Biostatistics Homework 6

By 𝕃umi (张鹿鸣 12112618)

1. Gas Mileage

1.1) C

1.2) B

2. Old and New Machines

2.1)

H_0 : The new machine is not faster than the old machine on average ($\mu_{\text{new}} \geq \mu_{\text{old}}$).

H_1 : The new machine is faster than the old machine on average ($\mu_{\text{new}} < \mu_{\text{old}}$).

2.2)

It means that there is a 0.05 chance to reject the null hypothesis when the null hypothesis is true.

2.3)

The q2_table.txt:

Oldmachine 42.7 43.8 42.5 43.1 44.0 43.6 43.3 43.5 41.7 44.1

Newmachine 42.1 41.3 42.4 43.2 41.8 41.0 41.8 42.8 42.3 42.7

```r
machines <- as.data.frame(t(read.table('q2_table.txt', sep = '\t', row.names = 1)))
mean_old <- mean(machines$Oldmachine)
mean_new <- mean(machines$Newmachine)
var_old <- var(machines$Oldmachine)
var_new <- var(machines$Newmachine)
statistics <- data.frame(
  mean = c(mean_old, mean_new),
  variance = c(var_old, var_new)
)
rownames(statistics) <- c('Old', 'New')
statistics
```

```
##      mean  variance
## Old 43.23 0.5623333
## New 42.14 0.4671111
```

So the statistics are:

$$\bar{x}_{\text{new}} = 42.14, s^2_{\text{new}} \approx 0.5623$$

$$\bar{x}_{\text{old}} = 43.23, s^2_{\text{old}} \approx 0.4671$$

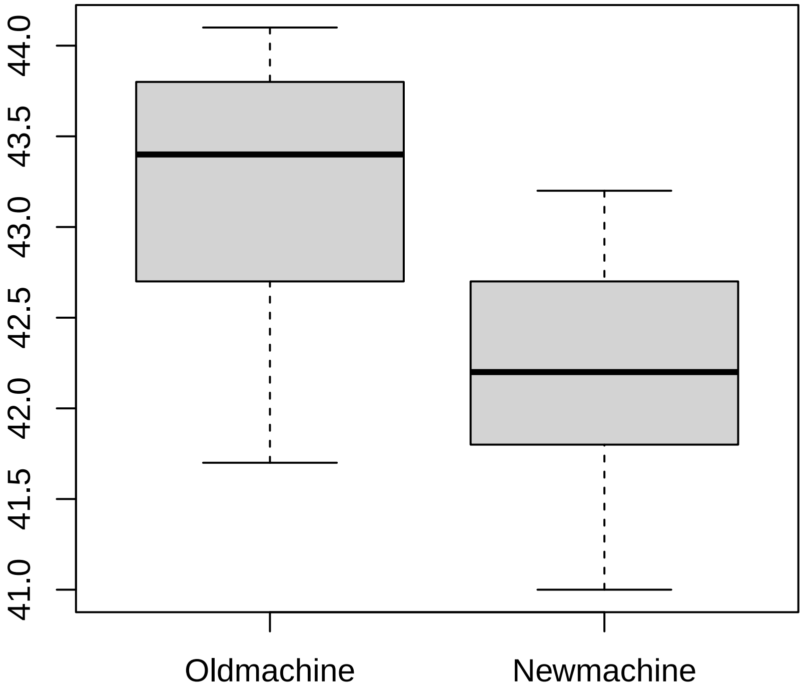
2.4)

We should use an independent two sample T-test. Since we have independent samples that are not paired.

We should check our samples are randomly and independently selected.

It is also important that the samples follow T-distributions. Or that the population follows a normal distribution.

```
boxplot(machines)
```



From the plot, I think we can say that our sample roughly follows a T-distribution, so using a T-test is reasonable.

We should also make sure that the variance between the two samples don't vary too much. In our samples, it doesn't.

2.5)

So the common variance is :

$$s^2 = \frac{(n_{\text{new}}-1)s^2_{\text{new}} + (n_{\text{old}}-1)s^2_{\text{old}}}{n_{\text{new}}+n_{\text{old}}-2} = \frac{9 \times 0.5623 + 9 \times 0.4671}{18} = 0.5147$$

The T-score is:

$$t = \frac{\bar{x}_{\text{new}} - \bar{x}_{\text{old}}}{\sqrt{s^2(\frac{1}{n_{\text{new}}} + \frac{1}{n_{\text{old}}})}} = \frac{42.14 - 43.23}{\sqrt{0.5147 \times (\frac{1}{10} + \frac{1}{10})}} \approx -3.397$$

The p-value is:

```
pt(-3.397, 18)
```

```
## [1] 0.001606392
```

```
t.test(machines$Newmachine, machines$Oldmachine, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  machines$Newmachine and machines$Oldmachine
## t = -3.3972, df = 17.847, p-value = 0.001621
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.5333684
## sample estimates:
## mean of x mean of y
##     42.14     43.23
```

From the `t.test()` function in R, we can see that our answers are quite right.

2.6)

D

(I don't know if B is correct or not, we reject the null hypothesis if the significance level is 0.05, but it doesn't mean the null hypothesis is wrong.)

3. ANOVA

C

4. Another ANOVA

B

(Since the variance of the data is mainly contributed by the variance within the groups (MSW).)

5. Another ANOVA

C

6. ANOVA and T-test

B

(Using multiple T-tests increase the chance of making a Type I error in your study, I don't think it is very good to say that it increases the probability of making a type I error)

7. post-hoc ANOVA

A (or between multiple groups)

8. Fisheries

8.1)

H_0: The mean weights of fish caught from the three lakes are not different.

H_1: The mean weights of fish caught from the three lakes are different.

8.2)

```
anova <- data.frame(d.f. = c(2,9,11),
                    SS = c(17.04, 14.19, 31.23),
                    MS = c(17.04/2, 14.19/9, NA))
row.names(anova) <- c("Between", "Within", "Total")
anova
```

```
##          d.f.    SS       MS
## Between     2 17.04 8.520000
## Within      9 14.19 1.576667
## Total      11 31.23       NA
```

So the F-score is:

$$F = \frac{\text{MSB}}{\text{MSW}} = 8.52 \div 1.5767 = 5.404$$

So the p-value is:

```
pf(5.404, 2, 9, lower.tail = F)
```

```
## [1] 0.0287282
```

8.3)

The p-value is 0.0287, so we should reject the null hypothesis.

9. Tar in Cigarettes

```
tar <- data.frame(
  BrandA = c(10.21,10.25,10.24,9.80,9.77,9.73),
  BrandB = c(11.32,11.20,11.40,10.50,10.68,10.90),
  BrandC = c(11.60,11.90,11.80,12.30,12.20,12.20)
)
```

9.1)

H_0: The tar contents for the three different brands of cigarettes are not different.

H_1: The tar contents for the three different brands of cigarettes are different.

9.2)

```
# We need to melt the data to perform anova test
library(reshape2)
molten_tar <- melt(tar)
```

```
## No id variables; using all as measure variables
```

```
res <- aov(value~variable, data = molten_tar)
summary(res)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## variable     2 12.000   6.000   65.46 3.89e-08 ***
## Residuals   15  1.375   0.092                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from the result summary that we have a F-value of 65.46 and a p-value of 3.98×10^{-8}. Thus we should reject the null hypothesis.

9.3)

```
mean_a <- mean(tar$BrandA)
mean_b <- mean(tar$BrandB)
mean_c <- mean(tar$BrandC)
s2_a <- var(tar$BrandA)
s2_b <- var(tar$BrandB)
```

```
s2_c <- var(tar$BrandC)
s2_ab <- (5 * s2_a + 5 * s2_b)/10
s2_ac <- (5 * s2_a + 5 * s2_c)/10
s2_bc <- (5 * s2_b + 5 * s2_c)/10
t_ab <- (mean_a - mean_b)/sqrt(s2_ab * (1/6 + 1/6))
t_ac <- (mean_a - mean_c)/sqrt(s2_ac * (1/6 + 1/6))
t_bc <- (mean_b - mean_c)/sqrt(s2_bc * (1/6 + 1/6))
c(t_ab, t_ac, t_bc)
```

```
## [1]  -5.491522 -13.000542  -5.358510
```

I just realized after I did this that the Fisher's Least Significant Difference requires to use MSW as the common variant. No matter, we show here that the resulting t-score using MSW and standard deviation between two samples are different.

```
t_ab_msw <- (mean_a - mean_b)/sqrt(0.092 * (1/6 + 1/6))
t_ac_msw <- (mean_a - mean_c)/sqrt(0.092 * (1/6 + 1/6))
t_bc_msw <- (mean_b - mean_c)/sqrt(0.092 * (1/6 + 1/6))
c(t_ab_msw, t_ac_msw, t_bc_msw)
```

```
## [1]  -5.710402 -11.420805  -5.710402
```

```
table <- data.frame(
  Comparison = c("A vs. B", "A vs. C", "B vs. C"),
  t_score = c(t_ab, t_ac, t_bc),
  p_value = c(pt(t_ab, 15) * 2, pt(t_ac, 15) * 2, pt(t_bc, 15) * 2),
  t_score_msw = c(t_ab_msw, t_ac_msw, t_bc_msw),
  p_value_msw = c(pt(t_ab_msw, 15) * 2, pt(t_ac_msw, 15) * 2, pt(t_bc_msw, 15) * 2),
  p_adj = c(pt(t_ab_msw, 15) * 2, pt(t_ac_msw, 15) * 2, pt(t_bc_msw, 15) * 2) * choose(3, 2)
)
table
```

```
##   Comparison    t_score      p_value t_score_msw  p_value_msw        p_adj
## 1    A vs. B  -5.491522 6.203182e-05   -5.710402 4.128135e-05 1.238440e-04
## 2    A vs. C -13.000542 1.436101e-09  -11.420805 8.473751e-09 2.542125e-08
## 3    B vs. C  -5.358510 7.969684e-05   -5.710402 4.128135e-05 1.238440e-04
```

Thus the table should be:

```
table[, c(1,4,5,6)]
```

```
##   Comparison t_score_msw  p_value_msw        p_adj
## 1    A vs. B   -5.710402 4.128135e-05 1.238440e-04
## 2    A vs. C  -11.420805 8.473751e-09 2.542125e-08
## 3    B vs. C   -5.710402 4.128135e-05 1.238440e-04
```

```
# Trying to write a all-in-one function to calculate the pairwise t-tests
get_msw <- function(data){
  msw <- 0
  for(i in 1 : dim(data)[2]){
    msw <- msw + (length(data[,i]) - 1) * var(data[,i])
  }
  return(msw / (dim(data)[1] * dim(data)[2] - dim(data)[2]))
}

pairwise_t <- function(data){
  msw <- get_msw(data)
  row_names <- c()
  t_score <- c()
  p_value <- c()

  for(i in 1 : (dim(data)[2] - 1)){
    for(j in (i + 1) : dim(data)[2]){
      a <- data[,i]
      b <- data[,j]

      # Add two column names to the row name
      row_names <- append(row_names,paste0(colnames(data)[i], " vs. ", colnames(data)[j]))

      # Calculate the t_score
      t <- (mean(a) - mean(b)) / sqrt(msw * (1 / length(a) + 1 / length(b)))
      t_score <- append(t_score, t)

      # Calculate the p_value
      p_value <- append(p_value, pt(t, dim(data)[1] * dim(data)[2] - dim(data)[2]) * 2)
    }
  }
  table <- data.frame(
    Comparison = row_names,
    t = t_score,
    p = p_value,
    p.adj = p_value * choose(dim(data)[2], 2)
  )
  return(table)
}
pairwise_t(tar)
```

```
##           Comparison          t            p        p.adj
## 1 BrandA vs. BrandB  -5.721192 4.046768e-05 1.214030e-04
## 2 BrandA vs. BrandC -11.442383 8.259836e-09 2.477951e-08
## 3 BrandB vs. BrandC  -5.721192 4.046768e-05 1.214030e-04
```

Which give identical results, Hooray!