) +
  geom_histogram(binwidth = 0.5, color = "black") +
  facet_wrap(~Type, ncol = 2)
g
```




### Sample Variance versus Theoretical Variance

The distribution of simulation variances are more skewed than the theorical one.

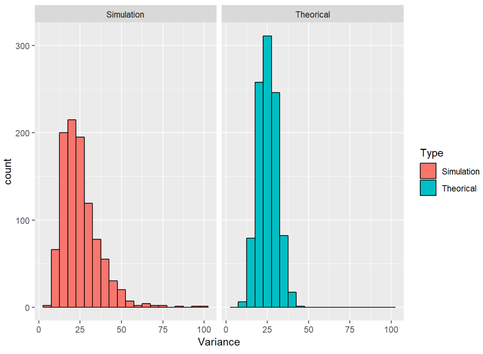
```
summary(Simulation_vars)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.153  16.912  22.739  25.065  30.465  99.828
```

```
summary(Theorical_vars)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.79   21.08   24.90   25.14   29.01   46.90
```

```
g2 <- ggplot(dat, aes(x = Variance, fill = Type)) +
  geom_histogram(binwidth = 5,  color = "black") +
  facet_wrap(~Type, ncol = 2)
g2
```




### Approximately normal

The red curve represents the density of normal distribution with the mean and standard deviation of simulation means. Except the distribution of simulation variances, all the distributions seem to be distributed normally.

```
scaled_dat <- as.data.frame(apply(
  data.frame(Simulation_means, Theorical_means, Simulation_vars, Theorical_vars),
  2, scale))

library(tidyr)
scaled_dat_updated <- tidyr::gather(scaled_dat, key = Distribution_type, value = z_score)

g5 <- ggplot(scaled_dat_updated, aes(x = z_score, fill =Distribution_type)) +
  geom_histogram(aes(y = ..density..), color = "black") +
  facet_wrap(~Distribution_type, ncol = 2, nrow = 2) +
  stat_function(fun = dnorm, color = "brown", args = list(mean = 0))
g5
```

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