Final project: Statistical Inference

Kwatanwa17

4/9/2018

Overview

This report 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

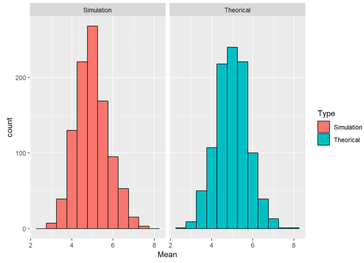
```
summary(Simulation_means)
```

```
##    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.

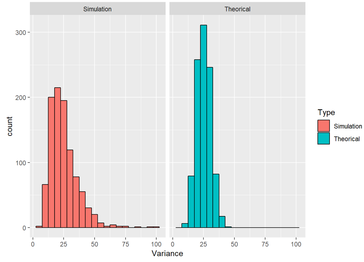
```
summary(Simulation_vars)
```

```
##    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

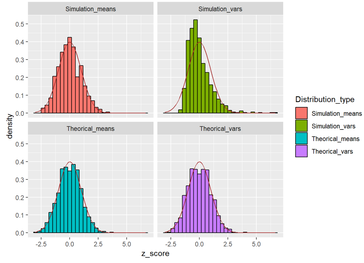
The red curve represents the density of normal distribution with the mean and standard deviation of simulation means. Except the distribution of simulation variances, all the distributions seem to be distributed normally.

```
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`.
```



Part 2: Basic Inferential Data Analysis

Load the data set.

```
library(datasets)
data("ToothGrowth")
```

Summary of the data

Notice that the variable dose is the quantity of dose which ranges from 0.5 to 2. There is the same number of subjects with respect to the supp variable (OJ/VC).

```
summary(ToothGrowth)
```

```
##       len        supp         dose
##  Min.   : 4.20   OJ:30   Min.   :0.500
##  1st Qu.:13.07   VC:30   1st Qu.:0.500
##  Median :19.25           Median :1.000
##  Mean   :18.81           Mean   :1.167
##  3rd Qu.:25.27           3rd Qu.:2.000
##  Max.   :33.90           Max.   :2.000
```

T test two sided test

```
VC <- subset(ToothGrowth, supp == "VC")
OJ <- subset(ToothGrowth, supp == "OJ")
```

```
t.test(VC$dose, OJ$dose, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  VC$dose and OJ$dose
## t = 0, df = 58, p-value = 1
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3278171  0.3278171
## sample estimates:
## mean of x mean of y
##  1.166667  1.166667
```

```
t.test(VC$len, OJ$len, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  VC$len and OJ$len
## t = -1.9153, df = 55.309, p-value = 0.06063
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.5710156  0.1710156
## sample estimates:
## mean of x mean of y
##  16.96333  20.66333
```

As a result of the analysis of the experiment data, there is no difference of tooth length by the supplement type (orange juice vs ascorbic acid). On the other hand, it is quite clear that there would not be difference by levels of Vitamin C (variable dose).For more information, please see R help page ("The Effect of Vitamin C on Tooth Growth in Guinea Pigs"; http://www.is.titech.ac.jp/~mase/mase/html.jp/temp/ToothGrowth.jp.html).