DATA 557 Homework Assignment 3

Anuhya B S

February 5, 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

```
leadData <- 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?

For the goal to compare the mean IQ scores, the null hypothesis is:

$$H_0: \mu_M = \mu_F$$

where μ_M is the mean IQ score for male and μ_F is the mean IQ score for female.

```
m = with(leadData,tapply(IQ, SEX, mean))
s = with(leadData,tapply(IQ, SEX, sd))
n = with(leadData,tapply(IQ, SEX, length))
data.frame(m,s,n,(s^2))
```

```
##          m        s  n   X.s.2.
## 1 91.23684 14.93083 76 222.9298
## 2 90.87671 13.58507 73 184.5540
```

Since the variances are nearly equal, we consider the 2 sample equal variance T-test.

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

```
##
##  Two Sample t-test
##
## data:  leadData$IQ[leadData$SEX == 1] and leadData$IQ[leadData$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
```

The p-value is **0.878**.

2. State the conclusion from your test.

Since the p-value is greater than the level of significance 0.05, we **do not have enough evidence to reject the hull hypothesis** of equal mean IQ scores for males and females.

3. Are the independence assumptions valid for the t-test in this situation? Give a brief explanation.

The independence assumptions are valid for the t-test in this case as the group of male and female are independent of each other. The data collected is not paired so the data collected from one group would not have an effect on 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 is:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

where μ_1, μ_2, μ_3 are the mean IQ scores of group 1, 2 and 3 respectively. The null hypothesis is that the mean IQ scores are all equal for groups 1, 2 and 3.

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

The alternative hypothesis is that the mean IQ scores are not all equal for groups 1, 2 and 3.

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

```
m = with(leadData,tapply(IQ, GROUP, mean))
s = with(leadData,tapply(IQ, GROUP, sd))
n = with(leadData,tapply(IQ, GROUP, length))
data.frame(m,s,n,(s^2))
```

```
##          m         s  n    X.s.2.
## 1 93.72414 15.570313 87 242.43464
## 2 87.65625  9.502493 32  90.29738
## 3 86.96667 12.962767 30 168.03333
```

The most appropriate test to perform the comparison of two or more groups is the **ANOVA test**. In this case, we use the ANOVA test to test equal mean IQ scores in all 3 groups.

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

```
summary(aov(leadData$IQ~leadData$GROUP))
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## leadData$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
```

The p-value is **0.0102**.

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

We **reject the null hypothesis** as the p-value is much smaller than the value of level of significance (0.05). There is an association between lead exposure and IQ.

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

All the following assumptions must be met for ANOVA test:

1. Independence (of samples and of observations within each sample)
2. Equal variances
3. Large sample sizes or normal distributions

For the previous test, all the three groups do not have equal variances (242,90,168 - seen as a part of Q7). Thus there is not strong reason to believe the assumption of the test is not met. We can assume that the data for Group 1, Group 2 and Group 3 are independent of each other, however since not all the assumptions of ANOVA test are met, the F-test may not have the right type I error probability.

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

Since the variances are not relatively equal for all three groups, we would perform the Welch t-test for the pairwise comparison.

```
p12 = t.test(leadData$IQ[leadData$GROUP=='1'],leadData$IQ[leadData$GROUP=='2'
],var.equal = F)
p12
```

```
##
##  Welch Two Sample t-test
##
## data:  leadData$IQ[leadData$GROUP == "1"] and leadData$IQ[leadData$GROUP =
= "2"]
```

```
## t = 2.5622, df = 90.607, p-value = 0.01205
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   1.363464 10.772312
## sample estimates:
## mean of x mean of y
##  93.72414  87.65625
```

```
p13 = t.test(leadData$IQ[leadData$GROUP=='1'],leadData$IQ[leadData$GROUP=='3'
],var.equal = F)
p13
```

```
##
##  Welch Two Sample t-test
##
## data:  leadData$IQ[leadData$GROUP == "1"] and leadData$IQ[leadData$GROUP =
= "3"]
## t = 2.3333, df = 60.024, p-value = 0.023
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.9643448 12.5505977
## sample estimates:
## mean of x mean of y
##  93.72414  86.96667
```

```
p23 = t.test(leadData$IQ[leadData$GROUP=='2'],leadData$IQ[leadData$GROUP=='3'
],var.equal = F)
p23
```

```
##
##  Welch Two Sample t-test
##
## data:  leadData$IQ[leadData$GROUP == "2"] and leadData$IQ[leadData$GROUP =
= "3"]
## t = 0.23761, df = 52.997, p-value = 0.8131
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.131548  6.510715
## sample estimates:
## mean of x mean of y
##  87.65625  86.96667
```

The p-value for Group 1 and 2 comparison is **0.01205**.

The p-value for Group 1 and 3 comparison is **0.023**.

The p-value for Group 2 and 3 comparison is **0.8131**.

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.)

We **reject the null hypothesis** as the p-value is smaller than the level of significance for two of the pairwise comparisons.

If we consider the issue of multiple testing and apply Bonferroni's correction on the data, the new level of significance would be 0.05/3 = 0.0167. After comparing with the new significance level, we **reject the null hypothesis** as the p-value is smaller than the new level of significance for one group.

Thus, there is an association between lead and IQ.

After applying the Bonferroni correction, the conclusion is not different from conclusion of Q8.

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

Shown below Descriptive statistics summarize and organize characteristics if a dataset where dataset is nothing but a collection of responses or observations from a sample or an entire population. The 3 main types of descriptive stats:

1. The distribution concerns the frequency of each value.
2. The central tendency concerns the averages of the value.
3. The variability concerns how spread out the values are.

Shown below are some descriptive stats (mean, SD, frequency and variance) for males:

```
m_1 = with(leadData,tapply(IQ[leadData$SEX==1], GROUP[leadData$SEX==1], mean)
)
s_1 = with(leadData,tapply(IQ[leadData$SEX==1], GROUP[leadData$SEX==1], sd))
n_1 = with(leadData,tapply(IQ[leadData$SEX==1], GROUP[leadData$SEX==1], lengt
h))
data.frame(m_1,s_1,n_1,(s_1^2))

##        m_1      s_1 n_1 X.s_1.2.
## 1 92.93478 15.42351  46 237.8845
## 2 90.17647 11.13124  17 123.9044
## 3 86.61538 17.32791  13 300.2564
```

Shown below are some descriptive stats (mean, SD, frequency and variance) for females:

```
m_2 = with(leadData,tapply(IQ[leadData$SEX==2], GROUP[leadData$SEX==2], mean)
)
s_2 = with(leadData,tapply(IQ[leadData$SEX==2], GROUP[leadData$SEX==2], sd))
n_2 = with(leadData,tapply(IQ[leadData$SEX==2], GROUP[leadData$SEX==2], lengt
h))
data.frame(m_2,s_2,n_2,(s_2^2))
```

```
##        m_2       s_2 n_2  X.s_2.2.
## 1 94.60976 15.877465  41 252.09390
## 2 84.80000  6.471917  15  41.88571
## 3 87.23529  8.898942  17  79.19118
```

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.

Performed the ANOVA test to compare the mean IQ scores in the 3 groups for males and females separately.

```
summary(aov(leadData$IQ[leadData$SEX==1]~leadData$GROUP[leadData$SEX==1]))

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

summary(aov(leadData$IQ[leadData$SEX==2]~leadData$GROUP[leadData$SEX==2]))

##                                   Df Sum Sq Mean Sq F value Pr(>F)
## leadData$GROUP[leadData$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
```

The p-value from the test to compare the mean IQ scores for males is **0.168**.

The p-value from the test to compare the mean IQ scores for females is **0.0243**.

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.)

From the above questions, we **do not have enough evidence to reject the null hypothesis** for males but we **reject the null hypothesis** for females.

Thus, we can conclude that there is no association between lead and IQ for males however there is an association between lead and IQ for females.

The conclusions agree with the results of Q8 and Q11 for females but not for males.

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

Since the variance are not equal for the three groups, 2 sample t-test for unequal variances (Welch test) has been performed for the pairwise comparisons.

```
p12_m = t.test(leadData$IQ[leadData$GROUP=='1' & leadData$SEX==1],leadData$IQ
[leadData$GROUP=='2' & leadData$SEX==1],var.equal = F)
p12_m

##
##  Welch Two Sample t-test
##
```

```
## data:  leadData$IQ[leadData$GROUP == "1" & leadData$SEX == 1] and leadData
$IQ[leadData$GROUP == "2" & leadData$SEX == 1]
## t = 0.78142, df = 39.661, p-value = 0.4392
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.377699  9.894323
## sample estimates:
## mean of x mean of y
##  92.93478  90.17647
```

```
p13_m = t.test(leadData$IQ[leadData$GROUP=='1' & leadData$SEX==1],leadData$IQ
[leadData$GROUP=='3' & leadData$SEX==1],var.equal = F)
p13_m
```

```
##
##  Welch Two Sample t-test
##
## data:  leadData$IQ[leadData$GROUP == "1" & leadData$SEX == 1] and leadData
$IQ[leadData$GROUP == "3" & leadData$SEX == 1]
## t = 1.1886, df = 17.738, p-value = 0.2503
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.862555 17.501351
## sample estimates:
## mean of x mean of y
##  92.93478  86.61538
```

```
p23_m = t.test(leadData$IQ[leadData$GROUP=='2' & leadData$SEX==1],leadData$IQ
[leadData$GROUP=='3' & leadData$SEX==1],var.equal = F)
p23_m
```

```
##
##  Welch Two Sample t-test
##
## data:  leadData$IQ[leadData$GROUP == "2" & leadData$SEX == 1] and leadData
$IQ[leadData$GROUP == "3" & leadData$SEX == 1]
## t = 0.64603, df = 19.325, p-value = 0.5259
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.963105 15.085277
## sample estimates:
## mean of x mean of y
##  90.17647  86.61538
```

```
p12_f = t.test(leadData$IQ[leadData$GROUP=='1' & leadData$SEX==2],leadData$IQ
[leadData$GROUP=='2' & leadData$SEX==2],var.equal = F)
p12_f
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  leadData$IQ[leadData$GROUP == "1" & leadData$SEX == 2] and leadData
$IQ[leadData$GROUP == "2" & leadData$SEX == 2]
## t = 3.2807, df = 53.22, p-value = 0.001831
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   3.812848 15.806664
## sample estimates:
## mean of x mean of y
##  94.60976  84.80000
```

```
p13_f = t.test(leadData$IQ[leadData$GROUP=='1' & leadData$SEX==2],leadData$IQ
[leadData$GROUP=='3' & leadData$SEX==2],var.equal = F)
p13_f
```

```
##
##  Welch Two Sample t-test
##
## data:  leadData$IQ[leadData$GROUP == "1" & leadData$SEX == 2] and leadData
$IQ[leadData$GROUP == "3" & leadData$SEX == 2]
## t = 2.2433, df = 50.748, p-value = 0.02927
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.7739536 13.9749703
## sample estimates:
## mean of x mean of y
##  94.60976  87.23529
```

```
p23_f = t.test(leadData$IQ[leadData$GROUP=='2' & leadData$SEX==2],leadData$IQ
[leadData$GROUP=='3' & leadData$SEX==2],var.equal = F)
p23_f
```

```
##
##  Welch Two Sample t-test
##
## data:  leadData$IQ[leadData$GROUP == "2" & leadData$SEX == 2] and leadData
$IQ[leadData$GROUP == "3" & leadData$SEX == 2]
## t = -0.89218, df = 29.016, p-value = 0.3796
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.017810  3.147222
## sample estimates:
## mean of x mean of y
##  84.80000  87.23529
```

The p-value from the test to compare the mean IQ scores for Group 1 and 2 males is **0.4392**.

The p-value from the test to compare the mean IQ scores for Group 1 and 3 males is **0.2503**.

The p-value from the test to compare the mean IQ scores for Group 2 and 3 males is **0.5259**.

The p-value from the test to compare the mean IQ scores for Group 1 and 2 females is **0.001831**.

The p-value from the test to compare the mean IQ scores for Group 1 and 3 females is **0.02927**.

The p-value from the test to compare the mean IQ scores for Group 2 and 3 females is **0.3796**.

16. What do you conclude about the association between lead and IQ from the results in Q15? Does your conclusion change from previous conclusions made in Q8, Q11 and Q14?

We **reject the null hypothesis** for female but **do not have evidence to reject the null hypothesis** for male based on the conclusions from Q15.

The conclusions from the previous questions are as follows:

Q8 -> reject the null hypothesis

Q11 -> reject the null hypothesis

Q14 -> reject the null hypothesis for females and do not reject the null hypothesis for males

Q16 -> reject the null hypothesis for females and do not reject the null hypothesis for males

Thus, we can conclude that there is an association between lead exposure and IQ values.(from Q8 and Q11).

However, when we compare the mean IQ scores in the 3 groups for males and females separately we see that there is no association between lead exposure and IQ values for males but there is an association between lead exposure and IQ values for female. (from Q14 and Q16).

It can also be noted in all the four questions there is an association between the lead exposure and mean IQ scores for females.