Homework Assignment 2

Shrusti Ghela

February 09, 2022

Data: ‘lead.csv’

The data are from a study of the association between exposure to lead and IQ. The study was conducted in an urban area around a lead smelter. A random sample of 124 children who lived in the area was selected. Each study participant had a blood sample drawn in both 1972 and 1973 to assess blood concentrations of lead. The children were grouped based on their blood concentrations as follows:

Group 1: concentration < 40 mg/L in both 1972 and 1973 Group 2: concentration > 40 mg/L in both 1972 and 1973 or > 40 mg/L in 1973 alone (3 participants) Group 3: concentration > 40 mg/L in 1972 but < 40 mg/L in 1973

Each participant completed an IQ test in 1973. (A subset of the IQ scores from this study were used in HW 1, Question 3.) The variables in the data set are listed below.

ID: Participant identification number SEX: Participant sex (1=M or 2=F) GROUP: As described above (1, 2, or 3) IQ: IQ score

```
lead <- read.csv("lead_study.csv")
```

1. The first goal is to compare the mean IQ scores for males and females. Use a 2-sample t-test for this comparison. What is the p-value?

```
m = with(lead, tapply(IQ, SEX,mean))
s = with(lead, tapply(IQ, SEX,sd))
n = with(lead, tapply(IQ, SEX,length))
data.frame(m,s,n)
```

```
##          m        s  n
## 1 91.23684 14.93083 76
## 2 90.87671 13.58507 73
```

Now to compare the mean IQ squares for 1 and 2.

```
with(lead, t.test(IQ[SEX==1], IQ[SEX==2], var.equal=T))
```

```
##
##  Two Sample t-test
##
## data:  IQ[SEX == 1] and IQ[SEX == 2]
## t = 0.15381, df = 147, p-value = 0.878
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -4.267092  4.987351
## sample estimates:
## mean of x mean of y
##  91.23684  90.87671
```

2. State the conclusion from your test.

The p-value is greater than 0.05, we would not reject the null hypothesis for equal means for males and females.

3. Are the independence assumptions valid for the t-test in this situation? Give a brief explanation. Yes, the independence assumptions would be valid in this situation as the data for males and females are independent of each other, where data for one group would not affect the other group.

4. The second goal is to compare the mean IQ scores in the 3 groups. State in words the null hypothesis for this test.

The null hypothesis for this case could be defined as the mean IQ scoreS for the 3 groups are equal.

5. State in words the alternative hypothesis for this test.

We define the alternative hypothesis as: Not all mean IQ scores for 3 groups are equal.

6. What method should be used to perform the test?

ANOVA ANOVA is simply an extension of the 2-sample equal-variance t-test to the comparison of 3 or more population means. Since we have 3 groups to test, ANOVA would be the best choice.

7. Perform the test. Report the p-value.

```
summary(aov(lead$IQ~lead$GROUP, data=lead))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)  
## lead$GROUP    1   1321  1320.6   6.766 0.0102 *
## Residuals   147  28692   195.2                 
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

8. State your conclusion about the evidence for an association between lead exposure and IQ.

The p-value is less than 0.05, and so we reject the null hypothesis of equal means for 3 groups. This means that there is an association between lead and IQ scores.

9. Are there strong reasons to believe that the assumptions of this test are not met? Briefly justify your answer.

```
M = with(lead, tapply(IQ, GROUP, mean))
S = with(lead, tapply(IQ, GROUP, sd))
N = with(lead, tapply(IQ, GROUP, length))
data.frame(M,S,N)
```

```
##          M         S  N
## 1 93.72414 15.570313 87
## 2 87.65625  9.502493 32
## 3 86.96667 12.962767 30
```

From above, we can clearly say that the equal variance assumption is not met. So, not all the required assumptions for ANOVA are met.

10. Conduct all pairwise comparison of group means. Report the p-values.

```
A.vs.B=t.test(lead$IQ[lead$GROUP==1],
lead$IQ[lead$GROUP==2],var.equal=F)
A.vs.B$p.value
```

```
## [1] 0.0120506
```

```
A.vs.C = t.test(lead$IQ[lead$GROUP==1],
lead$IQ[lead$GROUP==3],var.equal=F)
A.vs.C$p.value
```

```
## [1] 0.02300319
```

```
B.vs.C = t.test(lead$IQ[lead$GROUP==2],
lead$IQ[lead$GROUP==3],var.equal=F)
B.vs.C$p.value
```

```
## [1] 0.8131036
```

11. What conclusion about the association between lead and IQ would you draw from the pairwise comparisons of group means? Does it agree with the conclusion in Q8? (Consider the issue of multiple testing in your answer.)

Considering the issue of mutltiple testing, we use Bonferroni correction. And we change the significance level for each individual test to be 0.05/3 so that the overall significance level is maintained to 0.05. So, applying that the p-values calculated above are not all greater than 0.01667, and so we would reject the null hypothesis of equal means for 3 groups. This means that there is an association between lead and IQ scores.

12. Now we wish to compare the 3 group means for males and females separately. Display some appropriate descriptive statistics for this analysis.

```
m1 = with(lead, tapply(IQ[lead$SEX==1], GROUP[lead$SEX==1], mean))
s1 = with(lead, tapply(IQ[lead$SEX==1], GROUP[lead$SEX==1], sd))
n1 = with(lead, tapply(IQ[lead$SEX==1], GROUP[lead$SEX==1], length))
m2 = with(lead, tapply(IQ[lead$SEX==2], GROUP[lead$SEX==2], mean))
s2 = with(lead, tapply(IQ[lead$SEX==2], GROUP[lead$SEX==2], sd))
n2 = with(lead, tapply(IQ[lead$SEX==2], GROUP[lead$SEX==2], length))

data.frame(m1,s1,n1,m2,s2,n2)
```

```
##         m1       s1 n1       m2        s2 n2
## 1 92.93478 15.42351 46 94.60976 15.877465 41
## 2 90.17647 11.13124 17 84.80000  6.471917 15
## 3 86.61538 17.32791 13 87.23529  8.898942 17
```

13. Perform tests to compare the mean IQ scores in the 3 groups for males and females separately. Report the p-values from the two tests.

```
summary(aov(lead$IQ[lead$SEX==1]~lead$GROUP[lead$SEX==1], data=lead))
```

```
##                            Df Sum Sq Mean Sq F value Pr(>F)
## lead$GROUP[lead$SEX == 1]   1    427   427.5   1.942  0.168
## Residuals                  74  16292   220.2
```

```
summary(aov(lead$IQ[lead$SEX==2]~lead$GROUP[lead$SEX==2], data=lead))
```

```
##                              Df Sum Sq Mean Sq F value Pr(>F)  
## lead$GROUP[lead$SEX == 2]  1    922   922.1   5.295 0.0243 *
## Residuals                 71  12366   174.2                 
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

14. What can you conclude about the association between lead and IQ from these tests? Does it agree with the result in Q8 and Q11? (Consider multiple testing.) We see that the null hypothesis of equal means of three groups for male(1) is not rejected. However, the null hypothesis of equal means of three groups for female(2) is rejected. This means that there is an association between lead and IQ for females and there is no association between lead and IQ for males. When we performed ANOVA to test the null hypothesis of equal means of three groups, we reject the null hypothesis Which means that there is an association between lead and IQ (Q8) When we performed pairwise test to test the null hypothesis of equla means of three groups, we reject the null hypothesis which means that there is an association between lead and IQ (Q11)

15. Now perform all 3 pairwise comparisons of groups for males and females separately. Report the p-values from these tests?

```
Am.vs.Bm = t.test(lead$IQ[lead$GROUP==1 & lead$SEX==1],
lead$IQ[lead$GROUP==2 & lead$SEX==1],var.equal=F)
Am.vs.Bm$p.value
```

```
## [1] 0.4391938
```

```
Am.vs.Cm = t.test(lead$IQ[lead$GROUP==1 & lead$SEX==1],
lead$IQ[lead$GROUP==3 & lead$SEX==1],var.equal=F)
Am.vs.Cm$p.value
```

```
## [1] 0.2502756
```

```
Bm.vs.Cm = t.test(lead$IQ[lead$GROUP==2 & lead$SEX==1],
lead$IQ[lead$GROUP==3 & lead$SEX==1],var.equal=F)
Bm.vs.Cm$p.value
```

```
## [1] 0.5258587
```

```
Af.vs.Bf = t.test(lead$IQ[lead$GROUP==1 & lead$SEX==2],
lead$IQ[lead$GROUP==2 & lead$SEX==2],var.equal=F)
Af.vs.Bf$p.value
```

```
## [1] 0.001830775
```

```
Af.vs.Cf = t.test(lead$IQ[lead$GROUP==1 & lead$SEX==2],
lead$IQ[lead$GROUP==3 & lead$SEX==2],var.equal=F)
Af.vs.Cf$p.value
```

```
## [1] 0.02927457
```

```
Bf.vs.Cf = t.test(lead$IQ[lead$GROUP==2 & lead$SEX==2],
lead$IQ[lead$GROUP==3 & lead$SEX==2],var.equal=F)
Bf.vs.Cf$p.value
```

```
## [1] 0.3796352
```

16. What do you conclude about the association between lead and IQ from the results in Q13? Does your conclusion change from previous conclusions made in Q8, Q11 and Q14?

From Q15, the p-values for pairwise test for three groups for males are greater than 0.05/3, so we do not have enough evidence to reject the null hypothesis of equal means for three groups for males at 0.05 significance level. This means that there is no association between lead and IQ for males. However, the p-values for pairwise test for three groups for females are not all greater than 0.05/3, so we reject the null hypothesis of equal means for three groups for females at 0.05 significance level. This means that there is an association between lead and IQ for females. From Q14, we arrived at a similar conclusion where, there is an association between lead and IQ for females and there is no association between lead and IQ for males. From Q11, we arrived at a result that there is an association between lead and IQ. From Q8, we arrived at a result that there is an association between lead and IQ. So, our conclsuion changes from previous questions for males, however it doesn't change for females.