大數據統計分析與預測 第十一章作業 (Two Sample Test)

Create by 黃彥鈞 Weber YC, Huang (m946108006)

Q3. When should you use the two-sample t-test? When must the modified version of the test be applied?

A two-sample t-test is used when you want to compare two independent groups to see if their **means** are different.

In the situation that the variance σ^2 of two populations are **NOT** assumed to be equal, a modification of the two-sample t-test must be applied.

Q4. What is the rationale for using a pooled estimate of the variance in the two-sample t-test?

Only at the time that we **KNOW** the variance for both populations **are equal** or they **are assumed to be equal** can we use the pooled estimate of the variance in the two-sample t-test.

Q7. The following data come from a study that examines the efficacy of saliva cotinine as an indicator for exposure to tobacco smoke. In one part of the study, seven subjects -- none of whom were heavy smokers and all of whom had abstained from smoking for at least one week prior to the study -- were each required to smoke a single cigarette. Samples of saliva were taken from all individuals 2, 12, 24, and 48 hours after smoking the cigarette. The cotinine levels at 12 hours and at 24 hours are shown below :

Cotinine Levels (nmol/l)

SUBJECT	AFTER 12 HOURS	AFTER 24 HOURS
1	73	24
2	58	27
3	67	49
4	93	59
5	33	0
6	18	11
7	147	43

Let μ_{12} represent the population mean cotinine level 12 hours after smoking the cigarette and μ_{24} the mean cotinine level 24 hours after smoking. It is believed that μ_{24} must be lower than μ_{12}.

(a) Construct a one-sided 95% confidence interval for the true difference in population means $\mu_{12} - \mu_{24}$.

$H_0 : \mu_{24} \geq \mu_{12}$
$\Rightarrow \mu_{12} - \mu_{24} \leq 0$

$H_a : \mu_{24} < \mu_{12}$
$\Rightarrow \mu_{12} - \mu_{24} > 0$

> *R-code :*

```
# -------------------------------------
# Q7.a
paired_t <- function(len_1,len_2){
  l1 = c()
  cat(' plz fill in list one ...','\n')
  for (i in c(1:len_1)) {
    question1 <- as.numeric(readline(prompt="Enter number:
"))
    l1 <- append(l1,question1)
  }
  cat('list one :', l1, '\n')

  cat('\n', 'plz fill in list two ...','\n')
  l2 = c()
  for (j in c(1:len_2)) {
    question2 <- as.numeric(readline(prompt="Enter number:
"))
    l2 <- append(l2,question2)
  }
  cat('list two :', l2, '\n')

  sided <- as.character(readline(prompt="Enter sided
(less(left) or two.sided or greater(right)) :"))

  t.test(l1, l2, paired=TRUE, alternative = sided)
}
paired_t(7,7) # Run code ...
```

> *Output :*

```
 plz fill in list one ...
Enter number: 73
Enter number: 58
Enter number: 67
```

```
Enter number: 93
Enter number: 33
Enter number: 18
Enter number: 147
list one : 73 58 67 93 33 18 147

 plz fill in list two ...
Enter number: 24
Enter number: 27
Enter number: 49
Enter number: 59
Enter number: 0
Enter number: 11
Enter number: 43
list two : 24 27 49 59 0 11 43
Enter sided (less(left) or two.sided or greater(right)) :
greater

    Paired t-test

data:  l1 and l2
t = 3.3228, df = 6, p-value = 0.007974

alternative hypothesis: true difference in means is
greater than 0

95 percent confidence interval:
 16.37073      Inf

sample estimates:
mean of the differences
```

- Answer : 16.37

(b) Test the null hypothesis that the population means are identical at the $\alpha = 0.05$ level of significance. What do you conclude?

$H_0 : \mu_{24} \geq \mu_{12}$
$\quad => \mu_{12} - \mu_{24} \leq 0$

$H_a : \mu_{24} < \mu_{12}$
$\quad => \mu_{12} - \mu_{24} > 0 (right(upper)\ tail)$

$\Delta : \mu_{24} = \mu_{12}$
$\quad => \mu_{12} - \mu_{24} = 0$

$t = \frac{\bar{d} - \Delta}{S_d / \sqrt{n}}$

> *R-code :*

```r
# Q7.b
paired_t <- function(len_1,len_2, Delta, n){
  l1 = c()
  cat(' plz fill in list one ...','\n')
  for (i in c(1:len_1)) {
    question1 <- as.numeric(readline(prompt="Enter number:
"))
    l1 <- append(l1,question1)
  }
  cat('list one :', l1, '\n')

  cat('\n', 'plz fill in list two ...','\n')
  l2 = c()
  for (j in c(1:len_2)) {
    question2 <- as.numeric(readline(prompt="Enter number:
"))
    l2 <- append(l2,question2)
  }
  cat('list two :', l2, '\n')


  dif <- abs(l1-l2)
  dif_bar <- mean(dif)
  dif_sd <- sd(dif)
  cat('\n','Difference :', dif, '\n')
  cat('Difference mean :', dif_bar, '\n')
  cat('Difference sd :', dif_sd, '\n')

  t = (dif_bar-Delta)/(dif_sd/sqrt(n))

  cat('\n', '==== Answer ====', '\n', 'T :', t, '\n')

  pval = pt(t, df=n-1, lower.tail=FALSE)
  cat('\n', '==== Answer ====', '\n')
  cat('P-value :', pval)
}
paired_t(7,7,0,7) # Run code ...
```

> *Output :*

```
 plz fill in list one ...
Enter number: 73
Enter number: 58
Enter number: 67
Enter number: 93
Enter number: 33
Enter number: 18
Enter number: 147
list one : 73 58 67 93 33 18 147

 plz fill in list two ...
```

```
Enter number: 24
Enter number: 27
Enter number: 49
Enter number: 59
Enter number: 0
Enter number: 11
Enter number: 43
list two : 24 27 49 59 0 11 43

 Difference : 49 31 18 34 33 7 104
Difference mean : 39.42857
Difference sd : 31.39457

 ==== Answer ====
 T : 3.32281

 ==== Answer ====
P-value : 0.007974272
```

Since P-value is lower than $\alpha = 0.05$, we reject the H_0, and conclude that the true difference in population mean cotinine levels is not equal to 0. Mean cotinine level decreases significantly between 12 and 24 hours after smoking.

Q8. A study was conducted to determine whether an expectant mother's cigarette smoking has any effect on the bone mineral content of her otherwise healthy child. A sample of 77 newborns whose mothers smoked during pregnancy has mean bone mineral content $\overline{x_1} = 0.098\ g/cm$ and standard deviation $S_1 = 0.026\ g/cm$; a sample of 161 infants whose mothers did not smoke has mean $\overline{x_2} = 0.095\ g/cm$ and standard deviation $S_2 = 0.025\ g/cm$. Assume that the underlying population variances are equal.

(a) Are the two samples of data paired or independent?

Definitely, smoking and non-smoking won't influence each other, so these two samples of data independent.

(b) State the null and alternative hypothesis of two-sided test.

$H_0 : \mu_1 - \mu_2 = 0$

$H_a : \mu_1 - \mu_2 \neq 0$

(c) Conduct the test at the 0.05 level of significance. What do you conclude?

$$t_0 = \frac{\overline{x_1} - \overline{x_2} - \Delta}{\sqrt{S_p^2 \times (\frac{1}{n_1} + \frac{1}{n_2})}}$$

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

$$\Delta = 0 => hypothesized\ if\ testing\ for\ equal\ means$$
$$df = n_1 + n_2 - 2$$

(c-1) Method 1 (手刻達人):

> *R-code :*

```
# Q8.c Q11.b
ind_t <- function(){
  x1 <- as.numeric(readline(prompt="x1 ? :"))
  x2 <- as.numeric(readline(prompt="x2 ? :"))
  s1 <- as.numeric(readline(prompt="s1 ? :"))
  s2 <- as.numeric(readline(prompt="s2 ? :"))
  n1 <- as.numeric(readline(prompt="n1 ? :"))
  n2 <- as.numeric(readline(prompt="n2 ? :"))
  var_equal <- as.character(readline(prompt="Are
population variances assume to be equal (y/n) ? : "))
  tails <- as.character(readline(prompt="What tails of
test would you like to perform (1/r/2) ? : "))

  if (var_equal == 'y') {
    sp_square <- (((n1-1)*s1*s1)+((n2-1)*s2*s2))/(n1+n2-2)
    t0 <- (x1-x2)/(sqrt(sp_square)*sqrt((1/n1)+(1/n2)))
    df=(n1+n2-2)
    cat('sp square :', sp_square, '\n')
    cat('T :', t0, '\n')
    cat('df :', df, '\n')

    if (tails == '2'){
        p = 2 * pt(-abs(t0), df=df)
        cat(' ==== Answer ====','\n')
        cat('P-value :', p)
    }else{
        p = pt(-abs(t0), df=df)
        cat(' ==== Answer ====','\n')
        cat('P-value :', p)
    }
  }
  else{
    t0 <- ((x1-x2)-0)/sqrt((s1*s1/n1)+(s2*s2/n2))
    df <- ((s1*s1/n1)+(s2*s2/n2))^2/(((s1*s1/n1)^2/(n1-
1))+((s2*s2/n2)^2/(n2-1)))
    cat('T :', t0, '\n')
    cat('df :', df, '\n')

    if (tails == '2'){
        p = 2 * pt(-abs(t0), df=df)
        cat(' ==== Answer ====','\n')
        cat('P-value :', p)
    }else{
```

```
        p = pt(-abs(t0), df=df)
        cat(' ==== Answer ====','\n')
        cat('P-value :', p)
    }
  }
}
ind_t()
```

> *Output :*

```
x1 ? :0.098
x2 ? :0.095
s1 ? :0.026
s2 ? :0.025
n1 ? :77
n2 ? :161
Are population variances assume to be equal (y/n) ? : y
What tails of test would you like to perform (l/r/2) ? : 2
sp square : 0.0006414237
T : 0.8549064
df : 236
 ==== Answer ====
P-value : 0.39347
```

(c-2) Method 2 (不重複造車...):

> *R-code :*

```
# Q8.c Q11.b
library(BSDA)

summary_ttest <- function(){
    x1 <- as.numeric(readline(prompt="x1 ? :"))
    x2 <- as.numeric(readline(prompt="x2 ? :"))
    s1 <- as.numeric(readline(prompt="s1 ? :"))
    s2 <- as.numeric(readline(prompt="s2 ? :"))
    n1 <- as.numeric(readline(prompt="n1 ? :"))
    n2 <- as.numeric(readline(prompt="n2 ? :"))
    alter <- as.character(readline(prompt="alternative
(greater/less/two.sided)? :"))
    mu <- as.numeric(readline(prompt="H0 ? :"))
    vare <- as.character(readline(prompt="Are population
variances assume to be equal (y/n) ? : "))
    alpha <- as.numeric(readline(prompt="Alpha ? :"))
    if(vare == "y"){
        tsum.test(x1, s1, n1, x2, s2, n2, alternative =
alter, mu = mu,
                  var.equal = TRUE, conf.level = 1-alpha)
    }else{
        tsum.test(x1, s1, n1, x2, s2, n2, alternative =
alter, mu = mu,
```

```
                    var.equal = FALSE, conf.level = 1-alpha)
    }
}
```

> *Output :*

```
x1 ? :0.098
x2 ? :0.095
s1 ? :0.026
s2 ? :0.025
n1 ? :77
n2 ? :161
alternative (greater/left/two.sided)? :two.sided
H0 ? :0
Are population variances assume to be equal (y/n) ? : y
Alpha ? :0.05

    Standard Two-Sample t-Test

data:  Summarized x and y
t = 0.85491, df = 236, p-value = 0.3935
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -0.003913272  0.009913272
sample estimates:
mean of x mean of y
    0.098     0.095
```

Since P-value is lower than $\alpha = 0.05$, we reject the H_0, and conclude that the true difference in population mean bone mineral is not equal to 0.

> *遭遇問題 (已解決) :*
>
> *Q. 為何雙尾檢定會出現 P value 大於 1 的情況 ?*
>
> *A. 因為雙尾 t 值於分布曲線上可能是正值也可能是負值，因此必須要先將 t 值取絕對值後，應用高尾 (upper.tail) 來計算；或絕對值後取負值，應用低尾 (lower.tail) 計算，方不會出錯。*
>
> *參考 : Manually calculating p-value for t-test: How to avoid values bigger than 1?*

Q11. The table below compares the levels of carboxyhemoglobin for a group of non-smokers and a group of cigarette smokers. Sample means and standard deviations are shown. It is believed that the mean carboxyhemoglobin level of the smokers must be higher than mean level of the nonsmokers. There is no reason to assume that the underlying population variances are identical.

GROUP	N	CARBOXYHEMOGLOBIN(%)
Non-smokers	121	$\overline{x} = 1.3, s = 1.3$
Smokers	75	$\overline{x} = 4.1, s = 2.0$

(a) What are the null and alternative hypotheses of the one-sided test?

$H_0 : \mu_2 \leq \mu_1 => \mu_1 - \mu_2 \geq 0$

$H_a : \mu_2 > \mu_1 => \mu_1 - \mu_2 < 0 \; (left(lower)\; tail)$

(b) Conduct the test at the 0.05 level of significance. What do you conclude?

$$t_0 = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$df = \frac{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})^2}{(\frac{(\frac{S_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{S_2^2}{n_2})^2}{n_2 - 1})}$$

R-code :

```
# Q8.c Q11.b
# code is in Q8. c
summary_ttest() # Run code ...
ind_t() # for re-check
```

Output :

```
x1 ? :1.3
x2 ? :4.1
s1 ? :1.3
s2 ? :2.0
n1 ? :121
n2 ? :75
alternative (greater/less/two.sided)? :less
H0 ? :0
Are population variances assume to be equal (y/n) ? : n
Alpha ? :0.05

    Welch Modified Two-Sample t-Test

data:  Summarized x and y
t = -10.793, df = 113.05, p-value < 2.2e-16
```

```
alternative hypothesis: true difference in means is less
than 0
95 percent confidence interval:
        NA -2.369762
sample estimates:
mean of x mean of y
      1.3       4.1
```

Since P-value is extremely lower than $\alpha = 0.05$, we reject the H_0, and conclude that the mean carboxyhemoglobin level of the smokers must be higher than mean level of the nonsmokers.

Q12. Suppose that you wish to compare the characteristic of tuberculosis meningitis in patients infected with HIV and those who are not infected. In particular, you would like to determine whether the two populations have the same mean age. A sample of 37 infected patients has mean age $\overline{x_1} = 27.9$ years and standard deviation $s_1 = 5.6$ years; a sample of 19 patients who are not infected has mean age $\overline{x_2} = 38.8$ years and standard deviation $\overline{s_2} = 21.7$ years.

(a) Test the null hypothesis that the two populations of patients have the same mean age at the 0.05 level of significance.

$H_a : \mu_1 - \mu_2 = 0$

$H_a : \mu_1 - \mu_2 \neq 0$

> *R-code :*

```
summary_ttest() # Run code ...
```

> *Output :*

```
x1 ? :27.9
x2 ? :38.8
s1 ? :5.6
s2 ? :21.7
n1 ? :37
n2 ? :19
alternative (greater/less/two.sided)? :two.sided
H0 ? :0
Are population variances assume to be equal (y/n) ? : n
Alpha ? :0.05

    Welch Modified Two-Sample t-Test

data:  Summarized x and y
t = -2.153, df = 19.241, p-value = 0.04421
```

```
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -21.487445  -0.312555
sample estimates:
mean of x mean of y
     27.9      38.8
```

(b) Do you expect that a 95% confidence interval for the true difference in population means would contain the value 0? Why or why not?

I do not expect that a 95 % confidence interval for the true difference in population means would contain the value 0, since the confidence interval for 95 % between the assumption of population variances are same or not are c(-21.487445 , -0.312555)、c(-18.44668 , -3.35332). Both of them are not including 0.

Extra Questions :

Data : CVD_ALL.csv

Table in need : 64489 × 4 (sample head(6))

CVD	WAIST	SBP	DBP
0	81.0	138.0	87.0
0	79.0	98.0	66.0
0	86.5	135.0	97.0
0	84.0	117.5	88.5
1	96.0	153.0	91.5
1	94.0	191.0	135.0

```
# preprocessing...
# set working dir
setwd('C:/Users/doudi/Downloads')

# load packages
library(dplyr)
library(magrittr)

# load file
data <- read.csv('CVD_ALL.csv', encoding = 'utf-8')
df <- data[,c(2,6,7,8)]
colnames(df) <- c('CVD', 'waist', 'SBP', 'DBP')

# drop NA
sapply(df, function(x) {sum(is.na(x))})
df = df %>% na.omit()
```

```r
# Calculating difference
df$Diff <- df$SBP - df$DBP

# Seperate waist data as CVD 0,1 to respective variables
cvd0 = df %>% filter(CVD==0) %>% select(waist)
cvd1 = df %>% filter(CVD==1) %>% select(waist)
```

Extra_01. SBP 收縮壓和 DBP 舒張壓，差異是否大於 50？

(Hint. Use pair, this is not a independent case.)

$H_0 : \mu_{SBP} - \mu_{DBP} \leq 50$

$H_a : \mu_{SBP} - \mu_{DBP} > 50$

> *R-code :*

```r
# Perform two samples
t.test(df$SBP, df$DBP, mu = 50, paired = TRUE, alternative
= "greater")

# Calculate difference first and perform one sample
t.test(df$Diff, mu = 50, alternative = "greater")
```

> *Output :*

```
Paired t-test

data:  df$SBP and df$DBP
t = -84.39, df = 62336, p-value = 1
alternative hypothesis: true difference in means is
greater than 50
95 percent confidence interval:
 45.06764      Inf
sample estimates:
mean of the differences
               45.16194

#----------------------

One Sample t-test

data:  df$Diff
t = -84.39, df = 62336, p-value = 1
alternative hypothesis: true mean is greater than 50
95 percent confidence interval:
 45.06764      Inf
sample estimates:
mean of x
```

```
 45.16194
```

Since the p-value of both test are in 1, we do not reject $H_0 : \mu_{SBP} - \mu_{DBP} \leq 50$, and conclude that the difference in population means of SBP and DBP is lower or equal than 50.

Extra_02. waist(CVD=1)和waist(CVD=0) 是否有差異?

(Hint. It is a independent case.)

$H_0 : \mu_0 - \mu_1 = 0$

$H_a : \mu_0 - \mu_1 \neq 0$

> *R-code :*

```
t.test(cvd0, cvd1, alternatice = 'two.sided')
```

> *Output :*

```
Welch Two Sample t-test

data:  cvd0 and cvd1
t = -33.457, df = 7051.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -5.084332 -4.521517
sample estimates:
mean of x mean of y
 77.89709  82.70002
```

As the p-value is lower than $\alpha = 0.05$, we reject the $H_0 : \mu_0 - \mu_1 = 0$, and conclude that the population means of waist data in CVD=0 and CVD=2 are difference with each other.