STAT 614 - HW 5

By Sihyuan Han

Due: Thursday, October 29, 2020 in Blackboard by 11:59pm.

Instructions: Please type your solutions in a separate document and upload the document in Blackboard. Include supporting work (plots, etc.) when appropriate, but do not copy all computer output. Select only relevant output. I will not be collecting syntax for this assignment.

Notes:

- For this HW you will need some concepts from chapter 5 on the ANOVA model.
- HW 6 will finish out the ANOVA section.

The effects of exposure to lead on the psychological and neurological well-being of children were studied by Landrigan et al. (1975). Complete raw data for this study are in the data set lead.csv in Blackboard. The data describe a group of children who lived near a lead smelter in El Paso, Texas. Two exposed groups of children were identified who had blood-lead levels $\geq$ 40 □g/ml in 1972 or in 1973. Because neurological and psychological tests were performed in 1973, researchers argued that it would be better to define an exposure group based on blood-lead levels in 1973 only. For this purpose, the variable `lead_typ` in the data file gives three exposure groups:

If `lead_typ` = 1, then the child had normal blood-lead levels (<40 μg/ 100 mL) in both 1972 and 1973 (control group).
If `lead_typ` = 2, then the child had elevated blood-lead levels ($\geq$40 μg/100 mL) in 1973 (the currently exposed group).
If `lead_typ` = 3, then the child had elevated blood-lead levels in 1972 and normal blood-lead levels in 1973 (the previously exposed group).

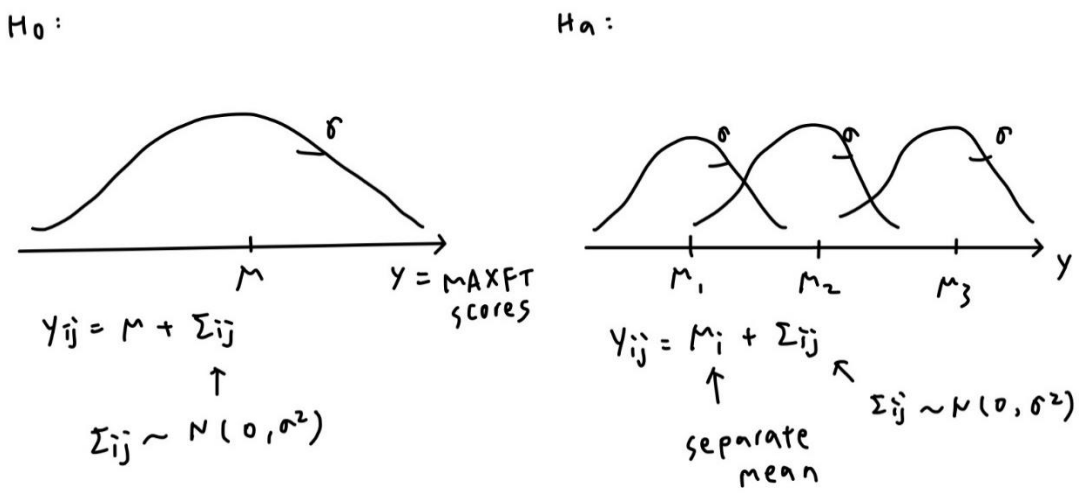
One important measure of neurological function studied was `MAXFT` = the number of finger-wrist taps in the dominant hand. Researchers are interested in whether there is evidence of differences in neurological function, as measured by `MAXFT`, on average, between the three exposure populations. They would also like to test and estimate the average difference in `MAXFT` between each pair of exposure populations, **with the expectation that populations with normal blood-lead levels will have higher average `MAXFT` scores**. **It is unclear if previously exposed populations will have "recovered" any function as compared to a currently exposed population**. Address these research questions by answering the following questions.

1. State the hypotheses of interest to be tested. Include the overall test of group differences in addition to all possible pairwise comparisons of interest.
   Ans:
   RQ1- Normal blood-lead levels will have higher average MAXFT scores
   H0: mean MAXFT group1 = group2
   Ha: mean MAXFT group1 > group2

   RQ2- Previously exposed populations will have "recovered" compared to a currently exposed population
   H0: mean MAXFT group2 = group3
   Ha: mean MAXFT group2 < group3

   Overall test of equal population means
   H0: mean MAXFT group1 = group2 = group3
   Ha: Not all means are equal (at least one pair of means is not equal)

2. Write the ANOVA model to be fit.
   Ans:

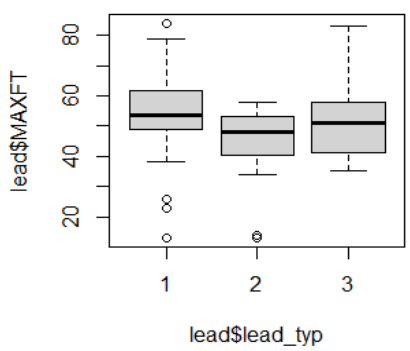


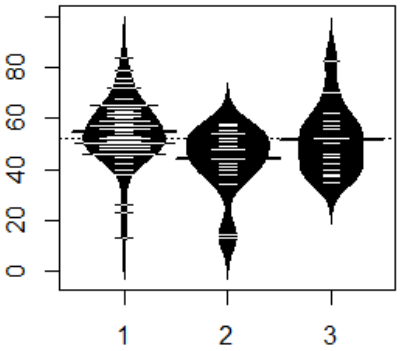
3. Conduct a brief exploratory analysis of the `MAXFT` variable by exposure group (`lead_typ`). Give supporting graphs, descriptive statistics, and interpret these results.
   Ans:

```
> favstats(MAXFT~lead_typ, data = lead)
  lead_typ min   Q1 median    Q3 max    mean       sd  n missing
1        1  13 49.0   53.5 61.25  84 54.4375 12.05658 64      14
2        2  13 40.5   48.0 53.00  58 44.0000 12.65350 19       5
3        3  35 41.5   51.0 57.50  83 51.5000 12.94604 16       6
```

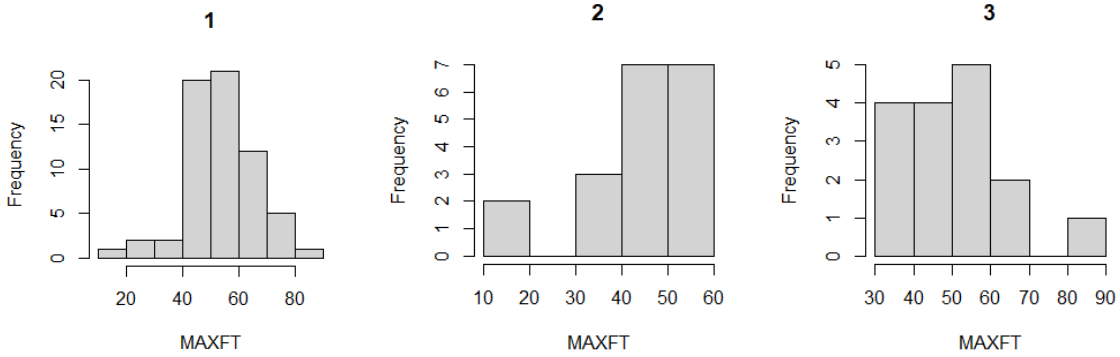
   We can see that mean of group1 ≈ 54.4, group2 has mean ≈ 44, and group3 ≈51.5.

As we can see below of the boxplot and beanplot, there are some outliers for group 1 and 2

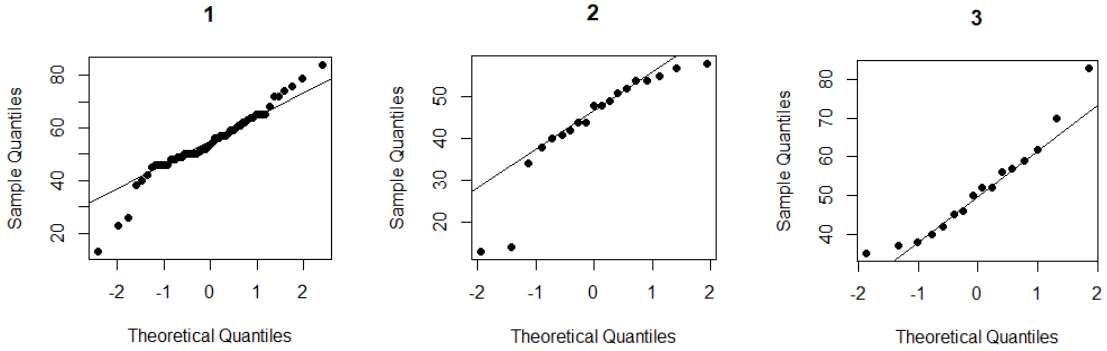




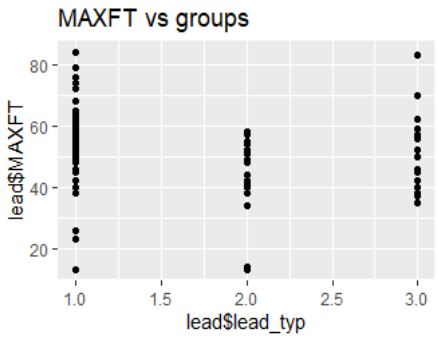
The histogram also shows some outliers



The qqplot shows the data is not mostly linear and some outliers as well



Residuals vs groups



4. What are the assumptions of the model (and corresponding hypothesis tests)? Based on the exploratory analysis in (3), are the assumptions reasonably met for this data? If not, what adjustments should you make in your analysis? (You don't need to use residuals here – you'll do that on HW 6 and then for the rest of the semester!!!)

Ans:

(1) Independent observations with groups (study design)
(2) Independent samples (between groups) (study design)
(3) Normally distributed – based on the QQ plot, it is mostly linear but have several outliers, so remove some outliers and check again, yes!
(4) Equal variances- as it shows in the favstats and boxplot above the variances are close, checked!
(5) Influential outliers assessed– as show in the plot, we can see there are outliers, but can further test if they are influential outliers.

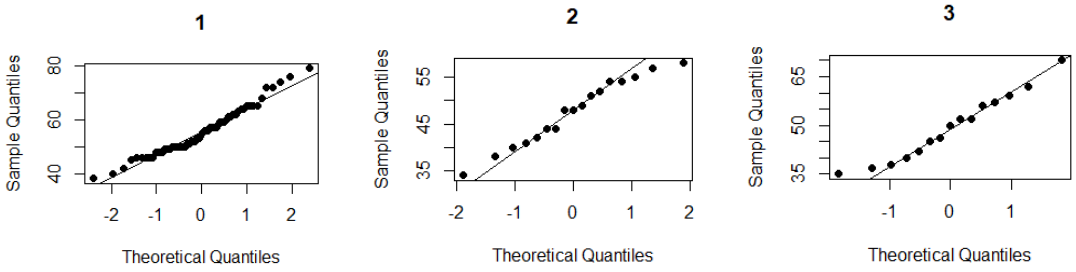
Check for normal distribution

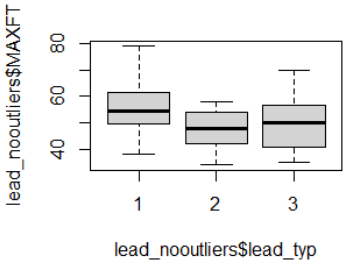
```
        Shapiro-Wilk normality test

data:  lead$MAXFT[lead$lead_typ == "1"]
W = 0.94641, p-value = 0.007647

> shapiro.test(lead$MAXFT[lead$lead_typ=="2"])

        Shapiro-Wilk normality test

data:  lead$MAXFT[lead$lead_typ == "2"]
W = 0.83912, p-value = 0.004485

> shapiro.test(lead$MAXFT[lead$lead_typ=="3"])

        Shapiro-Wilk normality test

data:  lead$MAXFT[lead$lead_typ == "3"]
W = 0.93698, p-value = 0.3137

>
```

```
        Shapiro-Wilk normality test

data:  lead_nooutliers$MAXFT[lead_nooutliers$lead_typ == "1"]
W = 0.96743, p-value = 0.1091

> shapiro.test(lead_nooutliers$MAXFT[lead_nooutliers$lead_typ=="2"])

        Shapiro-Wilk normality test

data:  lead_nooutliers$MAXFT[lead_nooutliers$lead_typ == "2"]
W = 0.96317, p-value = 0.6916

> shapiro.test(lead_nooutliers$MAXFT[lead_nooutliers$lead_typ=="3"])

        Shapiro-Wilk normality test

data:  lead_nooutliers$MAXFT[lead_nooutliers$lead_typ == "3"]
W = 0.96578, p-value = 0.7915

>
```

As show in the graph above, in the left the data is not normal distrbuted of group 1 and 2. In the right, I remove some outliers and each group shows normal distribuuted!

The under QQ plots are outliers removed, we can see they are mostly linear!





boxplot- outlier removed

5. Conduct the appropriate analysis (i.e. incorporate any recommended adjustments from (d) if you had them). Clearly and briefly state the conclusions of your analysis. Be sure you address the researcher's questions.

Ans:

(1) Overall test

F-test for equal means

```
> anova(model.fit)
Analysis of Variance Table

Response: MAXFT
          Df  Sum Sq Mean Sq F value   Pr(>F)
lead_typ   2  1600.1  800.04  5.2773 0.006692 **
Residuals 96 14553.8  151.60
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Total (SST)- sum of squares: 16153.9/ df: 98

P(F>=5.2773) ≈ 0.006692

The p-value is small enough so we have evidence to reject H0 (null hypothesis) which states equal means. We can conclude that there is evidence to say that the mean of MAXFT differs between at least one pair of groups.

(2) RQ

PosHocTest “Bonferroni”

```
  Posthoc multiple comparisons of means : Bonferroni
    95% family-wise confidence level

$lead_typ
        diff     lwr.ci    upr.ci   pval
2-1 -10.4375 -18.275728 -2.599272 0.0049 **
3-1  -2.9375 -11.323225  5.448225 1.0000
3-2   7.5000  -2.679885 17.679885 0.2273

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

PosHocTest “Bonferroni” (data remove outliers)

```
  Posthoc multiple comparisons of means : Bonferroni
    95% family-wise confidence level

$lead_typ
         diff     lwr.ci      upr.ci   pval
2-1 -8.045098 -13.965949 -2.12424724 0.0040 **
3-1 -6.233333 -12.454162 -0.01250499 0.0494 *
3-2  1.811765  -5.822087  9.44561677 1.0000

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For RQ1- Normal blood-lead levels will have higher average MAXFT scores

H0: mean MAXFT group1 = group2

Ha: mean MAXFT group1 > group2

For group1 vs group2, the p-value is approximately 0.0049 < 0.05, so we have evidence to reject the null hypothesis.
we can conclude that there is evidence to say that children with normal blood-lead levels will have higher average MAXFT scores.
The p-value ≈ 0.004 in outliers removed PosHocTest “Bonferroni”, we can say that the conclusion is the same.

For RQ2- Previously exposed populations will have “recovered” compared to a currently exposed population
H0: mean MAXFT group2 = group3
Ha: mean MAXFT group2 < group3
For group2 vs group3, the p-value ≈ 0.2273 > 0.05, so we have not enough evidence to reject the null hypothesis.
So there is no enough evidence to say that previously exposed populations will “recovered” any function compared to currently exposed population.
The p-value ≈ 1 still > 0.05 in outliers removed PosHocTest “Bonferroni”, the result is the same!