Applied Statistics - Homework 5

Sai Nikhil
0D1564864

(1)

This is an observational study.

There we cannot comment on Causation.

Hence, the conclusion that a woman should postpone childbearing until later in life to ensure a high IQ for her offspring seems inappropriate.

[P.T.O.]

(2)

12.4.8

n	$\bar{x}$	s
73	6.22	1.62
105	5.81	1.43
240	5.77	1.24
1080	5.47	1.31

a) $$s_w^2 = \frac{(73-1)\,1.62^2 + (105-1)\,1.43^2 + (240-1)\,1.24^2 + (1080-1)\,1.31^2}{73+105+240+1080-4}$$

$$\boxed{s_w^2 \simeq 1.75}$$

$$\bar{X} = \frac{73 \times 6.22 + 105 \times 5.81 + 240 \times 5.77 + 1080 \times 5.47}{73+105+240+1080}$$

$$\boxed{\bar{X} \simeq 5.58}$$

$$S_B^2 = \frac{73(6.22-5.58)^2 + 105(5.81-5.58)^2 + 240(5.77-5.58)^2 + 1080(5.47-5.58)^2}{4-1}$$

$$\boxed{S_B^2 \simeq 19.06}$$

$$F = \frac{S_B^2}{S_W^2} = \frac{19.06}{1.75} \simeq 10.89$$

$k-1 = 4-1 = 3$ and $n-k = 1498-4 = 1494$

b) $F_{3,1494} \simeq 2.61$, $P = 0.05$

$F_{3,1494} \simeq 5.43$, $P = 0.001$

$\Rightarrow \quad P < 0.001$

∴ We reject $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$

$\Rightarrow$ Mean changes in LDL cholesterol levels are different among four different populations

c)

Assumptions:

1) Each sample is an independent random Sample.

2) Samples are drawn from normal populations.

3) Variances of populations are equal.

d) $H_0: \mu_1 + \mu_2 + \mu_3 - 3\mu_4 = 0$

$$\vec{L} = \vec{X_1} + \vec{X_2} + \vec{X_3} - 3\vec{X_4} = 1.39$$

$$Var(\vec{L}) = 1.754 \times \left(\frac{1}{73} + \frac{1}{205} + \frac{1}{240} + \frac{(-3)^2}{1080} \right)$$

$$\simeq 0.0627$$

$$t = \left| \vec{L} / \sqrt{Var(\vec{L})} \right| = \left| \frac{1.39}{\sqrt{0.0627}} \right| = 5.55$$

$$t_{1494,\, 0.0975} = 1.96 < 5.55$$

$\Rightarrow$ Reject H_0 & conclude that cholesterol levels for people with no disease is different from people with disease.

The conclusion doesn't Change with Bonferroni because the α-level is same.

Scheffe's method:

$$T_{obs} = 5.55 > \sqrt{(k-1) F_{k-1, n-k, \alpha}}$$

$$= \sqrt{3 \times 2.61} = 2.798$$

$$T_{obs} > 2.388$$

$\Rightarrow$ Reject H_0. No Correction is required as conclusion is same from Bonferroni's & Scheffe's

e) $$H_0 : \mu_1 - \left(\frac{\mu_2 + \mu_3 + \mu_4}{3} \right) = 0$$

$$\vec{L} = \vec{x_1} - \left(\frac{\bar{x_2} + \bar{x_3} + \bar{x_4}}{3} \right) = 0.5367$$

$$\Rightarrow Var\left(\vec{L}\right) = 0.0268$$

$$t_{obs} = \left| \frac{0.5367}{\sqrt{0.0268}} \right| = 3.27 > 1.96 \Rightarrow \text{Reject } H_0$$

Multiple testing adjustment is needed as it is a post-hoc contrast.

Scheffe's method:

$$T_{obs} = 3.27 > 2.798$$

We reject H_0 and conclude that the group with intermittent claudation has higher LDL Cholesterol levels than other groups.

3

a)

```
> attach(airquality)
> airquality[is.na(airquality)] = 0
>
>
> pairwise.t.test(Ozone, Month, p.adjust.method = "bonf")

   Pairwise comparisons using t tests with pooled SD

data:  Ozone and Month

  5       6       7       8
6 1.00000 -       -       -
7 0.00029 0.10225 -       -
8 0.00019 0.08312 1.00000 -
9 1.00000 1.00000 0.00697 0.00485

P value adjustment method: bonferroni
> pairwise.t.test(Ozone, Month, p.adjust.method = "fdr")

   Pairwise comparisons using t tests with pooled SD

data:  Ozone and Month

  5       6       7       8
6 0.76096 -       -       -
7 0.00015 0.01704 -       -
8 0.00015 0.01662 0.91744 -
9 0.46493 0.91744 0.00174 0.00162

P value adjustment method: fdr
>
> airquality$Month<-as.factor(airquality$Month)
> p<-aov(Ozone~Month,data=airquality)
> TukeyHSD(p,conf.level = 0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Ozone ~ Month, data = airquality)

$Month
           diff          lwr       upr     p adj
6-5 -10.9731183 -32.27095900 10.324722 0.6139469
7-5  29.7741935   8.65164668 50.896740 0.0013894
8-5  30.4838710   9.36132410 51.606418 0.0009868
9-5  10.5935484 -10.70429233 31.891389 0.6454439
7-6  40.7473118  19.44947111 62.045153 0.0000044
8-6  41.4569892  20.15914853 62.754830 0.0000029
9-6  21.5666667   0.09496314 43.038370 0.0484120
8-7   0.7096774 -20.41286945 21.832224 0.9999830
9-7 -19.1806452 -40.47848588  2.117196 0.0990957
9-8 -19.8903226 -41.18816330  1.407518 0.0795001
```

Bonferroni

FDR

Tukey

For

Bonferroni, May-July, May-August, July-September, August-September

FDR, May-July, May-August, July-September, August-September,

Tukey, May-July, May-August, July-September, August-September, June-July, June-August

the mean ozone levels are different

b)

Bonferroni & Tukey gives same result. FDR is different.

Tukey is more appropriate in this case as the number of comparisons is large.

(4)

a)

```
> data <- read.csv(file="lowbwt.csv", header = TRUE, sep = ",")
> data$sex<-as.factor(data$sex)
> data$tox<-as.factor(data$tox)
> summary(aov(sbp~sex+tox, data=data))
            Df Sum Sq Mean Sq F value Pr(>F)
sex          1     48   48.25   0.367  0.546
tox          1     67   66.76   0.508  0.478
Residuals   97  12758  131.53
> summary(aov(sbp~sex,data=data))
            Df Sum Sq Mean Sq F value Pr(>F)
sex          1     48   48.25   0.369  0.545
Residuals   98  12825  130.87
>
```

b) p-value for gender

with blocking : 0.546

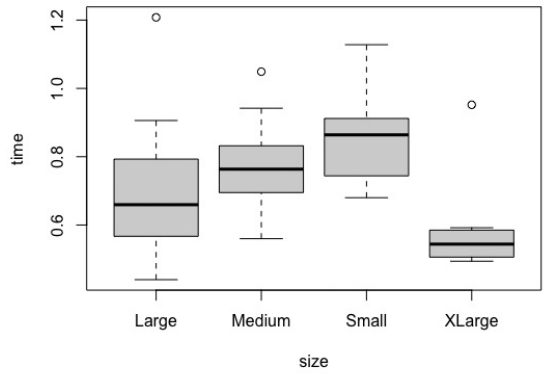
without blocking : 0.545

c) Yes, ANOVA without blocking is same as independent two-sample t-test.

Also,
ANOVA with blocking is same as paired two-sample test.

5

```
> library(agricolae)
> rt <- read.csv("response_times.csv", header = TRUE, sep = ",")
> boxplot(time~size, data = rt)
```




```
> res.aov <- aov(time ~ size, data = rt)
> summary(res.aov)
            Df Sum Sq Mean Sq F value  Pr(>F)
size         3 0.3826  0.1275   4.705 0.00716 **
Residuals   56 0.9757  0.0271
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> TukeyHSD(res.aov, confidence.level=0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = time ~ size, data = rt)

$size
                  diff        lwr          upr     p adj
Medium-Large    0.0580 -0.1402918  0.256291795 0.8595149
Small-Large     0.1367 -0.0615918  0.334991795 0.2644982
XLarge-Large   -0.1312 -0.3294918  0.067091795 0.2983930
Small-Medium    0.0787 -0.1195918  0.276991795 0.7103448
XLarge-Medium -0.1892 -0.3874918  0.009091795 0.0660379
XLarge-Small  -0.2679 -0.4661918 -0.069608205 0.0045301

> pairwise.t.test(rt$time, g=rt$size, p.adjust.method = 'none')

   Pairwise comparisons using t tests with pooled SD

data:  rt$time and rt$size

       Large   Medium  Small
Medium 0.43599 -       -
Small  0.07155 0.29222 -
XLarge 0.08319 0.01446 0.00085

P value adjustment method: none
```

we fail to reject H_0 in both cases.