DATA 557: HW Assignment 3

Hriday Baghar

February 10, 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

```
# Loading dataset
library(dplyr)
data <- read.csv('lead_study.csv')
str(data)
```

```
## 'data.frame':    149 obs. of  3 variables:
##  $ SEX  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ GROUP: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ IQ   : int  70 85 86 76 84 96 94 97 99 80 ...
```

```
summary(data)
```

```
##       SEX            GROUP             IQ        
##  Min.   :1.00   Min.   :1.000   Min.   : 46.00  
##  1st Qu.:1.00   1st Qu.:1.000   1st Qu.: 82.00  
##  Median :1.00   Median :1.000   Median : 90.00  
##  Mean   :1.49   Mean   :1.617   Mean   : 91.06  
##  3rd Qu.:2.00   3rd Qu.:2.000   3rd Qu.: 99.00  
##  Max.   :2.00   Max.   :3.000   Max.   :141.00
```

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?

```
males <- filter(data, SEX == 1)
females <- filter(data, SEX == 2)
cat(var(males$IQ), var(females$IQ))
```

```
## 222.9298 184.554
```

```
#We assume equal variances since the two groups don't seem too far off
t.test(x = males$IQ, y = females$IQ, paired = FALSE, var.equal = TRUE)$p.value
```

```
## [1] 0.8779726
```

2. State the conclusion from your test.

We fail to reject the null hypothesis.

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

The independence assumptions are valid because each sample from the male and female group should be independent from each other. We have no indication of whether the observed groups have the same underlying causes due to which each person is placed in the same group as defined by the GROUP variable.

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 mean IQ scores of the 3 groups are equal.

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

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

The mean IQ scores of the 3 groups are not all equal.

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

We should use ANOVA.

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

```r
summary(aov(IQ ~ factor(GROUP), data = data))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## factor(GROUP)   2   1491   745.4   3.816 0.0242 *
## Residuals     146  28522   195.4
## ---
## 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.

There is a relationship between lead exposure and IQ. We reject the null hypothesis.

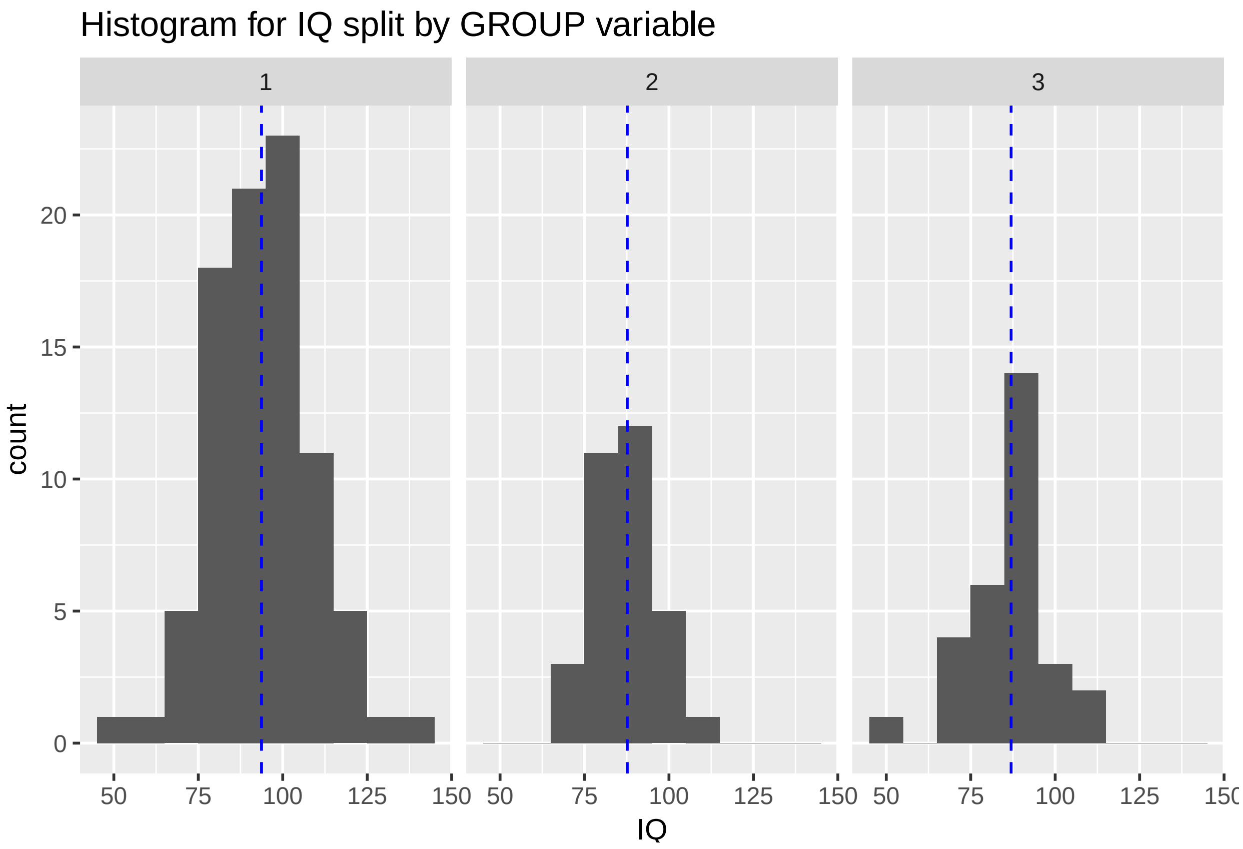
9. Are there strong reasons to believe that the assumptions of this test are not met? Briefly justify your answer.

```r
#Evaluating assumptions of test

#Checking for equal variance and sample size
data.stats <- data %>% group_by(GROUP) %>% summarise(sample.size = n(), mean = mean(IQ), variance = var(IQ))
data.stats
```

```
## # A tibble: 3 x 4
##   GROUP sample.size  mean variance
##   <int>       <int> <dbl>    <dbl>
## 1     1          87  93.7    242.
## 2     2          32  87.7     90.3
## 3     3          30  87.0    168.
```

```r
#Checking for normality
library(ggplot2)

#plot IQ histogram for each group
```

```
hist <- ggplot(data, aes(IQ)) +
  geom_histogram(binwidth = 10) +
  facet_grid(.~GROUP) +
  geom_vline(data = data.stats, aes(xintercept = mean), color = 'blue', linetype = 2) +
  ggtitle("Histogram for IQ split by GROUP variable")
hist
```



The 3 assumptions for ANOVA and an assessment of whether they are being met:

- Independence ✓
  - Each child's observation is independent of others hence it holds
- Equal variance ×
  - As we can see in the above table, variances of the 3 groups are not very close - this assumption might not hold true
- Large sample size or normal distribution of population ✓
  - There are at least 30 samples in each group and population distribution of IQ tends to follow a normal distribution

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

```
pairwise.fn <- function(data) {
  combs <- combn(unique(data$GROUP), 2)
  pairwise.df <- data.frame(group.1 = combs[1,],
                            group.2 = combs[2,],
                            p.value = rep(NA, length(combs[1,])))

  for(row in 1:length(pairwise.df)){
    x <- data %>% filter(GROUP == pairwise.df[row,1]) %>% select(IQ)
    y <- data %>% filter(GROUP == pairwise.df[row,2]) %>% select(IQ)
    pairwise.df[row, "p.value"] <- t.test(x, y, paired = FALSE, var.equal = FALSE)$p.value
  }
  return(pairwise.df)
}

pairwise.fn(data)
```

```
##   group.1 group.2    p.value
## 1       1       2 0.01205060
## 2       1       3 0.02300319
## 3       2       3 0.81310355
```

Since we observed that the variances for each of the samples are not close, we perform a Welch t-test for the pairwise comparisons for all pairwise comparisons going forward.

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 must apply a Bonferroni correction for the 3 tests that we conduct. The new significance level is $\alpha_{corrected} = 0.05/3 = 0.01667$.

Based on this, we fail to reject the null hypothesis of 2 of the 3 tests (1 of 3 if we choose not to apply corrections for inflate type 1 error). ANOVA's results suggest that all 3 of the means are not equal and it appears to be consistent with the pairwise Welch t-tests.

It appears that lead exposure does in fact affect IQ. The effect on group 1 seems to be different from the effect on group 2 and 3.

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

```
data %>% mutate(SEX = if_else(SEX==1, "Male", "Female")) %>%
  group_by(SEX, GROUP) %>%
  summarise(sample.size = n(), mean = mean(IQ), var = var(IQ))
```

```
## # A tibble: 6 x 5
## # Groups:   SEX [2]
##   SEX    GROUP sample.size  mean   var
##   <chr>  <int>       <int> <dbl> <dbl>
## 1 Female     1          41  94.6 252.
## 2 Female     2          15  84.8  41.9
## 3 Female     3          17  87.2  79.2
## 4 Male       1          46  92.9 238.
## 5 Male       2          17  90.2 124.
## 6 Male       3          13  86.6 300.
```

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.

```
#males
summary(aov(IQ ~ factor(GROUP), data = males))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## factor(GROUP)  2    429   214.7   0.962  0.387
## Residuals     73  16290   223.2
```

```
#females
summary(aov(IQ ~ factor(GROUP), data = females))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)  
## factor(GROUP)  2   1351   675.3    3.96 0.0235 *
## Residuals     70  11937   170.5                 
## ---
## 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 reject the null hypothesis for females but fail to reject it for males.

It seems there is no effect on IQ and lead exposure for males, however for females we see that there is an effect on IQ due to lead exposure. There is some agreement between this result and those in Q8 and Q11, i.e. amount of lead exposure has an effect on IQ, the difference is that we have found this to be true only for females.

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

```
#males
pairwise.fn(males)
```

```
##   group.1 group.2   p.value
## 1       1       2 0.4391938
## 2       1       3 0.2502756
## 3       2       3 0.5258587
```

```
#females
pairwise.fn(females)
```

```
##   group.1 group.2     p.value
## 1       1       2 0.001830775
## 2       1       3 0.029274571
## 3       2       3 0.379635167
```

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?

We conclude that there is no effect of lead exposure on male IQ. However, for females we find that group 1 is significantly different from group 2 and 3 (if we were to look merely at the p-value without accounting for inflated type 1 error).

We apply a Bonferroni correction for the significance level $\alpha_{corrected} = 0.05/3 = 0.01667$ since we are conducting 3 pairwise tests for each of the groups. The pairwise results are in agreement with ANOVA for both males and females.

Our conclusion narrows down compared to previous results. First, we found that lead exposure has an effect on IQ. Then, we found that this is true only for females (based on the observed results).

For the tests performed, it is important to decide the appropriate significance levels, power, assumptions we make about the population distributions and equality of variances between groups. These decisions will greatly impact the hypotheses that we accept and reject, especially given that the sample size is not very large for some of the subsets we created.