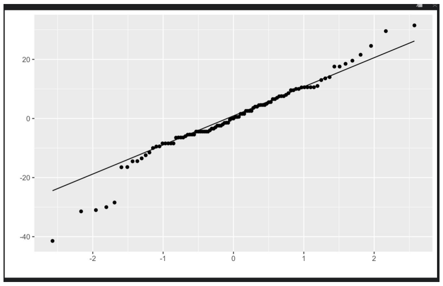
STAT 614 - HW 6 - Yunting Chiu

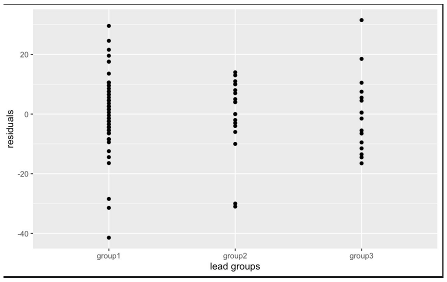
Due: Thursday, November 5, 2020 in Blackboard by 11:59pm.
Instructions: Please type your solutions in a separate document and upload the document in Blackboard as a pdf. Include supporting work (plots, etc.) when appropriate, but do not copy all computer output. Select only relevant output. I will not be collecting syntax for this assignment.

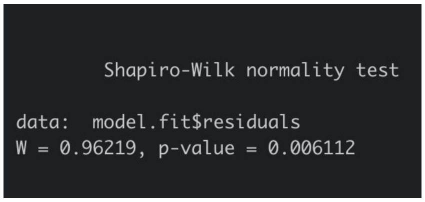
1. Revisit the analysis of the data from the study of effects of exposure to lead on the psychological and neurological well-being of children from the previous HW. (Recall that the data are given in the lead.csv data set from HWs 1 and 5.)

a. Use a residual analysis to assess whether the assumptions of the ANOVA model (on the untransformed, full data set) are met. What remedies do you recommend and why (Note: you will use a nonparametric method in the next part of this problem so there is no need to complete the analysis using your recommended remedy). (See my notes on the next page for addressing the missing observations on the MAXFT variable.)

   Normality assumption for residuals:
   1. Normal distributed (residuals)
   2. Equal variance (residuals)
   3. Influential outliers assessed (residuals)
   4. Independent observations with groups (study design)
   5. Independent samples between groups (study design)







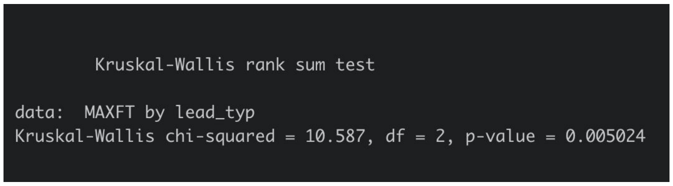
```
        Shapiro-Wilk normality test

data:  model.fit$residuals
W = 0.96219, p-value = 0.006112
```

According to qqplot, we can see there is a skewness / deviations in the tail and the head, and p-value is < 0.05 based on the Shpiro-Wilk test. These two points can conclude the residuals are not following normal distribution. Also, in the second plot, we also find some potential outliers and the variances are not equal. Therefore, we should use a nonparametric method or transform the data (square root, log) for this study.

b. Use nonparametric methods to address the research questions of interest. Restate how the hypotheses of interest from the last HW will be addressed using the nonparametric methods. Carry out the analysis (on untransformed data) and clearly state the conclusions. Conduct two- sided pairwise comparisons using the Bonferroni method. How would you conduct a one-sided test?

Overall test of equal population means:
H0: the distribution of MAXFT score is the same of all lead groups.
Ha: Not all population means are equal.

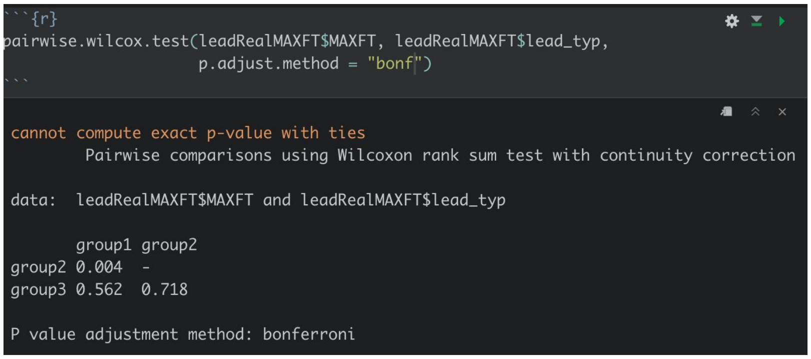


```
        Kruskal-Wallis rank sum test

data:  MAXFT by lead_typ
Kruskal-Wallis chi-squared = 10.587, df = 2, p-value = 0.005024
```

Based on the Kruskal-Wallis test, the small p-value 0.005024 indicates 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.

Pairwise comparisons using wilcoxon rank sum test, significant level is 0.05 with Bonferroni.



~~~
```{r}
pairwise.wilcox.test(leadRealMAXFT$MAXFT, leadRealMAXFT$lead_typ,
                     p.adjust.method = "bonf")
```
~~~

```
cannot compute exact p-value with ties
        Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  leadRealMAXFT$MAXFT and leadRealMAXFT$lead_typ

       group1 group2
group2 0.004  -
group3 0.562  0.718

P value adjustment method: bonferroni
```

RQ1 - Normal blood-lead levels will have higher average MAXFT scores.
H0: mu1 = mu2
Ha: mu1 > mu2
Ans: The small one-sided p-value 0.004/2 = 0.002 (<0.05) indicates that the mu1 and mu2 are significant differences with 95% confidence interval. We can conclude that there is conclusive evidence that the children with normal blood-lead levels will have higher average MAXFT scores.

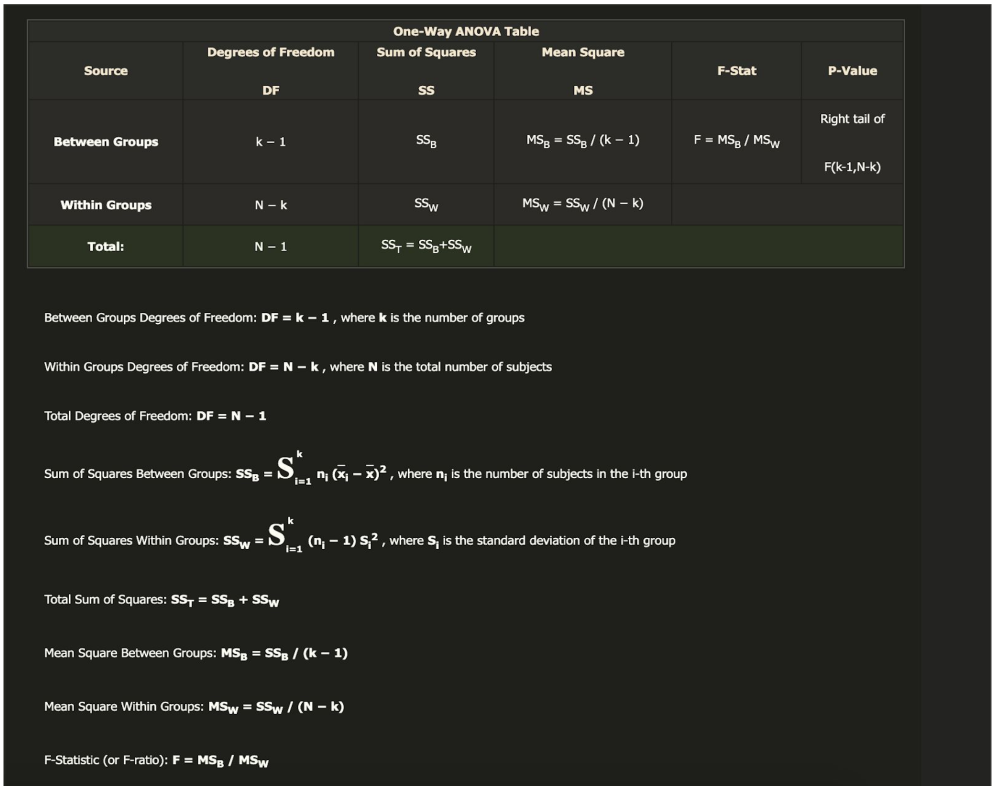
RQ2 - Previously exposed populations will have higher MAXFT scores because they have “recovered” in the following year.
H0: mu2 = mu3
Ha: mu2 ≠ mu3

Ans: We fail to reject the null hypothesis that the two-sided p-value is 0.718 which is greater than significant level (0.05) with 95 % confidence interval. Therefore, we have no evidence to indicate that the previously exposed populations lead_typ3 will have higher MAXFT scores than lead_typ2.

RQ3 - Normal blood-lead levels will have higher average MAXFT scores.
H0: mu1 = mu3
Ha: mu1 > mu3
Ans: We fail to reject the null hypothesis concluding that the one-sided p-value is 0.562/2 = 0.281 which is greater than 0.05, on average, between lead_typ1 and lead_typ3 with 95% confidence interval. We can conclude that there is conclusive evidence that children with normal blood-lead levels will NOT have higher average MAXFT scores.

2. From The Statistical Sleuth, Third Edition, Chapter 5, problem 17. Note that to get the p-value you will need to find the probability from an F-distribution. In R the function pf(x,numdf, denomdf) gives the probability P(X < x) from an F(numdf, denomdf) distribution, where numdf denotes the numerator degrees of freedom (between groups df) and denomdf the denominator degrees of freedom (within groups df) of the F-statistic.

Reference: ANOVA Calculator: One-way analysis of variance calculator. (n.d.). Retrieved from https://goodcalculators.com/one-way-anova-calculator/



One-Way ANOVA Table

Source	Degrees of Freedom	Sum of Squares	Mean Square	F-Stat	P-Value
	DF	SS	MS		
Between Groups	$k - 1$	SS_B	$MS_B = SS_B / (k - 1)$	$F = MS_B / MS_W$	Right tail of F(k-1,N-k)
Within Groups	$N - k$	SS_W	$MS_W = SS_W / (N - k)$		
Total:	$N - 1$	$SS_T = SS_B+SS_W$			

Between Groups Degrees of Freedom: **DF = k − 1** , where **k** is the number of groups

Within Groups Degrees of Freedom: **DF = N − k** , where **N** is the total number of subjects

Total Degrees of Freedom: **DF = N − 1**

Sum of Squares Between Groups: $\mathbf{SS_B} = \mathrm{S}_{i=1}^{k}\, \mathbf{n_i\, (\bar{x}_i - \bar{x})^2}$, where $\mathbf{n_i}$ is the number of subjects in the i-th group

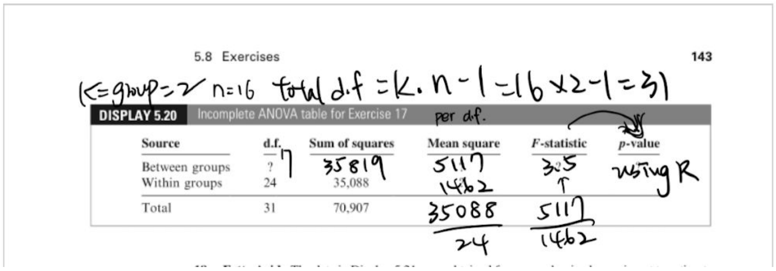
Sum of Squares Within Groups: $\mathbf{SS_W} = \mathrm{S}_{i=1}^{k}\, \mathbf{(n_i - 1)\, s_i^2}$, where $\mathbf{s_i}$ is the standard deviation of the i-th group

Total Sum of Squares: $\mathbf{SS_T = SS_B + SS_W}$

Mean Square Between Groups: $\mathbf{MS_B = SS_B / (k - 1)}$

Mean Square Within Groups: $\mathbf{MS_W = SS_W / (N - k)}$

F-Statistic (or F-ratio): $\mathbf{F = MS_B / MS_W}$



5.8 Exercises 143

K=group=2 n=16 total d.f =K. n-1 =16×2-1=31

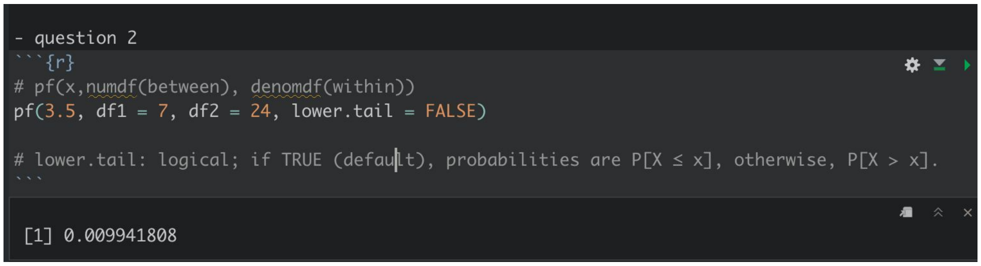
DISPLAY 5.20 Incomplete ANOVA table for Exercise 17 per df.

Source	d.f.	Sum of squares	Mean square	F-statistic	p-value
Between groups	? 7	35819	5117	3.5	using R
Within groups	24	35,088	1462	↑	
Total	31	70,907	$\frac{35088}{24}$	$\frac{5117}{1462}$	

Between + within = total

numerator = between

denominator = within



~~~
- question 2
```{r}
# pf(x,numdf(between), denomdf(within))
pf(3.5, df1 = 7, df2 = 24, lower.tail = FALSE)

# lower.tail: logical; if TRUE (default), probabilities are P[X ≤ x], otherwise, P[X > x].
```

[1] 0.009941808
~~~

1 - 0.9900582 = 0.009941808.

Explanation: There are 8 groups. As the p-value is less than 0.05, we have evidence to conclude that there is at least one pair mean in the groups.

3. In comparing 6 groups a researcher notices that the sample mean for the 6th group, $\bar{y}$, is the largest 6 and that the sample mean for the 3rd group, $\bar{y}$, is the smallest. The researcher then decides to test 3 that μ6 = μ3. Is it appropriate to conduct this test? Or, can any of the multiple comparison methods be used to test this hypothesis? If so, which method? If it is not appropriate, explain why not.

We can only do the pairwise comparison test after the assumptions are met. When we have several groups (more than two groups) to compare, use ANOVA, checking all means are equal (H0) or at least one pair of means is not equal (Ha). If the null hypothesis is not ture, then we use Post-Hoc Tests (such as LSD, Bonferroni, Duncan, ...) for multiple comparisons. The p-value of Post-Hoc Tests's result can be addressing the research questions. If not met the assumptions, we can consider to transform the datasets, use the nonparametric method, and further check if they have influential outliers.