单因素方差分析&
多重检验矫正

杨俊晨、魏杨蓁、张睿涵

1. One-way ANOVA　问题描述

水平	A1	A2	…	As
观察值	X_{11}	X_{12}	…	X_{1s}
	X_{21}	X_{22}	…	X_{2s}
	…	…	…	…
	$X_{n_1 1}$	$X_{n_2 2}$	…	$X_{n_s s}$
样本总和	T_1	T_2	…	T_s
样本均值	$\overline{X_{.1}}$	$\overline{X_{.2}}$	…	$\overline{X_{.s}}$
总体均值	μ_1	μ_2	…	μ_s

假定在水平 A_j 下的多次观察的结果为 $X_{1j}, X_{2j} \cdots X_{n_j j}$，视为来自正态总体 $N(\mu_j, \sigma^2)$ 的一个简单随机样本。

1. One-way ANOVA

变量定义

水平	A1	A2	…	As
观察值	X_{11}	X_{12}	…	X_{1s}
	X_{21}	X_{22}	…	X_{2s}
	…	…	…	…
	$X_{n_1 1}$	$X_{n_2 2}$	…	$X_{n_s s}$
样本总和	T_1	T_2	…	T_s
样本均值	$\overline{X}_{\cdot 1}$	$\overline{X}_{\cdot 2}$	…	$\overline{X}_{\cdot s}$
总体均值	μ_1	μ_2	…	μ_s

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{s}\sum_{i=1}^{n_j} X_{ij}$$

$$\overline{X}_{\cdot j} = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij} = \mu + \alpha_j + \varepsilon_{\cdot j}$$

记$\varepsilon_{ij} = X_{ij} - \mu_j$ 为随机误差, 则$\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$

上述单因素模型可表示为$\begin{cases} X_{ij} = \mu_j + \varepsilon_{ij} \\ \varepsilon_{ij} \sim N(0,\sigma^2) \end{cases}$

ANOVA的任务是检验在各水平下的均值是否相等，即$H_0 : \mu_1 = \mu_2 = \ldots = \mu_s$ ，$H_1 : \mu_1, \mu_2 \ldots \mu_s$ 不全相同

记 μ 为总平均，记 $\alpha_j = \mu_j - \mu$，则

上述单因素模型可表示为$\begin{cases} X_{ij} = \mu + \alpha_j + \varepsilon_{ij} \\ \varepsilon_{ij} \sim N(0,\sigma^2) \end{cases}$

ANOVA的任务是检验在各水平下的均值是否相等，即$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_s$ ，$H_1 : \alpha_1, \alpha_2 \ldots \alpha_s$ 不全相同

1. One-way ANOVA

平方和分解

$$S_T = \sum_{j=1}^{s} \sum_{i=1}^{n_j} (X_{ij} - \overline{X})^2$$

$$= \sum_{j=1}^{s} \sum_{i=1}^{n_j} (X_{ij} - \overline{X}_{\cdot j})^2 + \sum_{j=1}^{s} \sum_{i=1}^{n_j} (X_{\cdot j} - \overline{X})^2$$

$$= S_E + S_A$$

$$S_E = \sum_{j=1}^{s} \sum_{i=1}^{n_j} (X_{ij} - \overline{X}_{\cdot j})^2 = \sum_{j=1}^{s} \sum_{i=1}^{n_j} (\varepsilon_{ij} - \overline{\varepsilon}_{\cdot j})^2$$

$$S_A = \sum_{j=1}^{s} \sum_{i=1}^{n_j} (\overline{X}_{\cdot j} - \overline{X})^2 = \sum_{j=1}^{s} n_j(\alpha_j + \overline{\varepsilon}_{\cdot j} - \overline{\varepsilon})^2$$

$$E(S_E) = (n - s)\sigma^2$$

$$E(S_A) = \sum_{j=1}^{s} n_j \alpha_j^2 + (s - 1)\sigma^2$$

1. One-way ANOVA

平方和分解

$$S_E = \sum_{j=1}^{s}\sum_{i=1}^{n_j}(X_{ij} - \overline{X_{\cdot j}})^2 = \sum_{j=1}^{s}\sum_{i=1}^{n_j}(\varepsilon_{ij} - \overline{\varepsilon_{\cdot j}})^2$$

$$E(S_E) = (n-s)\sigma^2$$

无条件

$$E\left(\frac{S_E}{n-s}\right) = \sigma^2$$

$$S_A = \sum_{j=1}^{s}\sum_{i=1}^{n_j}(\overline{X_{\cdot j}} - \overline{X})^2 = \sum_{j=1}^{s} n_j(\alpha_j + \overline{\varepsilon}_{\cdot j} - \overline{\varepsilon})^2$$

$$E(S_A) = \sum_{j=1}^{s} n_j\alpha_j^2 + (s-1)\sigma^2$$

当且仅当H0成立时

$$E\left(\frac{S_A}{s-1}\right) = \sigma^2$$

$$F = \frac{\frac{S_A}{(s-1)}}{\frac{S_E}{(n-s)}}$$

1. One-way ANOVA

F分布

$$E(S_E) = (n-s)\sigma^2$$

无条件

$$E\left(\frac{S_E}{n-s}\right) = \sigma^2$$

$$\frac{S_E}{\sigma^2} \sim \chi^2(n-2)$$

$$E(S_A) = \sum_{j=1}^{s} n_j \alpha_j^2 + (s-1)\sigma^2$$

当且仅当H0成立时

$$E\left(\frac{S_A}{s-1}\right) = \sigma^2$$

$$\frac{S_A}{\sigma^2} \sim \chi^2(s-1)$$

$$F = \frac{\frac{S_A}{(s-1)}}{\frac{S_E}{(n-s)}}$$

1. One-way ANOVA 方差分析表

方差来源	平方和	自由度	均方	F值
因素A	S_A	$s-1$	$\overline{S_A}=\dfrac{S_A}{s-1}$	$F=\dfrac{\overline{S_A}}{\overline{S_E}}$
误差	S_E	$n-s$	$\overline{S_E}=\dfrac{S_E}{n-s}$	
总和	S_T	$n-1$		

2. aov 函数 举例

```
#tell where the data come from
datafilename= "http://personality-project.org/R/datasets/R.appendix1.data"

#read the data into a table
data.ex1=read.table(datafilename,header=T)
```

Dosage	Alertness
a	30
a	38
a	35
a	41
a	27
a	24
b	32
b	26
b	31
b	29
b	27
b	35
b	21
b	25
c	17
c	21
c	20
c	19

2. aov 函数 举例

```
#do the analysis of variance
aov.ex1 = aov(Alertness~Dosage,data=data.ex1)

#show the summary table
summary(aov.ex1)
```

```
> summary(aov.ex1)
            Df Sum Sq Mean Sq F value  Pr(>F)
Dosage       2  426.2  213.12   8.789 0.00298 **
Residuals   15  363.8   24.25
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

2. aov 函数 举例

```
#report the means and the number of subjects/cell
print(model.tables(aov.ex1,"means"),digits=3)
```

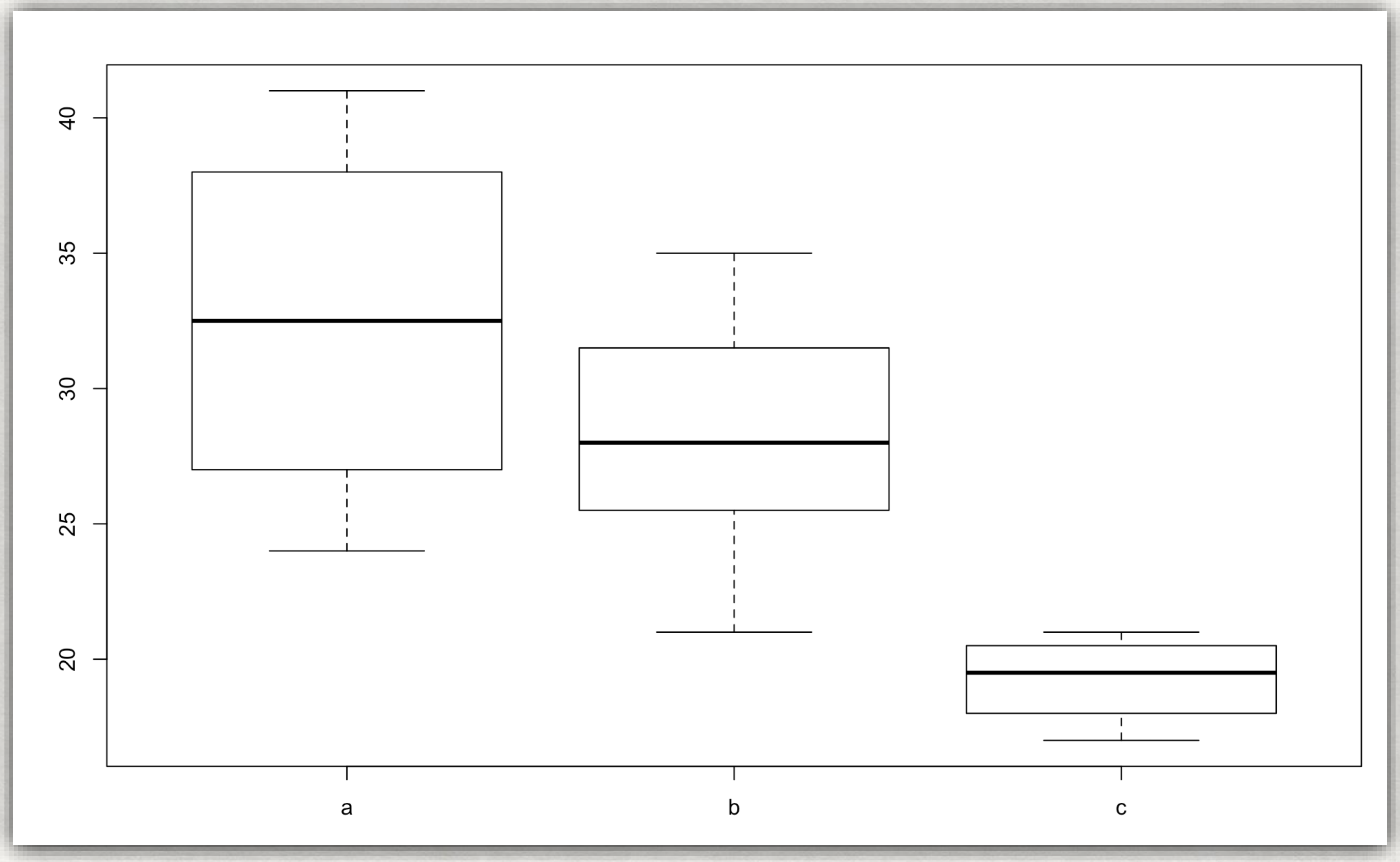
```
> print(model.tables(aov.ex1,"means"),digits=3)
Tables of means
Grand mean

27.66667

 Dosage
        a    b    c
     32.5 28.2 19.2
rep   6.0  8.0  4.0
```

2. aov 函数 举例

```
#graphical summary
boxplot(Alertness~Dosage,data=data.ex1)
```



3. 多重检验矫正

- Suppose the treatment is a new way of teaching writing to students, and the control is the standard way of teaching writing. Students in the two groups can be compared in terms of grammar, spelling, organization, content, and so on. As more attributes are compared, it becomes increasingly likely that the treatment and control groups will appear to differ on at least one attribute due to random sampling error alone.
- Suppose we consider the efficacy of a drug in terms of the reduction of any one of a number of disease symptoms. As more symptoms are considered, it becomes increasingly likely that the drug will appear to be an improvement over existing drugs in terms of at least one symptom.

$$\bar{\alpha} = 1 - \left(1 - \alpha_{\{\text{per comparison}\}}\right)^m.$$

4. p. adjust 函数　　　FWER

Bonferroni

Let $H_1, \ldots, H_m$ be a family of hypotheses and $p_1, \ldots, p_m$ their corresponding p-values. Let m be the total number of null hypotheses and m_0 the number of true null hypotheses. The familywise error rate (FWER) is the probability of rejecting at least one true H_i, that is, of making at least one type I error. The Bonferroni correction rejects the null hypothesis for each $p_i \leq \frac{\alpha}{m}$, thereby controlling the FWER at $\leq \alpha$. Proof of this control follows from Boole's inequality, as follows:

$$\text{FWER} = P\left\{\bigcup_{i=1}^{m_0}\left(p_i \leq \frac{\alpha}{m}\right)\right\} \leq \sum_{i=1}^{m_0}\left\{P\left(p_i \leq \frac{\alpha}{m}\right)\right\} = m_0\frac{\alpha}{m} \leq m\frac{\alpha}{m} = \alpha.$$

This control does not require any assumptions about dependence among the p-values or about how many of the null hypotheses are true.[6]

pairwise.t.test (x, A, p.adjust.method = "Bonferroni")

p.adjust(p, method = "Bonferroni")

4. p. adjust 函数　　FWER

Bonferroni

e. g.

```
> height<-data.frame(
+   X<-c(
+     176,178,159,165,167,
+     180,177,169,165,172,
+     168,174,162,156,167,
+     189,185,179,178,179
+   ),
+   A<-gl(4,5)
+ )
> pairwise.t.test(height$X,height$A,p.adjust.method = "bonferroni")

        Pairwise comparisons using t tests with pooled SD

data:  height$X and height$A

  1      2      3
2 1.0000 -      -
3 1.0000 0.5864 -
4 0.0352 0.2131 0.0055

P value adjustment method: bonferroni
```

4. p. adjust 函数 FWER Holm && Hochberg

The method is as follows:

- Let $H_1, \ldots, H_m$ be a family of hypotheses and $P_1, \ldots, P_m$ the corresponding P-values.
- Start by ordering the p-values (from lowest to highest) $P_{(1)} \ldots P_{(m)}$ and let the associated hypotheses be $H_{(1)} \ldots H_{(m)}$
- For a given significance level α, let k be the minimal index such that $P_{(k)} > \frac{\alpha}{m+1-k}$
- Reject the null hypotheses $H_{(1)} \ldots H_{(k-1)}$ and do not reject $H_{(k)} \ldots H_{(m)}$
- If $k = 1$ then do not reject any of the null hypotheses and if no such k exist then reject all of the null hypotheses.

The Holm–Bonferroni method ensures that this method will control the $FWER \leq \alpha$, where $FWER$ is the familywise error rate

1. Order the p-values $P(1), P(2), \ldots, P(n)$ and their associated hypotheses $H(1), \ldots, H(n)$
2. Reject all hypotheses $H(k)$ having $P(k) \leq \frac{\alpha}{n+1-k}$ where $k = 1, \ldots, n$

4. p. adjust 函数 HOMMEL

Let j be the largest integer for which

$$p_{n-j+k} > \frac{k\alpha}{j}$$

for all $k = 1, \ldots, j$.

If no such j exists, reject all hypotheses; otherwise, reject all H_i with $p_i \leq \frac{\alpha}{j}$. Both j and i, btw, go from 1 to n.

4. p.adjust 函数 FDR

BH / BY

The *Benjamini–Hochberg procedure* (BH step-up procedure) controls the FDR at level α.[1] It works as follows:

1. For a given α, find the largest k such that $P_{(k)} \leq \frac{k}{m}\alpha$.
2. Reject the null hypothesis (i.e., declare discoveries) for all $H_{(i)}$ for $i = 1, \ldots, k$.

- The *Benjamini–Hochberg–Yekutieli* procedure controls the false discovery rate under positive dependence assumptions.[13] This refinement modifies the threshold and

$$P_{(k)} \leq \frac{k}{m \cdot c(m)}\alpha$$

- If the tests are independent or positively correlated: $c(m) = 1$
- Under arbitrary dependence: $c(m) = \sum_{i=1}^{m} \frac{1}{i}$

In the case of negative correlation, $c(m)$ can be approximated by using the Euler–Mascheroni constant.

$$\sum_{i=1}^{m} \frac{1}{i} \approx \ln(m) + \gamma + \frac{1}{2m}.$$

4. p. adjust 函数　　　　　SUMMARY

Adjust P-values for Multiple Comparisons

Description

Given a set of p-values, returns p-values adjusted using one of several methods.

Usage

```
p.adjust(p, method = p.adjust.methods, n = length(p))

p.adjust.methods
# c("holm", "hochberg", "hommel", "bonferroni", "BH", "BY",
#   "fdr", "none")
```

Arguments

`p`	numeric vector of p-values (possibly with NAs). Any other R is coerced by `as.numeric`.
`method`	correction method. Can be abbreviated.
`n`	number of comparisons, must be at least `length(p)`; only set this (to non-default) when you know what you are doing!

Pre lab 1

Q1:

A large randomized trial compared an experimental drug and 9 other standard drugs for treating motion sickness. An ANOVA test revealed significant differences between the groups. The investigators wanted to know if the experimental drug ("drug 1") beat any of the standard drugs in reducing total minutes of nausea, and, if so, which ones. The p-values from the pairwise t tests (comparing drug 1 with drugs 2-10) are below.

Drug 1 vs. drug …	2	3	4	5	6	7	8	9	10
p-value	.05	.3	.25	.04	.001	.006	.08	.002	.01

a. Which differences would be considered statistically significant using a Bonferroni correction? or Holm-Hochberg correction?

Pre lab 2

Q2:

In the grade three of high school X, there are four kinds of classes using various teaching methods to teach mathematics. To identify whether the teaching method makes sense, five students' math scores are randomly chosen from the classes after the final exam.

Class 1	75	77	70	88	72
Class 2	83	80	85	90	84
Class 3	65	67	77	68	65
Class 4	72	70	71	65	82

Pre lab 1

- Answer:

```
> P_set<-data.frame(row.names =c(1:9),p_value=c(0.05,0.3,0.25,0.04,0.001,0.006,0.08,0.002,0.01))
> ordered_set<-data.frame(row.names = order(P_set$p_value),original_p = P_set[order(P_set$p_value),])
> ordered_set$bonferroni<-p.adjust(ordered_set$original_p,method = "bonferroni" )
> ordered_set$holm<-p.adjust(ordered_set$original_p,method = "holm" )
> ordered_set$hochberg<-p.adjust(ordered_set$original_p,method = "hochberg" )
> ordered_set
  original_p bonferroni  holm hochberg
5      0.001      0.009 0.009    0.009
8      0.002      0.018 0.016    0.016
6      0.006      0.054 0.042    0.042
9      0.010      0.090 0.060    0.060
4      0.040      0.360 0.200    0.200
1      0.050      0.450 0.200    0.200
7      0.080      0.720 0.240    0.240
3      0.250      1.000 0.500    0.300
2      0.300      1.000 0.500    0.300
```

Pre lab 2

- Answer:

```
> math<-data.frame(
+   X<-c(75,77,70,88,72,
+        83,80,85,90,84,
+        65,67,77,68,65,
+        72,70,71,65,82
+        ),
+   A = gl(4,5)
+ )
> summary(aov(X~A,data = math))
            Df Sum Sq Mean Sq F value  Pr(>F)
A            3  712.6  237.53   7.547 0.00229 **
Residuals   16  503.6   31.47
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Our question 1 p.adjust

```
Input = ("
Food               Raw.p
 Blue_fish         .34
 Bread             .594
 Butter            .212
 Carbohydrates     .384
 Cereals_and_pasta .074
 Dairy_products    .94
 Eggs              .275
 Fats              .696
 Fruit             .269
 Legumes           .341
 Nuts              .06
 Olive_oil         .008
 Potatoes          .569
 Processed_meat    .986
Proteins           .042
Red_meat           .251
Semi-skimmed_milk  .942
```

You need to order the data by their p value and adjust their p value with Bonferroni, BH, Holm, Hochberg, Hommel, BY.

a) Save your result in a table (see format below)

b) Plot new p — old p for all the methods in the same coordinate.

Food	raw.p	Beferroni	BH	Holm	Hochberg	Hommel	BY

Our question 2 anova

Using the following data, perform a oneway analysis of variance using α =.05. Write up the results in the following format.

Group1	Group2	Group3
51	23	56
45	43	76
33	23	74
45	43	87
67	45	56

ANOVA table

source	SS	df	MS	F
group	[redacted]	[redacted]	[redacted]	[redacted]
error	[redacted]	[redacted]	[redacted]	
total	[redacted]			

Effect size

η^2 = [redacted] = [redacted]

η^2相当于R^2

APA writeup

F([redacted])=[redacted], *p* [redacted].05, η^2=[redacted].