大數據統計分析與預測 第十二章作業 (ANOVA)

Created by Weber YC, Huang (m946108006) 2019-12-13

Q2. What is the idea behind the one-way analysis of variance? What two measures of variation are being compared?

It is aimed to compare *k* groups' difference of *means*. It is a test of ratio of **within-group variance** and **between-group variance.**

Q5. Consider the F distribution with 8 and 16 degrees of freedom.

(a) What proportion of the area under the curve lies to the right of F = 2.09?

> R-code :

```
pf(2.09, 8, 16, lower.tail = FALSE)
```

> Output :

```
0.09971543
```

Answer : approxmately **10% of the area** under the curve lies to the right of F = 2.09.

(b) What value of F cuts off the upper 1% of the distribution?

> R-code :

```
qf(.01, 8, 16, lower.tail = FALSE)
```

> Output :

```
3.889572
```

Answer : F approxmately **3.89**

(c) What proportion of the area under the curve lies to the left of F =4.52?

> R-code :

```
pf(4.52, 8, 16, lower.tail = FALSE)
```

> Output :

```
0.005003502
```

Answer : **0.5%**

Q7. A study of patients with insulin-dependent diabetes was conducted to investigate the effects of cigarette smoking on renal and retinal complications. Before examining the result of the study, you wish to compare the baseline measure of systolic blood pressure across four different subgroups : nonsmokers, current smokers, ex-smokers, and tobacco chewers. A sample is selected from each subgroup； the relevant data are shown below. Means and standard deviations are expressed in $mm\ Hg$. Assume that systolic blood pressure is normally distributed.

	n	$\overline{x}$	s
Nonsmokers	269	115	13.4
Current Smokers	53	114	10.1
Ex-smokers	28	118	11.6
Tobacco Chewers	9	126	12.2

(a) Calculate the estimate of the within-groups variance.

> R-code :

```
group_var <- function(k){

  cat('==== fill in the sums of the obs ====', '\n')

  N <- c()
  for (i in c(1:k)) {
    n <- as.numeric(readline(prompt="Enter n: "))
    N <- append(N, n)
  }
  cat('Sums of the obs are : ', N ,'\n')

  cat('\n', '==== fill in the means of the obs ====', '\n')
  M <- c()
  for (j in c(1:k)) {
    m <- as.numeric(readline(prompt="Enter means: "))
    M <- append(M, m)
  }
  cat('Means of the obs are : ', M ,'\n')

  cat('\n', '==== fill in the sds of the obs ====', '\n')
  S <- c()
  for (l in c(1:k)) {
    s <- as.numeric(readline(prompt="Enter sds: "))
    S <- append(S, s)
  }
  cat('Sds of the obs are : ', S ,'\n')

  in_group_var <- sum(((N-1)*(S^2)))/(sum(N)-k)

  cat('\n', '==== within group var ====', '\n', in_group_var, '\n')

  xbar <- sum(N*M)/sum(N)

  be_group_var <- sum(N*((M-xbar)^2))/(k-1)
```

```
  cat('\n', '==== between group var ====', '\n', be_group_var, '\n')

}
group_var(4) # Run code ...
```

> Output :

```
==== fill in the sums of the obs ====
Enter n: 269
Enter n: 53
Enter n: 28
Enter n: 9
Sums of the obs are :  269 53 28 9

 ==== fill in the means of the obs ====
Enter means: 115
Enter means: 114
Enter means: 118
Enter means: 126
Means of the obs are :  115 114 118 126

 ==== fill in the sds of the obs ====
Enter sds: 13.4
Enter sds: 10.1
Enter sds: 11.6
Enter sds: 12.2
Sds of the obs are :  13.4 10.1 11.6 12.2

 ==== within group var ====
 164.0857

 ==== between group var ====
 448.9749
```

Answer : **164.1**

(b) Calculate the estimate of the between-groups variance.

Answer : **449.0**

(c) At the 0.05 level of significance, test the null hypothesis that the mean systolic blood pressures of the four groups are identical. What do you conclude?

$F_0 = \frac{S_B^2}{S_W^2}$

> R-code :

```
pf((449/164.1), 3, 355, lower.tail = FALSE)
```

> Output :

```
0.04347148
```

Answer : We reject H_0 at the level of significance and conclude that there is a difference among the mean systolic blood pressures of the four groups.

(d) If you find that the population means are not all equal, use the Bonferroni multiple comparisons procedure to determine where the differences lie. What is the significance level of each individual test?

> R-code :

```
group_var <- function(k, afa){

  cat('==== fill in the sums of the obs ====', '\n')

  N <- c()
  for (i in c(1:k)) {
    n <- as.numeric(readline(prompt="Enter n: "))
    N <- append(N, n)
  }
  cat('Sums of the obs are : ', N ,'\n')

  cat('\n', '==== fill in the means of the obs ====', '\n')
  M <- c()
  for (j in c(1:k)) {
    m <- as.numeric(readline(prompt="Enter means: "))
    M <- append(M, m)
  }
  cat('Means of the obs are : ', M ,'\n')

  cat('\n', '==== fill in the sds of the obs ====', '\n')
  S <- c()
  for (l in c(1:k)) {
    s <- as.numeric(readline(prompt="Enter sds: "))
    S <- append(S, s)
  }
  cat('Sds of the obs are : ', S ,'\n')

  in_group_var <- sum(((N-1)*(S^2)))/(sum(N)-k)

  cat('\n', '==== within group var ====', '\n', in_group_var, '\n')

  xbar <- sum(N*M)/sum(N)

  be_group_var <- sum(N*((M-xbar)^2))/(k-1)

  cat('\n', '==== between group var ====', '\n', be_group_var, '\n', '\n')

  com_n <- dim(combn(k, 2))[2]
  C <- as.data.frame(combn(k, 2))


  for(col in C){
    X <- c(M[col[1]],M[col[2]])
    Num <- c(N[col[1]],N[col[2]])
    afa_new <- afa/com_n
    cat('new alpha :', afa_new, '\n','\n')
    t <- (X[1]-X[2])/sqrt(in_group_var*((1/Num[1])+(1/Num[2])))
    df <- sum(N)-k
    cat('==== T value of condition ====', '\n', col[1], '&', col[2], ':', t,
'\n')
    pval <- 2*pt(-abs(t), df=df)
```

```
    cat('==== P value of condition ====', '\n', col[1], '&', col[2], ':', pval,
'\n','\n')

    if (pval < afa_new) {
      cat('Reject H0 in the condition of', col[1], '&', col[2], '\n')
    }else{
      cat('Do not reject H0 in the condition of', col[1], '&', col[2], '\n')
    }
    cat('\n', '\n')
  }
}
group_var(4, 0.05) # Run code ...
```

Output :

```
==== fill in the sums of the obs ====
Enter n: 269
Enter n: 53
Enter n: 28
Enter n: 9
Sums of the obs are :  269 53 28 9

 ==== fill in the means of the obs ====
Enter means: 115
Enter means: 114
Enter means: 118
Enter means: 126
Means of the obs are :  115 114 118 126

 ==== fill in the sds of the obs ====
Enter sds: 13.4
Enter sds: 10.1
Enter sds: 11.6
Enter sds: 12.2
Sds of the obs are :  13.4 10.1 11.6 12.2

 ==== within group var ====
 164.0857

 ==== between group var ====
 448.9749

new alpha : 0.008333333

==== T value of condition ====
 1 & 2 : 0.5194583
==== P value of condition ====
 1 & 2 : 0.6037649

Do not reject H0 in the condition of 1 & 2


new alpha : 0.008333333

==== T value of condition ====
 1 & 3 : -1.179404
==== P value of condition ====
```

```
 1 & 3 : 0.239027

Do not reject H0 in the condition of 1 & 3


new alpha : 0.008333333

==== T value of condition ====
 1 & 4 : -2.53415
==== P value of condition ====
 1 & 4 : 0.0117013

Do not reject H0 in the condition of 1 & 4


new alpha : 0.008333333

==== T value of condition ====
 2 & 3 : -1.336592
==== P value of condition ====
 2 & 3 : 0.1822117

Do not reject H0 in the condition of 2 & 3


new alpha : 0.008333333

==== T value of condition ====
 2 & 4 : -2.598419
==== P value of condition ====
 2 & 4 : 0.009755654

Do not reject H0 in the condition of 2 & 4


new alpha : 0.008333333

==== T value of condition ====
 3 & 4 : -1.629874
==== P value of condition ====
 3 & 4 : 0.1040151

Do not reject H0 in the condition of 3 & 4
```

Answer : Since all comparison is do not reject the H_0, there is no difference among the mean systolic blood pressures of the four groups.

Q8. One of the goals of the Edinburgh Artery Study was to investigate the risk factors for peripheral arterial disease among persons 55 to 74 years of age. You wish to compare mean LDL cholesterol levels, measured in $mmol/liter$, among four different populations of subjects : patients with intermittent claudication or interruptions in movements, those with major asymptomatic disease, those with minor asymptomatic disease, and those with no evidence of disease at all. Samples are selected from each population； summary statistics are shown below.

	n	$\overline{x}$	s
Intermittent Claudication	73	6.22	1.62
Major Asymptomatic Disease	105	5.81	1.43
Minor Asymptomatic Disease	240	5.77	1.24
No Disease	1080	5.47	1.31

(a) Test the null hypothesis that the mean LDL cholesterol levels are the same for each of the four populations. What are the degrees of freedom associated with this test?

Answer :

$df_n = sum(n) - k = 73 + 105 + 240 + 1080 - 4 = 1494$
$df_k = k - 1 = 4$

(b) What do you conclude?

> R-code :

```
group_var <- function(k, afa){

  cat('==== fill in the sums of the obs ====', '\n')

  N <- c()
  for (i in c(1:k)) {
    n <- as.numeric(readline(prompt="Enter n: "))
    N <- append(N, n)
  }
  cat('Sums of the obs are : ', N ,'\n')

  cat('\n', '==== fill in the means of the obs ====', '\n')
  M <- c()
  for (j in c(1:k)) {
    m <- as.numeric(readline(prompt="Enter means: "))
    M <- append(M, m)
  }
  cat('Means of the obs are : ', M ,'\n')

  cat('\n', '==== fill in the sds of the obs ====', '\n')
  S <- c()
  for (l in c(1:k)) {
    s <- as.numeric(readline(prompt="Enter sds: "))
    S <- append(S, s)
  }
  cat('Sds of the obs are : ', S ,'\n')

  in_group_var <- sum(((N-1)*(S^2)))/(sum(N)-k)

  cat('\n', '==== within group var ====', '\n', in_group_var, '\n')

  xbar <- sum(N*M)/sum(N)
```

```
  be_group_var <- sum(N*((M-xbar)^2))/(k-1)

  cat('\n', '==== between group var ====', '\n', be_group_var, '\n', '\n')

  f = be_group_var/in_group_var

  p <- pf(f, sum(N)-k, k-1, lower.tail = FALSE)

  cat('\n', '==== p-value for f-distribution ====', '\n', p, '\n', '\n')


  # multi comparison
  com_n <- dim(combn(k, 2))[2]
  C <- as.data.frame(combn(k, 2))


  for(col in C){
    X <- c(M[col[1]],M[col[2]])
    Num <- c(N[col[1]],N[col[2]])
    afa_new <- afa/com_n
    cat('new alpha :', afa_new, '\n','\n')
    t <- (X[1]-X[2])/sqrt(in_group_var*((1/Num[1])+(1/Num[2])))
    df <- sum(N)-k
    cat('==== T value of condition ====', '\n', col[1], '&', col[2], ':', t,
'\n')
    pval <- 2*pt(-abs(t), df=df)
    cat('==== P value of condition ====', '\n', col[1], '&', col[2], ':', pval,
'\n','\n')

    if (pval < afa_new) {
      cat('Reject H0 in the condition of', col[1], '&', col[2], '\n')
    }else{
      cat('Do not reject H0 in the condition of', col[1], '&', col[2], '\n')
    }
    cat('\n', '\n')
  }
}
group_var(4, 0.05) # Run code ...
```

Output :

```
==== fill in the sums of the obs ====
Enter n: 73
Enter n: 105
Enter n: 240
Enter n: 1080
Sums of the obs are :  73 105 240 1080

 ==== fill in the means of the obs ====
Enter means: 6.22
Enter means: 5.81
Enter means: 5.77
Enter means: 5.47
Means of the obs are :  6.22 5.81 5.77 5.47

 ==== fill in the sds of the obs ====
Enter sds: 1.62
```

```
Enter sds: 1.43
Enter sds: 1.24
Enter sds: 1.31
Sds of the obs are :  1.62 1.43 1.24 1.31

 ==== within group var ====
 1.754207

 ==== between group var ====
 19.06123

 ==== p-value for f-distribution ====
 0.03555153

new alpha : 0.008333333

==== T value of condition ====
 1 & 2 : 2.031372
==== P value of condition ====
 1 & 2 : 0.04239393

Do not reject H0 in the condition of 1 & 2


new alpha : 0.008333333

==== T value of condition ====
 1 & 3 : 2.54195
==== P value of condition ====
 1 & 3 : 0.01112387

Do not reject H0 in the condition of 1 & 3


new alpha : 0.008333333

==== T value of condition ====
 1 & 4 : 4.682519
==== P value of condition ====
 1 & 4 : 3.090762e-06

Reject H0 in the condition of 1 & 4


new alpha : 0.008333333

==== T value of condition ====
 2 & 3 : 0.2581133
==== P value of condition ====
 2 & 3 : 0.796355

Do not reject H0 in the condition of 2 & 3


new alpha : 0.008333333

==== T value of condition ====
 2 & 4 : 2.511227
```

```
==== P value of condition ====
 2 & 4 : 0.01213607

Do not reject H0 in the condition of 2 & 4


new alpha : 0.008333333

==== T value of condition ====
 3 & 4 : 3.174033
==== P value of condition ====
 3 & 4 : 0.00153401

Reject H0 in the condition of 3 & 4
```

Answer :

The p-value of mean LDL cholesterol levels is less than $\alpha = 0.05$, so we reject the H_0 that the mean LDL cholesterol levels are the same for each of the four populations. And according to the multi comparison, the mean of Intermittent Claudication and the mean of Minor Asymptomatic Disease are less than the mean of no disease.

(c) What assumptions about the data must be true for you to use one-way analysis of variance technique?

Answer :

- Response variable residuals are normally distributed (or approximately normally distributed).
- Variances of populations are equal.
- Responses for a given group are independent and identically distributed normal random variables (not a simple random sample (SRS))

(d) Is it necessary to take an additional step in this analysis? If so, what is it? Explain.

Answer : Since the outcome of the anova is to reject the H_0, so we have to perform multiple comparison to further test the difference among all set.