Northeastern
College of Science

Applied Statistics
Homework-5

Anjali Patil
NUID: -00109806
3/15/2021

PROBLEM 1:-

Applied Statistics

HW-5

Anjali Patil

NUID 001074806

3/13/21

Q1) I don't think this conclusion is appropriate. We can say that older mothers & higher IQ are associated, But I think we should also consider other factors that might help depend on the higher IQ. We can also take into account factors like education ~~to~~ quality, medical facility rather than just considering age.

PROBLEM 2:-

Q.2) 12.4.8 :-

$n_1 = 73$ $\qquad \bar{x}_1 = 6.22$ $\qquad s_1 = 1.62$

$n_2 = 105$ $\qquad \bar{x}_2 = 5.81$ $\qquad s_2 = 1.43$

$n_3 = 240$ $\qquad \bar{x}_3 = 5.77$ $\qquad s_3 = 1.24$

$n_4 = 1080$ $\qquad \bar{x}_4 = 5.47$ $\qquad s_4 = 1.31$

a)

$$s_w^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_3-1)s_3^2 + (n_4-1)s_4^2}{n_1+n_2+n_3+n_4-4}$$

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

$$s_w^2 = 1.75$$

Now, $$\bar{X} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3 + n_4\bar{x}_4}{n_1+n_2+n_3+n_4}$$

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

$$= 5.58$$

$$S_B^2 = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + n_3(\bar{x}_3 - \bar{x})^2 + n_4(\bar{x}_4 - \bar{x})^2}{k-1}$$

$$= \frac{73(6.22-5.88)^2 + 105(5.81-5.58)^2 + 240(5.77-5.58)^2 + 1080(5.67-5.58)^2}{3}$$

$$= \underline{\underline{19.06}}$$

$$F = \frac{S_B^2}{S^2 w} = \frac{19.06}{1.75}$$

$$= \underline{\underline{10.89}}$$

Degree of freedom $= k-1 = 4-1$

$$= \underline{\underline{3}}$$

and $n-k = 1498-4$

$$= \underline{\underline{1494}}$$

b) For ~~$F_{3,1494}$~~

b) From the table, the critical value,

$F_{3,1494} = 2.6108$ at signiface level $= 0.05$

$\Rightarrow$ P-value < 0.001

Hence we reject H_0, so we conclude that the mean LDL cholestrol is differenet from the four population.

c) The assumptions are:-

The data within one group represents a random sample from a ~~po~~ population. The population are independent within the group, the observations are normally distributed with mean μ Variance σ^2 is same for all groups.

PART D:- R CODE AND OUTPUT:-

```
library("readxl")
library("psych")
#contr1 <- c(1,-1/3,-1/3, -1/3)
contr1 <- c(1, 1,1,-3)
ldl.data <- read_excel("hw5.xlsx")
k<-4
(contr1.est <- sum(contr1*ldl.data[,'mean']))
(MSE <- sum((ldl.data[,'n']-1)*ldl.data[,'sd']^2)/sum(ldl.data[,'n']-1)
contr1.se <- sqrt(sum(contr1^2/ldl.data[,'n'])*MSE)
contr1.t <- contr1.est/contr1.se
contr1.p <-2*pt(-abs(contr1.t),df=sum(ldl.data[,'n'])-k)
c(contr1.est,contr1.se, contr1.t,contr1.p)
```

13:1 (Top Level)

Console Terminal Jobs

~/LAB1.1/

```
> library("readxl")
> library("psych")
> #contr1 <- c(1,-1/3,-1/3, -1/3)
> contr1 <- c(1, 1,1,-3)
> ldl.data <- read_excel("hw5.xlsx")
> k<-4
> (contr1.est <- sum(contr1*ldl.data[,'mean']))
[1] 1.39
> (MSE <- sum((ldl.data[,'n']-1)*ldl.data[,'sd']^2)/sum(ldl.data[,'n']-1))
[1] 1.754207
> contr1.se <- sqrt(sum(contr1^2/ldl.data[,'n'])*MSE)
> contr1.t <- contr1.est/contr1.se
> contr1.p <-2*pt(-abs(contr1.t),df=sum(ldl.data[,'n'])-k)
> c(contr1.est,contr1.se, contr1.t,contr1.p)
[1] 1.390000e+00 2.503289e-01 5.552696e+00 3.324252e-08
>
```

d) From R,

est	se	t.test	p value
1.39	2.50×10^{-1}	5.55	3.324×10^{-8}

Reject H_0 at $\alpha = 0.05$, since $3.324 \times 10^{-8} < 0.05$.

With Bonferoni,

$$= \frac{0.05}{\binom{4}{2}} = 0.0083$$

Hence we still reject.

For Scheffe,

$t_{1494, 0.25} = -1.96$ & $5.55 > -1.96$

we reject.

Hence No correction at all.

PART E:- CODE AND OUTPUT:-

```
library("readxl")
library("psych")
contr1 <- c(1,-1/3,-1/3, -1/3)
#contr1 <- c(1, 1,1,-3)
ldl.data <- read_excel("hw5.xlsx")
k<-4
(contr1.est <- sum(contr1*ldl.data[,'mean']))
(MSE <- sum((ldl.data[,'n']-1)*ldl.data[,'sd']^2)/sum(ldl.data[,'n
contr1.se <- sqrt(sum(contr1^2/ldl.data[,'n'])*MSE)
contr1.t <- contr1.est/contr1.se
contr1.p <-2*pt(-abs(contr1.t),df=sum(ldl.data[,'n'])-k)
c(contr1.est,contr1.se, contr1.t,contr1.p)
```

13:1 (Top Level)

onsole | Terminal × | Jobs ×

-/LAB1.1/

```
library("readxl")
library("psych")
contr1 <- c(1,-1/3,-1/3, -1/3)
#contr1 <- c(1, 1,1,-3)
ldl.data <- read_excel("hw5.xlsx")
k<-4
(contr1.est <- sum(contr1*ldl.data[,'mean']))
1] 0.5366667
(MSE <- sum((ldl.data[,'n']-1)*ldl.data[,'sd']^2)/sum(ldl.data[,'n']-1
1] 1.754207
contr1.se <- sqrt(sum(contr1^2/ldl.data[,'n'])*MSE)
contr1.t <- contr1.est/contr1.se
contr1.p <-2*pt(-abs(contr1.t),df=sum(ldl.data[,'n'])-k)
c(contr1.est,contr1.se, contr1.t,contr1.p)
1] 0.536666667 0.163948582 3.273384000 0.001086971
```

e) From R,

est	se	t.test	p.value
0.536	0.16	3.27	0.0010

Because $0.0010 < 0.05$, we reject H_0

For Bonferri, as above $0.0083 > p$, so we again reject.

For Scheffe, $t_{1494, 0.25} = -1.61$, so we still reject as $-1.61 < 3.27$.

Hence we reject the null hypothsis

PROBLEM 3:-

CODE:-

```
attach(airquality)
airquality[is.na(airquality)] = 0


pairwise.t.test(Ozone, Month, p.adjust.method = "bonf")

pairwise.t.test(Ozone, Month, p.adjust.method = "fdr")

airquality$Month<-as.factor(airquality$Month)
p<-aov(Ozone~Month,data=airquality)
TukeyHSD(p,conf.level = 0.95)

detach()
```

OUTPUT:-

Bonferroni:-

```
> 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 1.0000 -       -      -
7 0.0015 4.4e-06 -      -
8 0.0011 2.9e-06 1.0000 -
9 1.0000 0.0625  0.1399 0.1088

P value adjustment method: bonferroni
```

FDR:-

```
> 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 0.19069 -       -       -
7 0.00037 2.2e-06 -       -
8 0.00035 2.2e-06 0.92620 -
9 0.19069 0.01251 0.01998 0.01813

P value adjustment method: fdr
> airquality$Month<-as.factor(airquality$Month)
```

TUKEY:-

```
  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
```

No, the results are not the same. I think Tukey is the more appropriate conclusion because we are using a pairwise comparison.

PROBLEM 4:-

CODE:-

```
#import data set
data <- read.table(file="lowbwt.txt", header = TRUE)

data$sex<-as.factor(data$sex)
data$tox<-as.factor(data$tox)

##Anova with bocking
x <- aov(sbp~sex+tox, data=data)
summary(x)

##Anova without bocking
x<-aov(sbp~sex,data=data)
summary((x))

```

OUTPUT:-

```
> data$tox<-as.factor(data$tox)
> ##Anova with bocking
> x <- aov(sbp~sex+tox, data=data)
> summary(x)
            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
> ##Anova without bocking
```

```
> ##Anova without bocking
> x<-aov(sbp~sex,data=data)
> summary((x))
            Df Sum Sq Mean Sq F value Pr(>F)
sex          1     48   48.25   0.369  0.545
Residuals   98  12825  130.87
> 
```

a) **We fail to reject Null hypthosis at .05 significant level. There is no significant difference for the mean systolic blood pressure for low birth weight boys and girls.**
b) **p-value =0.5461 for gender effects on blood pressure with blocking**

p-value =0.5451 for gender effects on blood pressure without blocking.

c) **The ANOVA F test without blocking equivalent to the independent two sample test. The ANOVA F-test with blocking is same as paired two sample test.**

PROBLEM 5:-

CODE for a,b,c:-

```
rt <- read.csv("response_times.csv", header = TRUE, sep = ",")
#a
anova_result<-aov(time~size, data=rt)
summary(anova_result)

#b
TukeyHSD(anova_result)

#c

pairwise.t.test(rt$time,rt$size, p.adjust.method = "none")

```

PART A):-

```
> rt <- read.csv("response_times.csv", header = TRUE, sep = ",")
> #a
> anova_result<-aov(time~size, data=rt)
> summary(anova_result)
            Df Sum Sq Mean Sq F value  Pr(>F)
size         3  1.929  0.6431   4.234 0.00882 **
Residuals   60  9.113  0.1519
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #b
```

Based on the ANOVA test, we fail to reject the null hypothesis. There is no significant difference among response times.

PART B) : -

HSD :-

```
> #b
> TukeyHSD(anova_result)
  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   0.1295625 -0.23454756  0.49367256 0.7833229
Small-Large    0.3094375 -0.05467256  0.67354756 0.1227683
XLarge-Large  -0.1641250 -0.52823506  0.19998506 0.6347743
Small-Medium   0.1798750 -0.18423506  0.54398506 0.5631358
XLarge-Medium -0.2936875 -0.65779756  0.07042256 0.1549644
XLarge-Small  -0.4735625 -0.83767256 -0.10945244 0.0057723
```

LSD :-

```
> pairwise.t.test(rt$time,rt$size, p.adjust.method = "none")

        Pairwise comparisons using t tests with pooled SD

data:  rt$time and rt$size

       Large  Medium Small
Medium 0.3508 -      -
Small  0.0284 0.1967 -
XLarge 0.2383 0.0372 0.0011

P value adjustment method: none
>
```

We came to same conclusion in both tests. Fail to reject null hypothesis.