STAT 614 - Yunting Chiu

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.sav 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 > 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 (>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.

   We typically start with an overall test of equal population means versus any difference in population means. This test allows us to access the evidence for whether the variability in the data comes mostly from variability within groups or can be attributed to variability between groups.

Overall test of equal population means:
H0: mu1 = mu2 = mu3 (the difference means is equal to 0)
Ha: mu1 ≠ mu2 or mu1 ≠ mu3 or mu2 ≠ mu3 (at least one pair of means is not equal)

RQ1 - Normal blood-lead levels will have higher average MAXFT scores.
H0: mu1 = mu2
Ha: mu1 > mu2

RQ2 - Previously exposed populations will have higher MAXFT scores because they have "recovered" in the following year.
H0: mu2 = mu3
Ha: mu2 < mu3

RQ3- Normal blood-lead levels will have higher average MAXFT scores.
H0: mu1 = mu3
Ha: mu1 > mu3

2. Write the ANOVA model to be fit.

ANOVA is used to determine whether there are any statistically significant differences between the means of two or more independent groups. In this study, because we have three group's means, and the parametric method is better than the non-parametric method. Thus, we choose the ANOVA model (parametric method) for the study.

ANOVA notation:

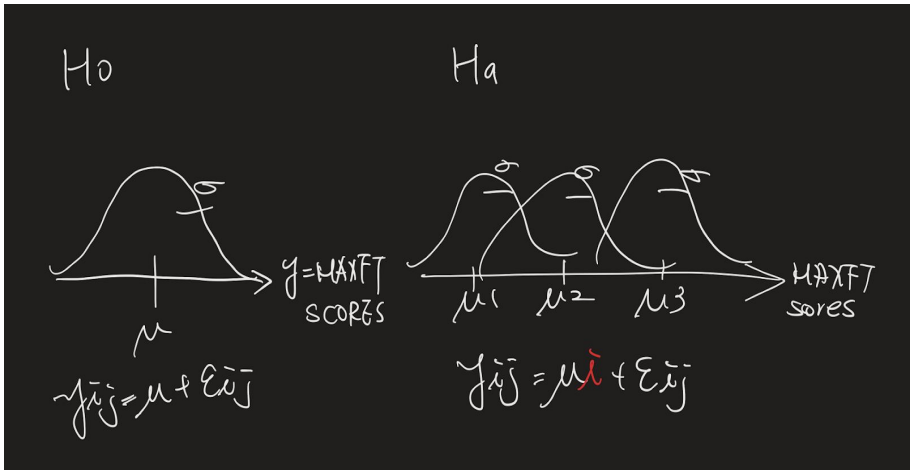
$$y_{ij} = \mu_i + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

i = lead group = 1, 2, 3 $\quad \varepsilon_{ij}$ = deviation from the mean for MAXFT$_j$ in lead group i

j = MAXFT score $\quad \sigma^2$ = population variance from the mean in ij.

μ_i = pop^n mean of lead group



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

With the exploratory data analysis of this study, we need to make sure the lead_tpe and MAXFTwe could see group 1 was missing 14 people of MAXFT score, group 2 was missing 5 people of MAXFT score, and group 3 was missing 6 people of MAXFT score, respectively. Therefore, we need to remove the NAs in the data frame first.

```
- EDA
```{r}
# summary(lead$lead_typ)
favstats(MAXFT ~ lead_typ, data = lead)
```
```

lead_typ <chr>	min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>	missing <int>
1	13	49.0	53.5	61.25	84	54.4375	12.05658	64	14
2	13	40.5	48.0	53.00	58	44.0000	12.65350	19	5
3	35	41.5	51.0	57.50	83	51.5000	12.94604	16	6

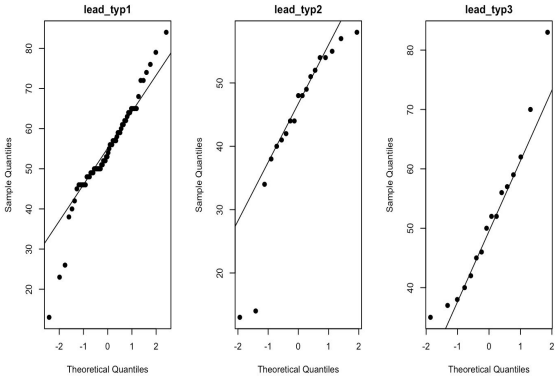
Removed NAs of MAXFT (more clear)

lead_typ <chr>	min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>	missing <int>
1	13	49.0	53.5	61.25	84	54.4375	12.05658	64	0
2	13	40.5	48.0	53.00	58	44.0000	12.65350	19	0
3	35	41.5	51.0	57.50	83	51.5000	12.94604	16	0

3 rows

According to the revised data, the mean of group 1 is 54.44, the mean of group 2 is 44.00, and the mean of group 3 is 51.50, separately. And the sample sizes are 64, 19, 16. The group 1 has the most sample sizes.The variance (sd*sd) of three groups are approximately equal.

Then, We use qqplot to detect it. According to the graph below, we could see some points do not match up along a straight line, especially in the head and tail. The qqplot of three groups does not follow a normal distribution.



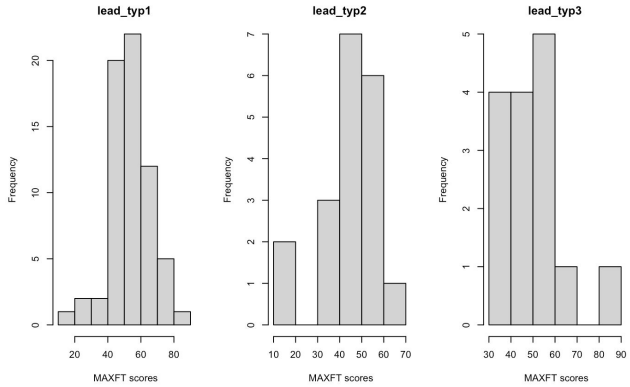
We also use the Shapiro-Wilk Normality Test to check the data. The output obtains that we can assume normality. The p-value (0.002108) is smaller than 0.05. Hence, the distribution of the given data is different from normal distribution significantly.

```r
shapiro.test(leadRealMAXFT$MAXFT)
```

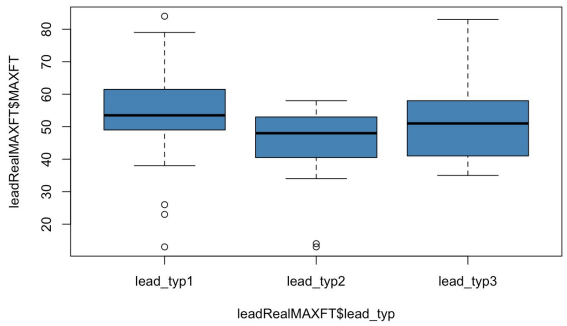
```
        Shapiro-Wilk normality test

data:  leadRealMAXFT$MAXFT
W = 0.9556, p-value = 0.002108
```

Based on the histogram below, we can see there three groups still have a few outliers.



The below boxplot indicates there are some outliers in group 1 and group 2.



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!!!)

ANOVA assumptions:
Study design:

1. Independent populations for lead groups.
2. Independent sample from each population group.

Data finding:

- The distribution of MAXFT with-in each group follows a normal distribution.
- The population distributions have the same variance.
- No influential outliers.

Based on the result of question3, we can conclude that the dataset does not follow the ANOVA assumptions, so we try to remove the outliers from the original data frame.

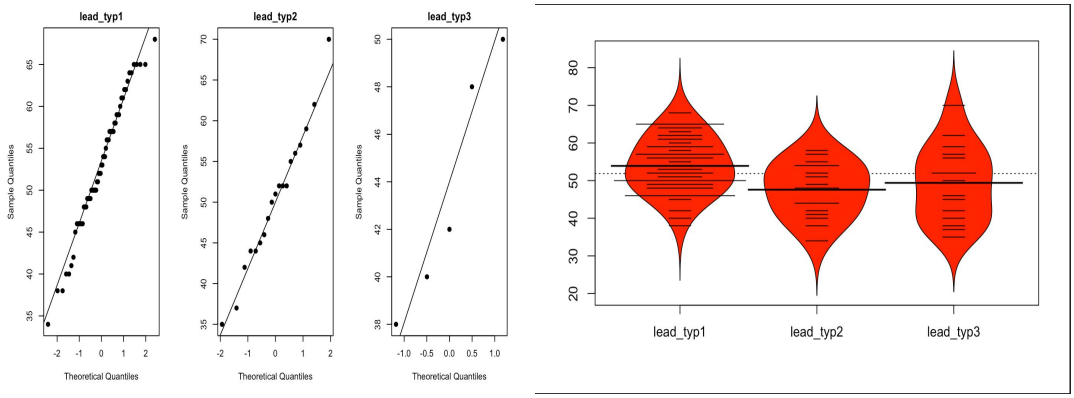
Then I took out the outliers of three groups. After that, the variance of three groups are different, but it is okay.

```r
leadRealMAXFT %>%
  filter(MAXFT >= 30 & MAXFT <= 70) -> leadRmOutliers
favstats(MAXFT ~ lead_typ, data = leadRmOutliers)
```

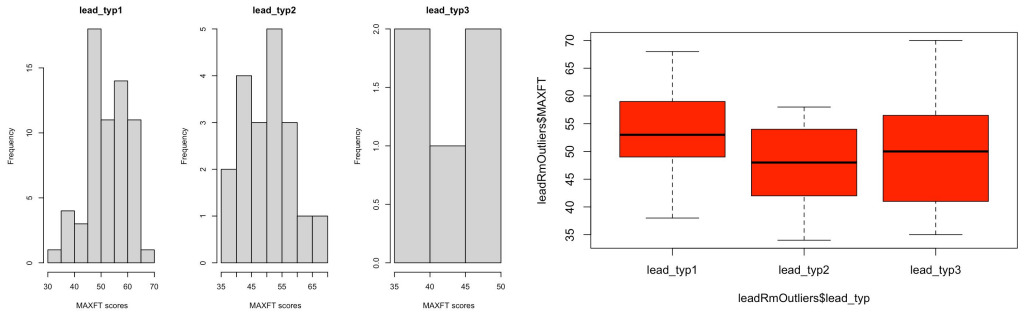
lead_typ	min	Q1	median	Q3	max	mean	sd	n	missing
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
1	38	49	53	59.0	68	53.90909	7.053428	55	0
2	34	42	48	54.0	58	47.58824	7.080420	17	0
3	35	41	50	56.5	70	49.40000	10.196638	15	0

3 rows

We can see the revised qqplot is following the normal distribution. And the beanplot's mean is approximately equal to the median.



The revised histogram and boxplot have no influential outliers now.



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.

   Above all, these conditions are approved by the assumptions of the ANOVA model. Therefore, we can start to analyze the case study.

   - F-test for equal means

```
Analysis of Variance Table

Response: MAXFT
          Df Sum Sq Mean Sq F value   Pr(>F)
lead_typ   1  425.2  425.21  7.0169 0.009625 **
Residuals 85 5150.9   60.60
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

   Based on the ANOVA result, because the F value is greater than 6.94 (critical value for 1% significance level), so we reject the null. Also, the small p value 0.009625 is smaller than 1%. These two reasons indicate that we reject the null model (equal means) in favor of the unequal means and conclude that there is evidence that the population mean of MAXFT scores across at least one pair of lead groups.

   For the multiple comparisons, we use LSD (Least Significant Different) test.

   - Original data frame vs. Removed outliers dataframe

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

$lead_typ
       diff      lwr.ci    upr.ci   pval
2-1 -10.4375 -16.8227981 -4.052202 0.0016 **
3-1  -2.9375  -9.7688090  3.893809 0.3955
3-2   7.5000  -0.7928946 15.792895 0.0758 .

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

```
- All pairwise comparisons
```{r}
library(DescTools)
PostHocTest(model.fit, method = "lsd")
```


  Posthoc multiple comparisons of means : Fisher LSD
    95% family-wise confidence level

$lead_typ
         diff   lwr.ci   upr.ci   pval
2-1 -6.320856 -10.554566 -2.08714536 0.0039 **
3-1 -4.509091 -8.953180 -0.06500204 0.0468 *
3-2 1.811765 -3.592861 7.21638999 0.5068

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

Because we use LSD test, so the significant level is p-value/3 (0.05/3 = 0.0166) for each pairwise comparison.

RQ1 - Normal blood-lead levels will have higher average MAXFT scores.
H0: mu1 = mu2
Ha: mu1 > mu2
We reject the null hypothesis concluding that the p-value is 0.0039 which is smaller than 0.05/3 = 0.0166, on average, between lead_typ1 and lead_typ2 with 95% confidence interval. We can conclude that there is evidence that children with normal blood-lead levels will have higher average MAXFT scores. On the other hand, based on original data we also can reject the null hypothesis(0.0016), that means the removed outliers are not influential outliers.

RQ2 - Previously exposed populations will have higher MAXFT scores because they have “recovered” in the following year.
H0: mu2 = mu3
Ha: mu2 < mu3
We fail to reject the null hypothesis that the p-value is 0.5068 which is greater than 0.05/3 = 0.0166 with 95 % confidence interval. Therefore, previously exposed populations lead_typ3 will NOT have higher MAXFT scores than lead_typ2.
On the other hand, based on original data, the p value 0.0758 also fails to reject the null hypothesis, that means the removed outliers are not influential outliers.

RQ3- Normal blood-lead levels will have higher average MAXFT scores.
H0: mu1 = mu3
Ha: mu1 > mu3
We fail to reject the null hypothesis concluding that the p-value is 0.0468 which is greater than 0.05/3 = 0.0166, on average, between lead_typ1 and lead_typ3 with 95% confidence interval. We can conclude that there is evidence that children with normal blood-lead levels will NOT have higher average MAXFT scores. Moreover, based on original data we fail to reject the null hypothesis (0.3955), that indicates the removed outliers are not influential outliers.